

Proceedings of the International Symposium on Advanced Radio Technologies

March 1–3, 2005

**J. Wayde Allen, General Chair
Jeanne Ratzloff, Publications**



**U.S. DEPARTMENT OF COMMERCE
Carlos M. Gutierrez, Secretary**

Michael D. Gallagher, Assistant Secretary
for Communications and Information

March 2005

DISCLAIMER

Certain commercial equipment, components, and software are identified to adequately present the underlying premises herein. In no case does such identification imply recommendation or endorsement by the National Telecommunications and Information Administration, nor does it imply that the equipment, components, or software identified is the best available for the particular applications or uses.

**INTERNATIONAL SYMPOSIUM
ON
ADVANCED RADIO TECHNOLOGIES**

March 1 - 3, 2005 • Boulder, Colorado

Sponsored by:

Institute for Telecommunication Sciences (ITS)
National Telecommunications and Information Administration (NTIA)

National Institute of Standards and Technology (NIST)

U.S. Department of Commerce, Boulder Laboratories

RF Globalnet

2004 ISART Technical Committee:

J. Wayde Allen, General Chair, NTIA/ITS

Brent L. Bedford, NTIA/ITS

William J. Ingram, NTIA/ITS

John J. Lemmon, NTIA/ITS

Robert J. Matheson, NTIA/ITS

Fredrick Matos, NTIA/OSM

Jeffery A. Wepman, NTIA/ITS

PREFACE

This is the third issue of the Proceedings of the International Symposium on Advanced Radio Technologies (ISART) and the value of these proceedings is starting to become evident. I am aware of at least one case where a paper published in these proceedings has been referenced in other publications. I can only hope that the proceedings will continue to serve as a repository of peer-reviewed literature spanning a large range of topics related to the use and development of emerging radio technology.

The purpose of ISART is to explore the future use of existing and emerging radio technologies. To this end, the conference brings together a diverse collection of people from academia, business, and government to discuss not only the technical aspects of the technology, but also economic pressures and regulatory realities. This year's agenda is no exception. There are presentations on radio spectrum usage, ideas for efficient spectrum sharing, the use of radio for public safety applications, voice over IP (VoIP), wireless networking, location tracking, ultrawideband, RFID, and the generation and use of terahertz signals.

The ultimate goal for ISART is that it become a kind of technological “crystal ball,” giving attendees a glimpse into the use of radio and related electromagnetic technologies in the years to come.

J. Wayde Allen, NTIA/ITS
ISART General Chair

CONTENTS

	Page
<i>Wireless Broadband in the UK</i> Andrew Muir	1
<i>Analysis and Comparison of Spectrum Measurements Performed in Urban and Rural Areas to Determine the Total Amount of Spectrum Usage</i> Allen Petrin, Paul G. Steffes	9
<i>RSMS Measurement and Analysis of LMR Channel Usage</i> J. Randy Hoffman, Robert J. Matheson	13
<i>Technical Challenges to Spectrum Sharing Between Radars and Non-Radar (Communication) Systems</i> Frank Sanders	21
<i>Nation Building and Spectrum Management in Iraq</i> Fredrick Matos	31
<i>Propagation and Throughput Study for 802.16 Broadband Wireless Systems at 5.8 GHz</i> Thomas Schwengler, Niranjan Pendharkar	45
<i>A Full Scale Wireless Ad Hoc Test Bed</i> Timothy X Brown, Sheetal Kumar Doshi, Sushant Jadhav Daniel Henkel, Roshan-George Thekkekkunnel	51
<i>Radio Propagation Measurements During a Building Collapse: Applications for First Responders</i> Christopher L. Holloway, Galen Koepke, Dennis Camell Kate A. Remley, Dylan Williams.	61
<i>Estimating the Demand for Voice Over IP Services</i> Paul Rappaport, Lester D. Taylor, Donald Kridel, James Alleman	65
<i>A High Grade Secure VoIP Using the TEA Encryption Algorithm</i> Ashraf D. Elbayoumy, Simon J. Shepherd	75
<i>Wireless Local Area Network Security: A Framework for Repairing the Broken WEP Protocol</i> Russ Housley, Jesse Walker, Nancy Cam-Winget	81

<i>Time Division Hashing (TDH): A New Scheduling Scheme for Wireless Ad-Hoc Networks</i> Winnie Cheng, I-Ting Angelina Lee, Neha Singh	91
<i>Digital Radio Mondiale Applications and Architecture</i> Edmund Coersmeier, Marc Hoffmann, Martin Kosakowski, Maxim Lobko, Yuhuan Xu	101
<i>A Low Power Methodology for Portable Electronics</i> Dae Woon Kang, James T. Doyle, Mark Hartman, Sandeep Dhar, Marty B. Dermody, Robert C. Woolf, Ravindra S. Ambatipudi, Yong-Bin Kim	109
<i>Sensor Fusion for UWB and WiFi Indoor Positioning Systems</i> Frederic Evennou, Francois Marx, Simon Nacivet	117
<i>Using Standard 802.11 Networks for Location Tracking</i> Arttu Huhtiniemi, Ekahau Oy	125
<i>DSRC Technology and the DSRC Industry Consortium (DIC) Prototype Team</i> John Freund, Randy Roebuck	129
<i>On the Performance of Ultra Wideband Radio in Stochastic Tapped Delay Line Model of the Ultra Wideband Channel</i> Kazi M. Ahmed, Mohammad Upal Mahfuz, Rabindra Ghimire, Mohammad E. R. Khan	137
<i>Performance Evaluation of Coded UWB-IR on Multipath Fading Channels</i> Michal M. Pietrzyk, Jos H. Weber	145
<i>Simulation of Interference Effects from UWB Sources on a Narrowband Digital Transmission System</i> Idnin Pasya, Atsushi Tomiki, Takehiko Kobayashi	151

Tutorials

Tutorial A: <i>Flexible Spectrum Use Rights</i> Robert J. Matheson	157
Tutorial C: <i>Detection and Measurement of Radar Signals</i> Frank Sanders	169

Poster Papers

<i>Adverbs and Adjectives: An Abstraction for Software Defined Radio</i> Troy Weingart, Doug Sicker, Dirk Grunwald, Michael Neufeld	183
<i>High Data Rate SATCOM On-the-Move for an Ambulance</i> Ivan Corretjer, Dave Minerath	193
<i>Measurement of Weak Signals Using a Communications Receiver System</i> Marc Rütchlin, Kate A. Remley, Robert T. Johnk, Dylan F. Williams, Galen Koepke, Chris Holloway, Andy MacFarlane, Mike Worrell	199
<i>Using a PCS Self-Interference Model to Evaluate the Effects of Cell Damage or Failure</i> Timothy J. Riley, Teresa L. Rusyn	205

Wireless Broadband In The UK

Dr Andrew Muir

Regional Manager, Mason Group, Scotland

Tel: +44 1854 622400, Fax: +44 131 443 9944

Email: andrew.muir@mason.biz

This paper provides an overview of the broadband wireless market in the UK covering the competitive position, the regulatory background, and the technology developments that are driving the interest in this area. Building on this overview, a focused case study is described which shows how wireless is providing broadband access to some very remote communities. The author of this paper is the Regional Manager of a major UK independent telecoms consultancy, who also lives and works from a home office in one of the remote communities discussed.

1 The UK Broadband Market

Increasing demand for bandwidth, coupled with growing availability of DSL services and British Telecom (BT) price reductions, has continued to drive the demand for broadband services in the UK. Despite a slow start, BT has accelerated its DSL exchange upgrade programme, which once complete, will allow an estimated 99.6% of UK homes and businesses to be connected to a DSL-enabled exchange. Coupled with announcements in summer 2004 on extended reach of the ADSL product, this has meant that BT's DSL product has consolidated its position as the main wholesale delivery mechanism and continues to account for 65% of the broadband market in the UK¹.

Whilst the delivery of broadband services over Fixed Wireless Access (FWA) networks has not been widespread in the UK to date, demand for wired broadband delivery networks, notably cable and xDSL, has continued to grow. The latest figures published by the telecom regulator, Ofcom, (September 2004)¹ show the total broadband subscriber base in the UK at 5.3 million, representing around 22% of all households. This includes an estimated 1.8m end-users of broadband cable services and 3.5m users of DSL services.

The European industry for wireless services is an interesting market, with both established cell phone groups and newer entrants offering services such as WiFi medium-range, high-speed data connections for laptops. However, the success of FWA networks, as a means of providing the 'last mile' connection to the home has been limited, due mainly to high equipment costs and technical issues such as line of sight limitations. This has limited the success of fixed wireless deployments to date in the UK, with only two

networks currently in operation (Your Communications at 28 GHz and Firstnet, now Pipex Communications, at 3.6 – 4.2 GHz), which are both now focusing on gaining business customers in provincial towns and through limited urban rollout plans. More recently, the huge growth in use of licence-exempt systems, for both WiFi and community broadband schemes, has been a significant factor in raising awareness of wireless solutions in the broadband market. There are a growing number of community wireless networks emerging in the 2.4 GHz and 5 GHz bands, providing broadband access to homes and small businesses in areas where ADSL and cable modem services are not available.

2 UK Spectrum

The 3.4 GHz, 3.6 GHz and 10 GHz frequency bands have been earmarked for FWA services by the UK and other European regulators for a number of years, although the market has been slower to develop than anticipated. To encourage the supply of equipment for operation in these bands, the European Telecommunications Standards Institute (ETSI) has developed various FWA standards for operation across these frequency bands, with standards aimed both at achieving interoperability (defining a standardised air interface for operation over a range of FWA frequency bands) and co-existence (so that different systems can co-exist in a shared band).

Although the intention has been to encourage the supply market, the range of different access technologies and configurations (TDMA, CDMA, FDD/TDD etc.) has meant that this goal has not been achieved. Coupled with the limited number of commercial FWA systems in operation in licensed bands in Europe, this has meant that the cost of equipment has not fallen to levels where economies of scale are being achieved. This is likely to

remain a key challenge in successful deployment of commercial FWA services and will be further challenged by the emergence of FWA equipment for the 5.8 GHz band, where vendors are co-operating on open standards and interoperability. This is likely to drive down equipment costs for those products compared to the higher cost of other proprietary FWA equipment. The licensed bands at 3.4 GHz, 3.6 GHz and 10 GHz were originally envisaged to accommodate narrowband FWA targeting the consumer market, although now there has been a shift away from this towards business data services.

It is likely that the role of these frequency bands in the broadband market will be to support a combination of end-user broadband services plus wireless backhaul, possibly in combination with 28 GHz or other microwave bands. However, this is likely to be more relevant to urban and suburban environments in particular, whilst in remote areas, the cost of deployment is likely to be higher than the revenues gained, which will make the business case unviable. An example of such a development is UK Broadband who, operating in the 3.4 GHz band under licences awarded by auction in June 2003, is delivering broadband to the residential market. They state that they have plans to offer services elsewhere in the UK following a launch in the highly populated Thames Valley in Southern England.

Ofcom has outlined its blueprint for bringing about an open market for the trading of radio frequencies in a move that will see spectrum openly traded on a secondary market². The new regime aims to encourage owners of spectrum to sell their unwanted capacity to interested buyers, releasing value for the owners and potentially spurring the introduction of new technologies and services. It is hoped that prices could fall and innovation grow if the regulator hands over more responsibility to the market for deciding how scarce spectrum should be used. Under the regime, the eventual aim is that more than 70% of the radio spectrum would be liberalised, allowing it to be freely traded. The regulator would retain control of just over 20% of all available spectrum either for security reasons or to avoid interference across international borders. One of the drivers of this reform is the acknowledgment that the central command approach to spectrum management traditionally used by regulators is slower than alternatives such as trading and liberalisation in enabling new applications. In environments where innovation can be applied more quickly, such as the US, new applications such as WiFi, WiMAX and UWB have emerged many years before the UK.

The overall view of Ofcom is that there should not be any regulation specifically aimed at providing an

advantage for broadband fixed wireless over other uses of spectrum. Instead, through trading and change of use, Ofcom aims to allow potential broadband fixed wireless operators access to the widest possible range of frequency bands and technologies. In accordance with Ofcom's general approach licences in future will, as far as practicable and subject to protecting existing users, be awarded on a technology and service neutral basis.

Some wireless technologies may also be used to provide next generation broadband access. These are likely to use higher frequencies, and possibly mesh architectures due to the propagation characteristics of these frequencies.

3 Technology Developments

It is recognised that wireless technologies may play a significant role in the promotion of effective and sustainable competition in the broadband market and encourage the investment necessary for continued rollout and upgrading of infrastructure. As an example, WiFi has become the fastest growing technology in the telecommunications market in the last two years. Several factors have contributed to its success, including low cost equipment (driven down by the global availability of the 2.4 GHz and 5 GHz bands on a licence-exempt basis, open standards and strong manufacturing support), its widespread availability and ease of installation. In addition to being used alongside fixed and wireless data networks, WiFi is also used to provide services that substitute entirely for other delivery platforms. WiFi is now being used in various deployments of broadband community networks, typically to serve remote areas not covered by ADSL or cable. By the use of directional antennas, it is possible to link fixed points over fairly large distances using standard WiFi equipment. An example of this will be discussed later. This growing interest in the use of 802.11 technologies to provide both wireless Internet access in hotspots and community broadband schemes has also been widely recognised in recent UK Government Reports³.

While WiFi technology was originally limited to use in the private domain (i.e. on a non-commercial basis), the UK regulator's decision in 2002 to enable commercial systems to operate in the band has created a range of new 'public' WiFi systems as a means of delivering localised wireless broadband access services. There are now well over 9,000 'hotspot' Internet access points in the UK with well established services in public areas within airports, railway stations, retail locations and hotels.

Use of 2.4 GHz and 5 GHz systems for 'last drop' broadband connectivity could generate further demand

for spectrum in licensed FWA bands, to provide the backhauling of traffic from community broadband schemes, for example. There are several alternative means of backhaul provision, however, typical wireless scenarios envisaged are the use of WiFi in the local area, with either point-to-point or point-to-multipoint links providing the backhaul to the nearest point of interconnect. This could be provided either by FWA at 5.8 GHz, or in licensed bands at 3.4, 3.6, 10 or 28 GHz, depending on availability. Whether an FWA solution is the most appropriate will depend on existing infrastructure available in the area and a consideration of the most feasible and cost effective way to achieve the backhaul.

Recent developments suggest that the wireless industry is increasingly looking to the 5.8 GHz band for future WiFi and wireless metropolitan area (MAN) deployment. Particular interest is focused on 5 GHz Band C, which has been opened for commercial use in the UK, and which will be important for products based on WiMAX IEEE 802.16 and other standards. The benefits of the 5.8 GHz band over 2.4 GHz include the significantly larger bandwidth availability and superior service quality. Since this should provide a cost effective means of providing FWA, due to economies of scale and low licence fees, this is likely to be a key development affecting future demand for other spectrum such as the licensed FWA bands at 3.4, 3.6 and 10 GHz, where equipment costs are higher. An example of the activity at 5.8GHz is BT's trial of 'wireless DSL' services in Cornwall, Wales and Scotland using point-to-multipoint technology operating in the 5.8 GHz band using equipment supplied by Alvarion. This aims to extend reach of DSL type services from an enabled exchange. This solution is currently being rolled out in Northern Ireland where BT have won a Government contract to provide 100% broadband coverage to households and businesses. BT predict that only around 40 locations in the entire area will be satellite based and not served though their solution of traditional ADSL and wireless extension, making them the leading region in the EU.

There are interesting developments in the shift from traditional point-to-multipoint architecture of FWA networks to ad-hoc and mesh networking solutions, which potentially offer a number of benefits to operators in deployment in terms of cost, scalability and improved quality of service. These products are typically designed for operation in lower frequency bands (less than 10 GHz), including the licence-exempt 2.4 and 5 GHz bands. As an example, the UK company LocustWorld produces the LocustWorld MeshAP software⁴ (free downloaded) which dynamically configures multiple wireless access points into a wireless network, or mesh,

running at 2.4 GHz. Each node in a LocustWorld Mesh consists of an access point and a mesh box, which is a PC running the MeshAP software. An example of a device that uses this software is the iMesh, a mesh-networking computer, from US-based DeFacto Wireless.

The first commercial broadband wireless WiMAX network in Britain has been kicked off by Telabria to deliver high-speed wireless broadband services to residential, business and enterprise customers in the south east of England, and provide back-haul for the company's installed base of WiFi hotspots in the region. This backhaul is currently achieved through ADSL but WiMAX will replace these links so that the services will completely cut out any reliance on the copper local loop. The service is being trialled in January 2005, with a commercial launch planned for mid-2005.

Another broadband wireless start-up in the UK is Libera that has recently announced a technology change from building its network in licensed 28 GHz spectrum to focus its efforts on creating a national WiMAX network for businesses. This is reported to be largely down to the reduced costs of standards-based equipment.

The potential for rollout of higher frequency technologies in the access network in the UK, such as 28 GHz systems, is likely to be limited beyond urban environments and limited 'hotspot' areas where there is a concentration of business demand. The high cost of rolling out networks to support broadband delivery into more remote areas of the UK will be a barrier for these higher frequencies. In the UK it is likely that services in this band will, therefore, be targeted at the high end of the broadband market, competing predominantly with the leased line market for high-quality, high bandwidth, symmetric data services to larger businesses plus the SME market. This is predominantly due to: the cost of CPE equipment making deployment economically unviable unless to high-bandwidth users; the difficulty in competing on price in the consumer market against established platforms; and the niche nature of the technology, which makes network deployment complex.

Whilst there have been equipment developments in this band, the price of customer premises equipment remains relatively high, which will always be a key limiting factor for these technology solutions.

Touching on mobile communications as an alternative access medium, the European 3G/UMTS standard provides for maximum theoretical rates of up to 2 Mbps for stationary phones, 344 kbps for a person walking and 144 kbps in a moving vehicle, although realistic rates at launch for UK 3G networks are less than this. The five operators awarded licences in the UK, are slowly rolling

out their networks but in these early years of deployment, availability is not expected to be universal; outside of urban areas it is likely that the networks will, for instance, fall back to GSM/GPRS coverage. A number of analysts cite the potential of 802.16e as a possible rival to the services offered by the 3G operators; a major concern to the five operators who together paid just under \$40 billion for 3G licences to the UK Government through the spectrum auction in 2000. Although WiMAX will need new infrastructure, once deployed, it should offer very high bit rates and the possibility for new entrants to compete either using licensed or unlicensed spectrum. Together with the Ofcom position of awarding licences on a technology and service neutral basis, it will be interesting to see how these markets develop in the UK.

The critical factors in the viability of FWA rollout are the cost of subscriber equipment, and the cost of backhauling traffic to interface at an operator Point of Presence (PoP). The backhaul cost varies on a regional basis and depends on the particular backhaul strategy that an operator follows. However, in general, the rollout of FWA networks in licensed spectrum will only be viable in certain areas of the UK, and there will be differences in markets between regions and nations of the UK. In rural and remote areas, the very low density of subscribers spread over a relatively wide area suggests that FWA services in licensed spectrum will not be viable. Rural demand will, therefore, probably be met by FWA systems operated in licence-exempt spectrum that can be more readily configured to connect geographically dispersed customers more cost effectively. The increasing numbers of community networks being deployed, initially at 2.4 GHz, to provide the 'last drop' connection to consumers in the more remote areas, is already being demonstrated. An example of such is now described.

4 Community Networks – A Case Study

Community networks have been reasonably well established in a number of countries, including Canada and the US, and interest has now spread to other countries, including the UK. These networks cover a range of sophistication, from do-it-yourself efforts sharing a broadband connection between immediate neighbours, to those more commercial services offered by new, alternative, service providers.

In the Scottish Highlands and Islands, access to broadband services is limited in terms of competitive supply, affordability and availability. This is an area that truly is a challenge when it comes to affordable broadband services. With a population density of just

11/km² - equal to the sparsest in Europe - innovative solutions have to be found.

A recent community broadband project has been developed to address the requirement for affordable broadband services in the most remote communities. A project run by the UK Government's Regional Development Agency (RDA) in the area and joint-funded with the European Regional Development Fund has been set up to manage the supply and rollout of broadband services to these communities. The funding covers the capital cost of infrastructure to bring low-cost solutions to a number of the most remote communities. The ongoing costs to sustain the community networks are not covered and, therefore, each community must be able to cover these costs through its own subscription levels. Monthly tariffs must be set to be equivalent to mainstream broadband services, such as ADSL, in order to make them attractive. It is clear, therefore, that the fundamental requirement is for the complete service to be affordable and sustainable in revenue terms through the community subscribers.

The community broadband wireless solution launched in the area is a challenge on a number of fronts. Communities in this region are small; often subscriber numbers within a community will not reach more than 20. The terrain is extremely challenging for traditional types of technology with mountains, lochs, coastline and scattered households. The availability of affordable backhaul is limited. This all adds up to a significant challenge in making any network service sustainable. Keeping capital investment in infrastructure to reasonable levels by avoiding long distance underground cable routes, and maintaining low ongoing costs dictates, to a large extent, the use of wireless networks. From a technology perspective, this can be provided in the most cost-effective manner using unlicensed WiFi equipment.

To keep ongoing costs low, site fees, licences and other revenue costs must be kept to an absolute minimum. Households and other suitable community property must be used to host the equipment, and this means working very closely with the community to ensure buy-in from the start. Community meetings have to be held and local champions used to generate and retain the enthusiasm. Bearing in mind that previously unreliable dial-up modem access at less than 28kbps may have been the norm in these areas, this enthusiasm is not hard to find and it is noticeable how readily broadband is being embraced within these localities.

The backhaul connection to the Internet may be achieved using technologies such as satellite to offer a relatively low cost solution with easy deployment, or

fixed line options where affordable. However, there are performance considerations with satellite that make it unsuitable for certain uses, and this may prove problematic to subscribers in the longer term. Fixed line services, although more expensive, offer improved performance that will be an important consideration when offering a commercial service. To be successful, after all, the service offered to the subscribers must be comparable with other broadband services available from providers elsewhere in the UK. The subscriber must be prepared to pay his monthly rental and poor service will not suffice.

Due to the nature of the end solution, it is important that the communities are involved in the programme from an early stage. Project funding provides the wireless infrastructure required and sets up the ISP services overlaying the infrastructure. Options for continued maintenance and support are made available with the level of ongoing support being dependant on the community revenue generated to pay for it. End users pay an installation cost and a monthly rental for the service which is comparable to that offered by ADSL services available elsewhere. This means that defined service packages are required, with offerings at 512 kbps and 1 Mbps download, plus the usual ISP services of email and web hosting. Costs of these services to the end user are typically set at \$250 installation and \$45/month. All revenues are paid into a community fund that is then used to pay for backhaul provision, a small amount of spares and, possibly, a maintenance contract with either the supplier or a locally trained individual.

The initial phase of the project provided broadband connection to five remote communities each with approximately 20 subscribers. At an early stage, the decision was taken to provide 2.4 GHz FHSS (Frequency Hopping Spread Spectrum) technology, due to its perceived suitability for low-cost, outdoor provision. The communities' that were chosen were typical of the area and included very remote mountainous and coastal locations plus island communities. Within each of the communities identified, a wireless exchange area was defined to cover the user registrations within that community, and to represent the extent of the broadband coverage required.

The process of bringing a community online was developed through a number of steps, including community consultations and public meetings. The principal aims of the community consultation phase were to: maximise the numbers of potential subscribers; ensure that as many people (including non subscribers) knew that a wireless broadband rollout was happening; to make all aware of the general community benefits. In some instances, the network may be entirely dependent

on nodal infrastructure located on third party (i.e. non subscriber) land or property and so a general review of all properties in a community must be made.

One of the pilot communities is Achiltibuie on the north west coast of Scotland. This is a small community situated along a stretch of coastline with households spread out along a coastal road for several kilometres. There is no simple notional design for 'linear coastal' communities such as Achiltibuie. Buildings are not evenly spread in a radial layout from the village centre, and so it is not feasible to connect all the subscribers by deploying just a few high hub sites across the area. Instead, the most practical method for efficient connectivity is to deploy a string of hub sites along the coast, with spurs to connect in subscribers who are located away from the main spread of buildings. This is shown in the following diagram:

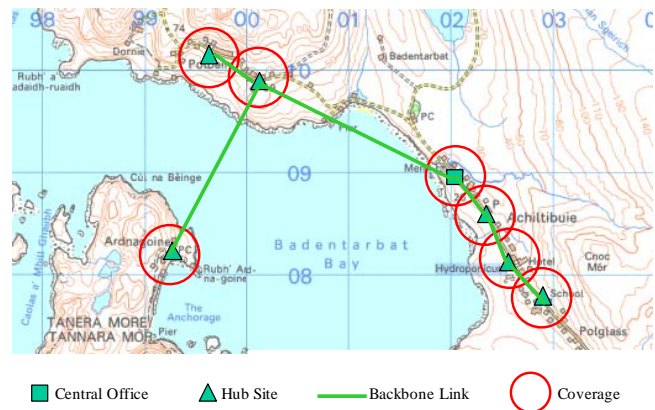


Figure 1 Community Network Design

This installation shows seven coverage sites, serving subscriber locations using point-to-multipoint 2.4 GHz technology and interconnected using the same point-to-multipoint technology. This shows that there are a number of radio hops per subscriber back to the central office location where a 2 Mbps broadband satellite provides the backhaul Internet connectivity.

Based on the technical specifications for the FHSS equipment, and observation of the typical installation environments (antenna types, feeder types and lengths, use of surge protectors etc), a theoretical link budget for the system was created. The link budget calculated the maximum permissible link range against a set of target data rates, in both dry and wet conditions.

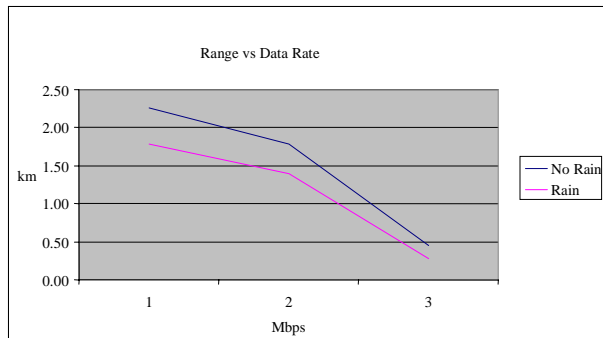


Figure 2 Link Range

Based on this analysis, it was concluded that the maximum practical link length of typical equipment was around 1.5km (with 15dBi antenna at one end, 18dBi antenna at other end).

In the backhaul, broadband satellite provides a low cost solution that is often targeted at rural broadband projects such as these. As a result, broadband satellite is becoming a feasible option for many organisations, either as direct access to subscribers, or as the backhaul element of a terrestrial wired or wireless scheme. One of the main advantages of satellite is its almost universal reach, with service offerings being available virtually anywhere within the UK. Against this there are some performance issues that can affect some applications.

Latency is perhaps the most obvious issue for a satellite system. A signal travelling to a Geostationary Earth Orbit (GEO) satellite at 37,800km will require approximately 260ms to reach the satellite and return to Earth. This gives a round-trip time of at least 520ms. This is a fundamental limitation in receiving satisfactory quality of service for some applications. The latency will have an impact on transport protocols, but for some aspects, mitigations are possible. For others (e.g. the response time from a remote application server) there is little or nothing that can be done.

There are a number of developments underway to mitigate the latency performance issues with satellite and these are centred on Performance Enhancing Proxies (PEPs) and other strategies such as caching. PEPs work by hiding the long-latency satellite hop from the end-systems by ‘spoofing’ TCP so that all recovery to loss and congestion is performed locally, improving the performance. Other types of PEPs work around compression and packet prioritisation. However, use of PEPs restricts the implementation of certain security mechanisms, ultimately meaning that IPsec VPN solutions cannot be used effectively. This is due to the PEP being unable to access the TCP header, which is protected, and so, the traffic cannot benefit from the acceleration features of the PEP and the resultant

performance is unlikely to be tolerated. Alternative security measures can be implemented, but end-end VPN tunnelling, which is a popular VPN deployment, remains an issue over satellite.

Due to the availability and relatively low cost of broadband satellite, this technology is being increasingly adopted as the backhaul mechanism for small community networks. As a result of this increasing take-up, satellite companies continue to look to address the key issues faced by the operators of these networks, namely the performance limitations on applications, and also to reduce the ongoing costs.

Some interesting lessons have been learnt from this community project. From an installation perspective, it is clear that, once a community knows that broadband delivery is potentially imminent (rather than simply a theoretical aspiration), there is a clear surge in interest. There is also a high degree of goodwill amongst communities in general towards the provision of broadband, with even residents who are not actually interested in broadband being willing to help out by allowing their properties to be used to host equipment.

It is evident in the UK that community broadband networks such as these described, only appear where traditional market services do not provide. There are no known examples in the UK of community networks being set up in competition to services offered by telecom operators. As such, community networks face the difficult dilemma of how to provide a quality service to a low numbers of subscribers. The service currently expected is an ‘ADSL-equivalent’ in terms of bandwidth, contention and price. Due to their very nature, potential subscriber numbers in community networks are limited, and there tends to be a real community spirit of developing something themselves, with the associated relaxation that comes with a ‘best effort’ approach. However, once the subscriber starts to pay a monthly rental charge, service quality must be on a par with other broadband services, and this is an area that still needs to be proved in many cases.

It does not seem to be an obviously attractive area in which the traditional telecom players could participate directly, bearing in mind the low subscriber numbers, reliance on community goodwill and the use of unlicensed spectrum as the basis for a network. Perhaps a crucial indicator is the likelihood that, even if a community network operating on the current basis were in place, if a more ‘commercial’ provider introduced a service, such as DSL, it is very likely that the majority of users would migrate, and this would pretty much kill-off any other business plan.

The strengths of community networks are that they provide a service for which an element of demand exists, but for which market forces do not provide. As such they are welcomed by the communities who provide support and encouragement, backed up by high levels of take-up. Community properties are often willingly made available, even by non-subscribers, who offer sites for the benefit of other users. This level of support is quite unique to the community network scheme and would not be available to larger, commercial telecom companies.

The potential weaknesses are largely around the sustainability of the network once the initial funding packages and enthusiasm have ceased. It is clear that the service must be reliable and affordable and a premium on the monthly rental is not likely to be sustained. Whilst some leeway is given in terms of getting the system up and running, once the rental is being paid, the service must perform or users will not continue with payment. Aside from the technology, a structure must be put in place to provide continued support that does not just rely on the continuing goodwill of community activists. This overhead cost must obviously be factored into the business model.

Wireless technology, if properly engineered and installed, should not have difficulties in meeting the community network requirement; however, satellite in the backhaul could cause problems with users. The availability of 5.8 GHz unlicensed wireless should provide greater opportunities for community broadband, and this is perhaps the greatest area of interest in the development of these projects. Additionally, this technology can now also be considered, for example, for the provision of backhaul links to neighbouring communities or to nearby PoP.

It is worth bearing in mind again that community networks have their position in the market where there are no likely opportunities for alternative supply. The launch of ADSL products for smaller exchanges has been targeted in the past at the small town level. With recent tariff reductions, these products now start to have potential in even the smallest of communities. Provided sufficient subscriber numbers can be generated, and ADSL distance limitations can be met, communities with around 30 subscribers now have alternative options through these products. The role of the community network is pushed further out to the very rural and remote areas, where wireless provision is still the most realistic alternative. As these communities become smaller, long-term sustainability becomes even more of a challenge. Pressure will increase around the affordability of backhaul, with affordable terrestrial 2Mbps backhaul remaining a big challenge.

Community networks have often operated as a catalyst to raise interest and attract commercial suppliers. Their role must change as the market develops and as wireless solutions improve over existing DSL services, it is likely that community networks will remain a useful solution.

References

¹ 'The Communications Market 2004, October 2004 Quarterly Update' Ofcom, October 2004

² 'Spectrum Framework Review' Ofcom, November 2004

³ 'Options for accelerating the deployment of terrestrial fixed and portable wireless broadband services by 2005' BSG Wireless Working Group, November 2003.

⁴ www.locustworld.com

Analysis and Comparison of Spectrum Measurements performed in Urban and Rural Areas to Determine the Total Amount of Spectrum Usage

Allen Petrin⁽¹⁾, Paul G. Steffes⁽²⁾

⁽¹⁾ Georgia Institute of Technology School of Electrical and Computer Engineering, 777 Atlantic Dr., Atlanta, GA, USA, 30332-0250; me@allenpetrin.com; 404-509-4501 (phone); 404-894-5935 (fax),

⁽²⁾ Georgia Institute of Technology School of Electrical and Computer Engineering, 777 Atlantic Dr., Atlanta, GA, USA, 30332-0250; ps11@prism.gatech.edu; 404-894-3128 (phone); 404-894-5935 (fax)

This paper will introduce the analysis of two broadband spectrum studies, presenting and comparing the level of spectrum usage. Two broadband spectrum studies were performed in urban Atlanta, GA and rural North Carolina over the frequency range from 400 MHz to 7.2 GHz. Over several months spectrum usage was measured in several azimuthal directions and in different linear polarizations for each of the two spectrum studies. These studies produced a database with five billion spectrum measurements. This database was analyzed with an advanced algorithm to determine the level of active spectrum usage. The detection method employed was designed at the Radio Spectrum Engineering Lab (RSEL) located at Georgia Tech and is optimized to detect signals that are marginally above the receiving system's thermal noise floor. The sophisticated algorithm uses the spectrum study's multidimensional aspect to achieve significantly better performance than conventional threshold detection. The results of this analysis can be used to identify areas of the spectrum that are unused or underused, creating opportunities for the use of frequency agile radio technologies.

1. Introduction

The shift from static spectrum licenses to a more flexible framework offers the possibility of increased utility from this limited resource. The legacy method of assigning spectrum to a user for a band of frequencies in one geographic area, with a specified intended use, has resulted in far from optimal spectrum usage. To quantify this inefficiency, several spectrum studies were performed and analyzed. These studies took place in both urban and rural locations and covered the spectrum from 400 MHz to 7.2 GHz. This study improved on past ones by resolving spectrum usage azimuthally, in polarization, and in time. The often dynamic nature of spectrum usage necessitates the analysis of its usage over time. To provide accurate and substantive information on spectrum usage more than five billion data samples were taken.

In the studies spectrum usage was measured as a function of frequency, time, polarization, azimuth, and location type. The continuous frequency range from 400MHz to 7.2 GHz was measured; this covers emitters from UHF TV, several land-mobile communication systems, radars (both air search and weather), satellites (uplink and downlink), fixed microwave services, and several passive bands. To measure spectrum usage in the time dimension, two time scales were employed. One measured short-term usage of the spectrum; this provided a metric of spectrum usage over a span of a few minutes. With this, the short-term percentage time

usage can be determined. This metric also aids the identification of periodic spectrum users. The other method is used to determine the usage of spectrum over the course of the day. It has been thought likely that spectrum usage is significantly less at 12am than at 12pm. To evaluate this theory, spectra were re-measured six times over a 24 hour period.

All spectra were measured in both linear polarizations (vertical and horizontal). If both linear polarizations have equal magnitude, slant and circular polarization can be inferred from these measurements. Finally, the azimuthal distribution of emitters was measured across the horizon. This facilitates the sorting between multiple transmitters operating on the same frequency. These measurements also determine the effects of one other variable, location type. Two types of locations were defined: urban, and rural. The urban measurements were taken at a site in downtown Atlanta, GA. For the rural measurements, a radio astronomy facility located in North Carolina was chosen. This site known as Pisgah Astronomical Research Institute (PARI) shown in figure 1 is located in the Pisgah National Forest. At this facility the local population density is less than 17 persons per km², and it is over 50 km from any city with a population greater than 50,000.[1]



Figure 1: Pisgah Astronomical Research Institute (PARI) Measurement Site.

2. Spectrum Usage Detection

The analysis of the spectrum studies to find active users resulted in the development of a spectrum occupancy model. An algorithm developed using this model has the ability to detect spectrum users that are marginally above the receiving system's thermal noise floor. The sophisticated algorithm employed to do this uses the study's multidimensional aspect to achieve significantly better performance than a conventional threshold detector.

Displayed in Figure 2 is a comparison of the advanced spectrum usage detection algorithm and threshold detection at a level which is 3 dB above the system noise floor of -133 dBm normalized to isotropic antenna gain. This plot depicts inferred spectrum usage in the urban area from 1.5 to 1.6 GHz; a band that is occupied by INMARSAT downlinks and electronic aids to air navigation. The algorithm developed exhibits a much lower false alarm rate while retaining a probability of detection similar to the threshold method.

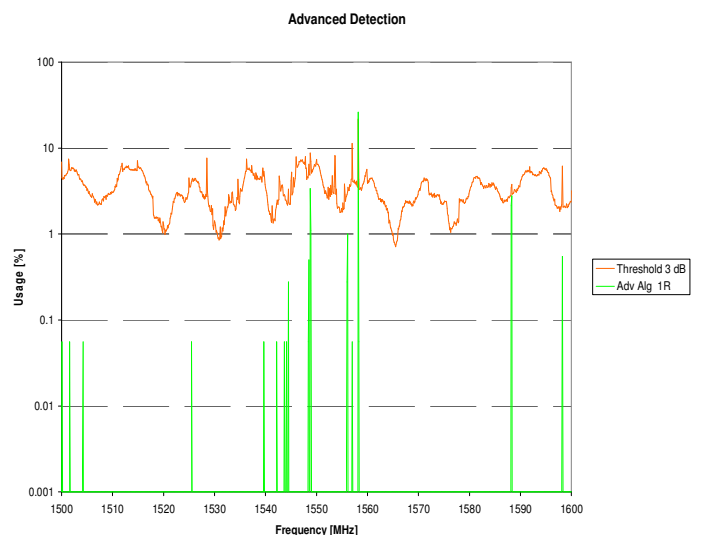


Figure 2: Urban spectrum usage from 1500 to 1600 MHz derived using a threshold detector and the advanced spectrum usage algorithm

The database of spectrum measurements was data mined to identify the characteristics of spectral emitters after propagating through the environment. A detection method was designed with this signature to identify spectral emitters. This advanced detection has the ability to detect several types of emitters including radar, broadcast and intermittent communications.

3. Measurement Results

Since our presentation at ISART 04 [2], we have completed measurement and analysis of both urban and rural spectrum usage. The urban usage levels in the 400 MHz to 7.2 GHz range were found to be much higher than expected. Few areas of unused spectrum were found, although significant portions of the spectrum were measured to be used less than 1% of the time. The rural study showed much lower usage of the spectrum.

For example, as shown in Figure 3, the measured urban usage levels in the 6600-6700 MHz range are quite significant, consistent with interference measurements from spaceborne microwave radiometers operating in that wavelength range [3].

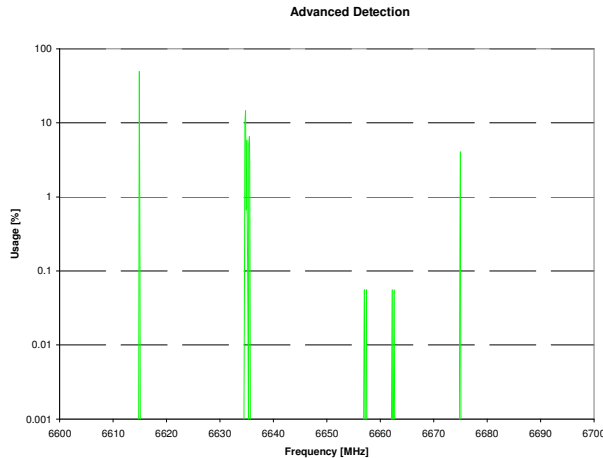


Figure 3: Urban Spectrum Usage, 6600-6700 MHz

By comparison, the usage level measured at the rural location in the 6600-6700 MHz range was below 0.001%, measured relative to the -125 dBm system sensitivity. This is presumably due to a paucity of fixed microwave service links in rural areas. There were however, numerous segments of the rural spectrum which exhibited high occupancy levels, mainly due to satellite downlinks (e.g. 1500-1600 MHz or 2300-2400

MHz) or air search radars (1300-1400 MHz) as shown below in Figure 4.

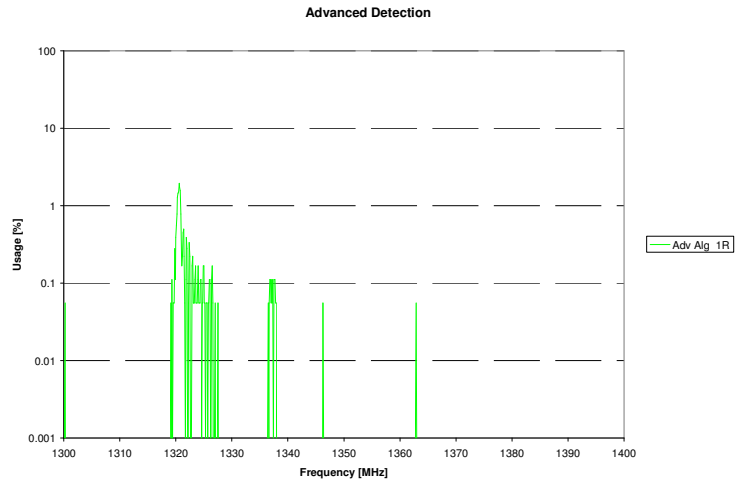


Figure 4: Rural spectrum usage 1300-1400 MHz.

Another aspect of the spectrum study with high relevance to communications applications relates to the concept of “interference temperature,” or the equivalent elevation of the noise floor by broadband interference sources. Measurements of such an elevation has been claimed by some authors [4], but appear to result from intermodulation products generated within the measurement systems, rather than from actual spectrum occupancy. No elevation of the noise floor beyond the thermal noise generated by the spectrum monitoring system was observed.

4. Conclusions

Extensive measurements of spectrum use in the 400 MHz to 7.2 GHz frequency range have been conducted in both urban and rural environments. The data from these measurements have been used in the creation of a spectrum usage model, which has enabled the development of an advanced usage detection algorithm, which would have great value in the implementation of a frequency-agile cognitive radio. While urban usage levels were found to be significant, the potential for sharing is still high, given advanced techniques for spectrum monitoring and transmitter control. An additional spectrum survey is currently being conducted at a suburban location, so as to better characterize the usage levels near a major metropolitan area.

This research has been supported by the National Science Foundation under grant number AST-0309469.

5. References

- [1] U.S. Bureau of the Census “US Census 2000 Summary File 1,” available: <http://factfinder.census.gov/>
- [2] A.J. Petrin and P.G. Steffes, “Measurement and Analysis of Urban Spectrum Usage,” Proceedings of the 2004 International Symposium on Advanced Radio Technologies, *NTIA Special Publication SP-04-409*, 2004, pp. 45-48. Presented at the 2004 International Symposium on Advanced Radio Technologies, Boulder, CO, March 3, 2004.
- [3] L. Li, E. Njoku, E. Im, P. Chang, and K. St. Germain, “A preliminary survey of radio-frequency interference over the U.S. in Aqua AMSR-E data,” *IEEE Trans. on Geoscience and Remote Sensing*, vol. 42, pp. 380-390, Feb. 2004.
- [4] M.-H. Chang and K.-H Lin, “A Comparative Investigation on Urban Radio Noise at Several Specific Measured Areas and Its Applications for Communications,” *IEEE Trans. on Broadcasting*, vol. 50, pp. 233-243, September 2004.

RSMS Measurement and Analysis of LMR Channel Usage

J. Randy Hoffman

rhoffman@its.bldrdoc.gov

303-497-3582, Fax: 303-497-3680

Robert J. Matheson

rmatheson@its.bldrdoc.gov

303-497-3293, Fax: 303-497-3680

National Telecommunications and Information Administration

Institute for Telecommunication Sciences

Abstract: The Radio Spectrum Measurement System (RSMS) is used to make a wide range of radio measurements to help manage the federal portion of the radio spectrum. This paper describes a recent set of land mobile radio (LMR) channel occupancy measurements in the Washington, DC, area. These RSMS measurements were made to provide data in support of several projects related to long-term planning of ways to use federal radio bands more efficiently. The measurements were made using new equipment and techniques that digitized spectrum in 5-MHz swaths and processed it to obtain simultaneous signal levels in 400 individual LMR channels. These techniques provided faster measurements than conventional swept-frequency techniques and also allowed enhanced post-processing of the data to remove measurement defects like intermodulation products and impulsive noise.

1. Introduction

Since 1973, the Institute for Telecommunication Sciences (ITS) has been operating various versions of the Radio Spectrum Measurement System (RSMS) to obtain real-world data on radio signals used by the Federal Government under the direction of the National Telecommunications and Information Administration (NTIA). The RSMS-4 – the most recent of the RSMS series – made measurements of land mobile radio (LMR) channel usage in the Washington, DC, area in the fall of 2004, as shown in Figure 1. This paper reports on those measurements, with specific emphasis on the measurement and analysis techniques used in obtaining and refining this data.

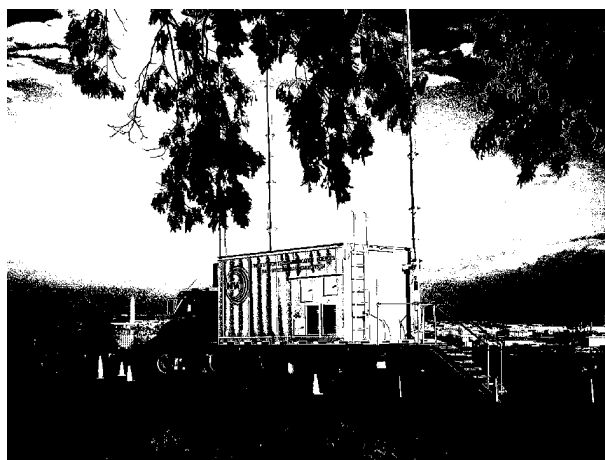


Figure 1 - RSMS-4 measurement site overlooking Washington, DC.

These measurements were made in the 138-174 MHz band, the 225-400 MHz band, and the 406-420 MHz band. Although similar techniques were used in each band, the measurements in the 138-174 MHz band will be used as an example throughout this paper. These measurements were made specifically to obtain data on usage of LMR channels by various federal agencies, for studies on how federal radio operations might be made more efficient in the future [1]. In addition, the measurement results were used to update usage trend information collected beginning in the mid-1970's, showing how usage levels are changing, and providing one basis for predicting future levels of use [2].

Although functionally-similar measurements of LMR channel occupancy have been made by earlier models of the RSMS, the recent RSMS-4 measurements incorporated some new measurement techniques, which seem to offer considerable advantages.

2. LMR Measurement Fundamentals

Federal agencies use the 162-174 MHz band for many LMR functions. Traditionally, most federal agencies used analog FM voice systems with a 25-kHz channel spacing. Recently, the spacing between channels was reduced to 12.5 kHz, though it is believed that many federal agencies have not yet converted their systems to the narrower channel spacing and modulation. Since a typical conversation on these federal channels will last several seconds or more, there is little advantage in measuring each channel more often than once every few seconds

(since the additional measurements will mostly give statistically-redundant information). In most (but not all) of these mobile radio systems, the transmitter is turned on only when the channel is carrying a message. For these signals, the presence of a message can be detected by tuning to the channel and noting whether a transmitted signal is present. Usually, this can be done most easily by simply measuring the amount of RF energy received within the channel bandwidth, and inferring the presence of a signal if the measured power is sufficiently larger than receiver noise. Based mainly on measurement system noise figure, a specific “usage threshold” was calculated, such that any measured channel amplitude higher than the usage threshold was assumed to indicate the presence of a transmitted signal.

The percentage of time that a given LMR channel is in use will be expected to vary drastically, depending on the specific function served by that channel, the time of day and day of week, and the random occurrence of emergencies and other events. Therefore, a minimum set of measurements for LMR channel usage should include measurements on each channel over a period of at least one week (including weekends) on a 24-hr per day basis. Although it is not required that each channel be measured continuously, it would be desirable to include sufficient measurements to characterize every channel on at least an hourly basis. In Washington, DC, the set of measurements made in the 138-174 MHz band included a 4-minute block of data for each channel, repeated every 28 minutes for a period in excess of 7 days. During the 4-minute measurement block, each channel was measured once every second.

Since independent LMR channels are allocated on 12.5 kHz spacings (other bands also use 15 kHz, 25 kHz, or 30 kHz spacings), the ability to separate the signal in one channel from activity in the adjacent channels means that the measurement system must have at least 60- to 80-dB rejection in adjacent channels while simultaneously measuring the signal in a substantial portion of the 12.5 kHz bandwidth of the desired channel. This is achieved with a so-called “rectangular” response bandpass filter, which is flat in response across the desired bandwidth and which falls off sharply outside the desired channel bandwidth.

Various types of modulations are expected in the measured LMR bands. Since some of these modulations may cause substantial variation in the instantaneous received power level measured by the RSMS, it will usually be desirable to incorporate a measurement algorithm that averages out at least some of the possible variation in instantaneous amplitude. The technique used for these measurements involved taking the median value

of a group of five independent closely-spaced amplitude measurements – called “median-of-5” measurements hereafter. In these specific measurements, the groups of five measurements were evenly spaced over about a 750 ms period. This technique averaged out much of the modulation-dependent amplitude variation of signals, and reduced the apparent incidence of broadband impulsive noise (which might otherwise masquerade as real signals), and it reduced the apparent variation in amplitudes produced by system noise. The reduction in the apparent variation of system noise meant that the “usage threshold” mentioned earlier could be set closer to the average system noise level without incurring an unacceptable incidence of false signal indications.

The measurement of LMR signal usage is technically demanding. The detection of very weak LMR signals from distant transmitters requires the use of the lowest possible system noise figure. However, the likely presence of strong signals from close-in transmitters can overload the measurement system electronic circuits and produce intermodulation (IM) products. These IM products appear like real signals to a measurement system. To eliminate the undesired generation of false signals from IM effects, a measurement system that can linearly measure very large signals is required. Large signals can also mix with local oscillator (LO) noise sidebands (what remains after the pure CW signal is subtracted from a local oscillator). These LO noise sidebands create a region of higher system noise for many channels on either side of the high-level signal. The major remedy for effects from LO noise sidebands is to obtain “cleaner” LO signals or to somehow avoid large signals in the receiver.

Because of the numerous detrimental effects of very strong LMR signals, major efforts are made to keep strong signals out of the measurement receivers. Measurement sites are selected on the basis of having at least a minimum geographical separation from known sites with strong transmitters. Measurement systems are normally used with narrowband preselection filters, whose function is to limit the frequency range over which strong signals can enter the measurement system. Known strong signals (e.g., transmitted from a nearby transmitting tower) can be specifically rejected by tunable notch filters or other techniques. Unfortunately, all of these techniques to control unwanted strong signals also tend to increase the measurement system noise figure, so the actual performance of the measurement system is often determined by a difficult set of compromises between noise figure, preselector filtering, and imperfect measurement locations.

3. RSMS-4 LMR Measurement System

The Washington, DC, measurements were made from a site at St. Elizabeth’s Hospital, at the edge of a large hill with line-of-sight coverage to the north and west over most of the District of Columbia, and beyond. Separate omni-directional, vertically-polarized, discone antennas were placed on the tops of three telescoping 35-ft antenna towers to give a dedicated antenna for each of the three measured frequency bands. The three bands were measured simultaneously, using three similar independent measurement systems. Only the details of the 138-174 MHz band measurements will be given here.

The block diagram of the RSMS-4 measurement system is shown in Figure 2. A preselector-preamplifier package included individual low-loss bandpass filters with a 5-6 MHz bandwidth (1-dB-down points) and a low-noise, wide-dynamic-range preamplifier. The 7 bandpass filters provide six contiguous 5-MHz (and one 6-MHz) RF measurement bandwidths to match seven discrete measured frequency ranges which cumulatively cover the 138-174 MHz band. The spectrum analyzer provides high-dynamic-range down-conversion to an IF bandwidth up to 8 MHz wide.

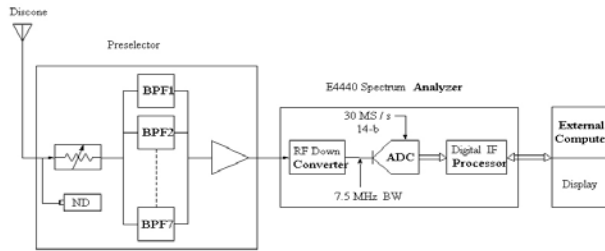


Figure 2 - Block diagram of measurement system.

The spectrum analyzer is used in a so-called “basic” mode, which allows direct access to a number of specialized software analysis routines. The digitizer provides 14 bits of resolution at a 30 MS/s rate, digitizing the entire 8 MHz bandwidth IF signal. A new data block is measured and processed every 1 second to obtain data on 400 measurement channels, each spaced 12.5 kHz away from adjacent channels. Figure 3 shows the effective processed 5.5 kHz “flattop” bandwidth that was chosen as the measurement bandwidth for each channel. This measurement bandwidth provides a receiver bandwidth of 5.5 kHz, while adjacent channel responses are at least 60 dB down. (Note that some of the measured bands utilize slightly different channelization bandwidths, so measurement algorithms were slightly adjusted for those bands.)

The 1-second, median-of-5 measurement routine for 400

channels includes several data-processing steps, as well as several tests to minimize the effects of any measurement problems. Once the spectrum analyzer is tuned to the center frequency of the 5-MHz RF measurement bandwidth, the 14-bit digitizer operates for about 1 ms, producing a record of about 32 K samples. If a strong signal ever overloads the digitizer, an alarm flag causes the 32K samples to be discarded and the digitizing process will repeat. These 32 K samples are processed inside the spectrum analyzer with software that produces a 400-element array, containing a single amplitude reading for the RF power contained within each of the 400 measurement channels.

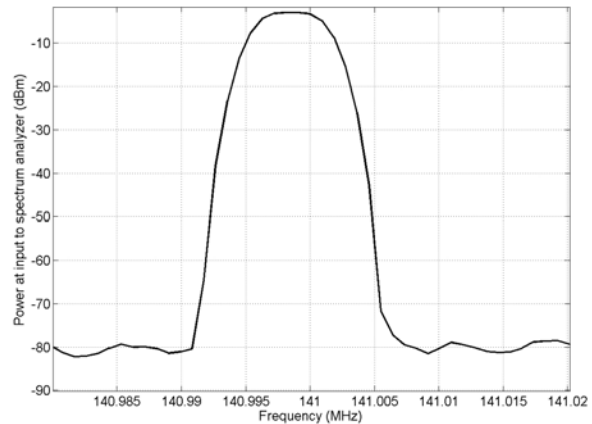


Figure 3 - Software filter bandpass shape.

The 400 readings are displayed on the front panel of the spectrum analyzer and also transferred via an Ethernet LAN to an external controller/computer. The process of digitizing, computing, and transferring 400 readings requires about 175 ms, with a majority of the time used by the data transfer process. The digitizing, computing, and transfer process continues until a total of five measurements for each of the 400 points have been transferred to the external computer. At this point the external computer saves the median value from the 5 measurements for each channel, and the 400 median-of-5 values become part of the data set measured every 1 second.

The 400 median values are also incorporated into a continuously updated (once every second) data display on the external computer. This external display shows the most recent 1-second, median-of-5 data for each measured channel, as well as a peak value for each channel updated over the current 4-minute data block. The operator can also position a cursor over a particular frequency and obtains readouts of these quantities, as well as the exact frequency of the selected channel. The display on the external computer is continually updated with data from the 1-second median values, while the spectrum analyzer

shows a display of the 5-per-s instantaneous values from the measurement process.

At the end of a 4-minute measurement period (240 median-of-5 values for each of 400 channels), the data is saved, the spectrum analyzer is tuned to the next 5-MHz measurement block, and a new 4-minute measurement is begun. Since the 138-174 MHz band was broken into six 5-MHz blocks and one 6-MHz block, a total of 7 blocks were used to measure this band. Every block was measured for 4 minutes out of every 28 minutes. This measurement routine was continued for at least 7 consecutive days, 24 hours/day. At the end of the entire measurement period, the usage on each channel was described by about 86,000 readings throughout the week-long measurement period.

4. Post-Measurement Defect-Reduction Techniques

Although very significant efforts were made to minimize measurement defects caused by IM, LO noise sidebands, broadband impulsive noise, and other problems, we believe that the measurements still contained a substantial amount of defective data. Although it would have been possible to eliminate all of these defects by the use of sufficient RF attenuation to reduce the amplitude of occasional high-amplitude signals, such a measurement configuration would have missed many of the weaker signals. Therefore, we adopted the strategy of choosing a measurement system configuration that would operate near the best theoretical sensitivity (low noise figure), knowing that some defects would be included in the measurements. We took this course of action because we believed that most of those measurement defects could be removed from the data by further post-measurement processing. This section describes identification and removal of the measurement defects.

All of the post-measurement defect-reduction processing is performed on the saved 1-second, 400-channel median data. A key factor in being able to remove these defects from the good data is that the entire set of 1-s, 400-channel data was computed from the same 1-ms digitizer record. That means that the IM products were computed from the same samples as the high-amplitude signals.¹

¹The situation for a traditional scanning spectrum analyzer is quite different, since IM products that showed up near the beginning of a scan might have been caused by strong signals that were gone by the time that their frequencies were measured several seconds later in the scan.

Broadband impulsive noise. Broadband impulsive noise from electrical machinery or automobile ignition systems is usually seen on scanning spectrum analyzers as a sequence of impulses appearing at somewhat regularly spaced frequencies across the analyzer display. Actually, the noise impulses are very broadband, but they show up on whatever frequency the (sweeping) spectrum analyzer was tuned to at the instant when the noise impulse occurred. However, the “simultaneous” nature of the spectrum analyzer used in these LMR measurements means that either all frequencies show the noise, or none of them do. If a noise impulse occurred during the 1 ms digitizing period, the effect will be present at each of the 400 frequencies. If a noise impulse did not occur during the 1 ms digitizing period, it will occur in none of the 400 frequencies. Therefore, impulsive noise can be eliminated merely by noting which 1-sec data blocks are contaminated with noise, and deleting them from further data analysis. The algorithm checked whether at least 300 of the total 400 frequencies were above the normal RMS system noise level. If so, that set of 400 frequencies was judged contaminated by broadband noise and removed from further processing. Figures 4 and 5 show two sequential measurements, one with noise and one without, taken 1 second apart.

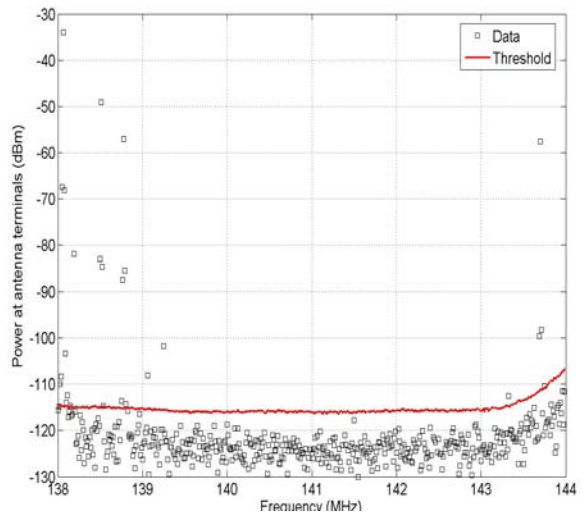


Figure 4 - 1-second sample without impulsive noise.

Note that the median-of-5 measurement mode is also designed to eliminate impulsive noise. Unless impulsive noise shows up on three or more of the five independent samples, it will tend not to affect the median data. Therefore, the only cases of broadband noise that remain in the data occurred when at least three of the five 1-ms sampling periods coincided with noise impulses.

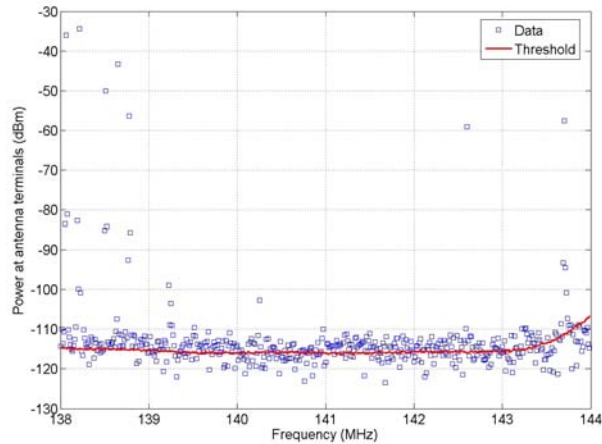


Figure 5 - 1-second sample with impulsive noise.

LO noise sidebands. Noise sidebands on the receiving system local oscillator (or “jitter” on the ADC sampling clock) can cause apparent signals adjacent to the channel occupied by a strong signal. In the case of these measurements, the LO noise sidebands produce additional apparent adjacent-channel signals for any real signal at least 70-80 dB above the system noise level, as shown in Figure 6. These additional responses decrease at the rate of about 0.5 - 1 dB per channel, for channels further away from the strong signal, and eventually these additional responses disappear below the measurement system noise. These additional responses are caused by the mixing (frequency down-conversion) process and cannot be “fixed” by using IF bandpass filters with sharper cutoff shapes.

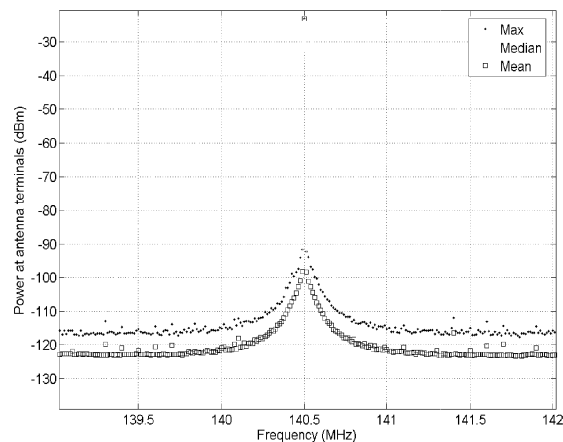


Figure 6 - LO noise sideband responses.

Since the noise sideband responses can affect a large number of additional channels, it is especially important to remove these artifacts. They are eliminated by measuring the levels of a typical noise sideband response and generating a “mask” that shows how much signal

power to subtract from each adjacent channel. This mask is typically quite stable and predictable. However, the mask is probabilistic – having an amplitude distribution similar to Gaussian noise, instead of giving a single value at each frequency. This means that 5-10 dB extra power must be subtracted to reliably discard the LO noise sideband responses. The major problem with subtracting so much additional signal power is that this process may also discard some real (but weak) signals near the strong signal. Figure 7 shows the effect of noise sideband responses on measured signal levels, as well as the effectiveness of the algorithm in removing this signal contamination.

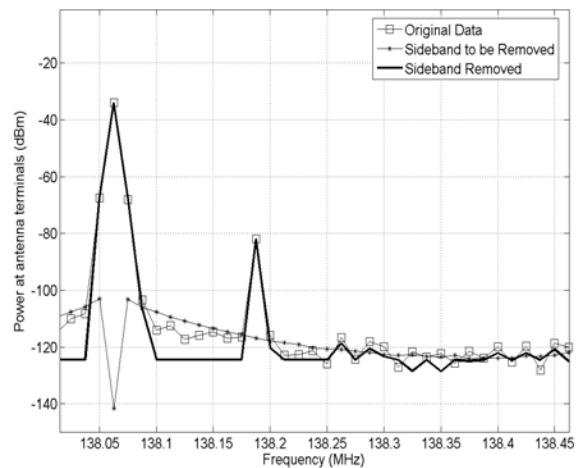


Figure 7 - Removal of LO noise sidebands.

Intermodulation products. Intermodulation (IM) products are probably the most difficult of the measurement defects to reliably remove. Intermodulation products are caused by multiple strong signals. When two strong signals with frequencies F_1 and F_2 are simultaneously processed by the same amplifiers or mixers (active circuits), non-linearities produce addition energy at frequencies $F_{m+n} = +/-mF_1 +/-nF_2$, where m and n are integers. In general, when $m+n$ is an odd number, the frequency F_{m+n} is separated from F_1 or F_2 by exact multiples of the frequency difference between F_1 and F_2 . (When $m+n$ is an even number, the frequency F_{m+n} is usually greatly separated in frequency from F_1 or F_2 and causes no problems.) By proper modeling of the measurement system characteristics, one can calculate what signal combinations would cause IM products, as well as the amplitude and frequency of those products. The predicted IM products can be subtracted from the measured data.

Figure 8 shows an example of the process involved in identifying and deleting IM responses. The figure shows two sets of traces. The upper two traces show strong

signals $s_2 = 138.0625$ MHz and $s_1 = 138.5875$ MHz, as well as other apparently real received signals. However, the IM analysis program shows corresponding predicted IM products, m_2 and m_1 , at 139.1125 MHz and 139.6375 MHz respectively. The lower trace shows other IM products, coming from strong signals outside of the frequency range of this graph. These predicted IM products – m_2 and m_1 – coincide closely with apparently real measured signals. However, when the predicted IM products are subtracted from the “measured” signals producing the heavy trace, the two “measured” signals disappear, showing that “signals” m_1 and m_2 were completely generated by IM in the measurement system. The other predicted IM products were all below the system noise and no significant changes in the measurements resulted from subtracting those IM products.

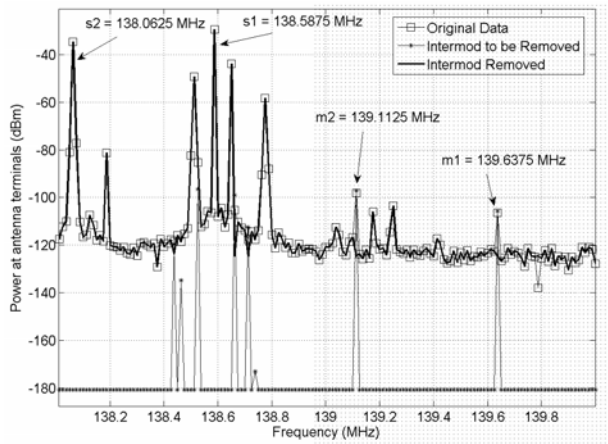


Figure 8 - Example of IM removal process.

As mentioned earlier, the ability to work with a complete set of simultaneous data is crucial for successful removal of IM products. The LMR preamplifier bandpass filters were designed to precisely match the 5-MHz or 6-MHz measurement segments, providing passbands that were narrower than the 8-MHz bandwidth of the final IF in the spectrum analyzer. Thus, IM products could not be caused by strong signals in the wideband RF stages (which were outside the 8-MHz measurement bandpass, and therefore invisible to the IM analysis program). In addition, the simultaneous digitizing of the whole IF bandpass meant that IM products could not be caused at one frequency by strong signals that had disappeared by the time that a scanning spectrum analyzer reached the frequency of the strong signal. The IM prediction process was fairly accurate here, because most LMR signals are FM-modulated (constant-amplitude) and the analysis did not need to account for rapidly-changing signal amplitudes.

Signal thresholds. When all of the measurement data have been processed as described above to purge the data from false signals, the next stage of processing is to determine which frequencies are occupied by signals, and how much of the time the signals are present. We decided to use a simple amplitude threshold; a channel with an amplitude reading larger than “X” is considered to be occupied by a signal. Therefore, we needed to find a value for X, such that measurement system noise did not indicate too many false signals, but was as low as possible to give maximum detection of actual weak signals. Figure 9 shows a cumulative amplitude probability distribution (APD) of Gaussian noise, which is representative of the measurement system noise.

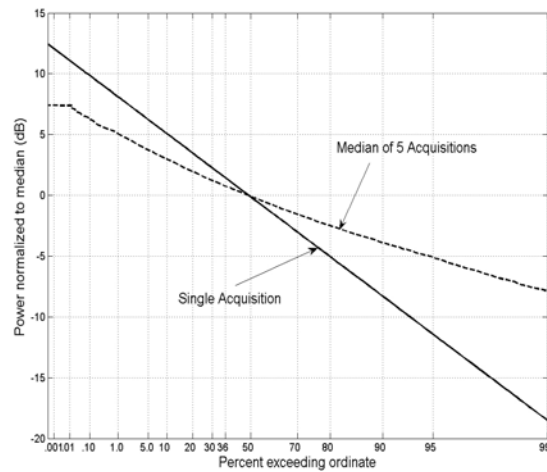


Figure 9 - APD of Gaussian noise with single-sample and median-of-5 sampling.

The solid line in Figure 9 shows the predicted distribution of noise amplitudes, plotted relative to the median power of a channel without any signal (0 dB). This graph shows how often the Gaussian noise in a channel without any signal would exceed X, for various possible values of X. For example, following the solid straight line, we see that Gaussian noise exceeds a level 8 dB above the median noise level about 1% of the time. Selecting a threshold 10 dB above median system noise would give a false signal indication about 0.1% of the time. However, our LMR data is not a set of single samples, but rather a set of median-of-5 samples. The dashed line shows an APD for Gaussian noise as measured with a median-of-5 process. To achieve a 0.1% false signal indication for a median-of-5 process would require that the threshold be set about 6.5 dB above the median power. Thus, the use of the median-of-5 process means that a signal threshold can be set about 3.5 dB closer to system noise (i.e., 3.5 dB weaker signals can be analyzed).

5. Summary

This paper describes a process for measuring the usage of LMR channels and the post-measurement processing of the data to remove certain measurement defects caused by limitations on the performance of measurement systems. Intelligent trade-offs can be made between measurement configurations that detect the most signals (but cause more defects in the data) versus measurement configurations that miss many weaker signals (but cause fewer measurement defects). The best compromise seems to be a measurement configuration that detects the largest possible number of signals, but aggressively removes most of the associated measurement defects in post-measurement processing.

Eliminating impulsive noise considerably decreases the number of channels with signals. In addition, the “wide pedestals” from LO noise on either side of strong signals are eliminated. Usage on multiple channels on either side of strong signals created by LO noise sidebands will not be counted as the presence of signals close to the strong signals, nor will they be interpreted as unwanted modulation sidebands on the respective strong signals. Finally, individual signals identified as IM products during post-measurement processing have been eliminated from the data. Eliminating them prevents later concerns

about whether there was unauthorized use of these frequencies.

This approach to LMR measurement and analysis makes the data set more complete and more trustworthy. Although the removed defects might not have contributed to major statistical differences in the total amount of signal occupancy in these LMR bands, these defects tended to make the occupancy data much more likely to be wrong at specific frequencies. Since all of these defects caused the apparent presence of a signal when no signal was actually present, the removal of these defects made the whole data set much more trustworthy.

References

1. G. Patrick, C. Hoffman, and R. Matheson, “Signal capacity modeling for shared radio system planning,” in “Proceedings of the International Symposium on Advanced Radio Technologies: March 2-4, 2004,” J.W. Allen, T.X Brown, D.C. Sicker, and J. Ratzloff (Eds.), NTIA Special Publication SP-04-409, Mar. 2004, pp 77-86.
2. F. H. Sanders, G. R. Hand, and V. S. Lawrence, “Land Mobile Radio Channel Usage Measurements at the 1996 Summer Olympic Games,” NTIA Report 98-357. Sep. 1998.

Technical Challenges to Spectrum Sharing Between Radars and Non-Radar (Communication) Systems

Frank Sanders
Institute for Telecommunication Sciences

303.497.5727; fax 303.497.3680
fsanders@its.blrdoc.gov

Abstract. *To partially satisfy a voracious worldwide appetite for additional spectrum allocations for data and voice communication systems, proposals have been put forth for such systems to share spectrum with radars by operating in bands that have previously been allocated on a primary or co-primary basis for radars alone. Technical justifications for sharing proposals typically include the following claims: Radar systems make little use of existing spectrum allocations, and so there is much radar spectrum available for other uses; radar receiver performance is inherently robust against interference from signals of other services, and therefore radar receivers can operate co-channel with, or at least in the same band as, communication signals; to the extent that interference to radars might occur from non-radar services, it can be limited in principle on some acceptable statistical basis; and finally, if interference to radar receivers due to spectrum sharing in fact is found to be intolerable, it is possible in principle to design and deploy communication systems that will mitigate interference by detecting locally utilized radar frequencies and avoiding operations on those frequencies. All of the above statements contain either technical flaws or implementation challenges that need to be understood by decision makers who must grapple with the radar spectrum sharing issues. This paper discusses the technical challenges of allowing non-radar communication systems to operate in radar spectrum.*

1. Introduction

At both the national level within the United States and the international level in such organizations as the International Telecommunications Union Radiocommunication Sector (ITU-R), substantial demands for increased spectrum allocations for data and voice communications systems have been made in recent years. As a result of this pressure, a variety of proposals have been put forth for such systems to share spectrum with radiodetermination systems (radars) by operating in bands that have previously been allocated on a primary or co-primary basis for radars alone.

Radar band allocations have typically excluded most communication and data systems because of perceived, fundamental electromagnetic incompatibility between radars and communication systems. But in recent years arguments have been put forth in a variety of forums to the effect that it should be technically feasible to operate communication systems in the same bands as radars, and sometimes even on the same fundamental frequencies as local radar signals. The technical justifications for these sharing proposals typically include the following claims: 1) Radar systems make little use of existing spectrum allocations, so therefore some radar spectrum should be available for other uses; 2) radar receiver performance is inherently robust against interference from signals of other services, and therefore radar receivers can operate co-channel with,

or at least in the same band as, communication signals; 3) to the extent that interference to radars might occur from non-radar services, it in principle can be allowed on a non-zero, statistical basis; and 4) if interference to radar receivers due to spectrum sharing in fact is found to be intolerable, it is still possible in principle to deploy communication systems in radar bands, but such systems will need to mitigate interference by detecting locally utilized radar frequencies and then avoiding operations on those frequencies.

These claims, if substantiated, could revolutionize this aspect of spectrum engineering. But if such claims are flawed, then egregious harm potentially could occur if radar bands were opened to use by communication systems without adequate technical consideration. It is therefore imperative that the assertions behind radar spectrum sharing proposals be carefully considered. This paper provides technical examination of the claims listed above; it illuminates the technical challenges of allowing non-radar communication systems to share spectrum with radars. It demonstrates that implementation of sharing in radar bands poses major technical challenges, and that these challenges need to be better understood and addressed by spectrum engineers.

2. Radar Utilization of Spectrum Allocations

Radar systems usually include high power, pulsed transmitters paired with highly sensitive receivers. Table 1 lists some typical radar transmitter parameters for several important radar missions. The effective isotropic radiated power (EIRP) of many radars exceeds 1 GW, while receiver noise figures are often only a few decibels above the theoretical thermal noise limit, and are sometimes even below the limit for certain types of target processing. Given operational requirements for ubiquitous and continuous surveillance of airspace, surface waters, weather, and other missions, it would seem that there would be no question as to whether allocated radar spectrum is heavily utilized. Nevertheless, the author's correspondence with some national and international authorities, as well as presentations made in some forums in recent years, indicate that some spectrum engineers are not convinced that radar emissions produce significant spectrum occupancy.

Table 1. Typical radar emission parameters as a function of mission*

Mission	Pulse width (us)	Pulse rate (Hz)	Peak power (MW)	Antenna gain (dBi)	Peak EIRP (GW)
Short range air search	1	1000	0.8	33	1.6
Long range air search	3-10	300	1	33	2
Maritime navigation	0.08-0.8	10000	0.02	30	0.02
Weather	1-5	300-1300	0.75	45	24

*These values are only representative of generic radar groups; individual radar systems will be expected to have different parameter values.

Such beliefs have hinged on apparent failures to observe radar signals with a variety of test and measurement equipment during attempted spectrum surveys. Indeed, most commercially available radio measurement equipment is not designed to observe radar emissions, and will fail because of the pulsed nature of the emissions (with a typical 0.1% duty cycle) combined with radar beam-scanning (typically 20 ms out of a 5-sec scan interval, or 0.004% duty cycle). A typical radar's signal therefore only will illuminate a measurement station at any given location for about 0.0004% of the time. Conventional digital sampling and analog swept-frequency techniques used for spectrum surveys (often combined with averaging-type detection)

rarely will intercept these low duty-cycle signals. Thus, failure to observe radar signals under such circumstances likely is due to shortcomings in the survey techniques. When effective survey approaches are used in which peak-hold detection is combined with stepped-frequency algorithms [1], then radar emissions are seen to substantially occupy allocated bands, as shown for example in Figures 1 and 2, and demonstrated further in a series of NTIA spectrum survey reports for four major American cities [1-4].

A conclusion that can be drawn from survey data in these reports is that radars generally do produce substantial, wideband occupancy of allocated bands. Conversely, failures to observe radar during attempted surveys are often more likely due to the use of inadequate measurement approaches rather than a lack of radar signals at the survey locations.

3. Radar Receiver Performance in the Presence of Interfering Radio Signals

When observing radar signals, it is natural to think that the signals needing protection are the very visible high-power impulses generated by radar transmitters, typically present about 0.1% of the time. However, in reality, the radar signals that need protection against interference are the very-low-power echoes reflected from targets (e.g., aircraft and ships). These reflected signals occur throughout the 99.9% remainder of the period when radar pulses are not being transmitted, and they are usually invisible to even the most sensitive monitoring systems. They can be seen by very sensitive radar receivers only because of the 30-40 dB gain employed by radar receiving antennas. This same 30-40 dB radar receiving antenna gain also, unfortunately, works just as well when the radar is receiving very low levels of interference from distant radio transmitters, often at distances beyond the normal operating ranges of most of the interferers.

Popular conceptions of radar systems have fostered the notion that most radar receivers are highly resistant to the effects of co-channel interference from other radio sources. It is sometimes assumed that many radar receivers are routinely equipped with anti-jam features that make them highly resistant to the effects of interference. In actuality, many radar receivers lack anti-jam features and even those that incorporate anti-jam capabilities are typically ineffective against jamming from high duty cycle (communication-type) signals; often the only effective expedient to mitigate such interference is for the victim radar to change its operating frequency (if possible).

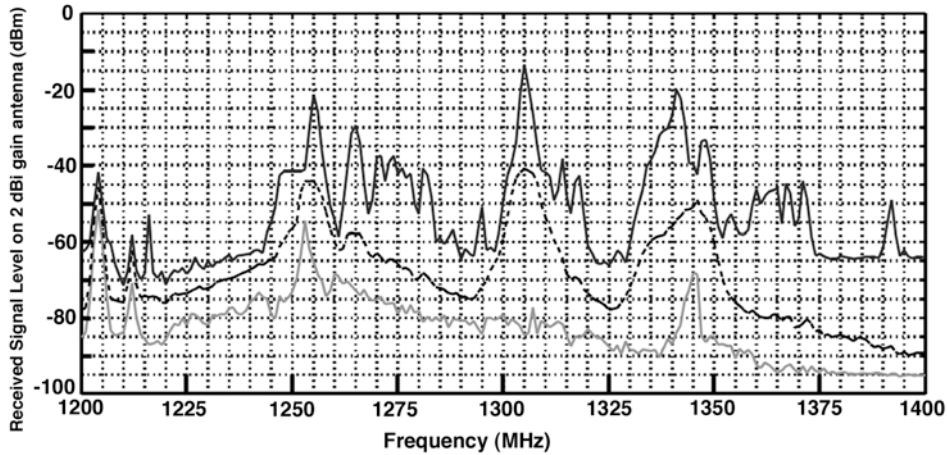


Figure 1. Spectrum survey in radar L-band at San Diego, CA using stepped-frequency algorithm and peak detection. The highest curve represents the maximum power measured in a two-week period on any given frequency, caused by at least eight radars. The two lower curves represent the average (decibel) and minimum power level measured on each frequency in the same two-week period. Detailed descriptions of the stepped-frequency algorithm are in NTIA spectrum survey reports [1-4].

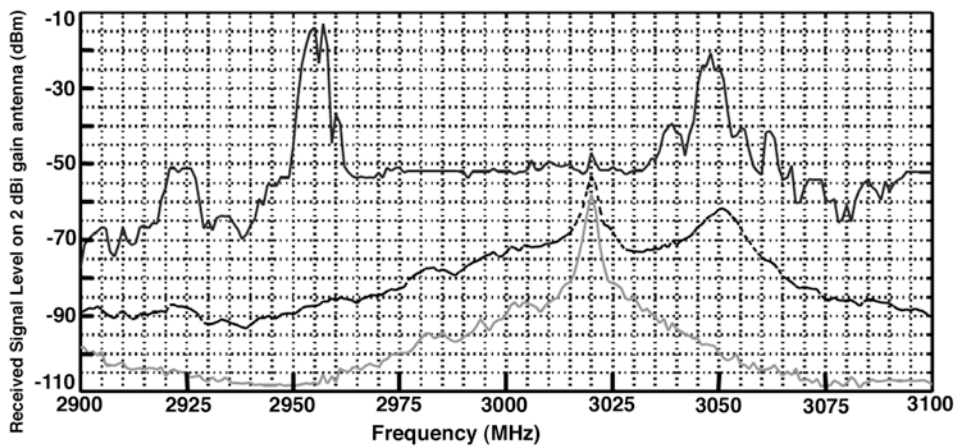


Figure 2. Spectrum survey in radar S-band at San Diego, CA. The flat line and the peaks on the maximum curve represents high occupancy by radar systems; peaks at 3025 MHz and 3050 MHz are generated by maritime navigation radars.

Unfortunately, existing technical literature lacks quantitative engineering studies investigating the effects of low-level interference in radar receivers. To address this gap, NTIA engineers have pursued a multi-year effort to assess interference effects in a wide variety of radar receivers. A summary of some of the results of this study is presented here.

To date NTIA engineers have measured the interference effects of the following types of modulated signals in selected radar receivers: pulsed CW, BPSK, QPSK, CDMA, and carrier-wave. Types of radars that have

been tested against these types of interference include short-range ground-based airport surveillance, long-range ground-based air surveillance (3 models), maritime surface search and navigation (several models), and a ground-based weather surveillance radar.

The protocols for the radar interference measurements have been as follows: The radar under test is placed in a normal operational mode with sensitivity time control (STC) and automatic gain control (AGC) features disabled in the receiver. Although STC and AGC

suppress close-in clutter effects and overly strong target returns in operational radars, tests using artificially generated targets and lacking clutter returns do not need these features for proper radar performance. Furthermore, disabling these functions allows interference levels to be controlled more easily in the radar IF sections without changing radar target response to weak targets and targets beyond clutter range. Performance of radars against strong targets and targets within clutter was not tested

The radar transmitter operates normally but the receiver side of the radar system is disconnected from the antenna at the RF circulator (or T/R switch) and a set of synthetically generated target returns are injected into the receiver at that point (at the radar's RF frequency). The synthesized target levels are adjusted until they are observed on the radar output screen at a probability of about 0.9 or a little greater. That is, roughly 90% of the generated targets are observed over long time intervals. This probability of detection (P_d) strikes a balance between making the targets strong enough to be easily observable while still being susceptible to interference effects. (The target P_d in the NTIA tests is substantially higher than the 0.8 P_d that is sometimes listed in radar manuals and engineering criteria for minimum target visibility.) Figure 3 shows a typical radar plan position indicator (PPI) display when such targets are generated.

With the target levels set appropriately, interference is then injected into the radar receiver at the same point as the targets (that is, on the radar RF frequency at the circulator or T/R switch.) The average (RMS) interference level in the radar receiver is calibrated by observing the radar IF with a spectrum analyzer. When the injected interference is set at an amplitude that causes a 3 dB rise relative to the radar's own IF noise (that is, the interference-plus-noise level is 3 dB higher than the inherent IF noise), then the ratio of interference-to-noise (I/N) is 0 dB. Other interference amplitudes are set relative to this level for the remainder of the tests on the radar.

With the I/N levels in the radar IF calibrated, the measurements proceed by counting the number of targets that are lost on the radar output display as a function of the input interference amplitude. Typically, over 200 desired targets are injected for each interference level and each interference modulation. Both human-operator counting and automatic radar target counting have been used in the tests, with consistent results. At the end of each set of tests on each radar, interference effects curves are produced. Examples of such curves are shown in Figures 4 and 5; the results of the interference tests are summarized in Table 2.

In summary, radar receivers tested by NTIA have been found to generally lose targets at I/N levels of about -10 dB in the receiver IF sections. That is, when interference signals are coupled into a radar receiver at an RMS level 10 dB *below* that of the receiver's inherent noise, target losses can be expected to begin. The only exception is pulsed interference, which can usually be sustained at tens of decibels (RMS average) above the radar receiver IF noise without causing target losses. Radar receivers are highly robust against pulsed interference because interfering pulse sequences are not ordinarily coherent with the radar's own pulse sequence. This is, essentially, why radars can coexist with each other within spectrum bands.

Table 2. Interference RMS threshold (dB relative to IF noise) causing measurable target loss in radar receivers.

Radar type	CW interference	QPSK interference	Pulsed interference
Airport air surveillance	-10 dB	-10 dB	+30-40 dB
Long-range air route surveillance	-10 dB	-10 dB	+30-40 dB
Maritime navigation	-10 dB	-10 dB	+30-40 dB
Weather	-12 dB	-12 dB	+30-40 dB

Effects of low-level interference in NTIA tests have been found to be insidious because they are manifested only by target losses; no other overt indications are observed. Only interference at high average levels (at amplitudes around +10 dB I/N) seems to cause overt indications such as strobe lines on the radar displays. This means radar operators usually cannot know whether low-level interference is occurring; it also means that lack of interference reports from radar stations does not necessarily mean that interference is not occurring.

As noted above, anti-jam capabilities in radar receivers tend to lack effectiveness against high duty cycle interference such as CW, BPSK, QPSK, CDMA, etc. This vulnerability is one reason that sophisticated (and costly) military radar designs often incorporate frequency diversity capability in their designs: so that they can change frequencies in jamming environments. But we are not aware that frequency diversity capabilities, when available, have ever been intended to accommodate increased amounts of intentional radio interference from other services. Rather, frequency diversity is a feature that is provided to meet tactical military needs in hostile operational zones.

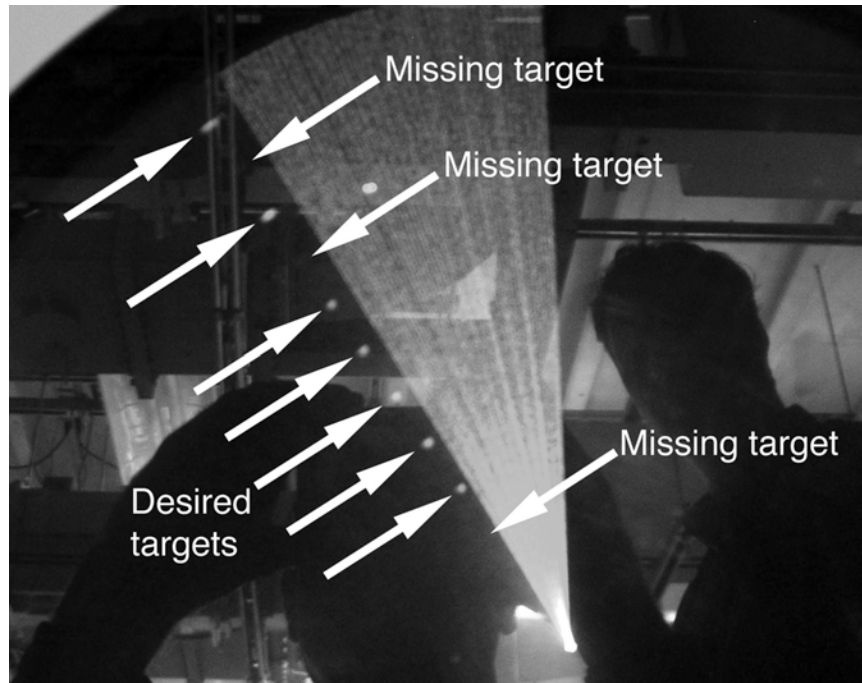


Figure 3. Radar display during interference testing (at an I/N of about -6 dB). On this radar scan, 3 targets are missing out of 10 that were injected. This exemplifies the insidious manifestation of low level interference in radars: targets disappear but there are no overt indications of interference for an operator to observe. (The bright wedge in this picture is the radar beam sweeping during the camera exposure.) Note that targets can disappear at any range, not just at the edge of the radar coverage.

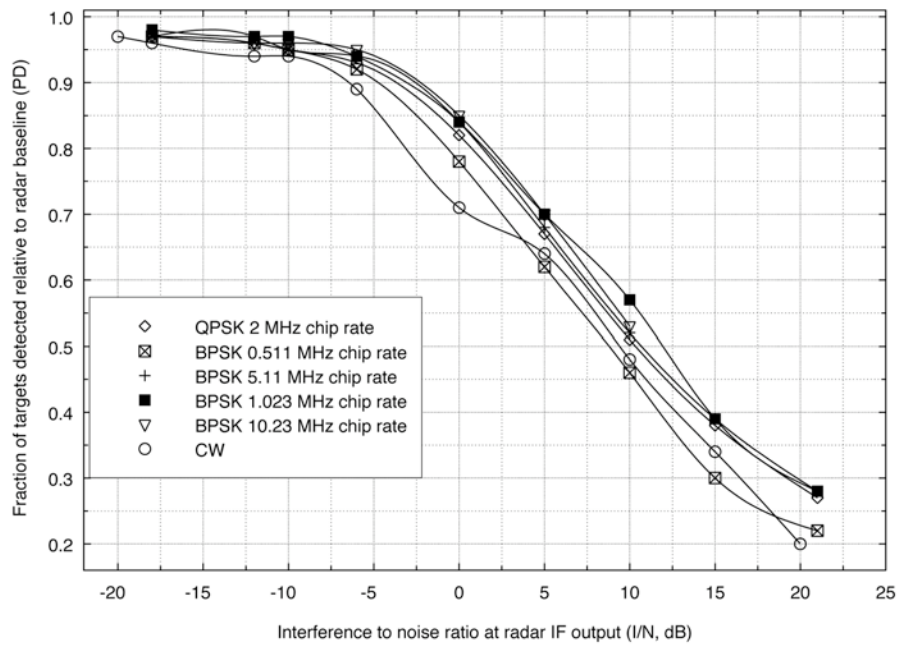


Figure 4. Interference curves for an air search radar. Target losses begin when average I/N ratio of interference exceeds about -10 dB in the receiver IF.

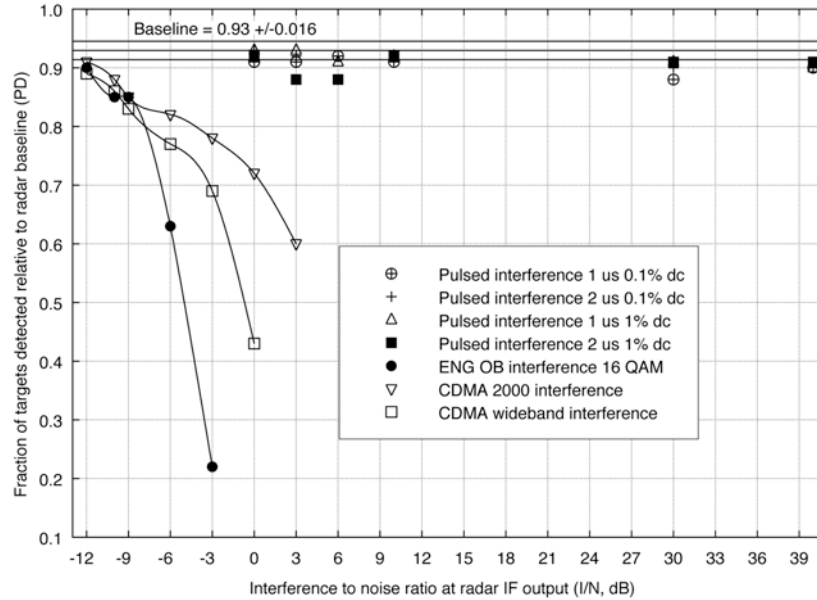


Figure 5. Interference curves for a maritime navigation radar. Target losses begin when average I/N ratio exceeds about -12 dB in the receiver IF. Note that the radar does not experience target losses at I/N ratios of as much as 40 dB when interference is pulsed.

In summary, radar receivers tested by NTIA have been found to be vulnerable to interference by communication-type signals at I/N levels in radar IF sections on the order of -10 dB; radar receivers generally are not robust against such interference. The only successful mitigation technique for radars experiencing such interference is to change operating frequencies. Worse, low-level interference effects are insidious because they only cause target losses and no other visible manifestations occur that might warn of the occurrence of interference.

4. Consideration of Radar Target Losses as a Function of Range

Target loss due to interference at levels of -10 dB or higher in radar receiver IF sections can be regarded as equivalent to loss of radar sensitivity in some relatable number of decibels. For example, some I/N level of interference will translate into an equivalent loss of 1dB in radar receiver sensitivity.

A persistent concept regarding radar target losses due to effective loss of radar receiver sensitivity is that this equates in turn to a range reduction for the radar. For example, it is commonly stated that 1 dB of sensitivity degradation equates to a 6% loss in radar range (since radar range varies as the fourth root of received power) and 11% of radar coverage area. This result begs the question: Why any given radar cannot afford to lose

targets in the outermost 6% (or some other non-zero percentage) of its coverage area?

One answer would be that it is unclear why any radar covering a full range of X km for given targets should be expected to lose some percentage of range capability for no offsetting advantage in national security or safety. But, in fact, the assumption that only range is reduced by interference is an erroneous simplification. This is because any observable target that was within X dB of not being seen in the absence of interference will disappear *at any range* if interference degrades radar receiver performance by X dB or more. The only factor that determines target observability is the echo level relative to the radar's inherent IF noise; to the extent that weaker echoes occur mainly for the most distant targets, then X dB of performance loss will affect distant targets more than closer targets. But since target cross sections vary, it is not true that reductions in radar performance can *only* affect targets at the edge of the coverage areas, because some weak echoes can and do occur at short ranges. This result has been documented during NTIA interference tests, in which desired targets have been vulnerable to loss so long as they are within the X dB detection margin, regardless of their range on the radar output display. This is shown, for example, by the varying ranges of missing targets in Figure 3.

5. Developing a Statistical Basis for Interference into Radar Receivers

Despite the susceptibility of radar receivers to low-level interference from communication-type signals, some spectrum sharing proposals have been put forward by other Administrations in the ITU-R to legally permit interference into radar receivers for specified percentages of time and at specified levels so that sharing can be allowed. This is the so-called statistical approach to spectrum sharing.

The argument behind the statistical approach is as follows: Radar receivers are known to lose some targets due to unavoidable environmental factors such as, for example, solar noise (which produces strobe lines on radar displays under certain circumstances). Since some targets are sometimes unavoidably lost due to naturally occurring noise, then radar operations should be expected to allow additional targets to be lost due to man-made interference. Such losses would be permissible if the percentages of lost targets were kept sufficiently low.

This approach raises two fundamental questions: First, how many targets can acceptably be lost under any given circumstances? Second, how can communication systems coordinate activities with radar operations to assess the rate at which targets are being lost at any given moment and how can the causes of target losses at any given moment be determined (i.e., how can natural losses be distinguished from man-made interference losses)?

In answer to the first question, no criteria have yet been identified by any Administration that would allow any authority to say that certain radar targets can afford to be lost under any circumstances. There is currently no way to know *a priori* whether any given target is sufficiently unimportant to not need to be observed and tracked.

Further, as noted above, weak targets are vulnerable to loss at any range due to interference, not just targets at the edge of radar coverage areas. Small-cross-section targets at critically short ranges from radar receivers can be vulnerable to loss due to even low levels of interference. Disastrous consequences could include collisions, failures to locate people in search-and-rescue missions, or failures to respond adequately to close-in military threats.

It is unclear why the loss of some targets due to unavoidable natural causes such as solar noise should be used as a justification to lose additional targets due to intentional man-made interference. But even if this

logic were justified, the problem remains that such interference somehow would have to be coordinated with radar operations. Such coordination would have to be performed in one of two ways. One method would be for all communication transmitters to know where all vulnerable radar receivers are located on a real-time basis, and adjust their operations accordingly. The other way would be to equip radar receivers with sensors that would identify, in real time, the amounts and sources of local interference and then communicate with those sources to mitigate the interference as necessary. Neither of these approaches are likely to be technically feasible with existing technology and resources in the near future. Further, the lack of overt signs of low-level interference in radar receivers (as discussed above) means it would be difficult to verify the performance of radars under such a regime. But without a verification capability, it would be impossible to know how well the sharing system was performing, and thus it would be difficult to accept such a regime as reliable and safe in its implementation.

6. Mitigation of Interference by Sensing Radar Signals in Communication Systems and then Avoiding Radar Frequencies

Given the sensitivity of radar receivers to low-level interference from communication signals, and further considering the difficulties inherent in a statistical approach to limiting such interference under spectrum sharing, proposals have been made to design and deploy communication systems that will mitigate interference by detecting locally utilized radar frequencies and avoiding operations on those frequencies. This approach, called dynamic frequency selection (DFS), is currently being tested in the 5 GHz part of the spectrum in the US and Europe.

No fundamental reasons why DFS cannot work have been demonstrated, but technical challenges of implementation are substantial. One major challenge is the need to somehow distinguish radar pulses from communication traffic, which itself may be packetized into pulses that may have lengths similar to those of radars in the band.

One possible solution is for DFS communication systems to turn off transmissions for substantial fractions of time and listen for radar pulses during those intervals. Figure 6 shows an algorithm that accomplishes this goal. However, implementation of the algorithm with a nearly 100% chance of observing radar emissions in typical radar bands on a single radar-scan basis would require that communication systems turn themselves off for monitoring about every 20 ms (since that is the duration of a typical radar beam

sweeping past any given location in space), and would need to monitor for about 4-8 ms during every such 20 ms interval (since it will take at least a few milliseconds for several radar pulses to occur within the radar beam). This would represent a downtime of 20-40% out of every 20 ms for communication systems implementing this approach.

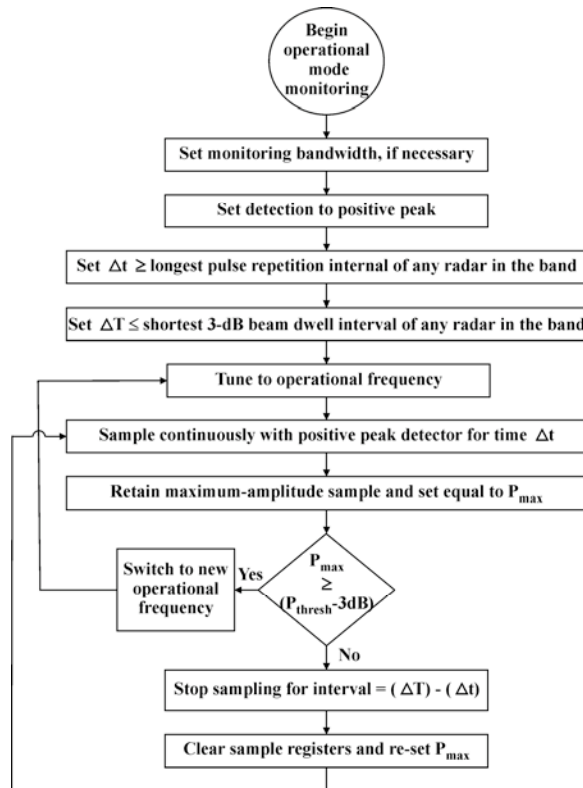


Figure 6. An algorithm for reliable detection of radar signals on a selected frequency.

Lower percentages of monitoring time could be used, but it then would take longer for radar signals to be detected. It would be up to communications operators to decide whether they would be willing to pay such a price to share spectrum with radars. Also, DFS is infeasible with bistatic radars as their receivers and transmitters are not collocated.

7. Summary and Conclusion

The problem of sharing radar spectrum with communication systems is very technically challenging. First, NTIA spectrum surveys have shown that radar spectrum is heavily utilized. This means that spectrum-sharing communication systems do not necessarily have as many opportunities to use radar spectrum in as many places as might be hoped.

Second, NTIA studies have shown that radar receivers typically begin to lose targets at I/N ratios in the receivers that are on the order of -10 dB RMS (for communication signals—pulsed interference can be tolerated at far higher I/N ratios). This means that communication signals are very likely to interfere with radars under almost any conceivable spectrum sharing deployment scheme. So far, no radars have been identified that are robust against interference from communication signals, including short-range and long-range air search radars, maritime navigation radars, or weather surveillance radars. Furthermore, low-level interference can cause target losses at all ranges, not just at the edge of radar coverage zones.

The only possible technical ways to deal with this inherent vulnerability are either to allow a certain amount of interference to radars on a statistical basis, or to design communication systems that can sense the presence of locally transmitted radar signals and systematically avoid the radars' frequencies.

The first solution would require the development of criteria for acceptable percentages of target losses in radars at all operational ranges. This is probably unlikely to happen in the near future and may never occur at all.

The second solution can be implemented, but could require up to 20-40% loss of operational time for communication systems if they are to listen often enough and long enough to reliably detect radar emissions with a high probability of intercept.

The technical challenges inherent in spectrum sharing between radars and communication systems are formidable and they need to be studied further. The problems inherent in this type of spectrum sharing are not necessarily impossible to solve from a technical standpoint, but they do pose substantial technical difficulties.

8. References

[1] F.H. Sanders and V.S. Lawrence, "Broadband spectrum survey at Denver, Colorado," NTIA Report 95-321, Sep. 1995.

[2] F.H. Sanders, B.J. Ramsey, and V.S. Lawrence, "Broadband spectrum survey at San Diego, California," NTIA Report 95-334, Dec. 1996.

[3] F.H. Sanders, B.J. Ramsey, and V.S. Lawrence, "Broadband spectrum survey at Los Angeles, California," NTIA Report 97-336, May 1997.

[4] F.H. Sanders, B.J. Ramsey, and V.S. Lawrence, "Broadband spectrum survey at San Francisco, California May-June 1995," NTIA Report 99-367, Jul. 1999.

Acknowledgments

The author wishes to thank Robert Sole of NTIA's Office of Spectrum Management and Brent Bedford of NTIA's Institute for Telecommunication Sciences for their technical participation and support.

Nation Building and Spectrum Management in Iraq

By

Fredrick Matos
National Telecommunications and Information Administration
Washington, DC
202-482-6493
Fax: 202-566-2440
fmatos@ntia.doc.gov

Formerly with the
Coalition Provisional Authority
Baghdad, Iraq

ABSTRACT

The paper describes the author's experiences as the Iraqi radio frequency spectrum manager and a telecommunications adviser during his nine months, June 2003-March 2004, in Iraq, acting for all practical purposes as a combined Iraqi national FCC and NTIA. He was on leave from NTIA, and worked for the Coalition Provisional Authority (CPA) Ministry of Communications that advised the Iraqis. The first action was to develop a national table of frequency allocations, working in conjunction with the Iraqi spectrum managers and the Coalition military spectrum managers. The joint plan was necessary to avoid interference and chaotic situations, and to bring some order to the spectrum management process. The joint table divided the spectrum bands into three categories: 1) bands for exclusive use by the military; 2) bands for exclusive use by civilians; and 3) shared bands. There was tremendous demand for broadcasting stations, so the next action was to develop national allotment tables for FM radio and television broadcasting, with the plans loosely following the ITU regional plans that were developed in 1984 and 1986. It was then necessary to develop license application forms for the various services such as land mobile, FM radio and TV broadcasting, aeronautical radionavigation services, fixed services, and fixed and mobile satellite Earth stations. The license forms were drafted, and brief rules were developed for the major services such as broadcasting and land mobile stations. The three national cellular telecommunications licenses, a complex 30-page document, were drafted and issued to the winning bidders. Daily activities included providing frequency assignments to the many private security firms in Iraq. Weekly meetings were held with the Iraqi spectrum managers to coordinate frequency assignments, exchange information, and discuss and develop spectrum management and telecommunications policies.

BACKGROUND

United Nations Security Council Resolution 1483, adopted on May 22, 2003, provided for the establishment of an "Authority" to "...promote the welfare of the Iraqi people through the effective administration of its territory..."¹ This led to the creation of the Coalition Provisional Authority (CPA) as the administrator, with 37 nations participating in the CPA at one time.

President Bush appointed Ambassador L. Paul Bremer as the Administrator of the Coalition Provisional Authority. The Coalition Provisional Authority replaced the Office of Reconstruction and Humanitarian Assistance (ORHA).

The CPA was organized to advise the Iraqis by establishing internal CPA "ministries" parallel to what were Iraqi Ministries, in most cases.² A few

examples are the Ministry of Education, Ministry of Health, Ministry of Justice, and the most relevant ministry to telecommunications, the Ministry of Transportation and Communications. The director of each CPA ministry carried the title of "Senior Adviser" with the principal responsibility of advising the Iraqi counterpart.

The Ministry of Transportation and Communications was a carryover from the previous regime, and it combined regulatory and managerial responsibilities of airports, seaports, railroads, telecommunications, broadcasting, and the post office. Since telecommunications and broadcasting are considered so important in the 21st century and to Iraq's future, it was decided in October 2003 to create a separate ministry by splitting the old ministry into two new ministries: the Ministry of Transportation responsible for airports and air travel, railroads, and seaports; and the Ministry of Communications, responsible for all

telecommunications and broadcasting, with the Ministry of Communications also responsible for the postal service.

CPA Order 11

The CPA was authorized to issue various regulations and orders in its nation building efforts. Most pertinent to telecommunications and spectrum management are Regulations and Orders defined as:

Regulations are instruments that define the institutions and authorities of the Coalition Provisional Authority (CPA).

Orders are binding instructions or directives to the Iraqi people that create penal consequences or have a direct bearing on the way Iraqis are regulated, including changes to Iraqi law.

Order 11 was drafted by the CPA Ministry of Transportation and Communications, with approval by the CPA Office of General Counsel, and signed by Ambassador Bremer and issued on June 8, 2003. It is titled "Licensing Telecommunications Services and Equipment."³ Order 11 consisted of two pages, with the paragraphs three and four providing for licensing and spectrum management:

3) The Ministry shall control, plan, administer, manage, and license the radio frequency spectrum. The Ministry shall honor Iraqi international commitments with respect to radio communication and telecommunication matters, unless prevented by security considerations or overriding international humanitarian law obligations. Accordingly, in exercising its responsibilities under this Order, the Ministry shall comply with the applicable standards and requirements of the International Telecommunication Union and its Radio Regulations, as agreed to or adopted by Iraq, unless prevented by reason of security considerations or overriding international humanitarian law obligations.

4) Applications to provide commercial telecommunications services, or for civil use of the frequency spectrum, shall be made to the Ministry.

Since the Iraqi Ministry of Transportation and Communications was non-existent following the immediate cessation of hostilities with the Saddam Hussein regime, the licensing and spectrum management responsibilities fell to the CPA Ministry because there was no Iraqi ministry to "advise."

The CPA Office of Spectrum Management

The first CPA spectrum manager was Fredrick Matos, on a detail assignment from the National Telecommunications and Information Administration, who arrived in Baghdad on June 18, 2003. It was a one-person office for several months until additional staff was added in August 2003 when a military spectrum manager was added, resulting in an experienced U.S. military spectrum manager being assigned to the CPA to assist in the day-to-day frequency assignment activities, to act as a liaison with Coalition forces, and to maintain the Spectrum XXI spectrum management software. The military spectrum manager liaison post was normally a three to four month assignment, necessitating some training of the news staff. Military spectrum managers are usually not experienced in civilian frequency management such as broadcasting, so some training was necessary.

The CPA spectrum management office was located in a small room in the Republican Palace within the fortified Green Zone. The Republican Palace was the CPA headquarters building.

Incumbent Iraqi Spectrum Management

In the pre-war Iraq, the spectrum management and licensing responsibilities were carried out by a small group of seven spectrum managers within the Iraq Telephone and Postal Company (ITPC).

The director of the ITPC spectrum management was Dr. Nasi Abachi, a very capable and experienced spectrum manager who received his PhD in electrical engineering, with a specialty in communications satellites, from the University of Essex in the United Kingdom. Dr. Nasi has more than 30 years experience in spectrum management.

Dr. Nasi had three very capable assistants in Engineers Maha Badir, Soumar Abdulkader, and Wesall Ali. It should be noted that Iraq was and remains a secular nation and women such as the three spectrum managers could obtain university educations and enter into professional careers.

The ITPC spectrum managers were wise in taking all computer equipment home with them during the war, and this proved to be an excellent move because their office was completely looted. The Iraqi spectrum management office had no wire telephones, and no Internet service, presenting even greater challenges in

getting things accomplished. Frequency assignment and coordination activities were delayed because of the poor communications.

The Iraqi spectrum management database of frequency licenses consisted of 400,000 assignment records. However, these were contained in three or four Microsoft databases, with the record of an assignment contained in part in all three databases. Thus it was difficult to review a single assignment record because the data was contained in three different databases.

Regarding the selection of frequency assignments, war-time damage was extensive, so the operational status of the 400,000 assignments was unknown in almost all cases. Nevertheless, to preclude interference, the frequency selection and coordination was conducted considering that all assignments were actually operating.

The CPA spectrum managers held weekly meetings with the Iraqi spectrum managers to select frequency assignments, to coordinate assignments, and to develop spectrum policies.

Iraqi Table of Frequency Allocations

The Coalition forces were extensive users of the spectrum, and the Iraqi civilian spectrum usage was increasing, with both segments seen as greatly expanding their spectrum usage in the near term. To provide managerial order and to prevent chaotic situations and serious interference, it was determined that a table of frequency allocations was necessary to allocate the spectrum to the two segments.

In late June and early July 2003, the CPA Ministry of Communications convened and chaired meetings of representatives from the Iraqi spectrum managers, the Coalition forces spectrum managers, the Central Command from Tampa, and the Joint Spectrum Center from Annapolis. Chaired by Fred Matos, the group developed a table of frequency allocations that divided the allocation table into two main parts: 1) military; and 2) civilian. The CPA Ministry of Communications retained all "licensing" authority, but "coordination" responsibilities for frequency assignment applications within their respective allocations were delegated to the Coalition forces for the military and to the Iraqis for civilian applications.

The table of frequency allocations followed the Table of Frequency Allocations contained in the Radio

Regulations of the International Telecommunication Union (ITU). The Table also emphasized the spectrum that was being used at the time, or was anticipated to be used in the near future. This emphasis stressed allocations such as land mobile and fixed service allocations, while making no provisions for the more esoteric services such as the Earth Exploration Satellite Service for which there were no requirements.

The allocation table underwent several minor changes, but overall it worked out very well, and there were no reported cases of interference.

(See appendix for the table of frequency allocations)

Regulations, License Application Forms, Licenses, and Application Fees

It quickly became apparent that there were no regulations, license application forms, licenses, or application fees. People were beginning to apply for broadcasting stations and other types of licenses, and the necessary forms and other documentation needed to be developed very quickly. Application forms and licenses were drafted for stations operating as amateur, land mobile, fixed, FM and TV broadcasting, AM broadcasting, aeronautical mobile, radionavigation, broadcasting satellite uplinks, and Fixed Satellite Earth Stations including VSATs. These application forms were coordinated with Dr. Nasi who suggested some additions in the forms, especially in expanding the questions on why the service was necessary and what it would be used for.

Application fees were developed, with exemptions made for CPA and Coalition member nations, and for contractors to the CPA. The broadcasting station fee was set at \$500 for a 6-month license. The short license term was established because the broadcasting regulations were minimal, and if a broadcaster created problems of any type, the license would not be renewed.

National Allotment Plans for FM and TV Broadcasting

Broadcasting prior to the hostilities was limited to government stations, and these were very few. Saddam Hussein's son was the broadcasting media director.

The exception to this is in the Kurdish area that was semi-autonomous prior to the war. The Kurdish

territory had two provincial capitals, each with a form of telecommunications licensing “authority.” This “authority” issued licenses of various types, including radio and television broadcasting.

A policy decision was made by the CPA to license commercial broadcasting stations.

FM and television broadcasting lend themselves to national allotment planning, with cities allotted various channels to provide for a national broadcasting plan to provide broadcasting for all of the citizens. National allotment plans are an excellent spectrum management tool, and it was decided to develop national tables as a part of a national spectrum management plan.

The ITU developed FM and TV plans for Region 1 in 1986 and 1988, with channels allotted in geographic areas. Unfortunately, some of the geographic areas are sparsely populated, resulting in fewer channels for the densely populated cities. The CPA worked with the Iraqis to develop national allotment plans, using the ITU plans as a foundation. Adjustments were made to the ITU plans, and to reach more Iraqi citizens, various channels were moved from rural areas to the more populated larger cities. This effort took a number of weekends, and was completed on August 10, 2003.

The allotment plans presented new challenges because the FM channels are spaces every 100 kHz, rather than the 200 kHz spacing used in the United States. Furthermore, some TV channels are 8 MHz wide while others are 6 MHz wide.

(See Appendix for copies of the allotment plans.)

The overall results were national allotment plans that were used when an applicant desired an FM or TV channel in a city.

Cellular Telecommunications Licensing

Wire telecommunications services were very poor prior to the war, with about two percent of the population having telephones. There were no cellular telecommunications services. The CPA made a policy decision to issue three regional cellular licenses, each licensed to a specific geographic area, with specific build-out requirements in the license. If the build-out milestones were met, then the company would be issued a national license thereby providing

competition that would keep prices at affordable levels.

A number of firms bid on the cellular licenses, and an Iraqi panel was convened to evaluate the bids and select the winners. The winners were Orascom (now known as Iraqna) for the central region including Baghdad; AsiaCell for the northern region; and Atheer for the southern region. Licenses were drafted, and licenses were formally granted to the three companies on December 22, 2003. The licenses themselves were complex documents of about 20 pages each.

The cellular service is very popular, and as of January 1, 2005, there are over 1,500,000 cellular subscribers in Iraq.

Day-to-Day Frequency Management Activities

Much of the normal day-to-day activities were concerned with the processing of land mobile and other types of license applications.

The widespread looting in Iraq resulted in the establishment of numerous private security companies to protect individuals and facilities, most of which required land mobile radio communications capabilities. License applications of the private security firms were the most prevalent of any entity in Iraq.

The CPA frequency management office selected the frequencies for various applicants. In most cases, the frequency band of the equipment that was to be used by the applicant conformed to the allocation table.

The most challenging spectrum management project was assigning frequencies to police communications radios operating in the 450-470 MHz band in 92 cities and communities. This was an urgent project because, with the exception of the Baghdad police, the police in other cities had no communications capabilities. Major cities were planned to have as many as 10 land mobile repeaters and numerous simplex operations, presenting challenges to avoid self-interference from mechanisms such as intermodulation and adjacent channel. Inter-city interoperability was integrated into the plan to provide communications capabilities to police in adjacent cities in those cases where they had to assist the police in the adjacent city. Frequency re-use was used as much as possible, and provisions for future growth and expansions were built into the frequency

plans with all radios factory programmed to include the future channels.

Amateur Radio in Iraq – New Freedoms

Amateur radio operations were permitted in Iraq prior to the war, but it was very limited and all operations had to be from the national radio club station, Y11BGD, located in Baghdad. Operations from amateur's homes were not permitted, nor were any types of mobile or portable operations. Operations are permitted only in the amateur HF bands.

The CPA spectrum management office licensed about 20 foreign amateurs to operate in Iraq. The licenses did not restrict operations any locations, and permitted operations in all frequency bands following the ITU allocations.

Diya Alasadi, the president of the Iraq Amateur Radio Society, observed that the CPA was issuing amateur licenses, and called and inquired if the CPA was writing amateur rules and regulations.

Fred Matos replied, "No, we are not writing any amateur rules and regulations. You must write your own rules and regulations. This is a new era for Iraq, with new freedoms, and the amateurs will be largely self-regulating."

Diya was both surprised and dumfounded by this newly found freedom that Iraq amateurs never had before. Diya later energized a group of Iraqi amateurs and drafted the rules and regulations.

(See July 2004 issue of QST magazine for additional information on amateur radio activities in Iraq.⁴)

The Independent Telecommunications and Broadcasting Commission

It was decided very early that Iraq should have an independent regulatory commission for telecommunications and broadcasting. Such a commission would make national policy, and conduct all spectrum management activities such as licensing, monitoring, policy development and implementation, and participation in regional and worldwide bodies where technical standards, agreements, and treaties are developed.

Order 65 was issued by the CPA and signed by Ambassador Bremer on March 20, 2004, in a document titled "Iraqi Communications and Media Commission"⁵ consisting of 14 pages.

Order 65 provided that the Iraqi Communications and Media Commission (ICMC) consist of nine part-time commissioners and a full time Director General.

One of the major problems confronting the ICMC is populating the various regulatory bureaus and offices. Other than a small spectrum management group, there is no telecommunications or broadcasting regulatory experience in Iraq, and the human resources and training problem is a major one. For example, there is no one to establish domestic telecommunications network interconnection rules, to draft broadcasting rules and regulations, or to develop the rules for Internet services, if any are needed.

United Nations Security Council Resolution 1483 (2003), adopted May 22, 2003

² There were some exceptions to this paralleling because in some cases the Iraqi pre-war ministry did not exist. One example of this was the Ministry of Human Rights that was newly created.

³ Coalition Provisional Authority Order "Licensing Telecommunications Services and Equipment," CPA/ORD/8 June 2003/11, June 8, 2003.

⁴ "Rebuilding the Iraq Amateur Radio Society," QST, July 2004, pages 86-87, American Radio Relay League

⁵ Coalition Provisional Authority Order Number 65, "Iraqi Communications and Media Commission," CPA/ORD/20 March 2004/65, March 20, 2004.

Draft Frequency Allocation Table for Iraq - 30 August 2003

Frequency (MHz / GHz)	Major Allocation	Main User and Coordinator	Notes
3-26.175 MHz	Fx/Mob/ Amateur	Shared band	Military to co-ordinate use and take responsibility for deconfliction for Fx/Mob. ITPC to provide details of current users and CPA to provide details of other known users (e.g. UN)
26.175-28.0	Fx/Mob	Civ	Citizens Band and small unlicensed use
28-30	Amateur	civ	
30-47	Fx/Mob	Mil	Freqs around 43 and 49 MHz are also used for domestic indoor cordless phones.
47-68	Fx/Mob	Mil	Possible future use for TV ch 3, 4 & 5. Cordless phones – see above note.
68-87.5	Fx/Mob	Mil	ITPC will forward known assignments to mil for inclusion in XXI database who will then work around them.
87.5-138	Broadcasting/ Mob	Civ	
87.5-108	FM Radio B/C	Civ	Possible mil assignments
116-138	Mob	Mil	Some civ assignments at airports
138-144	Mob	Mil	Possibly some illegal use of 143-147 from illegally imported equipment.
144-146	Amateur	Civ	
146-148	Mob	Mil	
148-174	Mob	Civ	Many civilian assignments; ITPC to provide 90 assignments (25 kHz) to Military. Any further military needs to be requested through CPA
174-230	Broadcasting	Civ	TV Channels 6,7, 8,9,10,11,12,13
230-400	Aeronautical Mobile and others	Mil	Civilian use of 253-263 and 379-389 MHz for long range cordless phones. Up to 8000 licensees though not clear how many still operational. Mil to work around these bands. Mil to be receptive to any future civilian requirement (e.g. Tetra 380-410).
400-403	Met Aids	Civ	
403-410	Met Aids Fixed Mobile Mobile Sat	Civ	
410-430	Mob	Mil	
430-470	Mob	Civ	Includes Baghdad Police radios etc.
470-510	Broadcasting	Civ	TV Chs 21,22, 23, 24, 25...each 8 MHz wide.
510-702	Broadcasting	Civ	TV chs 26-50
702-790	TV (8 MHz wide	Civ	Generally used for TV so military will make

Appendix A

	channels)		every effort to leave this band as soon as possible. In meantime, mil to protect ch 38 at 606-614 MHz used by PUK media office. 3 contiguous chs previously used for digital TV service by Rafidain Co. Service had about 10,000 subscribers – it may be about to relaunch – investigation required.
790-800	Broadcasting	Civ	
800-830	Fx	Mil	
830-960			MIL will coordinate through CPA with ITPC for assignments between 830-870MHz.
960-1215	Aero RadioNav	Mil	Except civil aviation requirements which will need to be assigned by mil
1215-1350	AeroRadioNav	Mil	Except civ air traffic control in the future (already assigned by mil)
1350-1465	Fx/Mob	Civ	
1465-1490	Fx/Mob/BC/BCSat	Mil	Possible future use by digital sat and terrestrial broadcasting
1490-1559	Fx/Mob/Sat	Civ	
1559-1610	AeroRadioNav	Mil	
1610-1660.5	Mob Sat	Civ	
1660.5-1700	Fx, Mob, Met Aids, MetSat	Mil	Except 1670,1691,1694.5 MHz for MetAids. BW unknown – mil will avoid assignments
1700-1710	Met Sat	Civ	
1710-1785	Fx/Mob	Civ	Lower end of GSM band (see note 1)
1785-1805	Fx/Mob	Mil	The guard band between lower and upper ends of GSM. Mil will make efforts to vacate the band as soon as possible. Possible future use for WLL.
1805-1880	Mob	Civ	Upper end of GSM band (see note 1)
1880-1956	Fx/Mob	Civ	Part of 3G band
1956-1980	Fx/Mob	Mil	Part of 3G band
1980-2075	Fx/Mob	Civ	
2075-2100	Fx/Mob	Mil	
2100-2200	Fx/Mob	Civ	
2200-2220	Fx/Mob/others	Mil	
2220-2300	Fx/Mob/others	Civ	
2300-2380	Fx/Mob/others	Civ	At the meeting the military was given the

Appendix A

		UNDER REVIEW	option to use 20MHZ bandwidth; either 2370-2390 or 2340-2360. Can this band be assigned as originally discussed (2370-2390MHZ) .
2380-2400	Fx/Mob/others	Mil	
2400-2500	ISM	Civ	Wi Fi Band
2500-2520	Mob Sat	Mil	ITPC to supply details of 6 or 7 fixed microwave links, assigned to oil companies, now most likely destroyed.
2520-2670	Fx/Mob	Civ	
2670-2690	Mob Sat	Mil	ITPC to supply details of 6 or 7 fixed microwave links, assigned to oil companies, now most likely destroyed.
2690-2700	Sat	Civ	
2700-3400	Radiolocation & radionav	Mil	Any Civ requirements such as WLL will be coordinated with Mil
3400-3625	Fx/FxSat	Civ	
3625-4200	Fx/FxSat	Civ	Mil has satellite downlinks. New satellite gateway to be installed soon. Paired with 5850-6425. Mil will request freqs
4200-4400	Aeronautical radionav	Civ	
4400-5000	Fx	Mil	Tropo band
5000-5250	Fx/Mob/ Radionav	Civ	
5250-5725	Radiolocation	Mil	
5725-5850	Radiolocation	Civ	New unlicensed Wi Fi band
5850-6425	Fx/FxSat	Civ	Mil to request freqs for satellite uplinks (Paired with 3625-4200 downlink)
6425-7125	Fx/FxSat	Civ	
7125-8500	Fx & FxSat	Civ	Mil will coordinate freq requests with Civ. Mil will check satellite Earth station use in the region.
8500-9000	Radiolocation	Mil	Assigned to civilian use at meeting but need to confirm in light of military needs.
9000-9990	Radiolocation & Radionav	Mil	Mil to avoid 9375 (actually 9355-9395) to avoid airborne weather radars and maritime radars.
9.9-10.7	Radiolocation	Civ	
10.7-12.5		Civ	10.7-12.5 is satellite uplink. Iraq has some satellite usage. Mil will coordinate through CPA/ITPC.
12.5-13.4	FxSat	Civ	
13.4-14.0	FxSat	Mil	

Appendix A

14.0-14.4	FxSat	Civ	
14.4-14.5	FxSat	Civ	Mil will request use as necessary
14.5-15.3	Fx	Civ	Managed by Civ – new requests to CPA. MCI use for GSM connections; and Civ use by mobile ENG.
15.3-15.7	AeroRadionav, FxSat in 15.43- 15.63	Mil	
15.7-16.0	Radiolocation	Mil	
16.0-16.5	Radiolocation	Mil	
16.5-16.7	Radiolocation	Civ	
16.7-17.25	Radiolocation and others	Mil	Region 1 Bands are 16.6-17.1, 17.1-17.2, and 17.2-17.3
17.25-17.7	FxSat	Civ	
17.7-23.0	Various	Civ	Military will coordinate with CPA and ITPC.
>23 GHz		Shared band	Except for the 38 GHz band which is allotted for GSM microwave backhaul links.

Assumptions and Notes

The table is the result of discussion between the Coalition Forces and the Iraqi Spectrum Management Team at the Ministry of Transport and Communications. The discussions took place at a conference held in Baghdad on 21 and 22 June 2003. A list of the attendees is set out at Annex A [*Grateful if Centcom could provide the list of attendees which was prepared and handed out in hard copy at the meeting*].

The table represents the recommendation for allocations to military and civilian users that was agreed at the conference. These recommendations are subject to approval from senior officials in Centcom and CPA. It was agreed that this approval should be obtained by 15 July.

The CPA is the body responsible for the overall management of the spectrum. The table sets out which frequency bands have been allocated to the military and which are reserved primarily for civilian use. In each case, the party assigned the frequency band (i.e. military or civilian) will be responsible for de-confliction and co-ordinating use of that band.

The above allocation does not take any account of any future frequency use by a future Iraqi army. At such time that an Iraqi army is established, further discussions will need to be held to identify what frequencies they should be assigned. At this time, it is expected that some of the frequencies currently assigned to the coalition will be re-assigned to the Iraqi army.

The table describes who (from civilian and military) is allocated each frequency band. It is to be expected that military users may, from time to time, need to request use of frequencies that have been assigned as a civilian band and vice versa.

Notes on Table

- (1) GSM bands. CPA/ITPC aware that military may make requests to use frequencies in this band – particularly in the short-term. Such requests will need to be balanced against commercial use of the band.

NATIONAL TELEVISION ALLOTMENT PLAN
Developed by the CPA, the Iraqi Spectrum Management Staff, and the I.M.N.
10 August 2003

ITEM	CITY	VHF CHANNELS				UHF CHANNELS							
		5	7	9	12	22	45	48	51	54	57	59	
1	BAGHDAD	5	7	9	12	22	45	48	51	54	57	59	
2	SINJAR	5				49	52	55					
3	MOSUL	7	9	12		22	25	27	57				
4	KIRKUK	6	8	10		32	35	50	59				
5	BAQUBA	11				23	26	39	42				
6	AL-QAIM	7	9	12		32	35	38					
7	HADITHA	6	8	10		30	33	36					
8	KUT(WASIT)	2	6*	8*	10*	31	34	37					
9	BABIL	4				35	38	41					
10	AL-NASIR					33	36	40	43				
11	KARBALAA	11				40	43	46	57				
12	AL-SAMAWA	6*	8*	10*		21	24						
13	MAYSAN	5	8	11		29							
14	BASRAH	3	7	9	12	22	25	27	41	44	54	57	
15	UM QASIR	5	8	11		48	51						
16	ALI AL-GARBI	7	9	12		48	51						
17	DEWANIA	7*	10*	12*		50	53						
18	RUTBA	3	7	9	12	48	51	54					
19	TIKRET	2				25	27	29	38				
20	ARBIL	4	11			21	24	27	31				
21	KHANAQEN					29	31	34	37				
22	DAHUK	3	10			30	33	36	43	47			
23	SUQ SHIOKH	11				26	28	49					
24	RAMADI	3	8	11		50	53	56					
25	SULAIMANIA	2	12			30	33	36	40	43			
26	NAJAF	3	5*	8*	11*	25	28						
27	NASIRYA	7	9	12									

* Low powered station not more than 40 dBW Effective Radiated Power (ERP)

Plan is based on the ITU "GE89" regional plan of 1989. Except for the low-powered stations, all others are high power as high as 50-60 dBW

Other stations may be added to the allotment plan pending an engineering electromagnetic compatibility analysis showing that no harmful interference will occur.

Iraqi Media Network (IMN)
OPERATING/RESERVED CHANNELS
10 August 2003

ITEM	CITY	VHF CHANNEL				UHF CHANNEL							
1	BAGHDAD		7	9	12	22							
2	SINJAR	5				49	52						
3	MOSUL	7	9	12									
4	KIRKUK	6	8	10									
5	BAQUBA	11				23	26						
6	AL-QAIM	7	9	12									
7	HADITHA	6	8	10									
8	KUT(WASIT)		6*	8*	10*								
9	BABIL					35	38	41					
10	AL-NASIR					33	36	40					
11	KARBALAA	11				40	43						
12	AL-SAMAWA	6*	8*	10*									
13	MAYSAN	5	8	11									
14	BASRAH		7	9	12								
15	UM QASIR	5	8	11									
16	ALI AL-GARBI	7	9	12									
17	DEWANIA	7*	10*	12*									
18	RUTBA		7	9	12								
19	TIKRET					25	27	29					
20	ARBIL					21	24	27					
21	KHANAQEN					29	31	34					
22	DAHUK					30	33	36					
23	SUQ SHIOKH	11				26	28						
24	RAMADI		8	11		50							
25	SULAIMANIA					30	33	36					
26	NAJAF		5*	8*	11*								
27	NASIRYA	7	9	12									

FM BROADCASTING
NATIONAL ALLOTMENT TABLE
9 August 2003

1	Baghdad	88.6B, 89.4C*, 91.4C*, 92.3C, 94.6C*, 95.5B, 98.3C, 100.4B, 101.4B, 102.4B, 103.2B, 104.1B, 106.9B, 107.7B
2	Baqubah	90.2B*, 93.1B*, 96.3B*, 105.4B
3	Falluja	93.8B, 97.1B
4	Tikrit/Bayji/Samara	87.2B*, 88.0B*, 94.0A, 96.0A, 98.0A*, 105.2A, 107.2A
5	Karbala	89.7B*, 96.0B*, 99.3B*, 102.8B,
6	Al Hillah/Babil	90.4A*, 97.9B*, 104.2B*, 105.2A, 106.3B, 107.2 B
7	Bayji	107.2A
	Ramadi	88.9A*, 90.7A*, 96.6A, 100.1B*, 106.0A
		89.0B*, 98.3B*, 102.4B*, 103.6B, 107.8B
10	Mosul	87.2B, 90.7B, 92.4C*, 95.6C*, 97.2B, 98.9C*, 105.1B, 106.6B, 107.4B
11	Dahuk	88.2B, 94.0B, 97.5B, 99.3B*, 100.8B*, 102.8B*, 104.0B,
12	Irbil	89.8B, 93.6B*, 95.0B, 96.7B*, 100.1B*, 101.9B, 103.2B, 104.7B, 105.9B
13	Sulaimania	92.0A, 93.0A, 94.5C*, 96.0A, 97.0A, 98.0A, 101.3C*, 102.0A, 103.0A, 103.8B*, 104.6B, 105.8B, 107.2B
14	Kirkuk	87.9C*, 90.0B, 91.5C*, 93.3B, 96.4A, 97.8B, 98.8A, 101.0C*, 107.8A
15	Khanaqen	88.7B*, 90.4B*, 93.7B*
16	Basrah	88.5C*, 90.4A, 96.8C*, 98.1C*, 101.0A, 103.0B, 104.2B, 105.0B, 107.0B
17	Um Qasr	87.2A, 93.4B, 98.9B*, 105.8B*, 107.8B*
	An Nasir	90.6B*, 92.2B*, 101.2B*, 104.4B
	Al Garbi	90.0A, 98.1A, 99.6A
	Al Amarah	87.5A, 102.0B*, 103.8A*, 105.9B*
	Al Nasiriyah	89.4B*, 93.2B*, 95.0B*, 103.6A, 107.8B
21	Samawa	89.6B*, 92.7B*, 93.6B, 95.6B, 99.2B*, 102.7B, 106.3B
22	Al Diwaniya	88.0B*, 92.2B*, 95.4A*, 98.7A, 102.2A, 105.8A
23	Al Kut-Wasit	91.7B*, 94.9B*, 96.2B, 98.2B*, 101.7B
24	Rutbah	90.5B*, 93.6B*, 100.1B*
25		90.3B*, 93.4B, 98.2B*, 99.9B
26	An Najaf	91.9B*, 94.2B*, 95.1A*, 98.4A, 101.9A, 105.5A
	Al Qaim	90.0C, 93.1C, 96.3C, 99.6A, 103.1A
28	Suq Al-Shioki	94.5C, 99.0C, 104.9C, 107.5C
29		

* Currently operated by the IMN, or reserved for the IMN.

FM-Radio Broadcasting Station Classes – Definitions

All station power ratings are in Effective Radiated Power (ERP) consisting of the combination of transmitter power and antenna gain.

Class A – Low powered station

Maximum ERP of 6 kW.

Maximum antenna height of 100 meters above average terrain level.

Appendix C

Class B - High powered station

Maximum ERP of 50 kW.

Maximum antenna height of 150 meters above average terrain level.

Class C Very high powered station

Maximum ERP of 100 kW.

Maximum antenna height of 200 meters above average terrain level.

Minimum Distance Separation Distance in km

	Co-channel	st Adj Ch 200 kHz	^{2nd} Adj Ch 400 kHz	^{3rd} Adj Ch 600 kHz	10.6/10.8 MHz
Class A to A	132	85	45	37	8
Class A to B	206	132	76	69	16
Class B to B	237	164	94	74	24
Class A to C	242	177	108	100	32
Class B to C	274	209	125	106	40
Class C to C	306	241	153	113	48

Same city channels

Separate stations, either A or B, by a minimum of 800 kHz.

Example for City XYZ frequencies 101.4, 102.6, 103.4, 104.2

Propagation and Throughput Study for 802.16 Broadband Wireless Systems at 5.8 GHz

Thomas Schwengler, *Member IEEE*
Qwest Communications, 1860 Lincoln street
11th floor, Denver CO 80295 USA.
(phone: +1 720-947-1184; fax: +1 720-947-1194;
email: thomas.schwengler@qwest.com).

Niranjan Pendharkar
University of Colorado at Boulder
530 UCB, CO 80309-0530 USA
(phone: +1 720-352-2227; email:
niranjan.pendharkar@colorado.edu)

Abstract— This paper presents propagation studies and analyses of OFDM signals, following the 802.16-2004 standard, at 5.8GHz. Throughput measurements are conducted first in a controlled faded multipath environment in the lab, then in a suburban area. Results are analyzed and compared.

Index Terms—SUI propagation models, fading channels, OFDM.

I. INTRODUCTION

THIS paper presents the results of a study measuring data throughput of an OFDM radio system through various fade models. The radio system is 802.16-2004 compliant [1], using 256 FFT at 5.8 GHz. Only one sector is used, therefore no other cell interferences are considered. A single 20MHz channel is used for the sector and multiple access is obtained by time slot allocation to all units within the sector.

The first part of the paper presents a study of this system through a controlled environment, where radio multipaths and their resulting fades are generated by a channel emulator. The fading models are based on Stanford University Interim (SUI) channel models.

The second part of the paper presents the same radio system tested for throughput in a suburban area in Denver. In that case radio interferences are verified to remain consistent and fairly minimal in order to focus on channel variations similar to those considered in the controlled environment.

II. LAB TESTING

A. Test Setup

The radio system under test comprises one base station sector (BS) and several subscriber units (SU's). This study is interested in fixed broadband wireless communications in various propagation environments; consequently fixed models are considered, rather than the usual mobile propagation models. Tests were

conducted to measure the throughput in different modulations. Devices were tested in a part cabled environment and part unbounded media as shown below. The cabled environment undergoes different fading channels programmed in a fading emulator. The air interface is a short direct line of sight with the BS and the SU's at a distance of 10 feet. This is done in order to couple signals of four SU's over the air onto one sector.

The Fading emulator allows us to emulate two separate channels, each comprised of several multipaths, each of which is independently faded and delayed. Fade statistics for the direct path are either Rayleigh or Ricean, fade statistics for the delayed paths are all Rayleigh. Finally additive white Gaussian noise is added to the overall channel ($C/I=30\text{dB}$).

As in many wireless LAN devices, our radio devices are TDD and have duplex ports: transmit and received signals go to one unique antenna. In our test, the fading emulator fades the transmit and receive paths independently, the two paths are therefore separated by circulators. Finally the fading emulator required some careful calibrating of power levels (especially due to the high peak to average ratio of OFDM signals). Radio transmit power levels were adjusted and additional attenuation (pad) was added where necessary. Figure 1 shows the detailed setup.

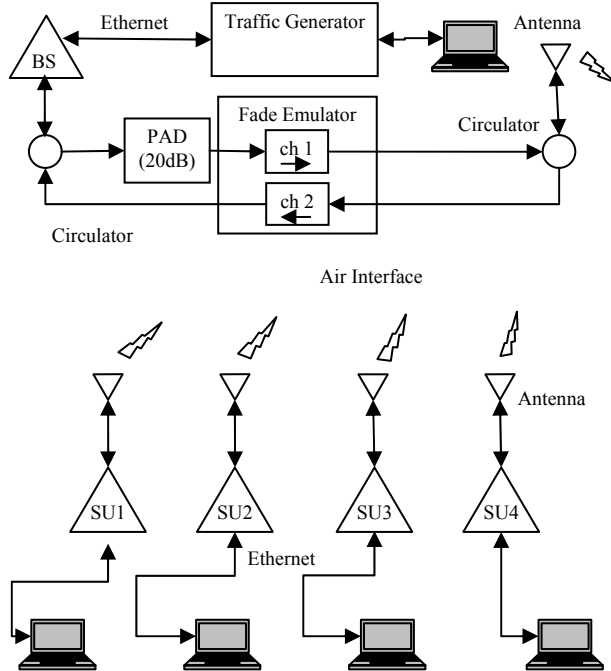


Fig. 1. Test Setup.

B. Channel Models

Different channel models are emulated using the modified Stanford University Interim (SUI) channel models. In particular we focus on fixed access and consider SUI-1, 3, and 5, described in table I below (SUI-4 & 6 have high Doppler spread and are not relevant to our study. Sui-2 is fairly similar to SUI-1). We therefore have a model for different terrain types A, B, and C, as described below (for more details, see [2], [3]).

TABLE I
FADING CHANNEL MODELS

Model	SUI-1	SUI-3	SUI-5
Terrain Type	A: Flat, light tree density	B: Hilly, light tree density or Flat, moderate heavy tree density	C: Hilly, moderate to heavy tree density
Doppler	Low	Low	Low
Delay spread	Low	Low	High
Ricean K of direct path	4 (High)	1 (Low)	0 (Rayleigh)
Multi-path (delay & atten.)	3 paths, 1: direct 2: 0.4 μ s, -21dB 3: 0.9 μ s, -30dB	3 paths, 1: direct 2: 0.4 μ s, -11dB 3: 0.9 μ s, -22dB	3 paths, 1: direct 2: 14 μ s, -11dB 3: 20 μ s, -22dB

Throughput results are measured for these different SUI models and different modulations and coding: BPSK, QPSK and 16QAM, with forward error correction coding (convolutional coding) with a coding

rate of 1/2, 2/3, or 3/4. (Tests were also conducted at 64QAM, but that throughput was perturbed by other system limitations rather than propagation fading; and these results are therefore not meaningful.)

C. Results

Overall throughput is very steady and reliable with – as expected – increasing degradation as the index n of the SUI model (SUI- n). In BPSK, the SUI model has barely any impact on throughput, at higher modulations, a slight degradation is noticeable.

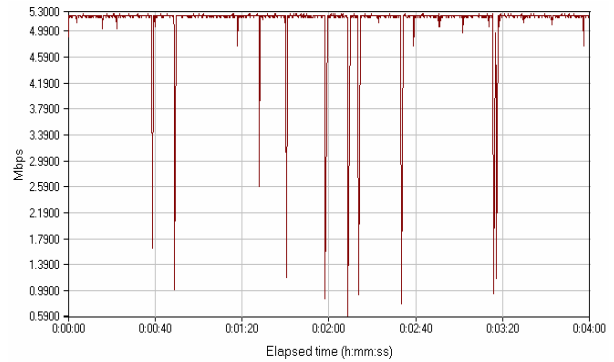


Fig. 2. Throughput vs. time for BPSK modulation in SUI-1 channel model.

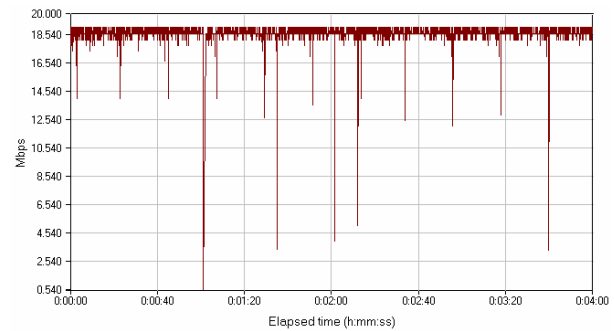


Fig. 3. Throughput vs. time for 16 QAM modulation in SUI-1 channel model.

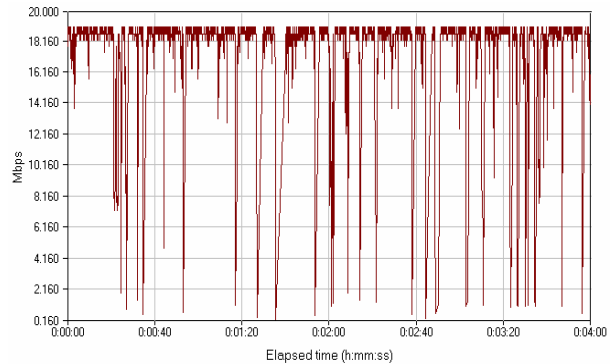


Fig. 4. Throughput vs. time for 16 QAM modulation in SUI-3 channel model.

Probability analysis of the throughput levels show a

very steep cumulative distribution function for SUI-1, less so for higher models.

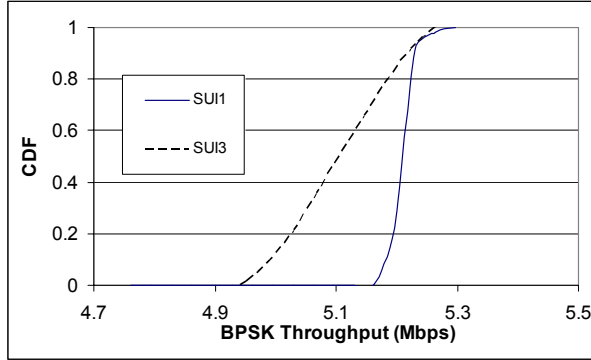


Fig. 5. Cumulative distribution of throughput in BPSK modulation, for SUI-1 and SUI-3.

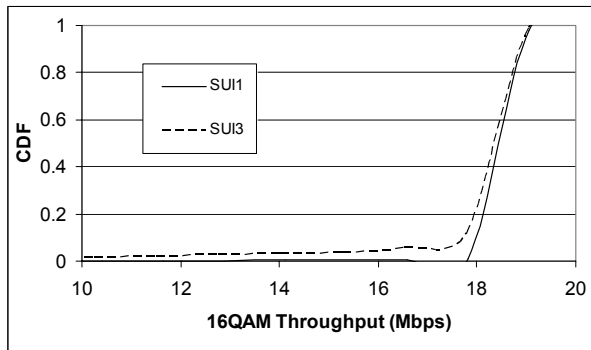


Fig. 6. Cumulative distribution of throughput in 16QAM modulation, for SUI-1 and SUI-3.

Still, in spite of these differences, average throughput comparison shows no significant degradation as modulation increases. Our three SUI models are represented on Figure 5 and 6 and compared to the “bypass” setup which is simply cabled through the fade emulator in bypass mode.

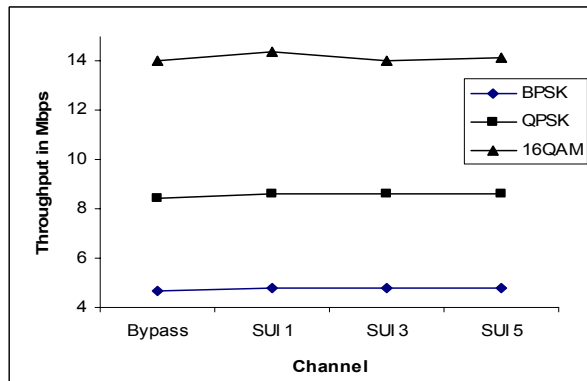


Fig. 7. Average throughput in Mbps for various channel models at different radio signal modulations.

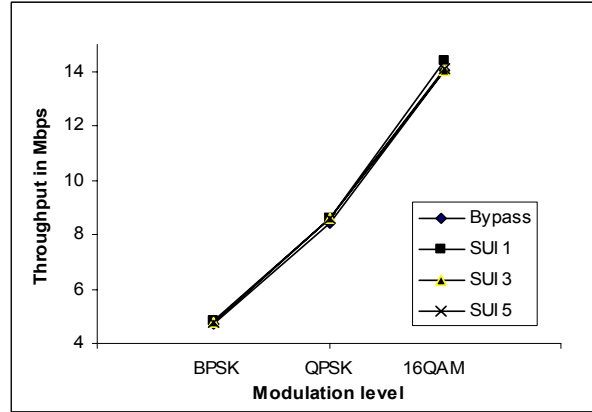


Fig. 8. Average throughput in Mbps for various channel models at different radio signal modulations.

D. Comments

The difference in propagation models does not significantly impact average throughput. Different SUI models present significant differences in fading, and these differences in fading statistics were observed; but the various coding schemes of an 802.16 radio system deal with these fades efficiently.

A further aspect that should be tested in future experiments is the impact of S/N or signal strength in those various SUI models.

III. URBAN TESTING

A. Test Setup

After the lab study we take the same equipment and conduct true field testing in a suburban area in Denver. The equipment used is similar than that of section I and Figure 1, but the circulators, padding and fade emulator are removed. The BS is placed on top of a 13-floor-high building, and the SU’s are placed 6 to 8 feet off the ground, each on a small pedestal atop a vehicle roof.

In this case the system is configured differently from the lab setup in one important aspect: a modulation on demand is allowed where each SU is allowed to choose a specific modulation according to its SNR. Unlike the lab test, the BS is communicating with SU’s at different modulations.

B. Single Unit

We first test throughput with one single SU at various locations within the sector. All locations are in obstructed line of sight, some only by minor foliage, some completely shadowed by buildings. In many cases insufficient signal was obtained to establish data link

reliably, these cases are not plotted but should be kept in mind: although our data points are very impressive for obstructed links at 5.8GHz, service is not ubiquitous.

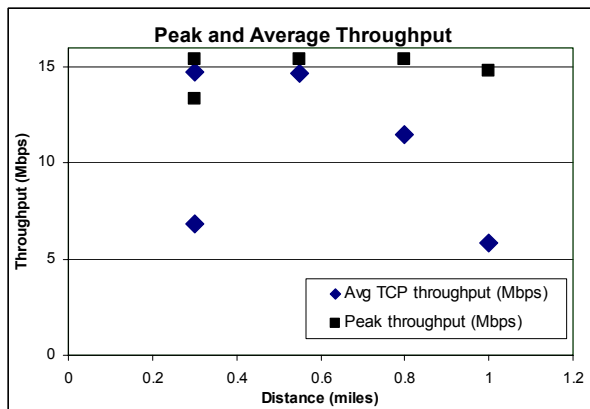


Fig. 9. Average and peak throughput in Mbps for various locations within a sector in actual field testing.

To compare to lab experiment, we represent the cumulative distribution of all data points.

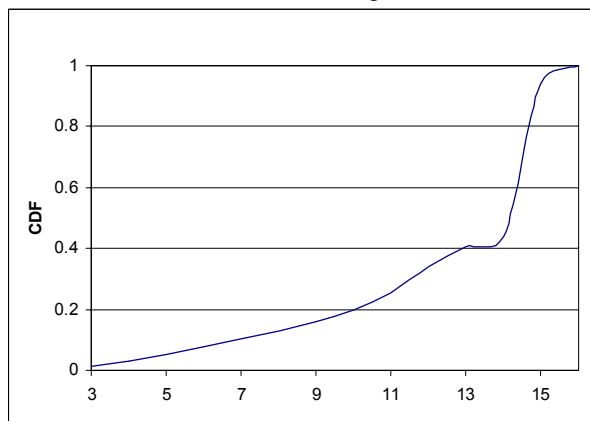


Fig. 10. Cumulative distribution of throughput in various locations measured in a suburban area.

C. Multiple Units

Finally for practical considerations of several simultaneous users, we then test throughput with several SU's at various locations within the sector. All locations are again in obstructed line of sight, some only by minor foliage, some completely shadowed by buildings.

Three sets of measurements are collected, with three

to five SU's in different areas (represented by a contour on the map). In each contour donut-shaped symbols represent the location of the SU. A map representing the topography of the setup is shown at the end of the paper.

D. Comments

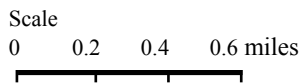
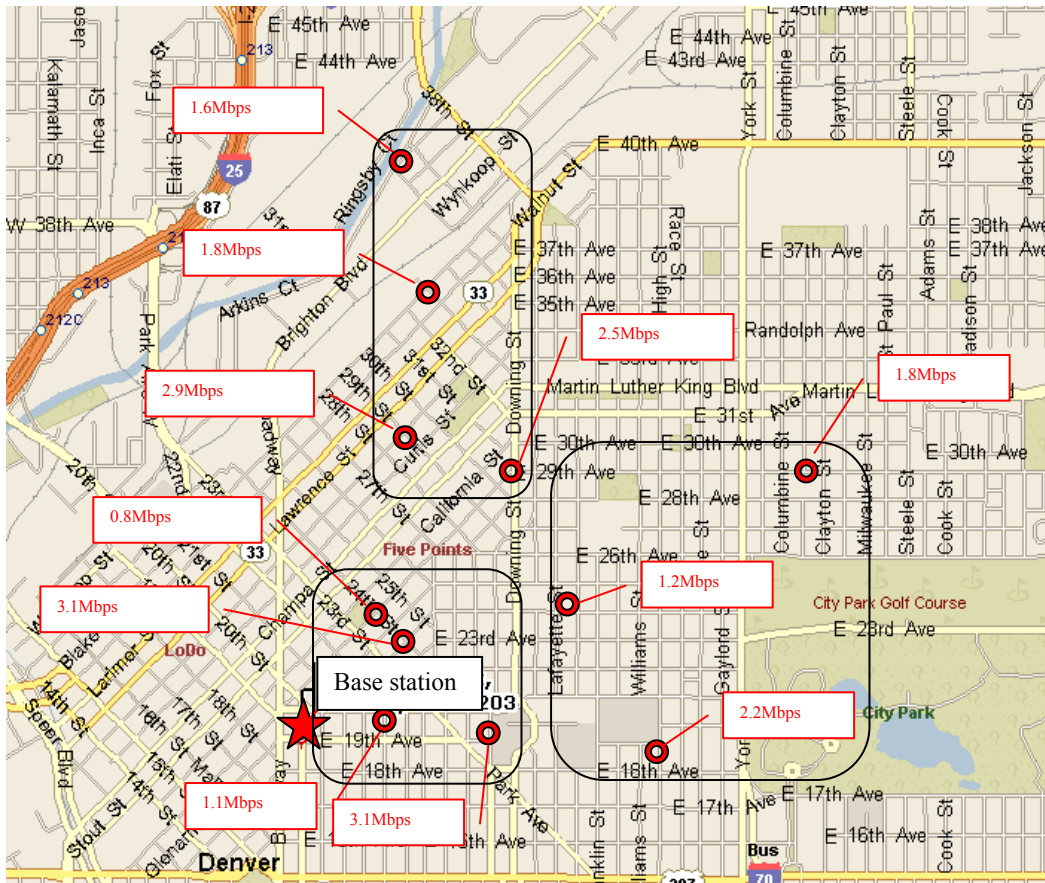
Throughput with one SU in different locations throughout the sector show good results in spite of obstacles such as trees, homes and urban traffic. When Several SU's are combined in one sector, details of the scheduling algorithm between SU's prevent us from analyzing these data point in too many details; but we nevertheless verified that performance is maintained when several SU's are used within one base station sector.

IV. CONCLUSION

We presented throughput measurements in three different SUI models for use in fixed broadband wireless access for rural and suburban areas. We then tested the system in a Denver suburb and observed similar throughput where signal strength was sufficient. Field tests reminded us that shadowing and obstruction effects are the most important point for any broadband radio deployment.

REFERENCES

- [1] IEEE 802.16-2004, May 2004. "Air Interface for Fixed Broadband Wireless Access Systems"
- [2] IEEE 802.16 Broadband Wireless Access Working Group, "Channel Models for Fixed Wireless Applications" Contribution to 802.16a, 2003, available at http://wirelessman.org/tga/docs/80216a-03_01.pdf.
- [3] V. Erceg et al., "An Empirically Based Path Loss Model for Wireless Channels in Suburban Environments", IEEE Journal on Selected Areas in Communications, Vol. 17, No. 7, July 1999.
- [4] M. K. Simon, M.-S. Alouini, Digital Communications over Fading Channels. New York: John Wiley & Sons, 2000, ch. 2.
- [5] C. Chrysanthou, H. Bertoni, "Variability of Sector Averaged Signals for UHF Propagation in Cities," IEEE Trans. Vehic. Tech., Vol. 39, No. 4, November 1990, pp. 352-358. H. Suzuki: "A Statistical Model for Urban Radio Propagation" IEEE Trans. Comm, Vol. COM-25, No. 7, July 1977, pp. 673-680.



A Full Scale Wireless Ad Hoc Network Test Bed

Timothy X Brown, Sheetakumar Doshi, Sushant Jadhav, Daniel Henkel, Roshan-George Thekkekunel
University of Colorado, Boulder, CO 80303
{timxb,doshi,sushant.jadhav,daniel.henkel,thekkeku}@colorado.edu

This paper describes a wireless mobile ad hoc network test bed developed at the University of Colorado, Boulder. The test bed is a framework on which one can run any ad hoc routing protocol implementation and collect performance statistics using benchmark tests. The results are reproducible and the performance statistics collected can be analyzed at a minute level. The test bed closes the gap between simulation and real life implementation and allows performance comparison of different ad hoc routing protocols on a common platform. In this paper we enumerate the design choices for the testbed and highlight its effectiveness by presenting an evaluation of the Dynamic Source Routing(DSR) protocol on the testbed .

1. Introduction

Most research in the area of ad hoc wireless networking has been conducted using simulation software. Current simulation software is known to have unrealistic hardware, propagation, interference, and mobility models. There is a significant gap between ad hoc protocol development and careful realistic studies of ad hoc network behavior.

Earlier efforts to address this gap include ad hoc network test beds that are bench top, indoor, fixed outdoor, and mobile outdoor. Bench top test beds employ MAC filtering, RF attenuators, or other emulation techniques to shrink the wireless range so that meaningful experiments can be performed within a single room [8][10][11][20][24]. These allow protocol development and testing in an easy to operate environment that is more realistic than simulation but does not capture all of the significant behaviors. Indoor test beds within a building provide more complex and realistic environments especially when the intended application is indoor [2][3][18]. These do not fully capture the mobility and propagation of the outdoor environment. Full scale outdoor test beds are often restricted to fixed sites [1][13][22]. These efforts provide insights into a full-scale outdoor environment, but ignore mobility. Outdoor mobile efforts include [14][15][16].

We envision the ideal test bed to be outdoor and to have the following features:

1. Test bed results are reproducible.
2. The test bed provides a common platform for testing different routing protocols.
3. Test scenarios are repeatable.
4. Testing is comprehensive in terms of test scenarios and traffic patterns.

We incorporate these features in the design of our ad hoc network test bed. Our test bed is designed to accommodate communication among arbitrary combinations of fixed nodes, mobile nodes on ground vehicles, and highly mobile nodes fixed on Unmanned Airborne Vehicles (UAVs).

In the following sections we describe our design choices for the components required to construct the test bed. We conclude the paper with the results obtained on evaluating the Dynamic Source Routing (DSR) protocol on the test bed.

2. Components of the testbed

The test bed consists of four basic elements: the ad hoc networking node components (hardware and software), the monitoring architecture, the database and graphical user interface (GUI) for storing and analyzing results, and the benchmark tests for evaluating the ad hoc routing protocols. Each of these is discussed in more detail in the following sections.

2.1 Nodes

Here we enumerate the design choices for constructing the ad hoc networking nodes. The main goal was to have a uniform network hardware and software that could be mounted in different types of nodes such as in Figure 1. First, we discuss the hardware choices, followed by the choices for the software running on the nodes, which includes the operating system and the routing infrastructure.

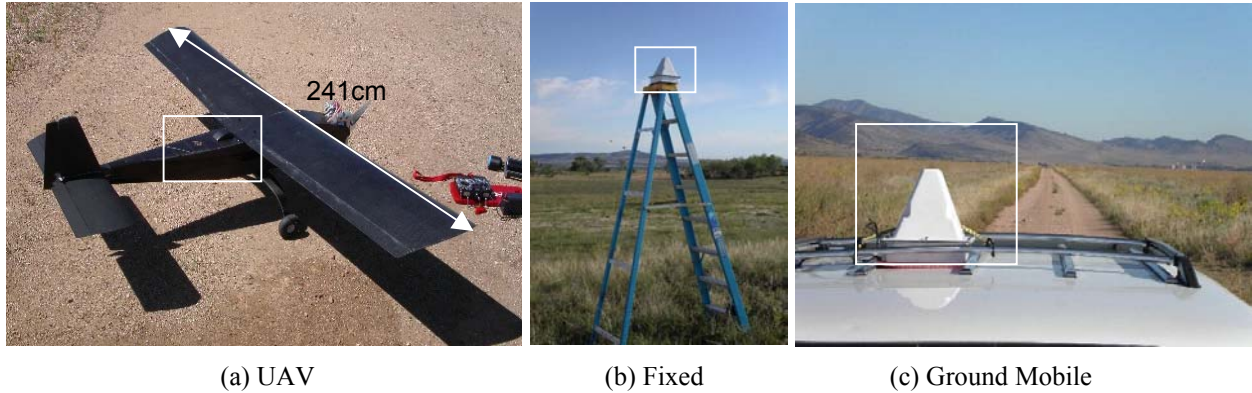


Figure 1. Heterogeneous Ad Hoc Radio Nodes

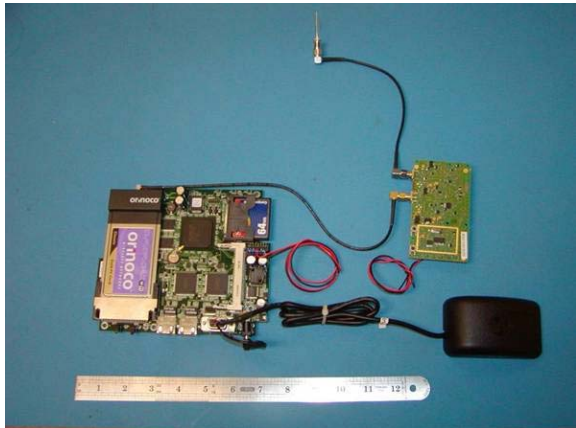


Figure 2: Node Hardware Components

2.1.1 Hardware

Due to the mobile outdoor environment, the ad hoc networking nodes must be tough, small and power efficient. Yet another design goal is that the nodes be built from low-cost commercial off the shelf (COTS) components, computers and electronics. The ad hoc networking node is a compact package that consists of computing hardware, a wireless radio interface, a bi-directional amplifier, and a GPS unit as shown in Figure 2. The nodes are packaged in environmental enclosures with an integrated antenna. These enclosures are mounted at fixed sites or on mobile nodes. Non-enclosed nodes are placed inside the UAVs. The fixed sites are connected to small lead-acid batteries, which power the node for seven to eight hours. The mobile nodes on the ground vehicles are powered off cigarette lighter sockets in the vehicles or batteries.

2.1.1.1 Computing Hardware

We chose the Soekris net4511 as the computing hardware for the ad hoc network nodes. The Soekris net4511 single board computer features a 486-class processor running at 100 MHz. It was selected since

it is commonly used in outdoor environments similar to the test bed. It has 64MB of RAM and a Compact Flash socket for flash memory storage up to 256MB. The Soekris computer can run a variety of operating systems, including OpenBSD, NetBSD, Linux, and a number of real-time operating systems. Sockets for PCMCIA and miniPCI cards make it suitable for test bed purposes. Two LEDs indicate power status and proper operation or failure. Two Ethernet ports as well as one RS232 serial interface allow for easy configuration and upgrade of the system.

2.1.1.2 Wireless Radio Interface

The ad hoc radio is based on the 802.11b MAC protocol because it is low-cost, has known behavior, is readily available and is commercial off the shelf. The channel rate was fixed at 2Mbps since this communicates over longer ranges than the higher rates while meeting a design target of one-hop throughputs in excess of 1 Mbps. We used the Orinoco 802.11b Gold cards with the PCMCIA form factor as the wireless radio interface.

2.1.1.3 Power Amplifier

The Orinoco cards have a maximum transmit power of 15 dBm (30 mW). This translates roughly into a range of 150m in an open area. In order to increase this working range, we employed a Fidelity Comtech RF amplifier package to amplify the signal from the Orinoco cards. The RF amplifier output power is adjustable from 20dBm (100mW) to 30dBm (1 W).

2.1.1.4 GPS Receiver

The role of the GPS receiver is to track the location of the node, as well as provide the current UTC time. This information is essential for the analysis of the routing statistics collected during the testing. We used a Garmin GPS 35 Smart Antenna for this purpose. The GPS receiver powers off the Soekris

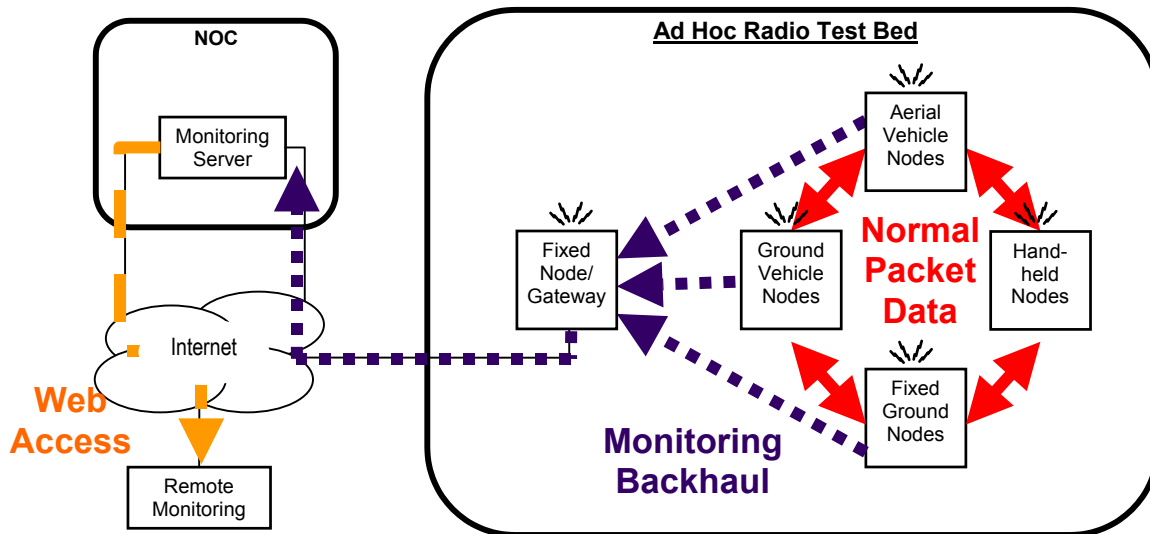


Figure 3. Normal data traffic (red solid) is monitored by each node. Periodically each node sends a report on the data (blue dotted) to the monitor server. This data can be viewed remotely over the Internet (yellow dashed) via a web-based GUI.

board and has a serial interface to the board. The GPS receiver sampling rate is 1 sample per second. The interfacing software extracts the location coordinates and the UTC time information from the NMEA stream of the GPS receiver.

2.1.2 Software

Here we enumerate the software choices for the ad hoc networking nodes. We first discuss the operating system for the nodes and then focus on the routing infrastructure employed for the ad hoc routing protocols.

2.1.2.1 Operating System

The aim behind selecting the operating system on the nodes is to make the system robust to power interruptions and other disturbances during operation. We chose a RAM based file system, so that no permanent state related to the working of the system is stored on the node. This allows the nodes to boot up cleanly in case of power failure. We chose the WISP-Dist distribution, a stripped down version of Linux whose size is 8MB and is suitable for Soekris board based routers.

2.1.2.2 Routing Infrastructure

We used the Click Modular Router as the common implementation infrastructure for the ad hoc routing protocols. Quite a few ad hoc routing protocol implementations have been done using Click e.g. the AODV implementation by Neufeld et. al. [17], the DSR implementation by Doshi et. al. [4] and the GRID project at MIT [21]. The effectiveness of Click as an implementation infrastructure for ad hoc

routing protocols has been proven in these studies. Neufeld et al. [17] from the Computer Science Department of the University of Colorado have developed a tool, so-called *nsclick*. This tool allows ad hoc routing protocols implemented on Click to be interfaced with the network simulator (ns2) tool so that the same Click implementation code can be used in simulating mobile ad hoc networking scenarios in ns2. This reinforced our decision to choose Click as the routing infrastructure.

The Click router software can be configured to run at the user level using a driver program or in the Linux kernel as a kernel module. When Click runs in the user level mode, it requires a kernel tap that captures packets that are destined to or from the kernel. This allows the packets to be manipulated in the user-space and also allows for the re-insertion of the packets into the kernel stack. When Click runs as a kernel module, it can steal intercept packets from the network devices before Linux gets an opportunity to handle them. It sends packets directly to the devices as well as to Linux for normal processing. The kernel module version of Click was used on the nodes. This gives higher performance because the router runs as a part of the Click kernel.

2.2 Monitoring

The monitoring mechanism is responsible for the collection of routing statistics and topology information of the ad hoc network and storing them so that the testing results can be replayed and analyzed. The monitoring must achieve several goals in order to be effective. The monitoring must provide sufficiently complete information to analyze network

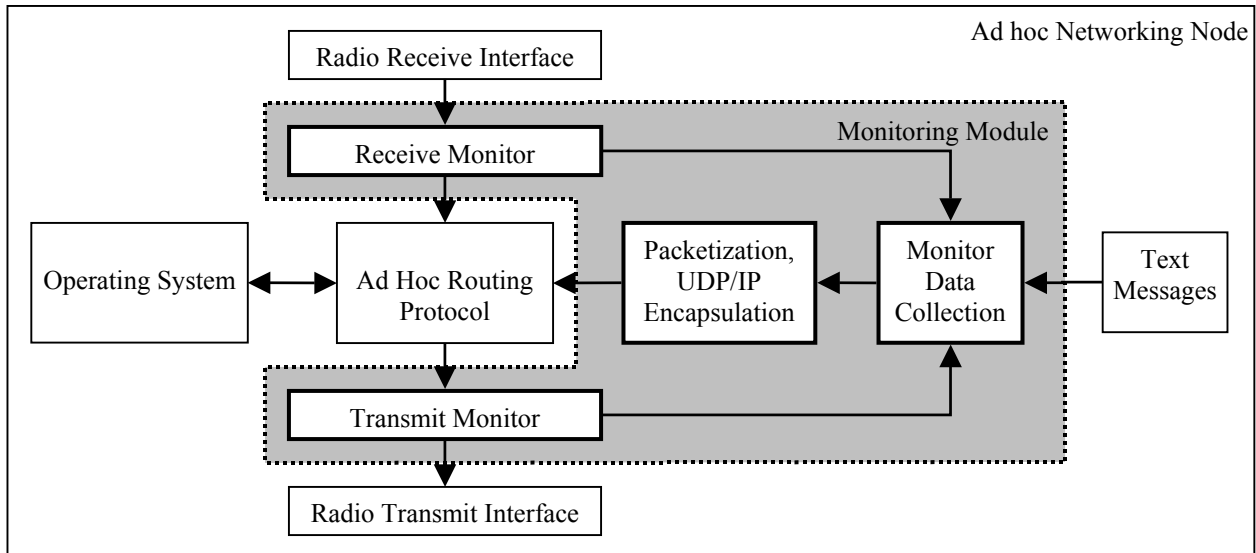


Figure 4. The monitor software (shaded) collects per packet data as packets pass in and out of the radio. This is periodically packetized and sent back to the monitor server.

behavior in detail. The test bed data should be available in real time to provide a situational awareness and feedback as testing progresses. The test bed should scale to 10's of monitored nodes. The monitoring should have minimal impact on the normal operation of the network. In reaching these goals, the monitoring must solve several challenges. The ad hoc networking nodes are subject to sudden power interruption and may shut off severing communication in the network. The ad hoc networking is complex with control distributed across the ad hoc nodes. Nodes may be disconnected for long periods of time during experimentation and the monitoring should be reliable to these disconnects. These constraints limit some approaches. The real time collection requirement precludes simply storing monitoring data on each node to be collected after the experiment. The distributed behavior suggests that data has to be centrally collected and correlated between nodes. The scaling and interference constraints imply that the monitoring should use minimal computing, storage, and bandwidth resources. The monitoring approach is shown in Figure 3. The pieces are described in the following sections.

2.2.1 Monitoring Architecture

Monitoring is done at a routing level. Running on each node is a monitoring process inserted into the radio packet processing as shown in Figure 4. The monitoring collects the packet statistics of time of arrival, type of packet, packet sequence number, and packet size. The type of packet field indicates if the packet has been received or sent to be transmitted and the transport layer type (UDP/TCP/ICMP). It also

logs information about the control packets of the ad hoc routing protocols. The monitoring also collects information about UTC time, latitude, longitude and altitude of the node's current position from the GPS attached to the serial port of the node. A unique feature of the monitoring is the inclusion of optional user defined text messages to annotate events during operation of the network. Test scripts running on the node can send messages to the monitoring module to be included in the monitoring information.

2.2.2 Packetization and Routing

The information collected by the monitoring module is packetized and a monitor sequence number is added to the packet that is unique per node. Packetization is triggered every 10 seconds or whenever the estimated packet size of the monitoring information equals 1000 bytes. The module now buffers the monitoring packet and a packet copy is passed on as an application layer packet to a module that adds to it a UDP/IP header with its destination as a fixed ad hoc networking node connected to the Internet through its wired interface, also known as the gateway node. This packet now is passed to the ad hoc routing protocol as a UDP/IP data packet and the ad hoc router finally routes this packet to the gateway node over the mesh network.

2.2.3 Reliable Delivery to the Gateway

The gateway node router receives the routed monitoring packet, strips off the routing headers if any and recognizes the packet as a monitoring packet. It then sends back a Monitoring ACK packet to the node that sourced the monitoring packet. The node on

receiving the ACK removes the corresponding monitoring packet from its buffer and is clear to transmit the next monitoring packet it has lined up in the buffer. If the node does not receive the ACK packet, it keeps retransmitting the same monitoring packet till it eventually gets an ACK from the gateway for that packet. Each retry occurs every 10 seconds. If the packet is buffered for over 1 hour, the packet is dropped and the next packet in the buffer is passed on for transmission.

2.2.4 Gateway to monitoring server

The gateway strips off the UDP/IP header of the received monitoring packet and adds a new UDP/IP header with the destination as the monitoring server, which is located on the University campus. It then forwards on the packet on its wired interface to the test-site router that routes the packet to the monitoring server through the Internet.

The test site has a T1 backhaul to the Internet. Alternate backhauls were considered if the test bed were to move to a test site without Internet connectivity. For this purpose, an Iridium satellite link was tested. A single Iridium phone has a nominal 2.4kbps data throughput. The data throughput can be increased by aggregating multiple phones using a Multi-Link Point-to-Point Protocol (MLPPP) connection. MLPPP is natively supported in Microsoft Windows XP. Using the Internet Connection Sharing feature in Windows any computer can be configured to act as a gateway, providing Internet access to any connected node on the Test Bed. The Internet service is through Iridium Satellite LLC, which provides access to their Internet gateway.

An Iridium phone connects to the computer through a serial port. Multiple Iridium phones are connected via serial cables to a four port serial to USB adapter. This allows for any laptop with one available USB port to control up to four Iridium phones for an MLPPP connection. The Iridium phone is Motorola model number 9505. The Motorola 9505 functions exactly like a standard 2400bps modem requiring only an additional initialization string. Windows can then use modems attached to each of the unique COM ports to create an MLPPP connection. Based on file transfer tests, the resulting four phone system supported an effective 6.7kbps data throughput. This was found to be sufficient for recording monitoring information.

2.3 Database and GUI

The function of the database is to store the routing and topology information collected by the monitoring mechanism. The GUI is responsible for

setting up queries for data analysis and presenting the test results to the user in a simple, easy to interpret format. We discuss each in turn.

2.3.1 Database

The monitoring server receives the monitoring packets from the gateway, parses them, and inserts them into a database. The database is both a data archive and an analysis tool. The database stores four types of data, per-packet data, per-node data, per-monitor-packet data and application messages. The per-packet data is the packet data recorded at each node. Note that a single data packet will appear several times in the database since it is transmitted and received by different nodes on its path across the network. Each entry is associated with the point on the path where it was recorded. This level of detail enables each packet to be tracked as it crosses the network and either its successful delivery at the destination or the point where it was lost can be determined. The per-node data is the GPS time and position data included in the monitor packet. The position of every node at every time during the experiment can be determined. In turn, the distance between any two nodes at any time can be determined. When combined with the per packet data, it allows packet losses to be correlated with node separations. The per-monitor-packet data records the sequence number of monitor packets received by the database, the time they were sent by a node, and the time they were received in the database. The application messages contain both free text and a numeric type to ease sorting and display. Examples include the start and stop times for experiments, the results of the experiments, and notification messages such as when a node powers up or the radio interface is turned off. By embedding this information in the database, the database becomes the complete archived repository of all test bed activities. This information is stored in a central ODBC-compliant relational database. We are currently using MySQL version 1.4.3 as our database engine since it is open source, freely available, and can be ported to many different platforms (www.mysql.com). The relational database enables complex queries for detailed network performance analysis.

2.3.2 GUI

The monitoring design also includes real time remote access and data visualization via a Web-based graphical user interface (GUI). A screen shot of the interface is shown in Figure 5. The GUI is a Java applet (version 1.4.2) using Sun's standard GUI library Swing to display and analyze network state and performance both post-test and real time. The

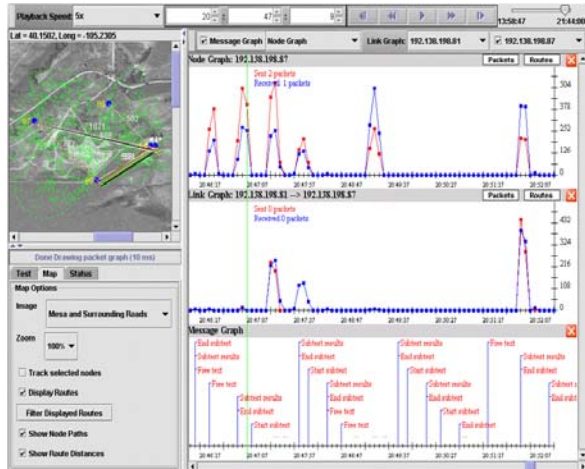


Figure 5. Screenshot from the remote monitoring GUI. Situation map is on the top left showing ad hoc radio positions, current routes, link lengths, and prior position tracks. Status messages and control panel are on the bottom left. Performance and message graphs are shown on right. Time control is at the top.

GUI shows the position of nodes and routes being used. Graphs versus time can be called up showing the traffic sent and received by a node; the traffic between two nodes, and the text messages in the database. All graphs share the same x-axis, and therefore the same time frame length and current position. This horizontal alignment of the graphs facilitates graph comparison. The traffic graph data can be filtered by the routes that packets take and by the packet types (TCP receive, TCP transmit, UDP receive, etc.).

The GUI serves several purposes; experimentation support, data dissemination, and data analysis. The experiments take place over a large area and situational awareness is limited. The GUI enables experimenters at the test bed site to observe node and traffic activity. For instance, when a radio and its GPS are properly functioning, they appear on a situational map in the GUI. Traffic and routing can be monitored during experiments for anomalies. By making the GUI Web-based, the data can be readily viewed by other observers and researchers. Finally, an ad hoc network has many simultaneous activities. The GUI provides a tool for comprehending the big picture and isolating specific events.

2.4 Benchmark Tests

The last component of our test bed is the benchmark test suite used to evaluate the ad hoc routing protocols running on the test bed. While measuring throughput and delay gives good quantitative values for network performance, day-to-

day applications stress the network in a different way, which reflects the ultimate test criteria for a network. We developed tests that measured throughput, latency, and congestion measures, and also designed tests to capture subjective impressions, namely web-browsing tests and voice quality tests. This suite of benchmark tests are:

- 1 Throughput: Purpose – to test the throughput that can be achieved with a TCP connection when no other traffic is present.
- 2 Latency: Purpose – to measure the ability for node pairs to send packets to each other when the network is lightly loaded.
- 3 Congestion: Purpose – to measure throughputs when there are competing data flows in the network.
- 4 Subjective: Purpose – to assess the performance of typical network applications as perceived by a user.

3. Test Bed Evaluation of DSR

This section summarizes the results obtained for the benchmarks tests on the test bed for the Dynamic Source Routing Protocol. We used a Click-based implementation of DSR [4] with modifications to include the monitoring components. The benchmark tests were performed for two types of scenarios: a fixed scenario and a mobile scenario. Other scenarios were tested which included using UAVs, but, for clarity these results are not included here.

3.1 Fixed Scenario

In this scenario all the mesh network nodes were fixed atop a ladder at a height of 7 feet (2m). They were arranged on the test bed to form a five-hop chain network.

Throughput Tests: These tests were conducted by running a network performance tool called Netperf from each node to every other node in the network. The results are shown below categorized by the number of hops in the route. Throughput degradation with increase in the number hops is a widely studied phenomenon in wireless networks. Processing delays at each node and bandwidth constraints are the two primary reasons for this. Processing delay is directly dependent on the protocol being used. Our results for the DSR protocol indicate that the throughput degrades by roughly a factor of two for every hop in the network. The measurement errors were small indicating good reproducibility.

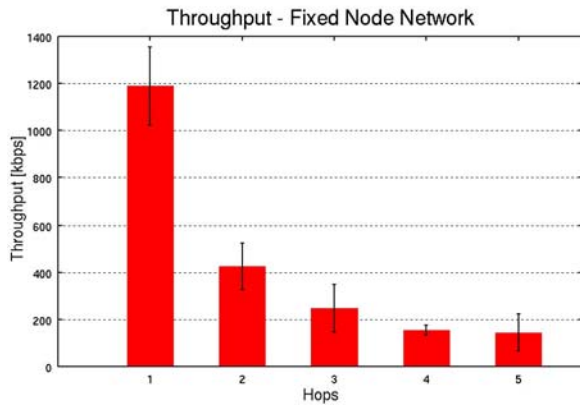


Figure 6. Fixed Scenario: Throughput vs. number of hops

Latency Tests: In this set of tests, for each source-destination pair at a time, 1 sec interval ping packets were sent from source to destination for 20 seconds. The results are presented below. The delays to nodes increase as the number of hops increases by about 13 msec per hop. The measurement error on the two-hop link clearly indicates an outlier event which raised the mean latency value. This outlier event can be attributed to the variability in the wireless environment. Again one sees high variability in the data for 4 hops. As the number of hops increases, there is a higher probability of variance in the measured data as it is direct function of the number of links.

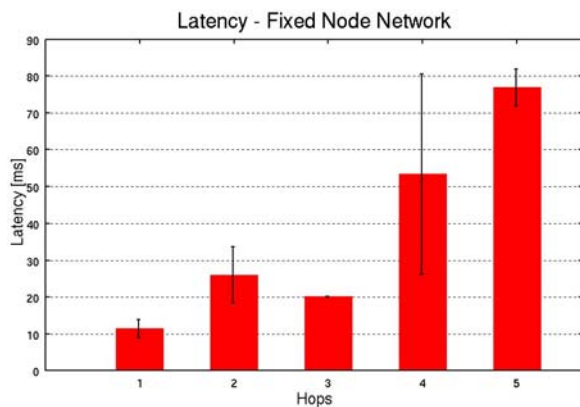


Figure 7. Fixed Scenario: Latency vs. Number of hops

Congestion Tests: In this test we start two simultaneous TCP flows between two random pair of nodes in the network using the Netperf tool. The flows are categorized into four groups:

1. The source-destination pairs were selected such that the flows did not interfere with each other (i.e. their routes had no common nodes). Neither

- were any of the source-destination pair nodes adjacent to each other.
2. The source-destination pairs were selected such that the flows did not interfere with each other. However at least one of the source-destination pair nodes were adjacent to each other.
3. The source-destination pairs were selected such that the flows were overlapping. Also, this group did not have any of the source-destination pair nodes that were identical.
4. The source-destination pairs were selected such that the flows were overlapping. The source-destination node pairs were selected such that at least two nodes were identical.

The results for these four groups are shown below. The results clearly indicate that interfering flows increase the packet loss as the flow overlap increases with a factor of four more losses between group 1 and group 4. Throughput was also measured. Overall the reduction in throughput was 25%. One observation was that the throughput losses were not uniform. One of the two competing flows was often much more affected than the other flow, even when the two flows had the same number of hops and relative interference. This suggests that the ad hoc routing combined with 802.11 is inherently unfair.

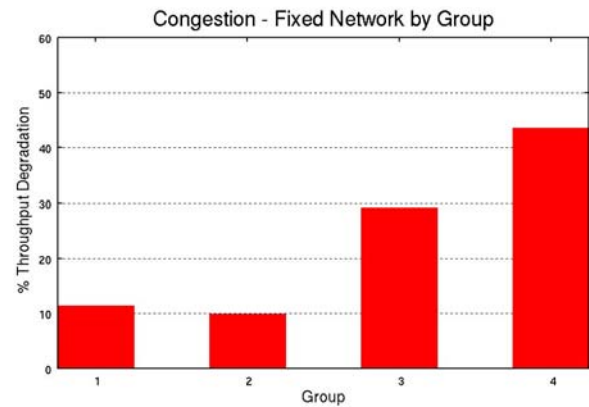


Figure 8. Fixed Scenario: Congestion Results

Subjective Tests:

We have designed two tests to capture subjective impressions: a web-browsing test and a voice quality test. For the web-browsing test we have a user chosen from the research group browse a website consisting of several pages with a different size image on each of them, namely 10kB, 100kB, 300kB and 500kB. The web pages are served by the small-footprint, single-threaded web-server Boa (www.boa.org) installed on the gateway minimizing impact on gateway performance. Candidates report their experiences browsing the pages compared to

browsing the Internet from their home connections. The voice quality test evaluates the subjective perception of a voice conversation carried out between two test candidates using laptops associated to one of the nodes in the test bed or the gateway. The open-source, Linux-based SIP-softphone Linphone (www.linphone.org) proved to be stable and user-friendly. It supports several voice codecs and enables adjustment of SIP and RTP parameters to compensate for changes in network performance.

With a well setup network of fixed nodes browsing, webpages from as far as six hops away could be compared to surfing the Internet on a fast dial-up connection. Picture rendering was more and more visibly slow with increasing hop count but was still acceptable.

Voice quality as tested from the gateway to a laptop moving around the test bed was found to be exceptionally good up to three hops and no noticeable end-to-end voice delay could be observed. New routes formed automatically as the laptop moved around the test bed and voice contact was re-established without having to re-dial or restart the phone application, although there were gaps in the speech when the laptop was not in range of any node. At a distance of four hops, voice streams became choppy and meaningful conversation was not possible anymore.

3.2 Mobile Scenario

The benchmark tests are repeated for the Mobile scenario. In this scenario we mounted two of the intermediate fixed nodes on top of vehicles. The vehicles were then driven at speeds of 20-30 miles per hour on designated paths. Care was taken that the paths of the vehicles were retraced in order to ensure identical test conditions for repeatability.

Throughput and Latency Tests:

The results for the throughput tests, latency tests and congestion tests are shown below. For the mobile scenario throughput and latency tests we grouped the routes into the following three categories:

1. Within Fixed nodes: routes where the source and destination nodes are fixed and there are no mobile nodes relaying packets in between
2. Mobile source-destination: routes where either the source or destination or both are mobile nodes.
3. Fixed nodes-mobile relay: routes where the source and destination nodes are fixed and they have mobile nodes acting as relays in between them.

The results below show that for both the throughput tests and the latency tests only the group 1 nodes reproduced the results of the fixed nodes scenario. The performance is severely degraded for nodes of the other two groups. The fixed nodes group was expected to be the most stable. In case of group 2 and there was always at least one link which was associated with a mobile node. Group 3 routes had at least two links which were mobile. The mobile nodes were not always in positions that could provide the necessary connectivity, hence the degradation in performance.

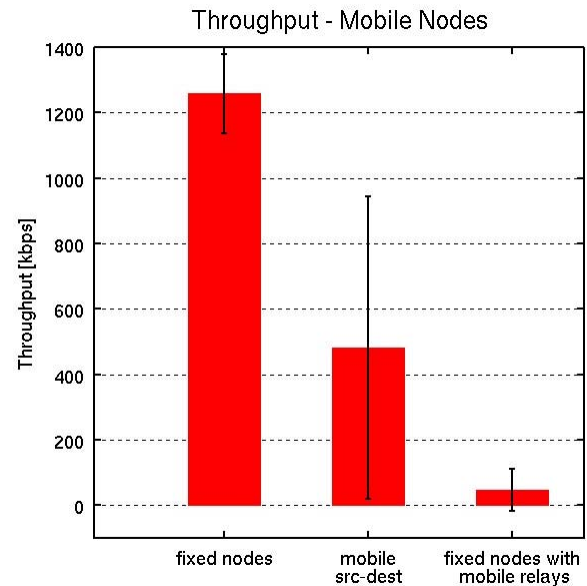


Figure 9. Mobile Scenario: Throughput Results

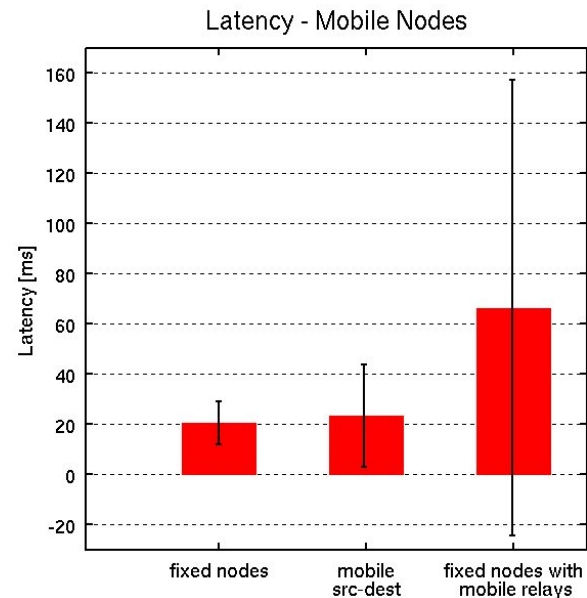


Figure 10. Mobile Scenario: Latency Results

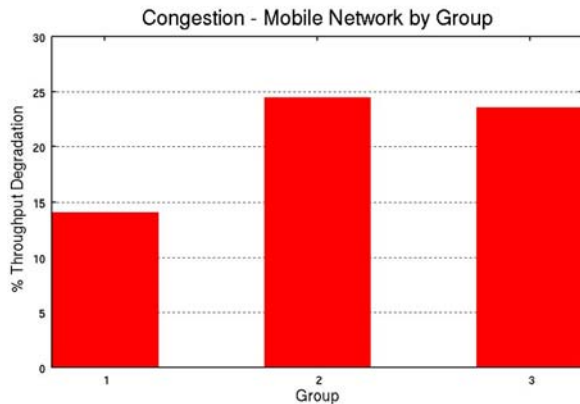


Figure 11. Mobile Scenario: Congestion Results

Congestion Tests:

The congestion test results are shown versus the same groups. The notion of interference was not relevant since the routing was dynamic over the testing. These results again indicate that the fixed nodes in group 1 have the best performance. Overall the congestion caused by two simultaneous TCP flows reduced the throughput by 20.7%.

Subjective Tests:

With a hybrid network of stationary and mobile nodes browsing became choppier as nodes moved out of reach and new routes to the web-server had to be discovered. Voice conversations were adversely affected due to mobility of nodes. Thus mobile scenarios did not work well for the subjective test applications of web browsing and voice. The users considered performance on par with a dial-up connection.

4. Conclusion & Future Work

The results in the previous section indicate the usefulness of the test bed for measuring performance of ad-hoc networks and ad-hoc routing protocols. For a test case we have used the DSR protocol. Definite numbers are obtained which can be used for comparative studies. In this case we have seen that for the given test setup the DSR protocol can give a max throughput of 1300 Kbps for 1 hop and around 100 kbps for 5 hops. These numbers suggest that using the DSR protocol for ad-hoc networks will limit us to using networks of diameter 5. Beyond this the throughput values would lie below 100 kbps which in turn would hamper real-time applications like video-streaming. The congestion tests illustrate that interfering wireless data flows is still a problem and the DSR protocol cannot circumvent this. Hence it is imperative to keep in mind that the baseline

results for fixed scenario throughput and latency are reproducible only under the condition of ideal node placement. Finally the subjective tests show us that web-browsing is achievable with tolerant delays approximately comparable to browsing on a dialup connection. The SIP protocol employed proved to be viable for establishing voice conversations in ad-hoc networks. However the end to end delays incurred due to this specific DSR implementation result in a maximum network diameter of 3 hops for a clear voice communication.

Another unique feature of our test bed was the mobile scenario. The very purpose of many ad-hoc network is to have mobile connectivity and this test bed allows us to evaluate performance in such scenarios. The performance of the DSR protocol was less than satisfactory for the mobile case. It had difficulty in routing through the highly mobile nodes. On going through the data obtained via our monitoring process we learnt that the minimum-hop route choice for the DSR protocol was chiefly responsible for the degradation of performance in the mobile scenarios. This metric would make the nodes send packets through weak transient links created during movement. Timing parameters were highly critical, since the network should react very quickly to failed links in mobile scenario. We observed that the protocol did not react quickly to broken links and this was adversely affecting the performance. Replaying the test results we found that there were instances when it took nearly 5s for the network to realize that a particular link was down. To address this we did further tests in the lab where we found two critical timing parameters that could be optimized: The DSR routing layer ACK mechanism was set for every 10th packet to save on processing overhead. This was reduced to every 3 packets. This significantly improved the reaction time of the network at the cost of a 10% degradation in throughput values. Also the timeout for packet retransmission was kept at 5s to account for RTT delays in heavily loaded links. However this was too conservative an estimate and we found that a RTT of 500ms would suffice. With these parameters changed we could obtain a network reaction time of roughly 2 seconds.

The scenarios all used the DSR routing protocol. One goal of the testbed is to use other protocols such as AODV. Integrating AODV into the test bed would require simply running the AODV router implemented on Click on the nodes instead of the DSR router.

The test bed enables concrete data to be collected in realistic but controlled conditions. In this paper we report on results with fixed and mobile scenarios. We have also used the test bed to

investigate networks with nodes mounted in UAVs and are developing further scenarios.

References

- [1] Chambers, B.A., "The Grid Roofnet: a Rooftop Ad Hoc Wireless Network," Master's Thesis, Massachusetts Institute of Technology, MA June 2002
- [2] Chin, K., Judge, J., Williams, A., Kermod, R., "Implementation Experience with MANET Routing Protocols", ACM SIGCOMM Computer Communications Review, 32(5):49-59, Nov. 2002.
- [3] Desilva, S., Das, S., "Experimental Evaluation of a Wireless Ad Hoc Network," Proceedings of the 9th Intl. Conf. on Computer Communications and Networks, Las Vegas, October 2000.
- [4] Doshi, S. Bhandare, S., Brown, T. X , "An On-demand minimum energy routing protocol for a wireless ad hoc network," Mobile Computing and Communications Review, vol. 6, no. 2, July 2002
- [5] Feeney, L., Nilsson, M., "Investigating the Energy Consumption of a Wireless Network Interface in an Ad Hoc Networking Environment," IEEE INFOCOM 2001
- [6] Gu, D.L., Pei, G., Ly, H., Gerla, M., Zhang, B., Hong, X., "UAV aided intelligent routing for ad-hoc wireless network in single-area theater," WCNC 2000 - IEEE Wireless Communications and Networking Conference, no. 1, September 2000, pp. 1220 – 1225
- [7] Heusse, M.; Rousseau, F.; Berger-Sabbatel, G.; Duda, A.; "Performance anomaly of 802.11b," INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications Societies. IEEE, Volume: 2, 30 March - 3 April 2003 Page(s): 836-843
- [8] Jin, Z., Liang, B., Shu, Y., Yang, O.W.W., "Designing and Implementing A Mobile Ad hoc Network Testbed," Proceedings of the IEEE Canadian Conference on Electrical and Computer Engineering, pp.1559-1564, Winnipeg, Manitoba, Canada, May 12-15, 2002
- [9] Johnson, D., Maltz, D., "Dynamic Source Routing in Ad Hoc Wireless Networks," Mobile Computing, Chapter 5, pp. 153-181, Kluwer Academic Publishers, 1996
- [10] Kaba, J.T., Raichle, D.R., "Testbed on a Desktop: Strategies and Techniques to Support Multi-hop MANET Routing Protocol Development," Proceedings of the 2nd ACM international symposium on Mobile ad hoc networking & computing 2001, Long Beach, CA
- [11] Ke, O., Maltz, D.A., Johnson, D.B., "Emulation of Multi-Hop Wireless Ad Hoc Networks," Proceedings of the Seventh International Workshop on Mobile Multimedia Communications (MOMUC 2000), IEEE Communications Society, Tokyo, Japan, October 2000
- [12] Kohler, E., Morris, R., Chen, B., Jannotti, J., Kaashoek, M.F., "The click modular router," ACM Transactions on Computer Systems, vol. 18, no. 3, pp. 263–297, August 2000. <http://www.pdos.lcs.mit.edu/click>
- [13] Lundgren, H., Lundberg, D., Nielsen, J., Nordstrom, E., Tschudin, C., "A Large-scale Testbed for Reproducible Ad hoc Protocol Evaluations," 3rd annual IEEE Wireless Communications and Networking Conference (WCNC 2002)
- [14] Maltz, D.A., Broch, J., Johnson, D.B., "Experiences Designing and Building a Multi-Hop Wireless Ad Hoc Network Testbed," CMU School of Computer Science Technical Report CMU-CS-99-116, March 1999
- [15] Maltz, D.A., Broch, J., Johnson, D.B., "Quantitative Lessons From a Full-Scale Multi-Hop Wireless Ad Hoc Network Testbed," Proceedings of the IEEE Wireless Communications and Networking Conference, IEEE, Chicago, September 2000
- [16] Morris, R. Jannotti, J., Kaashoek, F., Li, J., De Couto, D., "CarNet: A Scalable Ad Hoc Wireless Network System," 9th ACM SIGOPS European Workshop, Kolding, Denmark, September 2000
- [17] Neufeld, M. , Jain, A., Grunwald, D., "Nslclick: Bridging Network Simulation and Deployment," MSWiM 2002
- [18] Raychaudhuri, D., Seskar, I., Ott, M., Ganu, S., Ramachandran, K., Kremono, H., Siracusa, R., Liu H., Singh, M., "Overview of the ORBIT Radio Grid Testbed for Evaluation of Next-Generation Wireless Network Protocols," Proceedings of the IEEE Wireless and Networking Conference (WCNC 2005)
- [19] Royer, E., Toh, C., "A Review of Current Routing Protocols for Ad Hoc Mobile Wireless Networks," IEEE Personal Communications, April 1999, pp. 46–55
- [20] Sanghani, S., Brown, T.X, Bhandare, S., Doshi, S., "EWANT: The Emulated Wireless Ad Hoc Network Testbed," IEEE Wireless Communications and Networking Conference (WCNC), 16-20 March, 2003
- [21] The Grid Ad Hoc Networking Project at MIT, <http://www.pdos.lcs.mit.edu/grid/index.html>
- [22] Weber, S., Cahill, V., Clarke S., Haahr, M., "Wireless Ad Hoc Network for Dublin: A Large-Scale Ad Hoc Network Test-Bed," ERCIM News, vol. 54, 2003.
- [23] Xu, K., Hong, X., Gerla, M., Ly, H., Gu, D.L., "LANDMARK Routing In Large Wireless Battlefield Networks Using UAVs," Proceedings of IEEE Military Communications Conferences (MILCOM 2001), McLean, VA, October 2001.
- [24] Zhang, Y., Li, W., "An Integrated Environment for Testing Mobile Ad-Hoc Networks," Proceedings of 3rd ACM International Symposium on Mobile Ad-Hoc Networking and Computing (MobiHoc'02), pp. 104-111, Lausanne, Switzerland, Jun 2002

Radio Propagation Measurements During a Building Collapse: Applications for First Responders

Christopher L. Holloway, Galen Koepke, Dennis Camell, Kate A. Remley, and Dylan Williams

National Institute of Standards and Technology (NIST)
Electromagnetics Division, U.S. Department of Commerce, Boulder Laboratories
325Broadway, Boulder, Colorado 80305, email: holloway@boulder.nist.gov

Abstract

The National Institute of Standards and Technology is involved in a research project to improve wireless communications for first responders (firefighters and police) in large structures (i.e., large apartment and office buildings, supermarkets, sports stadiums, warehouses, convention centers, etc.). Part of this effort involves assessing communication problems in large-scale disaster situations (i.e., collapsed buildings). This work utilizes buildings that are scheduled for implosion. In this paper we present preliminary results of radio-propagation measurements obtained before, during, and after an apartment-building implosion.

INTRODUCTION

When first responders enter large structures (such as apartment and office buildings, sports stadiums, stores, malls, warehouses or convention centers) communication using portable radios to individuals on the outside of these large structures can be problematic [1]. Unreliable communications may occur due to decreased signal strength brought about by losses through structural materials. Reports published on the rescue efforts at the World Trade Center Towers [2, 3] highlighted this difficulty.

The National Institute of Standards and Technology (NIST) is investigating communications problems experienced by first-responders (firefighters and police) in disaster situations (i.e., collapsed buildings). In this effort we are investigating the propagation and coupling of radio-waves into and out of large structures. We are also investigating various schemes for improving detection of radio signals from firefighters and civilians who may have portable radios or cell phones and are trapped in voids in collapsed and partially collapsed buildings [4]. However, understanding propagation issues are the focus of the present paper.

Buildings scheduled for implosions provide the ideal research environment to investigate radio-wave propagation issues in collapsed buildings. We place portable radios similar to those used by first responders in various locations in the building. The radios are tuned to transmit at frequencies near public safety and cell phone bands (approximately 50 MHz, 150 MHz, 250 MHz, 400 MHz, 900 MHz, and 2 GHz). Once the radios are in the building, the building is imploded. We measure the received signals, before, during, and after the building is imploded.

This paper discusses one such set of experiments carried out in a 14-story apartment complex near New Orleans, LA (see Figure 1).



Figure 1: New Orleans apartment building.

EXPERIMENT

Two types of data were collected in the experiment. The first set of data, which is referred to as “radio mapping,” was collected a few days before the building was imploded. This involved carrying radios tuned to various frequencies through the building while recording the received signal from a site located outside the building. Figure 2 shows a typical radio

that was used. The radio was placed in a protective case to improve survivability after the implosion.

In the second type of data collection, radios were placed in fixed sites throughout the building. Received signals were collected before, during, and after the implosion. Our receiving sites in this case were both fixed and mobile. The mobile site consisted of a measurement system placed on a cart (see Figure 3). The cart was pulled around the perimeter of the building both before and after the implosion, enabling direct comparison of signal strength through the building and through rubble.



Figure 2: Typical Transmitter.



Figure 3: Mobile receiving cart.

Figures 4 and 5 show typical sets of data collected during the radio-mapping experiments (moving transmitter, fixed receivers). We see that propagation through the building can reduce the radio signal by as much as 50 dB, depending on the location of the transmitter.

Figures 6-8 show typical data collected before, during, and after the implosion. The implosion event can readily be seen in the figures. The results in each figure correspond to receivers at three different locations in the building. Note that at some locations the signal loss increased after the collapse, while at other locations the loss decreased. Figure 9 shows

results of signals from the mobile cart obtained before and after the implosion. After the implosion, the radios in this figure had most of the building lying on top of them. From the comparison, we see that the collapsed building caused between 50 to 70 dB reduction of the signals.

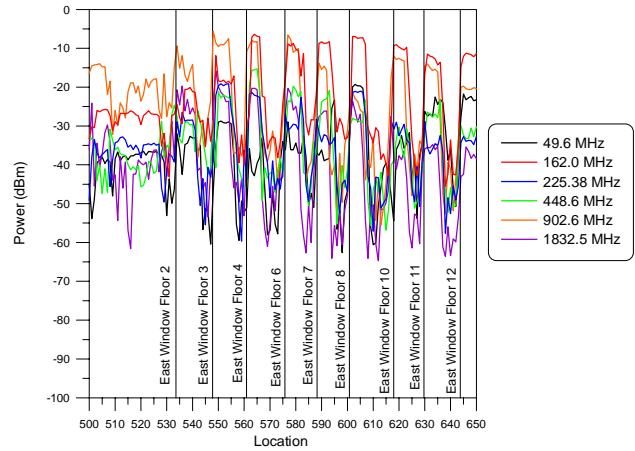


Figure 4: Typical radio-mapping result

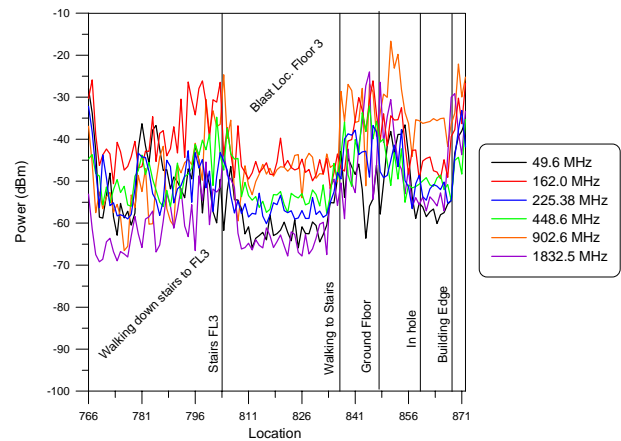


Figure 5: Typical radio-mapping result.

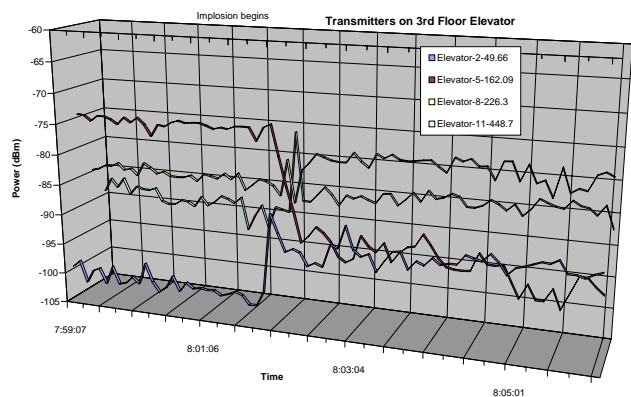


Figure 6: Radios located in elevator.

CONCLUSIONS AND DISCUSSION

In this paper we presented selected results of radio propagation data collected before, during and after the implosion of a 14-story apartment building near New Orleans, LA. The preliminary results of this experiment show that this type of building can reduce the radio signal by as much as 50 dB by just entering the building. Once the building collapsed, attenuation can increase much more than this. This type of data helps us understand the communication problems with which first responders are confronted when they enter large structures, and the changes in propagation that occur when a building collapses.

We have carried out similar sets of experiments during the implosion of a large sports stadium and a convention center. The initial findings in these data sets are very similar, that is, attenuation by as much as 70 dB may be encountered when communicating in these large structures. Details of these additional experiments will be published later.

Acknowledgements: This work was sponsored by the U.S. Department of Justice and the U.S. Department of Homeland Security through the Office of Law Enforcement Standards of NIST. We thank Jim Redyke of Dykon Explosive Demolition and John Angelina of D.H. Griffin Demolition Contractors for their assistance during this experiment. Without their help, this experiment would not have been possible.

REFERENCES

- [1] "Statement of Requirements: Background on Public Safety Wireless Communications," *The SAFECOM Program*, Department of Homeland Security, v.1.0, March 10, 2004.
- [2] "9/11 Commission Report", *National Commission on Terrorist Attacks Upon the United States*, 2004.
- [3] "Final report for September 11, 2001 New York World Trade Center terrorist attack.", *Wireless Emergency Response Team (WERT) WERT*, October 2001.
- [4] M. Rütshlin, K. A. Remley, R. T. Johnk, D. F. Williams, G. Koepke, and C. L. Holloway, A. MacFarlane, and M. Worrell, "Measurement of weak signals using a communications receiver system," *Proc. Intl. Symp. Advanced Radio Technology*, Boulder, CO, March 2005, accepted for publication.

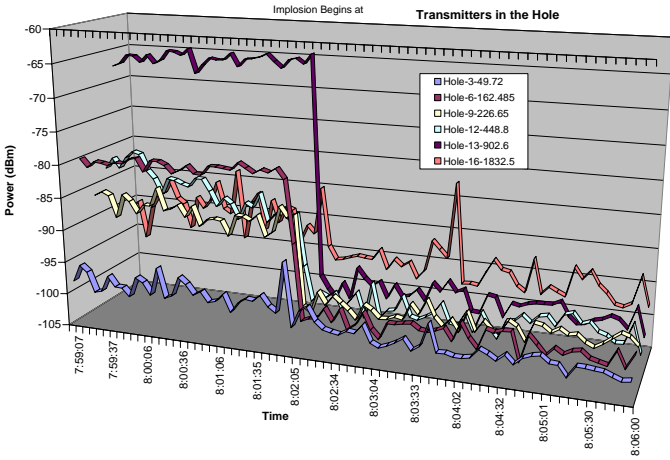


Figure 7: Radios located at bottom of building.

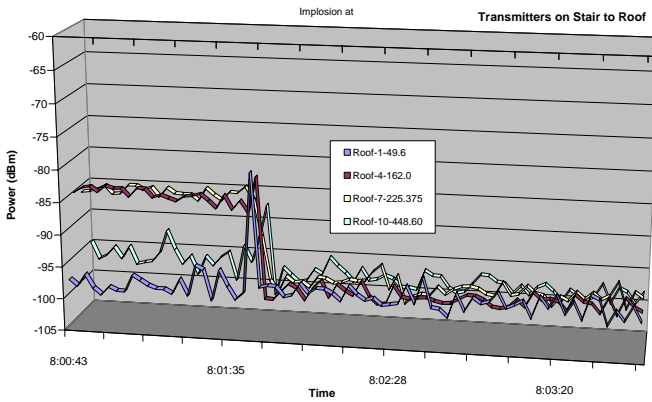


Figure 8: Radios located at top of building.

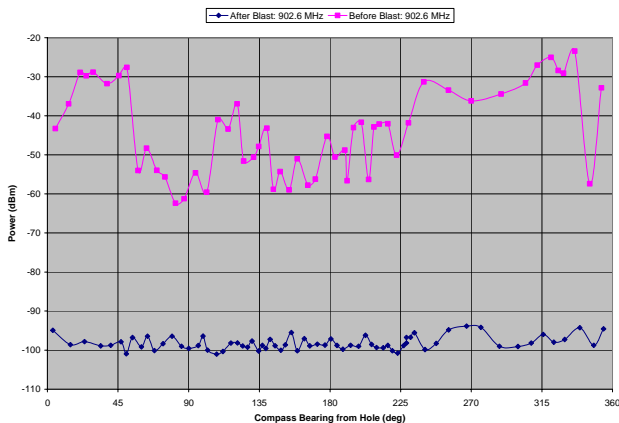


Figure 9: Mobile cart measurements: both before and after the implosion. Radios located at bottom of building.

Estimating the Demand for Voice over IP Services^{*}

Paul Rappoport
Temple University
Lester D. Taylor
University of Arizona
Donald Kridel
University of Missouri at St. Louis
James Alleman
University of Colorado¹
Contact author: 303 443-4465
James.Alleman@Colorado.edu

The demand for Voice-over-IP (VoIP) services is receiving increasing attention as an alternative to traditional switched access based telephone services. This paper focuses on the underlying determinants of demand to assess the potential for VoIP. The authors utilize a model of demand based on a consumer's willingness to pay and provide estimates of the elasticity of demand for VoIP. The intent of the paper is to add to the discussion of VoIP and to stimulate analysis.

I. Introduction

This focus in this paper is on what can seem the logical final chapter for consumers in the “convergence” of the computer and telecommunications industries, namely, the residential market for VoIP (Voice over Internet Protocol) services. Although at present this market is scarcely an infant, a number of companies (Vonage, AT&T, and Qwest, for example) foresee it as potentially very large, and are investing accordingly.² The purpose of the present exercise is to take a sober look the market for VoIP by using data on willingness-to-pay from a representative survey of U.S. households to provide estimates of the underlying price elasticity of demand for ‘best-effort’ VoIP services, as well as initial estimates of the size of the ‘best-effort’ VoIP market.³

^{*} This paper is forthcoming in *Teletronikk*; an updated version of the paper will be available from the authors after the ISART 2005 Conference.

¹ The authors thank Dale Kulp, president of Marketing Systems Group for access to the CENTRIS omnibus survey.

² For a comprehensive look at VoIP providers see <http://VoipWatch.com>

³ ‘Best-effort’ refers to VoIP plans that provide voice services over the Internet. This offering requires potential customers to have or be willing to have a broadband connection. ‘Primary line’ quality VoIP is provided by a service provider who owns or controls the infrastructure between the MTA (telephone enabled DOCSIS modem) and the gateway.

VoIP is a common term that refers to the different protocols that are used to transport real-time voice and the necessary signaling by means of Internet Protocol (IP). Simply put, VoIP allows the user to place a call over IP networks. “Best-effort” VoIP is the provisioning of voice services using broadband access (cable modem or DSL). It is referred to as ‘best-effort’ because service quality and performance cannot be guaranteed by the provider. The traditional voice telephony system meshes a series of hubs together using high-capacity links. When a call is placed, the network attempts to open a fixed circuit between the two endpoints. If the call can be completed, a circuit that stretches the entire length of the network between the two endpoints is then dedicated to that particular call, and cannot be used by another until the originating call is terminated.

The basic architectural difference between traditional telephony and IP telephony is that an IP network such as the Internet is inserted between the telephony endpoints, typically central offices. IP networks are packet-switched, as opposed to circuit-switched traditional telephony. Unlike circuit-switched networks, packets networks do not set up a fixed circuit before the call begins. Instead, the individual voice packets are sent through the IP network to the destination. Each packet may traverse an entirely different path through the network; however, the conversation is reassembled in the correct order before being passed on to the VoIP application. The “glue” that ties together the PSTN (Public Switched Telephone Network) with the IP network is known as an IP gateway. IP gateways perform many of the traditional telephone functions such as terminate (answer) a call, determine where the call is to be directed, and perform various administrative services such as user verification and billing before passing the call on to a receiving IP gateway. The receiving IP gateway, which may also be interconnected with the PSTN, dials the destination and completes the call.⁴

Pricing a new service is mostly a trial and error process, and the pricing of Best-Effort VoIP service is obviously no exception. Judging from the number of recent press releases, financial analyses, and articles written on VoIP, estimation of market size, consumer interest, and willingness-to-pay for VoIP services is a hot subject. A recent Goldman Sachs telecom services report notes that, as the VoIP threat evolves, it should not be viewed as catastrophic by the incumbent local exchange

⁴ http://www.cse.ohio-state.edu/~jain/cis788-99/ftp/voip_products/

carriers (ILECS).⁵ Business 2.0 published a story “Beware the VoIP Hype” in its December 9, 2003 issue describing a mismatch between expectations of investors and realities of the market. The author of that story noted that “... the big winners are likely to be the established companies that are already profitable and can afford to spend money on research and development and marketing. Most companies are not making money off the technology.”⁶

Before turning to technical details, it is useful to note just what it is that VoIP represents. Unlike some services that have emerged out of the electronic revolution, VoIP does not involve a new good *per se*, but rather a new way of providing an existing good at possibly lower cost and in a possibly more convenient manner.⁷ The good in question, of course, is real-time voice communication at a distance. The word *possibly* is to be emphasized, for voice communication is a mature good in a mature market, with characteristics that for all practical purposes are now those of a commodity. The ultimate potential market for VoIP, accordingly, is simply the size of the current voice market plus normal growth. Hence the evolution of VoIP, is pretty much strictly going to depend upon the efficiency *vis-à-vis* traditional telephony that VoIP vendors can provision this market.

To our knowledge, the present effort, which builds upon a previous study of the demand for broadband access using models of willingness-to-pay [Rappoport *et al.* (2003c)], is the first to focus on the modeling of VoIP services. The analysis in this paper makes use of data from an omnibus survey conducted in March and April, 2004, by the Marketing Systems Group of Ft. Washington, PA,⁸ in which respondents were asked questions concerning their willingness-to-pay (WTP) for VoIP services. In the study of broadband access just referred to, price elasticities for broadband access were developed using extensions of a generally overlooked procedure suggested by Cramer (1969). The same analyses have been used in this study. Among other things, price elasticities for VoIP are

obtained that range from an order of -0.50 for a fixed price of \$10 to -3.00 for a fixed price of \$70.

In addition to the range of elasticities just mentioned, the principal findings of the paper are:

- (1). Market drivers include the distribution of total telephone bills (local and long distance); the distribution of WTP and the distribution of broadband access to the Internet.
- (2). The market size for best-practice VoIP is small. For example, at a price of \$30, the estimated consumer market size is 2.7 million households.
- (3). Households with access to the Internet, especially with broadband access, have a higher willingness-to-pay for VoIP services.

The format of the paper is as follows. The next section begins with a short descriptive presentation of factors that underlie the demand for VoIP services. Section III provides the underlying theoretical framework that guides the analysis, while Section IV presents price elasticities for best-practice VoIP services derived from kernel-smoothed cumulative distributions of willingness-to-pay. Market-size simulations are presented in Section VII. Conclusions are given in Section VIII.

II. Descriptive Analysis

The analysis of the demand for VoIP services can be viewed as the conjunction of the three forces or factors. These include the distribution of total telephone bills; the probability that a household has or is interested in getting broadband access to the Internet; and the household's willingness to pay for VoIP service. Figure 1 displays the distribution of telephone bills (local and long-distance). Of interest here is the assumption that a household's interest in VoIP -- and hence willingness-to-pay -- depends on the household's total telecommunication expenditures. Thus, households with large telephone bills will presumably be more interested in VoIP than households with a smaller telephone bill. The fall-off in telephone expenditures after \$50 shown in this figure suggests that the potential size of VoIP may be limited by the number of households that have monthly telephone bills greater than \$50.

⁵ Goldman Sachs, Telecom Services: Wireline / Broadband Competitive Analysis, April 16, 2004

⁶ Business 2.0
<http://www.business2.com/b2/subscribers/articles/0,17863,534155-2,00.html>

⁷ Cellular telephone provides an apt contrast with VoIP, for, while cellular, too, represents an alternative way of providing real-time voice communication, it also allows for such to take place at times not available to traditional fixed-line telephony, hence in this sense is a genuine new good.

⁸ www.m-s-g.com

Fig. 1. Distribution of total telephone bill

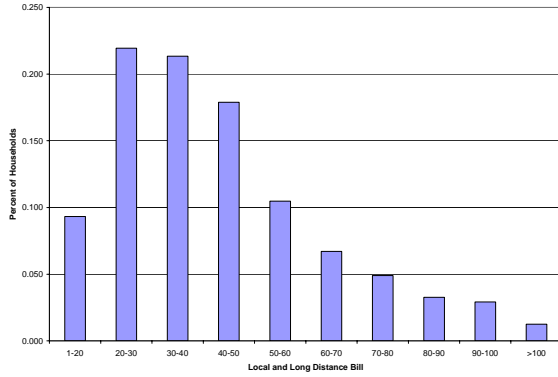


Figure 2 shows the distribution of willingness-to-pay for VoIP for households that already have broadband access to the Internet. Since these are the households that would seem to have the most potential for migrating to VoIP, the prospective size of the VoIP market suggested by this the numbers in this distribution would appear to be pretty modest. At a “price” of \$40 per month, the indicated size of market (as measured by the number of households with WTP greater than \$40) is seen to be about 2 million households, while at \$10 a month (which would almost certainly not be remunerative), the number is only 7 million.

Fig. 2. Distribution of Willingness-to-Pay for VoIP

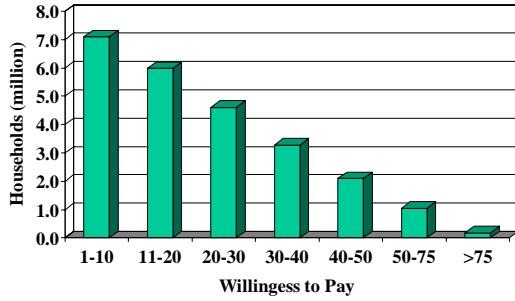
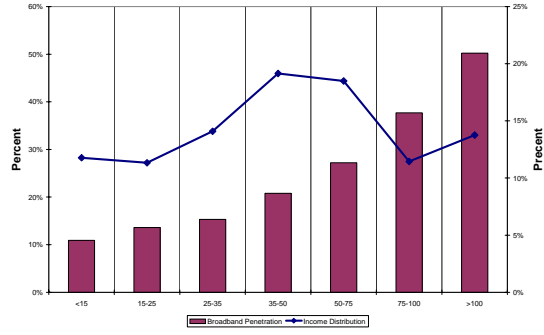


Figure 3 examines the relationship between the distribution of income (left scale) and the broadband penetration rate (right scale). Since broadband access is presumed to be a requirement for best-practice VoIP, the strong positive relationship that is indicated to hold between broadband penetration and income makes it clear that the distribution of income (especially the upper tail) is an important determinant of the potential VoIP market.⁹

⁹ The relationship between WTP and income will be examined in Section V below.

Fig 3. Demand for Broadband as a Function of Income



III. Theoretical Considerations

We begin with the usual access/usage framework for determining the demand for access to a network, whereby the demand for access is determined by the size of the consumer surplus from usage of the network in relation to the price of access.¹⁰ Accordingly, let q denote usage, and let $q(p,y)$ denote the demand for usage, conditional on a price of usage, p , and other variables (income, education, etc.), y . The consumer surplus (CS) from usage will then be given by

$$(1) \quad CS = \int q(z, y) dz .$$

Next, let \square denote the price of access. Access will then be demanded if

$$(2) \quad CS \geq \square$$

or equivalently (in logarithms) if

$$(3) \quad \ln CS \geq \ln \square$$

Alleman (1976, 1977) and Perl (1983) were among the first to apply this framework empirically. Perl did so by assuming a demand function of the form:¹¹

$$(4) \quad CS = \int_p^\infty A e^{-cp} y^\beta e^u dz ,$$

where y denotes income (or other variables) and u is a random error term with distribution $g(u)$. Consumer’s surplus, CS, will then be given by

¹⁰ See Chapter 2 of Taylor (1994).

¹¹ Since its introduction by Perl in 1978 in an earlier version of his 1983 paper, this function has been used extensively in the analysis of telecommunications access demand [see, e.g., Kridel (1988) and Taylor and Kridel (1990)]. The great attraction of this demand function is its nonlinearity in income and an ability to handle both zero and non-zero usage prices.

$$(5) \quad CS = \frac{Ae^{-\alpha p} y^\beta e^u}{\alpha} = \frac{P(WTP \geq \pi)}{1 - CDF(\pi)},$$

With net benefits from usage and the price of access expressed in logarithms, the condition for demanding access to the telephone network accordingly becomes:

$$(6) \quad P(\ln CS \geq \ln \pi) = P(a - \alpha p + \beta \ln y + u \geq \ln \pi) \\ = P(u \geq \ln \pi - a + \alpha p - \beta \ln y),$$

where $a = \ln(A/\alpha)$. The final step is to specify a probability law for consumer surplus, which, in view of the last line in equation (6), can be reduced to the specification for the distribution of u in the demand function for usage. An assumption that u is distributed normally leads to a standard probit model, while an assumption that u is logistic leads to a logit model. Empirical studies exemplifying both approaches abound in the literature.¹²

The standard procedure for estimating access demand can thus be seen in terms of obtaining information on the consumer surplus from usage by estimating a demand function, and then integrating beneath this demand function. In the present context, however, our procedure is essentially the reverse, for what we have by way of information are statements on the part of respondents in a survey as to the *most* that they would be willing-to-pay for a particular type of VoIP service. This *most* accordingly represents (at least in principle) the maximum price at which the respondent would purchase that type of service. Thus, for any particular price of VoIP, VoIP will be demanded for WTP's that are this value or greater, while VoIP will not be demanded for WTP's that are less than this value. Hence, implicit in the *distribution* of WTP's is an *aggregate demand function* (or more specifically, *penetration function*) for VoIP service. In particular, this function will be given by:

$$(7) \quad D(\pi) = \text{proportion of WTP's that are greater than or equal to } \pi$$

¹² Empirical studies employing the probit framework include Perl (1983) and Taylor and Kridel (1990), while studies using the logit framework include Bodnar *et al.* (1988) and Train, McFadden, and Ben-Akiva (1987). Most empirical studies of telecommunications access demand that employ a consumer-surplus framework focus on local usage, and accordingly ignore the net benefits arising from toll usage. Hausman, Tardiff, and Bellinfonte (1993) and Erikson, Kaserman, and Mayo (1998) represent exceptions.

where $CDF(\pi)$ denotes the cumulative distribution function of the WTP's. Once CDF's of WTP's are constructed, price elasticities can be obtained (without intervention of the demand function) via the formula (or empirical approximations thereof):

$$(8) \quad \text{Elasticity}(\pi) = \frac{d \ln CDF(\pi)}{d \ln \pi}.$$

IV. Data Employed in the Analysis

As noted, information on willingness-to-pay for VoIP service was collected from an omnibus national survey of about 8000 households in April and May, 2004, by the Marketing Systems Group (MSG) of Philadelphia. The omnibus survey, Centris¹³, is an ongoing random telephone survey of U. S. households. Each of the participants in the surveys utilized here was asked one (but not both) of the following two questions regarding their willingness-to-pay.

(a). What is the most you would be willing to pay on a monthly basis for a service that provides unlimited local and long distance calling using your computer?

(b). What is the most you would be willing to pay on a monthly basis for a service that provides unlimited local and long distance calling using your computer with internet connection at a cost of \$20 per month?¹⁴

The first question was asked of those households that currently have broadband access, while the second version was asked of those households that did not have broadband access.

V. Calculation of Price Elasticities

We now turn to the calculation of price elasticities in line with expression (8) above. The most straightforward way of doing this would be to define the elasticities as simple arc elasticities between selected adjacent points on the empirical CDF's. Unfortunately, however, because the survey-elicited WTP's tend to bunch at intervals that are multiples of 5 dollars, the values that emerge from this procedure are

¹³ www.Centris.com

¹⁴ \$20 was selected since dial-up prices were approximately \$20.

highly unstable, and accordingly of little practical use. To avoid this problem, elasticities are calculated using a kernel-based non-parametric procedure in which the “pileups” at intervals of 5 dollars are “smoothed out.”

Since kernel estimation may be seen as somewhat novel in this context, some background and motivation may be useful. The goal in kernel estimation is to develop a continuous approximation to an empirical frequency distribution that, among other things, can be used to assign density, in a statistical valid manner, in any small neighborhood of an observed frequency point. Since there is little reason to think that, in a large population, “pileups” of WTP’s at amounts divisible by \$5 reflect anything other than the convenience of nice round numbers, there is also little reason to think that the “true” density at WTP’s of \$51 or \$49 ought to be much different than the density at \$50. The intuitive way of dealing with this contingency (i.e., “pileups” at particular discrete points) is to tabulate frequencies within intervals, and then to calculate “density” as frequency within an interval divided by the length of the interval (i.e., as averages within intervals). However, in doing this, the “density” within any particular interval is calculated using only the observations within that interval, which is to say that if an interval in question (say) is from \$40 to \$45, then a WTP of \$46 (which is as “close” to \$45 as is \$44) will not be given weight in calculating the density for that interval. What kernel density estimation does is to allow *every* observation to have weight in the calculation of the density for *every* interval, but a weight that varies inversely with the “distance” that the observations lie from the center of the interval in question.

For the analytics involved, let $\hat{g}(x)$ represent the density function that is to be constructed for a random variable x (in our case, WTP) that varies from x_1 to x_n . For VoIP WTP, for example, the range x_1 to x_n would be 0 to \$700.¹⁵ Next, divide this range (called the ‘support’ in kernel estimation terminology) into k sub-intervals. The function $\hat{g}(x)$ is then constructed as:

$$(9) \hat{g}(x_i) = \sum_{j=1}^N \frac{K\left(\frac{x_i - x_j}{h}\right)}{Nh}, \quad i = 1, \dots, k.$$

¹⁵ This is for the households with broadband access. For those households without broadband access [i.e., for households responding to Question (b)], the range is from 0 to \$220.

In this expression, K denotes the kernel-weighting function, h represents a smoothing parameter, and N denotes the number of observations. For the case at hand, the density function in expression (9) has been constructed for each interval using the unit normal density function as the kernel weighting function and a ‘support’ of $k = 1000$ intervals.¹⁶

From the kernel density functions, VoIP price elasticities can be estimated using numerical analogues to expression (8).¹⁷ The resulting calculations, undertaken at WTP’s of \$70, \$60, \$50, \$40, \$30, \$20, and \$10 per month, are presented in Table 1.¹⁸ The estimated elasticities are seen to range from about -3.0 for WTP’s of \$60-70 to about -0.6 for WTP’s of \$10. Interestingly, the values in column 1 (for households that already have broadband access) for the most part mirror those in column 2 (which refer to households that do not). Since this appears to be the first effort to obtain estimates of price elasticities for VoIP, comparison of the numbers in Table 1 with existing estimates is obviously not possible. Nevertheless, it is of interest to note that the values that have been obtained are similar to existing econometric estimates for the demand for broadband access to the Internet.¹⁹ More will be said about this below.

¹⁶ Silverman’s rule-of-thumb,

$$\hat{h} = (0.9) \min[\text{std. dev.}, \text{interquartile range}/1.34](N^{-1/5}),$$

has been used for the smoothing parameter h . Two standard references for kernel density estimation are Silverman (1986) and Wand and Jones (1995). Ker and Goodwin (2000) provide an interesting practical application to the estimation of crop insurance rates.

¹⁷ The kernel-based elasticities are calculated as “arc” elasticities using points (at intervals of \pm \$5 around the value for which the elasticity is being calculated) on the kernel CDF’s via the formula:

$$\text{Elasticity}(x) = \frac{\Delta CDF(x) / CDF(x)}{\Delta WTP(x) / WTP(x)}.$$

Thus, for \$70, for example, the elasticity is calculated for x (on the kernel CDF’s) nearest to 75 and 65.

¹⁸ Since Question (b) postulates an access cost of \$20, the WTP’s for households without broadband access are assumed to be net of this \$20.

¹⁹ See, e.g., Rappoport, Taylor, Kridel, and Serad (1998), Kridel, Rappoport, and Taylor, (1999), Kridel, Rappoport, and Taylor (2002a), Rappoport, Taylor, and Kridel, (2002b),

Table 1
VoIP Elasticities Based on WTP
Kernel-Smoothed CDF

	With <u>WTP</u> <u>Broadband</u>	Without <u>Broadband</u>
\$70	-2.8616	-2.9556
60	-3.0217	-2.4730
50	-2.7794	-3.0093
40	-1.7626	-1.5630
30	-1.0753	-1.0527
20	-0.7298	-0.7564
10	-0.5454	-0.6025

VI. Modeling Willingness-to-Pay

An interesting question is whether willingness-to-pay, which in principle represents areas beneath demand curves, can in turn be “explained” in terms of the determinants of demand, that is, as functions of price, income, and other relevant factors. To explore this, we return to the expression for consumer surplus in equation (5), which we now express (in logarithms) as:

$$(10) \quad \ln CS = f(p, y, x, u),$$

where p , y , x , and u denote the price of usage, income, a variety of socio-demographic and other characteristics, and an unobservable error term, respectively. Since information on y and x is available (but obviously not vendor prices for VoIP) from the Centris survey, expression (10) can be estimated as a regression model with $\ln WTP$ as the dependent variable. However, before this can be done, the fact that some of the WTP’s are zero -- which creates an obvious problem in defining the dependent variable -- has to be dealt with.

Two solutions emerge as possibilities. The first solution is to use WTP as the dependent variable in place of $\ln WTP$ (in which case zeros are clearly not a problem), while the second solution is simply to eliminate all of the observations with zero WTP’s from the sample. We have opted for the second solution. However, we do this as part of a two-stage procedure, in which, in the first stage, a discrete-choice probit model is estimated that explains zero and non-zero values of WTP. The inverse of a “Mills ratio” is then constructed from this model and used as a “correction” term in a second-stage model, in which the logarithms of non-zero values of WTP are regressed on a set of

dummy variables representing income and various socio-demographic factors.²⁰ The “independent” variables that are available to the analysis include income (measured in terms income intervals); gender, age, education, and region of residence; satellite, cable, Internet, telephone, and cellular bills; forms of Internet access (including no access), and the number of cellular telephones.²¹ The statistical relationships should almost certainly be positive between WTP for VoIP and income and education, but probably negative for age.²² Obviously, a strong positive association should prevail between VoIP WTP and a household’s telephone bill.

The results for the first-stage probit models are tabulated in Tables 2 and 3. Table 2 refers to households that already have broadband Internet access in some form, while Table 3 refers to households that have either dial-up service or no Internet access at all. The results, though weak statistically,²³ are in keeping with the expectations just noted, especially with respect to the size of telephone bill. The effect of income is non-linear and positive, as is education, while age (especially for households that do not already have broadband access) is negative.²⁴ It is interesting that DSL and wireless Internet access, but not cable-modem access, are seen to be important for households that

²⁰ Although a value of zero is certainly a valid response to questions concerning willingness-to-pay, to put zero and non-zero values on the same footing in constructing CDF’s would seem to entail the assumption that penetration rates would be 100% at VoIP prices of zero. Obviously, this need not be the case. By specifying a first-stage model that explains the likelihood of a household having a non-zero WTP, and then incorporating this information as a “correction” in a second-stage model that explains the magnitude of (non-zero) WTP, penetration is thereby determined only with respect to those households that value VoIP positively.

²¹ Definitions of the variables are provided in the appendix.

²² We say this because of a finding in our earlier study of the demand for broadband access [Rappoport, et al. (2004)] of a strong negative relationship between age and the WTP for broadband access.

²³ That the results are generally weak statistically is evident in the paucity of p-values that are less than 0.10 and the fact that linear probability models (i.e., regression models with zero-one dependent variables) yield R^2 ’s that are of the order of 0.07. The strong importance of the dummy variable denoting whether the household was surveyed in April as opposed to March may well reflect the increased awareness of VoIP by households.

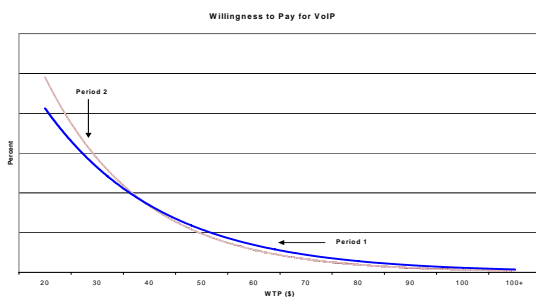
²⁴ For income, the left-out category is income less than 15 K, for education, the left-out category is less than high-school graduate, and for region, the left-out category is west.

Rappoport, Kridel, Taylor, Duffy-Deno, and Alleman (2003a) and Rappoport, Taylor and Kridel (2003b).

already have broadband access,²⁵ which suggests that households view VoIP, at this point anyway, primarily in terms of conventional telephony.

The first stage results suggest that the size of the telephone bill and whether a household has broadband access emerge as the key determinants of interest in (and willingness-to-pay) for VoIP telephony services. Since the effects of income and age are highly correlated with broadband demand²⁶ it is not surprising that these demographics enter into the first-stage equations. The dummy variable for April was included to test whether WTP changes when households have the ability to obtain more information on VoIP services. Other things equal, we would expect the WTP function to shift “clockwise” as households are exposed to more information on VoIP service: It is becoming harder to miss the Vonage banner adds! This shift is illustrated in Figure 4. Period 2 is the most recent period.

Figure 4: Shifts in the WTP Function Over Time



The estimated coefficients, standard errors, t-ratios, and p-values for the second-stage models are tabulated in Tables 4 - 6. The dependent variables in these models are the logarithms of the (non-zero) WTP's. The independent variables include all those that appeared in the first-stage models, plus the first-stage “Mills ratios” (LnMillsVoIP).²⁷ Tables 4 contains the second-stage

²⁵ Because broadband access is a prerequisite for VoIP, readers are cautioned to view these relationships as simply ones of association.

²⁶ See discussion in Rappoport, Taylor and Kridel, (1999), Kridel, Rappoport, and Taylor (2002a), Rappoport, Taylor, and Kridel, (2002a)

²⁷ The Mills ratio, it is to be noted, corrects for the fact that, because the second-stage model can be interpreted as the conditional expectation of WTP, given that WTP is greater than zero, the error-term in this model is “drawn” from a truncated distribution, and therefore does not have a mean of zero. The Mills ratios are accordingly calculated according to the formula $n(\pi_i)/N(\pi_i)$, where π_i denotes the predicted value (in the first-stage probit equation) of the probability that respondent i has a non-zero WTP and $n(\pi_i)$ and $N(\pi_i)$ represent the standard normal density and cumulative

model for households that already have broadband access, while Table 5 contains the same for the households that do not. An equation using a merged sample for the two groups is given in Table 6.

The results for the second-stage models, for the most part, parallel those for the first-stage models. Statistical significance is generally weak,²⁸ the effects of income and education are positive, and the effect of age is negative. VoIP WTP for both groups of households is positively related to telephone expenditures, but the effects, especially for households already with broadband access, are not as strong empirically as might have been expected. The Mills ratio is positive, but of little importance statistically.²⁹ The importance of the telephone bill and the internet bill is stronger for households with broadband access. This result is expected for “best-practice” VoIP, as we would expect households with broadband access to be better positioned to fully use VoIP services. The strongest results, quite clearly, are the ones for the merged sample in Table 6, especially regarding the effects of income, age, and telephone expenditures.³⁰

VII. The Potential Market for VoIP

As noted in the introduction, VoIP is not a new good per se, but rather provides a new way of supplying an existing good. Its success, consequently, is going to depend upon whether VoIP vendors can supply acceptable quality voice telephony at costs that are lower than those of the traditional carriers. VoIP is not a “killer-app” whose “explosion on the scene” will fuel a whole new industry. The voice telephony market is an old market, and the growth of VoIP is for the most part going to have to be at the expense of existing vendors.³¹ The purpose of this section is to take a sober

distribution functions of this event. For a derivation and discussion, see Chapter 6 of Maddala (1983).

²⁸ Reference here is to t-ratios and p-values for individual variables and coefficients. Low R-squares with cross-sectional survey data of the type being employed are normal.

²⁹ Unlike the usual procedure in two-stage models of this type, which is to include the Mills ratio as it stands, the logarithm is used instead. Use of the latter makes little difference with regard to fit and own significance, but does cause an increase in significance of several of the other independent variables.

³⁰ The format for this equation is to allow for the two subgroups of households (i.e., those already having broadband access and those without) to have separate intercepts and separate coefficients on type of Internet access. All other coefficients are constrained to be the same.

³¹ While the discussion here is couched in terms of new companies versus old, the argument is really with regard to technologies. If VoIP should in fact turn out to be superior in terms of quality and cost in relation to traditional circuit-

look at what VoIP vendors might accordingly reasonably expect as an initial potential market.

30

2,700,000

Despite the relatively weak statistical relationship that was found in the preceding section between a household's WTP for VoIP and the size of the household's telephone, rationally this has to be a key consideration, for we should not expect a household to demand VoIP unless doing so leads to a reduction in the cost of voice communication. Not unreasonably, therefore, the potential market for VoIP can be viewed as consisting of those households for which both its telephone bill and WTP for VoIP are greater than price of the service. Potential markets employing this criteria using information from the Centris survey have accordingly been constructed for three different VoIP prices, namely, 50, 40, and 30 dollars. The resulting households that are estimated to be candidates for demanding VoIP are presented in Table 7.

These numbers are small by any assessment, and stand in marked contrast with various estimates that have promulgated by industry analysts. There are approximately 23 million households with broadband access in the U. S., and to some, this is the size of the potential market for VoIP. However, if one simply looks at willingness-to-pay for VoIP that is greater than zero, the potential market drops to approximately 7 million households. A then closer look at the willingness-to-pay function suggests that at a price of \$30, the market size is less than 3 million households. There is room for some growth, especially if the number of households with broadband services grow. Nonetheless, the total size of the "best-practice" VoIP market is likely to remain at best modest.³²

Table 1
Potential Market for VoIP

<u>Price</u>	<u>Households</u>
\$50	810,000
40	2,170,000

switched technology, then existing telecommunication companies will have to adjust accordingly, which they almost certainly will do, rather than go the way of the dodo bird. The end result might be that traditional telcos simply transform themselves into full-scale internet service providers.

³² It is for these reasons that a number of large cable providers have opted to downplay best-practice VoIP in their telephony strategies and focus on providing basic telephony services over their IP-based network. Initial estimates for cable providers of the market size for IP-based telephone services are 20% of the current RBOCs market.

VIII. Conclusions

This paper has analyzed the consumer demand for best-practice VoIP service using information on willingness-to-pay for VoIP that was collected in early March and April, 2004, in an omnibus survey of some 8000 households. A theoretical framework has been utilized that identifies willingness-to-pay with consumer surplus from usage, which both allows for willingness-to-pay to be modeled as a function of income, education, and other socio-demographic factors, as well as the construction of a market demand function. The results of the exercise suggest that the demand for VoIP service is elastic (i.e., has an elasticity greater than 1 in absolute value) over the range of prices currently charged by VoIP service providers. The distribution of total telephone bills, the probability that a household has broadband access and the household's willingness to pay for VoIP service are used to simulate potential market size for various prices of VoIP service. In all simulations, the potential market size is estimated to be small.

Since the elasticities of the exercise are constructed from information elicited directly from households, and thus entail the use of contingent-valuation (CV) data, the seriousness (in light of the longstanding controversy surrounding the use of such data) with which our elasticities are to be taken might be open to question.³³ However, in our view, the values that we have obtained are indeed plausible and warrant serious consideration. Added credence for our results, it seems to us, is provided by the fact that, with VoIP service, we are dealing with a product (voice telephony) with which respondents are familiar and already demand, unlike in circumstances (such as in the valuation of a unique natural resource or the absence of a horrific accident) in which there is no generally meaningful market-based valuation can be devised.

It is interesting to note that our estimated elasticities suggest that at a price around \$30 demand shifts from inelastic to elastic. Vonage, the largest of the best-

³³ The critical literature on contingent valuation methods is large. See the NOAA Panel Report (1993), Smith (1993), Portnoy (1994), Hanneman (1994), Diamond and Hausman (1994), and McFadden (1994). On the other hand, particularly successful uses of CV data would seem to include Hammitt (1986) and Kridel (1988).

practice VoIP providers, recently announced a price reduction to \$30 for all new and existing customers.³⁴

References

Alleman, James. (1976), *The Demand for Local Telephone Service*, US Department of Commerce, Office of Telecommunications, OT Report, 76-24.

Alleman, James. (1977), *The Pricing of Local Telephone Service*, US Department of Commerce, Office of Telecommunications, OT Special Report, 77-14, pp. i-iv, 1-183.

Andersson, K. and Myrvold, O. (2002), "Residential Demand for 'Multipurpose Broadband Access': Evidence from a Norwegian VDSL Trial," *Telektronik*, Vol. 96, No. 2, pp. 20-25.

Andersson, K, Fjell, K., and Foros, O. (2003), "Are TV-viewers and surfers different breeds? Broadband demand and asymmetric cross-price effects," paper presented at the Norwegian Annual Conference in Economics, Bergen, 2003, Telenor R&D, 1331 Fornebu, Norway.

Bodnar, J., Dilworth, P., and Iacono, S. (1988), "Cross-Section Analysis of Residential Telephone Subscription in Canada," *Information Economics and Policy*, Vol. 3, No. 4, pp. 311-331.

Cramer, J.S.(1969), *Empirical Econometrics*, Elsevier Publishing Co., New York.

Diamond, P.A. and Hausman, J.A. (1994), "Contingent Valuation: Is Some Number Better Than No Number?," *Journal of Economic Perspectives*, Volume 8, No. 4, Fall 1994, pp. 45-64.

Erikson, R.C., and Kaserman, D.L., and Mayo, J.W. (1998), "Targeted and Untargeted Subsidy Schemes: Evidence from Post-Divestiture Efforts to Promote Universal Service," *Journal of Law and Economics*, Vol. 41, October 1998, pp. 477-502.

Hammit, J.K. (1986), "Estimating Consumer Willingness to Pay to Reduce Food Borne Risk," Report R-3447-EPA, The RAND Corporation.

Hannemann, W.M. (1994), "Valuing the Environment though Contingent Valuation," *Journal of Economic Perspectives*, Volume 8, No. 4, Fall 1994, pp. 19-44.

Hausman, J.A., Sidak, J.G., and Singer, H.J. (2001), "Cable Modems and DSL: Broadband Internet Access for Residential Customers," *American Economic*

Review Papers and Proceedings, Vol. 91, No. 2, May 2001, pp. 302-307.

Hausman, J.A., Tardiff, T.J., and Bellinfont, A. (1993), "The Effects of the Breakup of AT&T on Telephone Penetration in the United States," *American Economic Review Papers*, Vol. 83, No. 2, May 1993, pp. 178-184.

Ker, A.P. and Goodwin, B.K. (2000), "Nonparametric Estimation of Crop Insurance and Rates Revisited," *American Journal of Agricultural Economics*, Vol. 83, May 2000, pp. 463-478.

Kridel, D.J. (1988), "A Consumer Surplus Approach to Predicting Extended Area Service (EAS) Development and Stimulation Rates," *Information Economics and Policy*, Vol. 3, No. 4, pp. 379-390.

Kridel, D.J., Rappoport, P.N., and Taylor, L.D. (2001), "An Econometric Model of the Demand for Access to the Internet by Cable Modem," *Forecasting the Internet: Understanding the Explosive Growth of Data Communications*, ed. by D.G. Loomis and L.D. Taylor, Kluwer Academic Publishers.

National Oceanographic and Atmospheric Administration (NOAA;1993), 58, *Federal Register*, 4601, January 15, 1993.

Maddala, G.S. (1969), *Limited -Dependent and Qualitative Variables in Econometrics*, Cambridge University Press.

McFadden, D. (1994), "Contingent Valuation and Social Choice," *American Journal of Agricultural Economics*, Vol. 76, November 1994, pp. 695-707.

Perl, L.J. (1978), "Economic and Demographic Determinants for Basic Telephone Service," National Economic Research Associates, White Plains, NY, March 28, 1978.

Perl, L.J. (1983), "Residential Demand for Telephone Service 1983", prepared for the Central Service Organization of the Bell Operating Companies, Inc., National Economic Research Associates, White Plains, NY, December 1983.

Portnoy, P.R. (1994), "The Contingent Valuation Debate: Why Economists Should Care," *Journal of Economic Perspectives*, Volume 8, No. 4, Fall 1994, pp. 3-18.

Rappoport, P.N., Taylor, L.D., and Kridel, D.J. (1999), "An Econometric Study of The Demand for Access to The Internet," in *The Future of The Telecommunications Industry: Forecasting and Demand Analysis*, ed. by D.G. Loomis and L.D. Taylor, Kluwer Academic Publishers, Dordrecht.

³⁴ See http://www.citi.columbia.edu/voip_agenda.htm. See also <http://www.vonage.com>

- Rappoport, P.N., Taylor, L.D., and Kridel, D.J. (2002a), "The Demand for High-Speed Access to the Internet," in *Forecasting The Internet: Understanding the Explosive Growth of Data Communications*, ed. by D.G. Loomis and L.D. Taylor, Kluwer Academic Publishers, Dordrecht.
- Rappoport, P.N., Taylor, L.D., and Kridel, D.J. (2002b), "The Demand for Broadband: Access, Content, and The Value of Time," in *Broadband: Should We Regulate High-Speed Internet Access*, ed. by R. W. Crandall and J. H. Alleman, AEI-Brookings Joint Center for Regulatory Studies, Washington, D.C., 2002.
- Rappoport, P.N., Taylor, L.D., and Kridel, D.J. (2004), "Willingness-to-Pay and the Demand for Broadband Access," in *Down to the Wire: Studies in the Diffusion and Regulation of Telecommunications Technologies*, ed. by Allan Shampine, Nova Science Publishers.
- Rappoport, P.N., Kridel, D.J., Taylor, Duffy-Deno, K., and Alleman, J. (2003), "Forecasting The Demand for Internet Services," in *The International Handbook of Telecommunications Economics: Volume II*, ed. by G. Madden, Edward Elgar Publishing Co., London.
- Rappoport, P.N., Taylor, L.D., Kridel, D.J., and Serad, W. (1998), "The Demand for Internet and On-Line Access," in *Telecommunications Transformation: Technology, Strategy and Policy*, ed. by E. Bohlin and S.L. Levin, IOS Press.
- Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*, Monographs on Statistics and Applied Probability 26, Chapman and Hall, London.
- Smith, V.K. (1993), "Non-Market Valuation of Natural Resources: An Interpretive Appraisal," *Land Economics*, Vol. 69, No. 1, February 1993, pp. 1-26.
- Taylor, L.D. (1994), *Telecommunications Demand in Theory and Practice*, Kluwer Academic Publishers, Dordrecht.
- Taylor, L.D. and Kridel, D.J. (1990), "Residential Demand for Access to the Telephone Network," in *Telecommunications Demand Modeling*, ed. by A. de Fontenay, M.H. Shugard, and D.S. Sibley, North Holland Publishing Co., Amsterdam.
- Train, K.E., McFadden, D.L, and Ben-Akiva, M. (1987), "The Demand for Local Telephone Service: A Fully Discrete Model of Residential Calling Patterns and Service Choices," *The Rand Journal of Economics*, Vol. 18, No. 1, Spring 1987, pp. 109-123.
- Varian, H.R. (2002), "Demand for Bandwidth: Evidence from the INDEX Project," in *Broadband: Should We Regulate High-Speed Internet Access*, ed. by R. W. Crandall and J. H. Alleman, AEI-Brookings Joint Center for Regulatory Studies, Washington, D.C., 2002.
- Wand, M.P. and Jones, M.C. (1995), *Kernel Smoothing*, Monographs on Statistics and Applied Probability 60, Chapman and Hall, London.

A high grade secure VoIP using the TEA Encryption Algorithm

Ashraf D. Elbayoumy, Simon J. Shepherd
Advanced Signals Laboratory
School of Engineering Design & Technology
University of Bradford, BD7 1DP, UK
ademahmo, S.J.Shepherd @bradford.ac.uk

Abstract: *Unless a VoIP network is encrypted, anyone with physical access to the office LAN can potentially connect network-monitoring tools and tap into telephone conversations. Unfortunately, the price of this security is a decisive drop in QoS caused by a number of factors. In this paper we present the results of the experimental analysis of the transmission of voice over secure communication links. We present an efficient solution for securing VoIP using the TEA Encryption Algorithm.*

Key words: *VoIP, QoS, TEA.*

1. INTRODUCTION

The Internet community agrees that security is one of the key properties that should characterize any ICT (Information and Communication Technology) system and application, with particular emphasis on those that rely on the Internet for their nature, e.g., e-commerce. Unfortunately, security does not come for free and, in general, security and efficiency are conflicting requirements. Confidentiality, integrity and authentication can slow down packet transmission, which may not be acceptable by the application itself. Various aspects have to be considered in order to address the problem of real-time transmission over secure channels. The real-time nature of the problem poses some constraints.

In the case of voice transmission, the maximum acceptable delay in packet delivery for optimal voice quality is 150ms, which can be extended up to 200ms in case of encrypted communications. Thus, in a standard VoIP application, after the signal has been digitized, there are 150ms to code the signal using some standard scheme such as ITU standards G.723, G.729...etc, divide it into packets and encapsulate the packets into IP packets, then route the packets on the Internet, and reconstruct the original traffic stream at the destination, where it usually is buffered in order to smooth the jitter. Because of such a timing constraints, voice packets are small (10-50 bytes long payload) in order to guarantee that all above mentioned operations can be performed within the given time constraint.

Many factors determine voice quality, including the choice of codec, echo control, packet loss, delay, delay variation (jitter), and the design of the network. Packet loss causes voice clipping and skips. Some codec algorithms can correct for some lost voice packets. Typically, only a single packet can be lost during a short period for the codec correction algorithms to be effective. If the end-to-end delay becomes too long, the conversation begins to sound like two parties talking on a Citizens Band radio. A buffer in the receiving device always compensates for jitter (delay variation). If the delay variation exceeds the size of the jitter buffer, there will be buffer overruns at the receiving end, with the same effect as packet loss anywhere else in the transmission path.

As with traditional telephony, eavesdropping is a concern for organizations using VoIP—and the consequences can be huge. Because voice travels in packets over the data network,

hackers can use data sniffing and other hacking tools to identify, modify, store and play back voice traffic traversing the network. A hacker breaking into a VoIP data stream has access to a lot more calls than he would with traditional telephone tapping. As a result, one of the big differences is that a hacker has a much higher probability of getting intelligent information from tapping a VoIP data stream than from monitoring traditional phone systems. It might be a good idea to encrypt VoIP traffic flowing internally over a corporate network to prevent insider attacks.

There were many experiments to measure the effect of encryption and decryption on throughput [1][2]. Their results showed that the computationally lighter algorithms achieved better throughput than the more expensive ones.

The Tiny Encryption Algorithm is one of the fastest and most efficient cryptographic algorithms in existence. It was developed by David Wheeler and Roger Needham at the Computer Laboratory of Cambridge University. It is a Feistel cipher, which uses operations from mixed (orthogonal) algebraic groups - XOR, ADD and SHIFT in this case. This is a very clever way of providing Shannon's twin properties of *diffusion* and *confusion*, which are necessary for a secure block cipher, without the explicit need for P-boxes and S-boxes respectively. It encrypts 64 data bits at a time using a 128-bit key.

In this paper we present the results of the experimental analysis of the transmission of voice over secure communication links. We introduce the TEA encryption algorithm as a faster and powerful algorithm, which gives us the best compromise between security and efficiency. Our results show that using TEA encryption algorithm saves about 20% of the end-to-end delay compared to all available encryption algorithms.

This paper is organized as follows; Section 2 presents a quick overview of VoIP. Section 3 presents the Tiny Encryption Algorithm and summaries its security. Section 4 describes the testbed used for our experiments. Section 5 presents the experimental results of using TEA for securing VoIP. Section 6 concludes the paper and summarizes our findings.

2. VOICE OVER IP

In recent years, we have witnessed a growing interest in the transmission of voice using the packet-based protocols. Voice over Internet protocol (VoIP) is a rapidly growing technology that enables the transport of voice over data networks such as the public Internet. The following steps are performed:

- Digitization of the analog signal;
- Packet generation of the digital signal according to the TCP-UDP/IP protocols;
- Transmission of the packets on the network;
- Packet reception and analog signal reconstruction at the destination.

When sending voice traffic over IP networks, a number of factors contribute to overall voice quality as perceived by an end user. Some of the most important factors are end-to-end delay in the voice carrier path and degraded voice quality. Among the factors that degrade voice quality are packet loss, delay variation, or jitter, voice compression schemes (CODECS), echo cancellation algorithms. In this paper we focus on end-to-end delay and packet loss. Various factors influence signal delay during a VoIP transmission. The time spent by the CODEC, the device that performs the digitization process, may vary between 0.75-30ms, depending on the coding schemes adopted and the quality of the reproduced signal. The queuing delay (i.e., the time spent by a packet in the router buffers waiting for being routed) may add up to 30 ms. A further delay in the range of 40-70ms, called jitter delay, is introduced by buffering arriving packets so that they can be delivered at a uniform rate.

Table 1 reports the number of phone calls (using VoIP) that can be performed with up to date technology given channels with different bandwidth and different payload per packet.

TABLE 1. Number of telephone calls and average delay in ms as a function of channel bandwidth (B/W, from 32Kbps to 10 Mbps) and payload size (10,20, and 40 bytes).

	Payload Size					
	10		20		40	
B/W	#calls	delay	#calls	delay	#calls	delay
32	0	-	0	-	1	>200
64	0	-	1	100-150	2	150-200
128	1	<100	2	~100	4	150-200
256	2	<100	5	~100	9	~150
512	5	<100	10	<100	18	100-150
1024	11	<100	20	<100	36	100-150
10240	117	<100	214	<100	365	~100

Table 2 reports the main characteristic in terms of bit rate (in Kbps), compression delay (in ms) and Mean Opinion Score (MOS) for a set of algorithms that can be adopted by CODECS [2]. The MOS is a parameter used to measure the quality of the signal reproduced by such algorithms and ranges from 1 to 5, 1 being the worst case. As the table shows, the best algorithm is significantly better than the rest, which obtain a MOS in the surrounding of 3.5.

TABLE 2: CODEC algorithms and their characteristic bit rate in Kbps, compression delay in ms, and MOS.

Compression Algorithm	Bit Rate [Kbps]	Delay [ms]	MOS
G.711 PCM	64	0.75	4.1
G.726 ADPCM	32	1	3.85
G.728 LD-CELP	16	≤ 5	3.61
G.729 CS-ACELP	8	10	3.92
G.729a CS-ACELP	8	10	3.7
G.723.1 MP-MLQ	6.3	30	3.9
G.723.1 ACELP	5.3	30	3.65

Quality of Service is fundamental to the operation of a VoIP network. Despite all the money VoIP can save users and the network elegance it provides, if it cannot deliver at least the same quality of call setup and voice relay functionality and voice quality as a traditional telephone network, then it will provide little added value. Unfortunately, the implementation of various security measures can degrade QoS.

These complications range from delaying or blocking of call setups by firewalls to encryption-produced latency and delay variation (jitter). QoS issues are central to VoIP security. If QoS were assured, then most of the same security measures currently implemented in today's data networks could be used in VoIP networks. But because of the time-critical nature of VoIP, and its low tolerance for disruption and packet loss, many security measures implemented in traditional data networks just aren't applicable to VoIP in their current form.

3. THE TEA ENCRYPTION ALGORITHM

As shown in Fig. 1 it is a Feistel type routine although addition and subtraction are used as reversible operators rather than XOR. The routine relies on the alternate use of XOR and ADD to provide non-linearity. A dual shift causes all bits of the key and data to be mixed repeatedly. The number of rounds before a single bit change of the data or key has spread very close to 32 is at most six, so sixteen rounds will suffice (although the authors suggest 32!).

The key is set at 128 bits, which is more than enough to prevent brute force attack for the foreseeable future. The top 5 and bottom four bits are probably slightly weaker than the middle bits. These bits are generated from only two versions of z (or y) instead of three, plus the other y or z . Thus the convergence rate to even diffusion is slower. However the shifting evens this out with perhaps a delay of one or two extra cycles.

The key scheduling uses addition, and is applied to the unshifted z rather than the other uses of the key. In some tests $k[0]$ etc. were changed by addition, but this version is simpler and seems as effective. The number delta, derived from the golden number is used where $\text{delta} = (\sqrt{5} - 1)2^{31}$.

A different multiple of delta is used in each round so that no bit of the multiple will not change frequently. The algorithm is not very sensitive to the value of delta. It will be noted that delta turns out to be odd with truncation or nearest rounding, so no extra precautions are needed to ensure that all the digits of sum change.

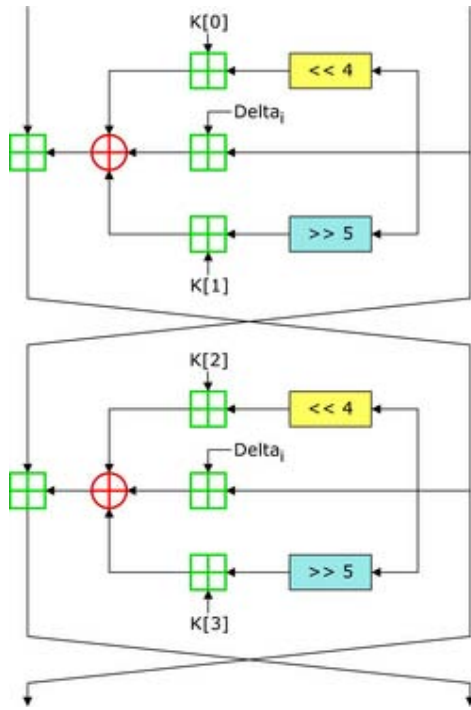


Figure 1. Two rounds of the TEA block cipher

The use of multiplication is an effective mixer, but needs shifts anyway. It was about twice as slow per cycle on our implementation and more complicated. The algorithm will easily translate into assembly code as long as the exclusive or is an operation. The hardware implementation is not difficult, and is of the same order of complexity as DES, taking into account the double length key.

In the last couple of years (X) TEA has been subject to several differential attacks, to which, surprisingly, TEA emerges less vulnerable than XTEA. Moon et al. [3] use impossible differentials to attack 14-round XTEA, requiring $2^{62.5}$ chosen plaintexts and the computation time of 2^{85} encryptions. The equivalent attack on TEA is on 11-rounds and needs $2^{52.5}$ chosen plaintexts and 2^{84} encryptions.

TEA and XTEA are analysed for differential and truncated differential weaknesses in [4]. An ordinary differential attack can break 15 rounds of XTEA with 2^{59} chosen plaintexts. Truncated differentials of probability 1 are used in an attack on 17-round TEA (1920 chosen plaintexts and $2^{123.37}$ time complexity) and 23-round XTEA ($2^{20.55}$ chosen plaintexts and $2^{120.65}$ time complexity).

The best attack to date on XTEA is a related-key differential attack on 27 rounds [5]. The attack requires $2^{20.5}$ chosen-plaintexts under a related key-pair and has a time complexity of $2^{115.15}$ 27-round XTEA encryptions.

4. EXPERIMENTAL ENVIRONMENT

In this section we describe the environment where most of the experiments described in section 5 were performed. An active VoIP QoS measurement is performed with software that simulates VoIP streams between two hosts connected via an IP network. The measurement data consists of a stream transmitted round-trip between two hosts, see “Fig. 2”. The stream consists of N packets of size S bytes that are transmitted at intervals of T milliseconds. This is very flexible approach since the parameters S and T can be chosen to correspond to any codec with any packetization scenario.

For example, by choosing S=200 bytes and T=20ms, we can simulate the transmission of G.711 speech data grouped into 20 ms frames, one frame per packet. By virtue of the round-trip test, the measurement simulates a full-duplex VoIP stream on the network level.

The measurement packets are marked with a transmission timestamp and sequence number (TS1, SEQ1) immediately prior to transmission. Once the packets are received by the host, another timestamp and sequence number (TS2, SEQ2) is added to the data packet and it is transmitted back. Upon reception of the packet, the original sender adds a third timestamp (TS3).

Using these data, the round-trip delay of a single packet is computed from the equation:

$$d = \alpha(TS_3 - TS_1) \quad (1)$$

where α is a scaling factor such that milliseconds result from the computation. Measurement software was implemented with C-language using windows sockets API and run on Pentium II laptops.



Figure 2. Packets are transmitted from Host 1 to Host 2 and back to Host 1.

5. EXPERIMENTAL RESULTS

We now investigate the impact of different encryption algorithms to encrypt the payload on the packet size and the delay. We consider the cryptographic algorithms DES, 3DES, IDEA, TEA (all implemented in software). Our experiments show that the crypto-engine is a serious bottleneck in the transmission of real-time traffic and the best performance is achieved by TEA.

A. Packet size:

Our results show that the impact of different encryption algorithms on the packet size is negligible; especially as the packet size increases. “Fig. 3” shows the percentage increase in packet size as a function of the original packet size for DES, 3DES, IDEA (top line) and for TEA (bottom line).

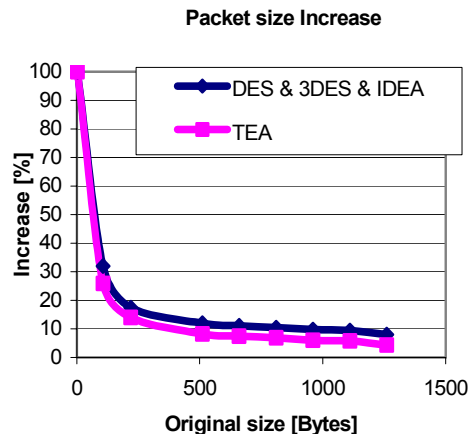


Figure 3. Packet size increase for two sets of cryptographic functions, as a function of packet size in bytes.

The packet size increase has negative effects not only on the bandwidth usage but it also impacts on the transmission delay, router internal delays, queuing delay, thus affecting jitter and overall packet delay.

B. Crypto-engine:

In order to measure the maximum encoding rate, when different algorithms are used, we performed the following experiments. We considered the cryptographic algorithms DES, 3DES, IDEA, TEA (all implemented in software) and for each case we generated 4 packet flows with packets of size 60, 100, 250, 1000 bytes, respectively. Each flow starts from 0 pps and increases its rate of 25 pps every 30 s in order to saturate the crypto-engine. "Fig. 4" graphs the measured throughput as a function of the global traffic flow.

The straight line is the throughput for transmission of packets in the clear, therefore it increases linearly with traffic. The figure shows that when encryption is performed, throughput levels off or decreases after reaching a maximum value, which depends on the algorithm. It also shows that longer packets significantly improve the crypto-engine performance. The best performance is achieved by TEA, then IDEA, DES, and the last is 3DES.

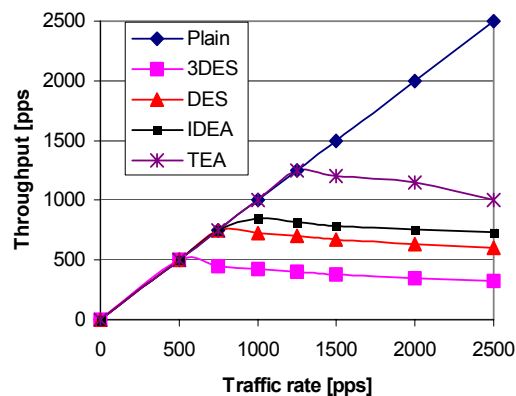


Figure 4. Throughput of the crypto-engine in pps as a function of linearly increasing traffic in pps for plain and encrypted traffic.

The negative slope throughput exhibits after reaching the maximum is due to packets discarded by the engine because it is saturated. Discarded packets contribute to lower the quality of the signal during the reconstruction phase.

In order to evaluate the effect of the payload encryption using different cryptographic algorithms on the end-to-end delay and the QoS degradation we implement the mean opinion score (MOS) test. In voice communications, particularly Internet telephony, MOS provides a numerical measure of the quality of human speech at the destination end of the circuit. The scheme uses subjective tests (opinionated scores) that are mathematically averaged to obtain a quantitative indicator of the system performance. It ranges from 1 to 5, 1 being the worst case. All traffic streams in our test are using G.711 PCM CODEC.

As Table 3 shows, the best MOS for a secure communication link is in the surroundings of 3.5 and is achieved by the TEA algorithm.

TABLE 3. Different cryptographic algorithms and their MOS.

Encryption Algorithm	MOS
NONE	4.1
DES	3.15
3DES	3.01
TEA	3.5
IDEA	3.25

6. CONCLUSION

Security is a serious bottleneck for the future of VoIP. Because of the time-critical nature of VoIP most of the same security measures currently implemented in today's data networks could not be used in VoIP networks. In this paper we have presented an efficient solution for securing VoIP using the TEA Encryption Algorithm. All our objective and subjective tests show that the best performance of transmitting voice over secure communications links is achieved by TEA.

REFERENCES

- [1] R. Barbieri, D. Bruschi, E. Rosti, "Voice over IPsec: Analysis and Solutions", Computer Security Applications Conference, 2002. Proceedings. 18th Annual, 9-13 Dec. 2002.
- [2] O. Elkeelany, Performance "Analysis of IPsec Protocol: Encryption and Authentication", University of Missouri-Kansas City, 2002.
- [3] D. Hong, Y. Ko, D. Chang, W. Lee, and J. Lim, "Differential cryptanalysis of XTEA", Technical Report TR03_13, Center for the Information Security and Technologies (CIST), Seoul, Korea, 2003a.
- [4] Y. Ko, S. Hong, W. Lee, S. Lee, and J. Lim, "Related key differential attacks on 26 rounds of XTEA and full rounds of GOST", In Proceedings of FSE '04, Lecture Notes in Computer Science, Berlin, Germany / Heidelberg, Germany / London, UK / etc., 2004.
- [5] D. Moon, K. Hwang, W. Lee, S. Lee, and J. Lim, "Impossible differential cryptanalysis of reduced round XTEA and TEA", Lecture Notes in Computer Science, 2365: 49-60, 2002. ISSN 0302-9743.
- [6] M. Marjalaakso, "Security Requirements and Constraints of VoIP", Technical Report, Dept. of Electrical Engineering and Telecommunications, Helsinki University of Technology, 2001.
- [7] R. Perlman, "Analysis of the IPsec Key Exchange Standard", Sun Microsystems Laboratories, 2001.
- [8] B. Schneier. Applied cryptography: "Protocols, Algorithms, and Source Code in C", John Wiley & Sons, Inc., 2nd Edition, 1996.
- [9] U.Black, "Voice over IP", Prentice Hall, 1999.

APPENDIX

Reference code of TEA:

Following is an adaptation of the reference encryption and decryption routines, released into the public domain by David Wheeler and Roger Needham:

```
void encrypt(unsigned long* v, unsigned long* k) {
    unsigned long v0=v[0], v1=v[1], sum=0, i;      /* set
up */
    unsigned long delta=0x9e3779b9;                /* a key
schedule constant */
    unsigned long k0=k[0], k1=k[1], k2=k[2], k3=k[3]; /*
cache key */
    for (i=0; i < 32; i++) {                        /* basic cycle
start */
        sum += delta;
        v0 += (v1<<4)+k0 ^ v1+sum ^ (v1>>5)+k1;
        v1 += (v0<<4)+k2 ^ v0+sum ^ (v0>>5)+k3;    /*
end cycle */
    }
    v[0]=v0; v[1]=v1;
}

void decrypt(unsigned long* v, unsigned long* k) {
    unsigned long v0=v[0], v1=v[1], sum=0xC6EF3720, i;
/* set up */
    unsigned long delta=0x9e3779b9;                /* a key
schedule constant */
    unsigned long k0=k[0], k1=k[1], k2=k[2], k3=k[3]; /*
cache key */
    for(i=0; i<32; i++) {                          /* basic cycle
start */
        v1 -= (v0 << 4)+k2 ^ v0+sum ^ (v0 >> 5)+k3;
        v0 -= (v1 << 4)+k0 ^ v1+sum ^ (v1 >> 5)+k1;
        sum -= delta;                               /* end cycle */
    }
    v[0]=v0; v[1]=v1;
}
```


Wireless Local Area Network Security: A Framework for Repairing the Broken WEP Protocol

Russ Housley
Vigil Security, LLC
housley@vigilsec.com

Jesse Walker
Intel Corporation
jesse.walker@intel.com

Nancy Cam-Winget
Cisco Systems
ncamwing@cisco.com

The IEEE 802.11 standard published in 1999 includes the Wired Equivalent Privacy (WEP) algorithm to protect communication from eavesdropping and to prevent unauthorized wireless network access; however, WEP has critical security flaws. A task group within the IEEE 802.11 working group, TG1, has developed standards for improved wireless local area network security. This paper exposes the security flaws in WEP and shows how IEEE 802.11i corrects them.

1. Introduction

In 1999, the IEEE 802.11 [6] standard introduced the Wired Equivalent Privacy (WEP) algorithm to protect communication from eavesdropping and to prevent unauthorized wireless network access. WEP has critical security flaws, as pointed out by Nikita Borisov, Ian Goldberg, and David Wagner [1] as well as Scott Fluhrer, Itsik Mantin, and Adi Shamir [2]. The flaws are the result of incorrectly using the RC4 stream cipher and choosing CRC-32 as a data integrity algorithm.

WEP constructs an RC4 per-frame key by simply concatenating a known initialization vector (IV) to the base key. This construction allows the attacker to easily identify frames encrypted with weak keys as described in [2], facilitating recovery of the base key. The lack of replay protection and the ability to repeatedly use the same IV values, coupled with the lack of a WEP key management protocol, also facilitates the recovery of the base key. Once the base key has been compromised, the system is wholly vulnerable. Finally, the inappropriate choice of CRC-32 as a data integrity mechanism trivializes bit-flipping attacks. With these vulnerabilities WEP is extremely susceptible to both passive and active attacks.

A task group within the IEEE 802.11 working group, TG1, has developed standards for improved wireless local area network (WLAN) security. TG1 was also faced with the reality of millions of deployed IEEE 802.11b units. Therefore, TG1 adopted a short-term solution to address WEP vulnerabilities in the deployed units and a long-term solution to fully address WLAN security. The short-term solution had to be easily deployed without requiring customers to discard their hardware. Nonetheless, both the short-term and the long-term solutions fit into a common framework and provide elements critical to WLAN security. This paper describes the framework used by TG1.

2. Overview of Wireless LANs and WEP

The fundamental building block of the IEEE 802.11 WLAN architecture is the Basic Service Set (BSS). The BSS is a group of stations (wireless network nodes) located within a limited physical area, where each station is capable of communicating with every other station. There are two WLAN design structures based on the BSS: infrastructure and ad hoc networks.

An infrastructure-based WLAN is composed of one or more BSS. Each station has exactly one BSS link to a connecting infrastructure, called the Distribution System (DS), which allows access to external networks. The station connects to the DS via an Access Point (AP), which relays frames from stations within the BSS to the DS as shown in Figure 1.

An ad hoc WLAN has no infrastructure, and therefore, no ability to communicate with external networks. An ad hoc WLAN permits multiple wireless stations to communicate directly with each other with minimal hardware or management support. The BSS of an ad hoc WLAN is referred to as an independent BSS (IBSS).

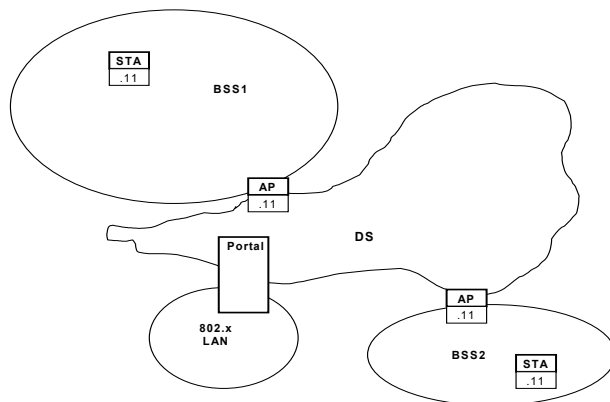


Figure 1. Typical Wireless LAN Configuration.

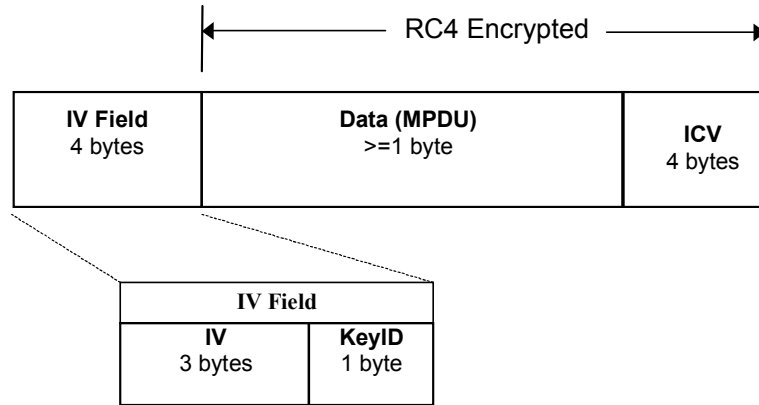


Figure 2. WEP Protocol Data Unit.

To transmit and receive data, stations compose datagrams called media access control (MAC) service data units (MSDUs). When transmitting data, the MAC layer determines whether the data in the MSDU should be partitioned into smaller frames called MAC protocol data units (MPDU) that are then processed for transmission. Conversely, when receiving data, the MAC layer determines whether the MPDU is a fragment, thus requiring reassembly of an MSDU. Each MPDU includes a frame check sequence (FCS); a CRC-32 is computed over the entire MPDU. The MAC uses the FCS to ensure that the frame has not been perturbed by noise.

Prior to communicating data, stations and APs must establish and validate access to the network as well as establish whether to communicate openly (open authentication) or securely (shared authentication). Open authentication is used to pass frames freely between the station and the AP, and shared authentication is used to verify that a station is authorized to use the network. Open authentication is really no authentication at all; it allows any requesting stations to enter the BSS. Shared authentication uses a challenge and response exchange along with a shared secret to authenticate the station to the AP, but it is easily compromised. TGi has replaced the flawed shared key authentication with authentication mechanisms running over IEEE 802.1X, but we will not discuss this effort further in this paper.

The IEEE 802.11 standard does not specify a means for obtaining the shared secret. The shared secret is typically a 40-bit key or a 104-bit key shared between many stations. A key that is shared between the AP and many stations is referred to as a default key. A key that is shared between the AP and only one other station is referred to as a key-mapping key. Both default keys and key-mapping

keys are subsequently used to protect communications between associated stations.

The WEP protocol is used to protect MPDUs produced by the IEEE 802.11 MAC from MSDUs. WEP uses the pre-established shared secret key and the RC4 algorithm for encryption, and it uses CRC-32 to compute an Integrity Check Value (ICV). The ICV is computed over the MPDU data. The resulting 32-bit ICV is appended to the MPDU prior to encryption. The RC4 key is composed of a 24-bit IV value concatenated with the shared secret key to form a per-frame key. The MPDU data and ICV are then encrypted under the per-frame key. The IV and a key identifier are prepended to the encrypted MPDU data field, and the resulting WEP Protocol Data Unit, shown in Figure 2, is ready for transmission to the peer.

3. Review of WEP Flaws

This section reviews the major problems with WEP. This summarizes the work reported in [1], [2], [3], and [4]. The WEP design exhibits both sins of commission and sins of omission.

Foremost among the sins of commission is the misuse of the RC4 stream cipher. RC4 is an excellent cipher, used widely in a range of modern security applications, largely because of the high degree of privacy it affords with a relatively low performance penalty. However, stream ciphers are difficult to use properly in frame protocols. RC4 was a particularly poor choice in the WEP context, as will be described later. To understand this, it is useful to review how stream ciphers operate.

By definition, a stream cipher generates a pseudo-random stream, called a key stream. The encryptor exclusive-ORs (XORs) the key stream with the plaintext to produce ciphertext. The decryptor generates the same key stream and XORs it with the ciphertext to recover the plaintext.

What happens if the encryptor XORs the same key stream with two different plaintexts? Data compromise. Suppose two plaintext byte sequences p_1, p_2, p_3 and q_1, q_2, q_3 are both encrypted with key stream k_1, k_2, k_3 . The corresponding ciphertexts are:

$$\begin{aligned} p_1 \oplus k_1, p_2 \oplus k_2, p_3 \oplus k_3 \quad \text{and} \\ q_1 \oplus k_1, q_2 \oplus k_2, q_3 \oplus k_3 \end{aligned}$$

If both of these two ciphertext streams are exposed to an attacker, then a catastrophic failure of privacy results, since:

$$(p_i \oplus k_i) \oplus (q_i \oplus k_i) = p_i \oplus q_i$$

That is, using a stream cipher to encrypt two plaintexts under the same key stream trivially leaks a great deal of information about plaintext. This is a property of all stream ciphers.

A second problematic stream cipher characteristic is the encryptor and decryptor must remain synchronized. That is, the decryptor must know the relative offset of each byte of ciphertext, otherwise it will use the wrong byte of key stream for decryption. Since the IEEE 802.11 MAC is neither reliable nor does it provide in-order delivery at the level at which WEP operates, a cipher with the random access property is needed. A cipher with the random access property can easily generate any specified byte of the key stream. This property is useful over an unreliable communications channel, because context accompanying a frame can specify the synchronization to the decryptor. RC4 does not have the random access property, but the WEP key derivation function (concatenation of the IV and the base key) simulates the random access property. Unfortunately, the RC4 key schedule is too lightweight to be used in this manner, and the resulting key streams have too many similarities.

The WEP designers had some intimation of these stream cipher limitations, and they tried to compensate for them. To avoid these problems they defined a per-frame RC4 key. This is a reasonable strategy, but the design introduces more problems in the way it implements the strategy.

The first per-frame key problem is the manner in which WEP constructs the per-frame key. The IV is concatenated with a base key. The encryptor selects the IV and transmits it as plaintext as described above. The decryptor uses the received IV to construct the same per-frame key. This construction should have been suspect at the outset, as it exposes the first part of the encryption key. In a paper presented in August 2001, Scott Fluhrer,

Itsik Mantin, and Adi Shamir [2] investigated the simplistic RC4 key schedule when a portion of the RC4 key is known, and they showed that this kind of construction leads to a class of RC4 weak keys. Patterns in the keys themselves are reflected in patterns at the beginning of the generated key stream. If the first two bytes of enough key streams can be observed, then the whole RC4 key can be recovered, including the base key used to construct the per-frame keys for other frames. This exploit is called an FMS attack.

Using such a weak method to construct per-frame keys would be a problem in any case, but the WEP design compounds this by an unrelated situation: the first few bytes of encrypted data in almost every frame are known. The SNAP-SAP header [12] is almost always at the beginning of the payload. This known value allows an adversary to recover the first two bytes of the generated key stream, which is precisely the information needed to mount the FMS attack and recover the base key. AirSnort, a public domain hacker tool, implements this attack. The first version of AirSnort appeared in August 2001, and it had to examine about one million frames to recover the base key. By January 2002, the AirSnort implementation had reportedly been improved to require only about 20,000 frames to recover the base key, which is about 11 seconds of IEEE 802.11b traffic under normal conditions.

IV usage also allows an attacker to recover all the plaintext without ever learning the base key. WEP uses a 24-bit IV, which means there can be a maximum of $2^{24} \approx 16$ million per-frame keys associated with any base key. Thus, to avoid duplication, the base key must be replaced at least once every 2^{24} frames. Since an IEEE 802.11b channel can sustain an average of about 1800 data frames a second, the base key must be replaced at least every 2.5 hours.

The collision of the 24-bit IVs suggests some very simple, low-tech attacks. An eavesdropper can record all the WEP encrypted traffic, group recorded frames by IV, and XOR frames encrypted under the same IV to learn a significant amount about the data itself. It is often feasible to use pattern recognition techniques to disentangle the two XORed plaintext frames. Once this is accomplished, the generated key stream can be exposed, permitting the eavesdropper to directly decrypt all subsequent frames until the base key changes. A less patient attacker can arrange to the transmission of known plaintext, such as SPAM email, to directly recover the key stream.

A third problem is that the WEP specification imposes no rules on IV selection; “frequent” change of the IV is recommended. As a result,

vendors implement their own IV strategies. Some implementations operate with a fixed IV, employing the same RC4 key to encrypt every frame! Other vendors selected the IV at random. After two frames the probability of an IV collision under this strategy is $1/2^{24}$, but the situation rapidly degrades. After only 4823 frames there is a 50% chance of collision. Other vendors pursued a third strategy, using the IV space as a circular counter, always starting at zero upon boot. This strategy guarantees a collision after two stations transmit a single frame protected by the same base key.

A passive eavesdropper can exploit the problems discussed thus far, but an active attacker can do even more damage. WEP fails to provide effective data integrity. The WEP designers thought they had designed a data integrity mechanism, but the specified algorithm fails to provide the intended protection. There are three problems with this design.

The first problem is the data integrity mechanism itself. The WEP transmitter computes a CRC-32 over the data payload, appending the resulting ICV to the data, and then encrypts the ICV along with the data. The idea was the receiver could detect data modifications by decrypting the data and the ICV and then verifying that the decrypted ICV matches the data. However, this algorithm does not prevent undetected data modification. An attacker can record a valid frame, create a zero pad with the same length of the encrypted data, flip one or more bits, and compute the ICV of this bit-flipped zero pad. Then the attacker can create a forgery by XORing both the bit-flipped zero pad and ICV and the encrypted data in the recorded frame, including the recorded encrypted ICV. This works because the CRC-32 construction and XOR-based encryption commute. That is, the same value results, regardless of the order of the operations. Further, the CRC-32 is linear over combinations of data it protects. After decryption, the modified ICV in the forgery will validate correctly, and WEP will accept the frame as genuine. If combined with traffic analysis, the attacker can use this technique to construct frames with correct application data.

Even if the encrypted CRC-32 construction provided the intended protection, the data integrity mechanism does not cover all the information that needs to be protected from modification. As an example, the ICV mechanism does not protect the frame destination address. An attacker can record a frame from a station to an AP, change the destination address, and then send the frame. When the AP receives this forgery, it dutifully decrypts the frame and forwards it to the “wrong”

address, perhaps to the attacker. A similar alteration of the source address of frames from the AP to another station allows the adversary to masquerade as any station.

The last data integrity mechanism problem: WEP provides no replay protection. An attacker can record any frame and then retransmit them later with or without alteration. Since each of the frames is encrypted under a valid key, they will be accepted at the IEEE 802.11 level as valid. Traffic analysis can reveal the use of various connectionless protocols, with the replayed data being accepted as authentic at the application layer.

All of these problems arise when an eavesdropper can collect a sufficient number of frames encrypted under the same base key. If the WEP base key were changed sufficiently often, then these attacks might afford an adversary significantly fewer options to compromise security. However, IEEE 802.11 provides no mechanism to replace keys, practically requiring customers to use static, manually configured keys. It is infeasible to manually change keys often enough to provide protection from these attacks.

The WEP architecture compounds this problem in two ways. First, WEP uses the same key to protect data in both directions over a link. Implementations often use a counter in each direction to generate the next IV, which guarantees immediate IV collision and data exposure. Second, IEEE 802.11 only provides a way to name default keys, thereby encouraging the use of a single group key within a WLAN. It is simply infeasible to manage quantities that cannot be named.

4. Solution Constraints

Millions of WEP-based devices have been deployed. The industry had an obligation to fix the security of these devices if at all possible. Like most modern communication equipment, IEEE 802.11 devices are comprised of hardware and software. WLAN hardware has been designed as a commodity, so it is not cost effective to add or swap out particular hardware chips in a WLAN device; instead, it is cheaper to replace the entire hardware unit. This implies that WEP patches operating on already-deployed IEEE 802.11 hardware will rely entirely on software upgrade. This is the first design constraint, and it poses a particularly sticky dilemma.

IEEE 802.11 APs present a computational bottleneck, as they have little spare processing capacity. Recall that in an infrastructure deployment, all stations link with the AP instead communicating directly among themselves, and the

AP handles every message exchanged within the BSS. In order to be competitive in a commodity market, APs are typically implemented with the cheapest hardware possible, using a microprocessor like an i486, ARM7, or PowerPC running at 40 or even 25 MHz. The load generated by normal WLAN traffic often consumes 90% or more of the microprocessor computational bandwidth, so very few cycles are available for new functions. In some cases, there may only be 2 million unused instructions per second available. This is the second major design constraint.

This is an impassible barrier for a traditional security design targeted for implementation in software. The cryptographic functions such a design would necessarily employ are processor intensive. At IEEE 802.11b data rates, standard cryptographic primitives can easily consume the entire AP processor.

Nearly all shipping APs have custom hardware to handle the RC4 encryption. Most of this hardware is tuned to construct per-frame keys according to the WEP algorithm: the per-frame key is a base key concatenated to an IV, which appears as plaintext in each frame. On transmit, the hardware expects the frame as it input, along with the base key and IV. The custom hardware constructs the per-frame key, encrypts the MPDU payload, inserts the IV, and passes result to the radio transmitter. On receive, the hardware extracts the IV, locates the base key, constructs the per-frame key, and decrypts the MPDU payload as it arrives from the radio receiver. In most receivers there is very little time between frame arrival and start of decryption. In this time interval, at most three hundred instructions can be executed. In some devices, some of this time is used to locate the base key. The hardwired encryption function represents a third major design constraint. The design affords few opportunities for software intervention into an outgoing frame after encryption and even fewer for an arriving frame prior to decryption.

On first analysis, therefore, fixing WEP with any cryptographically sound approach seems to be impossible without instantly obsolescing all existing hardware. TGi designed a long-term solution called CCMP that does precisely this. It is impossible to utilize standard cryptographic functions in any way to rescue WEP, at least on already-deployed hardware, because very few have sufficient spare processing capacity to accommodate the needed operations.

One alternative within present hardware was to do nothing. However, TGi designed a short-term solution called TKIP with vastly improved

security; however, the cost, performance, and security trade-offs required to support deployed hardware does not allow these WEP repairs to fully address the TGi security goals. The WEP repairs presented in this paper serve as a short-term solution to allow security improvements on currently deployed hardware until the long-term solution becomes available.

5. WEP Repairs

Four components comprise the WEP security repairs. Two components, session key derivation and an improved per-frame key derivation function, provide protection against passive attacks. The other two components, data integrity checking and replay prevention, provide protection against active attacks. All four components are necessary for a complete security solution.

The session key derivation component is called the 4-Way Handshake. In the next section, we discuss the 4-Way Handshake. Following that, we discuss the other elements of the design: the per-frames key derivation function, data integrity checking, and replay prevention.

5.1. 4-Way Handshake

A key must be refreshed when its lifespan has expired or when an attack is presumed. The lifespan of a particular key depends on the encryption algorithm and the way that the key is used. In the WEP protocol, the use of a 24-bit IV suggests a key lifespan of $2^{24} \approx 16\text{M}$ frames. However, because of the security flaws discussed above, some experts suggest that the maximum key lifespan for the WEP base key ought to be no more than about $2^8 \approx 256$ frames. Using IEEE 802.11b average frame and data rates, a key refresh every 256 frames would demand a new key every 0.2 seconds, a prohibitive rate. The short-term solution (using the per-frame key derivation function discussed in the next section) employs a 48-bit IV, allowing the key to survive up to $2^{48} \approx 2.8 \times 10^{14}$ frames, and the key lifespan is about 1 minute. The long-term solution uses AES [11] with a 44-bit IV, allowing the key to survive up to $2^{44} \approx 1.8 \times 10^{13}$ frames between rekeys, with the key lifespan exceeding 600 years.

Protocols for key management and key refreshment are well known and practiced today, typically in layers above the MAC. Mechanisms such as IKE [7] and TLS-EAP [8] facilitate the establishment of secret keys. While these protocols are well suited for their intended use, they lack some of the characteristics necessary to

establish MAC layer keys while minimizing disruption to the traffic flow. To overcome this shortcoming, TGi defined a new protocol, called the 4-Way Handshake, that uses IEEE 802.1X [9] as its transport.

IEEE 802.11i defines two types of keys, group keys and pairwise keys, corresponding to multicast and unicast traffic. Group keys can be shared between an AP and more than one station, while each pairwise key maps to a unique AP and a single station. Naturally, a group key and a pairwise key needs to be refreshed under different conditions. The conditions necessary to refresh a pairwise key are straightforward. Both the AP and the station must:

- Share the some common cryptographic keying material, necessary to authenticate the key refresh event;
- Construct the key(s) used to protect the association;
- Generate a replay protection value to thwart adversaries from desynchronizing and spoofing the link; and
- Assign the same key identifier to the resulting key, allowing the AP and station to agree when the new key will be used (and the old key will be discarded).

IEEE 802.11i uses a three tier pairwise key hierarchy. The first tier is a master key, established by IEEE 802.1X authentication between the station and an authentication server. Each party uses the master key to derive a fresh Pairwise Master Key (PMK), and possession of the PMK demonstrates authorization to access the wireless channel. The authentication server delivers the PMK to the AP. Next, the PMK is used in the 4-Way Handshake by the station and the AP, which results in a key confirmation key (KCK), a key encryption key (KEK), and a temporal key (TK). The KCK authenticates 4-Way Handshake messages, demonstrating possession of the PMK. The AP encrypts the group key under the KEK to securely distribute it to the station. The TK is used as an encryption and data integrity key for the session.

Each of these keys has freshness requirements. The station and the AP must establish a new master key each time the station comes into contact with the wireless LAN after being away long enough for its master key to lapse. This leads to a fresh PMK for each AP. Otherwise, the 4-Way Handshake establishes a fresh session keys for each reassociation between the station and the AP by exchanging random values and mixing them with the PMK to produce a fresh KCK, KEK, and TK.

5.2. Per-frame Key Derivation

As previously described, WEP generates a different RC4 key for each frame by concatenating the 24-bit IV and the 104-bit (or 40-bit) base key. This method, constructs per-frame keys that lead to a keystream correlated with the IV. As demonstrated by the FSM attack, the lightweight RC4 key-scheduling algorithm is vulnerable when the initial few bytes of plaintext are easily predictable, as is almost always the case with IEEE 802.11. The short-term solution therefore had to introduce an improved key derivation function.

The long-term solution, CCMP, uses AES, not RC4, for encryption. AES has significantly different properties that obviate the security requirement for per-frame keys, so there is no need for per-frame key derivation function in the long-term solution.

Ron Rivest, the author of RC4, suggests two ways to employ RC4 with per-frame IV values. He recommended (a) discarding the first 256 output bytes of the key stream, or (b) strengthening the key-scheduling algorithm by preprocessing the key and the IV by passing them through a one-way hash function such as SHA-1 [13]. However, discarding the first 256 output bytes is too expensive for existing equipment, and it is infeasible for some implementations. The use of one-way hash functions is also too expensive for deployed equipment.

Since the obvious fixes are too expensive, a new key derivation function was designed that is cheap enough to execute on existing hardware. It derives a per-frame key from the 128-bit TK. This solution is distributed as a firmware upgrade by vendors, allowing their customers to update existing vulnerable equipment.

The new per-frame key derivation function operates in two phases. In the first phase, the transmitter address (TA) and upper 32 bits of the frame sequence number are mixed into the TK. By including the TA, multiple stations can use the same TK, and each station (including the AP) generate a different key stream. This property is important in all networks. Consider the simple case where a station communicates only with one AP. Data sent by the station to the AP and data sent by the AP to the station will be encrypted with the same TK. If the TA were not mixed with the TK, the same series of RC4 key streams would be used by both the station and the AP, enabling data recovery attacks discussed above.

The output of the first phase will likely be cached; it can be reused to process subsequent frames that use the same TK and the same TA.

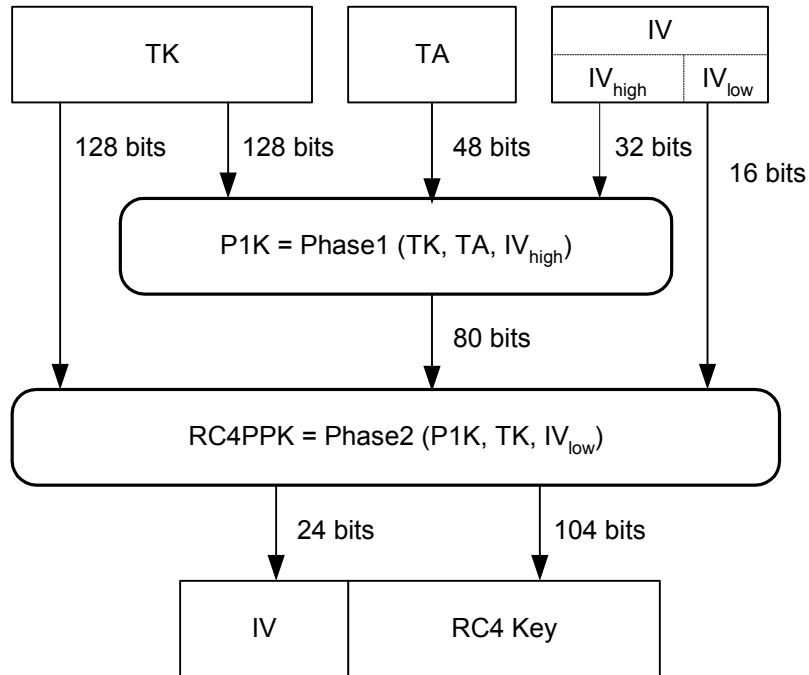


Figure 3. Per-frame RC4 Key Derivation.

The second phase uses a Feistel cipher to mix the IV and the output of the first phase. The IV is the low order 16 bits of the frame sequence number, initialized to zero when the TK is established. By including the IV, each frame will be encrypted with a unique key stream. Using a Feistel cipher to perform the mixing makes it difficult for an attacker to correlate the IV and the per-frame key. The two-phase per-frame key derivation process is summarized in Figure 3.

Since the key derivation function must easily integrate with existing hardware, the outputs are a 24-bit IV and a 104-bit RC4 key. These are concatenated to form the RC4 per-frame key, as was done in WEP

5.3. Data Integrity

WEP does not prevent frame forgery. The addition of a Message Integrity Code (MIC) attempts to correct this deficiency by ensuring that the data within the frame and selected portions of the frame header have not been modified. That is, the data sent by the originator and the data received by the recipient are the same. The MIC is a keyed cryptographic function, traditionally called a Message Authentication Code (MAC). Using the term “MIC” avoids confusion with Medium Access Control, with the same acronym.

The MIC value is computed from the key, the data within the frame and selected portions of the frame header, and then the MIC value is appended to the frame data. When the frame data is encrypted, the MIC value is also encrypted.

The near-term solution had to accommodate the very limited amount of processing power available in the MAC processor (as well as some APs), so a relatively weak integrity function is employed.

This is acceptable because:

1. The ICV (a CRC-32 checksum) is computed on the plaintext data payload. The ICV is part of WEP, and most implementations compute it in hardware.
2. The FCS (another CRC-32 checksum) is computed on the ciphertext. The FCS is used to detect transmission errors.
3. The MIC value is encrypted, which indirectly increases the overall strength.

To accommodate implementation of the MIC algorithm in the host processor or the MAC processor, which ever has the available processing capacity; the MIC is computed on the MSDU (before fragmentation into MPDUs). The host does not generally have access to MPDUs, and the host is completely unaware of any fragmentation or reassembly.

A new MIC algorithm, called Michael [14], was developed for the short-term solution. Michael was designed to be efficient on currently deployed MAC processors as well as processors typically used by first generation APs.

Performance and security are the two dominating concerns in Michael's design. Its inner loop uses only XORs, shifts, byte-swaps, and additions; all of these operations are cheap on the target processors. Michael costs about 3.5 cycles/byte on an ARM7, and about 5.5 cycles/byte on an i486. This means it will consume about 3.1 M cycles/second on an ARM7-based AP, and 4.8 M cycles/second on an i486-based AP. This processing ought to consume every unused processor cycle on many first generation IEEE 802.11 APs, resulting in some performance degradation in a fully loaded BSS.

It is easy to establish an absolute upper bound on the security afforded by any MIC. When an n -bit MIC algorithm is employed, the algorithm maps any message to one of 2^n possible values. As a result, MIC security level is usually measured in bits. If the security of a MIC is s bits, then by definition the probability an attacker can construct an acceptable forgery on the first frame is 2^{-s} , and by the birthday paradox, an adversary expects to produce an acceptable forgery after about $2^{s/2}$ frames. If the MIC algorithm is completely secure, then the number of bits in the MIC value is the number of bits of security provided. However, Michael, by design, sacrifices security for computational efficiency. Even though Michael has a 64-bit MIC value, it was designed with a target security level of 20 bits, and it is believed that it slightly exceeds this target. The best attack known against Michael is based on differential cryptanalysis, and it indicates that Michael offers 29 bits of security.

Accidental MIC check failures will occur very rarely. The FCS will detect noise or interference on the radio channel. A station receiving 100 randomly formatted frames per second can expect one to pass the FCS checks less than once a year. As discussed above, the FCS and the ICV CRC-32 checksum computations use the same polynomial. Therefore, if a modified frame passes the FCS check, it will most likely also pass the ICV check. However, the encryption between the FCS and the ICV do ensure that the receiver and the originator are using the same TK. Only frames that satisfy both the FCS and ICV checks will get to the point where the MIC is verified.

Given the very low rate of accidental MIC failures, it is reasonable to assume that an active attack in progress. Countermeasures to thwart the

active attacker are deployed. These countermeasures include discarding the current TK, closing the association, and notifying the system administrator. Discarding the current TK prevents the attacker from learning anything about the TK from the MIC failure. Shut down of the association introduces delay, and this slow down prevents the attacker from sending a large number of fraudulent frames in a short time. Notification of the system administrator allows a human to search for the location of the active attacker.

In the long-term solution, an AES-based data integrity mechanism will be used to detect MPDU modification. By protecting the MPDU, AES modes that provide both confidentiality and data integrity can be employed in the long-term solution. Such modes provide protection against modification for the encrypted data and for plaintext MPDU header fields. Also, by providing both services on the MPDU, replay detection can take advantage of an integrity protected sequence number. Replay detection is discussed further in the next section.

5.4. Replay Detection

In order to provide replay protection, the short-term solution uses the existing WEP IV field as a frame sequence number. Each TK has its own sequence number space.

The transmitter initializes the sequence number to zero whenever a new TK is set, and then increments the sequence number by 1 for each successive MPDU. If the TK is not refreshed prior to IV sequence space exhaustion, the transmitter must halt communication.

The receiver follows the same initialization rule, resetting a sequence number to zero when the TK is refreshed. A frame is considered to be out of order when its sequence number is the same or smaller than a previously received MPDU associated with the same TK. If a MPDU arrives out of order, then it is considered to be a replay; it is discarded, and a MIB counter is incremented. The receiver increments the replay counter only if the ICV of the MPDU is valid and the sequence number indicates in-order delivery. As described in the previous section, the long-term solution will provide data integrity protection for the MPDU header fields. This will provide modification detection for the IV field, which is used as the sequence number.

In the short-term solution, the data used by replay detection algorithm is protected indirectly. Since the IV is used to construct a per-frame key, modification of the IV will cause the receiver to

attempt decryption with the wrong key stream. Thus, the frame data and the MIC will decrypt incorrectly, leading to a MIC verification failure.

The replay detection algorithm relies on the fact that IEEE 802.11 preserves the frame sequence. However, IEEE 802.11 TGe is working on a quality of service (QoS) definition that obviates this assumption. The replay detection mechanism was therefore extended to include a per-traffic-class replay counter value.

5.5. Interdependence

A WEP security patch requires all of these enhancements. If the 802.11i key management is not implemented, then the resulting protocol is still subject to data compromise when IVs collide. If the per-frame key derivation algorithm is not implemented, then the protocol is still subject to the FMS attack. If the replay detection algorithm is not implemented, then the protocol is still subject to forgeries by replay. And if the message integrity check is not implemented, the protocol is still subject to frame forgery attacks. Trying to implement only some of these enhancements would be similar to closing only some of the hatches on a submarine before submerging: doing so fails to achieve the ultimate goal. Security becomes possible only when all the core deficiencies are addressed.

The replay detection and MIC together defend against active attacks. However counter-intuitive it might be, it is always necessary to try to defeat active attacks that undermine data integrity to achieve confidentiality guarantees, because they can be turned into attacks against the encryption itself, as they cause the protocol itself to reveal more about the encryption key than passive attacks.

The 4-Way Handshake and the improved key derivation function together restore the assumptions made by the encryption algorithm. Without these guarantees, the encryption function cannot do its job properly.

6. Interoperability Considerations

Interoperability must be preserved during the transition from WEP to the short-term solution to the long-term solution. To accomplish this, each station must know the protocol and algorithms that are being used by its peers.

In a BSS, the AP is the only peer. It offers its preferences for authentication algorithms and the cipher suite for the protection of unicast and multicast traffic in its Beacons and Probe Responses. If none of the alternatives offered are

acceptable, then the STA must not associate, or it will be rejected; otherwise the STA selects one of the authentication algorithms, one of the unicast cipher suites, and the multicast cipher suites. An EAP method over IEEE 802.1X is the only non-proprietary authentication algorithm specified for use with the short-term solution. EAP-TLS [8] is a suitable EAP method, as it provides mutual authentication and key establishment. EAP Authentication mechanisms without these properties should not be used, as they fail to provide fresh master keys. Further, use of the flawed legacy IEEE 802.11 authentication is prohibited with the short-term and long-term solutions. IEEE 802.11 TGi defines a similar solution for ad hoc networks.

7. Conclusion

IEEE 802.11 TGi defined an architecture that repairs the known deficiencies in WEP. Four components comprise the WEP security repairs. The 802.11i 4-Way Handshake and an improved per-frame key derivation function provide protection against passive attacks. Message integrity checking and replay prevention provide protection against active attacks. All four components are needed for a complete security solution.

IEEE 802.11 TGi developed a short-term solution to address the WEP vulnerabilities on currently deployed hardware. Though the short-term repairs do not provide the same security strength as that of the long-term solution, it allows currently deployed hardware to persist until the long-term solution becomes available.

IEEE 802.11 TGi developed a long-term solution to fully address WLAN security needs, adopting the AES algorithm.

These solutions were published in IEEE 802.11i in July 2004.

References

- [1] Borisov, N., I. Goldberg, and D. Wagner, "Intercepting mobile communications: the insecurity of 802.11," in *Proc. International Conference on Mobile Computing and Networking*, ACM, July 2001, pp 180-189.
- [2] Fluhrer, S., I. Mantin, and A. Shamir, "Weaknesses in the key schedule algorithm of RC4," in *Proc. 4th Annual Workshop on Selected Areas of Cryptography, 2001*.

- [3] Stubblefield, A., J. Ioannidis, and D. Rubin, "Using the Fluhrer, Mantin, and Shamir attack to break WEP", AT&T Labs Technical Report TD-4ZCPZZ, AT&T Labs, August 2001.
- [4] Walker, J., "Unsafe at any key size: an analysis of the WEP encapsulation," IEEE 802.11 doc 00-362, October 27, 2000.
- [5] Shirey, R., "Internet Security Glossary," RFC 2828, May 2000.
- [6] IEEE Std 802.11, Standards for Local and Metropolitan Area Networks: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, 1999.
- [7] Harkins, D., and D. Carrel, "The Internet Key Exchange (IKE)," RFC 2409, November 1998.
- [8] Aboba, B., and D. Simon, "PPP EAP TLS Authentication Protocol," RFC 2716, October 1999.
- [9] IEEE Std 802.1X, Standards for Local and Metropolitan Area Networks: Standard for Port based Network Access Control, 2001.
- [10] Bellare, M., and P. Rogaway, "Entity Authentication and Key Distribution", Crypto '93 Proceedings, August 1993.
- [11] National Institute of Standards and Technology. FIPS Pub 197: Advanced Encryption Standard (AES). 26 November 2001.
- [12] Postel, J., and J.K. Reynolds, "Standard for the transmission of IP datagrams over IEEE 802 networks," RFC 1042, February 1988.
- [13] National Institute of Standards and Technology. FIPS Pub 180-1: Secure Hash Standard. 17 April 1995.
- [14] Ferguson, N., "Michael: an improved MIC for 802.11 WEP," IEEE 802.11 doc 02-020, January 17, 2002.
- [15] IEEE Standard for Information technology—Telecommunications and information exchange between systems—Local and metropolitan area networks—Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications Amendment 6: Medium Access Control (MAC) Security Enhancements, July 23, 2004.

Time Division Hashing (TDH): A New Scheduling Scheme for Wireless Ad-Hoc Networks

Winnie Cheng, I-Ting Angelina Lee, Neha Singh

MIT Computer Science and Artificial Intelligence Laboratory

32 Vassar Street, Cambridge, MA 02139

Phone: (617) 253-5851, Fax: (617) 258-8682, Email: {wwcheng, angelee, nsingh }@mit.edu

Abstract

The current 802.11 MAC protocol uses RTS-CTS-DATA-ACK packet exchange and exponential backoffs to prevent collisions at both the sender's side and the receiver's side. While the protocol works well in prevention of collisions, it degrades throughput due to a significant amount of overhead produced by RTS/CTS message exchange. In addition, aggressive backoffs used in current 802.11 MAC protocol can lead to long channel idle time, resulting in the channel being under-utilized. In this paper, we propose a new scheduling scheme, TDH, for ad hoc wireless networks which minimizes the chances of collisions by eliminating RTS/CTS message exchange and the use of backoffs. We also study and compare the performance of our proposed scheme with the performance of MAC RTS/CTS scheme under different types of topologies. Our analysis leads to the conclusion that, our scheme has higher throughput in many cases compared to MAC RTS/CTS scheme, however, at the expense of a slightly higher collision rate.

I. INTRODUCTION

Shared media is a fundamental problem in wireless networks. Nodes within certain range can interfere with each other's communication. A substantial amount of research has been done to address this issue, and most working schemes are based on Carrier Sense Multiple Access/Collision Avoidance (CSMA/CA). Current 802.11 Media Access Control (MAC) protocol, for instance, uses RTS-CTS-DATA-ACK message exchange to establish connection between two nodes and prevent other nodes within range from transmitting and causing interference. When a node tries to establish a connection with another node and fails (either due to busy media or no CTS response from the other node) or when a collision happens, exponential backoff is performed. While it works well in preventing collisions at both the sender's side and the receiver's side, there are downsides to the scheme that can sometimes cause more harm than benefit and decrease overall throughput. For instance, the RTS/CTS message exchange can generate a lot of overhead – ideally, this should not be the case since RTS/CTS packet is only 30 bytes each in length. However, in reality, RTS/CTS packets are sent at a much slower rate than data packets (usually 10 or 11 times slower), and thus, cost a substantial amount of overhead, especially when the data packet sizes are small. Furthermore, while aggressive backoff provides a mechanism for channel contentions, it sometimes

causes long channel idle time (when multiple nodes decide to backoff simultaneously) or unfairness with transmitter-based contention (i.e. one transmitter gets considerably higher bandwidth compared to other transmitters).

These fundamental problems within the MAC RTS/CTS scheme motivated the design of our proposed scheme – Time Division Hashing or TDH. Our proposed scheme is based on dividing time into fixed-length slots and hashing these slot values to determine whether the node is in 'send' or 'receive' mode. Each node generates its own random but deterministic sending/receiving schedule by hashing with a pseudo random seed (details of the algorithm are described in section III). That means, as long as the random seed is known, a node can deterministically predict when any other node will be sending or receiving. While our proposed scheme does not prevent all types of collisions, it generates better overall throughput in many cases because there is no overhead associated with RTS/CTS message exchange or idle time caused by exponential backoffs. Furthermore, our scheme leads to better fairness within the network traffic, i.e. each node gets approximately fair share of bandwidth for transmission.

This paper is organized in the following manner: in section II, we briefly discuss some related work done previously and explain our motivations for designing the new scheme. In section III, we describe our proposed scheme, including pseudo code and how the

algorithm works. In section IV, we present our simulation results of comparisons between the two schemes. Finally, in section V, we summarize and conclude our results from the simulation, and propose possible future work.

II. MOTIVATION AND RELATED WORK

Some of the early work in Media Access Control schemes for wireless networks was based on the Carrier Sense Multiple Access/Collision Avoidance (CSMA/CA) technology. An example of such a scheme is Multiple Access with Collision Avoidance (MACA) [6]. Such schemes try to address the issues specific to wireless networks such as congestion at the receiver, collisions due to hidden terminal problems and inefficiency due to exposed terminals. To solve some of these problems, they use the RTS/CTS handshake mechanism to perform ‘Virtual Carrier Sensing’ and establish a connection between the two communicating nodes. Once the RTS/CTS messages have been exchanged, the communicating nodes cannot participate in any other transmission for the length of time specified in the control messages. In addition, neighboring nodes (within receive range) of the two communicating nodes also cannot transmit during that time.

These early schemes suffer from inherent problems such as unfairness and long delays during traffic congestion. This led to development of improved backoff algorithms and contention window management schemes such as MACAW [3], Estimation-based Backoff [5], Adaptive Backoff [4], Weighted Hierarchical Backoff [9] and Dual Stage Contention Resolution [16]. There are other approaches that borrow ideas from fair queuing such as Distributed Fair Scheduling [15] and fair-share estimates [2]. These schemes often require great implementation complexity to achieve reasonable fairness.

Other than unfairness and congestion, the RTS/CTS mechanism has some more fundamental problems. Firstly, it is ineffective in detecting hidden terminals that are present outside of the receive range, but in each other’s interference range. Virtual carrier sensing does not work in more realistic interference models, where the interference range is usually two or more times greater than the receive range. Another reason for performance degradation can be inefficiency caused by exposed terminals not being able to transmit simultaneously. Two neighboring nodes should be able to transmit at the same time as long as their receivers are not neighbors, but RTS/CTS messages prevent this from happening as no node that has heard an RTS from a neighbor will transmit until the first transmission is complete. These and other issues have been analyzed in some recent work that is aimed at evaluating performance of the RTS/CTS mechanism ([10], [13]).

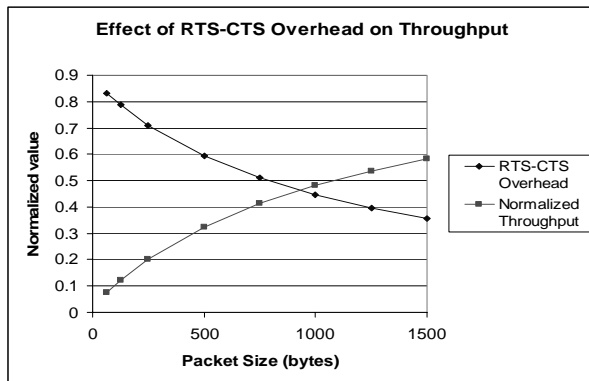


Figure 1. Effect of RTS-CTS Overhead on Throughput

Another problem with RTS/CTS as implemented in the 802.11 standard is the overhead caused by sending these control messages at a much lower bit rate than the rate at which data is sent [12]. This leads to observed capacities being significantly less than optimal even for simple chain and lattice networks [7]. The aggressive backoff mechanism can also cause long idle times [7]. Backoff occurs in 802.11 when two RTS messages collide. It also happens in the event of a data collision, which is detected by the sender upon not receiving acknowledgement for the packet that it transmitted.

To demonstrate some of the weaknesses of the RTS/CTS mechanism, we performed some simulations *in ns*. The first result (Fig. 1) shows the effect of the RTS/CTS control messages on the overall throughput.

The scenario above consists of two nodes, with one sending to the other. The normalized throughput decreases as the packet size decreases and reaches a value below 10% for a packet of size of around 64 bytes. The overhead is calculated as the fraction of the total transmission time taken up by the RTS/CTS messages. It can be seen from this result that the RTS/CTS overhead is significant, especially for small packet sizes. The normalized throughput is directly affected by this overhead.

The second experiment demonstrates the performance of RTS/CTS in a topology with interference. Each pair of nodes in Fig. 2 indicates a flow. The carrier sensing range is indicated by the circles, so none of the pairs directly hear other. But the interference range of the nodes is approximately twice the carrier sense range.

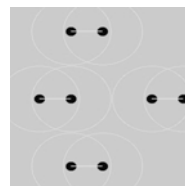


Figure 2. Interference Topology

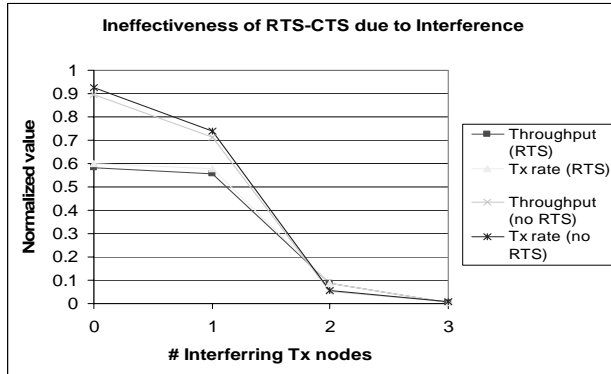


Fig 3. Ineffectiveness of RTS-CTS due to Interference

The results of measuring throughput as more such pairs are added are shown in the graph below (Fig. 3). In this case, RTS/CTS is not helpful in avoiding collisions at the receiver. The throughput drops to less than 10% when there are 2 or more pairs of interfering nodes. Disabling RTS/CTS does not help much either.

To address some of the above weaknesses of 802.11 caused by the RTS/CTS mechanism, we propose a new scheduling scheme, Time Division Hashing (TDH), for the MAC layer of 802.11. This scheme is based on the clock exchange scheduling algorithm for Packet Radio Networks proposed by Tim Shepard [11]. Tim’s thesis addresses the problem of designing a scalable multihop packet radio network that manages transmissions of spread spectrum radios. He analyzes propagation and interference models for local interference as well as the overall noise in the system, and proposes a scheme for scheduling packet transmissions to avoid collisions caused by interference from nearby stations without the need for global coordination or synchronization. TDH is based on this idea of forming deterministic schedules without the need for global coordination. However, Tim’s model cannot be used as such for the wireless networks using 802.11 because he makes some assumptions in deriving his results that are not true for current wireless networks in general. In his scheme, collisions at the receiver are avoided by the use of multiple receive channels and code division multiplexing. Most of today’s wireless radios do not use this spread spectrum technology and therefore, we still have to face the problem of collisions at the receiver. Also, his experimental results are derived with the assumption that minimum energy routing is used for routing packets in multihop networks. This assumption also does not hold for today’s wireless networks in general.

There has been other work in designing a better MAC scheduling scheme for wireless networks than the RTS/CTS mechanism. [14] proposes a topology-dependent transmission scheduling scheme, called collision-avoidance time allocation (CATA). CATA allows nodes to contend for and reserve time slots by

means of a distributed reservation and handshake mechanism. This scheme still suffers from the overhead of the handshake required for data transmission, which TDH is trying to avoid. [1] proposes a scheme called Neighbor-aware Contention Resolution (NCR), which generates a permutation of the contending members, the order of which is decided by the priority of all participants. This scheme relies on the availability of local topology information within at least two hops. Our work is different from these and other similar schemes because it tries to solve the scheduling problem with minimum overhead of control messages by having a deterministic schedule for a node that is known to all its neighbors. It tries to maximize throughput by probabilistically picking the time slots in which a node will be sending and receiving. We also simulate our scheme on a variety of different topologies, including some that are more realistic than the ones used for deriving results in most other schemes.

III. THE TDH ALGORITHM

Time Division Hashing (TDH) is a decentralized media access scheme that is based on each node in the wireless network having a deterministic schedule. Time is divided into fixed-length slots. Each slot is meant for transmitting a single packet. The schedule is represented in terms of the node being in either ‘send’ state or ‘receive’ state in each of these slots. Slots of either kind are picked according to some fixed probability. The mechanism for the nodes to come up with a schedule is as follows. Each node in the network maintains a value called the seed, which is randomly picked. Each node’s schedule is offset by this random value, and that is how each node has a different schedule. To determine the state of a slot, the value of the time slot is offset by the seed and is then hashed. Depending on the fixed transmit probability value, the value returned from the hash function is either chosen to be a ‘send’ slot or a ‘receive’ slot. Fig 4 shows an example schedule for a few nodes as returned from this algorithm.

The power of this scheme lies in the fact that a node can compute its neighbor’s schedule by knowing just one value, which is the seed. Each node needs to know the seed values of all its neighbors. Using this seed, it can determine whether its neighbor is in ‘send’ or ‘receive’ mode using the method above. The problem of scheduling now reduces to finding a slot in which the sender is in ‘send’ mode and the receiver is in ‘receive’ mode. For example, in Fig 4, if node 1 gets a packet destined for node 2 at time 0, the first available slot it would pick is slot 2 which is a ‘send’ slot for node 1 and a ‘receive’ slot for node 2.

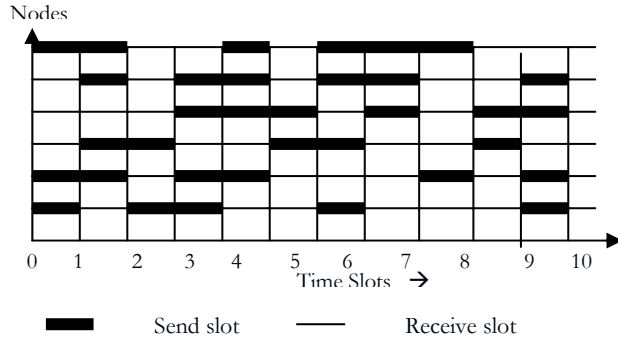


Figure 4. Example of a schedule for 6 nodes

The most important parameter to this algorithm is the ‘transmit probability’ p with which slots are picked to be transmission slots or reception slots. The optimal value of this parameter is dependent on the topology that the algorithm is being used for and on the traffic patterns. For some simple topologies, the optimal value of p that will maximize throughput can be derived theoretically. For more complicated scenarios, it is hard to get the best value for p theoretically. In such cases, a good value can be arrived at by estimation. In our scheme, all nodes share the same value of p . However, it is also possible to envision a scheme in which each node has its own p depending on its own traffic pattern. So if a node is going to be sending a lot of traffic, it should pick a high value of p so that it will maximize its chances of finding a match for a receive slot.

The second important parameter to the algorithm is the length of a time slot. In the TDH scheme, collisions are still possible at the receiver. This is because two nodes that are in ‘send’ mode in the same time slot could simultaneously schedule a packet for transmission to a third node that is in ‘receive’ mode in that slot. This is a side-effect of the fact that the scheme has been designed such that a sender only needs to know the receiver’s schedule in order to transmit a packet. Even if a sender checked all its neighbors’ schedules, collisions due to hidden terminals could not have been avoided. Therefore, it is required to detect collisions in this scheme and acknowledgements are necessary. The time slot has been designed so that both the packet and the ACK can be transmitted in one slot. The slots are fixed-size and are long enough to transmit a packet of the maximum size. Also since transmission and reception are being done in the same slot, it is necessary to switch the wireless radios between the two modes twice in the slot. So the length of a slot can be calculated with the following formula:

$$\begin{aligned} \text{slot length} = & \text{time to transmit packet} \\ & + \text{time to transmit ACK} \\ & + 2 * (\text{time to switch radio}) \end{aligned} \quad (1)$$

Fig. 5 outlines the TDH algorithm in pseudo-code. The first step in TDH is for each node to pick its seed s , which is a random value. The seeds are exchanged

Figure 5. TDH Algorithm

```

SchedulePacket (recvID, currrentTime) {
  t = beginning of next time slot from currentTime ;
  if ( PrevTimes(recvID) > t )
    then t = PrevTimes(recvID) + slotLength ;
  srecv = seed of receiver ;
  loop {
    increment t till a slot is found that is not
    in ScheduledTimes;
    senderMode = hash ( s + t ) ;
    destMode = hash ( srecv + t ) ;
    if ( senderMode <= p and destMode > p ) {
      Schedule packet for t ;
      Update ScheduledTimes, PrevTimes ;
      return t ;
    } else
      t = t + slotLength ;
  }
}

```

between all pairs of neighbors, so that a node knows the seeds of all its neighbors. The scheduler uses a few data structures to keep track of the state of the scheduled packets. *PrevTimes* is a table that keeps track of the time that the last packet for scheduled for each destination that the node has sent to. *ScheduledTimes* is a list that contains all the slots that are ‘occupied’, meaning that all the time slots beginning from the current time that already have packets scheduled to be sent in them.

The TDH algorithm is called to schedule a packet for a certain receiver at a certain time. The scheduler starts looking for a slot after the current time. It first checks to see if there is a packet already scheduled for that destination at a later time using *PrevTimes*. In this case, it only needs to start looking for a slot after that time. The algorithm then iterates on this time slot until it finds a suitable time. It first checks to see if the current slot has already been used by checking if it is in *ScheduledTimes*. If it is, then the current slot is incremented to the next slot. Once it finds a free slot, it finds out its mode in that slot by adding its seed to the current time and hashing that value. The hash function returns a value between 0 and 1. If the value returned is less than the ‘transmit probability’ p , then the node is in ‘send’ state in that slot. The same process is repeated for the receiver using its seed. If the hash function returns a value greater than p for the receiver, it means

the destination is in ‘receive’ mode at that time. If both these conditions are met, it is the ideal case and a packet can be scheduled for that time. If the result of the hash function is not less than p for the sender or if it is not greater than p for the destination, then the time is incremented to the next slot and the algorithm repeats the process until it finds a match.

As an optimization to the algorithm, if a node is in ‘send’ mode in a particular time slot but does not have any packets scheduled for that slot, it’s receiver is turned on to listen to broadcast traffic. Also, for the initial seed exchange phase of the algorithm when all nodes pick their random seeds and transmit it to all their neighbors, a modification is possible that will eliminate the need for this exchange. A unique ID of the node such as its MAC address can be used as the seed value of the node. As long as the hashing function is good, this should work as well as picking random numbers. In this case, there would be no seed exchange phase needed as the nodes could just discover their neighbor’s seeds from the MAC address on the packets sent by them.

Theoretical Analysis

As suggested previously, the transmit probability p affects the throughput of the network. Though the optimal value of p depends on the underlying topology and network traffic dynamics, the theoretical throughput can be derived analytically for the following multi-node topology to provide an upper bound on the achievable throughput. Assuming a topology in which one sender has the freedom to send to any of its k neighboring nodes. The maximum throughput is expressed as the probability of this sender being in ‘send’ state and the probability that at least one of its neighbors is in ‘receive’ state. That is,

$$T(p, k) = p(1 - p^k) = p - p^{k+1} \quad (2)$$

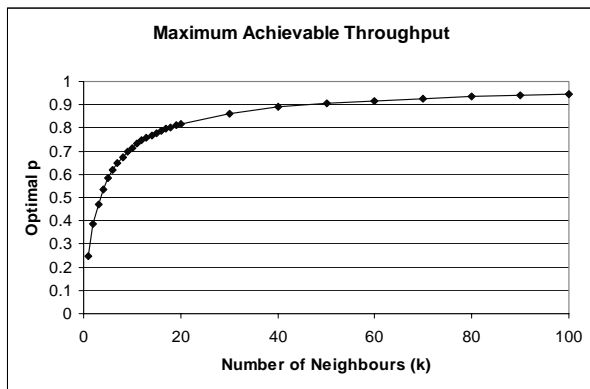


Figure 6. Theoretical Upper bound on Throughput

Implementation

To analyze the TDH scheme, we implemented and simulated it under NS-2. We built our scheme over the CMU Monarch protocol stack [8]. The algorithm described above is implemented as a standalone class, called ‘TDHScheduler’. The MAC layer code is replaced with the new scheduling algorithm. RTS/CTS message exchange and use of backoffs is eliminated.

A few implementation issues are discussed below. Firstly, with our scheme, when a packet needs to be scheduled, it is not necessary that the time chosen for it will be later than the time the previous packet was scheduled for. Because of this, the MAC layer needs to get multiple packets from the link layer at once for TDH to work efficiently. For instance, node A might be in the ‘send’ mode during time slots 2, 3, and 4. Assume it gets a packet from the link layer which should be delivered to node B, and it schedules the packet to be sent at slot 3 because that is the time when node B happens to be in ‘receive’ mode. If node A gets more packets from the link layer, it is possible to schedule them for the free slots before 3 if they are meant for other destinations. Doing so maximizes the utilization of node A’s transmission time. However, this also means that the MAC layer needs to be able to manage multiple outstanding packets at one instant. This includes buffering of the scheduled packets and managing their sending states (actual transmission time, status of outstanding acknowledgements, etc). Secondly, since we are buffering multiple packets at the MAC layer, flow control is required between the link layer and the MAC layer. Ideally, MAC layer would like to get as many packets as possible to maximize the utilization of its transmission time. However, it also needs to limit the amount of buffering since the latency seen by upper layers increases as more packets are buffered. As future work, it would be interesting to look into the relationship between the link layer latency and the amount of buffering at the MAC layer.

IV. SIMULATION RESULTS AND ANALYSIS

To analyze the performance of TDH, we compared the proposed scheme with 802.11 RTS/CTS using the network simulator NS-2. We studied 5 topologies in our simulation: *single sender*, *access-point*, *clique*, *chain* and *random topology*.

In these simulations, we used Constant-Bit-Rate (CBR) traffic with Dynamic Source Routing (DSR) as the underlying routing protocol. The data points are collected for simulation run time of 300s. The results are presented for channel bandwidth of 11Mbps and RTS/CTS transmission rate of 1Mbps. The interference range and carrier sensing range are set to 550m and

250m, respectively. Unless otherwise specified, the data packet size is 1500-bytes.

A. Single Sender Topology

This topology consists of N nodes in which one node is designated as the ‘sender’. All nodes are within carrier sensing range of each other. The sender node creates N-1 CBR flows sent at equal rate to each of its neighbors. The aggregate rate of all flows is equal to the channel capacity. The transmit probability p is set to the optimum value according to eqn 2.

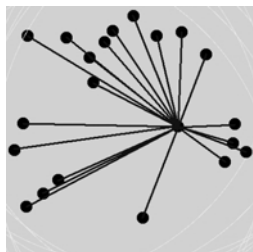


Figure 7. Single Sender Topology

Since there is only one sender, schedule collisions and interference are kept to a minimum. This should result in a throughput close to the theoretical maximum derived in Section III. Fig. 8 plots the observed normalized throughput against the theoretical maximum as a function of the total number of nodes in the topology. The observed throughput is defined as the sum of all received CBR bytes at the (N-1) CBR sink nodes normalized to the product of channel bandwidth and total simulation time. The experimental results are close to the theoretical values with slightly greater deviation as the number of nodes increases. This is due to increase in the overhead of routing messages.

The RTS/CTS throughput is shown in comparison with TDH in Fig 9. In this simulation run, RTS-CTS exchange is enabled. With 1500-byte data packet sizes and an 11:1 ratio in transmission rate, the expected overhead due to RTS/CTS is $11 * (40 + 39) / [11 * (40 + 39) + 1500 + 39] \approx 36\%$ compared with the measured value of 42% ($1 - 0.58$ from the graph) which also includes the overhead of routing messages and protocol maintenance. The throughput is almost constant regardless of the number of nodes in the topology. The contention scheme of 802.11 is sender-centric where transmitting nodes compete for channel access. With the aggregate rate being constant and only one sender in this topology, the normalized throughput should therefore be constant. For TDH, on the contrary, as the number of nodes increases, there is a higher probability of matching the sender’s transmit slot with any of the neighboring nodes’ receive slot. This should have an impact on scheduling latency and throughput. When the number of nodes exceeds 10, TDH gives higher throughput than RTS/CTS.

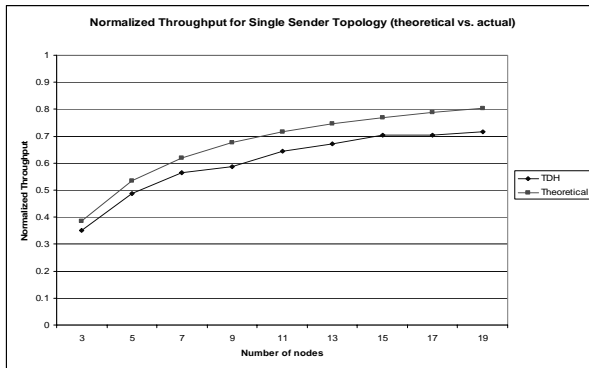


Figure 8. Comparison of Normalized Throughput with Theoretical value in Single Sender Topology

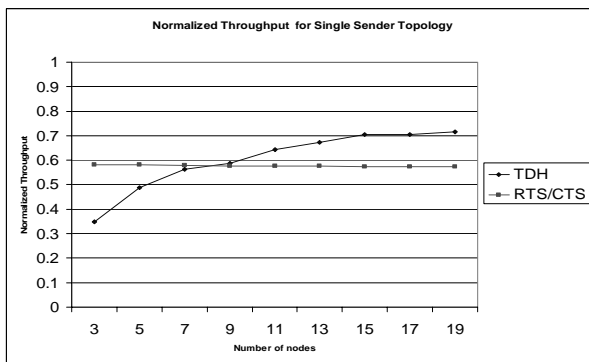


Figure 9. Comparison of Normalized Throughput with RTS/CTS for Single Sender Topology

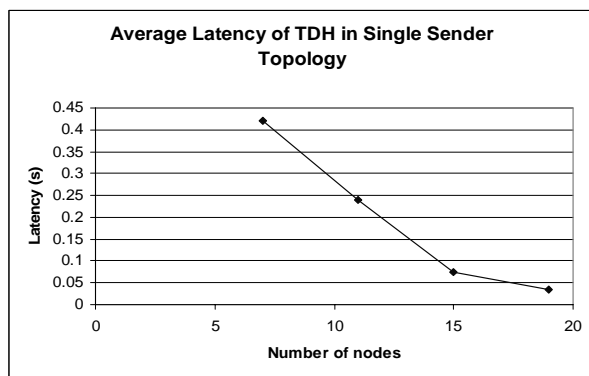


Figure 10. Average Latency in Single Sender Topology

From the latency plot in Fig. 10, it can be seen that average latency is decreasing as the number of nodes increases, though this includes some queuing delay. Here the latency is calculated as the difference between the time that the TDH algorithm is called to schedule a packet and the time that the packet is scheduled to be sent. The latency decreases because the time to find a match reduces when there are more receivers and more slots can be utilized.

B. Access-Point Topology

In this topology, we model an access point configuration where there is one node sending upstream with (N-2) downstream nodes receiving data from the access point. All flows have the same rate, and the aggregate transmission rate does not exceed the channel capacity. We investigate two questions with this topology:

- How “fair” are the schemes in allocating per-flow access?
- How is throughput affected when RTS/CTS is enables/disabled?

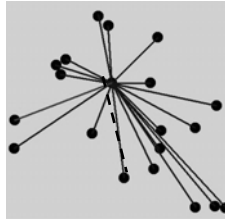


Figure 11. Access Point Topology. (Upstream flows are dashed. Downstream flows are solid.)

In evaluating per-flow fairness, we use the fairness index f as our metric. It is defined as follows:

$$f = \frac{(\sum_{i=1}^n x_i)^2}{n \sum_{i=1}^n x_i^2}$$

x_1, x_2, \dots, x_i refer to the throughput of the i^{th} competing flow.

TDH has a high fairness index between 0.9 and 1 as shown in Fig 12. It schedules all transmissions with equal probability, namely, $p^*(1-p)$. In RTS/CTS, when there are 2 flows (one upstream and one downstream), the fairness index is 1 as these two senders use RTS to reserve the channel for their respective flows. However, as the number of downstream flows increases, there are still 2 senders. The access point sender is now responsible for more flows but from the channel’s perspective, the competition is still between 2 senders. As a result, this causes significant unfairness to the downstream flows, in particular, when the number of downstream flows is small. As this number increases, the RTS sending rate of the access point also increases as a result of a higher aggregate transmission rate and the fairness index improves.

In practice, the RTS/CTS mechanism is sometimes disabled for access point configuration especially when the amount of upstream traffic is small. Fig. 13 illustrates the effect of disabling RTS/CTS on throughput in comparison with TDH and RTS/CTS enabled. The RTS/CTS throughput is almost identical to the throughput in Single Sender topology. This

suggests that the extra sender causes very few RTS collisions. Hence, as observed, it is reasonable that the throughput is much higher when RTS/CTS is disabled.

Comparing TDH with 802.11 RTS/CTS, the graph is not too different from the Single Sender topology with the exception of an earlier crossover point. With the extra sender in this topology, when the number of nodes is small, the overall probability of a transmission occurring is increased as it is equal to the sum of the each of the sender transmitting, resulting in higher throughput.

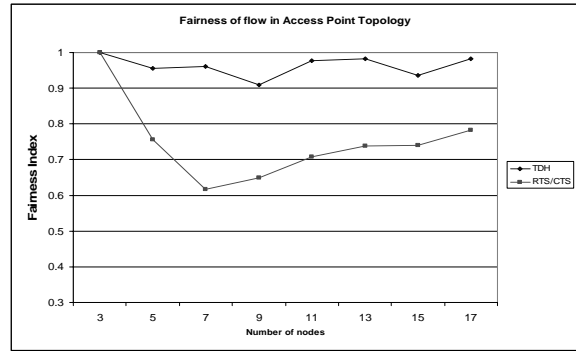


Figure 12. Fairness Indices for TDH and 802.11 RTS-CTS in Access Point Topology.

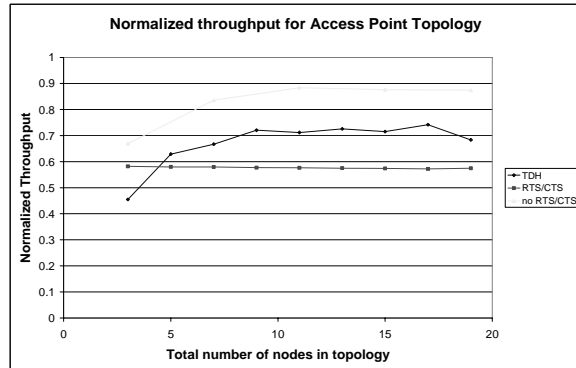


Figure 13. Comparison of Normalized Throughput for Access Point Topology.

C. Clique Topology

One of the drawbacks of TDH is that it does not have an explicit mechanism to avoid collisions from senders that have scheduled transmission in the same time slot. TDH uses a probabilistic approach to reduce the frequency of such events. The clique topology is devised to maximize the chances of such collisions. There are N nodes in carrier sensing range of each other. Each node acts as a sender of (N-1) CBR flows to its neighboring nodes. The aggregate flow rate of all senders is equal to the channel capacity.

The RTS/CTS mechanism avoids many of the collisions resulting in higher throughput than the other two schemes as shown in Fig. 15. Since the aggregate flow rate is set to the channel capacity, as the number of nodes increases, the total transmission rate of each node decreases by $(1 / N)$. The per-flow rate thus decreases approximately by $(1 / N^2)$. The RTS/CTS throughput drops only slightly because the increasing number of senders and associated RTS collisions are offset by the decrease in flow rate.

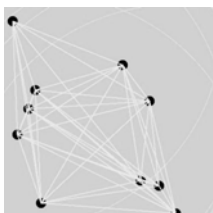


Figure 14. Clique Topology

Under the possibility of multiple senders colliding, TDH shows an average normalized throughput of about 0.45, around 10% less than that of RTS/CTS. One would expect that as the number of nodes increases, the probability of collision increases. Assume that the probability that a node has scheduled transmission in a ‘send’ slot is g . Then, for a clique with N nodes, the probability of collision in a given time slot is:

$$P(\text{collision}) = 1 - (1-g)^{N-1} \quad (3)$$

$(1-g)$ is the probability of no transmission in a ‘send’ slot and therefore, $(1-g)^{N-1}$ is the probability that none of the other $(N-1)$ nodes are also sending in that slot. Or put differently, the probability of successful transmission in a time slot diminishes with the power of $(N-1)$. Therefore, one may expect a much sharper decline in the throughput as the number of nodes increases. However, from Fig. 15, we see that this is not the case. There is no sharp drop due to the following two reasons. First, from previous results, as the number of transmissions increases, the slots are utilized more efficiently. Second, in this topology, the channel is saturated but not oversubscribed as the per-flow rate is set such that the aggregate rate is close to the channel capacity. This means that the per-flow transmission rate decreases by $1 / N^2$ as the number of nodes increases and therefore, a high number of collisions is not observed as shown in Fig 16. These two factors lead to a gradual decrease instead of a drastic drop in throughput, in a scenario where a large number of collisions are possible.

D. Chain Topology

It is important to examine multi-hop behavior in MAC protocols. As suggested by [7], interference several hops away can affect node-to-node transmission, and undesirable congestion can occur. The chain

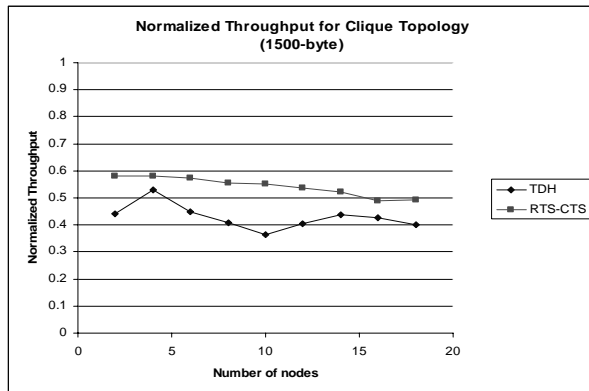


Figure 15. Normalized Throughput for 1500-byte packet size in Clique Topology.

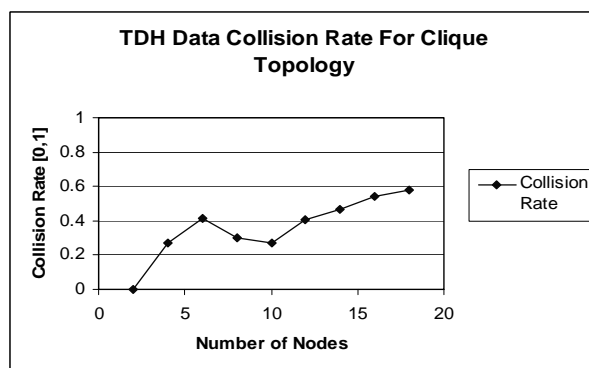


Figure 16. Data Collision Rate of TDH scheme in clique topology

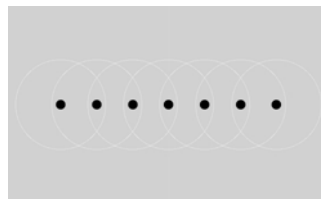


Figure 17. Chain Topology

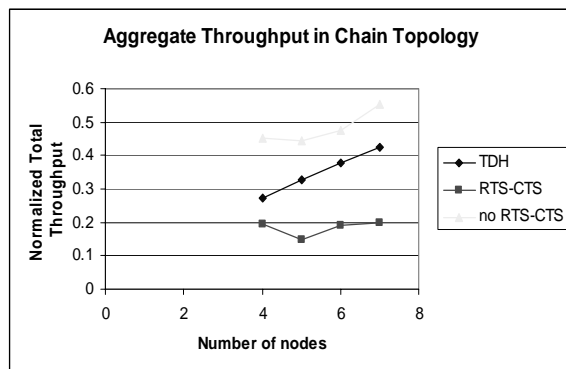


Figure 18. Aggregate Throughput in Chain Topology

topology arranges nodes in a line. Successive nodes are spaced 200m so that only adjacent nodes are within

carrier sensing range. The interference range, however, extends out to 550m. The transmit probability p is set to 0.5, the optimal value for two-node communication. The leftmost node sends at the maximum channel capacity rate to the rightmost node in the chain. All other nodes are passive participants in the forwarding path.

TDH suffers from having only one neighbor to try to match schedules with, so the node-to-node transmission rate is limited to 25%. Despite this, it is able to achieve a relatively high throughput compared to the RTS/CTS scheme. Fig. 18 shows the total throughput with the received bytes summed at all the nodes. RTS/CTS schedules transmissions over one in every three adjacent links. However, due to the larger interference range, collisions can happen between transmissions from nodes that are even two links apart. To achieve a collision free scenario, only one in four links should be used for transmission. With a maximum of 7 nodes in our topology, we expect no more than $(2 * \text{channel capacity})$. The aggregate throughput is normalized to this value in Fig. 18.

Ideally, in a perfect forwarding chain, the aggregate throughput should increase linearly with the number of nodes. TDH exhibits this trend with a small but steady slope. suboptimal schedules and long backoff in subsequent nodes.

With RTS/CTS disabled, it is able to push traffic through quite aggressively. On the contrary, the aggregate throughput of the RTS/CTS scheme is almost constant. As the number of nodes increases, transmission early on in the chain causes congestion in the middle of the chain which leads to suboptimal schedules and long backoff in subsequent nodes.

The chain throughput in Fig. 19 shows the result of the cumulative effect of interference and collisions at intermediate nodes. It is the observed throughput at the destination (rightmost) node. A quick examination of the TDH curve shows that extrapolating the line to a 2-node chain gives approximately 0.25 which is the throughput limit for node-to-node transmission.

The scheme with RTS/CTS disabled delivers the highest chain throughput. It is able to achieve a throughput of 30% for a 4-nodes 3-links topology, which is very close to 33% that can be obtained by using 1 out of the 3 links. For a 5-nodes 4-links topology, the value is around 22%, close to the expected 25%. However, this is at the expense of a high collision rate. The RTS/CTS disabled chain throughput curve shows the steepest gradient. As the number of nodes increases, the throughput drops significantly since the transmission of earlier nodes interferes with those later in the chain.

The throughput is worsened with RTS/CTS enabled, as it does not actually transmit data until the RTS/CTS messages are successfully exchanged. During

the process of RTS/CTS exchange, several tries might be needed to reserve the channel and each unsuccessful try triggers a binary exponential backoff. This results in the channel being greatly under-utilized. Even if the channel is successfully reserved, data collisions can still occur due to interference.

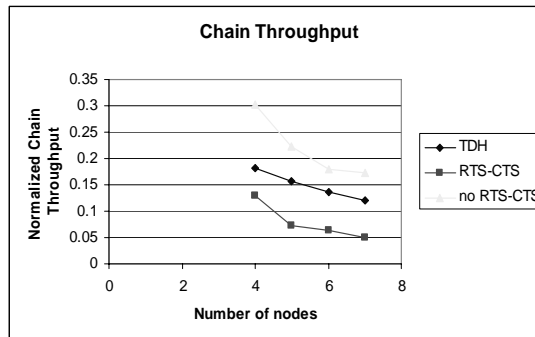


Figure 19. Chain Throughput

E. Random Topology

In an attempt to evaluate more realistic scenarios, the following random topology is generated. An area of 1000m by 1000m is divided into 16 squares. 10 nodes are randomly placed in each square, one of which is chosen as the sender. This sender creates some flows that are intended for nodes within the square and some intended for nodes in different squares. With this configuration, it reflects local traffic patterns as well as multi-hop behavior. The transmit probability p is set at 0.7 for all nodes.

The throughput of TDH turns out to be the highest compared to the other schemes as shown in Fig. 20 without RTS/CTS, many collisions occurred, significantly impacting throughput. However, with RTS/CTS, the situation is worse as RTS collisions triggered the backoff mechanism. About 75% of the transmission time, as illustrated in Fig. 21, is used for RTS/CTS transmissions. RTS collision rate is approximately $((54-19) / 54) \approx 65\%$ and therefore, the number of transmitted data packets is significantly lower than the other schemes. Time is wasted in leaving the channel idle as a consequence of backoff.

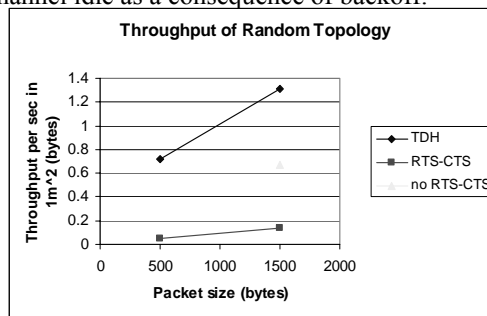


Figure 20. Throughput of Random Topology

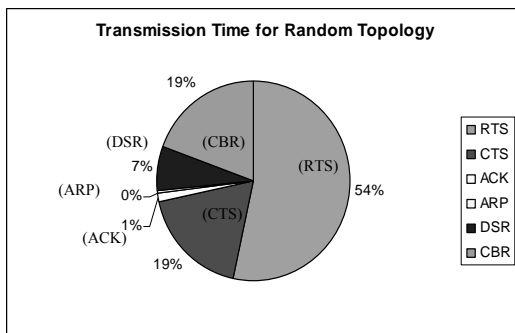


Figure 22. Breakdown of Transmission Time in the RTS/CTS scheme for Random Topology

V. CONCLUSIONS AND FUTURE WORK

The RTS/CTS mechanism in the 802.11 MAC layer incurs a lot of overhead, resulting in significant degradation of overall performance. The gain in throughput from collision avoidance does not always compensate for the harm caused by aggressive backoffs and the RTS/CTS overhead. Our proposed scheme is simple, does not cost additional overhead, and gives better performance in many cases, as the simulation results have shown. In addition, it has better fairness properties.

Some weaknesses of the TDH scheme we implemented are pointed out below. Firstly, the drop rate is high compared to the RTS/CTS scheme since TDH does not have a mechanism to avoid collisions on the receiver side. Secondly, performance degrades as the number of neighbors decreases. This is inherent in the algorithm because it gets harder for the sender to find a matching time slot (when the sender is in ‘send’ mode, and the receiver is in ‘receive’ mode) as there are less potential receivers. Also, we use static values for p (the percentage of time a node spends being in ‘send’ mode) and a fixed slot length in our simulations. This is less flexible and may not work too well with variable packet sizes.

Keeping these weaknesses in mind, there are some interesting topics to investigate and potential improvements that can be made to the TDH scheme. First, it will be interesting to investigate the impact of collisions on the transport layer, especially when the transport layer guarantees reliable and in-order delivery. Second, we can replace the fixed-length slots with multiple sub-slots to achieve more flexible time slot management when dealing with variable packet sizes. Third, we would like to experiment more and see how TDH performs under different network parameters (such as higher bit rate, delay and channel loss rate). Finally, machine learning techniques can be applied to dynamically assess and adjust the p value to optimize performance for different network topologies and traffic patterns.

REFERENCES

- [1] L. Bao and J.J. Garcia-Luna-Aceves. “Distributed Dynamic Channel Access Scheduling for Ad Hoc Networks”. In JPDC, Special Issue on Wireless and Mobile Ad Hoc Networking and Computing, 2002.
- [2] B. Bensaou, Y. Wang, C.Ko. “Fair medium access in 802.11 based wireless ad-hoc networks”. International Conference on Mobile Computing and Networking, 2000.
- [3] V. Bharghavan, A. Demers, S. Shenker, and L. Zhang,. “MACAW: A Media-Access Protocol for Packet Radio”. In Proceedings of ACM SIGCOMM, 1994.
- [4] F. Cali, M. Conti, and E. Gregori. “IEEE 802.11 Protocol: Design and Performance Evaluation of an Adaptive Backoff Mechanism”. IEEE Journal on Selected Areas In Communications, Vol. 18, NO. 9, September 2000
- [5] Z. Fang, B. Bensaou, and Yu Wang. “Performance evaluation of a fair backoff algorithm for IEEE 802.11 DFWMAC”. In Mobile Computing and Networking, pages 48-57, 2002.
- [6] P. Karn, “MACA – A New Channel Access Method for Packet Radio”, ARRL/CRRL Amateur Radio 9th Computer Networking Conference, September 1990.
- [7] J. Li, C. Blake, D. De Couto, H. Lee, R. Morris. “Capacity of Ad Hoc Wireless Networks”. In the Proceedings of the 7th ACM International Conference on Mobile Computing and Networking, July 2001.
- [8] CMU Monarch Group. CMU Monarch extensions to ns: <http://www.monarch.cs.cmu.edu>
- [9] T. Ozugur. “Weighted Hierarchical Backoff Algorithm for Wireless Ad Hoc Networks”. IEEE Global Telecommunications Conference, 2001.
- [10] K. Xu, M. Gerla. “Effectiveness of RTS/CTS Handshake in IEEE 802.11 based Ad Hoc Networks”. Mobicom 2003.
- [11] T. Shepard. “Decentralized Channel Management in Scalable Multihop Spread Spectrum Packet Radio Networks” MIT PhD Thesis. 1995.
- [12] ANSI/IEEE Std 802.11. “Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications.” 1999.
- [13] S. Xu, T. Saadawi. “Does the IEEE 802.11 MAC protocol Work Well in Multihop Wireless Ad Hoc Networks?” IEEE Communication Magazine, Vol 39, N. 6, June 2001, pp 130-137.
- [14] Z.Tang and J.J.Garcia-Luna-Aceves. “A protocol for topology-dependent transmission scheduling in wireless networks”. In Proc. of IEEE Wireless Communications and Networking Conference, 1999.
- [15] N.H. Vaidya, P Bahl, and S. Gupta. “Distributed fair scheduling in a wireless LAN. In Mobile Computing and Networking, pages 167-178, 2000.
- [16] Xue Yang and Nitin H. Vaidya. “DSCR: A More Stable MAC Protocol for Wireless Networks”. Technical Report, University of Illinois at Urbana-Champaign, August 2002.

Digital Radio Mondiale Applications and Architecture

**Edmund Coersmeier, Marc Hoffmann,
Martin Kosakowski, Maxim Lobko, Yuhuan Xu**
Nokia Research Center, Meesmannstrasse 103, 44807 Bochum, Germany

firstname.lastname@nokia.com

Abstract – Digital Radio Mondiale is an ETSI-based broadcast technology, which replaces traditional analog AM by digital techniques. The broadcasting takes place in the long, medium and short wave bands and thus provides a very large coverage per transmitter. The Digital Radio Mondiale belongs to narrow band systems and is designed for low data rates and low power consumption. Digital Radio Mondiale is favourable for portable and mobile devices and can enrich future media products with good sound quality and digital broadcast services. By means of clever assignment of adequate applications to Digital Radio Mondiale it is possible to accent the advantages of the OFDM technology and to utilize optimally infrastructure and bandwidth.

Index Terms – Digital Radio Mondiale, broadcasting, digital services, OFDM, direct conversion, IQ quadrature error

I. Introduction

Digital Radio Mondiale [1] is the successor of analog AM broadcast system. The technology can be implemented into portable and mobile devices to offer FM-like audio quality and digital broadcast services. Since several standards for digital communication have been established during the last years it is often difficult to decide, which service can be best applied to a certain technology. Digital Radio Mondiale has been designed for narrow band transmission in the long, medium and short wave bands in between 150kHz-30Mhz. The channel bandwidth can differ from 4.5kHz, 5kHz, 9kHz, 10 kHz up to 18kHz, 20kHz, respectively. For analog FM-like sound quality a Digital Radio Mondiale data rate about 20-24kBit/s is required [2]. Digital Radio Mondiale technology is favourable for portable and mobile devices because it is based on OFDM to allow mobility, low power consumption and reliable outdoor and indoor reception. Digital Radio Mondiale requires only a few transmitters to enable low data rate regional, national and international broadcasting with coverage of several hundred and thousand kilometer per transmitting site.

This paper analysis possible application categories advantageous to Digital Radio Mondiale, emphasis Digital Radio Mondiale technology ability and proposes a first step towards optimal low power, low cost receiver architecture. The Digital Radio Mondiale implementation proposal concentrates on the analog

front-end and digital compensation algorithm for IQ quadrature error.

II. Digital Radio Mondiale Application Categories

The applications for Digital Radio Mondiale are depending on the available data rate and thus might differ to applications known from high data rate technologies. Figure 1 positions Digital Radio Mondiale with regards to data rate versus coverage. Digital Radio Mondiale provides low data rates and a very large coverage. Instead of high data rate technologies have significant lower reach.

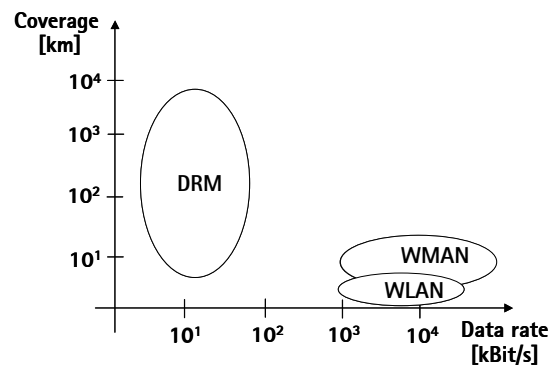


Figure 1 Data rate vs. coverage.

First, when mapping applications to a wireless technology the required data rate needs to be investigated. This parameter indicates the limits of

Digital Radio Mondiale. The next driving parameter concentrates on the target area. In case of regional, national or international broadcasting with only a few transmitters Digital Radio Mondiale is advantageous compared to other technologies.

Table 1 indicates which application categories are favourable for Digital Radio Mondiale and which services are available with traditional analog AM broadcasting..

Digital Radio Mondiale	Analog AM
Regional, national and international FM-like quality for voice and music broadcasting.	Regional, national and international low quality voice and music broadcasting.
Multilingual speech.	Monolingual speech.
Image- and text-based information services.	-
Low rate data broadcasting.	--

Table 1 Application categories for Digital Radio Mondiale and analog AM.

Digital Radio Mondiale concentrates on FM-like audio transmission and prioritizes first of all to deliver high quality audio for speech and music. Based on the advantages of long, medium and short wave the broadcasters address national and international audience and thus Digital Radio Mondiale supports multilingual speech transmission in parallel. Digital Radio Mondiale is not advantageous for video transmission but can deliver beside the audio signal information text, images and multimedia content. Also low rate data broadcasting is possible, which on the one hand can be addressed directly to the user's Digital Radio Mondiale device or on the other hand linked to other equipment as well. Analog AM is at first designed for large coverage with the disadvantage of low quality audio compared to analog FM or Digital Radio Mondiale.

Based on the advantages of Digital Radio Mondiale it is important to investigate the technology and possible implementation architectures. Thus the next sections summarize the Digital Radio Mondiale technology and present a first step towards low cost, low power receiver architecture

III. Digital Radio Mondiale Technology Overview

Digital Radio Mondiale technology is based on ETSI standard [1] and supported by ITU. In [2] a comprehensive technology overview from the ETSI standard perspective is given. Comparing Digital Radio Mondiale technology with analog AM broadcasting there is a significant improvement for the system robustness in fading channels, a much better audio quality because of MPEG-4 based Advanced Audio Coding (AAC), Code Excited Linear Prediction (CELP), Harmonic Vector eXcitation Coding (HVXC) and the enhancement Spectral Band Replication (SBR). CELP and HVXC coders are adequate for low and ultra low bit rates whereas AAC coder works in a wide range of bit rates [2].

Based on source or pre-encoding four different audio or data streams, respectively, can be transmitted via the logical Main Service Channel (MSC). In parallel two information channels Fast Access Channel (FAC) and Service Description Channel (SDC) provide decoding and additional service information.

In the physical layer an OFDM scheme operating in four different modes is applied. FFT sizes with 288, 256, 176 and 122 sub-carriers, respectively, are specified to tackle the different propagation conditions. Sub-carrier spacing is about 50Hz [2]. Symbol mapping is based on 4-, 16- and 64-QAM. Each of the four different modes corresponds to an own OFDM symbol duration. The system employs three different guard times.

Based on this general Digital Radio Mondiale technology overview the next section describes possible receiver architecture.

IV. Digital Radio Mondiale Receiver Architecture

Digital Radio Mondiale receiver architecture can be separated into the analog front-end, digital base band and finally the source decoding. Figure 2 shows the general setup.



Figure 2 Digital Radio Mondiale Receiver architecture

After down conversion of the received signal into the base band the OFDM demodulation starts. First the signal needs a coarse synchronization and is afterwards FFT converted. After fine synchronization and channel correction the channel decoding finalizes the digital base band activities. Source decoding output provides the audio and data information.

Concentrating on the analog front-end there can be assumed different architectures. Historically heterodyne architectures are good candidates for analog AM signal reception, shown in Figure 3.

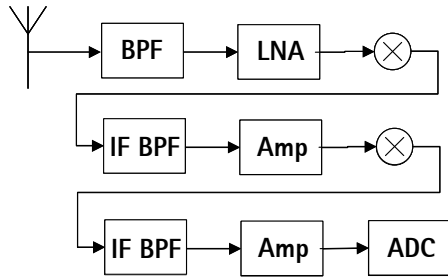


Figure 3 Heterodyne architecture for analog front-end.

Heterodyne architectures can guarantee good signal quality but generate high costs because several different, accurate components are required.

To come up with low cost and low power front-end one can investigate a direct-conversion architecture. Direct conversion front-end has reduced costs compared to heterodyne architectures because less complex components are required. Overall power consumption is lower compared to heterodyne architecture as well.

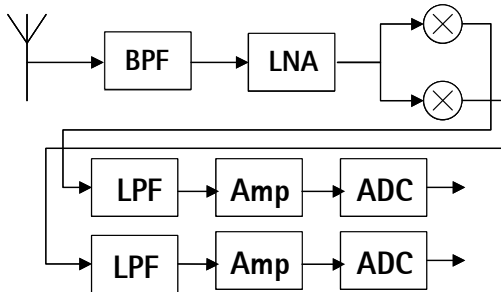


Figure 4 Direct conversion analog front-end.

In case of direct conversion architecture the image rejection requirements are more relaxed compared to traditional heterodyne topology, because the image and the desired signal are located at the same level. The received signal at the antenna is pre-filtered by the band pass filter. Further signal amplification follows by the low noise amplifier. Afterwards the signal is down converted to the IQ base band. If the mixers do not provide a high quadrature quality further digital signal processing for quadrature error correction is required.

In the base band additional amplification and low pass filtering are employed. Power efficient ADCs deliver the analog Digital Radio Mondiale signal to the digital

base band domain. Comparing the two approaches leads to Table 2, which lists advantages and disadvantages.

	Direct conversion architecture	Heterodyne architecture
Advantages	Image rejection relaxed	Good image rejection
	Lower costs	High selectivity
	Lower power consumption	
Disadvantages	DC offset	Higher costs
	1/f noise	Higher power consumption
	IQ quadrature error	
	IQ amplitude error	

Table 2 Comparison between direct conversion and heterodyne architecture.

Both approaches can sufficiently handle the image rejection. The direct conversion approach because the image rejection requirements are more relaxed, the heterodyne architecture because of its powerful filters. Additionally the heterodyne architecture provides a high selectivity because of different filter and amplification stages. Lower costs and lower power consumption give added value to the direct conversion architecture. Comparing the disadvantages the direct down conversion architecture provides several technical problems whereas the heterodyne architecture consumes more power and has higher costs. DC offset and 1/f noise are important parameters especially when talking about carrier frequencies below 30Mhz. IQ quadrature and amplitude errors can cause significant quality impairments.

However when talking about portable and mobile consumer devices power consumption and costs are important parameters for a successful product. Thus it needs to be investigated whether the technical disadvantages of the direct conversion architecture can be solved without significant costs or power increase.

Only if all technical imperfections have been solved either in the analog domain or with additional compensation algorithms in the digital domain the direct down architecture becomes a reliable candidate. Otherwise higher costs and more power consumption need to be accepted when employing a heterodyne architecture for a reliable implementation.

This paper provides in the next section results to remove frequency selective IQ quadrature imperfections.

V. IQ Quadrature Imperfection Compensation

Assuming that no other imperfection than IQ quadrature error exists in the OFDM demodulator then the received signal includes the IQ quadrature error of angle φ .

$$s(t) = I(t) + Q(t) \cdot \sin(\varphi) + j \cdot Q(t) \quad (1)$$

The closely related IQ amplitude imbalance error between the I-branch and the Q-branch of factor $\cos(\varphi)$ has been neglected. The adjustment procedure is a combination of two consecutive steps:

1. Error detection
2. Error correction

The here presented system [5], [7] will be installed as a feedback loop system, Figure 5.

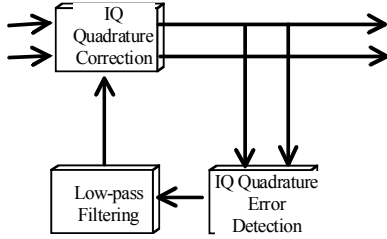


Figure 5 IQ quadrature error compensation loop.

The incoming IQ samples are corrected first. After that the remaining error is calculated and low-pass filtered. When the whole IQ quadrature error is compensated the loop remains in equilibrium.

The digital blind error detector will apply the following mathematical considerations. If the I- and Q-branch samples are statistically independent the expectation of their product equals zero.

$$E\{I[n] \cdot Q[n]\} = 0 \quad (2)$$

In that case the adjustment block will do no corrections at all. Equation (3) indicates when an IQ quadrature error appears.

$$\begin{aligned} E\{(I[n] + Q[n] \cdot \sin(\varphi)) \cdot Q[n]\} &= \\ E\{I[n] \cdot Q[n]\} + E\{Q[n] \cdot \sin(\varphi) \cdot Q[n]\} &= \\ E\{Q^2[n] \cdot \sin(\varphi)\} &= \\ E\{Q^2[n]\} \cdot \sin(\varphi) = \sigma_Q^2 \cdot \sin(\varphi) \approx \sin(\varphi) \end{aligned} \quad (3)$$

The first addend in the second line of equation (3) results in the value zero. The remaining expectation value will be proportional to the error value $\sin(\varphi)$. The expectation of the factor $Q^2[n]$ provides the Q-branch mean power and can be interpreted as an amplification factor, because it will have always a positive sign. This result will be used to correct the incoming signal stream. It is assumed that one or both analog base band filters provide imperfections depending on their time domain impulse response and their frequency transfer function, respectively. These imperfections could be one or more items like amplitude ripple, non-linear filter phase behavior or filter ISI. Thus it is necessary to implement an IQ quadrature error detector, which is frequency selective and capable of covering analog filter imperfection afflicted \tilde{I} and \tilde{Q} symbols. Equation (4) describes the mathematical operations.

$$\begin{aligned} e_i[n] &= \tilde{I}[n - (N-1)/2] \cdot \tilde{Q}[n - (i-1)] \\ i &= 1, 2, \dots, N \end{aligned} \quad (4)$$

In this paper it is assumed that N is an odd number and the index of the error value is valid from 1 to N . Figure 6 presents a possible implementation setup. The center-tap $(N-1)/2$ of the I-branch will be multiplied with N different values from the Q-branch.

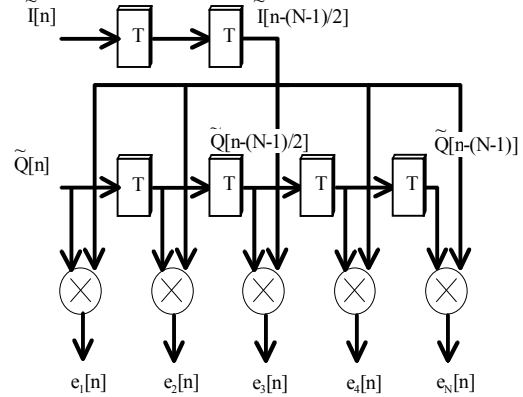


Figure 6 Frequency selective IQ quadrature error detector without integrators. $N = 5$.

Each error value $e_i[n]$ will be low-pass filtered by its own integrator.

$$c_i[n] = \mu \cdot \int_{k=0}^n e_i[k] \, , \quad i = 1, \dots, N \quad (5)$$

This is shown by Figure 7 as well.

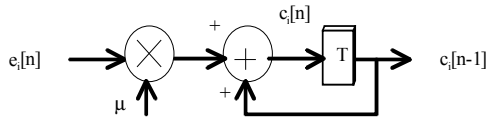


Figure 7 N different IQ quadrature integrator low-pass filters to provide N coefficients $c_i[n-1]$.

The multiplier in Figure 7 is able to define the overall loop bandwidth or loop adjustment speed, respectively. Typically the quadrature error adjustment speed can be chosen to a small value because quadrature error as well as analog base band filter properties will change slowly in time. The correction will be done by equation (6) and Figure 8.

$$\tilde{I}[n-(N-1)/2] = \left[\tilde{I}'[n-(N-1)/2] - \sum_{i=1}^N c_i[n-m] \cdot \tilde{Q}[n-(i-1)] \right] \quad (6)$$

$i = 1, 2, \dots, N$ and $m > 0$

The variable m describes the implemented loop latency. Similar to a channel equalizer the Q-branch values from the tap-delay line are multiplied with the corresponding correction coefficients c_i and are summed up. This result is subtracted from the imperfect I'-branch center tap. In the case the Q-branch instead of the I-branch contains the IQ quadrature error the whole setup can be used without any changes but the corrected IQ diagram in the frequency domain will be phase shifted. This problem can be solved by the OFDM phase tracker.

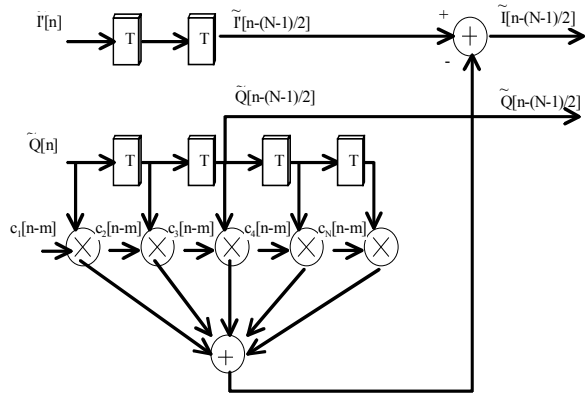


Figure 8 Freq. selective IQ quadrature correction. $N=5$.

The Figure 9 - Figure 11 show different IQ diagrams in an OFDM environment. Analog low-pass filters insert imperfections to the IQ data stream after an IQ quadrature error has been inserted from the down modulation process. Figure 9 shows the results of the non-frequency selective adjustment loop. The

adjustment loop with $N=1$ uses only one single coefficient. Thus the loop is not able to correct the incoming signal equally on all frequencies.

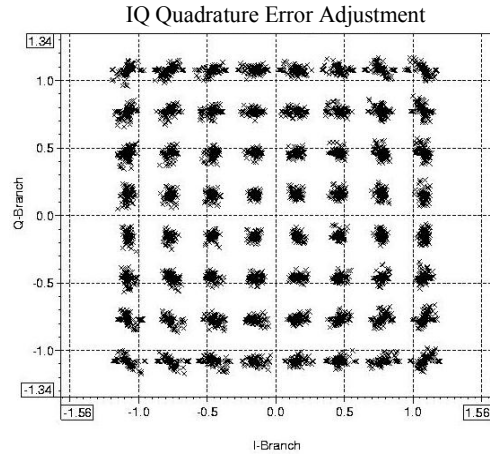


Figure 9 IQ diagram with the non-frequency selective IQ quadrature adjustment. IQ error $\phi = -10^\circ$.

This non-frequency selectivity results in sub-optimal 64-QAM constellation points. Figure 10 presents the adjustment with 3-tap quadrature loop. The loop is frequency selective and corrects the quadrature error at different frequencies with different correction values.

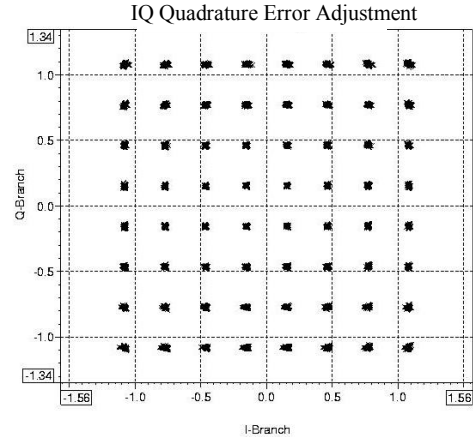


Figure 10 IQ diagram with 3-tap frequency selective IQ quadrature adjustment. IQ error $\phi = -10^\circ$.

One can see that compared to Figure 9 the 64-QAM constellation points have been improved but still are not optimal. Thus it requires more than 3-taps to reach optimal quadrature error compensation. Figure 11 presents the adjustment done with a 7-tap frequency selective compensation loop.

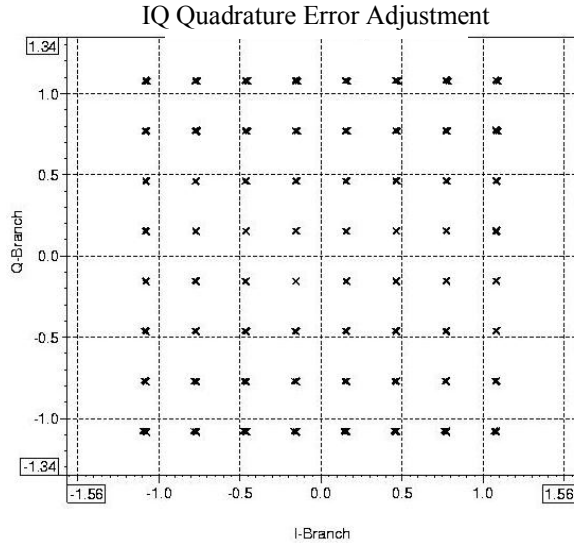


Figure 11 IQ diagram with 7-tap frequency selective IQ quadrature adjustment. IQ error $\varphi = -10^\circ$.

It can be seen that the constellation point precision has been optimized and no further quadrature impairments can be identified.

To achieve those constellation diagrams it is necessary to equalize the analog filters in the data stream itself, too. It means after the quadrature compensation there is applied a post-equalizer in the I- and Q-branch as well. This is true for Figure 9-Figure 11. Figure 12 and Figure 13 show the starting phase of the feedback loop by visualising the corresponding coefficient processing. Figure 12 provides the convergence behaviour of the non-frequency selective IQ quadrature loop (corresponding to Figure 9). Thus one coefficient adaptation curve is visible. The final value of the coefficient should be $\varphi = -10^\circ$ but because of the frequency selective influence from the analog base band filters the adjustment loop identifies a different value, about $\varphi = -8.7^\circ$. This value is about $\Delta\varphi = 1.3^\circ$ too large but the 1-coefficient loop is not able to recognize the analog base band filters' influence.

This changes with the frequency selective adjustment loop. Figure 13 shows a 7-coefficient setup, which corresponds to Figure 11. The most significant coefficient converges to the value about $\varphi \sim -9.0^\circ$, which is similar to the value of Figure 12. Additionally the residual coefficient values take values unequal to zero and thus provide frequency selective information to the adjustment loop. There are not 6 different curves visible because some of the coefficients converge towards equal values.

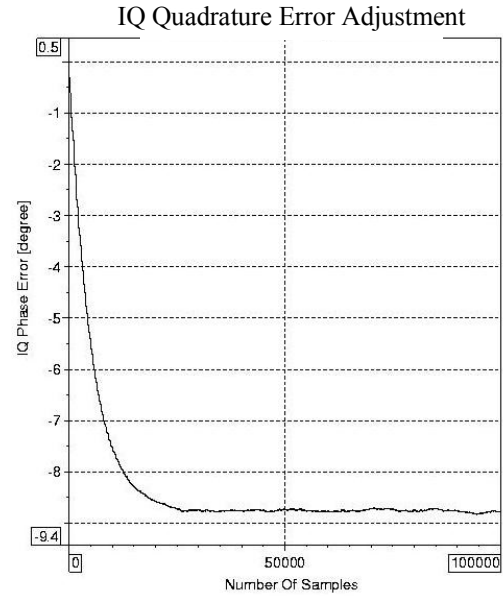


Figure 12 Non-frequency selective IQ quadrature imbalance error curve. Generated IQ error $\varphi = -10^\circ$. Imperfect analog filters in use.

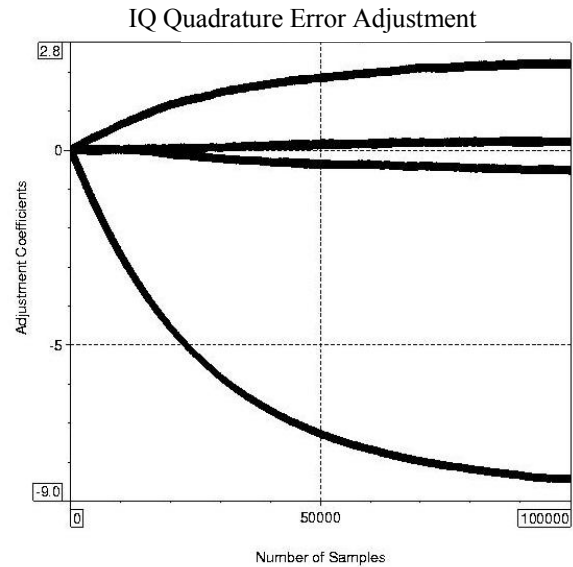


Figure 13 7-coefficients. IQ quadrature error $\varphi = -10^\circ$. Several coefficients have a very similar value thus only 4 different curves instead of 7 curves are visible.

VI. Conclusion

This paper gives an overview about the Digital Radio Mondiale technology and categorizes possible Digital Radio Mondiale applications, which are appropriate for this new digital broadcast technology. To achieve good cost and power consumption figures for a portable or

mobile receiver implementation it is necessary to choose best fitting architecture. With regards to the analog front-end a direct conversion front end provides low cost and low power figures but introduces several unwanted signal imperfections. Thus it has been shown a first step to eliminate some of the imperfections. Concentrating on IQ quadrature imperfections a frequency selective IQ quadrature compensation algorithm for an OFDM system has been presented. In future one needs to investigate how to remove the other imperfections successfully as well without increasing the costs or power consumption significantly.

References

- [1] Digital Radio Mondiale (DRM) System Specification, European Telecommunication Standards Institute (ETSI), ETSI TS 101980, 2001
- [2] Digital Radio Mondiale (DRM) Digital Sound Broadcasting in the AM Bands, F. Hofmann, C. Hansen, W. Schäfer; IEEE Transactions on Broadcasting, Vol. 49, No.3, September 2003
- [3] High Precision Analog Front-End Transceiver Architecture for Wireless Local Area Network, Edmund Coersmeier, Yuhuan Xu, Ludwig Schwoerer, Ken Astrof, 6th International OFDM-Workshop 2001, Hamburg
- [4] Adaptive Filter Theory, Simon Haykin, Prentice Hall, Third Edition, 1996
- [5] Frequency Selective IQ Phase Imbalance Adjustment in OFDM Direct Conversion Receivers, E. Coersmeier, IEEE ISCE 02, Erfurt, Germany
- [6] Performance of Differentially Detected $\pi/4$ DQPSK in the Presence of IQ Phase Imbalance, M.Scarpa, IEEE ISCAS 2000, Geneva, Switzerland
- [7] Frequency Selective IQ Phase and IQ Amplitude Imbalance Adjustments for OFDM Direct Conversion Transmitters, Edmund Coersmeier, Ernst Zielinski, ISART 2003, Boulder, USA

A Low Power Methodology for Portable Electronics

Dae Woon Kang¹, James T. Doyle¹, Mark Hartman¹, Sandeep Dhar¹, Marty B. Dermody¹,
Robert C. Woolf¹, Ravindra S. Ambatipudi¹, and Yong-Bin Kim²

¹Portable Power Systems, National Semiconductor, {dae.woon.kang, jim.doyle, mark.hartman, sandeep.dhar, marty.dermody, rob.woolf, ravindra.ambatipudi}@nsc.com,

²Dept. of Electrical and Computer Engineering, Northeastern University, ybk@ece.neu.edu

Tel. (303)-845-4064, Fax. (303)-845-4005

Abstract

This paper introduces a power savings methodology for portable electronics based on PowerWise™ Interface (PWI), leakage handling, and low power amplifier scheme. The PWI system dynamically monitors circuit performance with a slacktime detector, and provides a substantially constant minimum-supply voltage for digital processors to properly operate at a given frequency. Back-biasing preventing current from leaking becomes more important as the technology goes deeper. And the efficiency of power amplifiers affects battery lifetime directly. A dynamic power model is presented to calculate power savings and battery life time according to systems topology.

I. INTRODUCTION

The trend in electronics is to move more and more of the computing resources into portable and wireless applications. As a result, portable power handheld products and computers now account for nearly 10% of humanities' total power usage [1]. Battery charging efficiencies and cost of manufacturing power delivery systems and components represent a significant portion of this 10%. On the other hand, manufacturers are quickly realizing that size and run time demands of their portable equipment cannot be met by increasing energy density in batteries.

To achieve the leap in wireless devices' functionality and run time, first of all, manufacturers are turning to high efficiency energy management of system functions. Decreasing the amount of energy it takes to complete a function means the battery has more energy left for other processes. This energy management philosophy is being developed for several common wireless functional blocks such as DSP and micro-processor. Next, shrinking integrated circuit technology (<0.13μm) causes the amount of leakage to equal or exceed the portion of dynamic power dissipation. Unless the leakage is reduced, the power delivery in deep submicron era will ultimately restrict the ability of handheld to meet the customer demand for improved capabilities. Third, presently 2.5G GPRS/EDGE GSM which dominates the cellular market place uses relatively low efficiency linear power amplifiers (PAs) to achieve the required data rate as well as to meet multimode and multiband requirement. PAs play a critical part in determining

the power efficiency of a RF system because of their high output power levels, which can reach 3W for some cellular systems. As a result, the design of highly efficient PAs is of great importance to extend battery life time.

The focus of this paper is to address these power management approaches to save power for portable electronics at the aspects of dynamic power (Section II), static power (Section III), and RF PA power (Section IV). Section V shows a dynamic power model to demonstrate power savings and battery life time according to systems topology.

II. POWERWISE™ ADAPTIVE VOLTAGE SCALING

A. Overview

As semiconductor process technology has become lower-voltage and deeper sub-micron, and the number of transistors per chip has increased according to Moore's Law, two critical circuit design issues are presented: 1) the non-uniformity of process parameters within a single die; and 2) the increment of power consumption per die [2]-[3]. In deep sub-micron circuit design, variations due to the first issue cause differences in transistor and interconnect characteristics across a single die. They in turn impact the performance of circuit since they generate deviations in MOSFET drive current, resulting in propagation delay distributions of the critical path across a chip. Furthermore, the distribution of process parameters expands from die to die within a wafer as well as a lot. After fabrication, operating variations such as power supply voltage, chip temperature and across-chip temperature also affect the propagation delay. By combining both operational and process

induced variations, the propagation delay fluctuates to 18% ~ 32% [2]. The yield of CMOS logic circuits satisfying a specific performance requirement is significantly influenced by the magnitude of critical path delay deviations due to both operational and intrinsic parameter fluctuations. To compensate the impact of these parameter fluctuations and to achieve a desired yield, there are two approaches: 1) to reduce performance by operating at a lower clock frequency, and 2) to increase the supply voltage.

While the operating frequency limits allowable propagation delay, this delay strongly depends on intrinsic process parameters, supply voltage and junction temperature (PVT). The propagation delay in a MOSFET is proportional to the product of the active resistance of the MOSFET and load capacitance as in (1).

$$R_{ON} = \frac{V_{DD}}{\beta \cdot (V_{DD} - V_T)^\alpha} \quad (1)$$

$$C_L = C_D + C_G + C_W$$

where α is the velocity saturation term, β is the process transconductance parameter, V_{DD} is the supply voltage, V_T is the threshold voltage, C_D is the drain capacitance, C_G is the gate capacitance, and C_W is the interconnect capacitance.

If a design is fabricated as the best process corner, and is operating at low temperature, it needs less than $\frac{3}{4}$ of the minimum supply voltage required at the worst case [2]. Process parameters and operating junction temperature are not controllable, but supply voltage is. This results in chances to reduce power consumption by adjusting supply voltage with regard to process and temperature.

In many portable-computing devices such MP3-players and digital cameras, the full processing power of a processor is not required all the time. There are certain times when an operating frequency can be reduced, and a lower frequency means a longer allowable delay. This longer time margin also allows a supply voltage level to be lowered whereas the applied lower voltage increases the propagation delay. Since power consumption is quadratic with the supply voltage and proportional to operating frequency, reducing both operating frequency and supply voltage allows an excellent energy-efficient operation. This technique, adaptive voltage scaling (AVS), decreases power consumption without sacrificing performance provided performed tasks are finished within the allowed time. From the trade-off between performance and energy consumption, supplying just enough voltage to a system at a given frequency represents its optimum power consumption [4]-[7].

B. PowerWise™ Interface

National Semiconductor and ARM have teamed up to develop an AVS energy management system that reduces the energy consumption of a processor to the minimum amount for

a given clock frequency. This AVS method uses an open standard 2 wire, low power communication interface between the processor and the energy management unit (EMU) called PowerWise™ Interface (PWI) [8]. The two combined provide for an intelligent and aware energy management system.

AVS in the general sense refers to a power supply rail that adjusts its voltage corresponding to the demands of its load which could be any compliant electronic device. The enormous benefit of AVS is that for completing the same function, an AVS compliant processor will use 30% to 60% less energy than a fixed voltage processor. This is the kind of energy savings that will allow heavier use of our wireless devices while maintaining the operating time between charges. The way to reduce energy consumption in a processor, then, is to not only to reduce the clock frequency as low as possible, but, more importantly, to reduce the core supply voltage to the minimum amount for a given clock frequency.

The goal of the AVS system is to reduce the supply voltage to the minimum amount and still maintain critical path deadlines. Open loop AVS accomplishes this by regulating the supply voltage to a pre-characterized value that guarantees operation over process, temperature, and power supply variations. However, regulating to a pre-characterized voltage does not guarantee minimum energy consumption. This is guaranteed when the maximum propagation delay (and thus minimum voltage) is present for any given situation (frequency, process, temperature). Closed loop AVS in fact accomplishes this by regulating the propagation delay margin. In other words, no matter what lot the processor is from, nor the temperature or frequency it is operating at, the specified delay margin is maintained. Because of the voltage/propagation delay relationship, this condition necessarily requires that the supply voltage be at the acceptable minimum at all times.

As shown in Figure 1, the closed loop AVS system developed by National Semiconductor and ARM has two hardware components: the Intelligent Energy Manager™ (IEM) and Adaptive Power Controller (APC), located in the processor, and the AVS compliant energy management unit (EMU). The ARM IEM determines the minimum performance (clock frequency) required by the processor for given tasks. The APC accepts a performance request from the IEM and determines the minimum voltage the processor can operate at for that performance level. It also commands the EMU to attain the lowest supply voltage for a given clock frequency. It is important to realize that the APC is synthesizable code operating in the processor, and it manages the IEM requests and voltage control without any intervention from the processor. All the software hooks for controlling performance are contained in the IEM. The APC controls the supply voltage transparent to the IEM, however it is coupled to the external EMU. The AVS EMU is equipped to interpret commands from the APC through a new open standard interface, PowerWise™ Interface (PWI). PWI is a low power, 2 pin serial protocol specifically designed to meet the needs of next generation AVS portable systems.

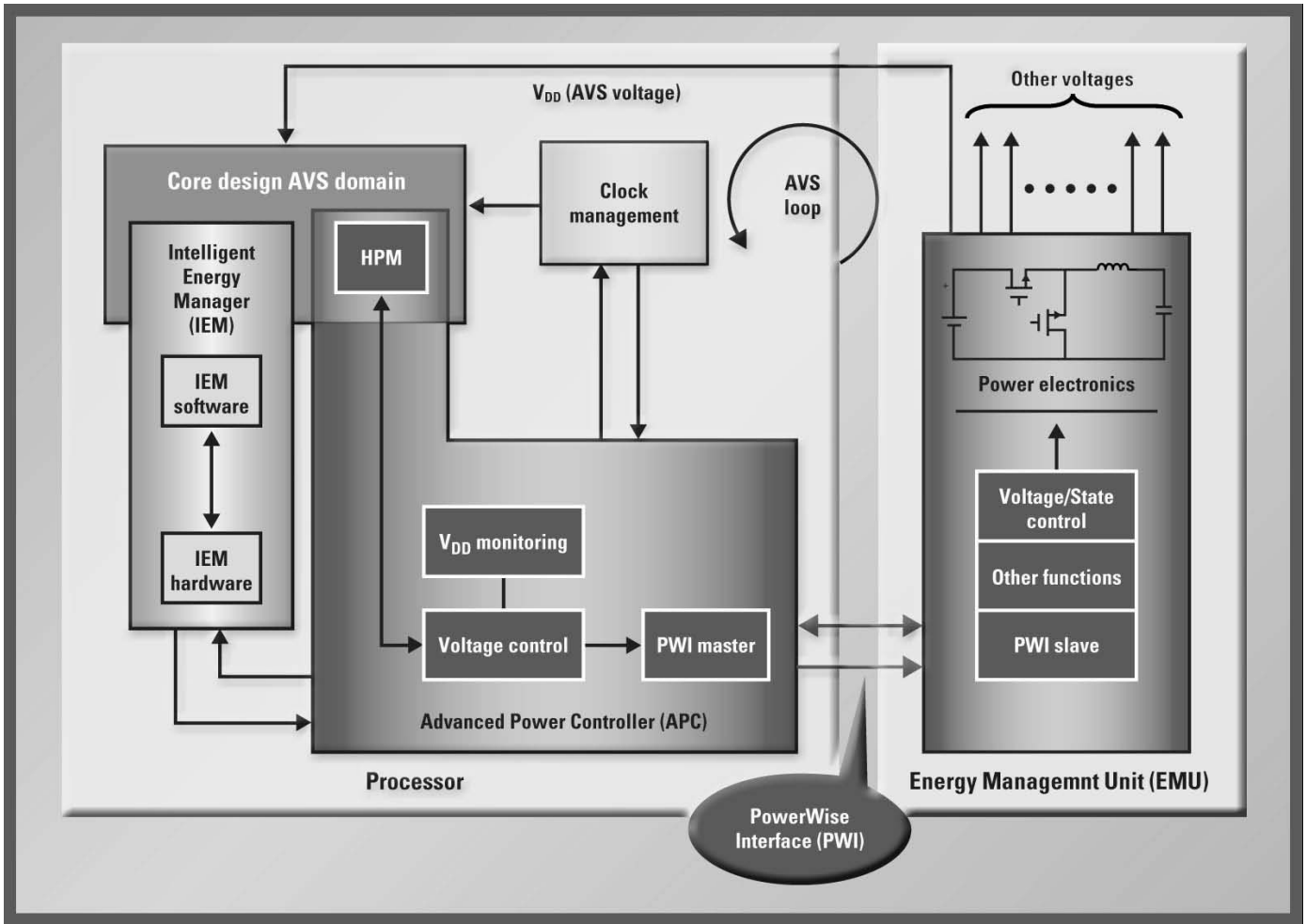


Figure 1: The closed loop AVS system consists of an Intelligent Energy Manager (IEM) and Adaptive Power Controller (APC) located in the processor, and an Energy Management Unit (EMU). The AVS loop regulates propagation delay margin to ensure the minimum supply voltage is achieved.

Closed loop AVS is distinguished from open loop, table based voltage scaling techniques in that it regulates the propagation delay margin in logic cells. In this system, the power supply voltage is a variable that increases or decreases, and the delay margin is a fixed parameter that is regulated over parts, temperature, and clock frequency. Many advantages arise from this methodology. Closed loop AVS relaxes the characterization process. There is no need for characterizing voltage/frequency tables because a delay margin is maintained by the AVS feedback loop. Another incentive is that less demand is placed on power supply regulation. The AVS loop adjusts the supply voltage as necessary, compensating for the $\pm 5\%$ tolerance typically allocated to power supply regulation. By and far the most beneficial advantage is that the minimum operating voltage is realized for all conditions, and can dynamically change as conditions change. For example, as temperature changes, the cell delays change, and the thus the voltage is adjusted to maintain the same delay margin – without using voltage/frequency tables. Figure 2 shows this temperature compensation. By using closed loop AVS, the designer can rest assured that the minimum voltage is being applied for any given clock frequency.

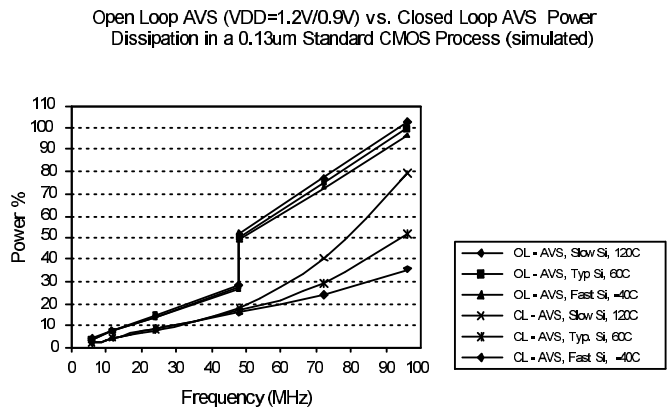


Figure 2: Power dissipation comparison between Open loop AVS and Closed Loop AVS. Closed loop AVS tracks temperature and process variations, and yields significant power savings over open loop AVS. The open loop method uses two voltages (1.2V between 48MHz and 96MHz, and 0.9V between 6MHz and 48MHz).

III. BACK BIAS LEAKAGE CONTROL

A. Overview

Static power dissipation is an important design criteria for modern battery powered portable devices since it can significantly impact battery life. The current trend is for increasing functionality in portable devices provided by complex system-on-chip ICs implemented in deep submicron processes. With each step improvement in transistor scaling the supply voltage V_{dd} is lowered. However in order to maintain performance, the threshold voltage V_t is also lowered to compensate for scaling in V_{dd} . This has the undesirable effect of increasing the drain-to-source leakage current I_{off} when the application IC is in standby. In some applications such as cell phones the static power dissipation of the application ICs can be as much as 40% of the total static power dissipation.

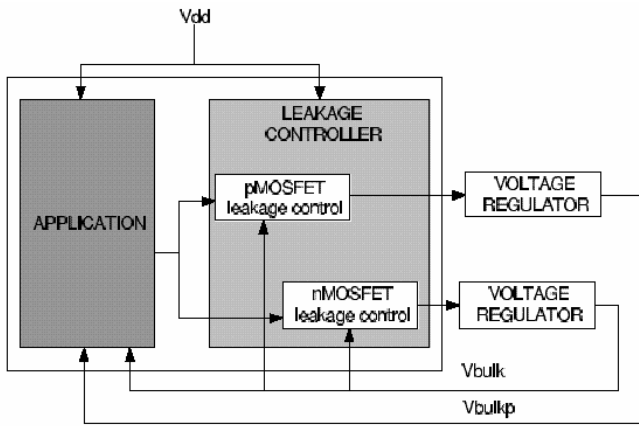


Figure 3: Block diagram of scheme with well regulators.

The use of reverse body bias has been demonstrated as an effective technique to reduce the total static power dissipation by reducing the total static (leakage) current I_{off} drained from the battery [9]. This technique is attractive to implement (as shown in the high level block diagram of Figure 3) because of the following reasons:

- it requires no changes to the process technology and can be used along with other leakage control techniques,
- the reverse body bias regulator can be implemented off-chip and incorporated into an external power management IC,
- it allows state retention upon standby (as opposed to another leakage control technique known as MTCMOS [10]),
- alternatively forward body bias can be applied to an application IC to reduce V_t variations at the expense of increased active power consumption [11].

B. Approaches on-chip leakage monitoring

1) Leakage controlled sawtooth oscillator

The circuit shown in Figure 4 can tightly track the on-chip leakage current. A simple sawtooth oscillator design is driven by the leakage current of an NMOS transistor¹. During each cycle, the capacitor C_{gs} is charged to V_{dd} . When transistor M_1

¹ for a PMOS transistor the leakage current source is.

is turned off, leakage current will discharge the capacitor until the threshold voltage of the inverter is reached. To prevent crowbar current between V_{dd} and ground, a positive feedback circuit rapidly switches the inverter output. After a small delay the capacitor is again charged to V_{dd} and the cycle repeats. The sawtooth input to a counter. The counter operates off a low-frequency clock (such as 32kHz). Hence, the output of the counter is a digital number that is proportional to the leakage current. Simple control logic sweeps the reverse body bias voltage and determines the smallest count that represents the optimal body bias voltage.

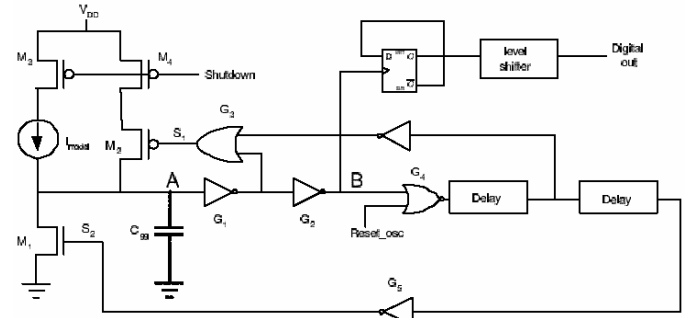


Figure 4: Leakage controlled sawtooth oscillator.

The circuit is capable of detecting leakage currents from a few pAs to several hundred μ As. At the lower end of the scale, the circuit can take long evaluation time (slow discharge of C_{gs}) and at the upper end of the scale, a high frequency oscillation consumes dynamic power from V_{dd} . Hence, the scheme is designed with split current sources that can be selected to provide the appropriate current input for a given operating condition (V_{dd} and temperature). The worst case average current consumed over $10ms^2$ evaluation period is determined by simulation to be $75 \mu A$.

2) Curve traversal algorithm for leakage detection

The circuit shown in Figure 5 will follow the leakage current curve until a reversal in the leakage current slope is observed (Figure 6). The circuit interprets this reversal to imply that (a) prior to the reversal, increasing the reverse body bias voltage decreases I_{off} and (b) after the reversal, increasing the body bias voltage increases I_{off} since the optimum body bias point has been crossed. Following this reversal the circuit will attempt to fine-tune the body bias voltage. The tradeoff is between a body bias voltage close to the optimum value and increased time for leakage current evaluation. The circuit is initialized at a body bias voltage which is lower than the optimum voltage. Traversal is simply done by determining whether the capacitor C_1 can be discharged in a given sample period. C_1 is comprised of individual capacitors and can be calibrated during initialization by the control logic. The sample period can be changed by factors of 2 to accommodate a larger variation in the leakage current. Once the reversal of leakage current is detected, the

² this simulation model assumes that temperature shifts on chip occur at a time constant of about 10ms. Hence the leakage current monitor could be activated every 10ms. Once the evaluation is complete, the leakage monitor is shut down for the remaining part of the 10ms time period.

sampling time is increased and the leakage current can be monitored with higher resolution. Since there is no oscillator circuit to consumer dynamic power, this circuit scheme consumes very little current. Circuit simulations show an average current consumption of $7\mu\text{A}$ over a period of 10 ms.

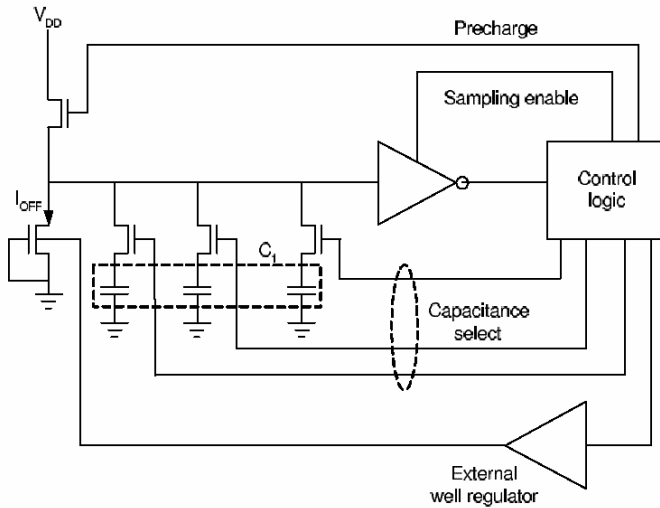


Figure 5: A circuit to traverse the leakage current curve.

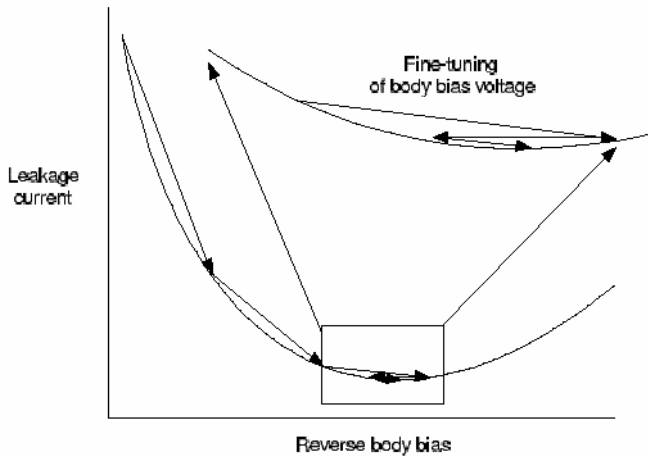


Figure 6: Illustration of the curve traversal.

IV. HIGH EFFICIENCY POWER AMPLIFIERS

A. Overview

The spread-spectrum wireless communications standard, 3GPP, requests stringent specifications for linearity and adjacent-channel power ratio (ACPR). To meet the 3GPP specifications, wideband-code-division multiple access (W-CDMA) wireless handsets require highly linear Class A or Class A-B RF power amplifiers (PAs). The power-added efficiency (PAE) for that type of PA, however, is about 35% maximum at an output power of +28 dBm. And the PA operates discontinuously in voice mode. Without speaking, it runs at a half rate (50 percent of the time) or at a one-eighth rate. However, in data mode the PA runs continuously during transmitting the data transmission. Combining the PA's low efficiency and continuous operation results in discharging the battery quickly, and in turn may cause the phone too hot.

Therefore, the most inefficient circuit in a cell phone is the RF PA.

On the other hand, demand for multi-mode and multi-band operation, longer battery life, and smaller size in cellular equipment is increasing [12]. To achieve multimode operation, transmitters must be able to accommodate constant envelope signals as well as non-constant envelope signals. Therefore, to avoid distortion of non-constant envelopes, conventional transmitters has to employ linear power amplifier (class-A) or use predistortion techniques to linearize slightly saturated amplifiers (class-AB). Linearity and power efficiency in transmit front ends are conflicting requirements demanding innovative solutions for present and future wireless mobile systems. Therefore, the design of PA linearising and efficiency improvement schemes, and architectures, is attracting much attention in recent years, including feedforward, envelope elimination and restoration (EER), Cartesian feedback, and predistortion techniques [13]-[16].

B. A Polar EDGE CMOS PA Controller with Flat Amplitude and Phase Response

A polar modulation is presented for the multimode multiband operation with high power efficiency because it uses amplifiers in the switched mode. The function of a PA control chip is to maintain an accurate control voltage when variations of process, supply voltage, and temperature are encountered. In the presented approach, an open loop method [17] is adopted in order to avoid the problematic issues of closed-loop power control such as non-linearity between PA and detector, small dynamic range, long calibration time, inaccurate current detector, and cost. The open loop design as shown in Figure 7 covers multiband and multimode such as GSM, DCS, PCS, GPRS and EDGE. The goal of this section is to design a high efficiency amplitude-tracking circuit to improve transmit power efficiency versus time. The incentive to use a CMOS technology is based on the per-unit-area cost which is roughly $\frac{1}{2}$ size of BiCMOS.

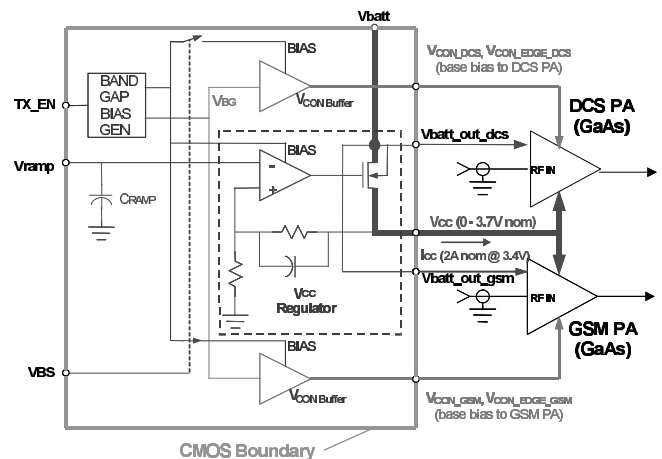


Figure 7: Block diagram of LM4401 with external PAs.

To reduce PSRR and circuit parameter sensitivities, each circuit module provides higher gain than its function requires. First of all, a bandgap ($100\mu\text{A}$) is optimized for both low noise and offset without a trim ($\pm 3\%$ maximum tolerance) where a

folded cascode P-channel input stage eliminates the potential for shut down due to the input stage being biased below the threshold ($V_{be} < V_t$). In order to reduce noise, the bandwidth of the reference is also reduced, and a grounding in the immediate location to the bandgap is utilized to minimize coupling noise. Second, a P-channel shunt circuit is required to make sure no charge on nodes driving outputs. The input circuit uses a trickle subthreshold current source pull-down which is operational during initial power on. The TX enable completely shuts off all supplies and references but allows the input inverter to sense the power. A Schmitt trigger master slave implementation overcomes high standby current ($\sim 10\mu\text{A}$) issues due to a high value resistor tied directly to the 2.9V supply. Third, a folded cascode LDO amplifier with an internal bandgap reference is utilized to make the input stage independent of supply voltage variations. The main power LDO is a circuit capable of delivering more than 4A of peak current to RF PA in a cell phone handset. And a high bandwidth bias amplifier slews in less than $1\mu\text{s}$ into a 56Ω load with a $0.1\mu\text{F}$ capacitance in order to keep the quiescent current relatively constant into a load. Its bandwidth is approximately 2 MHz. An input level shifter utilizes a “diode” connected P-channel in series with an inverter to provide a 1.5 volt to 5.5 volt interface.

The P-channel power driver FET is a unique bent gate approach to maximize conductance and in turn to provide the lowest R_{ds} on per unit area in CMOS without latchup. Next, a power down circuit technique separates the circuit from the supply by using an N-channel switch with all of the analog active circuitry (N-channel sources connected to ground). This technique eliminates the need for an input supply reference voltage required by an inverter input stage. This results in a standby current of less than $1\mu\text{A}$. The sum of the entire power down current and leakage current is less than $10\mu\text{A}$ over process voltage and temperature including pull up and pull down currents. Finally, a dedicate 1.6V reference is developed in order to provide an isolated direct reference for the input clamp. This clamps the input (V_{ramp}) from ever allowing the output to go beyond 3.7V. On the other hand, a separate 2.8V reference is used to set the input stage supply voltage of the VCC amplifier which drives the collector of the PA. Making this node supply independent (fixed 2.8V) results in more controllable operation over the input common mode range ($0\sim 1.6\text{V}$) as well as improved supply noise rejection.

The PA controller has a linear gain (2.6) transfer response from the DAC of a baseband to the output power of the PA to meet multi-mode specifications. The PA control loop is accurate, temperature-stable and fast enough to meet the turn-on time requirement ($< 3\mu\text{s}$). The bias amplifier is able to deliver 50mA with a load capacitance up to $0.1\mu\text{F}$. The supply controller precisely tracks the DAC’s input voltage and slews at better than $2\text{V}/\mu\text{s}$ into 1.7Ω and $0.1\mu\text{F}$ capacitive load. Group delay variation of better than $\pm 10\text{ns}$ over a bandwidth of 10 KHz to 3MHz without trim is realized with a potential to extend this to greater than 7MHz for WCDMA applications as shown in Figure 8. The gain flatness is also dramatically improved when the region of interest is moved near the band edge. This implies that bandwidth of the amplifier is increased. A reproducible and consistent gain characteristic is achieved as shown in Figure 9.

This eliminates the need for costly active trim or calibration. Amplitude variation of less than $\pm 0.3\text{dB}$ has been simulated to 3MHz which meets the requirements for WCDMA and polar EDGE. Since the bandwidth and gain of an amplifier are inversely related to the square root of the area divided by the current through the device, reducing the input channel length effectively increases the bandwidth without adversely affecting the gain. Interbusrt recovery time is less than $3\mu\text{s}$ settling to 10mV with typical off standby current less than $10\mu\text{A}$ including input control pull down current. Spot noise at 10MHz is less than $24\text{nv}/\text{rtHz}$ for both the bias amplifiers and the power LDO.

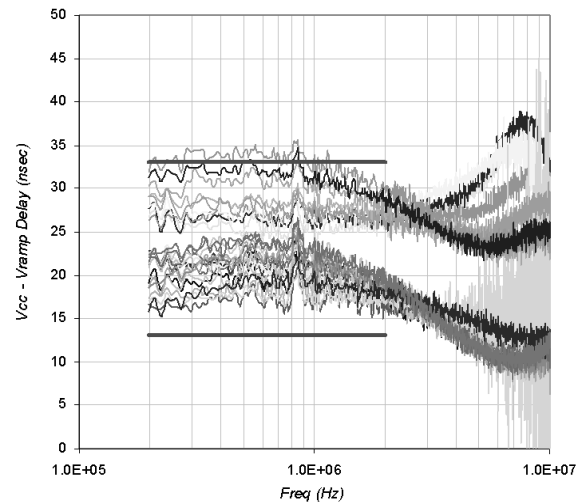


Figure 8: Group delay vs. frequency (Measured).

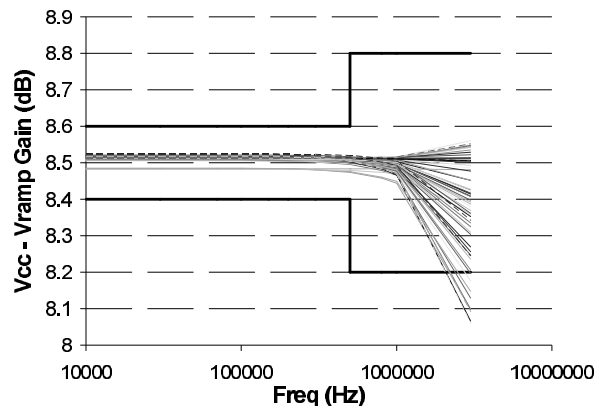


Figure 9: Gain vs. frequency.

Table 1. Features of LM4401.

Key Performance Limits	Min	Max	Units	Condition
Gain Flatness	-0.3	0.3	dB	1.5MHz
Group Delay Flatness		2	ns	
Group Delay Variation	10	30	ns	
LDO Bandwidth	25		MHz	
LDO Noise		24	nV/rtHz	10MHz
Power down Leakage		10	μA	
Supply Voltage	2.9	4.8	V	
Total Standby Current		25	mA	
Turn Time		3	μs	
Load Resistor		1	Ω	2A
Load Capacitor		0.1	μF	

And the linear CMOS controller is used with GaAs HBTs PA providing better than 37 dBc adjacent channel performance with peak efficiency of better than 40%. Not only does this power controller provide up to 4A supply current, it also is able to properly perform with temperatures exceeding 150°C. Significant improvement in performance, noise, and linearity are realized together with ruggedness enhancements including input clamping and offset techniques as shown in Table 1.

V. DYNAMIC POWER MODEL

A dynamic Excel spread sheet as shown in Figure 10 is developed to show the improvements as a function of system demand requirements. Sliders or radial buttons and be used to determine the battery life in Figure 11 and power savings in Figure 12.

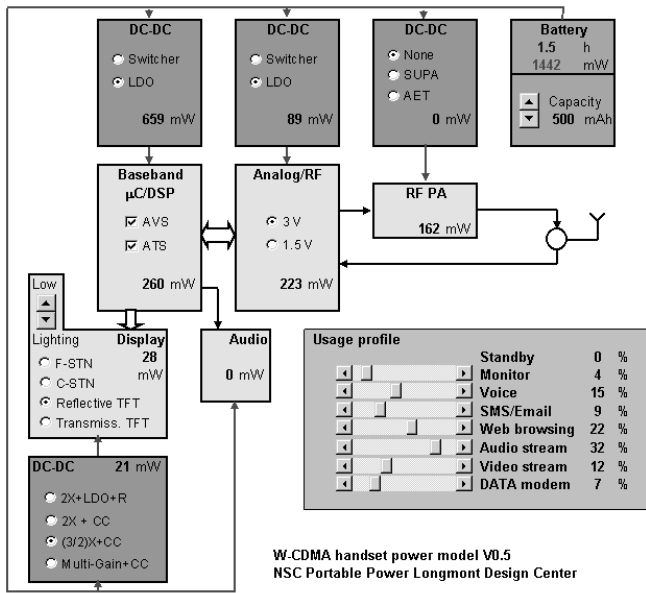


Figure 10: Dynamic Power Model.

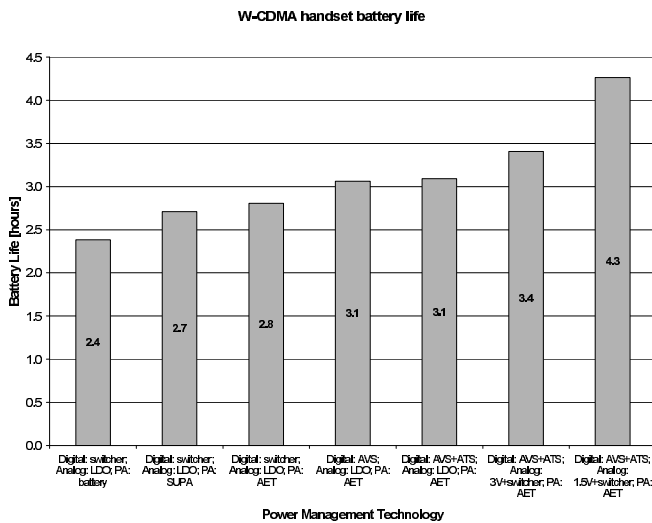


Figure 11: An example of batter life time calculation.

W-CDMA handset power consumption

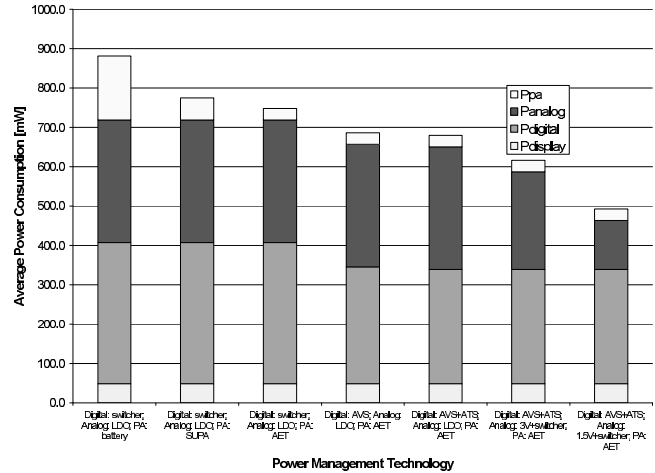


Figure 12: An example of power savings calculation.

VI. CONCLUSION

A power savings methodology for portable electronics has been described. Portable electronics are very vulnerable as the technology trend goes more complex and more power consuming because their power source depends on only a battery. Therefore, high energy-efficiency systems are required to extend the battery life time. Moreover, as RF products among consumer electronics are prevalent, their energy consumption becomes a major part of total home electronic products. As a result, the presented low power methodology will leverage the energy savings of total consumer electronics.

REFERENCES

- [1] Energy Use of U.S. Consumer Electronics at the End of the 20th Century, <http://eetd.lbl.gov/EA/Reports/46212>
- [2] K. A. Bowman, Xinghai Tang, J.C. Eble, and J.D. Meindl, "Impact of Extrinsic and Intrinsic Parameter Fluctuations on CMOS Circuit Performance," *IEEE J. Solid-State Circuits*, vol. 35, no. 8, pp. 1186-1193, Aug. 2000.
- [3] J. W. Tschanz, J.T. Kao, S.G. Narendra, Raj Nair, D.A. Antoniadis, A.P. Chandrakasan, and Vivek De, "Adaptive Body Bias for Reducing Impacts of Die-to-Die and Within-Die Parameter Variations on Microprocessor Frequency and Leakage," *IEEE J. Solid-State Circuits*, vol. 37, no. 11, pp. 1396-1402, Nov. 2002.
- [4] T. D. Burd, and R.W. Brodersen, "Design Issues for Dynamic Voltage Scaling," in *2000 Proc. ISLPED Conf.*, pp. 9-14.
- [5] G.-Y. Wei, and Mark Horowitz, "A Fully Digital, Energy-Efficient, Adaptive Power-Supply Regulator," *IEEE J. Solid-State Circuits*, vol. 34, no. 4, pp. 520-528, Apr. 1999.
- [6] D.W. Kang, "Low Power Digital Adaptive Voltage Controller Design Based on Hybrid Control and Reverse Phase Mode," Ph.D. dissertation, Dept. Elect. and Comp. Eng., Northeastern Univ., Boston, MA, 2003.
- [7] J. Kim, and Mark Horowitz, "An Efficient Digital Sliding Controller for Adaptive Power Supply Regulation," in

- 2001 Proc. VLSI Circuits Dig. Tech. Papers Conf., pp. 133-136.
- [8] PowerWise™ White Paper, <http://powerwise.national.com>
- [9] H. Mizuno et. al., *IEEE Journal of Solid-State Circuits*, vol. 24, 1999, pp. 1492
- [10] K.Roy et. al., *IEEE Journal of Solid-State Circuits*, vol. 91, 2003, pp. 306
- [11] S. Narendra et. al., *IEEE Journal of Solid-State Circuits*, vol. 38, 2003, pp. 696
- [12] Sander, W.B. Schell, S.V. Sander, B.L., "Polar modulator for multi-mode cell phones", *Proc. of the IEEE 2003 CICC*, pp. 439-445, Sept. 2003.
- [13] P. Kenington, *High Linearity RF Amplifier Design*, Artech House, 2000.
- [14] S. Cripps, *RF Power Amplifiers for Wireless Communications*, Artech House, 1999.
- [15] L. Larson, P. Asbeck et al. "Device and circuit approaches for improved wireless communications transmitters", *IEEE Personal Communications*, vol. 6, No. 5, pp. 18-23, Oct. 1999.
- [16] P. Kenington, "Mobile transmitter linearisation for spectrum-efficient modulation formats," *Radio Science Bulletin*, No.293, pp. 16-22, June 2000.
- [17] -, "Triple band dual mode power amplifier application board for GSM/DCS1800/DECT", Motorola, *Microwave Engineering Europe* February/March 1997, pp. 27-28.

Sensor Fusion for UWB and Wifi Indoor Positioning Systems

Frédéric EVENNOU, François MARX, Simon NACIVET

France Telecom Division R&D - Grenoble - France

E-Mail: {frederic.evennou, francois.marx, simon.nacivet}@francetelecom.com

Phone: +33 (0)4 76 76 40 90 - Fax: +33 (0)4 76 76 44 50

Abstract

This paper advocates the application of sensor fusion for location. More and more sensors, like video, RFID, Wifi, are available in those environments. Fusing all those information is becoming a major task in indoor positioning as all the measurements coming from the sensors are noisy. This noise introduces positioning errors that may vary from one technological system to another. Besides, the coverage area of each single system may not be well adapted for all the application so a multi-scale coverage area system may be defined. This paper presents a reliable mobile positioning system taking advantage of the Wifi and the Ultra Wide Band positioning systems. The first may provide a rough position whereas the second is expected to achieve sub-centimeter position in restrained area. Fusing those two systems should lead to a more accurate system enabling to track a device in a building with different scales of accuracy along the path.

I. Introduction

Mobile positioning becomes increasingly an interest for many applications. Many networks are deployed in public and private area. They become some very interesting sources of information for mobile positioning. Each one can provide the position of an equipment but with a certain accuracy. As no location sensor takes perfect measurements or work well in all situations, it becomes interesting to fuse live measurements from multiple location technologies. To achieve optimal performance, a tracking system must exploit all the information in order to compensate the weaknesses of the other sensors. Wifi positioning has recently been a point of interest. Many buildings are equipped with WLAN Access Points (shopping malls, museums, hospitals, airports, ...). The positioning method is based on the fingerprinting [1], [2] to localize the equipment. As the Access Points have a wide coverage area, it is possible to localize a mobile with little equipment. But the received signals fluctuate over time what introduces errors in the positioning.

Some other sensors, like ultra sound or infra red sensors have already been used for short positioning. UWB technology is now widely investigated in order to estimate the accuracy that can be awaited from this technology. Many experiments have been carried out in ranging in dense multipath environment. They have shown the importance of direct-path finding algorithm [9], [10]. But the accuracy of a positioning system based on this technology have not been investigated yet. Fusing those two-scale positioning systems becomes

interesting. When the UWB tracking system may lose the track of the object due to off-range position, the Wifi tracker could continue tracking the object by using its fingerprinting database. Conversely, UWB could help improving the Wifi positioning accuracy where this technology is available.

The main contribution of our paper is to investigate the performances that can be achieved in term of accuracy of the position estimation and coverage area. Then a multi-scale positioning infrastructure based on particle filters will be studied to fuse data coming from Wifi sensors for a wide area coverage technology, with the short range positions provided by a TDOA based UWB system.

This paper presents in a first section the two positioning systems, on one hand based on a Wifi sensor using fingerprinting, on the other hand on a TDOA based UWB system. Then a tracking particle filter will be discussed and modified to lead to a sensor fusion system fed by the data coming from the two previous systems. Finally, some results using physical measurements will illustrate an unprecedented scaling capability to indoor positioning.

II. Indoor mobile location

A. A fingerprinting Wifi based system

Many outdoor systems are based on time measurements, i.e. the mobile equipment and the network are synchronized, thus the mobile can calculate the distance that it is separated from the base stations (GSM) or the satellites (GPS).

However getting this kind of information with commercialized WLAN products is almost impossible. The only available information is the signal strength received from each access point (AP). With such information, it is necessary to find a way to estimate the distance. Using a propagation model ($P_s = f(d)$) might be practical. However, it is really difficult to find an accurate indoor propagation model due to complex RF waves propagation. Simple model (Motley Keenan) [8] has been tried out but lead to bad accuracy. The main source of error is the fluctuation of the RSS over the time.

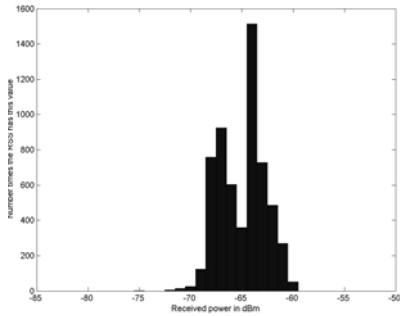


Fig. 1. RSS variations over the time

Finally, we opted for an on-site training to have a mapping between the position and the received signal strength (RSS) database. This method was introduced in [1] and consists in two steps. The first step creates a database of the RSS over the building. At some positions in the building is associated an n-uplet of power measurement. The kept value for each AP, is the mean of the RSS over 100 measurements. During the second step, the device samples the signal strength from each access point and finds its position by comparing its RSS to the ones recorded in the database. It looks for the n-uplet of RSS which is the closest to the instantaneous power measurement:

$$X = \underset{x_k, y_k}{\operatorname{argmin}} \left(\sum_{l=1}^N (P_{r_l}(x, y) - P_{r_l}(x_k, y_k))^2 \right)$$

with l the index over the received APs and (x_k, y_k) the positions recorded in the database.

In comparison with the use of the propagation model, constructing a database is a constraint for the system. However, the fluctuations in the measurements often lead to choose the wrong point in the nearest-neighborhood algorithm. For example the user's position can change even if the user stops or the trajectory of the mobile can become discontinuous. This kind of problems may be avoided with the use of estimating filters like the Kalman filter [3] or the particle filter [5]. A location based on a Kalman filter has been tested in

spite of the restrictions on the linear laws that match the prediction and the correction by a measurement. The Kalman filter delivers a continuous trajectory but cannot take into account the other information which are available like the map of the environment. The particle filter is a more generic filter and allows the use of different kinds of information. The price to be paid is a higher complexity of the implementation.

B. An UWB positioning system

1) *Overview of the system:* A 2D location experiment was constructed, consisting in four receivers and one mobile transmitter that should be localized. Figure 2 shows a high level block diagram of our UWB location experiment. The transmitter consists in a high speed pulse generator which generates a 300 ps width UWB pulse triggered by a pseudo random code generator designed on a FPGA. This code is modulated onto the pulse train using an on-off keying modulation. Then the signal is amplified and broadcasted through a transmitting diamond antenna [12]. The same kind of antenna is used at reception. A synchronous acquisition of the four received signals is made by using a digital sampling oscilloscope (4-channel Lecroy Wavemaster 8620A) after signal amplification.

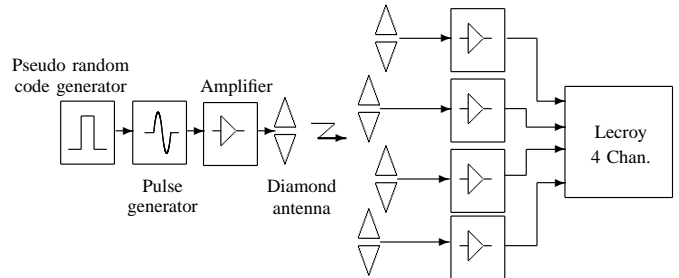


Fig. 2. High level block diagram of Ultra-Wideband location experiment

The sampling rate of the measured signal is 10 GS/s. No sweep averaging is used. The 8-bit pseudo random code is constructed from a 7-bit Barker sequence and has been chosen according to its good autocorrelation properties and shortness. The chip time is set to 200 ns to avoid ISI due to channel delay spread. The length of an acquisition is $5 \mu s$ and enables the capture of about 12 UWB pulses (i.e. 24 chips). The pseudo random code is known at reception but there is no synchronization between the transmitter and the receivers. Using four synchronous receivers completely resolves a 2D location of the transmitter thanks to the TDOA algorithm [13].

2) *TDOA algorithm:* The measurements consist in observing differences in the times of arrival of signal from the transmitter to the four receivers whose

locations are known. Each range (or time) difference determines an hyperbola, and the intersection point of the three hyperbolas is the estimated source location. For each receiver, the relative time of arrival is determined thanks to a correlation between the received signal and the template signal. The receiver that has the best SNR provides the reference time to get the three TDOA. Once the relative time of arrival of the reference antenna is determined, a validity window inferior to the pseudo random code length is defined for the three other receivers. This window avoids ambiguity as several periodic correlation peaks may appear. Three TDOA allow 2D positioning. To avoid error due to the fact that we are working in a 3D environment, we assume to know the height of the transmitter. So the location error only comes from direct-path signal missing or excessive propagation delay through materials. Let $[x_i, y_i, z_i]$ denote the coordinates of the i^{th} receiver, and $[x_M, y_M, z_M]$ the coordinates of the mobile transmitter. The range difference from transmitter to receivers i and j is r_{ij} . Let suppose that the reference receiver is the number 1. The 2D estimate of the transmitter is given by the following equation:

$$[x_M, y_M] = \underset{\hat{x}_M, \hat{y}_M}{\operatorname{argmin}} \left(\sum_{i=2}^4 \left(r_{i1} - \sqrt{(x_i - x_M)^2 + (y_i - y_M)^2 + (z_i - z_M)^2} \right)^2 \right) \quad (1)$$

Note that in equation 1, z_M is assumed to be known.

3) *The UWB Digital signal processing:* Experimental results show that finding the ideal template is difficult. The UWB pulse shape suffers from important distortions through the antennas and the amplifiers, which increase the pulse duration. So it is harder to separate the multipath signals. In our suboptimal but robust signal processing, the exact received pulse shape is assumed to be unknown. As figure 3 shows, the method consists in taking the absolute value of the signal and correlating it by a template whose basic pattern is a square wave. Each 2ns-wide square wave is coded by the value of the corresponding chip in the bipolar pseudo random sequence.

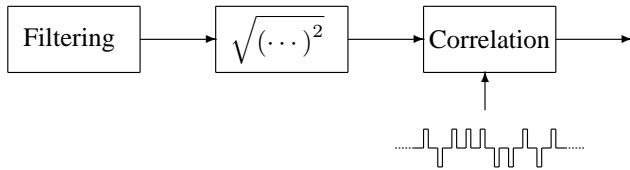


Fig. 3. Block diagram of digital signal processing at reception

The template length is chosen such as all the received energy contributes in the maximum correlation peak. A whitening filter is also implemented as the hypothesis

of an additive white Gaussian noise is needed to use the maximum likelihood criterion in the next section. A challenge of indoor UWB location is the multipath propagation in NLOS situations.

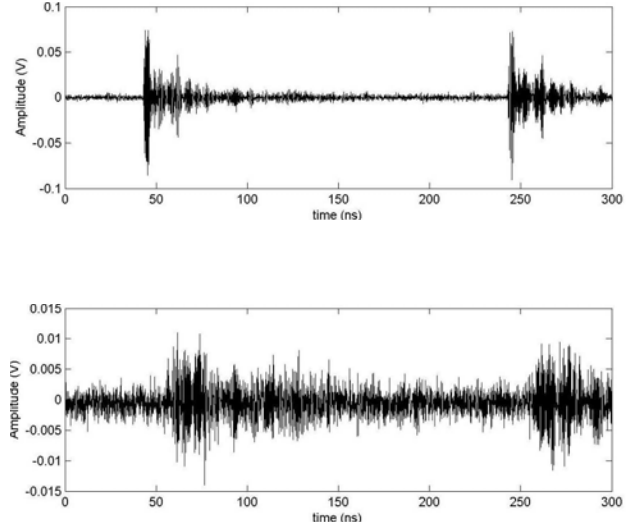


Fig. 4. Typical received signals in LOS and NLOS situations. The signal shown in the first plot was measured with a clear LOS and the other was measured in the presence of LOS blockages.

Figure 4 shows the acquisitions of two typical received signals in case of a LOS and NLOS situations. Two UWB pulses and their multipath replicas are distinguishable. The transmitted pulses are separated by a chip duration of 200 ns. It appears that in the absence of a clear line of sight (NLOS situation) the direct-path signal is not always the strongest one. So the accuracy of the location depends on the direct-path detection error. A GML (Generalized Maximum Likelihood) algorithm was proposed in [11] based on the CLEAN algorithm [14]. The hypothesis is that the received signal is a linear combination of replicas of the single path signal with different delays and amplitudes. The noise is assumed to be an additive white Gaussian noise. The basic steps are:

- 1) Compute the cross-correlation $R_{ST}(t)$ of the absolute value of the received waveform $S(t)$ with the template $T(t)$.
- 2) Find the strongest correlation peak in $R_{ST}(t)$. Keep the amplitude and time delay $\{a_{max}, \tau_{max}\}$.
- 3) Find the first correlation peak satisfying:

$$\frac{a_k}{a_{max}} \geq \theta_\rho \quad \tau_k \in [\tau_{max} - \theta_\delta, \tau_{max}]$$

where :

- θ_ρ is the threshold on the correlation peak amplitude. The range of values of θ_ρ are from 0 to 1 as the correlation peaks are normalized.

- θ_δ is the maximum delay relative to τ_{max} . All τ_k must be searched within $[\tau_{max} - \theta_\delta : \tau_{max}]$.

Those two last parameters need to be dimensioned. They determine two probabilities the False Alarm probability (P_{FA}) which is due to the detection of some noise, and the Missed-Path probability (P_M) which is the detection of a multipath signal for the direct-path.

Those two last probabilities must be minimized to get the optimum parameters. Here the determination of those parameters has been done given the following criteria:

- Find θ_δ such as the probability that the direct-path signal delay is not contained in $[\tau_{max} - \theta_\delta : \tau_{max}]$ is equal to α . Typically $\alpha = 0.001$.
- Choose θ_ρ minimizing the sum $P_{FA} + P_M$.

It is helpful to notice that P_{FA} only depends on the SNR and signal processing at the receiver. On the other hand, P_M is entirely determined by the channel statistics. In our experiment, the channel statistics realized by Cramer, Win and Scholtz [15] were used and the SNR is evaluated for each received signal in order to determine the optimal adaptive threshold θ_ρ .

III. Particle filter and sensor fusion

A. The particle filter

Nowadays, the map of all the public or companies buildings are available in digital format. The key idea is to combine the motion model of a person and the map information in a filter in order to obtain a more realistic trajectory and a smaller error for a trip around the building. In the following, it will be considered that the map which is available is a bitmap. So no information is available except the pixels in black and white that model the structure of the building. The particle filter tries to represent the density function of the mobile-position by a set of random samples with associated weights (i.e. a particle) [7]. Each particle explores the environment according to the motion model and map-information, the weight is updated each time a new measurement is received. It is possible to forbid some moves like crossing the walls by forcing the weight at 0.

The particle filter tries to estimate the probability distribution $Pr[x_k|z_{0:k}]$ where x_k is the state vector of the device at the time step k , and $z_{0:k}$ is the set of collected measurements until the $(k+1)^{th}$ measurement. When the number of particles (position x_k^i , weight w_k^i) is high, the probability density function can be assimilated to:

$$Pr[x_k|z_{0:k}] = \sum_{i=1}^{N_s} w_k^i \delta(x_k - x_k^i)$$

This filter comprises two steps:

- Prediction
- Correction

1) *Prediction*: During this step, the particles propagate across the building given an evolution law that assigns a new position for each particle with an acceleration governed by a random process:

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \\ V_{x_{k+1}} \\ V_{y_{k+1}} \end{bmatrix} = \begin{bmatrix} 1 & 0 & T_s & 0 \\ 0 & 1 & 0 & T_s \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_k \\ y_k \\ V_{x_k} \\ V_{y_k} \end{bmatrix} + \begin{bmatrix} \frac{T_s^2}{2} & 0 & 0 & 0 \\ 0 & \frac{T_s^2}{2} & 0 & 0 \\ 0 & 0 & T_s & 0 \\ 0 & 0 & 0 & T_s \end{bmatrix} \begin{bmatrix} \nu_{x_k} \\ \nu_{y_k} \\ \nu_{x_k} \\ \nu_{y_k} \end{bmatrix}$$

where $[x_k, y_k, V_{x_k}, V_{y_k}]^T$ denotes the state vector associated to each particle (position and speed), T_s the elapsed time between the $(k-1)^{th}$ and the k^{th} measurements. $[\nu_{x_k}, \nu_{y_k}]^T$ is a gaussian random process, which is realistic for a pedestrian move, that simulates the acceleration of the k^{th} particle. This last equation is often called the prior equation. It tries to predict a new position for all the particles.

When the new position of a particle is known, it is possible to include the map information, in order to remove the particles with an impossible move, like crossing a wall. An algorithm, using the previous known position of the particle, its new one, plus the map of the building, checks all the pixels between those positions to see if a wall has been crossed. When this checking is finished, it is possible to assign a weight $Pr[x_k|x_{k-1}]$ as follows:

$$Pr[x_k|x_{k-1}] = \begin{cases} 0 & \text{if a particle crossed a wall} \\ 1 & \text{if a particle did not cross a wall} \end{cases}$$

Then some particles disappear when they cross a wall.

2) *Correction*: When a measurement (n-uplet of RSS or n-uplet of UWB impulse responses) is available, it must be taken into account to correct the weight of the particles in order to approximate $Pr[x_k|z_{0:k}]$. As the measurement is a signal strength or UWB impulse responses n-uplet, and that particles are characterized by their position, the n-uplet must be translated into a position. For a Wifi RSS n-uplet, the mapping between the position and the signal strength is performed thanks to the empirical database. In fact, the algorithm presented in section II-A to find the position of the mobile given the RSS coverage in the building is used. For the UWB system, the channel impulse responses are transformed into a position with the algorithm presented in II-B.2. Then it is possible to estimate $Pr[z_k|x_k]$.

3) *Particles update and resampling*: The weight update equation is given in [4], [5]:

$$w_k^i = w_{k-1}^i \cdot Pr[x_k|x_{k-1}] \cdot Pr[z_k|x_k]$$

To obtain the posterior density function, it is necessary to normalize those weights. After a few iterations, when too many particles crossed a wall, just a few particles will be kept alive (non zero weight). To avoid having just one remaining particle, a re-sampling step is triggered.

The re-sampling is a critical point for the filter. The basic idea behind the re-sampling step is to move the particles that have a too low weight, in the area of the map where the highest weights are. This leads to a loss of diversity because many samples will be repeated. Various re-sampling algorithm were proposed. We did not choose the simple SIS (Sequential Importance Sampling) particle filter [4], but the re-sampling approach presented in [6], Regularized Particle Filter (RPF). The RPF adds a regularisation step. This approach is more convenient because it locally introduces a new diversity after the re-sampling. This may be useful in extreme situations when all the particles are trapped in a room whereas the device is still moving along a corridor. This method of re-sampling adds a small noise to the particle position and avoids this phenomenon.

B. Sensor fusion

The particle filter introduced in section III-A is the tool that enables to merge different information as it relies on the probability densities of the sensors. Combining the information can be done in the expression of the posterior law expressed in III-A.2:

$$Pr [z_k|x_k] = Pr [z_k^{wif}, z_k^{uwb}|x_k]$$

As a simplification, the hypothesis that the Wifi and UWB measurements are uncorrelated has been chosen. This is not true as the the received Wifi or UWB measurements condition one another. With this hypothesis, it becomes possible to write the posterior law as follows:

$$Pr [z_k|x_k] = Pr [z_k^{wif}|x_k] \cdot Pr [z_k^{uwb}|x_k]$$

where z_k^{wif} is the measurement coming from the Wifi sensor (here the position delivered by the database) and z_k^{uwb} the measurement from the UWB sensor. Here it is considered that z_k^{wif} and z_k^{uwb} will be the positions obtained from the Wifi sensor and the UWB TDOA based positioning system respectively.

It has been assumed that both $Pr [z_k^{wif}|x_k]$ and $Pr [z_k^{uwb}|x_k]$ are gaussian probabilities centered on the position delivered by the corresponding sensors. As the availability of the UWB positioning system is limited, it is necessary to select the frames which can deliver a coherent position. The most natural thing to use is an estimate of the SNR of the UWB channels to decide

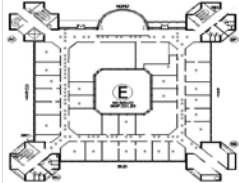
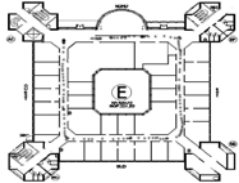
if the received channels must be taken into account to find the position of the mobile. In this system, the SNR that was used is the one estimated when a new frame is received. The higher the SNR on each channel is, the better the estimation of the position must be, and the greater the confidence in the measurement will be. The variance of this UWB gaussian law depends on the estimations of the SNR of each channel.

IV. Experimentations

A. The Wifi based positioning demonstrator

To experiment all those techniques and estimate their capabilities and accuracy to localize a device, a demonstrator has been built. The database is built with one measure in each room, and a measurement every two meters in the corridor. A single floor problem is considered. The criterion to define the error is the mean error over a trip in the building. A walk around the building is taken for the test. Some real measurements are collected along this path and then reused to estimate the performances of each technique (Table I).

TABLE I. Comparison of the different filters

	Database	Particle filter
Trajectory		
Mean error (m)	3.50	1.99

A large improvement may be noticed when a particle filter is applied. When the database is used without any filtering algorithm, it is impossible to determine the trajectory followed by the device. Moreover, many jumps between two measurements are observed. The accuracy with a full database is previously described. A temporal averaging filter (5 samples sliding average) is also used to smooth the variations of the instantaneous RSS. On the contrary, the particle filter succeed in giving a coherent trajectory. It removes most of the wall crossings due to the RSS variations. This can be noticed by observing the trajectory obtained when this kind of filter is used. Some few wall crossings may still be visible because it has been considered that the delivered position of the device would be the barycentre of all the particles. However, over the whole trajectory, the number of wall crossings decreases. Figure 5 gives more information about the performances of this filter.

It provides the cumulative distribution function of the root mean square errors over the trajectory.

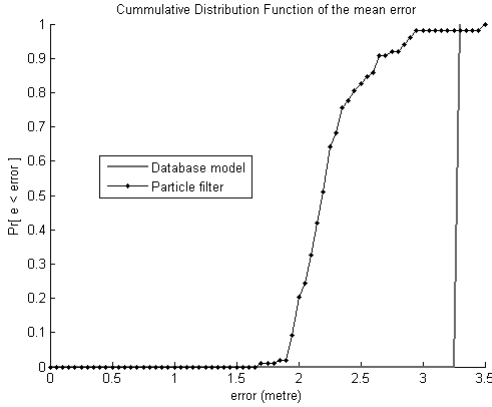


Fig. 5. Cumulative Distribution Function of the different filters

The performances achieved by the Wifi technology to localize a mobile can be sufficient to determine the room where it stands, but not accurately its position in that room. The performance that can reach an UWB system in positioning should help the Wifi tracker to determine the position of the mobile in a room, as long as this service is available. Some few results about this innovative system will be discussed next.

B. The UWB location experiment

Actual data was collected to test our direct-path search method to localize an UWB transmitter. An application was created to collect the data, process it and display the position on a map (real time). The data was recorded in a typical office environment. The four static receiving antennas formed a square of about 6×6 m and were approximately 2.3 m high. As fig. 6 shows, one of the antennas was placed in a room whose dimensions are 7×7 m. The three other antennas were in the corridor surrounding this room. So our UWB location system was conceived typically to test NLOS situations. Wherever the mobile was in the area, at least one of the four receiving antennas was not in line of sight with the transmitter. Note that no trigger signal was needed. We expected better location accuracy for mobile locations in the central room and the corridor because the surrounding zones suffered higher attenuations. Indeed the transmitted signal could have to go through two walls to reach one of the receiving antennas. As fig. 6 shows, 10 transmitter locations were tested in a zone whose dimensions are approximately 20×20 m. TABLE II gives the 75th best location estimation from the 100 sets of acquisitions taken for each location. Two methods of direct-path signal detection were tested. The adaptive threshold described in section

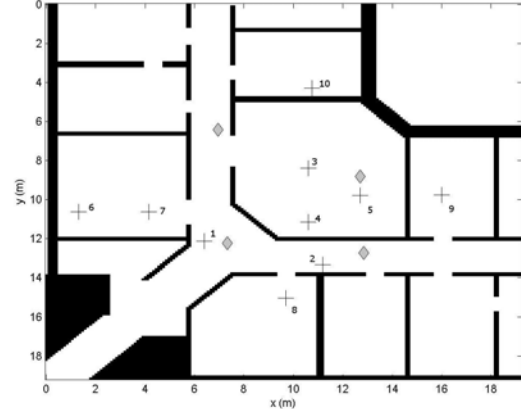


Fig. 6. Basement floor plan of the building where the experiments were conducted. Diamond marks stand for the locations of the receiving antennas and the cross marks indicate every transmitting antennas location.

TABLE II. location error versus location and method (cm).

Location number	1	2	3	4	5	6	7	8	9	10
Error with adaptive threshold	51	32	27	22	13	50	53	52	69	115
Error with invariant threshold	60	57	32	31	39	55	89	41	86	191
Error with maximum peak detection	428	350	214	313	98	681	355	366	341	493

II-B.3 was compared to an invariant threshold. This invariant threshold was intuitively chosen to work as well as possible. Moreover, the results from taking the maximum correlation peak are given. In this case, errors typically larger than one meter occurred. This shows the importance of the direct-path signal investigation. For most transmitter locations the adaptive threshold led to the best results. An analysis shows that it estimates much better the TDOA and prevents most large false alarm errors thanks to the SNR estimation. A tracking experiment was also conducted. The transmitter was carried by a user through the experimentation area. Figure 7 shows the results of the tracking experiment. This section has shown the performances of the UWB technology in positioning systems. A good accuracy can be obtained in the coverage area (about 0.5m). But with the same number of equipment in the network as with the Wifi technology (4 APs in Wifi and 4 receivers in UWB) the coverage area of UWB is far smaller. So the idea is to combine those two technologies, in order to improve the accuracy of the Wifi technology in the area where the UWB positioning is available, and then

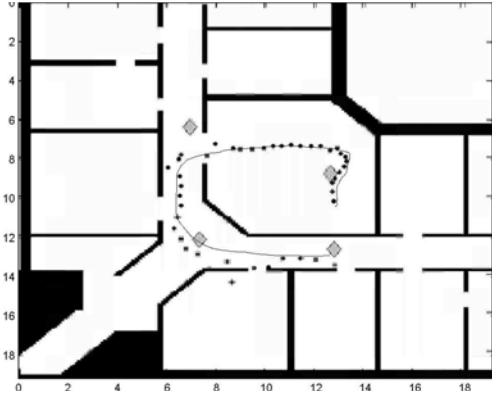


Fig. 7. Estimated user itinerary (Tracking experiment). Circular marks stand for the estimated mobile itinerary. Continue line stands for the real itinerary.

to enable the tracking all over the building even if the UWB positioning is not present. The following section will describe such a system taking into account those two positioning technologies.

C. A multi-scale system

As it was presented earlier, the key idea is to combine the two previous systems that commit some positioning errors due to a measurement noise. Fusing the information should lead to a better accuracy in the area where both technologies are available. On the other hand, as it was previously presented, the coverage area of each technology is not the same. On one hand, there is a wide area coverage ($1600m^2$) enabled by the Wifi technology, on the other hand the UWB covering a small area ($400m^2$). It can be noticed that the accuracy scale is not the same either. With the UWB system, it becomes possible to know the position of the mobile in a room, whereas the Wifi system could only provide the information of the room where the mobile was. To achieve the best performances, it is necessary to design a simple but robust algorithm allowing to take into account those information. As the Wifi positioning is always available, it is natural to use it all the time. But as the UWB system is not always available, it is necessary to define a criteria that will define the frames (and then positions) that must be taken into account to be fused with the Wifi information. Here the most natural and simple way to handle these information is to select the measurements depending on a SNR level. If one of the SNR is too low then it means the confidence in the delivered position must be low. On the contrary, if all the SNR are very high, it means extracting the direct path will be easy and a good positioning will result. In this experiment, it has been considered that the influence would be introduced by the variations of the variance of the gaussian law associated to this process.

This following law is given by:

$$\sigma_{uwb} = \begin{cases} \infty & \text{if } \min [SNR_i] < SNR_{low} \\ \sigma_{wifi} & \text{if } SNR_{low} \leq \min [SNR_i] < SNR_{high} \\ \frac{\sigma_{wifi}}{\alpha} & \text{if } \min [SNR_i] \geq SNR_{high} \text{ and } \alpha > 1 \end{cases}$$

Some measurements have been carried out to estimate the performances of this sensor fusion, compared to the one achieved with each single technology. A reference path has been considered. Figure 8 shows the results of this experimentation.

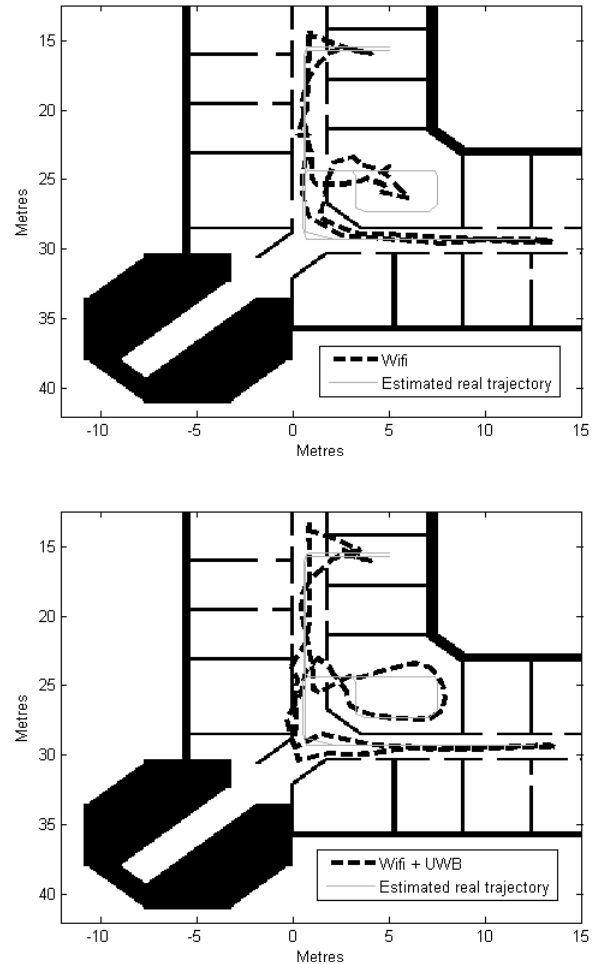


Fig. 8. Comparison of the estimated path when only using the Wifi (upper picture), and then when using the Wifi and the UWB systems (lower picture). The dot line represent the estimated real path whereas the two other path represent the one estimated by the particle filter.

Those last results show that a better estimation of the mobile can be reached when both technologies are available. In fact, with only Wifi, it is possible to detect where the mobile is with an accuracy of about 1.8m (particle filter). But when Wifi and UWB systems are available it is possible to detect the position of the

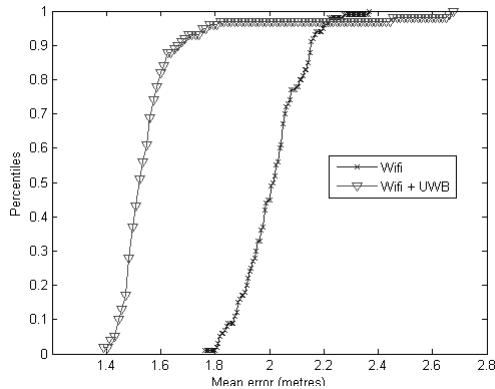


Fig. 9. Cumulative distribution function of the mean error over 100 estimations of the path presented above.

mobile with a 50cm accuracy in the room and fusing those two systems leads to a 1.4m accuracy location. Those performances can be noticed on the cumulative distribution of the mean error over the estimated path (see Fig. 9).

So combining UWB and Wifi positioning may be a good way to localize a mobile thanks to the Wifi network. This first infrastructure would provide a wide coverage area for positioning, but with a 1.80m accuracy, whereas locally, it would be possible to accurately foresee the position of the mobile when an UWB infrastructure would be available and combined with the previous network.

V. Conclusion

In this paper, we have presented a positioning and tracking system for indoor environments. The use of a particle filter which takes into account the human motion, the map information and the signal strength received, leads to a positioning accuracy of 1.8m. When combined with another technology, such as UWB, it is possible to locally and accurately detect the position of the mobile. The experiments carried out show an improvement of 40cm on the global mean error. This study also focused on the performances of UWB systems that can accurately find the position of the mobile

(about 50cm), even in NLOS environments. But those performances can be reached just on little coverage area. Moreover, the particle filter is a useful tool to introduce the sensor fusion as demonstrated earlier.

References

- [1] P. Bahl, V.H. Padmanabhan, "RADAR: an in-building RF-based user location and tracking system", Proceedings IEEE Infocom 2000, Tel Aviv, Israel, vol. 2, pp. 775-784, Mar. 2000
- [2] Y.Chen, H. Kobayashi, "Signal Based Indoor geolocation", Proc. IEEE International Conference on Communications, April-May 2002, New York
- [3] G. Welch, G. Bishop, "An introduction to the Kalman filter", University of North Carolina, Chapel Hill, 2001
- [4] S. Arulampalam, S. Maskell, N. Gordon, T. Clapp, "A tutorial on particle filters for on-line non-linear/nongaussian bayesian tracking", IEEE Transactions on Signal Processing, vol.50, no. 2, Feb. 2002
- [5] F. Gustafsson, F. Gunnarsson, N. Bergman, U. Forsell, J. Jansson, R. Karlsson, P.-J. Nordlund, "Particle filters for positioning, navigation and tracking", IEEE Transactions on Signal Processing, vol.50
- [6] C. Musso, N. Oudjane, F. Le Gland, "Improving regularized particle filters", in : Sequential Monte Carlo Methods in Practice, A. Doucet, N. de Freitas, and N. Gordon, editors, Statistics for Engineering and Information Science, Springer Verlag, New York, 2001, ch. 12, pages 247-271
- [7] A. Doucet, N. de Freitas and N. Gordon, "Sequential Monte-Carlo Methods in Practice", (Statistics for engineering and information Science), Springer-Verlag, 2001
- [8] A.J. Motley, J.M.P. Keenan, "Personal communication radio coverage in buildings at 900 MHz and 1700 MHz", *Electronics Letter*, 9th June 1988, Vol. 24 No.12
- [9] D.P. Young, C.M. Keller, D.W. Bliss, K.W. Forsythe, "Ultra Wideband (UWB) Transmitter Location Using Time Difference of Arrival (TDOA) Techniques". *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers*, Vol. 2, pp. 1225-1229, Nov. 9-12, 2003
- [10] Multispectral Solutions ([http://www.multispectral.com])
- [11] Joon-Yong Lee, Robert A Scholtz, "Ranging in a dense multipath environment using an UWB radio link", *IEEE JSAC*, Vol. 20, No. 9, pp 1677-1683, Dec. 2002
- [12] H.G. Schantz, L. Fullerton, "The diamond dipole : a Gaussian impulse antenna", *2001 IEEE Antenna & Propagation Society International Symposium*, Vol. 4, pp. 100-103, Jul. 8-13, 2001
- [13] B. Friedlander, "A passive localization algorithm and its accuracy analysis", *IEEE Journal of oceanic engineering*, Vol. OE-12, No. 1, Jan. 1987
- [14] J.A. Högborn. "Aperture Synthesis with a Non-Regular Distribution of Interferometer Baselines", *Astron. And Astrophys. Suppl. Ser.*, Vol. 15, 1974.
- [15] J-M Cramer, R. Scholtz, M. Win, "Evaluation of an indoor Ultra-Wideband Propagation Channel", IEEE P802.15-02/286-SG3a and IEEE P802.15-02/325-SG3a.

Using Standard 802.11 Networks for Location Tracking

Arttu Huhtiniemi

Ekahau Oy

Phone: +358 50 598 9153

Fax: +358 20 743 5919

Email: arttu.huhtiniemi@ekahau.com

Abstract

Finding the location of people or assets is needed in many logistical and industrial applications. Real time location systems provide a way to receive a location feed of tracked objects. The systems often require expensive infrastructure to measure the locations. Ekahau has developed a technology that makes use of standard 802.11 networks for accurate location tracking

1. Overview

Ubiquitous systems often need to be aware of the position of other people and assets to be able to make intelligent decisions. Real time location systems provide a way to receive a location feed of tracked objects. Traditional systems often require expensive infrastructure to measure the locations. Ekahau has developed a technology that makes use of standard 802.11 networks for accurate location tracking.

Real time location system is a fully automated system that continually monitors the location of assets and personnel. Network based tracking systems use a combination of tags, base stations and application software to locate, track and monitor assets and personnel in real time.

Some important features of positioning and tracking systems are:

- Track assets without being in 'line-of-sight' inside a building and outside on a campus area network
- Total cost of the system including the infrastructure, software and tags
- Real time information is available to authorized users via the corporate Intranet or through network software
- Solves expensive logistical problems by instantly locating assets or persons

- Helps in maintaining a complete log of movements for auditing, security, and usage analysis
- Generates real time inventory of all tagged assets, thus saving time and money used in locating an asset

2. Positioning and Tracking Technologies

All positioning technologies make use of the fact that certain quantities measurable in wireless networks vary with respect to the physical location where the measurement is done. This variability is caused by factors such as the distance and the angle between the location of the transmitter and the location of the receiver, and the properties of the surrounding physical environment, such as the location, shape and material of reflecting/absorbing surfaces like walls, furniture etc.

Unfortunately, measuring these location-dependent signals accurately is difficult mainly because of the so called **multipath problem** (*Rayleigh fading*)[1]: the signal measurements are inherently noisy as the radio signals travel between the transmitter and receiver along several alternative paths, and each path is affected by different environmental factors. What is worse, some of the environmental factors are dynamically changing due to presence of people, variation in air humidity and so on. The situation is particularly problematic when there is no line-of-sight between the transmitter and the receiver. In this case the signal travelling distance is not necessarily the same as the direct physical distance.

When building a system for positioning, the two most important technological issues are the following:

1. What location-dependent variables are to be used for positioning, and how to obtain the required measurements? The most commonly used location-sensitive variables in positioning are timing-based variables like time or time difference, angle of arrival, and signal strength related variables like the received signal strength indicator (RSSI).
2. How to use the received measurements for location estimation and to improve the received data in order to improve accuracy.

For solving these two sub-problems, Ekahau has developed and patented methods that are both theoretically and empirically validated.

3. Ekahau Positioning Technologies

Ekahau has developed the following technologies that are used in its positioning system.

3.1. Location-sensitive variables

Ekahau uses the RSSI as the basis for positioning. As the RSSI is provided standard in all 802.11 networks, the Ekahau positioning and tracking system can be used without any changes to an existing 802.11 network infrastructure. Another benefit is that the signal strength values change relatively smoothly with respect to changes in location, which means that the RSSI approach is not as sensitive to measuring errors as timing-based or angle-of-arrival approaches.

When using RSSI signals, two alternatives are possible: one can either measure at the mobile handset the signals transmitted by the access point, or one can measure at the access point the signals sent by the tracked devices. Ekahau uses the first approach since the signals reported by the tracked devices are stronger, more consistent and require no changes to access points used. The technology however is usable also when measuring signals from the tracked devices where special built network hardware is needed for measuring.

3.2. Probabilistic framework

For estimating the location based on the RSSI values Ekahau has developed a probabilistic positioning framework, where the world is taken to be stochastic, not deterministic, and one accepts the fact that the measured signals are inherently noisy.

A *probabilistic model* assigns a probability for each possible location (L) given observations (O) consisting of the RSSI of each channel:

$$P(L|O) = P(O|L) P(L) / P(O)$$

$P(O | L)$ is the conditional probability of obtaining observations O at location L.

$P(L)$ is the prior probability of location L.

$P(O)$ is a normalizing constant.

This formula is an example of an application of a mathematical theorem known as the Bayes rule[2]. Based on probability theory, this theorem gives a formal way to quantify uncertainty, and it defines a rule for refining a hypothesis by factoring in additional evidence and background information, and leads to a number representing the degree of probability that the hypothesis, which in Ekahau's case is the location estimate, is true.

The key issue in the probabilistic approach to positioning is to estimate the probability distributions of the measured signals O in different locations L; in other words, we need to determine the conditional probabilities $P(O|L)$. For building the distributions Ekahau uses probabilistic site calibration method, which is described below as one of the practical enhancements to the probabilistic framework.

4. Enhancements to the probabilistic framework

In order to make the theoretically elegant probabilistic framework to work in practice, a number of important problems needed to be solved. Ekahau has developed several innovative approaches that solve many of the practical problems and improve the overall accuracy of the system. The enhancements below illustrate three patented enhancements of the framework.

4.1. Ekahau Site Calibration

For building the probability distributions, Ekahau technology uses a representative sample of site-specific calibration measurements as input. By using this type of calibration data and elaborate state-of-the art machine learning algorithms, it is possible to construct a site-specific model of the environment very quickly and easily. Note that with this type of an approach, the system does not need to be told anything about the wireless environment explicitly, not even the locations of the base stations.

In Ekahau Site Calibration collecting sample points from different site locations creates a site-specific model of the radio network. In practice the calibration is made by measuring the signals strengths using a laptop or tablet PC and clicking the measurement location on a map. Each sample point contains RSSI and the related map coordinates, stored in an area-specific positioning model for accurate tracking.

Ekahau's probabilistic site calibration provides superior accuracy and reliability over any competing direct RSSI methods such as look-up-tables or any other matching algorithm.

4.2. Ekahau Rail Tracking

Another way of improving the accuracy is by making an assumption that the current location of the tracked person or asset is almost certainly near the place where the tracked device was one or two seconds ago. This allows the system to use the full history of observations is taken into account when locating the user which further enhances the positioning accuracy.

Additionally, if one wishes to record the position history of a tracked device, it would make sense to distinguish legal paths from illegal paths. With illegal paths we mean paths that the tracked asset or person cannot go through for instance walls. Similarly paths could be used for indicating the most common path a user would take when walking through an open area from A to B, or calculating the shortest routing between two locations.

Ekahau Positioning Engine supports adding legal paths to the positioning model by giving a set of drawing tools to the system manager. The method of defining legal paths is Called Ekahau Rail Tracking. Ekahau Rail Tracking improves the overall accuracy by directing the estimation to the legal coordinates. This method is optional and the choice of using it depends on application requirements.

4.3. Normalizing RSSI scales

The parameters used for estimating the location of tracked devices are the RSSI values read from the network interface cards (NIC). This means that also the accuracy of positioning depends on the accuracy of the RSSI values. As the quality and accuracy of the values depend on the WiFi chip set used, the choice of NIC card generates error that directly affects to the accuracy of the position estimates.

To overcome the differences between NIC cards Ekahau has invented and patented a technology for normalizing the RSSI values received. The normalization is done in the server and minimizes the error caused by the differences between the accuracy of the NIC cards. Additionally the same technology can be used for normalizing values received when the tracked devices is enclosed or attached to an object that may affect the signal such as hospital bed, forklift or any other large device.

5. Summary

Using the described technological approaches Ekahau has developed a positioning system that accurately tracks objects in real time. As the system makes use of standard wireless network infrastructure, it can be deployed over existing networks without affecting the use of the network for data transfers and with no need to deploy a secondary network for location tracking.

6. References

- [1] <http://encyclopedia.thefreedictionary.com/rayleigh%20fading>
- [2] <http://encyclopedia.thefreedictionary.com/Bayes'%20rule>



DSRC Industry Consortium (DIC)

DSRC Technology and the DSRC Industry Consortium (DIC) Prototype Team

White Paper

28 January 2005

Prepared for:

ARINC (Broady Cash) / U.S. DOT (Bill Jones)

Prepared by:

SIRIT Technologies

1321 Valwood Parkway, Suite 620

Carrollton, Texas 75006

(Randy Roebuck, rdroebuck@sirit.com)

Revisions

Revision Number.	Date	Description
0.1	12 Nov 2004	Draft
1.0	28 Jan 2005	Collection of DIC team comments

Table of Contents

Revisions.....	ii
Table of Contents.....	iii
DSRC Technology and the DSRC Industry Consortium Prototype Team	4
1 Introduction	4
2 Operation & Parameters	4
3 Term/Meaning	5
4 Industry Activities - Participants & Issues / Deployment.....	5
5 Service Categories	6
6 Regulations	6
7 Standardization	7
8 DIC Prototype Team.....	7
9 Summary.....	8

List of Figures

Figure 1 - Industry Activities Participants	5
Figure 2 - Frequency Channel / Antenna Plan.....	7

DSRC Technology and the DSRC Industry Consortium Prototype Team

1 Introduction

Dedicated Short-Range Communications (DSRC) is an emerging technology with intriguing performance and benefits that provides a critical communication link for future Intelligent Transportation Systems. DSRC technology will provide secure, reliable communication links between vehicles and infrastructure safety subsystems that can increase highway safety. Improved highway safety is the number one priority of the United States Department of Transportation (DOT). These DSRC-based systems may save lives by providing warnings to drivers of impending dangerous conditions or events, thereby providing drivers more time to take corrective or evasive actions. The 5.9 GHz DSRC link uses *digital radio* techniques to transfer data over short distances between roadside and mobile units, between mobile units themselves and between portable and mobile units. This link enables operations related to the improvement of traffic flow, highway safety, and other ITS applications in a variety of application environments called DSRC/WAVE (Wireless Access in a Vehicular Environment).

5.9 GHz DSRC system requires robust, fast, localized transmissions from vehicle-to-vehicle (V-V) and roadside-to-vehicle (R-V) to serve many public safety and private commercial applications (in-vehicle signage, collision avoidance, fee collection, internet access, etc). The technology draws upon the increasingly popular IEEE 802.11 “Wi-Fi” standard already widely deployed in businesses and homes. However, for high-speed vehicular applications, significant changes were required to provide latency minimization, authorization, prioritization and anonymity without compromising messaging integrity, correctness, privacy, & robustness attributes. This highly efficient system is complementary to existing cellular and satellite communications but does not give “2 Way Voice / Broadcast” or “Tracking” device capabilities.

2 Operation & Parameters

The 5.9 GHz DSRC system contains Roadside Units (RSUs) connected to a land-based infrastructure with ITS application interface and On-board Units (OBUs) integrated into the vehicle’s internal network (IVN) and supporting embedded vehicular applications. The DSRC/WAVE system supports communication links in the following parameters:

- 1) Vehicle speed (up to 120 mph)
- 2) Communication range (up to 1000 meters for special vehicles; nominal is 300 meters)
- 3) System Latency (< 50 ms)
- 4) Data rate (default is 6 Mbps; up to 27 Mbps)
- 5) Single transaction size (up to 20K bytes)

The system is based on “events and snapshots” in a read zone when an OBU enters the communication zone of an RSU. In this case, the RSU sends messages on the control channel and the OBU listens and then responds with public / private data. In the case of V-V communication, one of the OBUs will start the transaction by taking on many characteristics of an RSU (i.e., sending the initial interrogation). An OBU is a transceiver that is normally mounted in or on a vehicle, but may, in some instances, be a portable unit. An OBU can be operated while a vehicle or person is either in motion or stationary. OBUs receive and contend for time to transmit on one or more radio frequency (RF) channels. An RSU is a transceiver that is mounted

along a road or pedestrian passageway. In some cases, an RSU may also be mounted on a vehicle or can be hand-carried; however, such operations are only permitted when the vehicle or hand-carried unit is stationary. Furthermore, RSU operations are restricted to those locations where it is licensed to operate. An RSU also provides channel change control and operating instructions to OBUs within its communications zone along with exchanging data.

3 Term/Meaning

“DSRC” has different meanings, different technical characteristics and different operating frequencies around the world in the transportation sector. In most parts of the world, DSRC is generally used only for electronic tolling collection (ETC) & access control applications. In the United States, DSRC/WAVE operates at 5.9 GHz and supports a whole new range of vehicle communication uses with two-way high data rate capabilities and larger communication ranges. It offers superior vehicle-to-vehicle communications and will probably be the “Tag of the 2010 Decade” from the tolling perspective.

4 Industry Activities - Participants & Issues / Deployment

In order to deploy a nationwide technology, the U.S. Department of Transportation (DOT), including both NHTSA and FHWA, has been working with many groups to address public safety, regulations/licensing, standards, testing/compliance, certification, interoperability, networking, data security, electronic technology and system performance issues. Figure 1 below shows many of the companies and organizations involved:

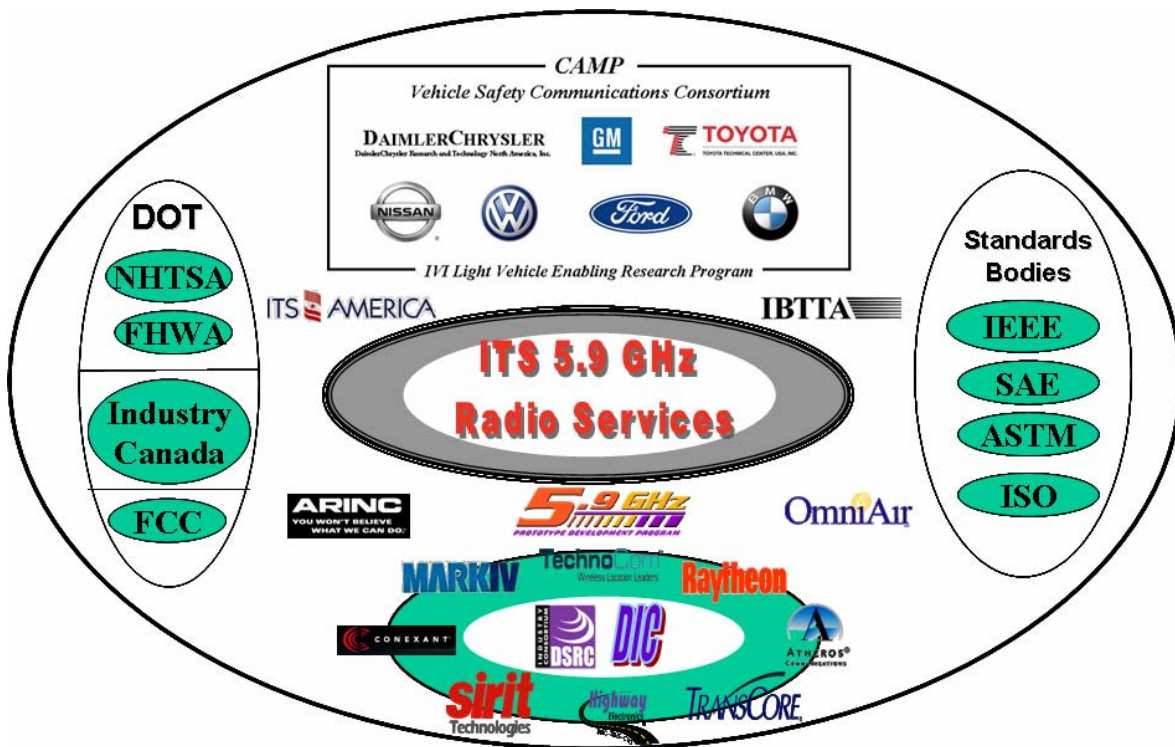


Figure 1 - Industry Activities Participants

The U.S. DOT is planning to make a strategic safety-system deployment decision in the 2008 timeframe. If positive, this decision would result in RSUs (roadside units) being deployed primarily at intersections and OBUs (on-board units) incorporated into new automobiles in the 2010 timeframe. Whether this deployment decision takes the form of a mandate is yet to be determined.

5 Service Categories

DSRC will incorporate a rigid message prioritization scheme to assure that the most important messages are always the first to be transmitted. DOT has dictated “Safety of Life” messages (e.g., messages related to an imminent collision) as top access priority followed by public safety messages (police, fire, ambulance) and then private commercial applications. The following eight service categories have been identified to increase safety and provide more efficiency in the ITS transportation system:

- | | |
|------------------------------------|--|
| 1) Travel and traffic management | 5) Maintenance construction operations |
| 2) Public transit management | 6) Electronic payment |
| 3) Commercial vehicle operations | 7) Emergency management |
| 4) Advanced vehicle safety systems | 8) Information management |

6 Regulations

The Federal Communication Commission (FCC) has completed the rule making and licensing policies for DSRC. The FCC issued Report & Order “03-324A1” in February 2004 and posted in the Federal Register August 2004 as a final rule. Licensing registration started October 2004 per FCC’s public notice DA-04-3165A1. The permanently fixed RSUs are registered and licensed per Part 90 and OBUs are licensed by rule per Part 95. FCC ruling was based on the ASTM standard E2213-03 where the 5.9 GHz (5.850-5.925) band is divided into seven 10 MHz channels (one control & six service) at power levels up to 44.8 dBm (30 Watts) EIRP for RSUs and 33 dBm (2 Watts) EIRP for OBUs. The antennas (omni-directional or directional) connected to RSUs must be mounted between 6 and 15 meters high. Each channel is allocated for specific application types and performance characteristics as depicted in Figure 2.

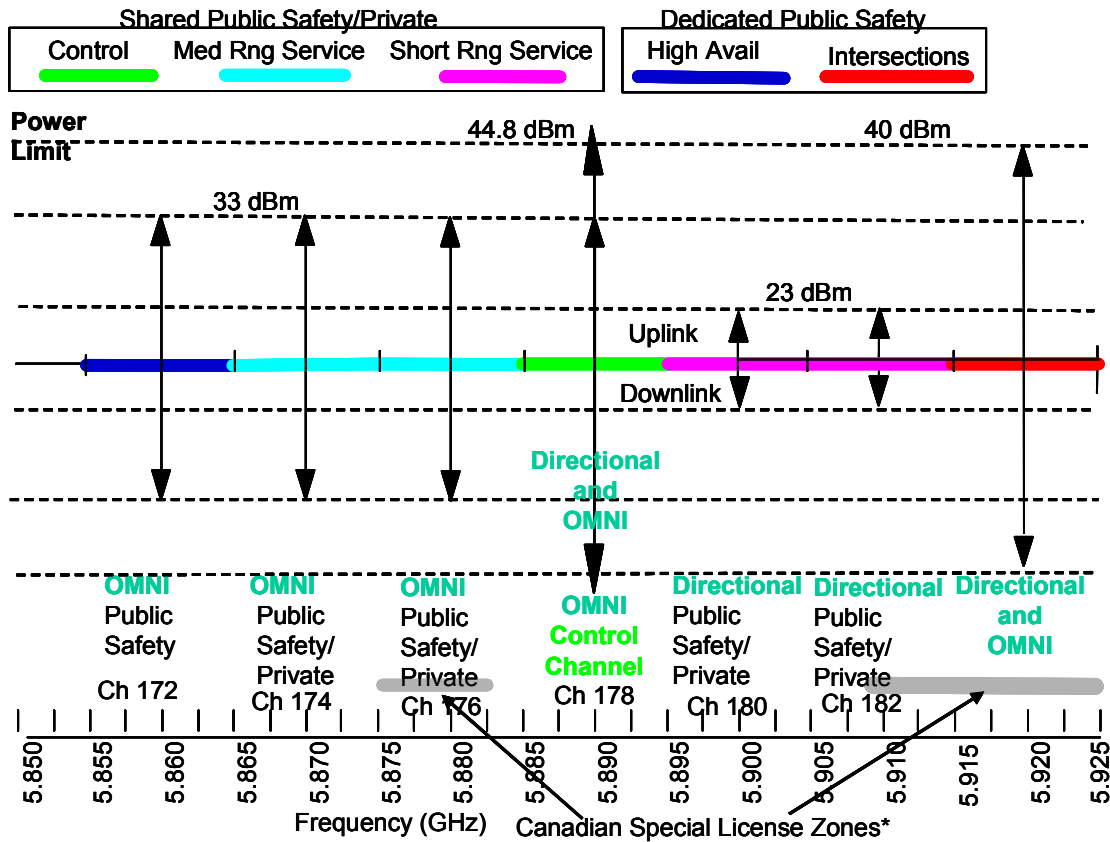


Figure 2 - Frequency Channel / Antenna Plan

7 Standardization

Standardization plays a very important role in the success of any potential large-scale deployment of DSRC technology. A national deployment requires interoperability of equipment and systems coming from many different manufacturers, hardware / software certifications, compliance testing and security. A complete suite of standards is currently under development within IEEE and these are expected to eventually migrate into ISO. IEEE 802.11p addresses the physical layer and medium access control layer (MAC) called 802.11p module. The upper layers (network & data) of the communication stack are being developed within IEEE 1609 (Wave Management, Channel Management, & Resource Manager) and IEEE 1556 (DSRC Security) through the normal IEEE committee process. The vehicle aspects are being developed and evaluated through VSCC / CAMP (represents seven major automotive manufacturers) and SAE is developing the message set, data dictionary and application framework standards. Certification / compliance processes are being worked through the OmniAir consortium.

8 DIC Prototype Team

The *DSRC Industry Consortium (DIC)* prototype team has been created by industry, funded by U.S. DOT and administrated through ARINC to develop / verify the standards and hardware in order to validate the technology & application parameters in its real-world environment. The prototype team includes Mark IV, Raytheon, Sirit Technologies and TransCore. These companies are the technology leaders in ITS systems in the United States and Canada. Prototype program tasks include the development of system architecture, standards,

hardware, software and testing. The first generation of hardware will validate the initial standards and test the prototype units for technical and operational parameters. Once the basic functions are verified, the prototype units may be used in larger pilot trials such as intersection testing model deployments and other robustness / system capacity tests and therefore provide a foundation for larger North American DSRC deployments.

9 Summary

5.9 GHz DSRC is the emerging communication technology that offers standardized ITS products and benefits in national large-scale deployments. U.S. DOT and the automotive OEMs will be the strategic players making deployment decisions in the year 2008 timeframe. 5.9 GHz DSRC systems provide a significant enhancement in communication capabilities over all previous ITS systems. DSRC will support multiple uses in vehicle / public safety and commercial applications that cannot be achieved today. DSRC is a cost-effective communications service, especially when compared with current cellular and satellite systems. The technology can be leveraged for Open Road ETC and mobile 802.11 Wi-Fi deployments, creating nationally interoperable systems and networks. DSRC is the technology for the 2010 decade and beyond.

On the Performance of Ultra Wideband Radio in Stochastic Tapped Delay Line Model of the Ultra Wideband Channel

Kazi M. Ahmed, Mohammad Upal Mahfuz*, Rabindra Ghimire, and Mohammad E. R. Khan
Telecommunications Program, Asian Institute of Technology, PO Box: 4,
Klong Luang, Pathumthani 12120, Thailand.
Email*: st100136@ait.ac.th

Abstract—This paper demonstrates the performance of Ultra Wideband (UWB) systems in the Stochastic Tapped Delay Line (STDL) model of the UWB channel with RAKE receiver. System models are based on Direct Sequence Code Division Multiple Access (DS-CDMA) principle. The system has been studied in presence of multiple user interference and Gaussian noise. During the analysis, various types of narrow, sub-nanosecond pulses with same power were used and extensive simulations were run in 3.1 to 10.6 GHz frequency band. The simulation results show that pulse shape has noticeable impact on the performance of the system. The performance of the pulse waveform based on third derivative of Gaussian pulse has proved to be better than other pulses that were used. With a thorough analytical approach, spectral behavior of higher order Gaussian pulses has been explained. Finally, it is concluded that the third derivative of Gaussian pulse is the most suitable pulse shape for fulfilling the FCC regulated power spectral density (PSD) in multi-user interference environment.

Index Terms—Ultra wideband, channel model, pulse shape.

I. INTRODUCTION

Ultra Wideband Impulse Radio (UWB-IR) is currently receiving a great deal of global attention because of its ability to provide higher data rate with low cost and relatively low power consumption. UWB radio communicates with baseband pulses of very short duration on the order of tenth of a nanosecond. According to the regulation of Federal Commission of Communications (FCC), a signal is defined as UWB signal if it has a -10 dB fractional bandwidth, F_{BW} greater than (or equal to) 0.20 or it occupies at least 500MHz of the spectrum [1] expressed as

$$F_{BW} = 2 \frac{f_H - f_L}{f_H + f_L} \geq 0.20$$

where f_H and f_L correspond to the -10 dB bandwidths. FCC has also regulated the spectral shape and maximum power spectral density (≈ -41.3 dBm/MHz) of the UWB radiation in order to limit the interference with other communication systems like UMTS or WLAN [2]. Although UWB impulse radio can be termed as ‘Spread Spectrum’ technique because of its extremely large operational bandwidth (7.5 GHz), the main difference is that conventional spread spectrum techniques utilize carriers for transmitting information whereas UWB radio transmits information without carrier. The waveforms

that are used for UWB radio are very short in duration, causing their energy to be spread across the frequency spectrum. With UWB signals the dense multipath can be resolved, allowing a rake receiver for signal demodulation [3].

In order to analyze the potential of UWB systems, models describing the UWB propagation channel must be properly known. The channel models help analyze system performance and make design trades. In [4] the Statistical analysis of indoor radio propagation and the modified Poisson process or (Δ - K) model for path arrival of UWB system was presented. Examination of the Spread Spectrum (SS) principle with time hopping technique in the presence of wideband interference with AWGN channel has been considered [5]. The modified Poisson process or Δ -K model of the channel was used to study the UWB system for various M-PAM signals [6], but the channel model was not suitable for the UWB system because the frequency band used for measurement was in the range from 900 MHz to 1300 MHz. The effectiveness of multiuser detection for UWB using DS-CDMA was shown in [7]. A channel model based on measurements made in the indoor environment was presented in [8]. The model presented in [8] was formulated as a stochastic tapped delay line (STDL) model of the UWB indoor channel.

In this paper, to develop the UWB system model, DS-CDMA spreading approach, Stochastic Tapped Delay Line (STDL) model of the channel [8] and RAKE receiver in presence of multiple user interference are considered. The Bit Error Rate (BER) performance of DS-CDMA systems in UWB using STDL model and RAKE receiver with multiple user interference is shown. The effect of pulse shapes on channel characteristics has also been indicated. A thorough analytical explanation supporting the fact that third derivative of Gaussian pulse is the most suitable pulse shape for simultaneously fulfilling the FCC regulated power spectral density (PSD) has also been illustrated. Finally, on the basis of extensive simulations of BER performance and fulfillment of FCC power emission regulation, a conclusion has been drawn to choose the third derivative of Gaussian pulse as the best pulse shape for UWB radio communication.

The paper is organized as follows: Section II provides with a brief description of the concept of DS-CDMA based multi-user interference. The discussion is followed by Section III describing the complete STDL system model under investigation. Simulation environment and corresponding results have been explained in Section IV. Finally, Section V concludes the paper.

II. CONCEPT OF MULTI-USER INTERFERENCE

Multiple User Interference (MUI) is modeled as signals coming from other UWB users having the same basic signal characteristics as those of the user of interest, but using different spreading codes. Assuming a total of Nu users, the received signal $r(t)$ can be written as [9]

$$r(t) = s^{(1)}(t) \otimes h^{(1)}(t) + i(t) + n(t)$$

where $s^{(1)}(t)$ is the signal of interest at the output of the transmitting antenna, $h^{(1)}(t)$ represents the impulse response of the channel corresponding to the user of interest, $n(t)$ is the Gaussian noise and $i(t)$ is the interference coming from the other $(Nu - 1)$ users. Here the interference, $i(t)$ can be defined as

$$i(t) = \sum_{n=2}^{Nu} s^{(n)}(t - \tau_n) \otimes h^{(n)}(t)$$

where $h^{(n)}(t)$ for $n = 2, 3, \dots, Nu$ are the impulse responses of the channels corresponding to the $(Nu - 1)$ interfering users and τ_n is the delay of the arrival time of the n^{th} user from the user of interest i.e. $s^{(1)}(t)$ for which $\tau_1 = 0$. Assuming DS-CDMA as the used multiple access technique, $s^{(n)}(t)$ can be written as [9]

$$s^{(n)}(t) = \sum_{k=-\infty}^{+\infty} \sum_{j=1}^{Np} w_{d_k^{(n)}}(t - kT_d - jT_c)(c_p)_j^{(n)}$$

where $w(t)$ is the used pulse waveform, $(c_p)_j$ is the j -th chip of the Pseudo-random Noise (PN) code, generated by using bipolar PN sequences with values $\{+1, -1\}$, d_k is the k -th data bit chosen which, in Pulse Shape Modulation (PSM) case, also defines the pulse waveform to be transmitted, Np represents the number of pulses per data bit, T_c is the chip length and T_d is the data bit length defined as $T_d = N_p T_c$. Each user experiences the channel with a different channel realization so that the channel impulse response

$$h^{(n)}(t) \neq h^{(m)}(t) \text{ where } n, m = 1, 2, \dots, Nu, n \neq m$$

Each MUI user has the same power as the user of interest. Interfering users are considered either synchronous or asynchronous with the user of interest. The 'synchronous' condition refers to 'frame synchronization' where the receiver is synchronized with all other users and can receive all users' data at the same time instant. On the other hand, in the asynchronous case,

the interfering signals arrive at the receiver with resolution of a time sample. The asynchronism is performed between every interfering user and the user of interest, and also among the interfering users.

III. SYSTEM MODEL

In this research study higher order derivatives of Gaussian and Rayleigh pulses have been used with STDL channel model. STDL channel model is a derivative of S-V model. STDL model has been chosen for channel representation because S-V channel has been very popular for indoor wireless channel modeling and many related studies have been based on this channel [9]. The time domain representations of some pulses are given in [10]. The overall block diagram of the proposed system is then shown in Figure 1.

A. Sub-nanosecond mono-pulses

The basic pulses used in simulation of this study are the n^{th} order derivatives of Gaussian pulse expressed as

$$\Omega_G^n(t) = \frac{d^n}{dt^n} \left(\frac{A}{\sqrt{2\pi}\sigma} e^{-\left(\frac{t^2}{2\sigma^2}\right)} \right)$$

where the basic waveform of Gaussian pulse is given by

$$\Omega_{oG}(t) = \frac{A}{\sqrt{2\pi}\sigma} e^{-\left(\frac{t^2}{2\sigma^2}\right)}$$

In addition to above, the channel effects using Rayleigh pulse as expressed below are also compared.

$$\Omega_{oR}(t, \sigma) = \frac{t}{\sigma^2} \exp(-t^2 / 2\sigma^2)$$

Truncated *Sinc Pulse* and *Dual Gaussian Monopulse* were also used for simulation.

B. Transmitter Section

In the transmitter section Direct Spread CDMA spreading approach has been used. The equiprobable binary bit stream $\{b_j(i)\} \in \{1, -1\}$ is first multiplied by the spreading code, $c_j(k)$, named Walsh Hadamard Code of length 16. This DS-CDMA spreading is then followed by BPSK modulator where the spread output is BPSK modulated by a signature waveform, $\Omega_c(t)$. The modulated signal for the j^{th} user is represented as [11]

$$x_j(t) = \sum_{i=-\infty}^{+\infty} \sum_{k=1}^N \Omega_c(t - iT_s - kT_c) c_j(k) b_j(i)$$

where, $b_j(i)$ and $c_j(k)$ represent the i^{th} information bit and the k^{th} chip of the spreading code with length, N and chip duration, T_c respectively for the j -th user. T_s is the

symbol duration and $\Omega_c(t)$ is the chip pulse waveform. The radiated waveforms for Rayleigh Pulse and Third derivative of Gaussian pulse are shown in Figure 2(a) and 2(b).

C. Channel Model

The transmitted BPSK modulated signals, $x_j(t)$ are then passed through the STDL propagation channel model, which is based on measurements made in a typical office environment for UWB radio channel [8]. The model characterizes the shape of the power-delay profile (PDP) in terms of path gains, G_K and delays, τ_K of multipath components. The delay axis was adjusted in such a way that the delay bin of the first quasi Line of Sight (LOS) path began at time instant $\tau = 0$. The path resolution of the considered system was $\Delta\tau = 2ns$. Before the signal is received by the receiver section, thermal noise, $n(t)$ and interference from narrow-band transmitters, $n_f(t)$ get added to it. The model represented the statistics of the path gains and its dependence on the delays.

The large-scale and the small-scale fading statistics were also distinguished. The Power Delay Profile (PDP) at one location and that averaged over all locations within the measurement area were referred to as ‘Local PDP’ and ‘Small Scale Averaged PDP (SSA-PDP)’ respectively. Large scale fading was investigated by considering the SSA-PDPs of different locations, i.e. after having removed partially the small scale statistics by spatially averaging the PDPs over the locations within one measurement area. The ‘global’ parameters referring to large scale fading effect of the channel have been shown in Table 1.

For large scale fading, the SSA-PDP exhibits an exponentially decay function of excess delay. The power ratio, r and the decay constant, ε , which vary from location to location, were treated as stochastic variables. Since there were limited values of r and ε , they were modeled as Log-Normal variables [8] as shown in Table 1. The path loss (PL) was defined as a function of distance also as shown in Table 1 [8]. The total average energy gain, \bar{G}_{tot} can be calculated by integrating the SSA-PDP of each place over all delay bins. The \bar{G}_{tot} is also lognormally distributed about path loss (PL) with a standard deviation of 4.3 [8]. The average PDP is specified according to [8]

$$\bar{g}(\tau) = \frac{\bar{G}_{tot}}{1 + r \frac{1}{1 - e^{(-\Delta\tau/\varepsilon)}}} \left\{ \delta(\tau - \tau_1) + \sum_{k=2}^{N_{bins}} [r e^{-(\tau_k - \tau_2)/\varepsilon}] \delta(\tau - \tau_k) \right\}$$

On the other hand, small-scale fading characterizes the changes in the received signal when the receiver position changes only by a small fraction of the distance between transmitter and receiver. This was derived by considering

the deviations of all the local PDP from the respective SSA-PDP. The ‘local’ parameters, referring to small scale fading effect, have also been shown in Table 1. The local path gain values, G_k are the superposition of large and small scale statistics and are Gamma distributed random variables with mean, \bar{G}_k and parameters, m_k . The values of m_k are independent truncated Gaussian random variables with parameters dependent on the delay, τ_k and mean, $\bar{G}_k = \bar{g}(\tau_k)$ [8].

The received signals are the sum of replicas of the transmitted signals. The received signal is, therefore, expressed as

$$r(t) = \sum_{i=1}^{N_{path}} (c_i x(t - \tau_i) + n(t) + n_f(t))$$

where, $n(t)$ is zero mean AWGN and $n_f(t)$ is the interference signal.

D. Receiver Section

For the simulation of this study RAKE receiver with a maximum of 12 arms are used. First arm is locked to the first multipath component, m_1 . Multipath component, m_2 arrive τ_1 time units later than m_1 and is captured by second tap and so on. The signals in the correlators are despread by Walsh Hadamard Code of length 16. All decision statistics are weighted by a weighting factor, α to form overall decision statistics. The signals are then integrated over the entire period. The integrated output signal is then compared with the appropriate threshold value to receive the better estimate of the transmitted signal.

Table 1: Expressions of global and local parameters

<u>Global Parameters:</u>	
Path Loss = $\begin{cases} 20.4 \cdot \log_{10}(d/d_0) & d \leq 11m \\ -56 + 74 \cdot \log_{10}(d/d_0) & d > 11m \end{cases}$	
Decay Constant $\varepsilon \sim L_N(16.1; 1.27)$	
Shadowing $\bar{G}_{tot} \sim L_N(-PL; 4.3)$	
Power Ratio $r \sim L_N(-4; 3)$	
<u>Local Parameters:</u>	
Energy Gains $G_k \sim \Gamma(\bar{G}_k; m_k)$	
m Values	$m_k \sim T_N(\mu_m(\tau_k); \sigma_m^2(\tau_k))$
	$\mu_m(\tau_k) = 3.5 - \frac{\tau_k}{73}$
	$\sigma_m^2(\tau_k) = 1.84 - \frac{\tau_k}{160}$

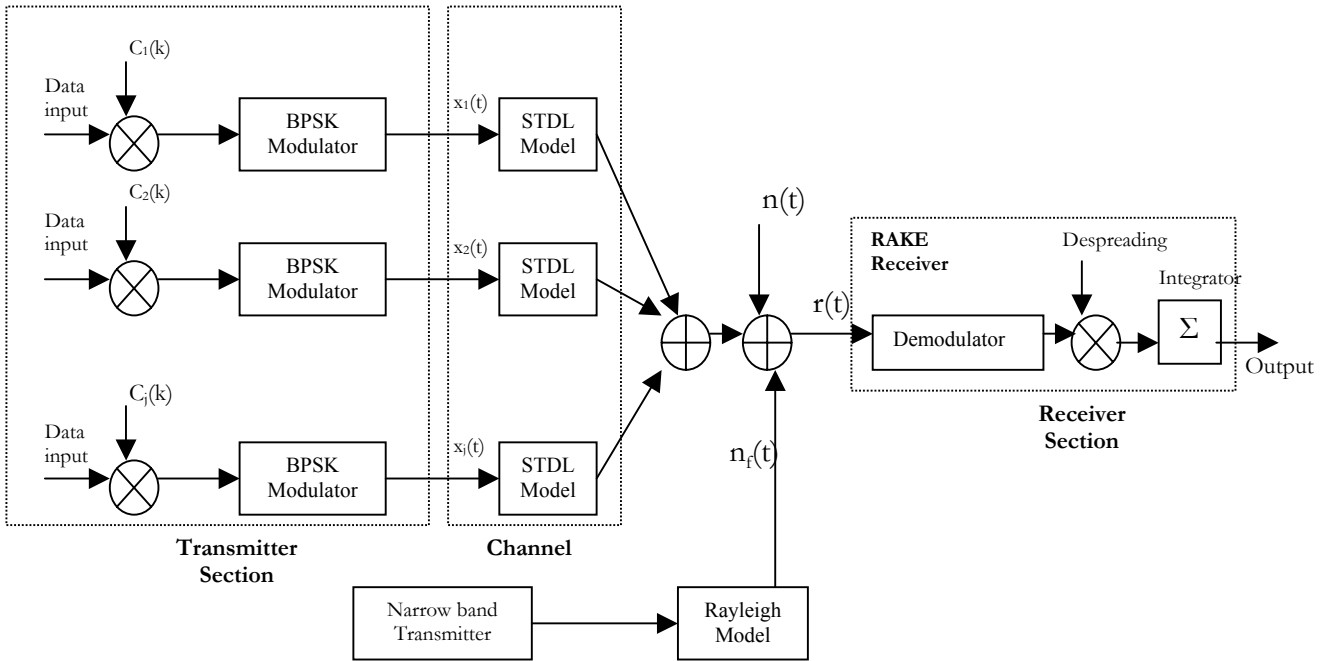


Figure 1: Block diagram of system model

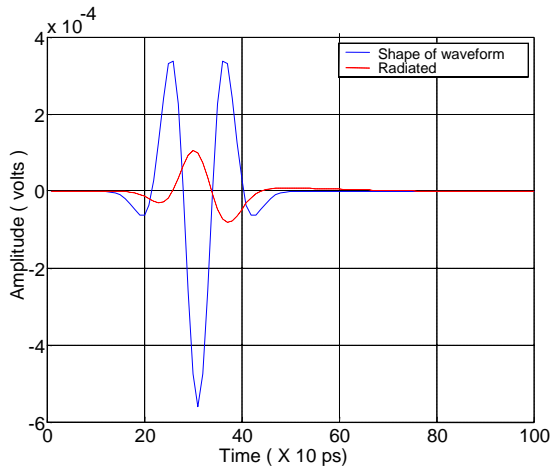


Fig. 2.b: Third Derivative of Gaussian Pulse

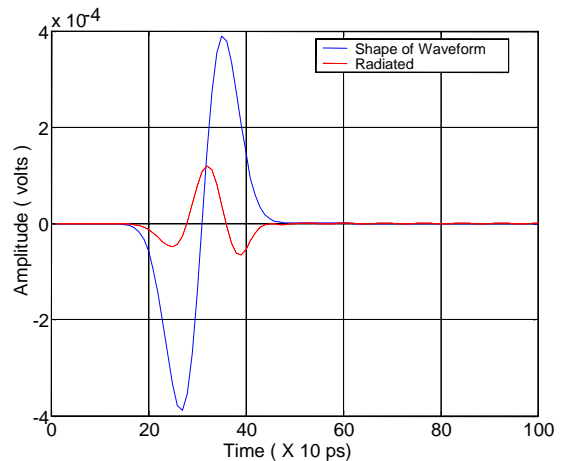


Fig 2.a: Rayleigh pulse

IV. RESULTS

In this section, results based on the simulation of performance of an ultra wideband system using DS-CDMA multiple access technique, Stochastic Tapped Delay Line (STDL) channel model and RAKE receiver with individual cases of 8 and 12 Rake arms and various types of sub-nanosecond pulses as well as multiple user interference (MUI) have been presented. The chip rate is 320 Mchip/s with spreading length of 16 (Walsh Hadamard code). The arrival of multiple users is assumed to be Gaussian in nature and exist over the entire period. Performance of various sub-nanosecond pulses in STDL model has been discussed in the next sub-section which is followed by a discussion on results of system performance in the presence of multiple user interference.

A. The effect of pulse shapes on channel model

Figures 3.a and 3.b show the results of simulation for different pulse shapes with a number of Rake receiver taps equal to 8 and 12 at a distance of 5m with STDL model of the channel and in the presence of white Gaussian noise. For each case, synchronization and channel estimation are assumed to be perfect. The results show that the pulse shape has noticeable impact on the performance of the system. It has been found in simulation that the performance of third derivative of Gaussian pulse is better than any other pulses.

For the Bit Error Rate (BER) of 10^{-4} , using RAKE arms equal to 8 at a distance of 5m as shown in Figure 3.a, the third derivative of Gaussian pulse has 3.5 dB E_p/N_0 gain over that value of Gaussian monopulse, whereas the fifth

derivative of Gaussian pulse has almost the same E_b/N_o ratio as that of Gaussian pulse. Similarly, for the BER of 10^{-4} , increasing the number of RAKE arms to 12, as shown in Figure 3.b at a distance of 5m the third derivative of Gaussian pulse has 2.9 dB E_b/N_o gain over that value of Gaussian monopulse, whereas the fifth derivative of Gaussian pulse needs 0.5 dB E_b/N_o more than what is required by the Gaussian pulse.

Pulse shape has noticeable impact because in the multipath environment there is more chance that the pulses will arrive with different delay, and the arrived components are out of phase with each other. In UWB systems, baseband pulses and BPSK modulation schemes have been used for simulation. If the arrived components are out of phase, in such case the resulting output of the receiver depends on pulse shape. A comparison between performances shown in Figure 3.a and Figure 3.b also indicates a performance improvement in multipath fading environment as a result of using an increased number of RAKE arms.

B. The effect of Multiuser Interference on System Performance

Figures 4.a, 4.b and 4.c show system performances in presence of multiple user interference at a distance of 5m. As shown in the above figures, it is found that degradation in the performance of third derivative of Gaussian pulse is less than the degradations of Rayleigh pulse and second derivative of Rayleigh pulse in the presence of multiple user interference. Simulations at a distance of 7m without interference and in the presence of interference have also been performed and the results have been shown in Figures 5.a, 5.b and 5.c. At a distance of 7m the degradation in the performance is more if the number of interferences is increased. At the distance of 7 meters, performance of third derivative of Gaussian pulse is also better than other sub-nanosecond pulses. The results at a distance of 10m and 12m without interference are shown in Figure 6 and in Figure 7 respectively. The result showed that third derivative of Gaussian pulse shows better performance all other pulses.

V. VALIDATION INTO FCC POWER MASK REQUIREMENT

In this section, an analytical approach has been shown for the validation that the third derivative of Gaussian pulse suits the FCC power emission requirement the best among all possible pulse shapes, especially when the wireless channel is impaired with the presence of multiple paths and Multi-user Interference (MUI). According to Part 15 of FCC Regulation on UWB emission limits in the form of a spectral mask for indoor and outdoor systems [12], UWB radios can emit a peak transmit power of -41.3 dBm/MHz in the frequency band from 3.1 GHz to 10.6 GHz. Outside this band, the Power Spectral Density (PSD) must be decreased in order not to interfere with other wireless communication systems operating in that frequency bands. From 0.96 GHz to 1.61 GHz range, the reduction in admissible transmitted power is necessary to protect GPS transmissions. Also, to

protect PCS transmission for outdoor systems in the band from 1.99 GHz to 3.1 GHz, the required backoff to be ensured is 20 dB for outdoor systems.

The PSD of the first derivative Gaussian pulse does not meet the FCC power emission requirement regardless of the pulsewidth being used. On the other hand, as indicated in Sect. IV, it has been found by simulation in our experiment that higher order derivatives of Gaussian pulse have significant E_b/N_o gain for the same BER value over all other pulses used. Therefore, an analytical investigation has been made to research into the PSD of higher order derivatives of Gaussian pulses, more specifically for third derivative of Gaussian pulse. In Figure 8, the normalized PSD has been drawn for first to fifth order derivatives of Gaussian pulse for a chosen value of σ . The value of σ has been calculated according to Bisection method by numerical analysis [14]. The normalization factor is the peak value allowed by the FCC, -41 dBm/MHz. It is clear from the figure that increasing order of derivative increases the center frequency and thus help to suit the FCC power emission regulation, while at the same time reduces the relative bandwidth [14]. Shifting the center (peak) frequency and adjusting the bandwidth so that the FCC requirements are met could be done by modulating the monocycle with a sinusoid by varying the values of σ . But since Impulse radio Ultra Wideband (IR-UWB) systems are carrierless systems, modulation will increase the cost and complexity in design. After an analytical explanation described in Appendix A, it can be concluded that higher order derivatives have a clear impact in fulfilling the FCC PSD mask for UWB transmission and more specifically, third order derivative of Gaussian pulse meets the FCC requirement quite well in 3.1 GHz – 10.6 GHz region while providing a good trade-off between PSD and operational bandwidth. The 3dB bandwidths of higher order Gaussian pulse have been calculated and shown in Figure 9 [14]. As order of derivative increases, the relative bandwidth decreases. This point is important for choosing the third order derivative, since its relative bandwidth is much larger than 4th or 5th order derivatives. Also, it has been found from simulation that for the same BER value to be ensured the required E_b/N_o value for the third derivative of the Gaussian pulse is the minimum among all. A more in-depth research is necessary to shape the third derivative of Gaussian pulse so that it fits the FCC PSD requirement more accurately in the 0.96 GHz-1.61 GHz and 1.99 GHz-3.1 GHz ranges. The effectiveness of matched filter in this case can also be investigated.

V. CONCLUSIONS

At present, Ultra Wideband (UWB) radio is an emerging technology especially in short distance indoor communications. In this study, the performance in the form of Bit Error Rate (BER) in MUI environment for Ultra Wideband radio has been described. Stochastic Tapped Delay Line (STDL) model derived basically from Saleh-Valenzuela model and accepted as 'modified S-V model' by IEEE 802.15.3a study group, has been chosen

here to investigate the system performance for various sub-nanosecond pulses and in the presence of multi-user interference. DS-CDMA, a comparatively less complicated and cheaper multiple access technique currently available, is used in the transmitter section. Till date, the most currently adopted and widely accepted pulse shape used for UWB indoor communication is modeled as the second derivative of Gaussian function. [13]. Experiments are still going on so that pulse shapes meet the FCC requirement more accurately. In our experiment, it has been found that third derivative of Gaussian pulse shows the best performance among all pulses used even in the presence of MUI environment. Gaussian pulse and its derivatives have been given more importance because they can be generated at the pulse generator at the easiest way. In our simulation, the best results have been obtained for the third derivative of the Gaussian pulse, where considering the antenna effects is referred as future work of the current research. Through spectral analysis it has also been proved that third derivative of Gaussian pulse fits the FCC power emission mask quite effectively in the 3.1 GHz – 10.6 GHz range where UWB devices are supposed to be used in indoor communications.

REFERENCES

- [1] Yomo, H., Popovski, P., Wijting, C., Kovacs, I. Z., Deblauwe, N., Baena, A. F., and Prasad, R. (2001), "Medium Access Techniques in Ultra-Wideband Ad Hoc Networks", paper in the context of the IST-2001-34157 Power aware Communications for Wireless OptiMised personal Area Network (PACWOMAN), the IST program partially funded by the EC. Web: www.imec.be/pacwoman/publications/ WP8-CPK-MAC-UWB-ETAI2003-09-07-2003-V1.0.pdf
- [2] D. Barras, F. Ellinger, H. Jäckel, "A comparison between ultra-wideband and narrowband transceivers", TRLabs/IEEE Wireless 2002, Calgary, July 2002
- [3] Ramrez-Mireles, F., (2001), "On the Performance of UWB Signals in Gaussian Noise and Dense Multipath," IEEE Transaction on Vehicular Technology, January, Vol. 50, Issue: 1, pp. 244 -249.
- [4] Hashemi, H., (1993), "Impulse Response Modeling of Indoor Radio Propagation Channels," IEEE Journal on Selected Areas in Communications, September, Vol. 11, No. 7, pp 967-978.
- [5] Win, M. Z. and Scholtz, R. A., (2000), "Ultra wide bandwidth Time Hopping Spread Spectrum Impulse Radio for Wireless Multiple Access Communications," IEEE Transactions on Communications, April, Vol. 48, No. 4, pp 679-691.
- [6] Foerster, J. R., (2001), "The Effects of Multipath Interference on the Performance of UWB Systems in an Indoor Wireless Channel," Proceedings of the IEEE Vehicular Technology Conference, VTC Spring 2001, Rhodes, Greece, 6-9 May, Vol. 2, pp. 1176 -1180.
- [7] Li, Q. and Rusch, L. A., (2002), "Multiuser Detection for DS-CDMA UWB in the Home Environment," IEEE Journal on selected Areas in Communications, December, Vol. 20, No. 9, pp 1701-1711.
- [8] Cassioli, D., Win, M. Z. and Molisch, A. F., (2002), "The Ultra-Wide Bandwidth Indoor Channel: From Statistical Model to Simulations," IEEE Journal on Selected Areas in Communications, August, Vol. 20, No. 6, pp. 1247-1257.
- [9] Tesi, R., Hämäläinen, M., Iinatti, J., Oppermann, I. and Hovinen, V. (2004), "On the Multi-User Interference Study for Ultra Wideband Communication Systems in AWGN and Modified

Saleh-Valenzuela Channel" in Proceedings of 2004 International Workshop on Ultra Wideband Systems joint with Conference on Ultra Wideband Systems and Technologies, May 18-21, Kyoto, Japan

- [10] Conroy, J. T., LoCicero, J. L. and Ucci, D. R., (1999), "Communication Techniques using Monopulse Waveform," Proceedings of the IEEE Military Communications Conference, MILCOM'99 Proceedings, Atlantic city, NJ, USA, 30 October - 3 November, Vol. 2, pp. 1181-1185.
- [11] Nee, R. V. and Prasad, R. (2000), "OFDM for Wireless Multimedia Communications", Artech House Boston, USA, ISBN 0-89006-530-6.
- [12] Federal Communications Commission, "Revision of Part 15 of the commission's rules regarding ultra-wideband transmission systems, FIRST REPORT AND ORDER," ET Docket 98-153, FCC 02-48, pp. 1-118, February 14, 2002.
- [13] Di-Benedetto, M.-G. and Giancola, G. (2004), "Understanding Ultra Wide Band Radio Fundamentals", 1st Edition, 2004, pg. 188, Prentice Hall PTR, New Jersey, USA, ISBN: 0-13-148003-0
- [14] Hongsan Sheng, P. Orlik, A. M. Haimovich, L. J. Cimini, Jr. and J. Zhang, "On the spectral and power requirements for Ultra-wideband transmission," in Proceedings of IEEE 2003 International Conference on Communications (ICC), vol. 1, May 2003, Anchorage, AK, pp. 738-742.

APPENDIX: A

Using the expression described in Section III.A n-th order derivative of Gaussian pulse can be obtained recursively from

$$\Omega_G^n(t) = -\frac{n-1}{\sigma^2} \Omega^{(n-2)}(t) - \frac{t}{\sigma^2} \Omega^{(n-1)}(t)$$

The Fourier transform of the n-th derivative of Gaussian pulse is

$$\Omega_n(f) = A(j2\pi f)^n e^{-\frac{(2\pi f\sigma)^2}{2}}$$

from where the amplitude spectrum can be considered as

$$|\Omega_n(f)| = A(2\pi f)^n e^{-\frac{(2\pi f\sigma)^2}{2}}$$

The peak emission frequency, f_M can be determined by equating $\frac{d|\Omega_n(f)|}{df}$ to zero. The value of f_M and the

corresponding $|\Omega_n(f_M)|$ are found as

$$f_M = \frac{\sqrt{n}}{2\pi\sigma} \quad \text{and} \quad |\Omega_n(f_M)| = A \left(\frac{\sqrt{n}}{\sigma} \right)^n e^{-\left(\frac{n}{2}\right)}$$

The normalized PSD, $|P_n(f)|$ can be defined as

$$|P_n(f)| \equiv \frac{|\Omega_n(f)|^2}{|\Omega_n(f_M)|^2} = \frac{(2\pi f\sigma)^{2n} e^{-(2\pi f\sigma)^2}}{n^n e^{-n}}$$

which has a peak value of 1 (0 dB). Considering the n-th derivative of Gaussian pulse as the transmitted pulse of UWB transmission where A_{\max} is the peak power spectral density that has been set as limit by FCC then the PSD of the transmitted signal can be expressed as

$$|P_n(f)| \equiv A_{\max} |P_n(f)| = \frac{A_{\max} (2\pi f \sigma)^{2n} e^{\{-2\pi f \sigma\}^2}}{n^n e^{(-n)}}$$

It has also been found from the expression of f_M that f_M varies proportionally with \sqrt{n} for a given decaying factor, σ . Gaussian derivatives of higher order are characterized by higher peak frequencies. Thus, differentiation is a way to move energy to higher frequency bands.

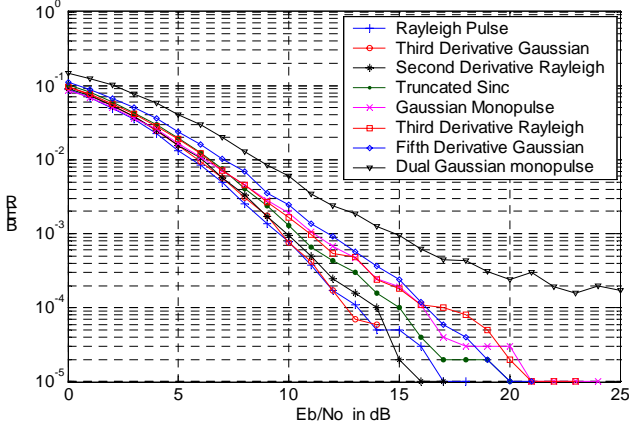


Fig. 3.a: BER performance of different pulses in STDL model of the UWB channel at a distance of 5m in the presence of AWGN with 8 arms

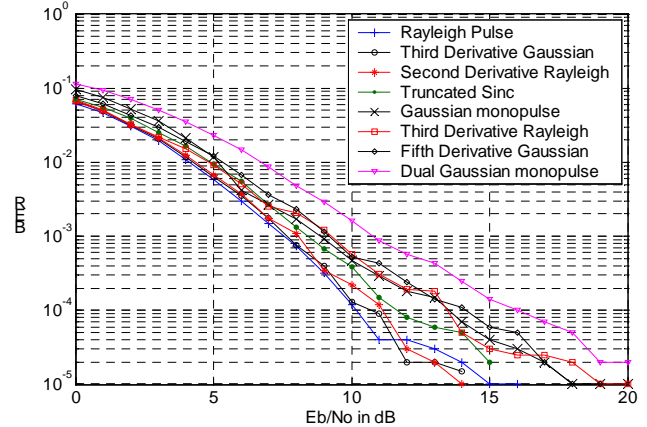


Fig. 3.b: BER performance of different pulses in STDL model of the UWB channel at a distance of 5m in the presence of AWGN with 12 arms

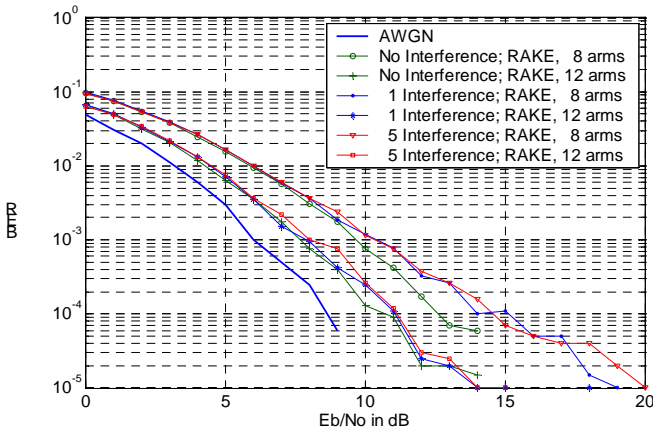


Fig. 4.a: Comparison of the performance at a distance of 5m with third derivative Gaussian pulse

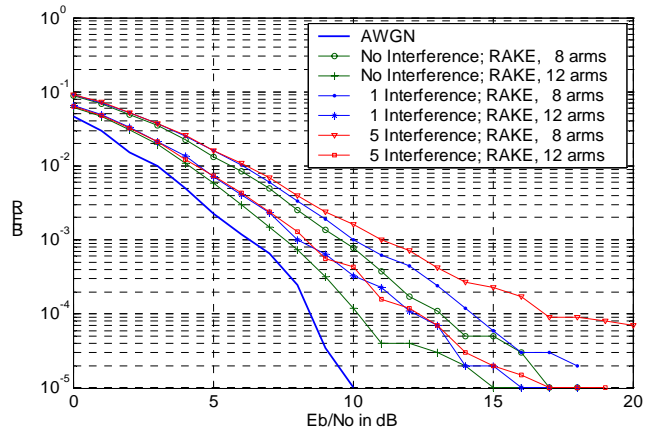


Fig. 4.b: Comparison of the performance at a distance of 5m with Rayleigh pulse

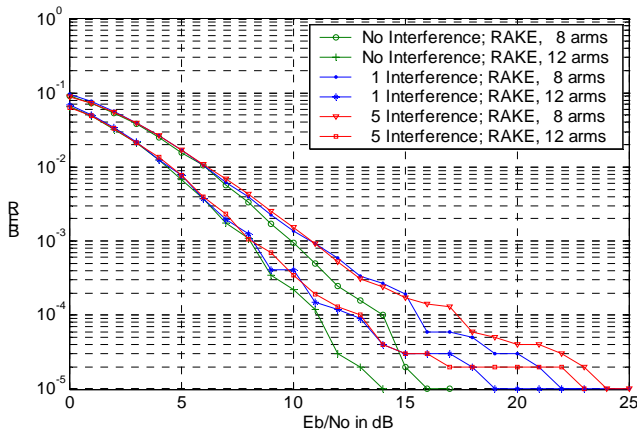


Fig. 4.c: Comparison of the performance at a distance of 5m with second derivative Rayleigh pulse

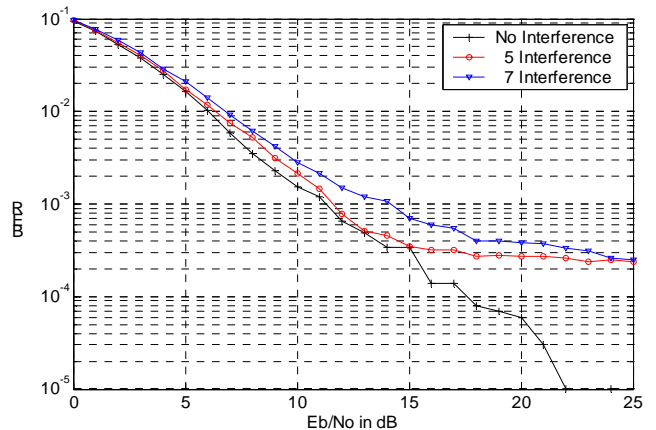


Fig. 5.a: Third derivative of Gaussian pulse at a distance of 7m with RAKE 12 arm

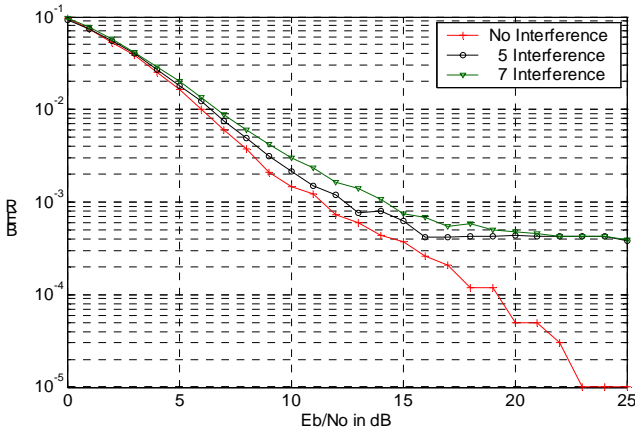


Fig 5.b: Rayleigh pulse at a distance of 7m with RAKE 12 arm

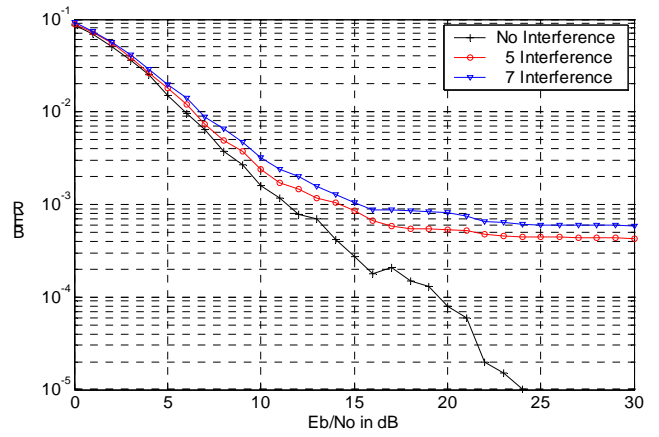


Fig 5.c: Second derivative of Rayleigh pulse at a distance of 7m with RAKE 12 arm

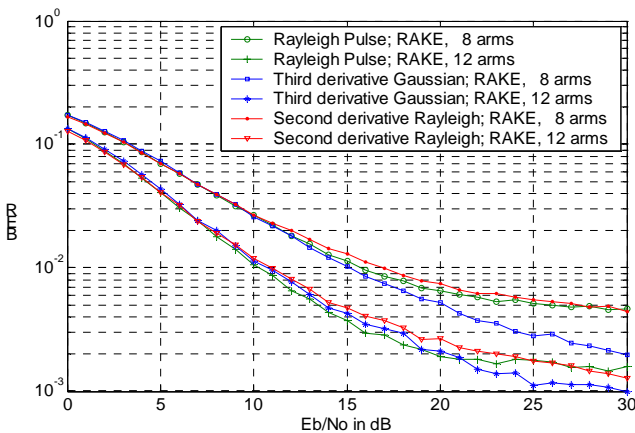


Fig 6: BER Performance of only one user at a distance of 10m

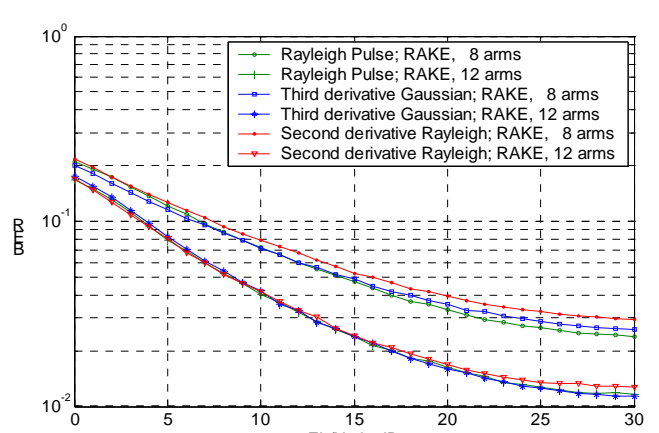


Fig 7: BER Performance of only one user at a distance of 12m

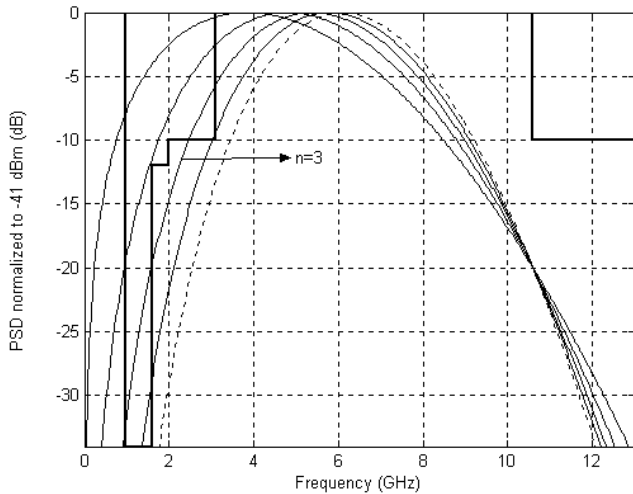


Fig 8: PSD of higher order derivatives of Gaussian pulse for UWB indoor communication

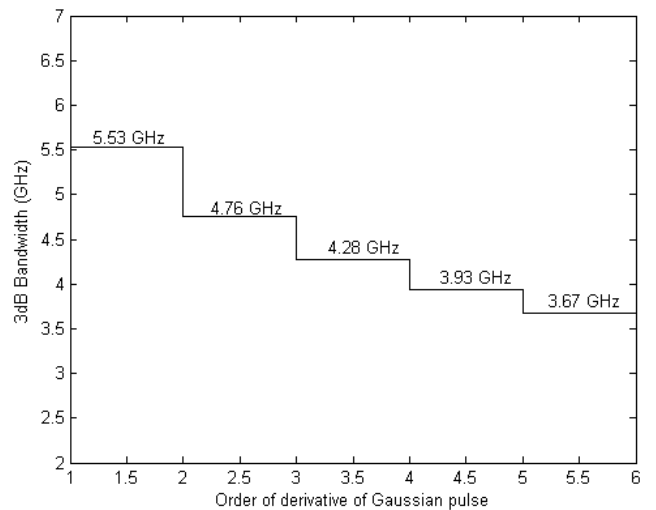


Fig 9: 3dB Bandwidths for higher order derivatives of Gaussian pulse

Performance Evaluation of Coded UWB-IR on Multipath Fading Channels

Michal M. Pietrzyk and Jos H. Weber

Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology
Mekelweg 4, P.O. Box 5031, 2600 GA Delft, The Netherlands
Telephone: +31 15 27 81609, Fax: + 31 15 27 81774
Email: M.M.Pietrzyk@ieee.org, J.H.Weber@ewi.tudelft.nl

Abstract—In most research on error correction coding for UWB techniques the channel is assumed to be Gaussian, whereas the multipath case is neglected. In this paper, we evaluate the performance of a realistic and feasible UWB-IR system in a severe multipath environment. We model the effect of the transmitter and receiver antennas on the pulse shape, by using their real characteristics obtained through the measurements. We present a general coding-modulation scheme for UWB communications and focus on two particular cases, namely, one using superorthogonal convolutional coding, and the other based on simple UWB frame repetition. Our theoretical results, confirmed by simulations, show that superorthogonal convolutional coding provides a more effective way of protection against errors than simple frame repetition.

Index Terms—Ultra-wideband, channel coding, frame repetition, multipath.

I. INTRODUCTION

Ultra-wideband Impulse Radio (UWB-IR) has several unique characteristics that make it a promising candidate for future wireless communications. Exceptionally low transmission power and very large available bandwidth enable a UWB system to co-exist with narrowband systems. The large bandwidth occupied by UWB systems allows for high data rate transmission. However, the interference issues pose restrictions on the maximum data rate. One possible solution to ensure a desired data rate and simultaneously maintain a certain performance level is to apply channel coding. Although several channel coding schemes have already been proposed [1], [2], [3], research into their performance under realistic UWB channel conditions is limited. Such performance evaluation is of great importance, since the investigations up to now, for instance in [1], [3], have been limited to the AWGN case, which does not correspond to the conditions in a typical indoor environment.

The goal of this paper is to evaluate the performance of UWB-IR systems incorporating superorthogonal convolutional (SOC) coding or a frame repetition scheme in the presence of severe multipath. The investigated UWB-IR system employs a differential autocorrelation receiver with a realistic and accurate UWB channel model. The channel model used in our simulations is a modified Saleh-Valenzuela model [4] that

has been recently proposed by the IEEE 802.15.3a channel modeling subcommittee for the evaluation of the UWB physical layer submissions. This model is based on measurements spanning the frequency spectrum from 2 to 8 GHz. In this model, the path resolution time equals 0.167 ns, enabling reliable estimation of the real UWB channel behavior.

We evaluate the performance of the UWB-IR system using theoretical analysis as well as Monte Carlo simulations. For the case of a multipath fading channel, both line-of-sight (LOS) and non-line-of-sight (NLOS) environments are considered. Our results show that the performance of the UWB-IR system can be significantly enhanced by the use of SOC coding instead of the frame repetition scheme, without costs in terms of additional bandwidth expansion.

This paper is organized as follows. Section II describes the structure of the considered UWB-IR system with insight into modulation format, pulse shaping, channel model, and receiver architecture. Furthermore, principles of the proposed coding-modulation scheme are given in detail. Section III focuses on the performance evaluation of the considered UWB-IR system by means of theoretical and numerical analysis. Finally, Section IV presents conclusions.

II. SYSTEM MODEL

A. General Coding-Modulation Scheme

The proposed general coding-modulation scheme for a UWB-IR technique is depicted in Figure 1. Every packet consists of a number of information bits, each of duration T_b . A selected channel coding scheme is applied on k information bits, resulting in n output code symbols. Every code symbol is then represented by N_f UWB frames, each of duration T_f . Every frame consists of one pulse that is pseudorandomly assigned to one of N_p time slots. In this paper, we consider a single user scenario. We focus on two particular cases of the general coding-modulation scheme, one further referred to as a UWB-IR system with a SOC code, for which $k = 1, n > 1$, and $N_f = 1$, and the other, further referred to as a UWB-IR system with frame repetition, in which there is no coding scheme applied, i.e., $k = n = 1$ and $N_f > 1$. In order to allow a fair comparison between the two schemes, we choose

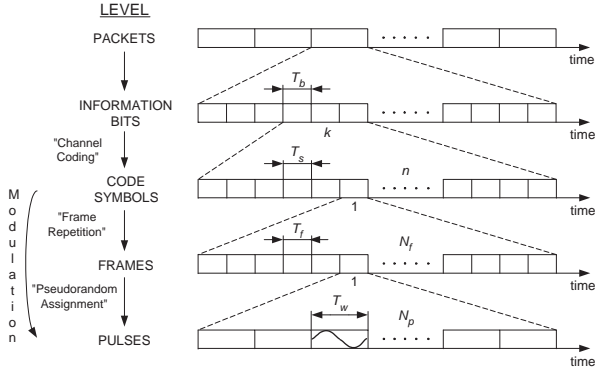


Fig. 1. Diagram showing a general coding-modulation scheme in a UWB-IR system.

the n of the SOC code equal to the N_f of the frame repetition scheme. In this way, an equal number of transmitted pulses per information bit is guaranteed.

B. SOC Coding

In the UWB-IR system with the SOC code, a data information bit is encoded by the SOC encoder with code rate of $R = 1/n$, where $n = 2^{K-2}$ and K is the constraint length. The SOC encoder consists of a K -stage shift register, a bit orthogonal block encoder, and a modulo-2 adder with 3 inputs, as it is shown in Figure 2. The block encoder is a Hadamard-Walsh encoder with length $K - 2$. The decoding process is performed with the use of the Viterbi algorithm with 2^{K-1} states. The branch metrics are calculated according to the soft output of the differential autocorrelation receiver. An important feature of the SOC decoder is that processing complexity of the decoder grows only linearly with K , making the decoder feasible even for high values of K [5].

C. Modulation Format

We consider a differential autocorrelation modulation format with the following set of signal waveforms [6]: $S = \{s_0(t) = s(t), s_1(t) = -s(t)\}$, where $s(t)$ is defined as

$$s(t) = \sum_{j=0}^{N_f-1} w(t - jT_f - c_jT_w), \quad 0 \leq t < T_s, \quad (1)$$

where $w(t)$ is the Gaussian monocycle, N_f is the number of pulses transmitted per code symbol, and T_f is the frame time, also known as the average pulse repetition time. The term c_jT_w determines the position of the pulse within a frame and T_w denotes the pulse duration. The pseudorandom code sequence c_0, \dots, c_{N_f-1} assigning the pulse within the frame is fixed for every packet and generated according to the uniform distribution in the range $0 \leq c_j \leq N_p - 1$. As in [6], we call the transmission of a logical code symbol "1" as H_1 and the transmission of a logical code symbol "0" as H_0 . When H_0 is true, the transmitter generates the same signal waveform as transmitted in the previous symbol time. Conversely, when H_1 is true, the transmitter switches to the antipodal signal waveform.

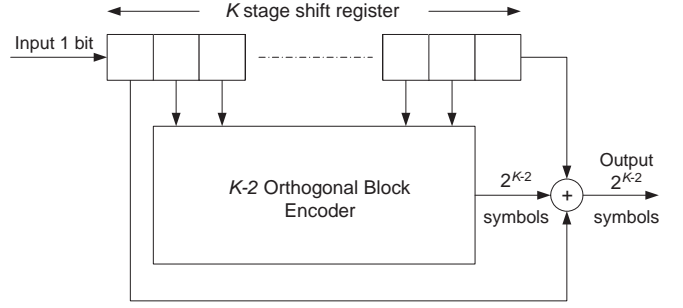


Fig. 2. Diagram showing a superorthogonal convolutional encoder architecture.

D. Pulse Shape

We model the transmitted pulse as a distorted Gaussian monocycle. The Gaussian monocycle is the first derivative of the Gaussian pulse and is given by

$$w(t) = \frac{2At}{\sigma^2} e^{-\left(\frac{t}{\sigma}\right)^2}, \quad (2)$$

where A is the amplitude and σ is the temporal width parameter. The practical advantage of the Gaussian monocycle, in comparison to the Gaussian pulse, is that it does not contain a DC component, allowing for simplified transmitter architecture. In order to characterize the UWB-IR system as accurate as possible, we model distortions introduced by a bandpass filter and amplifier by a third-order passband Chebyshev filter with the cutoff frequencies $f_1 = 2$ GHz and $f_2 = 8$ GHz, on which the magnitude response of the filter equals -0.2 dB. In Figure 3, the transfer function of this filter is denoted as $H_2(f)$. Moreover, we model the effect of the transmitter and receiver antennas on the pulse shape, by employing the data collected in [7]. In Figure 3, the transfer function of the antenna is denoted as $H_1(f)$. The width of the transmitted pulse T_w corresponds to the channel model time resolution and equals $T_w \cong 0.167$ ns. The original and modeled received pulses are depicted in Figure 4.

E. UWB Channel Model

Since the performance analysis of a UWB-IR system is based on statistics of the channel, we select a model providing an accurate description of the real UWB channel conditions. The chosen channel model was developed at Intel [4] and is a modified Saleh-Velenzuela (S-V) model. The main difference is that instead of a Rayleigh probability density function (p.d.f), the Intel model employs a lognormal p.d.f. for the fading channel coefficients. The impulse response is given by [4]

$$h(t) = \sum_{l=1}^L \sum_{m=1}^M \alpha_{m,l} \delta(t - T_l - \tau_{m,l}), \quad (3)$$

where M is the number of paths within a cluster, L is the number of clusters, $\alpha_{m,l}$ is the multipath gain coefficient, T_l is the delay of the l -th cluster, and $\tau_{m,l}$ is the delay of the

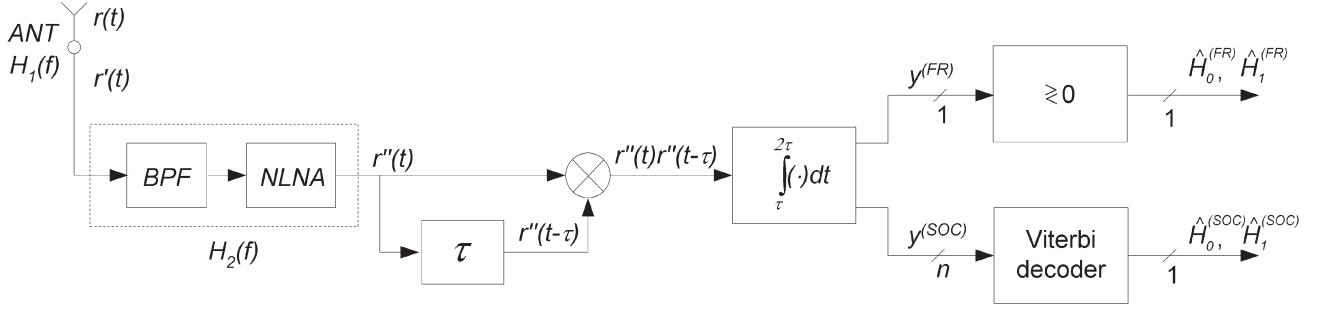


Fig. 3. Diagram showing the modeled UWB-IR receiver architecture.

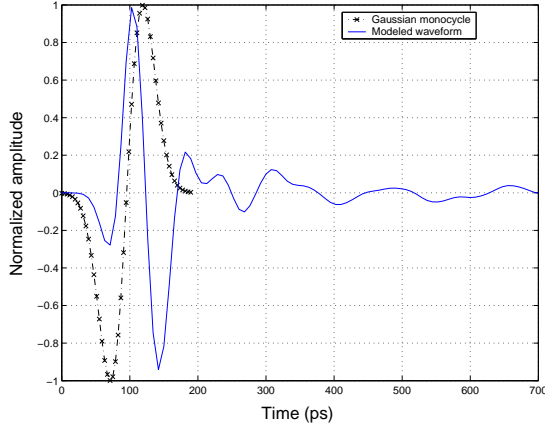


Fig. 4. The Gaussian monocycle and the modeled received waveform.

m -th multipath component relative to the l -th cluster arrival time T_l . The multipath channel coefficients are defined as follows: $\alpha_{m,l} = p_{m,l}\beta_{m,l}$, where $p_{m,l}$ denotes the sign of the coefficient and is equally likely to take values of ± 1 , and $\beta_{m,l}$ is the lognormal fading term where $20 \log(\beta_{m,l})$ follows a normal distribution. The inter-cluster and inter-path arrival times are exponentially distributed. The main characteristics of the model are RMS delay spreads and mean number of significant paths ranging from 5-25 ns and 20-120, respectively. Table I shows the set of parameters used in our model, as suggested in [4], for LOS and NLOS environments. The parameter NP_{10dB} denotes the number of significant paths that cross a 10 dB threshold.

F. Receiver Architecture

A simplified block diagram of the modeled UWB-IR receiver is shown in Figure 3. The input to the receiver is a signal $r(t)$. After passage through an antenna, the signal $r'(t)$ feeds a bandpass filter, and then a nonlinear amplifier. Next, a resulting signal $r''(t)$ is directed to a differential autocorrelator that correlates the signal with its symbol-delayed version. Depending on the coding scheme used, the results of correlation are directed to a threshold detector or a Viterbi decoder.

Receivers that are based on aurocorrelator are feasible and have numerous implementation advantages compared to other

TABLE I
CHANNEL CHARACTERISTICS

Environment	LOS	NLOS
RMS Delay Spread (ns)	9	15
NP_{10dB}	7	35

types of receivers including, for instance, RAKE receivers. Such receivers do not require a priori knowledge of the pulse to correlate and are less susceptible to jitter on the receiver clock. However, the price for all of these advantages is that BER performance is worse than that of the system employing the RAKE receiver. When H_0 is true, the received waveform can be expressed as [6]

$$H_0 : r(t) = (s_m(t) + s_m(t - T_s)) * h(t) + n(t), \quad (4)$$

whereas when H_1 is true, the received waveform is

$$H_1 : r(t) = (s_m(t) + s_n(t - T_s)) * h(t) + n(t), \quad (5)$$

where $m = 0, 1$, $n = (m + 1) \bmod 2$, $0 < t \leq 2T_s$, $n(t)$ is zero mean additive white Gaussian noise, and $*$ denotes the convolution. The autocorrelator output is given by

$$y = \int_{T_s}^{2T_s} r(t)r(t - T_s)dt. \quad (6)$$

III. PERFORMANCE EVALUATION

We compare the performance of the UWB-IR system incorporating superorthogonal convolutional coding with the performance of the UWB-IR system with frame repetition. The data rates of both systems are the same and the bandwidth expansion introduced by SOC coding and the frame repetition scheme is equal. We will show that superorthogonal convolutional coding provides significant coding gain in comparison with the simple frame repetition scheme. Table II shows the parameters of the considered UWB-IR system models.

A. Bounds on Bit Error Probabilities on AWGN Channel

The upper bound on the bit error probability of the UWB-IR system with the superorthogonal convolutional code is derived

TABLE II
SIMULATION PARAMETERS

Bandwidth		$B = 6$ GHz
Modulation		Differential Autocorrelation
Pulse Width		$T_w \simeq 0.167$ ns
Bit Rate		$R_b = 125$ Mbps
Processing Gain		$G_p = 48$
SOC Channel Coding	Coding Scheme	SOC
	Constraint Length	$K = 4, 5$
	Code Rate	$R = 1/4, 1/8$
	Decoding Algorithm	Soft-Input Viterbi Algorithm
Frame Repetition	Coding Scheme	None
	Number of Frame Repet.	$N_f = 4, 8$
Number of Pulse Positions		$N_p = 12, 6$
Channel Model		AWGN, LOS, NLOS

from the graph generating function of the code that is given by [5]

$$T_{SOC}(W, \beta) = \frac{\beta W^{K+2}(1-W)}{1-W[1+\beta(1+W^{K-3}-2W^{K-2})]}, \quad (7)$$

where $W = Z^{K-3}$. Expanding the above expression we get a polynomial in which the exponent of W gives the path weight and the exponent of β gives the path length, that is, the number of state transitions associated with the path. The parameter β denotes the information error weight. The parameter Z can be calculated from the Bhattacharyya bound as

$$Z = \int_{-\infty}^{\infty} \sqrt{p_0(y)p_1(y)} dy, \quad (8)$$

where $p_0(y)$ and $p_1(y)$ are the density functions of the receiver/channel output conditioned on the input symbol being 0 and 1, respectively. The upper bound on the bit error probability of the UWB-IR system is expressed as

$$P_b < \left. \frac{\partial T_{SOC}(W, \beta)}{\partial \beta} \right|_{\beta=1} = \frac{W^{K+2}}{(1-2W)^2} \left(\frac{1-W}{1-W^{K-2}} \right)^2. \quad (9)$$

For a Gaussian channel, the parameter W can be calculated as $W = \exp(-\gamma)$, where γ denotes the signal-to-noise ratio at the input of the SOC decoder. Since the relationship binding the input and output signal-to-noise ratio of the differential autocorrelation receiver for the Gaussian monocycle is compound, for simplicity, as in [6], we consider a rectangular monocycle waveform having

$$\gamma \cong \frac{G_p \gamma_{in}}{1 + (2\gamma_{in})^{-1}}, \quad (10)$$

where γ_{in} can be calculated from

$$\gamma_{in} = \frac{E_b}{N_0} G_p^{-1}. \quad (11)$$

The parameter G_p denotes the processing gain of the UWB-IR system and is defined as

$$G_p = \frac{B}{R_b} = BN_f N_p T_w \frac{n}{k}, \quad (12)$$

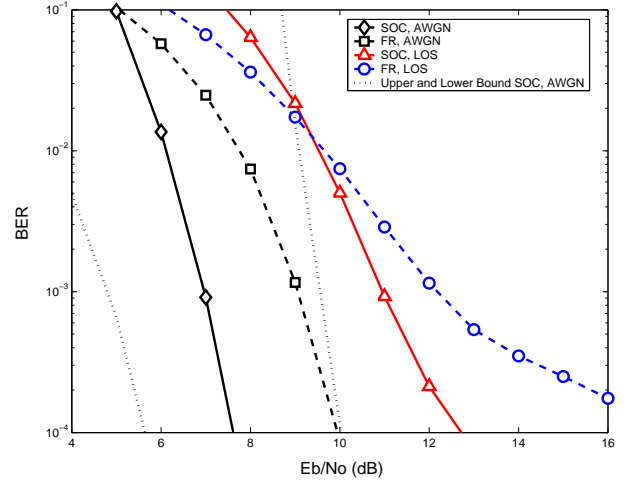


Fig. 5. BER performance of the UWB-IR systems with superorthogonal convolutional (SOC) coding and frame-repetition (FR) for $N_f = 8$, $N_p = 6$, $K = 5$ in AWGN and LOS environments.

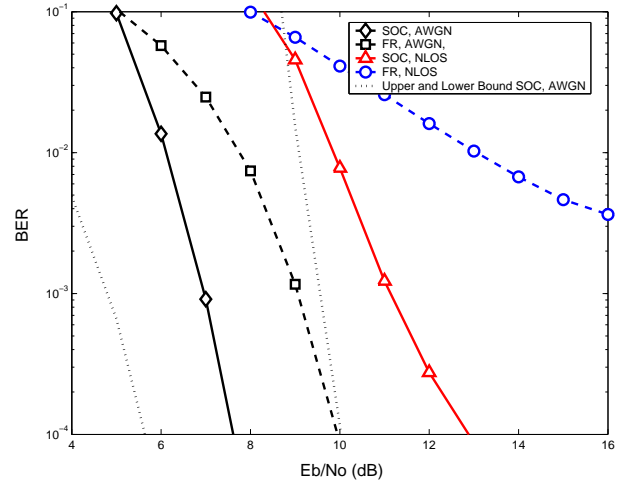


Fig. 6. BER performance of the UWB-IR systems with superorthogonal convolutional (SOC) coding and frame-repetition (FR) for $N_f = 8$, $N_p = 6$, $K = 5$ in AWGN and NLOS environments.

where B is the bandwidth and R_b is the bit rate. From (7) we can also compute free distance of the SOC code with the constraint length K as $d_f^{(SOC)} = 2^{K-3}(K+2)$. Comparing this value with the free distance of the simple frame repetition scheme $d_f^{(FR)} = 2^{K-2}$, it can be easily observed that SOC coding enables substantially better performance in comparison to frame repetition.

The lower bound on the bit error probability of the UWB-IR system incorporating the SOC code can be calculated as [1]

$$P_b \geq Q \left(\left(\frac{\mu^2}{\sigma^2} d_f \right)^{1/2} \right), \quad (13)$$

where μ and σ^2 are the mean and the variance of the autocorrelation receiver output conditioned on the input symbol

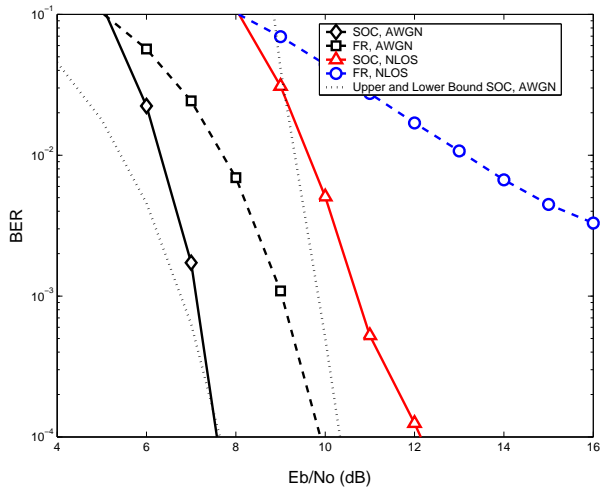


Fig. 7. BER performance of the UWB-IR systems with superorthogonal convolutional (SOC) coding and frame-repetition (FR) for $N_f = 4$, $N_p = 12$, $K = 4$ in AWGN and NLOS environments.

being zero. In Figures 5-7 the lower and upper bounds are represented as dotted lines without markers.

B. Simulation Results

Apart from the theoretical analysis, we evaluate the performance of the considered UWB-IR systems using Monte Carlo simulations. The BER performance is examined using 40 channel realizations and 2000 bits in every packet. Our assumption is that the channel is invariant during the duration of a single data packet and the synchronization is ideal. Figure 5 illustrates a comparison between the BER performance of the UWB-IR systems incorporating the superorthogonal convolutional code or the frame repetition scheme in two environments: AWGN and LOS. As can be seen from Figure 5, the performance of the UWB-IR system in a LOS environment is noticeably worse than that in AWGN. The difference in the bit energy between the UWB-IR systems incorporating the SOC code with $K = 5$ on a BER = 10^{-3} level, for AWGN and the case of LOS, equals circa 4 dB. Comparing the performance of the UWB-IR systems based on the SOC code and frame repetition in a LOS environment only, we observe the reduction of 1 dB of the bit energy on BER = 10^{-3} level that is introduced by the SOC coding scheme.

Figures 6 and 7 show the BER performance of the evaluated UWB-IR systems in AWGN and NLOS environments for different sets of system parameters. When considering a NLOS environment, we notice much larger coding gain that is introduced by the SOC coding scheme in comparison to the LOS case. The application of SOC coding in a NLOS environment enables the reduction of more than 6 dB of the bit energy when a considered bit error rate level equals BER = 10^{-3} .

IV. CONCLUSIONS

In this paper, we evaluated the performance of UWB-IR systems incorporating the superorthogonal convolutional coding or the frame repetition scheme using a realistic multipath channel model. We demonstrated that SOC coding significantly outperforms the frame repetition scheme. The coding gain introduced by the SOC scheme is noticeably higher in the NLOS environment. Due to the simple structure of the SOC encoder, decoder and the differential autocorrelation receiver, the UWB-IR system with the SOC scheme can be easily implemented into a hardware platform.

ACKNOWLEDGMENT

This work was partially funded by the Dutch Min. Econ. Affairs and via the *Airlink* project under the Freeband - Impulse Program.

REFERENCES

- [1] A. R. Forouzan, M. Nasiri-Kenari, J. A. Salehi, "Performance analysis of Ultra-wideband time-hopping code division multiple access systems: uncoded and coded schemes," IEEE Int. Conf. on Communications, vol. 10, pp. 3017-3032, Helsinki, Finland, June 2001.
- [2] N. Yamamoto, T. Ohtsuki, "Adaptive internally turbo-coded ultra-wideband impulse radio," IEEE Int. Conf. on Communications, vol. 5, pp. 3535-3539, Alaska, USA, May 2003.
- [3] K. Ikemoto, R. Kohno, "A coded modulation scheme using orthogonal pulses based on low density parity check codes for UWB communications," Int. Workshop on Ultra Wideband Systems, Finland, June 2003.
- [4] J. Foerster, Q. Li, "UWB Channel Modeling Contribution from Intel," IEEE P802.15-02/279r0-SG3a, 2002.
- [5] A. J. Viterbi, "Principles of Spread Spectrum Communications," Int. Workshop on Addison-Wesley Publishing Company, Massachusetts, 1995.
- [6] M. Pausini, G. J. M. Janssen, "Analysis and Comparison of Autocorrelation Receivers for IR-UWB Signals Based on Differential Detection," IEEE Int. Conf. on Acoust., Speech, and Signal Processing, vol. 4, pp. 513-516, Quebec, Canada, May 2004.
- [7] Z. Irahauten, A. Yarovoy, H. Nikookar, G. J. M. Janssen, L. P. Ligthart, "The Effect of Antenna and Pulse Waveform on Ultra-wideband Link Budget with Impulse Radio Transmission," Europ. Microwave Week, pp. 261-264, Amsterdam, The Netherlands, October 2004.

Simulation of Interference Effects from UWB Sources on a Narrowband Digital Transmission System

Idnin Pasya, Atsushi Tomiki, and Takehiko Kobayashi

Wireless Systems Laboratory, Tokyo Denki University
2-2 Kanda-nishiki-cho, Chiyoda-ku, Tokyo 101-8457 Japan.
idnin@grace.c.dendai.ac.jp Tel & Fax: +81-3-5230-3839

Abstract—This paper studies the interference effects from 4 types of ultra wideband (UWB) sources on a narrowband $\pi/4$ -shift differential quadrature phase keying (DQPSK) transmission system by simulation. The culprit UWB sources were: multi-band orthogonal frequency-division multiple-access (MB-OFDM), direct-sequence code-division multiple-access (DS-SS UWB), DS spread spectrum UWB (DS-SS UWB), and additive white Gaussian noise (AWGN). The MB-OFDM and DS-SS UWB were modeled based on the proposal specifications in the IEEE.802.15.3a to standardize high-speed wireless personal area networks. Average bit error rates (BER) degradation of the victim system was evaluated in the presence of the UWB signals as a source of interference. We propose a modified equivalent baseband system to accelerate the simulation speed. In the proposed system, the victim system was generated in the passband domain, while the UWB signals were generated at the equivalent baseband domain to lower the sampling rate of the simulation. It was found that the interference effects of the UWB signals vary according to their statistical characteristics entering the victim receiver. The MB-OFDM marks spectral peaks at every 3.2 MHz in the frequency spectrum, thus, would severely degraded the BER performance in the victim system. The amplitude probability distributions of the UWB signals entering the victim receiver were also investigated.

Index Terms—Ultra wideband, MB-OFDM, DS-SS UWB, interference, equivalent baseband, amplitude probability distribution

I. INTRODUCTION

UWB technologies have attracted considerable interest due to its potential to generate high data rates of communication. UWB systems are expected to coexist with conventional narrowband radio systems in the frequency domain. For this reason, the evaluation of interference effects from UWB systems to existing radio systems are essential for the commercialization of UWB technologies. The regulating authorities in many countries have authorized the emission limit mask for UWB communication systems to protect existing radio services. However, doubts and contradiction in the standardization of these regulations are limiting the realization of UWB systems in the near future.

Regarding coexisting problems, the studies concerning electromagnetic compatibility of UWB systems with other narrowband transmission are strongly encouraged by the Federal Communications Commission and other regulating authorities. Initial studies show that UWB signals closely resembles noise to narrowband receivers. Tesi *et al.* have evaluated the performance of an OFDM receiver under the presence of an impulse radio as an interference source using computer simulations [1], and pointed out that although UWB signals are not Gaussian signals, their interference effects on narrowband systems are equivalent to that of Gaussian noise. Supporting this result, the bit error rates degradation of a digital wireless transmission system caused by impulse radio and DS-SS UWB have been experimentally evaluated [2].

The present work studies the interference effects from 4 typical UWB signals on a narrowband transmission system. The victim system used was the $\pi/4$ -shift DQPSK transmission system, which is widely used modulation scheme in mobile communications nowadays. The simulated UWB signals were MB-OFDM [3] and DS-SS UWB [4],

which were modeled based on IEEE.802.15WPAN (TG3a)'s standard proposal specifications, DS-SS UWB [5], and AWGN.

This paper also investigates the statistical characteristics of the UWB signals by calculating their amplitude probability distribution (APD), which is useful for identifying the signals behavior in the victim receiver. Most receivers are designed to operate in bands with Gaussian noise, which is characterized by the average noise power statistics alone. However, the amplitude statistics of UWB signals are dependent upon their specifications and the frequency entering the band limited filter of the victim receiver. This induces the author to include the APD measurement of the UWB signals in this research.

The remainder of this paper is organized as follows. In Section II, we introduce the simulation model employed in this work. In Section III, we explain the simulated 4 types of UWB sources; The APD characteristics of UWB signals will be discussed in Section IV. In Section V, computer simulation results are presented in terms of BER performance of the victim system. Finally, we draw conclusions in the last section of this paper.

II. SIMULATION MODEL

Figure 1 depicts the simulation model used in this study, which was implemented using SPW® (Signal Processing Worksystem), a software designed for signal processing and numerical simulations. We define the victim system as a $\pi/4$ -shift DQPSK transmission scheme, assuming an ideal modulation at 400 kHz data rates, within a 300 kHz bandwidth. The modulated signal was shifted to the RF band with a carrier wave, and UWB sources are added as a source of interference. Next, thermal noise was added to the transmission signal before being demodulated at the

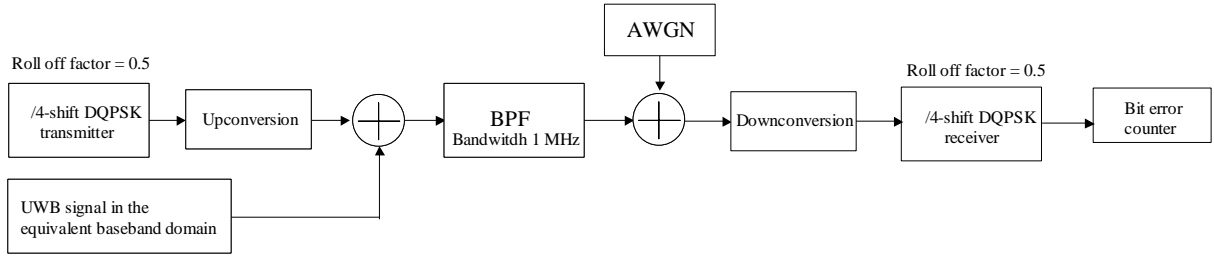


Fig. 1. Simulation diagram.

victim receiver. Note that the indoor multi-path fading was not a subject in this study. Finally the average bit error rates were calculated while verifying the desired-to-undesired signal power ratio (D/U), where D is the transmission signal's average power and U represents the power of the interference signal's average power occupying the same bandwidth.

In real environments, UWB systems would be occupying a high frequency bandwidth (generally 3.1 GHz to 10.6 GHz). However, to generate such high frequency signals, we would need a significantly high sampling rate, which must be twice the highest frequency according to the sampling theory. Thus, the simulation period will be significantly longer. Therefore, we propose a modified equivalent baseband system to speed up the simulation. The modified equivalent baseband system requires lower sampling frequency because the UWB signals are generated in the baseband domain. Figure 2 illustrates the spectral relation between the culprit UWB system and the victim narrowband system, whose center frequencies are f_c and f_v , respectively, in the real radio frequency domain and the modified equivalent baseband domain. The UWB signals have wide frequency spectrum, thus allowing them to overlay with the victim system in the frequency domain, although being generated in baseband. The victim system was shifted from f_v to $f_v - f_c$ in order to tune the center frequency to specific frequencies of the UWB signal.

III. UWB SOURCES USED IN THIS STUDY

This section briefly introduces the 4 types of UWB sources used in this study. As mentioned above, the UWB sources generate complex baseband signals in the equivalent baseband domain. The MB-OFDM and DS-CDMA use the parameters that are being proposed for the IEEE.802.15WPAN (TG3a) standards. The waveforms in the time domain and the spectra in the frequency domain of these signals are shown in Figs. 3, and 4.

A. MB-OFDM

The MB-OFDM is a multi-carrier transmission scheme, where the data bits are mapped to 128 sub-carriers, which are allocated at every 4.125 MHz. The UWB spectrum is divided into several 528 MHz bands, and frequency hopping within these bands are implemented to support multiple accesses. In this study, only one sub-band was used, which means that the frequency hopping was not applied. The total length of an OFDM symbol is 312.5 ns, where 242.2 ns

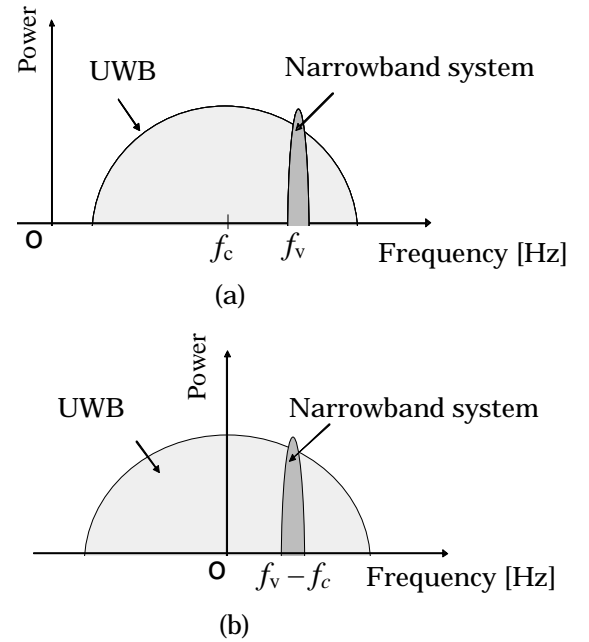


Fig. 2. Frequency spectra of the culprit UWB system and the victim narrowband system in: (a) the real radio frequency and (b) the modified equivalent baseband domain.

of them is the data length, 60.6 ns are the zero-padded prefix, and the other 9.2 ns are the guard interval. We have found that spectrum peaks appear at every 3.2 MHz, as shown in Fig. 4(a), due to the zero padding of the OFDM symbols. These peaks are about 15 dB above the average total power, resulting a series of peaks and valleys in the frequency spectrum. The interference effects from the MB-OFDM at these frequencies needed to be evaluated, and so did the statistical characteristics of the signal.

B. DS-CDMA

Concerning the DS-CDMA signal used in this study, ternary codes are assigned to the modulated symbols from the lookup table as defined in the proposal [3]. The modulation scheme used was binary phase shift keying (BPSK). The pulse was then filtered with a pulse-shaping filter, before being transmitted to the victim's channel. The frequency spectrum was spread to around 800 MHz of bandwidth. The length of each pulse is 9 ns, and the pulse repetition frequency (PRF) is 110 Mb/s.

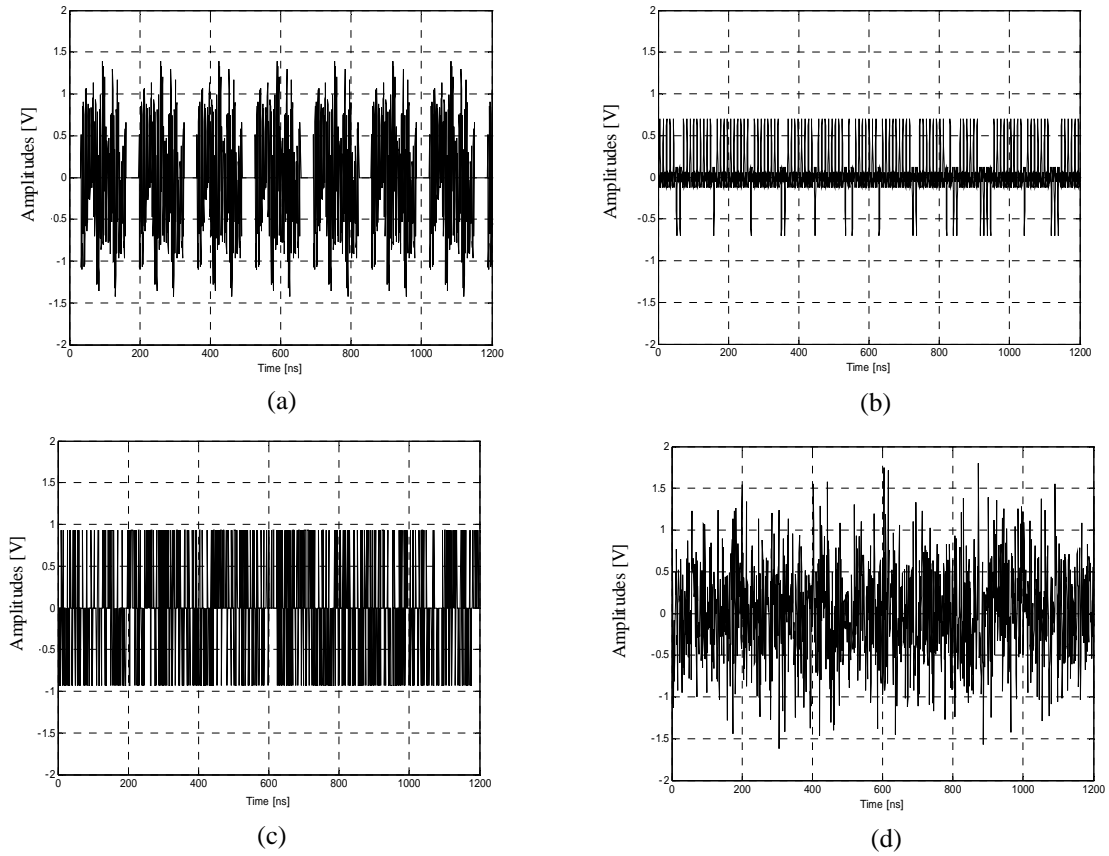


Fig. 3. The wave forms of the culprit UWB signals in the time domain: (a) MB-OFDM, (b) DS-SS UWB, (c) DS-SS UWB, and (d) AWGN.

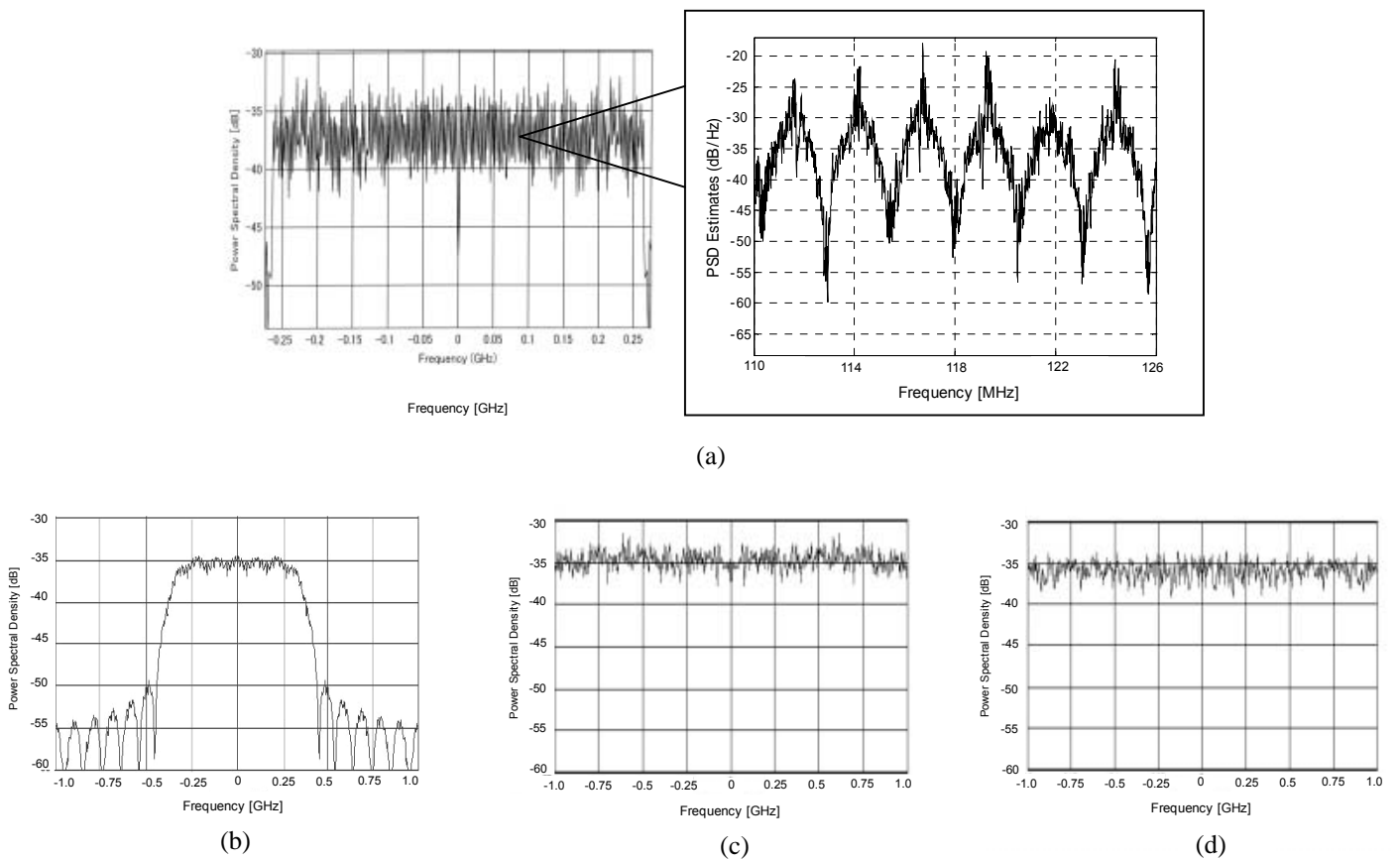


Fig. 4. The power spectra of the culprit UWB signals: (a) MB-OFDM, (b) DS-SS UWB, (c) DS-SS UWB, and (d) AWGN.

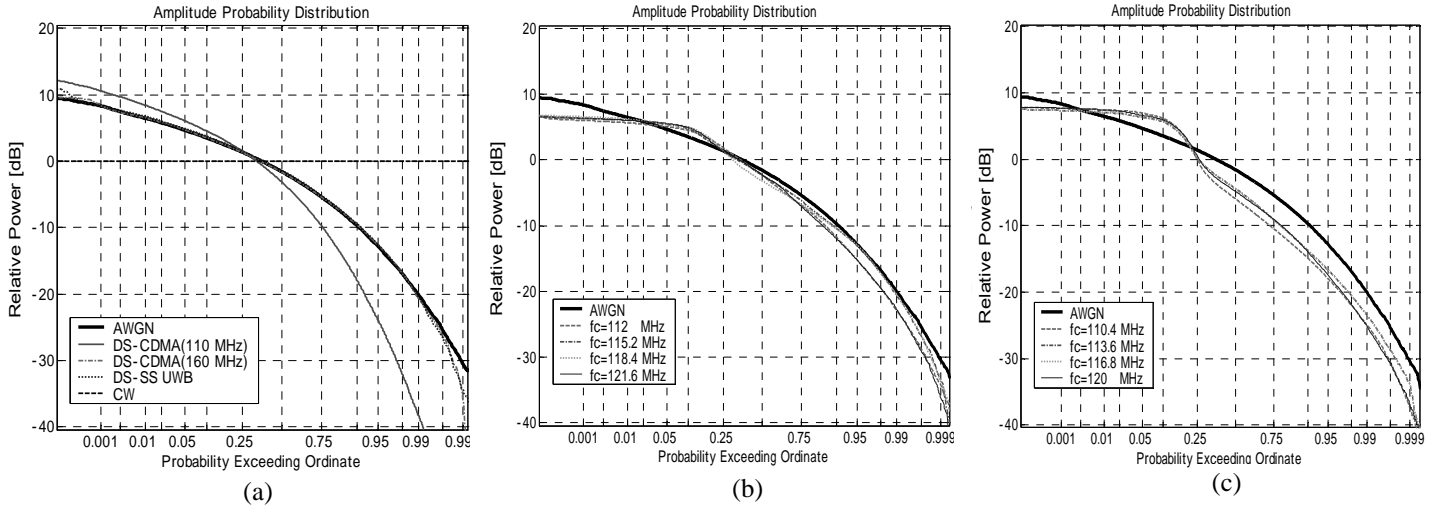


Fig. 5. The calculated APD of: (a) DS-CDMA and DS-SS UWB, (b) Example of peak frequencies of the MB-OFDM spectrum, and (c) Example of valley frequencies of the MB-OFDM spectrum.

The DS-CDMA power spectral density is shown in Fig. 4(b). Compared to the conventional impulse radio signal [2], spectral comb lines at the pulse repetition frequencies (PRF) were suppressed, generating a nearly flat spectrum, due to the spectrum spreading techniques.

C. DS-SS UWB

We modeled a DS-SS UWB signal similar to the signal developed in our laboratory using actual circuits, described in [5]. Basically, the DS-SS UWB uses a 15-step PN sequence generator with a chip rate of 2 GHz to spread a stream of BPSK modulated data bits. We can see in Fig. 4(c) that the frequency spectrum yielded a 2 GHz of bandwidth. The signal had a spectrum much wider than the conventional version of the conventional DS-SS transmission system.

D. AWGN

The AWGN signal was implemented by generating Gaussian-distributed random sequence at a sampling rate of 2 GHz sampling frequency to obtain a wide spectrum signal. Figure 4(d) illustrates the frequency spectrum of the signal.

IV. STATISTICAL PROPERTIES OF THE UWB SIGNALS

The measurements of UWB signals' APD are critical for characterizing their interference effects to narrowband receivers. The APD describes a statistical property of a signal's amplitudes, which may be used to predict the particular signal's behavior in the victim receiver. Figure 5(a) depicts the APD of each UWB source entering the victim's receiver, possessing 300-kHz bandwidth. The magnitude of each signal was calculated keeping the average power at 0 dB. A CW signal's APD was also plotted for reference.

The flat solid line represents the AWGN. The DS-SS UWB and DS-CDMA at 160 MHz yielded almost the same statistics as AWGN, as shown in Fig. 5(a). This is attributable to the randomized pseudo-noise sequence used for spectrum spreading. However, DS-CDMA at the PRF frequency (110 MHz), showed non-Gaussian APD. On the other hand, the MB-OFDM appears to have non-Gaussian APD, where the signal amplitudes exceeded the AWGN curve at 0.75% probability and above.

Since the spectrum of the MB-OFDM marks peaks at every 3.2 MHz interval, it is necessary to investigate the APD when the center frequency of the receiver is tuned to the peak or valley frequency. As shown in Fig. 5(b), we can see that all peak frequencies yield nearly the same statistical characteristics as their APD curves overlay with each other. Similarly, Fig. 5(c) shows that the APD of the valley frequencies also matched with each other. Thus, we can conclude that the interference effects from peak frequencies should be identical at any 3.2 MHz intervals. This goes the same for the valley frequencies case. Comparing both statistics, the valley frequencies marked slightly larger amplitudes at high probability.

V. SIMULATION RESULTS AND ANALYSIS

Results from the computer simulations are represented in Figure 6(a) to (f). The average BER was calculated while varying the D/U from 5 dB to 15 dB. It was found that the BER degrades from the theoretical value in every case. For most cases, the BER degrades significantly when the D/U is around 5 dB, producing floor characteristics of the BER. This means that the UWB signals cause immense interferences when the power is relatively high.

Figures 6(a) and (b) depict the BER performance of the AWGN and the DS-SS UWB. It can be seen that the interference effects of DS-SS UWB are similar to that of an AWGN of the same power. This corresponds to its APD, which closely resembles a Gaussian distribution. The BER performance of the DS-CDMA's case is represented in

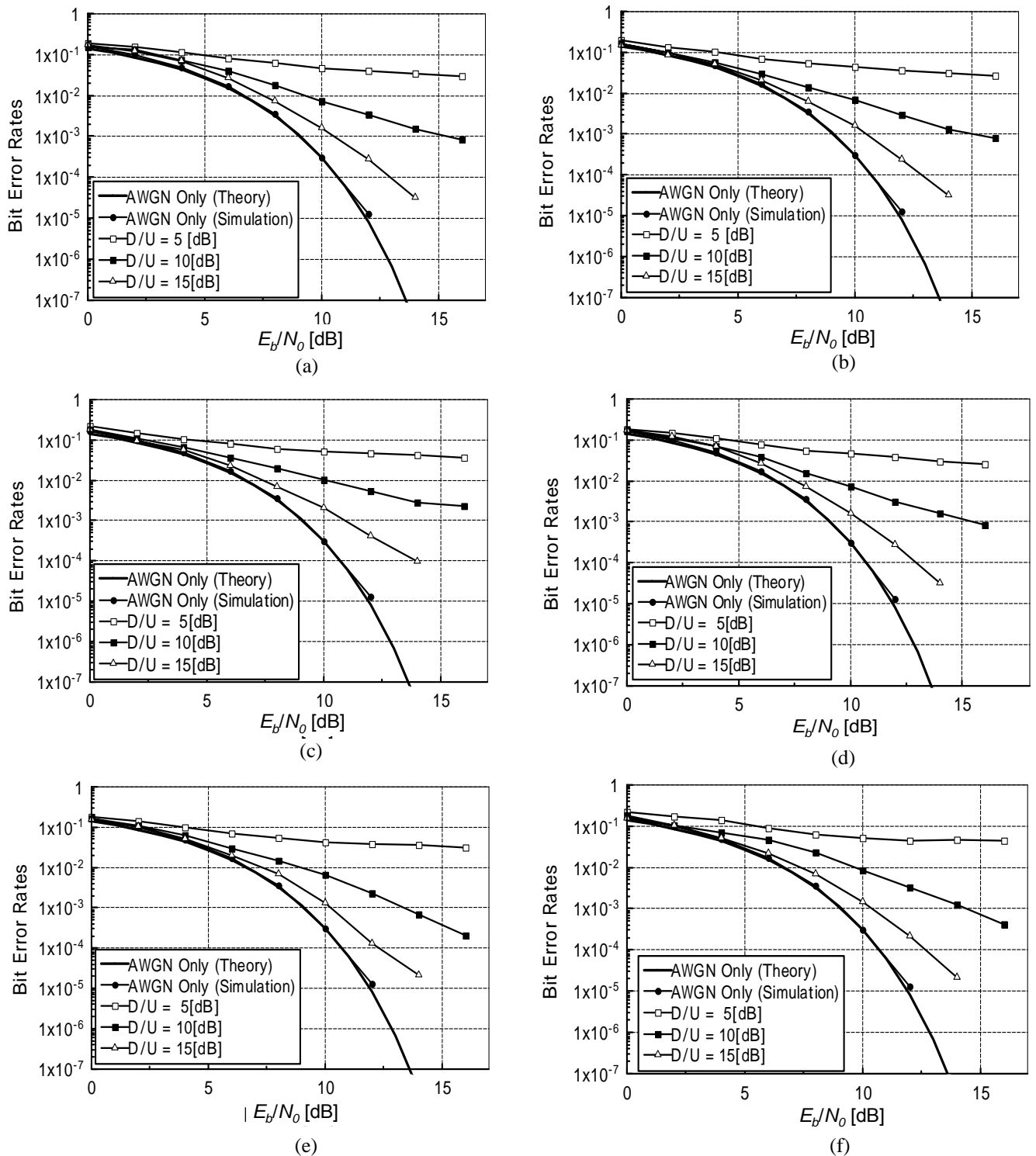


Fig. 6. The BER performance of the victim's receiver: (a) AWGN, (b) DS-SS UWB, (c) DS-CDMA (110 MHz from center frequency), (d) DS-CDMA (160MHz from center frequency), (e) MB-OFDM (3.2 MHz from center frequency), and (f) MB-OFDM (4.6 MHz from center frequency).

Figs. 6(b) to (c), where the center frequency of the victim system is tuned to a PRF frequency (110 MHz) and another frequency (160 MHz), respectively. The interference effects at the PRF frequencies are 3 dB worse than the AWGN in the $D/U \approx 15$ dB region, according to Fig. 6(c). At 160 MHz, the BER curve was identical to the AWGN one.

To summarize, the interference effects from the UWB signals to DQPSK transmission system approximated that of an AWGN, with slight difference in certain cases. We, therefore, concluded that the differences in BER performance are attributable to the non-Gaussian characteristics of the particular interference sources. In this evaluation, the interference level was verified as a function

of D/U , where the U was calculated as the power of the UWB signal occupying the same bandwidth— not the average total power. However, in practice, if it is assumed that the interference source is positioned at a certain distance from the receiver’s antenna, the interference signal’s power is defined by the average total power. Thus, the 20 dB spectral peaks at every 3.2 MHz intervals of the MB-OFDM will significantly degrade the BER, if tuned to the receiver’s center frequency. The calculated BER performance of the DS-SS UWB case agrees considerably with the results shown in [2], where actual measurements of the average bit error rates were done using an ideal software receiver. This shows that the results of this simulation are accurate.

VI. CONCLUSIONS

The interference effects from 4 types of UWB signals to $\pi/4$ -shift DQPSK transmission system were evaluated by simulation. It can be concluded that the interference effect from UWB signals varies according to their statistical characteristics entering the victim receiver. Moreover, the MB-OFDM showed spectral peaks at every 3.2 MHz that would degrade the BER further. We have also investigated the statistical characteristics of UWB by calculating their APD. The UWB signals characteristics depend upon the frequency.

Our future work will include the evaluation of the interference effects of UWB sources on a $\pi/4$ -shift DQPSK transmission system under a multi-path environment, and also simulations of interference effects to other narrowband digital transmission systems.

ACKNOWLEDGMENTS

This study has been in part funded by the Strategic Information and Communications R&D Promotion Scheme under the Ministry of Telecommunication of Japan and in part of the Research Institute of Science and Technology (grant no. 2Q240), Tokyo Denki University. The authors would also like to express special thanks to CoWare Co. for providing the simulation tool for this study.

REFERENCES

- [1] R. Tesi, M. Condreanu, and I. Opperman, “Inteferece effects of UWB transmission in OFDM communication systems,” in *Int. Workshop on Ultra Wide Band Systems*, Oulu, Finland, June 2003.
- [2] A. Tomiki, T. Ogawa, A. Fukuda, N. Terada., and T. Kobayashi, “Evaluation of interference from impulse-radio and direct-sequence-UWB sources to 2-GHz digital radio transmission,” in *IEEE Internat. Symp. on Electromag. Compat.*, TH-P-II.4, Istanbul, Turkey, May 2003.
- [3] A. Batra *et al.*, “Texas Instruments et. al., IEEE 802.15.3aUpdated MB-OFDM Proposal Specification (03/268r3),” Nov. 2003
- [4] M. Wellborn, “DS-UWB Physical Layer Submission to 802.15 Task Group 3a (04/137r),” Mar. 2004.
- [5] A. Tomiki, T. Ogawa, and T. Kobayashi, “Experimental Evaluation of Interference from UWB Sources to a 5-GHz Narrowband Digital Wireless Transmission System,” in *IEEE Conf. On Ultra Wideband Systems and Technologies*, Reston, VA, USA, Nov 2003.

Flexible Spectrum Use Rights Tutorial - ISART 2005

Robert J. Matheson,
National Telecommunications and Information Administration
Institute for Telecommunication Sciences¹
Matheson@its.bldrdoc.gov
303-497-3293, fax: 303-497-3680

Abstract: Although “command and control” spectrum management techniques have provided licenses for many specific services since the early days of radio, such licensing may not easily permit new technologies and new services. This paper describes the necessary principles of flexible use spectrum rights, which may allow a wide variety of spectrum uses in a single general-purpose band. Based on the electrospace description of the radio spectrum, these principles allow general aggregation or division of licensed electrospace regions via secondary markets, providing rules for how regulatory limits change under aggregation or division. These flexible-use principles limit transmitter behaviors that tend to create a more difficult operating environment for receivers, while making receivers responsible for handling any remaining interference. Flexible-use principles could provide a basis for real-world flexible-use frequency bands.

1. Introduction

The process of spectrum management determines what users and services can be provided at certain frequencies. In the past (and most of the present also) spectrum managers have assigned radio licenses that tightly prescribe exactly what frequency, bandwidth, modulation, transmitter power, geographical location, services, and type of user can be active on a specific frequency. Since a central authority assigned a specific use at each frequency, following a pre-engineered formula for that particular service, radio users could be squeezed together in a band just tightly enough for maximum efficiency, but not too tightly to cause interference. This method of spectrum management – called “command and control” – seemed to be the best way to simultaneously ensure high efficiency and freedom from interference.

The command and control technique has worked for many years, but it also has some major problems. The most obvious problem is in smoothly accommodating new services or new technologies, for which no pre-engineered formulas and associated frequency bands have been established yet. How can a new service, “A,” be offered, when there are no rules and frequencies at which this new service can operate? Equally frustrating, other services, “B” and “C,” may have less demand than anticipated, and their corresponding earmarked frequency bands will lie fallow for lack of users.

Keeping in mind the lack of application flexibility as the major problem of command and control regulations, one might ask if there could be a more flexible way to license

radio spectrum for a broad range of technologies and applications. This paper describes a set of rules and principles – called “flexible use spectrum rights” – that would allow the spectrum to be used for a wide range of user-selected services, as well as to be freely traded, aggregated, and divided via a secondary market. Throughout the remainder of this paper, the specific term “electrospace” will be used in place of “spectrum” to eliminate confusion with the more traditional use of the word “spectrum,” which usually refers to only the “frequency” dimension of the electrospace. The electro-space will be described in more detail in the next section.

The proposed flexible-use spectrum rights environment includes the following major features:

1. All signals must remain within their respective licensed electrospace region.
2. There are scalable limits on transmitter power or field intensity at ground level.
3. Receivers are unregulated, without any guarantee of freedom from interference.
4. Aggregation or division along any electrospace dimension can occur via secondary markets.

Definitions:

1. A licensed region of electrospace is a hyper-space volume described by dimensions of frequency, location, time, and angle-of-arrival.
2. A “signal” is defined to be present wherever spectral energy flux density is larger than X Watts per MHz per square-meter ($W/MHz/m^2$).

¹The views, opinions and/or findings contained in this report do not represent an official position of the National Telecommunications and Information Administration or of the U.S. Department of Commerce.

Although the flexible-use environment described here may be especially suitable for rapidly changing radio services requiring substantial infrastructure investment, it might have some disadvantages compared to traditional environments. In particular, since a flexible-use band might have a wide variety of possible services, a receiver might need to be capable of rejecting interference from a wide variety of different types of signals. This might require a more expensive receiver design, compared to receivers in traditional bands, which typically would need to reject interference from a much smaller variety of signal types.

It remains likely that the traditional regulatory environments will continue to be especially useful for some services. In particular, low-power non-licensed (Part 15) devices and spectrum for large government (especially defense) systems do not seem likely to benefit from flexible-use principles. Traditional command-and-control regulations will surely remain for many years, while low-power and “commons” environments will probably grow. Fortunately, there seem to be few impediments to establishing different regulatory environments in different frequency bands, so we could possibly select multiple sets of regulatory features, as needed to best serve various types of users and services. Therefore, this description of flexible-use spectrum regulations should be understood as applying only to (currently non-existent) flexible-use frequency bands, without prejudicing in any way the regulatory practices that apply to other bands.

2. The Electrospace as Property

This section shows how the electrospace reasonably fits into the category of “property,” so that a set of normal property transactions – including the unrestricted ability to buy or sell via a secondary market – can be established for the electrospace. It must be understood that the issues we will discuss here concern only the rights and obligations of the current holder of electrospace property, not whether the user holds the electrospace property permanently or temporarily. We describe the rights and obligations of the current property holder to use the electrospace – no matter whether the holder has a long-term lease (e.g., based on a 10-year license from the FCC), a short-term rental agreement (e.g., a temporary 1-week rental from the primary license-holder), or permanent ownership (via a permanent title transfer from the FCC or a future secondary market).

2.1 A Description of the Electrospace

The electrospace is a formalized description of the radio signal environment, as it might be seen by a hypothetical

ideal measurement receiver [1]. It applies to all types of radio systems and all regulatory environments. It can also be used to describe ways in which the radio environment can be shared among multiple radio systems. It is particularly useful in flexible-use environments, where it provides a straightforward basis for unambiguously describing licensed electrospace regions, as well as for aggregating and dividing electrospace regions.

The electrospace describes radio signals, which means that it describes the domain of transmitters and transmission paths. Any description of real receiver characteristics is totally separate from the electrospace description. Although any real radio system must consider all system components – i.e., the electrospace and the receiver – it is appropriate to divide these two components for regulatory purposes. The crucial regulatory difference between the two components is that the electrospace describes the ability of radio signals to cause interference to other users – which involves an externalized cost that must be regulated. The receiver domain, however, includes only components that do not cause interference to others. Therefore, the receiver domain has no associated externalized costs that need to be controlled by regulations; it can operate completely free of regulation.

An “ideal” electrospace model is based on the radio universe as seen from the viewpoint of an ideal receiver. In Section 3.3, various assumptions about the limitations of real receivers to reject unwanted signals will be used to modify electrospace management rules, giving interference rights that are more appropriate and efficient for real-world use.

The electrospace describes the radio field strength at a given electrospace “location” that is defined by the 7 electrospace dimensions. These 7 dimensions are all independent of each other, which means that the electrospace can be considered to be a 7-dimensional hyperspace. A “location” in the electrospace can be described by assigning specific values to several independent variables. It should be noted that different investigators have sometimes included other sets of variables in the electrospace. The set shown in Table 1 is a useful starting point, and probably no great harm is done by including or omitting some marginal variables such as polarization and modulation.

The physical location of a test point or hypothetical receiver is defined by the three spatial dimensions. The field strength characteristics at that location are described by the remaining variables, including the frequency, time of occurrence, and angle-of-arrival. In a frequency band whose licensing is based on the electrospace, a numerical limit will typically be established, such that field strengths in excess of X are considered to be signals, which are not permitted outside of the user’s licensed regions of the

electrospace. An electrospace “region” consists of all points within a described 7-dimension hyperspace volume. An electrospace region is typically used to denote the hyperspace volume defined by an electrospace license (e.g., the licensed electrospace region) or the hyperspace volume occupied by a signal (e.g., the region where field strength is greater than X).

Table 1 - Electrospace Dimensions

Quantity	Units	# of dimensions
Frequency	kHz, MHz, or GHz	1
Time	seconds, hours, or years	1
Spatial location (geography)	latitude, longitude, altitude	3
Angle-of-arrival	azimuth, elevation angle	2

One characteristic of the electrospace is that an ideal receiver can theoretically separate any radio signals that differ by at least one of their seven electrospace dimensions. For example, two co-located radio receivers could function without interference if the signals were at different frequencies, or if the signals occurred at different times, or if the signals came from different directions. Radio signals using the same frequency, operating time, and angle-of-arrival could be separated without interference if the receivers were present at different locations.

2.2 Comments on Electrospace Dimensions

Frequency. The frequency dimension of the electrospace has the standard meanings of the word, namely a description of the frequency or range of frequencies (bandwidth) at which field strength is being characterized. Frequencies can be divided over a wide range of increments, typically matching the channelization of particular services.

Time. The time dimension can be subdivided over a wide range of increments. Useful time divisions might include the several-year-duration of a licence, an agreement to allow a particular user to transmit regularly during the midnight-to-5 AM time block (when bandwidth would be inexpensively available to update computer files for the following day), or a one-time use during a 4-hour special events broadcast. On a much smaller time scale, a user could use a particular time slot on a TDMA system, to transmit during a 2.5-ms time slot that would be available

once every 20 ms, or transmit data during the vertical blanking interval of an NTSC television signal 30 times every second.

Spatial location. The spatial dimensions represent a physical (geographical) location. They can be problematic, because there is no practical way to confine radio signals within a desired region. In typical hilly terrain, there are many distant locations that have higher signal amplitudes than many closer locations. Therefore, although one might easily select an arbitrary spatial region, the selected region might be extremely inconvenient to use efficiently. In order to prevent excessive signal levels (larger than “X”) outside the boundaries of the selected spatial region, it might be necessary to greatly diminish signal amplitudes at many otherwise-useful locations within the spatial boundaries. Transmitter power, details of the terrain, and the use of directional transmitting antennas are operative in establishing the spatial boundaries of the electrospace associated with a given transmitter.

Angle-of-arrival. This factor describes the angle-of-arrival or direction of radio signals at a given location, including the possible effect of multipath components scattered from many objects in many different directions from the receiver location. Note that this factor is not created by physical antenna pointing angles. The pointing direction of transmitting antennas primarily affects the spatial dimensions of the occupied electrospace, i.e., the geographical areas where signals are larger than X. No aspect of physical receivers – including the pointing angle of receiving antennas – ever has any effect on the electrospace. Therefore, neither transmitting nor receiving antennas influence the angle-of-arrival factor. On the other hand, receivers that exploit the angle-of-arrival dimension will often employ directional antennas. Recently developed “multiple-input, multiple-output” (MIMO) technology exploits multipath reflections coming from different directions, handled by multiple transmitting and receiving antennas and mathematically processed to generate independent transmission channels. MIMO technology can be considered to be a generalization of the angle-of-arrival dimension of the electrospace.

3. Flexible-use Rights in the Electrospace

3.1 Use of the Electrospace to Divide the Radio Environment

The electrospace model can be directly applied to a flexible-use, market-based frequency management environment, since the model describes a way that the use of the radio spectrum can be unambiguously divided (shared) among multiple users. The only significant regulatory principle is that a licensee has the right to

radiate a signal within a licensed electrospacetime region. Outside the licensed region, signals must be kept below a specified very low spectral power flux density limit, X , in Watts/m²/MHz. X includes power coming from all directions, though in many cases the great majority of total power will arrive from the direction of the respective transmitter location. Note that X is a power that is proportional to bandwidth. There are no restrictions on type of service, transmitter power, bandwidth, modulation, antenna height, number of sites, etc., as long as the signal is kept lower than X at all points (including all dimensions of time, frequency, location, and direction-of-arrival) outside the licensed electrospacetime boundaries.

Ideal Electrospacetime Rules
(assumes ideal receivers)

- 1. Transmit without any restrictions inside your licensed electrospacetime region.**
- 2. Keep your signals below X outside licensed electrospacetime region.**

As described in greater detail in Section 4, an electrospacetime region is permitted unlimited aggregation or subdivision along all of its dimensions. This allows electrospacetime to be freely repackaged and resold as a market-based commodity, redistributing spectrum without requiring prior approval by a regulator. A 1-MHz bandwidth could be subdivided into 40 channels of 25 kHz each or augmented with 4 MHz of additional adjacent frequencies to make a single 5-MHz bandwidth. A given channel could be subdivided into TDMA time slots of 10 ms occurring once every second and rented to a hundred separate transmitters. A statewide geographic coverage area could be divided into much smaller geographical cells and rented to short-range neighborhood wireless ISPs. Multiple fixed transmitters could be allowed to radiate signals into a common receiver location, if the transmitters are arranged to provide signals that have different angles-of-arrival.

Although the electrospacetime model is critically based on a specified spectral power flux density limit, X W/m²/MHz, which cannot be exceeded outside the licensed region, it is not obvious what numeric value to choose for X . Presumably X will be chosen so that systems licensed outside the region will not receive interference from the signal. However, the minimum level of interfering signal for various types of systems varies over a wide range – perhaps 50-60 dB – depending on the system. Since all types of systems are assumed to operate in a flexible-use band, which type of system should X protect? One answer is that the selection of a specific value for X might

be done to make that band particularly suitable or unsuitable for various types of services; multiple bands could use different values of X to efficiently accommodate various services.

3.2 Practical Limitations on the Electrospacetime

Although the electrospacetime model is conceptually powerful and potentially very useful, there are a few important problems with its application to the real world. One major problem, non-ideal receivers, will be discussed in the next section. Other problems are discussed in this section.

The division of the electrospacetime along any selected dimensions – while theoretically possible – may or may not produce a useful division in the real world. Arbitrary spatial regions, for example, may not match easily achievable propagation/coverage areas. A more useful spatial division technique may be to use propagation models to determine easily achievable coverage areas and divide the electrospacetime regions in a corresponding way. The angle-of-arrival dimensions may be compromised by unintended scattering from the terrain or by lack of sufficiently-narrow-beamwidth receiving antenna performance (especially at lower frequencies). Division into very narrow time slots may produce systems that are difficult to synchronize properly. Division into very narrow frequency slots may produce unreasonable requirements for frequency stability and Doppler shift.

The spatial dimensions pose some other problems. The field strength at a particular location is often the vector sum of many multipath signals. These multiple signals can occasionally add up to a field strength that is larger than the average field strength in the general vicinity. Therefore, it may be desirable that the field strength limit contain a statistical parameter, which would allow the occasional presence of signals above the limit. However, the inclusion of a statistical limit might make it much more difficult to show that a user had violated the electrospacetime limits, since a single instance of excess field strength might not be sufficient proof of a violation.

One obvious application of spatial coordinates is to describe licensed regions using some imaginary lines drawn on the ground – e.g., lines described by latitudes and longitudes, a circle centered on a designated location, geographical boundaries, political boundaries, etc. For many applications, radio signals will be attenuated by buildings, terrain, and the earth's curvature, which all tend to give the greatest attenuation at ground level. Raising a receiving antenna farther above ground will usually increase the received signal level. Therefore, a transmitted signal that is below X at ground level will often increase greatly at higher elevations above ground.

Many radio systems have receiving antennas located on tall buildings, towers, or mountaintops. Therefore, a simple electrospatial boundary at ground level may describe only part of the real world; the success of a given application may depend on a much more complex understanding of how field strength changes with all three of the spatial electrospatial dimensions – possibly with much more complex 3-D descriptions of the associated electrospatial regions.

The frequency dimension can also cause problems. Although a transmitter can radiate any amount of power inside the licensed frequency range, the signal strength outside the licensed band must be less than X . Presumably this condition must be met at all locations – even very close to the transmitting antenna, where the in-band field strength is very high. To meet the “ X ” condition near a transmitter may involve a very high signal level inside the licensed region to drop below X immediately outside the licensed region (bandwidth) – requiring a very rapid decrease in signal strength over a small change in frequency. Therefore, the out-of-region absolute limit, X , may need to be supplemented by establishing an optional “relative-dB” emission mask that provides a legal “safe harbor.” The relative-dB emission mask would allow higher out-of-region levels in locations (e.g., near transmitters) where the field strength is very high. However, in areas where the in-region field strength was already low, a relative-dB emission mask would require out-of-region levels lower than X . Therefore, although the relative-dB emission mask could replace the absolute X criterion at any location. In practice the relative-dB emission would be invoked only near strong transmitters.

3.3 Receiver Regulatory Theory

The most serious limitation on the practical application of the electrospatial model to flexible-use spectrum management is that the electrospatial model assumes that all receivers are “ideal.” In this context, “ideal” means that the receiver has infinite rejection of unwanted frequencies (i.e., signal power at frequencies outside of the nominal receiver bandpass), infinite dynamic range (strong out-of-band signals will not cause intermodulation products or gain compression), and directional receiving antennas that have infinite rejection of signals coming from unwanted directions. At the end of Section 2.2, we stated that an ideal receiver could theoretically separate any two signals that were different in at least one of their electrospatial coordinates. Some electrospatial dimensions are easier to separate than others; even a simple non-ideal receiver could separate signals that were present in substantially different locations, frequencies, or times. However, an ideal receiver could even separate signals occurring at the same location, frequency, and time – as long as the signals came from different angles-of-arrival (separated through

the use of directional receiving antennas).

Two non-identical signals (i.e., two signals with different electrospatial coordinates) can always be separated and received without interference if the receiver is good enough, including receiving antennas as part of the receiver. This means that all interference is always caused because the receiver is not good enough. There is no theoretical line separating cases where interference is caused by poor receiver performance from cases where interference is caused by an “actual” interfering signal. Although an inadequate receiver is *always* the cause of interference, the required “adequate” receiver might be extraordinarily complex and expensive, and it might not actually be achievable with today’s technology. For example, an adequate receiver might require an elaborate adaptive antenna array to null out unwanted signals, while generating a high-gain receive beam in the direction of the desired signal. Such technology would be quite difficult today for even large fixed base stations; it would surely be completely impossible today for handheld portable radios. But tomorrow ... who knows?

If all receivers were ideal receivers, we would only need to worry about foreign signals that illegally intruded within our licensed electrospatial region to appear at the frequency of our desired signal – so-called “in-band” interfering signals. In-band interference is controlled chiefly through the electrospatial parameter “ X ,” which sets a limit on the level of signal that can be present outside its licensed electrospatial region. If all receivers were ideal receivers, the electrospatial rules that control “ X ” would be all that is needed to control interference. Unfortunately, none of the receivers that are available to users at reasonable prices are ideal receivers. Even worse, the most popular and rapidly growing class of receivers – handheld, multi-band cellphones – are especially non-ideal, with performance constrained by small size, low cost, and limited battery power. An important characteristic of real (i.e., non-ideal) receivers is that they can generate interference even when no unwanted signal is actually present at the tuned receiver frequency. Strong signals at close-in frequencies or very strong signals at frequencies further away from the tuned frequency can also cause receiver distortions that are seen as interference; this is known as “out-of-band” interference.

Fortunately, real radio systems do not usually require ideal receivers for satisfactory operation. Instead, they merely require “good-enough” receivers. A “good-enough” receiver is a receiver whose performance is at least good enough to achieve the desired system performance in the actual radio signal environment. The required level of performance for a receiver that is good enough to reject unwanted signals without experiencing interference will vary greatly, depending on the specific characteristics of the electrospatial environment in which

the receiver is operating.

The overall regulatory strategy is to supplement the electrospacetime rules to produce a more benign signal environment that allows the successful operation of less-expensive “good-enough” receivers. Hopefully, this will allow an improvement of the overall cost/benefits that can be achieved from operating radio systems. The various supplemental rules that are selected for each band (some of which are described in the following paragraphs) should be selected on a principle of maximizing overall benefits – balancing the benefits from less-expensive “good-enough” receivers with the disadvantages of adding some restrictions on transmitter characteristics. Presumably, different rules could be selected for different bands, since this will differentially maximize the benefits for various types of systems that could be built in each band. Under these supplemented rules, the important principles that regulate interference are now:

Practical Electrospacetime Rules
(assumes non-ideal receivers)

- 1. Transmit within power restrictions inside your licensed electrospacetime region.**
- 2. Keep your signals below X outside licensed electrospacetime region.**

Transmitters must still follow the electrospacetime rules, including the supplemental rules that make a more benign environment for receivers. As before, receivers are not regulated in any manner. They are allowed to be as-good-as or as-poor-as their owners permit. There is no implied protection against interference, except that there is the expectation that the radio environment will probably allow the use of cheaper receivers. The aforementioned limitations on transmitter power will be discussed in more detail in section 3.4, and the way that these limitations scale when electrospacetime regions are aggregated or divided will be discussed in section 4.

A major advantage of these principles (compared to the current command-and-control rules, which tend to try to guarantee interference-free performance) is that there is much less legal ambiguity about who is responsible for fixing interference situations. Assuming that transmitters obey the supplemented electrospacetime rules, the receiver owner is completely responsible for solving his own interference problems. There is never any assumption that a transmitter operating within these rules has any further obligations to prevent interference to any receiver. An exception to this general rule would apply to tightly-grouped transmitters and receivers, where site managers

could have the authority to adjust radio systems to reduce interference. Note that this situation is a well-known “exceptional” case in traditional frequency management, also.

In an interference situation, the receiver owner has several basic options to deal with the problem:

- Show that a specific transmitter is violating one of the applicable supplemented electrospacetime rules, and require that the offending transmitter change its operation to become compliant.
- Improve his own system, as needed, to eliminate the interference. Depending on the exact cause of the interference, the changes might involve improving the performance of the victim receiver, increasing desired transmitter power (if permitted under electrospacetime rules), adding better error correction, etc.
- Figure out how to tolerate the interference. This might involve changing operating procedures, restricting the operation to areas where interference is not a problem, ignoring the issue, issuing the customer a partial refund, etc.
- Negotiate with the interferer. This is a strictly voluntary negotiation for both parties. After investigation of the possible alternatives in (a) – (c), it might turn out that an adjustment of the interfering transmitter would be the best way to solve the problem. If so, negotiations between the parties might result in an appropriate mutually voluntary business arrangement that could become a legal attachment to the respective electrospacetime licenses.

The opportunity to select an appropriate receiver remains completely with the receiver owner. The supplemented electrospacetime rules help to define the statistics of the expected unwanted signal environment in which a receiver must operate. The receiver owner has complete flexibility to select whatever level of receiver performance that he has judged to be adequate to accomplish the mission of his radio system. One would expect that there will be a wide variation of performance requirements among the population of operating radio systems, and the radio system owner has much better knowledge of the specific economic and operational requirements of his own mission than any federal regulator. Moreover, the radio system owner is more highly motivated than anyone else to make a correct decision about how to get that required performance. Finally, the owner’s selection of receiver performance does not cause any additional interference to any other radio system. Therefore, this set of decisions can be left completely in the hands of the radio system owner.

3.4 A Limit on Power or Maximum Field Strength

The original electrospace rules control external signals at the receiver operating frequency (in-band interference) by requiring that they are always less than X outside of their own electrospace region. The supplemented electrospace rules are needed to control the presence of (legal) strong signals at frequencies outside the receiver tuned frequency (out-of-band interference). Strong transmitters that cause high signal levels within the relatively wideband first RF stages in a practical receiver are usually the major cause of out-of-band interference. Therefore, supplementing the electrospace rules by placing a limit on transmitter power (or EIRP) is one obvious approach to controlling the occurrence of strong signals in the radio environment and reducing out-of-band interference for practical receivers. Note that limiting transmitter power would not be expected to eliminate all out-of-band interference. However, out-of-band interference would tend to be limited to a much smaller set of circumstances where the victim receiver is located very close to a transmitter tuned to a nearby frequency. Therefore, the use of supplemented electrospace rules will tend to allow interference-free operation in more locations using cheaper receivers.

It should be noted, however, that transmitter power is not solely responsible for causing out-of-band interference in receivers; additional factors must also be present. Specifically, the direct cause of out-of-band interference to receivers is when receivers are subject to high-field-strength, out-of-band signals. The cause of the high-field-strength signal interference to receivers is the result of a combination of three factors:

1. High transmitter power,
2. Transmitter vertical antenna patterns that produce high-field-strength signals on the ground near the transmitter, and
3. The presence of susceptible receivers in the high field strength areas.

Presumably, the out-of-band interference could also be prevented if any suitable combination of these three factors could be arranged, including controlling transmitter power, controlling the transmitter antenna patterns underneath/nearby the transmitting antenna, or placing transmitting antennas in locations where receivers will only rarely be found in the nearby high field strength locations. Therefore, instead of controlling the interference only by limiting transmitter power, it would provide more user flexibility to also allow the control of interference by controlling transmitter antenna patterns, and/or by carefully separating transmitter sites from high concentrations of susceptible receivers.

Thus, a more effective supplementary rule to protect

receivers might include a limit on signal field strength at ground level instead of a limit on maximum transmitter power. This alternative rule would state that field strength at ground level must be less than E_{\max} , where E_{\max} corresponds to a maximum watts/m². Note that this limit is not bandwidth-dependent, since the total power at the receiver input is usually what causes the problems, and the receiver front-end circuits will tend to be much wider bandwidth than most transmitters. This limitation must be met only in areas where there is a likelihood that susceptible receivers will normally be found there. In some circumstances, it might be necessary to similarly protect additional not-at-ground-level outdoor locations where people are often found (e.g., elevated walkways, rooftop cafes on nearby buildings, etc.).

This maximum-field-strength rule would allow much more flexibility in building a wide variety of radio systems, and it would protect receivers better.² Although this supplemented electrospace rule would not limit maximum transmitter power, it would still ensure that receivers are protected from the high-level fields that can cause interference. A higher power transmitter will still need to stay below a fixed maximum field strength at ground level. Part of the “cost” of using a higher transmitter power is that the field strength at ground level will need to be suppressed relatively more, so that it still meets the E_{\max} field strength limit. In an economic sense, this rule would tend to ensure that the higher cost of using a more powerful transmitter is borne entirely by the transmitter owner, instead of being partly externalized to unrelated receiver owners.

In summary, two basic methods could be used to control out-of-band interference problems. A transmitter power limit (or EIRP, or equivalent) is the simpler rule to apply. This “EIRP” rule indirectly tends to control the high-field-strength locations that can cause interference in receivers. The alternative version of the rule directly establishes a maximum field strength, E_{\max} , and leaves the details up to the transmitter owner. Although the “ E_{\max} ” rule is more complex in application, it provides better protection to receivers and allows more freedom in designing transmitter systems. There is no reason to require that a single rule would need to be applied to all flexible-use bands. One rule could be applied to one band, the other rule to another band.

²Note that the interference to public safety LMR in the 800-MHz band was caused by allowing apparently reasonable changes in antenna locations, even without allowing any changes in transmitter power. This is an example where E_{\max} limits would have allowed greater flexibility in system architecture, while simultaneously providing better protection from interference.

There are some possible refinements to either version of this supplementary rule. Since the possibility of overload is actually caused by the total power into a receiver, the E_{\max} limit should be determined by the total equivalent power from all the fields from various transmitters in a given location. Thus, the rule will probably require an obvious adjustment in areas where multiple transmitters produce high field strengths. However, since most receiving antennas do not operate efficiently over a very wide frequency band, the total power counted at a location would include only transmitters within fairly close frequencies. The actual algorithm for computing the weighting of field strengths with frequency might change according to the typical receiver front-end or antenna technologies used in nearby frequency bands.

It will also probably prove useful to adjust the values of X (maximum signal leakage outside of licenced electrospace) and E_{\max} (maximum field strength) in various flexible-use bands to preferentially optimize their use for various types of service. Special consideration should be given to nearby bands that contain large numbers of receivers that might be particularly susceptible to strong signals, such as portable (cellular, PCS) transceivers. For operational simplicity, it may also be useful to include “safe harbor” rules, so that transmitters with EIRP below a certain power limit would be automatically assumed to meet the E_{\max} field strength rules.

Other limitations on flexible use might also be beneficial in certain frequency bands. For example, one large class of radio systems (including most LMR and cellular/PCS services) will benefit from frequency bands that are engineered into duplex band architectures, where base station receiver frequencies are systematically separated from base station transmitter frequencies. Therefore, although “maximum-flexibility-of-use” remains a key principle, some bands will benefit from a requirement that specific sets of frequencies can be used only for base station transmitters or only for mobile transmitters. Similar generic restrictions may prove useful for other flexible-use bands intended to efficiently support other types of services, though additional examples of such restrictions are not yet obvious.

The actual values of X will need to be determined according to the performance of receivers that operate within or nearby the various flexible-use bands. It seems reasonable to expect that the selection of different combinations of values for these parameters will create bands that have different “sweet spots” for systems of different bandwidths and services. It would seem useful to eventually allocate a variety of flexible-use bands having different “sweet spots,” which would be expected to differentially attract a varied mix of applications in each band.

Note that the value of the parameter E_{\max} is totally determined by current (and past) practical receiver technology; there is nothing theoretically binding about these values. If a future change in receiver technology causes the performance of receivers to change substantially, this numerical value should also be expected to change (presumably after sufficient discussion and rule-making). Over the years, receiver performance has occasionally changed dramatically. The development of the “superhet” receiver created a major improvement in receiver performance, including much better receiver selectivity. The recent development of the receiver-on-a-chip technologies have surely made receivers much smaller and cheaper, but not necessarily much better. Major changes in receiver performance may result from much smarter receivers (that figure out how to move to a better frequency or a better modulation), from receivers using digital RF or IF processing (where certain types of receiver distortions can be recognized and processed away), from room-temperature superconductors (producing very-narrow-band, very-high-Q, RF filters that could reject many of the signals that would cause out-of-band interference in today’s receivers), or from adaptive antenna technology (that nulls out many strong unwanted signals).

Possibly none or possibly all of these receiver changes will actually occur in the next few decades. Since there is a substantial possibility of change, however, it might be useful to figure out how to easily change the values of the operational parameters that regulate the use of flexible-use bands. This would allow the band “sweet spots” to track the changes in markets and technologies.

4. Freedom to Aggregate or Divide

An important feature of the flexible-use regulatory environment is freedom to aggregate or divide an electrospace region along any or all of the 7 electrospace dimensions, presumably according to a secondary market and without the permission of a regulator. If this freedom is permitted, it will be necessary to define how the rules (including the values for X and E_{\max}) can be made to scale in a reasonable manner for the resulting new electrospace regions.

4.1 General Principles

The applicable principle here is that aggregation or division of an electrospace region should not expose electrospace neighbors to any greater threat of interference after a “transaction” than existed before the transaction. All allowable transactions must meet this general principle.

Only electrospace regions regulated by identical sets of

rules can be aggregated. Whenever electrospacetime regions are combined, any original regional borders that are now interior to the new region can be ignored. No limits that were associated with these “interior” borders need to be obeyed anymore. When a single party owns multiple adjacent electrospacetime regions, the owner can decide whether to consider the two regions as a single region or as multiple independent regions. In most cases, the common owner simply chooses not to enforce (against himself) the rules associated with excessive signal levels (signals greater than X) leaking across interior borders.

Whenever electrospacetime regions are divided, the new borders associated with the new regions must now meet all of the conditions associated with the borders of the original electrospacetime regions. All external boundaries maintain the same set of rules as before the transaction. No sets of internal changes can be construed to change the rules or values associated with external regions.

Any set of electrospacetime regions can be joined together. Similarly, a given electrospacetime region can be sub-divided into multiple new electrospacetime regions – essentially without any constraints or limits. However, the mere ability to identify and create a new electrospacetime region should not be understood to imply that the resulting electrospacetime region will necessarily be useful for any specific job. This limitation is particularly important when dividing or combining along the geographical dimensions, where natural terrain and buildings will tend to set limits – instead of being set by any arbitrary latitude/longitude boundaries.

4.2 Rules for Scaling X

The limit, X, for the amount of signal allowed outside of a licensed electrospacetime region is scaled in terms of $W/\text{MHz}/\text{m}^2$. When geographical areas are added or subtracted from a region, the change merely affects the geographical position of the boundaries outside of which the signal must be suppressed below X. Similar effects are applied to changes in the time and angle-of-arrival boundaries. When the frequency boundary is changed, the bandwidth of the signal leaking across geographical boundaries will presumably change with the bandwidth of the primary signal, but the value of X at any particular frequency will remain the same.

The only complication of aggregating or combining electrospacetime regions comes from a very basic understanding of what constitutes a “signal.” In particular, each independent signal source is allowed to leak a very small amount of power (up to “X”) at any or all electrospacetime locations outside the licensed region. If a single owner had ten base stations within a geographical region, each using the same frequency, those ten base

stations would be part of the same electrospacetime license, and cumulatively they could not leak more than X signal at any given frequency outside of the licensed electrospacetime region. If the single owner divided his electrospacetime region geographically into ten regions (one base station in each region), each of the base stations would be part of a different electrospacetime region and could presumably leak X outside of its own electrospacetime region. Potentially, this could represent a cumulative leakage of 10X from the ten independent stations.

Similarly, a region with a 10-MHz bandwidth might be filled with a single 10-MHz bandwidth signal that leaked X power at various frequencies outside the licensed frequency region. The single owner might divide the electrospacetime region into ten frequency regions, each having a 1-MHz bandwidth and containing a 1-MHz portion of the previously described 10-MHz signal. The new owner of the ten 1-MHz regions could claim that each 1-MHz region could individually radiate X energy to various other frequencies outside the ten regions, producing as much as 10X cumulative energy at any frequency outside the licensed regions (assuming that one could show that each 1-MHz region independently produced X energy at a certain outside frequency).

In each of these cases, the owner of a single region could apparently get permission to leak more signal outside his electrospacetime region by simply claiming that a single electrospacetime region (and signal) had been divided into multiple electrospacetime regions (and signals) – each “signal” with a separate allowance for X. Therefore, it may be necessary to understand that an independent “signal” must actually be independent of other signals in order to qualify for a separate right to radiate X outside the region. A COFDM signal could not be arbitrarily split into a thousand multiple carriers and corresponding electrospacetime regions, if they shared error-correction mechanisms and data between the various carriers. The entire COFDM signal would have to be treated as a single signal, entitled to leak no more than X into neighboring regions.

The exact rules by which a single signal is defined may be a little tricky to define exactly, since many independent signals might be used in a network of various load-sharing paths using different signals. Some other independent signals are very closely coordinated (e.g., synchronized spreading codes in multiple CDMA signals at a single base station). Are simulcast transmitters one signal or many signals? One could imagine future systems where multiple independent signals are combined to be amplified by a single broadband transmitter power amplifier and radiated from a single antenna, or where a single signal is split among multiple transmitting antennas in an adaptive array (after each antenna feed signal is adjusted for gain and phase, amplified, and maybe more?).

In summary, X does not need to scale in any way under aggregation or division of electrospacetime regions. The arbitrary division of a signal into separate pieces to acquire a separate allowance for X for each divided portion is not permitted.

4.3 Rules for Scaling EIRP and E_{\max}

This section describes two possible rules for scaling the power of transmitters under aggregation or division – one for the transmitter power/EIRP case and one for the E_{\max} case. It will be noted that these two cases develop somewhat different sets of rules.

EIRP/transmitter power model. In the case of an electrospacetime model that includes a maximum transmitter power (or EIRP) = Y, the actual definition is in terms of $Y = \text{Watts/MHz}$. If a wider bandwidth is divided to give two smaller bandwidths, each of the smaller bandwidths will have a maximum transmitter power proportional to the relative bandwidths of the new regions. Moreover, the maximum transmitter power of the original bandwidth will be equal to the sum of the maximum power of the two smaller bandwidths.

An additional refinement could be added to the above rule, based on typical receiver performance. The probability of interference from a strong unwanted signal is affected by the total power of the unwanted signal and also by how close the unwanted signal is to the frequency of the desired signal. There is more chance of interference if the strong unwanted signal is close to the frequency of the desired signal. Therefore, a further rule could be proposed, which states that the radiated energy should be spread out evenly across the licensed bandwidth, instead of being allowed to be concentrated at the edges of the licensed bandwidth. Otherwise, the entire extra transmitter power allowed by aggregating more bandwidth could be placed in a CW signal at one extreme edge of the bandwidth, creating a stronger signal immediately next to the frequency range used by an electrospacetime neighbor. Therefore, the proposed rule states that the cumulative radiated power measured from the edge of the licensed bandwidth to any point inside the licensed bandwidth cannot be more than twice the total power that would result if the allowable average power/EIRP were totaled over that same frequency range. The factor of two allows a wide range of modulations to be used without any derating of total transmitter power.

This scaling rule for EIRP is a very natural rule for scaling spectrum use rights, since the total transmitter power does not change when a given transmitter is divided into smaller bandwidths. However, such a rule is not totally effective in preventing interference to receivers. As a transmitter becomes wider in bandwidth (presumably, by

aggregating additional frequencies), its maximum power can increase (proportional to bandwidth), until the transmitter is possibly powerful enough to cause out-of-band interference in a nearby victim receiver. This is a problem that is quite similar to the current FCC rules concerning the situation where the number of separate transmitters at a given base station increases until a certain total power threshold is crossed. Eventually, there is sufficient total power radiated from the base station transmitters that they will cause out-of-band interference in nearby receivers (actually, the FCC maximum-power rules are intended to prevent health dangers to people). At that point, the combination of transmitters becomes equally responsible to control the problem, with any transmitter that contributes more than a certain percentage (e.g., 10%) of the total power being responsible to decrease transmitter power as required to cause the total power to drop below a certain threshold.

Similarly, a useful set of rules for the E_{\max} model would be to scale maximum transmitter power proportional to bandwidth under normal conditions of division and aggregation. However, once a certain maximum power threshold had been crossed, power would be limited on an absolute basis.

In terms of scaling power along other electrospacetime axes, the dimensions of time, space, or angle-of-arrival do not cause any difference in transmitter power scaling. Of course, extending the geographical area of a region may allow more powerful transmitters to be employed, simply because the new regional boundaries are further from the transmitter site, permitting more transmitter power without violating the leakage of signals above X outside the new boundary.

Although the installation of additional transmitter sites within a region can increase the total power radiated at a given frequency, this will generally not increase the risk of interference to other users. The probability of interference from out-of-band signals is primarily related to the presence of *strong* unwanted signals, not by the total area over which a weaker unwanted signal is available. Therefore, there is no reason to limit the total power radiated by multiple sites, as long as the total power radiated by a single site is controlled.

E_{\max} model. An alternative model for scaling transmitter power under aggregation and division is used to control the maximum field strength = $E_{\max} = V/m$ at ground level where receivers will be present. Since the occurrence of out-of-band interference is mostly related to the total amount of signal within the very-wide-bandwidth electronic circuits at the receiver front end, it should be assumed that all of the energy from any transmitter at any nearby frequency will be available to cause out-of-band interference. Therefore, the E_{\max} limit does not scale with

transmitter bandwidth, but remains tied to a certain maximum field strength. Presumably, the cumulative power from multiple transmitters should be included within this limit, with some rules for requiring compliance by any group of multiple transmitters that cumulatively violates the field strength limit. No other constraints are imposed to prevent out-of-band interference from other transmitters.

5. Summary

The preceding sections have described a possible spectrum management approach to flexible-use spectrum rights. This set of concepts could provide a market-based method that would give great flexibility in the use of spectrum, while controlling most interference. In the few circumstances where interference might result, the model includes clear rules for assigning responsibility to mitigate the interference.

This flexible-use model is believed to provide a highly flexible environment in which new or modified services can be rapidly provided by following a very small number of rules. It should be noted that there are still some areas of ambiguity, but some of these can be resolved with “safe harbor” practices or more detailed rules. Also note that many of the ambiguities refer to situations that are also substantial ambiguities under current command-and-control spectrum management practices. The flexible-use rules do not somehow make the radio world simpler than it now is, and even under flexible-use rules it will be necessary to occasionally make complex and difficult technical trade-offs. However, unlike with the current command and control management, the spectrum owner would be directly authorized to immediately make and implement these decisions, instead of waiting for an expensive and problematic federal regulatory process.

It is likely that a flexible-use environment will also have some disadvantages. What appears as “freedom” to one licensee might appear as a “lack of needed guidance and prescribed practices” to another licensee. A higher degree of technical expertise might be required to put a new system in a flexible-use band. The possible lack of

expertise in flexible-use system design might lead to higher levels of interference in a flexible-use band. The lack of narrow standards in the band might mean that a new system would have to be designed to withstand interference from a wide variety of possible interferers. This might require a more expensive system design than would be necessary in a more traditional band (where only one type of interferer would usually be present).

All of this suggests that a traditional single-service frequency band might remain the most suitable band for radio systems that comfortably fit there. However, it is also expected that technological obsolescence may fairly rapidly create situations where new applications no longer fit the existing band allocations. A major question will then be whether additional uses can be painlessly grafted into existing band allocations, using some of the principles of flexible-use rights described here, or whether some other conversion technique will be more useful.

Some of the remaining questions about possible flexible-use bands include:

1. What specific values for X , Y , E_{\max} should be chosen for a specific flexible-use band? Which sets of values would best match specific technologies or services?
2. What are appropriate characteristics for relative-dB safe-harbor emission masks?
3. What is a usable definition of a single signal? (to prevent multiple- X emission limits)
4. What is the best way to describe the geographical limits – especially re the vertical dimensions?
5. Are there any other holes in the model?
6. Would this model be too difficult to administer or enforce? Who would be responsible for enforcement?

6. References

1. R. J. Matheson, “The Electrospace Model as a Frequency Management Tool.” Addendum to the Proceedings of the 2003 ISART Conference, March 4-7, 2003. J. W. Allen and T. X. Brown, editors. NTIA Special Publication SP-03-401, March 2003.

Detection and Measurement of Radar Signals: A Tutorial

Frank Sanders
Institute for Telecommunication Sciences

303.497.5727; fax 303.497.3680
fsanders@its.blrdoc.gov

Abstract. *The wide use of radars for various functions makes significant demands on the electromagnetic spectrum. Effective measurement and monitoring of radar emissions is necessary to verify compliance with the legal emission limits specified in the Radar Spectrum Engineering Criteria (RSEC), as set forth by NTIA. Detection and measurement of radar signals is necessary to ensure an acceptable degree of electromagnetic compatibility among radar systems, and between such systems and those of other radio services in the frequency spectrum. This tutorial describes techniques for detecting and measuring radar emissions for compliance with the RSEC and other spectrum management purposes. Techniques for both conventional and advanced radar types are addressed.*

1. Introduction

This tutorial is a condensed version of an NTIA Report [1] that describes techniques for radiated radar emission measurements in great detail. It is meant to assist spectrum managers and engineers in detecting and measuring radar emissions and in using those measurements for two major purposes: to verify compliance with the NTIA Radar Spectrum Engineering Criteria (RSEC) [2, primarily Chapter 5], and to analyze radar signals for electromagnetic compatibility studies. This tutorial describes techniques for measuring the following radar emission parameters: pulse width, pulse repetition rate, antenna pattern, and spectrum.

It is emphasized that swept-frequency techniques (such as those implemented in spectrum analyzers) are not very efficient at finding or observing radar signals. Nor are high-speed time-domain digitizers especially effective. This is because radars are (usually) low duty cycle, pulsed emitters that scan narrowly directed beams through space. Therefore, the effective monitoring for radar signals requires that fixed-frequency, time-domain scans be used. Effective spectrum measurements use a variant of this technique, progressing in discrete frequency steps that last slightly longer than the scan (beam movement plus frequency-hop) interval of the radar being observed.

2. Pulse Parameter Measurements

Radar pulse parameters that may need to be measured include pulse and sub-pulse duration (t), pulse rise and fall times (t_r and t_f), number of sub-pulses in coded pulses (N), bandwidth of frequency deviation, (B_c) for FM-pulses (chirped) and (B_d) for FM-CW radars, and compression ratio of FM-pulses (d). For NTIA

measurements, pulse width, t , is defined at the 6-dB points (50% voltage points) of radar pulses. The rise time, t_r , or fall time, t_f , is measured between the 10%–90% voltage (-20 to -0.9 dB) points on a pulse's leading or trailing edge, respectively, as shown in Figure 2. For coded pulses, t_r and t_f are the rise and fall times of the sub-pulses. If sub-pulses are not discernable, then t_r is defined to be 40% of the time required to switch from one phase or chip to the next.

2.1 Non-FM (CW) Pulses

Figure 1 shows the setup for making radiated measurements of the radar pulses. The discrete-component detector and oscilloscope are used to measure radar pulse envelope characteristics. A vector signal analyzer can be substituted for the discrete detector and oscilloscope for measurements of pulse phase coding and frequency modulation (chirp) characteristics.

The detector output is connected to an oscilloscope. The oscilloscope's bandwidth should be wide enough to ensure that pulse rise/fall time can be measured accurately (wider than $1/t_{rise}$ and $1/t_{fall}$). Measurement personnel should be aware that some oscilloscopes achieve their widest bandwidth performance in repetitive sampling modes, but that radar pulses need to be measured in single-trigger modes. For many radar pulses, it is desirable that the single-trigger bandwidth be at least a few hundred megahertz. Impedances should be matched appropriately; most modern oscilloscopes have selectable input impedance values; 50 ohms and DC coupling are typically correct. The oscilloscope is adjusted to display and record pulse envelopes.

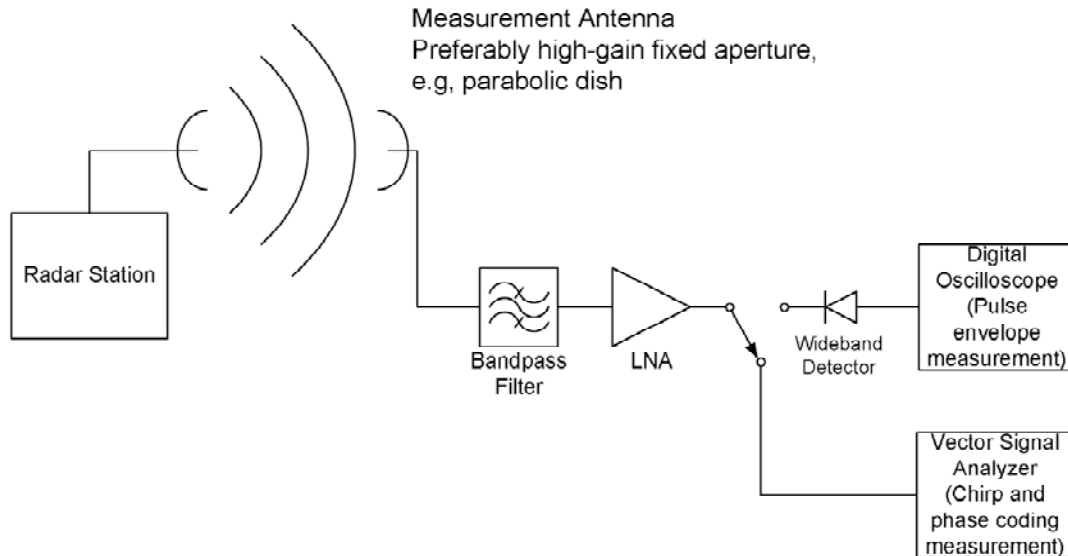


Figure 1. Block diagram schematic for measuring radiated waveform. If received pulses are sufficiently high in amplitude, then the LNA is not required. A spectrum analyzer may sometimes be substituted for the wideband detector and the oscilloscope, as described in the text.

A potential impediment to measuring pulse width via radiation is the effect of multipath energy, which typically causes distorted pulse envelope shapes. This effect can be minimized by the use of a narrow-beam receive parabolic reflector antenna on the measurement system. If multipath features are noted in the pulse envelope even when a parabolic antenna is being used, then slight adjustments in the vertical tilt angle of the antenna should be made, until the multipath features are minimized or eliminated.

2.2 Coded Pulses

There are two fundamentally different approaches to these measurements: envelope-detected and phase-response. Measurements of chip duration and rise time/fall time may be performed somewhat similarly to the procedure in section 2.1 above. However, measurement of these parameters is complicated with a wideband detector because although phase is shifted during each pulse (as a series of chips), only the power is observed at the output of a detector. This makes the edges of the chips unobservable, in principle. In application, transients may occur at the phase transitions between the chips, and these transients may be visible for some types of phase coding. When the transitions are observable, the chip width may be measured as the period between amplitude nulls in the transients. Chip rise/fall time may be taken to be the same as pulse rise/fall time, or else may be taken to be 40% of the time required to switch from one phase or

sub-phase to the next. Also, the number of chips may be estimated by counting the transients within the envelope (e.g., if 12 transients occur, then there are 13 chips in the pulse). Figure 2 shows a wideband detector measurement of a simple CW pulse; Figure 3 shows a similar measurement for a phase coded radar pulse.

For radar systems employing MSK or other phase-shifting technologies that eliminate transients between chips, it is impossible to determine the number of chips, their durations and their rise/fall times by measuring the detected pulse envelope. For phase-coded pulses, the chip duration and rise/fall time may be measured directly only if the waveform is sampled without envelope detection. A vector signal analyzer (VSA) can be used for this purpose. Current VSA technology does not always allow direct measurement of RF energy above about 6 GHz. If radar frequencies are too high for direct measurement with a VSA, then a spectrum analyzer may be used to downconvert the RF pulse energy to an IF frequency that can be fed to a VSA.

For FM-pulse radars, an additional pulse parameter might need to be measured, the bandwidth of the frequency deviation (chirp) (B_c). This parameter can be measured with a different instrument, a modulation analyzer, if it can operate at the RF frequency of the radar. Alternatively, this parameter can be measured with a vector signal analyzer with an operational setup as shown in Figure 1. An example measurement of B_c made with a VSA is shown in Figure 4.

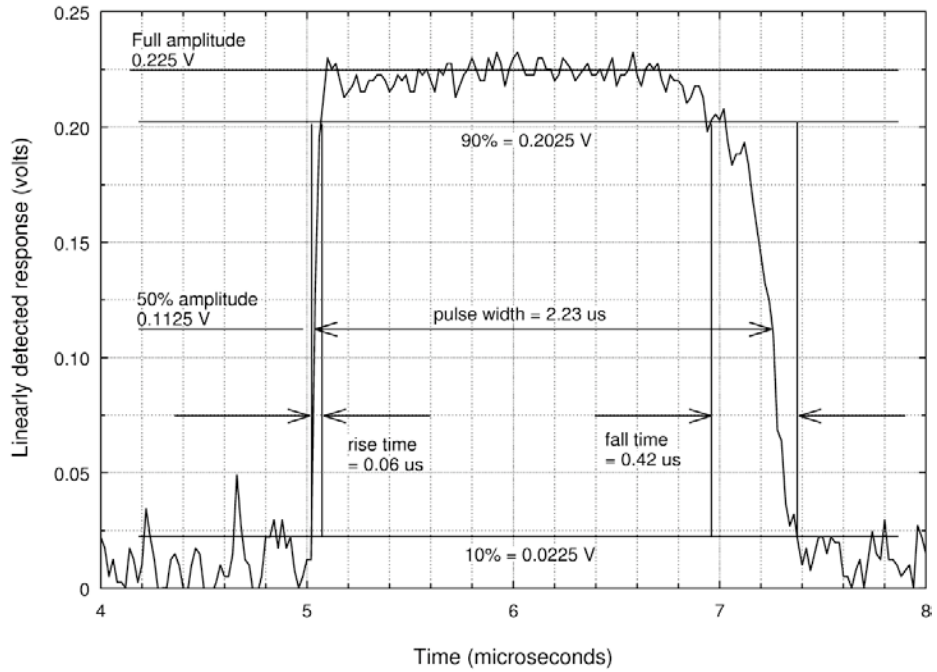


Figure 2. Diagram of parameters for a weather radar pulse, measured in radiated mode with a wideband detector.

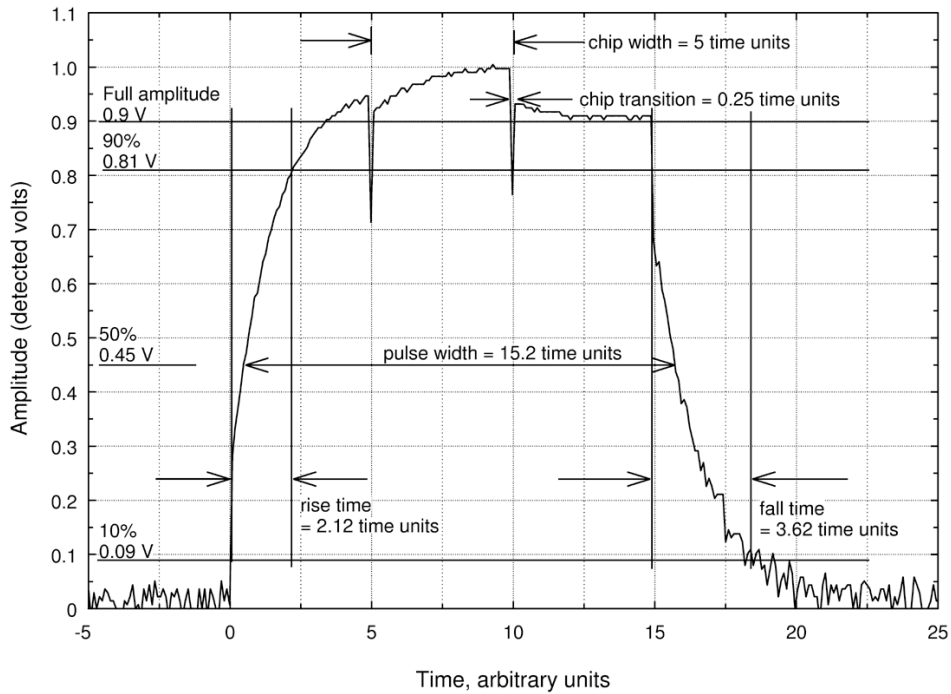


Figure 3. Diagram of pulse parameters for a phase-coded pulse with three chips, measured in radiated mode.

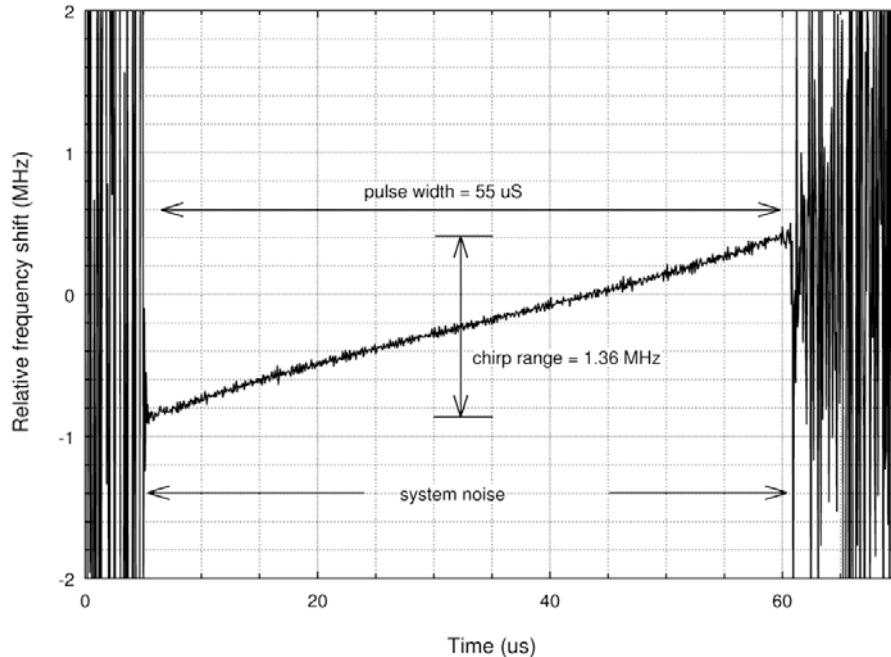


Figure 4. Measurement of the frequency deviation in time of a frequency-modulated pulse using a vector signal analyzer.

3. Pulse Repetition Rate

Pulse repetition rate (PRR) is measured with the same setup as shown in Figure 1 for pulse parameter measurements, although a spectrum analyzer may be substituted for the detector and digital oscilloscope. If a spectrum analyzer is used in lieu of the oscilloscope, then the radar pulses should be detected using the widest IF bandwidth available. The measurement system instrumentation is set up in a single-sweep mode, and initially a relatively low trigger threshold is used. A series of pulses are recorded, and then the trigger level is elevated and the operator waits for another set of pulses to activate the trigger. This process is continued until the threshold is so high that no more pulses exceed it. Then the trigger threshold is reduced slightly, and a pulse sequence is finally recorded. From this, the PRR is calculated.

If a radar produces a fixed-rate PRR at a single frequency, this procedure is trivial. Complications arise in radars with three types of complex PRRs. One is a non-uniform PRR, produced for example by staggered pulse trains emitted by air traffic control radars and some tactical radars. Another is the case in which a radar frequency-hops (and performs mechanical and/or electronic beam-steering) between pulses, but at a uniform PRR, resulting in only portions of the pulse train being produced at a single measurement

frequency. The pulse train is effectively fragmented, with pulses apparently missing in the measured train. The last is the case of frequency-hopping radars that also may perform mechanical and/or electronic beam-steering) between pulses, with a non-uniform (random or staggered) PRR.

The PRR depicted in Figure 5 is from a radar operating on a single frequency. The pulses were detected with a wideband discrete-component diode detector connected to the video output of a spectrum analyzer that was in turn connected to an antenna to measure radiated pulses; the spectrum analyzer was used in place of the oscilloscope in the setup of Figure 1. The radar measured in Figure 6 hops through sixteen frequencies that are spread across 500 MHz of spectrum, for a spacing of 31.25 MHz between frequencies. This measurement has been performed in a bandwidth of 3 MHz, which is an order of magnitude less than the channel spacing.

Despite the relative narrowness of the measurement bandwidth compared to the channel spacing, the out-of-band emission levels of off-tuned pulses on an adjacent channel are high enough to make them visible on the spectrum analyzer display, interleaved between the on-tuned pulses and two nearly on-tuned pulses (which are indicated with arrows). As a result, the apparent spacing between pulses as

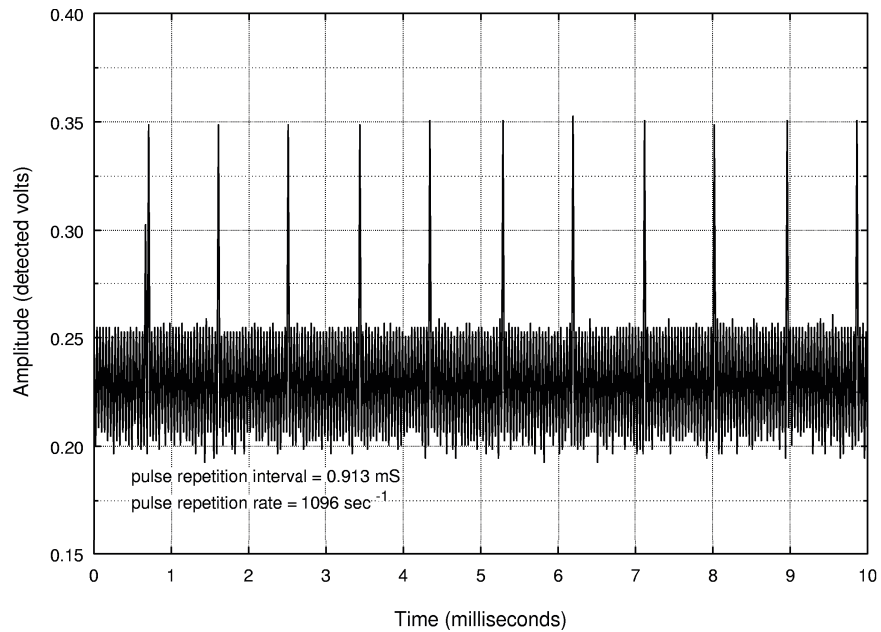


Figure 5. Example of a fixed-PRR radar pulse sequence.

Figure 6. Pulse repetition measurement on a single channel of a frequency-hopping radar made with a spectrum analyzer in a zero-Hertz span mode and positive peak detection. The line is an estimated threshold for on-frequency pulses.

measured by the spectrum analyzer is half of its true value (the apparent spacing being 0.5 ms, whereas the true pulse-to-pulse spacing on a single channel of this radar is nominally 1 ms). The same radar is measured with a wideband detector in Figure 7. The detector's

frequency response range far exceeds the 500 MHz frequency hop range of the radar. Therefore all pulses from all 16 radar channels are observed. The envelope of the pulse amplitudes is the beam shape of the main lobe of the radar antenna, observed as the beam sweeps

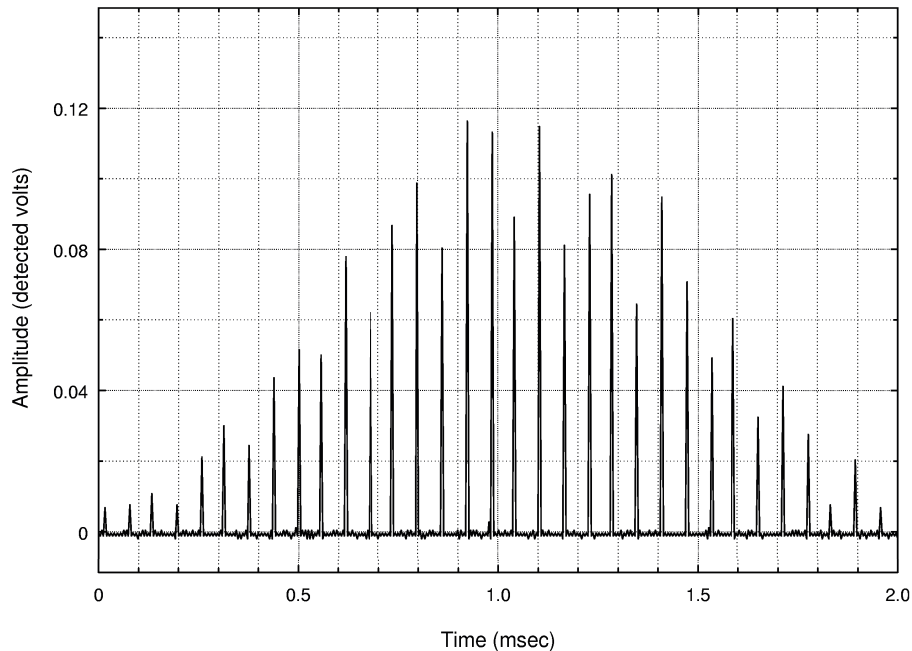


Figure 7. The pulse repetition rate of the same radar as that shown in Figure 6, but measured with a broadband detector configured as in Figure 1.

past the measurement location. (Note that there are about 18 pulses within the half-points of the beam, consistent with standard radar theory for reliable detections of targets.) The measured PRR is 16,470 pulses per second. Compare these data with those of Figure 6, in which pulses from a single channel are observed at 1/16 of the nominal rate, along with low-amplitude responses from adjacent channels.

4. Emission Spectra

Measurements of radar emission spectra are challenging. Difficulties that must be overcome include: the (typical) need to measure emission spectra over a wide dynamic range (90 dB or more); the need to measure spectra over a wide measurement frequency range (sometimes several gigahertz, plus harmonics); the need for high sensitivity (usually a 10 dB noise figure or less) in the measurement system; and the sometimes challenging task of assessing the proper bandwidth in which to perform the emission spectrum measurement.

Radiated emission spectrum measurements present complications due to the following factors: rotating radar antennas; radar beam scanning; and radar frequency hopping. Therefore, the stepped-frequency measurement procedure described below is desirable for such measurements.

4.1 Measurement Bandwidth for Radar Emission Measurements

The appropriate measurement bandwidth, B_m , is a function of the time waveform characteristics of the radar. Basic types of radar pulsed emissions are non-FM pulsed, phase-coded pulsed, FM-pulsed, CW, FM/CW, and phase-coded CW. The appropriate values of B_m for each waveform type are given in Table 1.

B_m may be confirmed empirically for any given radar system. This observation, called a bandwidth progression measurement, is performed as follows. The measurement system receiver is tuned to the fundamental frequency (or the frequency of a single channel if the radar frequency-hops), or within the chirp range if the radar chirps. The frequency-span range of the measurement system is set to zero hertz. The sweep time is set to a value somewhat greater than the radar beam-scanning and frequency-hopping interval, so that a maximum-amplitude peak power value is measured for each measurement system sweep. The measurement system IF bandwidth is set to the widest available value, and the received peak power level from the radar in this bandwidth is noted. The measurement bandwidth is then progressively narrowed, and the peak received power level is recorded as a function of the progressively

Table 1. Measurement bandwidths (B_m) for radar emission measurements.

Radar Modulation	Proper Bandwidth (B_m) for Spectrum Measurement
Non-FM pulsed and phase-coded pulsed	$B_m \leq (1/t)$, where t = emitted pulse duration (50% voltage) or phase-chip (sub-pulse) duration (50% voltage). Example for non-FM pulsed: If emitted pulse duration is 1 μ s, then $B_m \leq 1$ MHz. Example for phase-coded pulsed: If radar transmits 26- μ s duration pulses, each pulse consisting of 13 phase-coded chips that are each 2 μ s in duration, then $B_m \leq 500$ kHz.
FM-pulsed (chirped)	$B_m \leq (B_c/t)^{1/2}$, where B_c = frequency sweep range during each pulse and t = emitted pulse duration (50% voltage). Example: If radar sweeps (chirps) across frequency range of 1.3 MHz during each pulse, and if the pulse duration is 55 μ s, then $B_m \leq 154$ kHz.
CW	$B_m = 1$ kHz; See sub-paragraph 4.2 of [2, Chapter 5] for RSEC Criteria B, C and D. Example: $B_m = 1$ kHz.
FM/CW	$B_m = 1$ kHz; See sub-paragraph 4.2 of [2, Chapter 5] for RSEC Criteria B, C and D. Example: $B_m = 1$ kHz
Phase-coded CW	$B_m \leq (1/t)$, where t = emitted phase-chip duration (50% voltage). Example for phase-coded pulsed: If chip duration is 2 μ s, then $B_m \leq 500$ kHz.
Multi-mode radars	Calculations should be made for each waveform type as described above, and the minimum resulting value of B_m should be used for the emission spectrum measurement. Example: A multi-mode radar produces a mixture of pulse modulations as used in the above examples for non-FM pulsed and FM-pulsed. These values are 1 MHz and 154 kHz, respectively. Then $B_m \leq 154$ kHz.

narrower bandwidths. The end result is a graph showing measured power as a function of measurement system IF bandwidth, as in Figure 8.

The value of B_m will be the widest bandwidth that gives a peak power reading that is less than the full-peak power reading. If multi-mode radars emit different time waveforms on different frequencies (e.g., if non-FM pulses are emitted at frequency f_1 and chirped-pulsed emissions occur at frequency f_2), and if the calculated measurement bandwidths are different for these two waveforms, then the spectrum should be measured in *both* bandwidths. An example is shown in Figure 9.

4.2 Variation in Measured Spectra as a Function of Measurement Bandwidth

Measured radar emission levels are bandwidth-limited at their fundamental frequencies. That is, when measurement bandwidths equal or exceed the B_m values given in Table 1, the measured peak power will be

constant no matter how much the measurement bandwidth is increased. Conversely, in the radar unwanted (out-of-band and spurious) emission spectrum domain, measured levels of unwanted emissions will increase for measurement bandwidths that are wider than the B_m values given in Table 1, whereas the power measured at the radar fundamental will not increase for bandwidths wider than B_m . Measurements in bandwidths wider than B_m may result in apparent non-compliance with the RSEC or other standards, when the radar under test might have been in compliance had it been measured in accordance with the values of B_m given in Table 1.

4.3 Determination of Frequency-stepping Time Interval (Dwell Time)

As noted above, the most practical approach to measurement of radar emission spectra is to implement frequency-stepping rather than frequency-sweeping

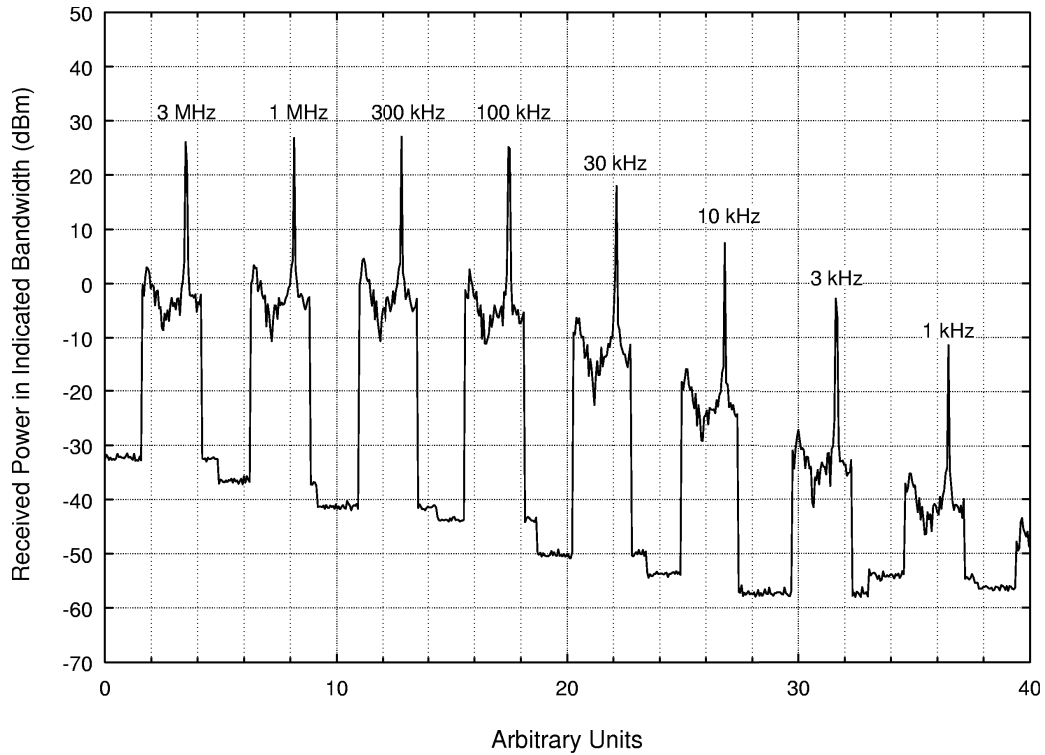


Figure 8. Example of a bandwidth progression measurement for assessment of the proper bandwidth in which to measure a radar spectrum. In this example, 100 kHz would be ideal because it is the widest bandwidth that gives less than a full-power response. The vertical drop-outs are due to sector blanking of the rotating radar beam in each scan.

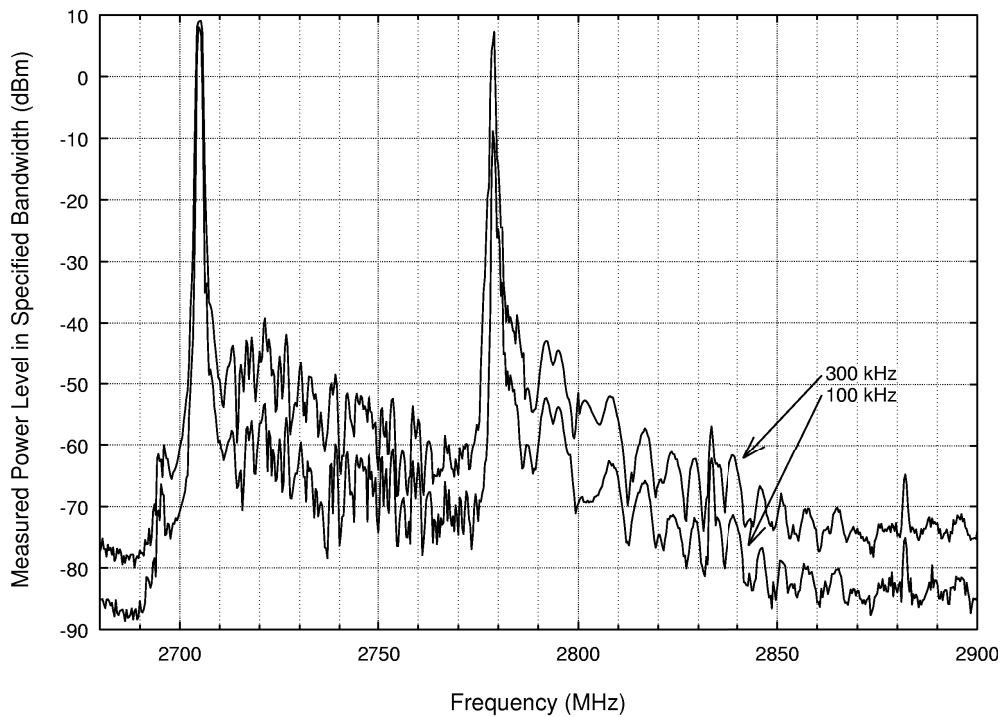


Figure 9. Emission spectrum measurement performed on a multi-mode chirped radar.

across the spectrum. This section describes the procedures for determining the time interval needed to determine peak power at each measured frequency (called dwell time) for radar emission spectra that are measured with the stepped-frequency approach.

Implementation of the dwell-time stepped-frequency measurement approach requires computer control of a spectrum analyzer. Alternative approaches that do not require computer control (such as sweeping across the spectrum in a maximum-hold trace mode) may be used, but with the caveat that they are inefficient and rather ineffective because they are probabilistic rather than deterministic in nature.

The necessary dwell time is a function of the radar antenna beam-scanning and frequency-tuning (fixed-tuned vs. hopping) characteristics. If the radar beam scanning can be stopped for the duration of the emission measurement, then the measurement step dwell time can be reduced to about 1 or 2 seconds, and the overall amount of time required to complete the measurement will be significantly reduced.

4.3.1 Conventional Beam-scanning, Fixed-tuned Radars

In this context, a conventional radar is one that scans a beam only in one dimension (usually azimuth), that repeats the scanning in a predictable, periodic manner, and that does not frequency-hop. Examples include air search radars with broad vertical beam patterns and mechanical azimuth scanning, sector-scanned radars (typically on aircraft), and phased-array radars that scan a beam only in azimuth.

The procedure for determining dwell time is as follows. First, a measurement location is identified. Then the measurement system is set up and tuned to the radar fundamental frequency and maximum attenuation is invoked in the RF front-end. The measured level is verified to be less than the saturation level of the measurement system. Attenuation (10 dB is suggested) may be inserted and removed in the measurement path to check for measurement system linearity.

The dwell time for each measured frequency needs to be slightly longer than the radar beam scanning interval. If the radar rotation interval is already known with certainty (e.g., 6 rpm = 10 seconds per rotation), then the dwell time may be set immediately, to a slightly longer interval (e.g., 11 seconds dwell time for a 10-second antenna rotation time).

If the beam scanning interval is not known, the following procedure may be used to measure it. The measurement system is tuned to the radar fundamental frequency. The frequency span of the measurement system is set to zero hertz, so that the radar beam scanning characteristic is now observed in the time domain on the system display. The sweep time of the measurement system is initially set to a few seconds, so that the radar beam is seen at least once on the measurement system display. Then the sweep time is gradually lengthened and additional sweeps are taken, until the radar main beam appears at least twice on the display. When such a display is achieved, a marker function is used to determine the time interval between the main beam features; this is the rotation (or sector-scan, if appropriate) interval for the radar. The dwell time of the spectrum measurement needs to be set to a value slightly longer than this.

4.3.2 Complex Beam-scanning and Frequency-hopping Radars

Some classes of radar scan space in elevation as well as azimuth, and may scan both these degrees of freedom with some amount of randomness; this is complex beam-scanning. Some radars change their tuned frequency on a pulse-to-pulse basis or at fixed or random intervals (i.e., they frequency-hop), and some radars combine complex beam scanning with frequency-hopping. The procedures for measuring the spectra of complex beam-scanning and frequency-hopping radars are nearly identical to those described above. The major difference is that the dwell time will need to be lengthened to ensure that a maximum peak level measurement will occur at each measurement step in the spectrum.

Complex beam scanning and frequency-hopping by a radar transmitter have the effect that maximum-amplitude pulses will not be directed toward the measurement system on its tuned frequency at *predictable* intervals, as is the case for conventional radars. Nevertheless, the antenna beam and the transmitted frequency will revisit the measurement system location and tuned frequency with a high probability within *some* interval; the problem is to determine that interval.

To do this, the measurement system is tuned to one of the radar fundamental frequencies in a zero hertz span and the RF front-end attenuation is adjusted to an appropriate value. The spectrum analyzer sweep time is set to a long interval, on the order of one minute. A single sweep is taken. The highest peak is identified, and then a delta marker is used to find the next-highest

peak. The delta marker is used again, to find the next-highest peak after that. The process is continued until all peaks with amplitudes within 2 dB of the highest peak have been catalogued. A pattern normally emerges. This pattern in time intervals between the highest peaks indicates the most probable interval that will elapse between radar main beam scans across the measurement location on the tuned measurement frequency. Unlike the situation for conventional radars, the dwell time for complex radars may require *two* or more antenna rotation periods (e.g., the radar may have a nominal 10-second antenna rotation period, but the dwell time required to measure a consistent peak value may be 20 or 30 seconds). In other words, the necessary time interval may be a random variable with a wide variance; in such cases, the dwell time needs to be long enough to assure that a valid peak is always measured, and this could turn out to be two or more complete radar scan periods.

The selected dwell time may be verified as correct by obtaining data in that interval a total of ten or twenty times, and noting the peak values returned from each of those individual times. If all these peak values are within 2 dB of each other, then the selected dwell time is adequate for the spectrum measurement.

As a matter of efficiency, it has been observed that this dwell time, while necessary for measurement of the radar spectrum at fundamental frequencies and within immediately adjacent out-of-band spectrum, is longer than what is required for measurement of the spurious spectrum. The dwell time required in the spurious domain for complex-beam-scanning and frequency-hopping radars is *less* than that required at the fundamental frequencies. For complex beam-scanning radars the reason is that the antenna does not generate a well-formed beam in the spurious domain. Also, frequency-hopping of the fundamental(s), which lengthens the step interval at those frequencies, has been observed during measurements as not presenting any issue in the spurious domain.¹

To determine how much shorter the dwell time can be in the spurious domain, measurement personnel should observe the radar's beam pattern carefully as the measurement progresses across the spurious domain.

¹The reason is that most radar pulses generate a spurious response that is relatively constant over a wide frequency range. Therefore, in the spurious domain, many different frequencies of fundamental pulses generate approximately the same amplitude at any given spurious frequency.

Eventually, they will observe that the radar peak amplitude is always repeated *twice* during each step. When this happens, the dwell time may be reduced by a factor of two. As the measurement progresses further through the spectrum, the phenomenon may occur again. If it does, the dwell time again may be reduced by half. This process may be continued as necessary, and will greatly reduce the overall measurement time without causing any degradation to the results.

4.4 Stepped-frequency Spectrum Measurement Procedure

The radar emission spectrum measurement is performed with a measurement bandwidth selected in accord with Table 1 and a dwell time selected as described above. A block diagram of the recommended measurement system is shown in Figure 10. To perform the measurement, the system is initially tuned to a selected frequency in a zero hertz span and the power level received from the radar beam is measured in the selected bandwidth with a peak-hold detector for the dwell interval. At the end of the dwell interval, the maximum power that has occurred is recorded. Then the measurement system is tuned higher in frequency by the amount of the measurement bandwidth (or perhaps slightly less). Again the received power is measured for a dwell interval and then the maximum power received in that interval is recorded. This *stepped-frequency* process is repeated until the desired amount of spectrum has been measured.

RF front-end attenuation is adjusted throughout the measurement, on a step-by-step basis, to keep the level of the received radar signal within the instantaneous dynamic range of the measurement system. It may be desirable to measure the emission spectrum in additional bandwidths. Such a set of spectra may be found useful at a later date, for they will show the progression of measured levels as a function of B_m across the spectrum. Example emission spectra measured with the stepped technique are shown in Figure 11. These spectra show the variation in unwanted emission levels as a function of B_m when it is less than, equal to, and greater than the value recommended in Table 1.

5. Antenna Patterns

Antenna patterns of radar transmitters are not generally measured in the radar main beam, because this energy is usually directed into space in directions inaccessible to terrestrial measurement systems. An example is the beam of a typical air search radar, the lower edge of which is ordinarily tilted about a degree above the horizon. However, horizontal plane radar patterns are

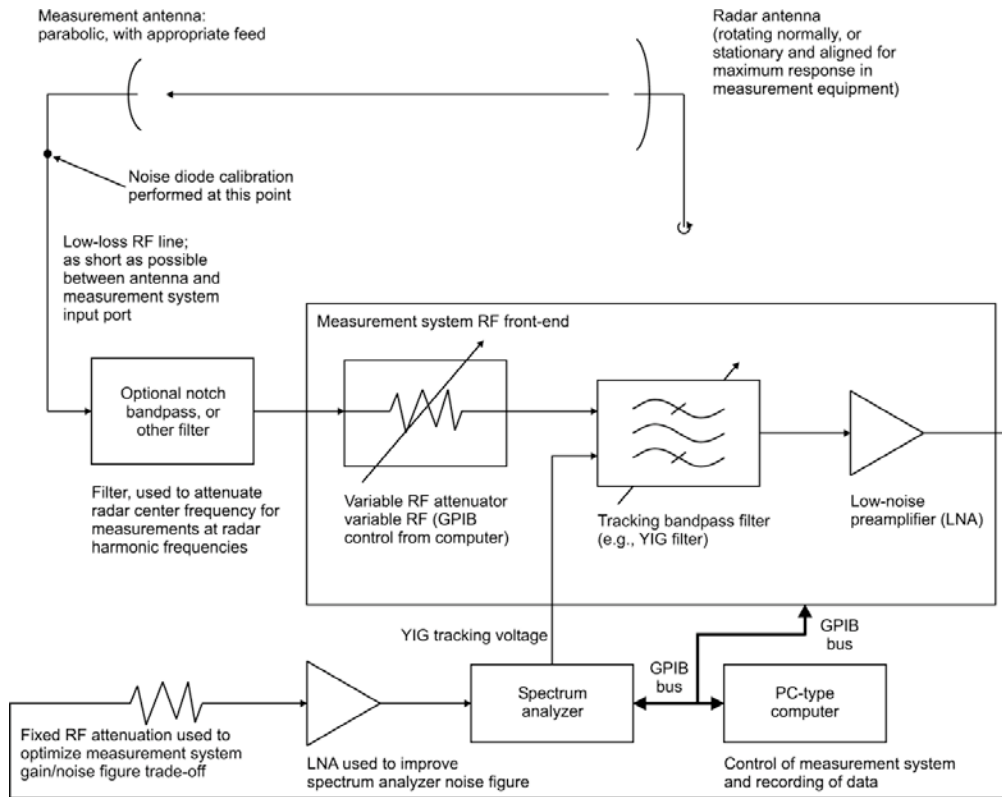


Figure 10. Block diagram of the RF front end and associated hardware recommended for radar emission spectrum measurements.

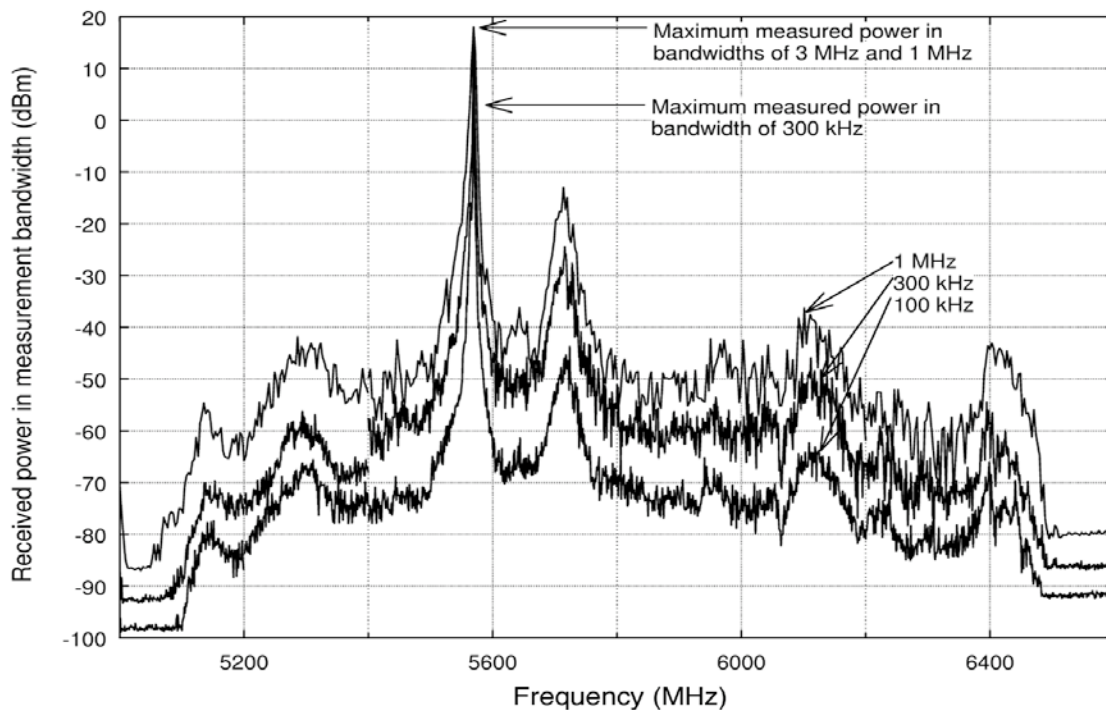


Figure 11. Three spectra for a single radar, measured in bandwidths less than, equal to, and greater than the value of B_m recommended in Table 1.

useful, especially for calculating interference potential to terrestrial systems. Radar antenna patterns should be measured from a location that maximizes the signal to noise ratio at the measurement system.

For antenna patterns of conventional radars, the measurement system is fixed-tuned to a radar fundamental frequency in a zero-hertz span. The sweep time is set to the dwell interval used for spectrum measurements (as described above). The same bandwidth should be used as for the radar spectrum measurement (Table 1), since this provides maximum dynamic range (signal-to-noise ratio) for the antenna pattern measurement. Positive peak detection should be used. Figure 12 shows an example antenna pattern measured in this manner.

For advanced radars, the electronic scanning and frequency-hopping of the beams cause antenna patterns to be measured as discontinuous points rather than smooth envelopes. To produce a smooth or nearly smooth envelope, the measurement should be repeated ten or twenty times. Subsequently, the resultant raw data should be normalized in time and added together digitally, to make the final pattern a reasonably smooth envelope.

A problem with all antenna pattern measurements on radars is that multipath-generating obstacles in the vicinity of the radar will cause nulls and peaks in any given pattern measurement. Variations such as multipath due to vehicles, buildings and other radio-reflective objects also will occur. To eliminate most of these features, perform the following procedures:

- Measure the radar antenna pattern several times at one location, and cross-correlate the results to eliminate temporal multipath effects at that location.
- Move the measurement system to another location and repeat the procedure to eliminate temporal variation at the second location.
- Find the median of the patterns from these two separate locations.
- If desired or necessary, repeat this procedure at a third measurement system location, and find the median of the three patterns.

The result is shown in Figure 13. This pattern will probably approach the result that would be obtained if the radar antenna pattern had been measured in an anechoic chamber.

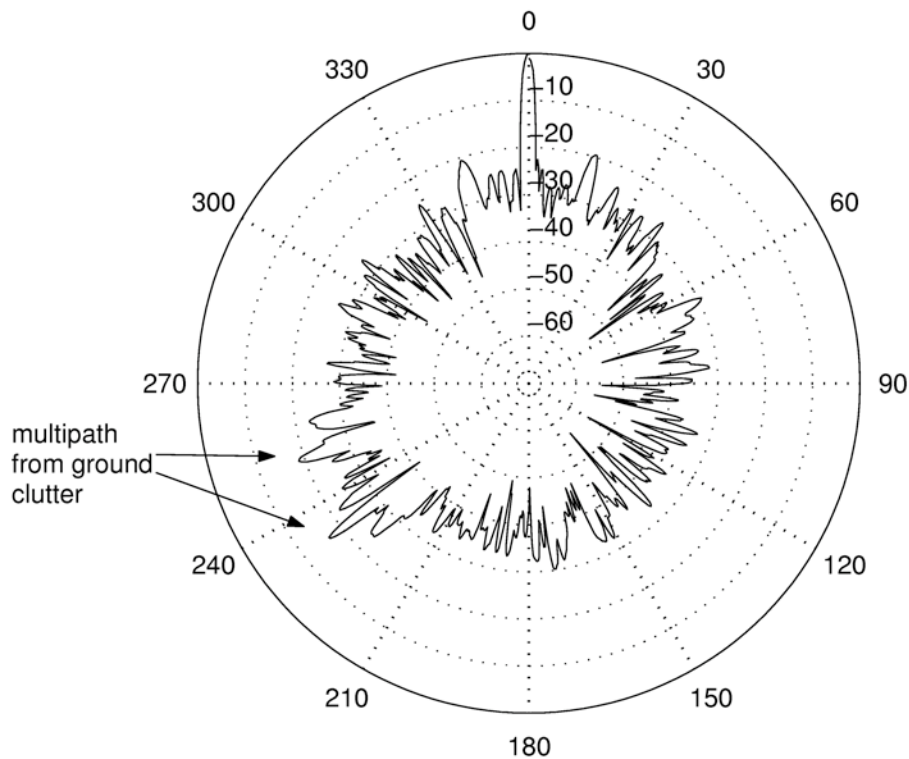


Figure 12. Example of a cluttered, raw radar antenna pattern for a maritime surface search radar.

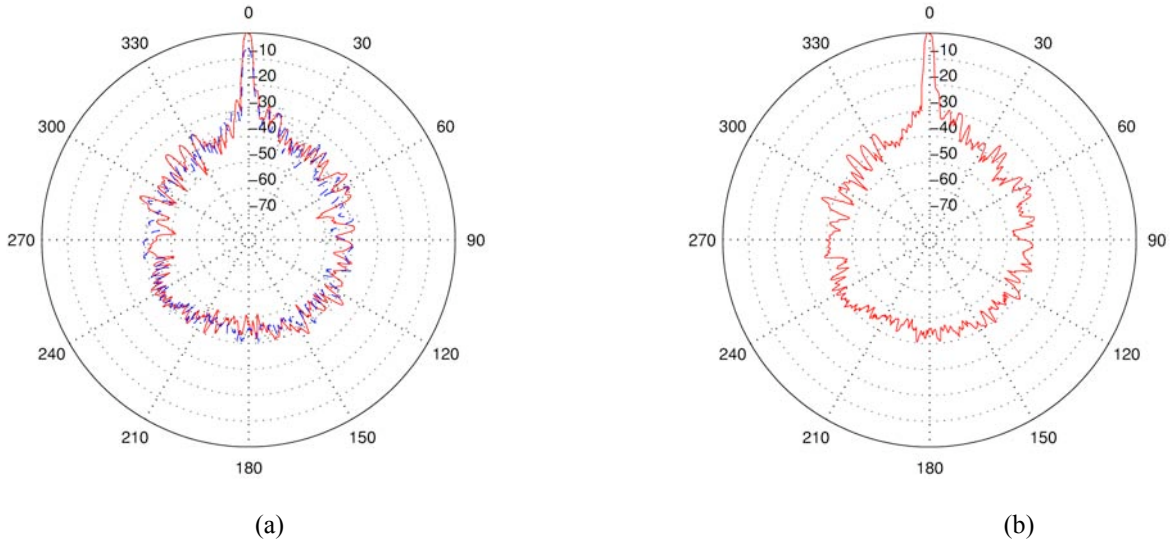


Figure 13. Three overlaid raw antenna patterns (a) of the radar of Figure 12, each containing multipath clutter. Median of patterns are shown in (b).

6. References

- [1] F.H. Sanders, R.L. Hinkle, and B.J. Ramsey, "Measurement Procedures for the Radar Spectrum Engineering Criteria," NTIA Report 05-xx, Feb. 2005.
- [2] *Manual of Regulations and Procedures for Federal Radio Frequency Management*, May 2003, Revised September 2004, NTIA Office of Spectrum Management, U.S. Government Printing Office, Stock No. 903-008-00000-8.

Adverbs and Adjectives: An Abstraction for Software Defined Radio

Troy Weingart, Doug Sicker, Dirk Grunwald and Michael Neufeld

Department of Computer Science

University of Colorado at Boulder

Boulder, CO 80309

Tel: 303.492.7514 Fax: 303.492.2844

{Troy.Weingart,Douglas.Sicker,Dirk.Grunwald,Michael.Neufeld}@colorado.edu

Abstract

Many wireless network and transport protocols take advantage of the interaction between the physical and link layers to achieve reasonable performance, reliability, and energy efficiency. Such “cross-layer” dependences are typically explicitly enumerated in order to cause predetermined behavior in the physical and link layers. The need for cross layer interaction will only increase as advanced physical and link interfaces are realized in software radio systems (SDRs). In order to meet the demands of large systems of heterogeneous software radios we will have to develop application and system programming interfaces that are highly flexible and portable. This paper demonstrates how a set of modifiers for the “verbs” and “nouns” in communication protocols can achieve the performance, flexibility, and portability improvements required. Adverbs in our abstraction, such as send locally and send reliably, are used rather than explicit directives. Modifiers like these provide freedom for each layer in the protocol stack to choose from a proven palette of cross-layer techniques. This work presents a subset of potential modifiers and their application. The promise of the abstraction is illustrated through simulation of the send locally adverb.

1 Introduction

Software Defined Radios (SDRs) promise to redefine wireless communications in numerous and profound ways. The ability to dynamically redefine the lower layers of a radio device offers tremendous opportunities to improve communication capabilities and efficiencies. This is in stark contrast to the static nature of traditional radio devices, which tend toward fixed operational modes and potentially inefficient use of the available RF spectrum. Beyond these technical (and regulatory) limitations, the static nature of the protocol stacks associated with these devices further limits their potential efficiency. This type of inefficiency is often due to the fact that higher layers make incorrect assumptions about lower layers and channel conditions. Such inefficiencies are further exposed when the protocols are evaluated against new metrics such as energy efficiency, overhead or impact on the noise floor. As a result, cross-layer approaches to overcome these deficiencies have become a common theme in the literature.

Such cross layer interactions occur at different layers of the network. For example, TCP may depend on the link layer for information about the cause of packet loss or expiration of timers. In the absence of such knowledge, TCP may relate the cause to network congestion. In reality it might be that transient noise introduced extra errors. Similarly, one may depend on the routing, link and physical

layer to provide the QoS. The routing layer may try to use multiple routes while the link layer may assist by choosing less congested links. Similarly the routing protocol also depends on the lower layers. Originally, many protocols were designed with little consideration of the properties of lower layer layers of the protocol stack; for example, application protocols largely viewed wireless networks as being similar to wired networks. However, lower layers (link and physical layer) play a significant role in achieving good performance in wireless networks. For example, choosing a higher capacity link at the physical layer or avoiding nodes with high link-layer contention can improve the throughput dramatically. Other desirable network performance metrics may also be met through cross layer interactions. For example, energy consumption, though a physical layer property, may depend on the needs of the higher layers. A routing protocol may vary transmission power depending on its need to reach just one or many nodes.

As a result, one may ask the question *how should such cross layer interaction be expressed?* In this paper, we propose a framework for cross layer interactions. In this, we can abstract the higher layer interaction from the lower layers using *adverbs*. In the traditional linguistic context, adverbs are used to modify verbs, adjectives or other adverbs. In our model, we apply the adverb analogy to modifying verbs associated with the communications. For ex-

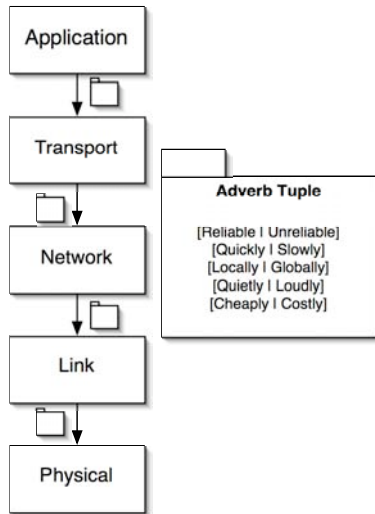


Figure 1: The adverb tuple is passed down the protocol stack. Each layer can select a mechanism to optimize performance according to the specified attributes.

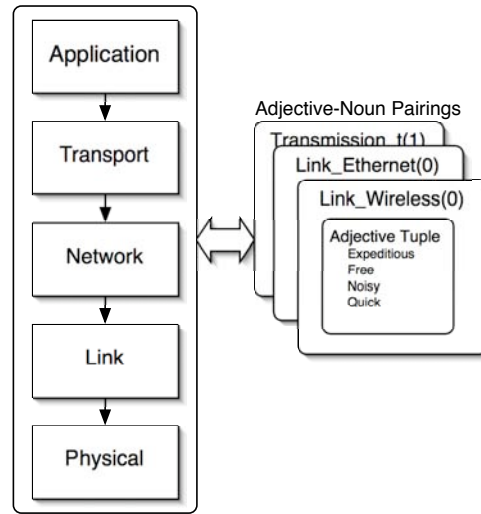


Figure 2: Each layer in the TCP/IP stack is able to access, modify, and/or act upon adjective-noun pairings optimizing performance or changing behavior according to the specified attributes.

ample, one might want the data to be send “quickly”, “reliably” or “locally”. Similarly, the properties of the layers can be abstracted using an *adjective*. Again, in the traditional context adjectives are used to describe nouns. In our model, an adjective is used to describe a communications attribute. For example, a network link can be “capacious”, the medium can be described as “noisy”.

The organization of the paper is as follows. We first describe our framework of adverbs and adjectives in Section 2. Then in Section 3 we show that using our more abstract framework allows us to improve the performance of a routing protocol using two distinct mechanisms available in the lower layers. One mechanism, transmission power control, is managed by the physical layer. The other mechanism, transmission rate control, is managed by the link layer. Either mechanism can be used to control the range of message transmissions. We also show that both mechanisms have similar effect on goodput in an 802.11b network and argue that they can be used interchangeably or in combination, resulting in improved routing algorithms that are more flexible than those that explicitly control a single mechanism. We then briefly review the considerable body of work that has explored cross-layer protocol design in Section 4. We conclude the paper by discussing our results and future work.

2 Conceptual Model

It is our belief that by modifying commands with information that is illustrative of the applications requirements we can improve the overall performance of the protocol stack. Standard communication protocols, routing algorithms and

applications send commands to affect lower layers. Our framework treats commands as verbs that can be modified through the use of adverbs. A tuple of adverbs is attached to the command.

As shown in Figure 1, an application generates an adverb tuple consisting of its communication requirements and passes this information down the stack. The layers read and autonomously act on the adverb tuple by selecting mechanisms that optimize performance according to tuple attributes. More specifically, the routing layer may instruct the link layer to send *quickly*. This can be interpreted at the link layer as choosing a link with more available bandwidth. Notice that the mechanism for how the link layer reacts to the modified command is not specified as part of the adverb tuple. The adverb abstraction allows the lower layer to choose any suitable mechanism to honor higher layer requests. In an advanced system that incorporates a software-defined radio (SDR), the adverb tuple could cause the link and physical layer to dynamically reconfigure the interface. The SDR would then be able to more efficiently use the spectrum available in the ISM band, possibly forming a single high-speed channel. This newly formed channel would greatly surpass the bandwidth offered by standard 802.11a/b/g solutions.

The methodology used in grafting a set of adverbs onto commands can also be applied to dynamically characterize properties of the communication environment. The adverb to command framework was extended to make use of adjectives to modify layer properties, expressed as nouns, in order to realize dynamism and performance gains. In Figure 2, the layers of the TCP/IP stack are able to access and modify the adjectives that pertain to a paired noun. Each of the layers can then take action to maximize performance

based on how a property is modified by its adjectives. One should note that the depicted noun and adjective tuple are one of many possibilities. In this instance the noun refers to the wireless link *w0*.

We contend that an adjective tuple can have a positive effect when paired with nouns that describe layer properties. For instance, a noun, *link status*, could be paired with adjectives like *busy*. In the 802.11 MAC protocol if a node wishes to transmit and another node is transmitting, the station attempting to communicate must defer its transmission. If we were to use the busy adjective with respect to link status, other layers could act upon this information. A system incorporating SDR could use this information to switch the transmitter to another channel and bypass the busy link without waiting. Thus, by using adjectives we are able to encapsulate communication and environmental properties without using explicit parameters or values.

It is important to note that adjective and adverb tuples can be interpreted and acted upon differently through and across the layers in the protocol stack. The adverb *locally* at the application layer could mean finding a printer physically near you. At the routing layer, it may mean finding a node fewer than two hops away.

The model we present here serves as a basis from which a more complete framework can be constructed and is not intended to be all encompassing. Rather, it was constructed to illustrate the viability of such an approach to improving performance, reliability, and energy consumption.

2.1 Adverbs

The following subsections serve to illustrate how adverb-command pairings may interact in our notional framework. Again, the adverbs discussed are not an exhaustive collection; rather, they serve to demonstrate the viability of our model.

2.1.1 Quickly vs. Slowly

The IEEE 802.11a/b/g standards are multi-rate in that they provide physical-layer mechanisms to transmit at higher rate than the base rate, if channel conditions permit. For example, the 802.11b standard offers different transmission rates such as 1, 2, 5.5 and 11Mbs. As a result, different link layer protocols were designed to exploit the availability of higher transmission rates. Auto Rate Fallback (ARF) was the first commercial implementation using this multi-rate capability at the MAC layer. With ARF, senders use the history of previous transmission error rates to adaptively select future transmission rates. Receiver Based Auto Rate (RBAR) is an enhanced protocol designed to exploit the multi-rate capabilities of the MAC layer [1]. RBAR lets the receiver control the sender's transmission rate through RTS/CTS negotiation.

RBAR and ARF are protocols that negotiate the transmission rates at the link layer. It is important to realize that the routing layer remains ignorant of such link layer properties. Alternatively, by using our framework, one could attach the adverb, *quickly*, to the routing layer tuple allowing the link layer to route data to the destination node using the faster link. This can be achieved if the routing layer can instruct the link layer to send the data rapidly. The link layer will choose to send the data at higher transmission rate if the channel conditions permit. Or the link layer could optionally send the data over a less congested link, honoring the higher layer request with an entirely different mechanism[2, 3].

The application layer can also have different QoS requirements; naturally, one of these is sending data quickly. The quickly adverb may be honored at the routing layer by choosing the routes that guarantee the latency specified[4, 5]. Subsequently, the link layer could negotiate between different nodes choosing the quicker link. Alternatively, different classes of traffic could be introduced by changing the link layer back off; thus offering another means to send data quickly.

The adverb *slowly* refers to the case where there is not a requirement to send data quickly. There are many classes of traffic that do not require low latency and high bandwidth. One could imagine a SDR that would dynamically select a slower, noisy link for a FTP session, or forwarding of email traffic, based on an adverb tuple that contains slowly.

2.1.2 Locally vs. Globally

It is widely believed that the key factor in building scalable network is the locality of the network traffic[6]. In other words, each node talks directly only to the nodes within a fixed radius, independent of network size. Using a large transmission range causes more interference with neighboring network traffic and reduces performance. Clearly, there is strong motivation for routing *locally*.

An obvious course of action to take in the presence of an adverb tuple which contains the *locally* adverb is for the physical layer to send the packet at a lower transmission power. This results in a shorter transmission range and reduces the amount of RF interference generated by the transmission of that packet.

Alternatively the link layer could change the packet transmission rate. Different transmission rates use different modulation techniques. Higher transmission rates use encodings that are faster to transmit, but are also more susceptible to error and hence require a better signal to noise ratio for successful reception. This is often expressed in terms of distance in open space, *i.e.*, an unobstructed, open field with minimal outside RF interference. Table 1 shows the specified ranges of communication at different trans-

Transmission rate	Open range
1 Mb/s	550m
2 Mb/s	400m
5.5 Mb/s	270m
11 Mb/s	160m

Table 1: Open range for a Orinoco Gold card

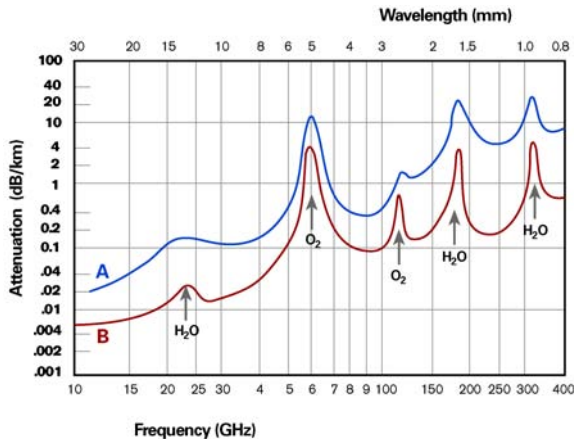


Figure 3: Signal attenuation across the millimeter-wave RF spectrum

mission rates for an Orinoco Gold 802.11b card. Thus when encountering the adverb, *locally*, the link layer can reduce the successful packet reception range by selecting a higher transmission rate. While this does not reduce the area over which the packet produces RF interference (as is the case when reducing transmission power), it does reduce the *time* required to transmit the packet. We will explore the details of how altering these lower layer network properties, *i.e.*, transmission power and symbol encoding, interacts with higher layers in the network stack in section 3.

Conversely the routing layer may want to route packets *globally*. In a less congested medium, one may want to send the packets globally if one is unable to find a route locally. If different traffic sources in the network are using separate physical channels, it may be more efficient to send data globally as it will reduce the number of hops without causing interference with neighboring traffic. Through the high-level abstraction provided by adverbs the lower protocol layers may achieve the globally goal using one or both of the mechanisms described.

At the application layer the adverb *locally* could have different implication. A user may wish to find a printer locally. The underlying routing protocol may honor it by supplying information about the printers in the same subnet. Alternatively, some geographical routing may help in choosing a physically close printer.

The power of our abstraction is even more clearly real-

ized when using SDR; many of the mechanisms available with a fully software defined radio would not have been considered by most protocol designers. If a radio has the ability to widely shift the frequency used, the *locally* adverb can be implemented by frequency shifting. As shown in Figure 3, low frequencies, such as in the short-wave band, have far reach but low bandwidth, whereas higher frequencies tend to have shorter range and increased bandwidth. Specific frequencies (such as that around 60GHz) have a pronounced attenuation due to absorption in oxygen. A SDR system, in order to fulfill the *locally* adverb, could select the 60GHz band to limit its transmission range. The framework frees an algorithm designer from encapsulating knowledge of oxygen absorption in their algorithms. The mechanism for meeting the intent of the adverb tuple is left in the hands of the expert. In this case, the designer of the SDR antenna hardware and software.

2.1.3 Cheaply vs. Costly

Reducing energy consumption by wireless communication devices is one of the most important concerns in designing an 802.11 solution. Although transmit power is a physical layer property, higher layers can also play an important role in determining the energy consumed by a node. For example, a proactive routing protocol will spend more energy than a reactive protocol, simply by virtue of sending more packets. By using the *cheaply* or potentially defining a *conservatively* adverb, the link and physical layers may chose to transmit and/or route in a manner that optimizes energy consumption. Additionally, the application layer may assist by using better compression on the data. Cross layer contributions can also be built upon at the transport layer. This layer, by being more cautious about initiating congestion control, can contribute to the overall goal of transmitting cheaply. Alternatively, the routing layer may choose to route data through energy-rich nodes [7]. It may also choose to send fewer routing packets, or the layer may choose to send larger sized packets rather than sending multiple small packets. Additionally, the link layer may determine an optimal combination of transmission rate and transmission power control to minimize the energy consumption [8]. Again, the abstraction provides a large amount of flexibility to enhance performance both within and across layers.

One could imagine a scenario where all links are congested, noisy or down. The application could then incorporate the costly adverb in the tuple. As a result the lower layers could dynamically select a link that requires the user to pay a fee for use. Also, when you have a powered base station, energy constraints may not be a concern and one may consider expensive transferring of data or possibly be more charitable about routing data for others.

2.1.4 Reliably vs. Unreliably

The dynamic nature of mobile networking is attributed to variable link characteristics, node movements, changing network topology, and variable application demands. As a result, it can be quite hard to guarantee reliable transmission of packets. Although reliable protocols such as TCP aim to provide end-to-end guarantees, the lower layers can also play a powerful role in improving network reliability and performance [9, 10, 11]. It is a logical assertion that the higher layer may wish to instruct the lower layer to help in sending data *reliably*. Again, the reliable attribute of the adverb tuple in our framework can have intra and cross layer benefits.

At the transport layer, this may be realized by choosing a reliable protocol such as TCP or SCTP. The routing layer may choose to use multiple routes to send data so as to ensure reliability [12, 13]. Further, the link layer may choose to send the data at a lower transmission rate while making maximum use of error correction. Additionally, a choice can be made to send data over different physical channels to minimize contention and reduce packet corruption and delay. An SDR has the added flexibility to redundantly send the data over multiple spectral ranges.

2.1.5 Quietly vs. Loudly

Due to the broadcast nature of the wireless media, when two hosts are communicating, all other hosts within the range of the two hosts must defer their transmissions in order to avoid a collision. Hidden terminals also complicate and contribute to the congestion problem. Performance degrades further when interfering with a bottleneck node. One can see the potential advantage in sending data quietly. Through the use of our framework one could gain a tremendous advantage by facilitating intelligent use of transmission space and spatial reuse. An SDR equipped with a narrow beam steerable antenna would have a dramatic impact on the ability of nodes to transmit concurrently. The flexibility of the framework is again realized across layers.

At the routing layer, this may be done by choosing maximally disjoint routes [14]. Through routing one may proactively try to avoid formation of bottleneck nodes. Additionally, the link layer may choose to reduce the transmission range, thereby reducing the number of nodes impacted by its transmission. Also, the link layer may choose to send the data in different physical channel, reducing the effect on other traffic [15, 16, 17]. At the physical layer, an SDR may also help by switching to a block of quiet spectrum to send the data.

2.2 Adjectives

The following discussion illustrates how adjective-noun pairings may interact in our notional framework. Like the

previous subsections, the adjectives discussed here are not an exhaustive collection; rather, they serve to demonstrate the viability of our model.

2.2.1 Link - Busy

The 802.11 MAC protocol employs carrier sense multiple access with collision avoidance (CSMA/CA). In this protocol, the node first senses the medium. If the medium is busy, i.e., some other node is transmitting, the station defers its transmission to a later time. This can often lead to packet delay and expiration of network timers. If the adjective *busy* were paired with the link, the transport layer via TCP, could interpret *busy* as congestion and take necessary corrective measures. In addition, the application layer could change its QoS requirements based on this information. A streaming video application could dynamically switch to buffering more data in light of the busy media.

2.2.2 Link - Noisy

Transient noise and inhospitable physical conditions may also cause packet loss, high errors and expiration of network timers. In this instance the medium is noisy rather than busy. Wireless media is more susceptible to transient noise. Without the availability of information afforded by the adjective framework, higher layer protocols such as TCP may incorrectly assume the cause of the packet loss was congestion. TCP congestion control over a noisy link may make the situation worse. An SDR may use spread spectrum transmission techniques to minimize the impact of noise. Some advanced antenna technologies are also able to emit energy patterns to cancel out noise.

2.2.3 Other Adjectives

One can imagine a myriad of additional adjectives that could serve to enhance our framework. In the adhoc networking domain one could imagine the noun *topology* being modified by adjectives like *mobile* or *stable*. An SDR acting upon this information could reconfigure to use favorable routing algorithms, more efficient symbol encoding, or exploit spatial reuse through antenna directionality. We believe that our framework when paired with an SDR has the potential to offer huge improvements in performance, energy use, and overall responsiveness to the users desires. The following section serves to demonstrate the potential of the framework through simulation.

3 In depth: Route “locally”

Network wide broadcasting is a fundamental operation in wireless *ad hoc* networks. Its goal is to transmit a message from a source node to many or all nodes in the network.

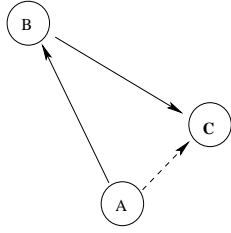


Figure 4: Formation of needlessly long routes in a pathological case

It is generally employed by the source node to search for a route to the destination node. Unfortunately, broadcasting increases congestion in the network while also causing interference with neighboring network traffic.

Due to increased interference, the performance of *ad hoc* networks can be significantly degraded. Li [6] argued that the key factor in building scalable *ad hoc* networks is the locality of network traffic. In other words, each node talks only to the nodes within a fixed radius, independent of the network size. In such a case, the per-node capacity of the network remains constant. Thus, there is a strong motivation for nodes to communicate mostly with local nodes and thus most routing will probably be local as well.

Another motivation for a node to route “locally” is to prevent formation of needlessly long routes. Figure 4 illustrates this situation. In this figure, node A is trying to find route to node C and broadcasts a “route request” message. That message reaches nodes B and C, but node B replies before node C and the route $A \rightarrow B \rightarrow C$ is formed at node A. Meanwhile, node C’s reply is lost due to congestion or is prevented from replying as the medium is busy with B’s reply. Such a situation can occur during a broadcast storm when a node is trying to find a route. Alternatively, if node A were able to limit the propagation range of a route request, it would only communicate with node C, resulting in the direct route of $A \rightarrow C$.

As mentioned, broadcast and unicast packets are usually sent at different transmission rates. The broadcast packets are usually sent at the base frequency (1Mb/s or 2Mb/s) and thus reach a greater number of nodes. Unfortunately, this can result in situations where a node can hear broadcasts from another node, but not be able to reply using a higher data rate. For example, node B may receive a route request from node A, but may not be able to directly reply at 11Mb/s. This problem can be mitigated by the use of link layer rate negotiation protocols such as RBAR (Receiver-Based AutoRate) [1]. The RBAR protocol establishes the optimal transmission rate to send the packet via RTS/CTS exchange. Although it prevents the situation described above, this negotiation takes time.

These link layer problems are exacerbated by the flooding nature of most route request mechanisms. If node B did

not know of a route to node C, it may re-broadcast the route request on behalf of node A. Many routing algorithms, including AODV (Ad Hoc On Demand Distance Vector Protocol) use such an “expanding ring” to try to locate routes. Hence, the notion of routing “locally” is appealing. The routing layer may achieve the notion of routing “locally” with some assistance from the link layer, as explained in Section 2.1.2. On the request of the routing layer, the physical layer may reduce the transmit power of the broadcast request packets or the link layer can send it at higher transmission rate. This technique reduces its transmission range but not the interference range. Alternatively, the system may use a combination of these two mechanisms as dictated by the channel conditions.

If the *ad hoc* routing protocols are oblivious to lower layer characteristics such as transmission range, it is hard or impossible for it to do “local” routing. This problem is further exaggerated by the fact that most routing protocols were designed with certain assumptions about the lower layers, such as a single-rate link layer or a single transmission power for all packets. Our abstraction frees the application or system level programmer from relying on specific lower layer capabilities in order to realize performance gains.

3.1 One adverb, two mechanisms, similar outcome

We conducted an experiment to evaluate the effect of changing transmission rate and transmission power individually. Two mechanisms are used to in this experiment to gauge the impact of the “local” adverb on routing performance. While using transmission rate control, each node sends the request packet at a fixed transmission rate (either 11, 5.5 2 or 1 Mb/s). Each transmission rate has varying transmission range, as shown in Table 1, resulting in varying effective throughputs at different transmission rates. In another set of simulations, we varied the transmission power. We selected four transmission power levels that would result in effective communication ranges approximating those of the transmission rates. For example, the 5.5mb/s transmission rate has an effective range of 270m, we selected a transmit power level that also had an effective range of 270m. The rate or power of both broadcast and unicast packets are controlled, but only the range of the broadcast packets is controlled by the routing layer. The unicast packets use the RBAR [1] adaptive rate control mechanism. By using this technique the differences in performance arise from route selection, not just from effective data transmission rates.

Transmit power control is more flexible than changing transmission rate since the power level can usually be more finely controlled, but not all wireless interfaces support transmission power adjustments. Similarly, not all

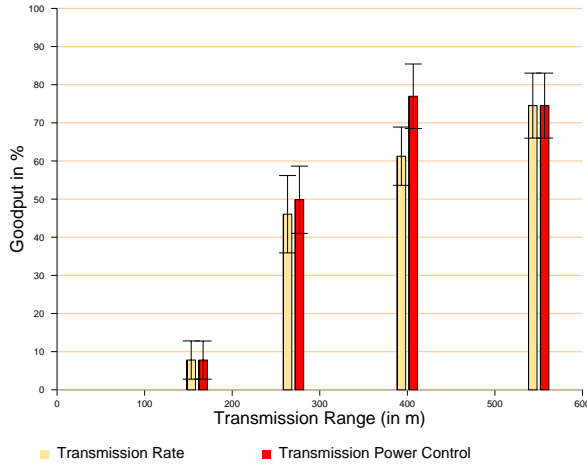


Figure 5: The effect of using transmission power control and transmission rate on goodput. The x-axis represents the transmission range, which is kept the same using power control and transmission rate in both cases.

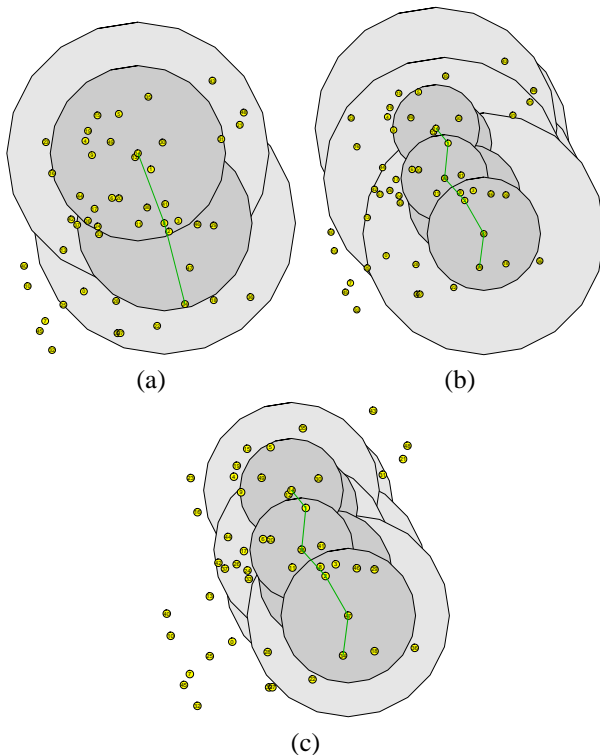


Figure 6: Visual representation of noise imparted by a single end-to-end transmission. The darker shade represents the transmission range while the lighter shade represents carrier sensing range in (a) Default AODV with 1Mb transmission rate for broadcast packets, (b) Stateful algorithm using transmission rate control and (c) Stateful algorithm using transmission power control

interfaces support multiple transmission rates. It should be noted that the ability to change transmit rate as well as transmission power of each individual packet is supported by many modern chipsets, e.g., the Atheros[18] 802.11a/b/g chips and the Intersil Prisms family of 802.11b chips.

Our results for a static wireless scenario are summarized in Figure 5. The horizontal axis in the graph corresponds to the transmission range achieved by increasing rate from 11Mbps to 1Mbps or by suitably changing transmission power. The error bars record the 95% confidence interval across different nodes. The two mechanisms (power vs. rate) result in statistically indistinguishable goodputs at each power level. The goodput achieved decreases for smaller transmission range because reducing the range of the broadcast request packets partitions the network. Intuitively, we would expect higher goodput with a higher transmission rate. However, in this simulation, we used a fairly low traffic injection rate (4 packets/s, 64 bytes/packet) thus did not gain from using the higher capacity links.

Reducing the transmission power reduces the carrier sensing range as well as reducing the transmission range. However, using transmission rate control only decreases the transmission range while the carrier sense range does not change. For the lowest transmission range, the difference in performance with both rate and power control is negligible since the dominant factor affecting the performance is network partitioning. Similarly, the default highest transmission range has similar behavior since both transmission range and carrier sense range are same.

The difference between the transmission range and carrier sense range is illustrated in Figure 6, which shows a single end-to-end (*i.e.*, multi-hop) transmission's effect on the network noise floor. The darker shade in the figure represents the transmission range while the lighter shade represents the carrier sensing range. Figure 6(a) shows standard behavior corresponding to broadcast packets being sent at the longest transmission range (*i.e.*, 1Mbps). Figure 6(b) illustrates the effect of transmission rate control. By reducing the transmission range we also reduce the active neighbor count. However, because power is unaffected, it still has the same carrier sensing range as that of Figure 6(a) thus it stops an equal number of nodes from transmitting. In Figure 6(c), using power control not only reduces the transmission range but also changes the carrier sensing range, allowing more nodes to transmit.

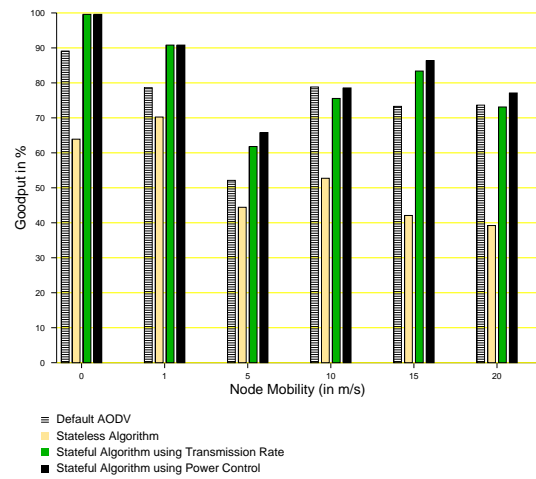
The impact of each mechanism, power vs. rate, depends on a number of factors including node density, node mobility, environmental factors and so on. This simple example illustrates that two differing mechanisms can have similar impact on performance. To fully compare the impact of both mechanisms, we turn to an in-depth simulation study of an *ad hoc* algorithm modified to use "local" routing.

3.2 Modifying AODV to use “local” routing

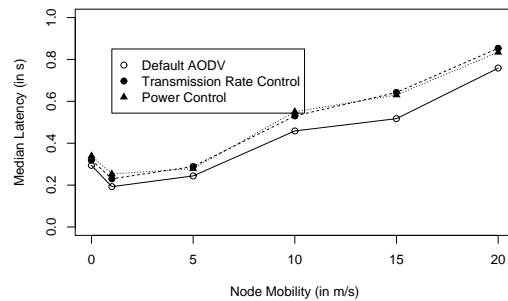
Several researchers have investigated using transmission power control to improve throughput in *ad hoc* networks [19, 20, 21, 22]. The general goal of this research is to limit interference between multiple communicating nodes. This is important both for route requests and standard unicast messages as mentioned above. By limiting route requests, the number of route replies is also limited, thus reducing the large amount of routing overhead seen in high mobility environments where routes are frequently broken. We are able to provide overall greater capacity by using the “locally” adverb to limit transmission power.

To realize the “locally” adverb we added cross-layer modifications to the AODV protocol. The routing layer adjusts the transmission rate or transmission power of the route request packets; as before, unicast packets use the RBAR mechanism to automatically select an optimal transmission rate for a specific link. We developed two variants of the AODV protocol using “local” routing. In the “stateless” mechanism, nodes forwarding a broadcast packet simply use the same transmission state (range or power) used by the original node; this resulted in overall poor performance. In the “stateful” mechanism, each node remembers what transmission range or power was needed for successful transmission. Both the methods can selectively use transmission rate or transmission power as means to achieve “local” routing. The ns-2 network simulator was used as the test bed for our modifications to AODV.

The results of our trials are compared to the default AODV settings which use maximum transmission rate and power. We found that reducing the transmission range, using either mechanism, without adaptation partitions the network. Figure 7(a) compares the goodput achieved with the “stateless” and “stateful” algorithms against the default AODV performance. The figure shows the end-to-end goodput achieved with different node mobilities. The “stateless” scheme has uniformly poorer behavior. On further inspection of the simulation traces, it was observed that re-initiating route discovery at the lowest transmission range causes needless route requests to be sent. The problem becomes quite dominant in pathological cases where the next hop can only be reached at the highest transmission range. Consequently, our “stateful” algorithm introduced the notion of soft-state such that every node remembers the transmission rate of the last successful packet. Hence, one can re-initiate the route discovery at a more appropriate range. Using transmission power control to route “locally” instead of transmission rate gives slightly better performance because reducing power limits the carrier sense range as well as limits the number of nodes that can respond to a route request. Figure 7(b) summarizes the corresponding effect on the latency. Using the stateful algorithms increases the latency compared to the base-



(a)



(b)



(c)

Figure 7: Figure (a) compares the “stateful” and “stateless” algorithms using both transmission power control and transmission rate with the default AODV. The “stateless” algorithm leads to degradation in throughput. The “stateful” algorithm performs significantly better. Figure (b) shows the corresponding effect on the latency. Figure (c) shows the how overall throughput is affected by the two range control mechanisms.

line AODV implementation because the stateful algorithms must spend time adapting to traffic conditions. The increased latency is balanced by the increased goodput. Although the stateful mechanisms both spend time adapting, that extra time results in less routing overhead and improved goodput.

In our approach, we have used transmission power control and transmission rate independently. We saw that using transmission power control reduces the noise floor as it decreases the carrier sensing range. However, adapting transmission rate also has the additional benefit of using higher bandwidth links. This is illustrated in Figure 7(b), which shows the throughput achieved as the message injection rate is increased. The transmission range for this experiment was set at that of an 11Mbps link. Thus, we see that by using transmission rate control we can achieve higher throughput by using higher capacity links.

We believe that this simple experimentation shows the promise of using our abstraction. The notion of “broadcasting locally” improves performance for an existing routing protocol. More importantly, when the lower layers in the protocol stack are free to choose one range control mechanism over another, the overall throughput of the network can be improved without encoding knowledge of the lower layers in our command and configuration interfaces.

4 Related Work

The properties of wireless networks make porting of the traditional protocols such as Transmission Control Protocol (TCP) difficult. While TCP is carefully calibrated to overcome the problems of stability and congestion control, wireless architectures introduce new challenges such as network partition and link failure due to mobility as well as different error characteristics. For example, traditional TCP error control is centered on congestion losses and ignores the possibility of transient random errors or temporary “blackouts” due to hand-offs and extended burst errors that are typical in wireless networks. As a result, different cross-layer approaches were introduced to overcome the deficiency of traditional TCP. A good summary of such techniques is given in [23, 24]. Balakrishnan [25] and Bakshi [26] explored different approaches including many cross-layer approaches involving the link layer. One such approach was the use of a *snoop agent* that monitors the traffic at the base station and caches the TCP segments. It retransmits the lost packet from its cache. This could be viewed as one implementation of a “*transmit reliably*” adverb, and could be combined with transmit power and modulation control.

All of these mechanisms, and more, can be classified and used to implement the more abstract notions of “adverbs” or modifiers on actions. The challenge will be to

have the different mechanisms be used at the appropriate time and at the correct level in the communication hierarchy.

5 Discussion

Our experiments have shown that control of transmission range (the *locally* adverb) may be achieved by either controlling transmission rate or power, but that these mechanisms have different effects. While the impact on goodput and latency is similar, the effect on *aggregate* throughput is markedly different. Utilizing higher transmission rates results in more efficient use of the spectrum and increased aggregate bandwidth. The experiments also showed that being able to route *locally* by either mechanism reduces in the number of routing messages generated, also improving the efficiency of the network.

From the perspective of the routing layer, sending a packet *locally* has the simple goal of reducing the number of control messages produced by limiting their scope of distribution. Since either mechanism achieves these goals, it is not productive for the routing layer to choose one over the other. We have shown that the routing protocol is better served by using an abstract interface specification, such as *locally*, which defers the decision to a lower layer that has more detailed knowledge of the hardware involved, its capabilities, and current network channel conditions. Such an approach was also suggested by Choudhury [2], specifically relying on the link layer to decide between two equally good routes. In our case, the choice of mechanism could depend on factors such as average message sizes in a flow. Message flows with large packets would use transmission rate control since that results in a higher bandwidth route. Flows with small messages could use transmit power control.

More importantly, as new RF interfaces become available the intent of *locally* will not change, although the physical mechanism to implement it may. Some hardware may be capable of large variations in data rate encoding but have very few power control levels, or vice versa. Tightly coupling the routing layer to such capabilities needlessly limits its ability to adapt to diverse hardware.

6 Conclusion

This paper is a first step toward defining a set of adverbs and adjectives suitable for flexible and intelligent utilization of available network resources. We have shown how it may be applied to *ad hoc* wireless network algorithms. Our survey of prior work indicates that most current cross-layer optimizations use limited information about the link layer and affect few physical mechanisms to control the RF layer. This implies that a careful “meta-protocol” design should

be able to define a framework that would provide sufficient information for these cross-layer algorithms. We are currently working on implementing such a framework on a networking research testbed, paying particular attention to the mechanisms made available by software radio.

7 Acknowledgments

The authors would like to thank Ashish Jain for his implementation of AODV and well as contributions to an earlier version of this paper.

References

- [1] Gavin Holland, Nitin H. Vaidya, and Paramvir Bahl. A rate-adaptive MAC protocol for multi-hop wireless networks. In *Mobile Computing and Networking*, pages 236–251, 2001.
- [2] Romit Roy Choudhury and Nitin H. Vaidya (UIUC). Mac-layer anycasting in wireless ad hoc networks. In *Second Workshop on Hot Topics in Networks (HotNets-II)*, 2004.
- [3] Siuli Roy, Dola Saha, S. Bandyopadhyay, Tetsuro Ueda, and Shinsuke Tanaka. A network-aware mac and routing protocol for effective load balancing in ad hoc wireless networks with directional antenna. In *Proceedings of the 4th ACM international symposium on Mobile ad hoc networking and computing*, pages 88–97. ACM Press, 2003.
- [4] Chenxi Zhu and M. Scott Corson. Qos routing for mobile ad hoc networks. In *Center for Satellite and Hybrid Communication Networks Tech Report - TR 2001-19*, 2001.
- [5] Prasun Sinha, Raghupathy Sivakumar, and Vaduvur Bharghavan. CEDAR: a core-extraction distributed ad hoc routing algorithm. In *INFOCOM (1)*, pages 202–209, 1999.
- [6] J. Li, C. Blake, D. DeCouto, H.I. Lee, and R. Morris. Capacity of ad hoc wireless networks. In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom 2001)*, Rome, Italy, July 2001.
- [7] S. Doshi, S. Bhandare, and T. Brown. An on-demand minimum energy routing protocol for a wireless ad hoc network. *ACM Mobile Computing and Communications Review*, 6(3), 2002.
- [8] Daji Qiao, Sunghyun Choi, Amit Jain, and Kang G. Shin. Miser: an optimal low-energy transmission strategy for ieee 802.11a/h. In *Proceedings of the 9th annual international conference on Mobile computing and networking*, pages 161–175. ACM Press, 2003.
- [9] R. Oliveira and T. Braun. TCP in Wireless Mobile Ad Hoc Networks. Technical report, University of Bern, Switzerland, March 2003.
- [10] S. Xu and T. Saadawi. Does the ieee 802.11 mac protocol work well in multihop wireless ad hoc networks? *IEEE Communications Magazine*, 39, 2001.
- [11] Gavin Holland and Nitin H. Vaidya. Analysis of TCP performance over mobile ad hoc networks. In *Mobile Computing and Networking*, pages 219–230, 1999.
- [12] W.H. Liao, Y.C. Tseng, S.L. Wang, and J.P. Sheu. A multipath qos routing protocol in a wireless mobile ad hoc network. In *Telecommunication Systems*, page 329347, 2002.
- [13] M. Pearlman, Z. Haas, P. Sholander, and S. Tabrizi. On the impact of alternate path routing for load balancing in mobile ad hoc networks. In *In Proceedings of the ACM MobiHoc*, pages 3–10, 2000.
- [14] Siuli Roy, Dola Saha, S. Bandyopadhyay, Tetsuro Ueda, and Shinsuke Tanaka. A network-aware mac and routing protocol for effective load balancing in ad hoc wireless networks with directional antenna. In *Proceedings of the 4th ACM international symposium on Mobile ad hoc networking & computing*, pages 88–97. ACM Press, 2003.
- [15] Jungmin So and Nitin H. Vaidya. Multi-channel mac for ad hoc networks: handling multi-channel hidden terminals using a single transceiver. In *MobiHoc '04: Proceedings of the 5th ACM international symposium on Mobile ad hoc networking and computing*, pages 222–233. ACM Press, 2004.
- [16] N. Jain, S. Das, and A. Nasipuri. A multichannel csma mac protocol with receiver-based channel selection for multihop wireless networks. In *IEEE IC3N, Phoenix*, October 2001.
- [17] A. Nasipuri, J. Zhuang, and S. Das. A multichannel cs-mamac protocol for multihop wireless networks. In *Proc. of IEEE Wireless Communications and Networking Conference (WCNC'99)*, September 1999.
- [18] Atheros. <http://www.atheros.com>.
- [19] Y. Tseng, Y. Chang, and B. Tzeng. Energy-efficient topology control for wireless ad hoc sensor networks. In *Int. Conf. Parallel and Distributed Systems (ICPADS 2002)*, 2002.
- [20] Ram Ramanathan and Regina Hain. Topology control of multihop wireless networks using transmit power adjustment. In *INFOCOM (2)*, pages 404–413, 2000.
- [21] Alaa Muqattash and Marwan Krunz. Power controlled dual channel (pdc) medium access protocol for ad hoc wireless networks. In *IEEE INFOCOM'03 Conference*, April 2003.
- [22] Roger Wattenhofer, Li Li, Paramvir Bahl, and Yi-Min Wang. Distributed topology control for wireless multihop ad-hoc networks. In *INFOCOM*, pages 1388–1397, 2001.
- [23] Vassilis Tsaoussidis and Ibrahim Matta. Open issues on TCP for mobile computing. Technical Report 2001-013, Northeastern University, March 2001.
- [24] Anurag Kumar. Comparative performance analysis of versions of TCP in a local network with a lossy link. *IEEE/ACM Transactions on Networking*, 6(4):485–498, 1998.
- [25] Hari Balakrishnan, Venkata N. Padmanabhan, Srinivasan Seshan, and Randy H. Katz. A comparison of mechanisms for improving TCP performance over wireless links. *IEEE/ACM Transactions on Networking*, 5(6):756–769, 1997.
- [26] Bikram S. Bakshi, P. Krishna, Nitin H. Vaidya, and Dhiraaj K. Pradhan. Improving performance of TCP over wireless networks. In *International Conference on Distributed Computing Systems*, 1997.

High Data Rate SATCOM On-the-Move for an Ambulance

Ivan Corretjer
Naval Research Laboratory
202-767-3677 phone
202-404-4050 fax
corretjer@nrl.navy.mil

Dave Minerath
Naval Research Laboratory
202-767-3540 phone
202-404-4050 fax
david.minerath@nrl.navy.mil

High data rate (HDR) communications between an accident/disaster scene and a trauma center can provide the EMS crew with assistance in severe trauma cases, enhancing the survival rate of the patients. Conventional SATCOM terminals are too large to be carried on a standard ambulance, which is height restricted by federal specifications. However, small-aperture terminals have beamwidths that may illuminate adjacent satellites, causing tracking difficulties and violating Intelsat interference standards for geosynchronous satellite usage.

In this paper we present the system design for a HDR SATCOM on-the-move system that addresses the conflicting operational and regulatory objectives of HDR SATCOM from small aperture terminals. The resulting system has been fielded on an ambulance as part of the University of Texas Health Science Center Houston DREAMS project. Successful demonstrations have shown the system to be capable of achieving a 1.544 Mbps full-duplex link from the ambulance to a hub site.

1. Introduction

Satellite communications (SATCOM) give network designers the capability to share data over long distances or to remote areas where traditional terrestrial network services are not available. The medical services community is currently developing technologies to connect their deployed EMS ambulances back to the trauma center. The technology and knowledge now exists that can greatly improve the chances of a trauma victim's recovery, provided that the knowledge and technology can be applied to the victim in a timely manner. Specifically, the Disaster Relief and Emergency Medical Services (DREAMS) Project sponsored by the University of Texas Health Sciences Center at Houston (UTHSCH) is bridging the gap between the accident scene and the initial evaluation by the trauma surgeon.

One of the goals of the DREAMS project is to develop medical technologies that allow doctors to perform remote triage using video, audio, and data links. Satellite communications can provide a vital link between the remote EMT crew and the trauma

center in these situations. By allowing the trauma center doctors to virtually assist the ambulance crew, some of the trauma center resources can be brought to bear on the patient's care, helping to stabilize the patient and increase the chances of a shorter and more complete recovery.

To support these types of applications, the ambulance satellite terminal will require a high data rate (HDR) on-the-move communications capability. The Ku-band commercial satellites offer these capabilities, but with stringent operational requirements. In this paper we present the challenges that motivate our system design as well as our results from fielding an HDR capable SATCOM on-the-move (SOTM) terminal.

2. Design Requirements and Problem Description

The DREAMS project objectives include HDR communications, minimal height extensions over the cab of the ambulance, and mobile on-the-move capabilities. The project objectives all offer conflicting design requirements when trying to operate a satellite terminal over commercial satellites.

Support of video, data, and other bandwidth intensive medical applications will require full-duplex megabit data rates. Current on-the-move systems, such as INMARSAT, only offer sub-megabit data services while higher data rate services, such as cellular, are only available in metropolitan areas. Other experimental or military on-the-move systems require specialized space segment resources [3,4]. In some cases half-duplex mega-bit data rates are available in these systems, but the mobile-to-hub (return) link is disadvantaged. For the DREAMS application it is this mobile user that requires the higher data rate return link. Ku-band commercial satellite services provide the ability to support these communications at commercially viable costs.

Another important design criterion to the medical community is ambulance height. GSA specification KKK-A-1822E defines the characteristics of an ambulance that may display the “Star of Life” symbol [1]. Paragraph 3.4.11.3 of the specification limits the height of the ambulance to 110 inches, “including roof mounted equipment, but excluding two-way radio antenna(s).” While a satellite antenna is technically a two-way radio antenna, the specification likely refers to flexible whip antennas that can flex when the ambulance drives under low overhangs. Thus the antenna must be as short as possible while still providing enough gain to support HDR capabilities.

The goal of keeping the doctor informed while the patient is en route requires continuous connectivity. Even in situations where the ambulance is under motion it is required that communications remain available. Use of an antenna stabilizing platform capable of maintaining pointing accuracy under vehicular motion is therefore required.

2.1. Small Aperture Antennas at Ku-band Frequencies

International Telecommunications Union (ITU) and Intelsat specifications [2] specify the radiation pattern envelope that must be met in order to operate over commercial Ku-band satellites. The acceptable radiation pattern envelope is described by Equation 1 and shown in Figure 1.

Equation 1 - IESS-601 Sidelobe Envelope [2]

$$G = 32 - 25 \log_{10} \theta, \quad 1 < \theta \leq 48$$

$$G = -10, \quad 48 < \theta$$

(Note that all gain values are expressed in dBi and all angles are expressed in degrees from the main lobe axis.)

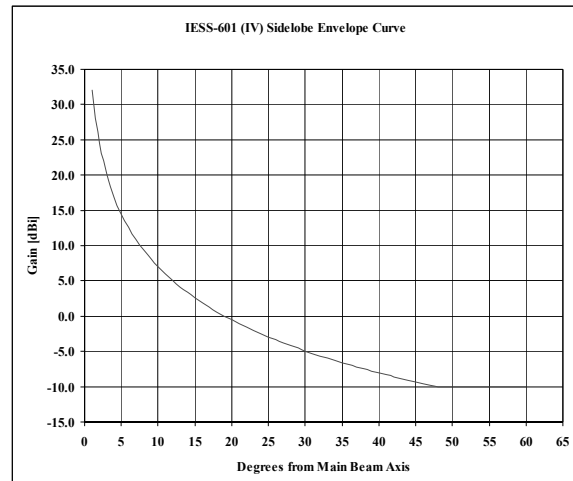


Figure 1 - IESS-601 Sidelobe Envelope Curve

In trying to approach the Star of Life height restrictions, a 60cm antenna was selected to keep the height of the ambulance at a minimum. Link budget calculations show that a 60cm antenna will support a HDR satellite link with a 3.7m antenna. However, when radiating a narrowband signal, the 60cm antenna would violate the standards and would not be approved for use on commercial Ku-band satellites.

2.2. On-the-move Motion Profiles

Motion studies were required to quantify how a road vehicle moves and vibrates over various road speeds and surfaces. This information was used to improve the ability to track an incoming signal.

Table 1 - Vehicle Roll (AVx), Pitch (AVy) and Yaw (AVz) (Nominal)

	AVx (deg/sec)	AVy (deg/sec)	AVz (deg/sec)
Max.	9.2	9.8	5.9
Median	0	0	0
Min.	-7.9	-8.8	-5.4
Mean	0.01	0.02	0.12
Std.			
Dev.	0.97	1.06	1.27

Key data points were the magnitudes and direction of vibration and vehicle turning rates. Two road courses were selected to provide typical motion data. The nominal course was selected to provide motion data over improved roads and the non-ideal to provide motion data over off-road conditions (See Tables 1 and 2).

Table 2 - Vehicle Roll (AVx), Pitch (AVy) and Yaw (AVz) (Off-Road)

	AVx (deg/sec)	AVy (deg/sec)	AVz (deg/sec)
Max.	25.1	21.7	30.6
Median	0.1	0	0
Min.	-14.9	-7.8	-10.4
Mean	0.01	0.03	0.66
Std.			
Dev.	2.88	1.97	4.49

3. Overall System Design

Figure 2 shows the system block diagram of the ambulance satellite terminal (AST). All components used in the system were selected as standard commercial off-the-shelf (COTS) equipment to both minimize the cost and need for custom equipment. Two exceptions to these design decisions were:

- Use of spread spectrum modems to allow the antenna radiation pattern to exceed the IESS sidelobe envelope without noticeable interference.
- Use of a customized tracking algorithm and hardware.

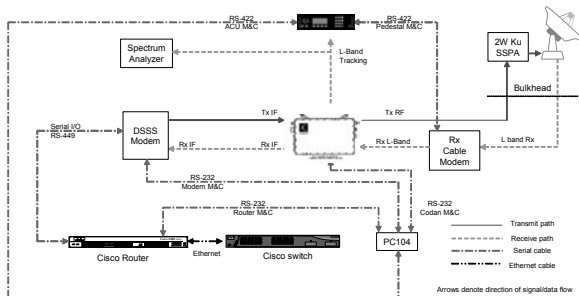


Figure 2 - AST System Block Diagram

3.1. Antenna Reflector Design

Assuming 50% efficiency, an ideal 60cm antenna operating at 14GHz has approximately 36dBi of gain with a 3dB beamwidth of 2.4 degrees. At the edge of the main beam, the antenna gain is 33dBi, well in excess of the 22dBi limit in the IESS specification. This limitation can be worked around by using a spread spectrum signal to reduce the power density per unit frequency to a level where an approved radiation pattern is not required. By using the spread signal, stray radiated energy would not cause noticeable interference on neighboring satellites. This has the added benefit of reducing the

carrier-to-noise ratio required for receiving the signal on a small antenna.

3.2. Pedestal and Tracking Algorithm Design

The original pedestal design was built for a marine platform. A stabilized antenna in a marine environment is designed around the motion characteristics of a waterborne vessel. Macroscopic ship motion is on the order of degrees per second and roll/pitch/yaw motion tends to be sinusoidal in nature. Ship's vibration, as well as heavy jarring as might be experienced in rough seas, is absorbed through the use of spring dampers and shock mounts. Mechanical control of the pedestal in such an environment was provided by the use of stepper motors with no motion feedback; when a command was sent to the motor to move, it was assumed to have moved as directed and the antenna's assumed position updated as appropriate.

Land-based vehicles have a very different motion profile. While vehicle motion on the three axes might have similar magnitudes to that of a ship, the angular rates and linear accelerations are significantly higher and likely to be outside the antenna control system's ability to compensate [5].

The tracking algorithm in the antenna control unit (ACU) was modified to utilize the closed-loop feedback provided by the upgraded pedestal servo motors with integrated encoders, replacing the open-loop control system that previously used stepper motors with no position feedback. Following successful modification of the antenna control system, system performance was verified to have been greatly improved.

3.3. Software System Manager

To provide a convenient and flexible interface to the developed AST for use by medical applications, an application programming interface (API) was developed. The application programmer is provided with a simple interface to the terminal that allows for control of all devices in the system, network connection setup/teardown, equipment configuration maintenance, and status monitoring. The monitoring function provides the medical applications a rating system regarding the health of the satellite link and when it is usable for HDR communications.

4. Results

Figure 3 shows the test ambulance with a 60cm antenna mounted forward of the patient compartment and over the cab.



Figure 3 - Ambulance with 60cm Ku-band antenna mounted over crew cab

The antenna radome extends over the patient compartment to enclose the entire antenna assembly. The antenna base was lowered to minimize the ambulance height; however, further lowering would affect the range of look angle elevations and/or azimuths caused by patient compartment blockage.

4.1. Radiation Pattern

The AST was tested at the Compact Range Antenna/RCS Test Facility at the Naval Research Laboratory in Washington, DC to determine the radiation pattern at Ku-band frequencies. A sample of the test results with the Intelsat mask overlaid is shown in Figure 4.

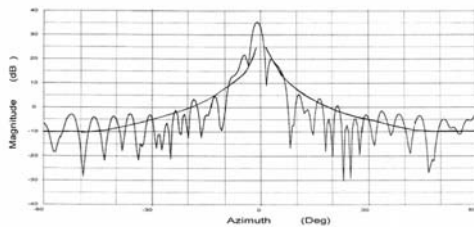


Figure 4 - AST radiation pattern at 14GHz

It can be seen that the 3dB beamwidth is only two to three degrees, with 15dB separation between the main lobe and first major sidelobes.

The effects of the wide beamwidth and relatively large sidelobes shown above are minimized by using the spread spectrum modem as the signal source. Testing with the commercial satellite provider, Panamsat, confirmed that the spread spectrum signal did not cause interference on adjacent satellites.

4.2. Antenna Tracking Performance

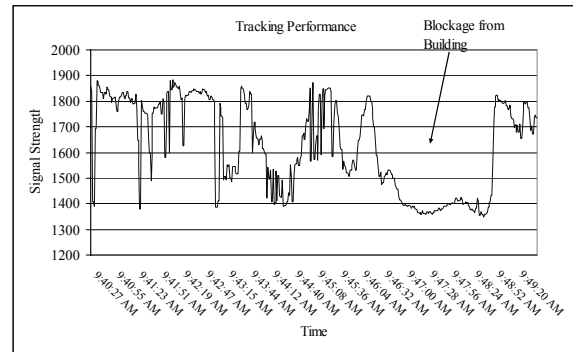


Figure 5 - Signal tracking performance nominal conditions

Figure 5 shows the relative signal strength and tracking ability of the modified terminal on a nominal road course. Typical signal variations under motion were 2-2.5dB on average. The large areas of blockage caused extended outages, but the use of servo motors with integrated position encoders allowed the ACU to know exactly where the antenna was pointed at all times. This greatly enhanced the ability to retarget the satellite when tracking was lost. Short periods of blockage from trees can be seen in the figure as sharp drops and rises in the received signal strength.

4.3. Network Integration

An AST test network was developed to provide Internet connectivity to the ambulance via the SATCOM link from Texas to NRL in Washington, DC (See Figure 6).

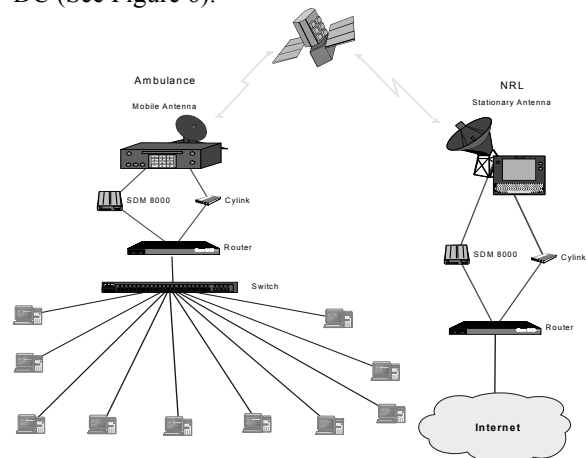


Figure 6 - AST test network

This network has been successfully used to test a satellite link between the ambulance and a hub station at data rates up to 1.544Mbps while stationary, and an asymmetric link (1.544Mbps

narrowband inbound, 300kbps spread spectrum outbound) in motion. The terminal is currently being fielded by UTHSCH as a development platform for the medical technology that will utilize the expanded capabilities offered by SOTM.

5. Conclusion

Land-based SOTM presents a unique environment for using a stabilized parabolic antenna. With the general vehicle motion profile understood and the pedestal control system modified, the antenna's ability to track under motion was greatly enhanced. Difficulties arising from the use of a small aperture antenna were successfully mitigated using spread spectrum technology. Use of the system in both mobile and stationary environments demonstrates that progress is being made in the development of land-based Ku-band SOTM technology.

6. Acknowledgments

We would like to thank the Disaster Relief and Emergency Medical Services Project sponsors from the University of Texas Health Science Center, and the Telemedicine and Advanced Technology Research Center. Funding was provided through the US Army Medical Research and Materiel Command, under contract: DAMD17-01-2-0054

7. References

- [1] KKK-A-1822E, Federal Specification for the "Star of Life" Ambulance
- [2] IESS-601, Intelsat Earth Station Standard
- [3] Dessouky, K. and Jedrey, T., "The ACTS Mobile Terminal (AMT)," American Institute of Aeronautics and Astronautics International Communication Satellite Systems Conference, Washington D.C., pp. 1758-1768, March 1992.
- [4] Sutton, F.S., Kim, J.H., Diener, D.F., Pounds, M.P., and Welling, R.E., "Wideband on-the-move satellite communications ground terminal," Proceedings of the Military Communications Conference (MILCOM), Vol. 2, pp. 770-774, Oct. 2001.
- [5] Ioakimidis, T.E. and Wexler, R.S., "Commercial Ku-band SATCOM on-the-move using a hybrid tracking scheme," Proceedings of the Military Communications Conference (MILCOM), Vol. 2, pp. 780-784, Oct. 2001.

Measurement of Weak Signals Using a Communications Receiver System*

Marc Rütshlin¹, Kate A. Remley, Robert T. Johnk, Dylan F. Williams, Galen Koepke, Chris Holloway
NIST Electromagnetics Division 818; 325 Broadway; Boulder, CO 80305

1. Corresponding author: ph. (303) 497 4674, email: rutschln@boulder.nist.gov

Andy MacFarlane, Mike Worrell
Phoenix Fire Department; Phoenix, AZ

We develop and characterize an inexpensive, reliable system for use in weak-signal detection. The system is implemented using widely available components, including a communications receiver and a computer sound card. Our characterization procedure allows the conversion of signals measured with the receiver system to electric field values. This enables comparison of measurements carried out on different receiver systems. The receiver system allows detection of signals up to several orders of magnitude weaker than is possible using handheld radio transceivers. This is of great use to the public safety community.

1. Introduction

A well-known problem facing first responders who rely on radio communications is the loss of signal in complex propagation environments such as large buildings, tunnels, basements, and collapsed structures. Reduced signal strength due to attenuation through building materials can significantly hamper communication. In the case of a dire emergency such as a collapsed building, the ability to detect a radio signal from a survivor may enable searchers to focus their efforts and may let the survivor communicate his or her status.

Here we describe a method that can be used to improve detection of weak signals by up to several orders of magnitude. The technique, sometimes known as joint time-frequency analysis [1,2], is particularly suitable for the detection of weak sinusoidal signals with time-varying frequency content such as those typically encountered in handheld radio communications. It has been used for years by ham radio enthusiasts, as well as in deep-space and other sciences that rely on weak-signal detection. Here we adapt the method to the unique needs of the public safety community where systems must be reliable, straightforward to implement, and easy to use in emergency scenarios. The system described here meets these objectives. Additionally, it is inexpensive and does not preclude the use of existing radio systems. At present, the method is limited to the detection of narrowband signals, meaning that its primary use is to determine only whether a radio signal is present and the strength of that signal rather than for voice communications.

A focus of this paper is the development of a characterization procedure that allows the calculation of the absolute electric field strength of received signals measured with the communications receiver system. This provides additional information on signal level for the operator, enables comparison of measurements that have been made on different systems, and makes it suitable for studying propagation in complex environments. Note that the procedure we discuss here does not increase the range of measurable signals; it merely allows us to calculate the absolute field strength.

Our approach in this paper will be to first describe the receiver-based measurement system, then to focus on the characterization procedure, which is divided into two steps: the quantification of the communication receiver's gain, and the measurement of the antenna factor. Finally, a case study implementing the system in the measurement of signal transmissions in a building will be described.

2. The Measurement System

The measurement system is shown in block diagram form in Fig. 1. It is based on a common communications receiver, which is used to downconvert a narrow band of radio frequencies, and a personal computer (PC) sound card, which is used to digitize this band of frequencies. The digitized frequency band is then amplified and/or graphically displayed, letting the operator know whether a radio signal is present and what the level of that signal is.

As shown in Fig. 1, the electric field corresponding to the signal is received by the antenna (labeled "handset" in the figure) and fed as P_{rec} to the RF input of a communications receiver (CR). The receiver is operated in its upper sideband (USB) mode at a frequency slightly below that of the transmitted signal (it is assumed that we have knowledge of that frequency).

* Partial work of the U.S. government, not subject to copyright in the U.S.

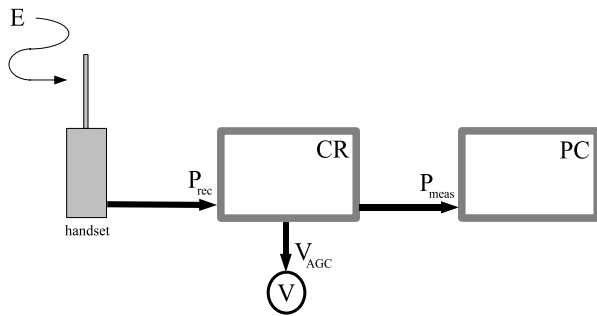


Figure 1: Block diagram of the measurement system. The electric field detected by the antenna (labeled “handset” in the figure) is related to the RMS voltage associated with the received power by an antenna factor: $E = AV_{rec}$, where $V_{rec} = (P_{rec}R)^{1/2}$; the power measured with the PC’s sound-card input is related to P_{rec} by a V_{AGC} -dependent power gain: $P_{meas} = G_p P_{rec}$, where $G_p = f(V_{AGC})$ and AGC is the automatic gain control of receiver.

In this way, the receiver operates as a simple frequency converter, down-converting an entire block of frequencies simultaneously. By setting the receiver’s frequency somewhat lower than the center frequency of the transmitted signal, the modulated received signal is converted down to the audio band. We may observe the upper and lower sidebands of the down-converted signal by setting the receiver’s

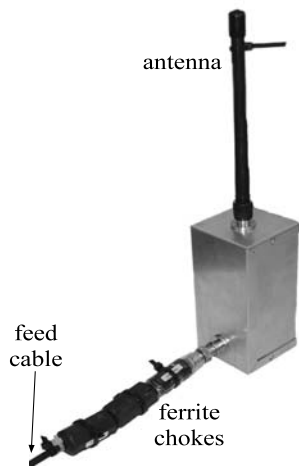


Figure 2: Photograph of the handset simulator used to develop the characterization. The receive antenna is of the helical type typically used with portable radio handsets. The effect of the feed cable is reduced through the use of ferrite chokes.

center frequency to approximately the middle of its passband. For example, a 100 MHz signal may be measured by a receiver with a 3 kHz passband by tuning the receiver to 99.9985 MHz. In this case, the receiver will display the 100 MHz signal at 1.5 kHz.

The received signal may consist of the unmodulated FM carrier or a frequency-modulated audio transmission. For public safety and other networked applications, the modulation may also correspond to a squelch tone. This sort of continuous tone is produced by many two-way radio handsets when the ‘push-to-talk’ button is depressed.

The down-converted signal is sampled by a sound-card connected to a PC running audio recording software. The communications receiver has an automatic gain control (AGC) circuit whose function is to control the receiver gain to produce a constant output signal regardless of the input power. In practice, the AGC is active only for signals within a certain power range, and does not modify weak signals (on the order of $P_{rec} < -90$ dBm). We monitor the level of feedback of the AGC, which is directly related to the input power, by measuring the voltage at the AGC jack on the back panel of the receiver at one-second intervals by use of a digital multimeter with a recording feature.

For the purpose of representing the signals measured by our system in terms of electric field strength, we developed a handset simulator, shown in Fig. 2, whose electrical properties emulate a typical handheld transceiver but is easier to characterize. A description of its construction is aided by the photograph shown in Fig. 2. It consists of an antenna attached to a metal box, fed through the box by a coaxial cable. Magnetic ferrite ‘chokes’ are placed near the point where the coaxial feed attaches to the box, disrupting common mode current flow and allowing the box to act as the second element of an asymmetric dipole, as it would in an isolated radio handset. The success of these chokes in removing the effect of the feed cable, at least at relatively low frequencies, has been demonstrated previously [3] and confirmed in our own tests. At higher frequencies, a narrow-band sleeve balun of the type proposed by Icheln *et al.* may also be used [4].

3. Receiver Characterization Procedure

Two steps are involved in the characterization procedure: first, apply a gain factor to convert the perceived power measured by the PC sound-card to the actual power at the RF input of the communications receiver; and second, use an antenna factor to convert this actual power to the electric field level present at the antenna.

The characterization procedure is carried out as follows:

1. Record the sound card input as a '.wav' file using commercially available audio recording software at the same time as the AGC voltage. The two measurements must be synchronized during post-processing. For long recordings such as field mapping in buildings, record time stamps corresponding to important events. For example, one might record the time of departure from a certain room. A sequence of this sort of time stamp allows for easier deciphering of the final recording.
2. Convert the signal to the frequency domain using successive N -point Fast Fourier Transforms (FFTs). Each FFT is carried out on a segment of the signal centered on a time that corresponds to when V_{AGC} was measured. The length of the FFT should be a power of two for greatest efficiency. A longer FFT will increase the frequency resolution of the results but will decrease the temporal resolution, which may cause loss of detail in a rapidly changing input.
3. Calculate the average power, P_{rec} , in the period chosen for the FFT by squaring and summing the magnitude of the frequency components over the frequency band of interest, as

$$P_{rec} = \frac{1}{G_p R} \sum_{i=\omega_1}^{\omega_2} |V_i(\omega)|^2, \quad (1)$$

where $V_i(\omega)$ is the root mean square (RMS) voltage of the i th spectral component, ω_1 and ω_2 are the lower and upper band-limiting frequencies, and R is the characteristic impedance of the system. The frequency band in (1) is chosen to incorporate as many transmitted signal components as possible. It will be limited by the communications receiver's IF filter bandwidth. The power gain, G_p , may be determined by the method described in Section 3.1. It is defined as

$$G_p = P_{meas} / P_{rec} = G_v^2 \quad (2)$$

$$G_v = V_{meas} / V_{rec}, \quad (3)$$

where G_v is the voltage gain that describes the ratio of the signal measured with the soundcard to the signal entering the

communications receiver. Each voltage in (2) and (3) is an RMS quantity associated with the relevant average power:

$$V_{rec} = (P_{rec} R)^{1/2}, \quad V_{meas} = (P_{meas} R)^{1/2}. \quad (4)$$

4. Obtain the electric field by multiplying V_{rec} by the antenna factor, A , derived in Section 3.2:

$$E = AV_{rec}. \quad (5)$$

Using (3), the electric field can be written simply as

$$E = A \frac{V_{meas}}{G_v}, \quad (6)$$

where V_{meas} is the voltage measured at the PC sound card port. The determination of the gain function G_v is described in Section 3.1, while the characterization of the antenna in a TEM cell with a characteristic impedance of 50 Ω is described in Section 3.2.

3.1. Gain Determination

The relation between the power measured with the sound card, P_{meas} , and that entering the communications receiver, P_{rec} , will be determined in this section. The setup of Fig. 1 is used, with the handset replaced by a signal generator that supplies a known signal. The V_{AGC} -dependent gain may be determined as follows:

1. Use a signal generator, or in our case a vector signal generator (VSG), to excite single-frequency signals over a desired range of power levels. This range is representative of the signal levels likely to be encountered in transmission scenarios where the set-up will be used. The minimum power level of the signal generator may be decreased by the use of attenuators.
2. The communications receiver output, operated as described above, provides the input to the sound card. After the signal generator output has been allowed to stabilize, the signal is recorded. Its spectrum is found using an FFT, and its average power, P_{meas} , is calculated from this spectrum. At the same time the AGC voltage is monitored and recorded.

- The ratio of measured voltage to input voltage, G_v , defined in (3), is plotted versus V_{AGC} to obtain a gain curve.

During a field measurement, the actual gain may be extracted by interpolating the curve based on the measured AGC voltage. We show gain curves at three frequencies in Fig. 3 for the particular receiver system that we characterized. Note how the gain increases fairly linearly with rising V_{AGC} , then flattens off when the voltage approaches its maximum of about 2.42 V. This maximum voltage is reached when the AGC is no longer active due to an insufficient strength of input signal.

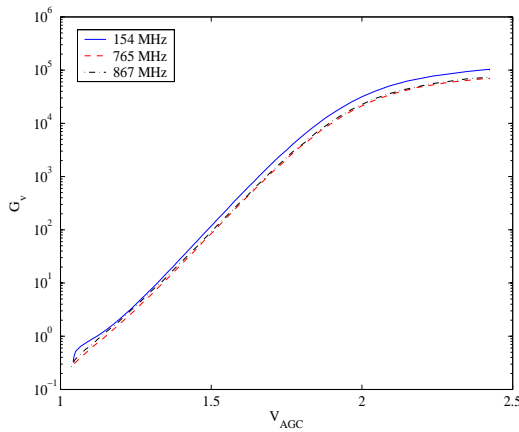


Figure 3: V_{AGC} -dependent voltage gain curves for the receiver system we characterized at three frequencies.

3.2. Antenna Characterization

A relatively simple way of characterizing the antenna is to determine its response to a known electric field. Such a field may be created to a good degree of accuracy—locally for physically small antennas such as ours—in a transverse electromagnetic (TEM) cell [5]. We used a broadband flared gigahertz TEM (GTEM) cell with a 50Ω characteristic impedance. In similar cells, measurements of small antennas’ gain factors have previously compared favorably to anechoic chamber measurements [6].

It is a simple matter to establish a specific potential difference between the conductors. Placing the antenna at the position where the plate spacing is x meters results in its exposure to an electric field of $1/x$ V/m if the potential difference between the plates is 1 V. The configuration we used is sketched in Fig. 4. A vector signal analyzer (or spectrum analyzer) is

used to measure the received signal as V_{VSA} , the RMS voltage corresponding to the measured power (equivalent to P_{rec} in Fig. 1). The antenna factor is thus defined as

$$A = \frac{E_{cal}}{V_{VSA}}, \quad (7)$$

where E_{cal} is the known electric field applied to the antenna.

Certain precautions are necessary when placing the antenna in the TEM cell so as to minimize impact on the field distribution in the immediate vicinity of the antenna. For example, we attach the feed cable to the antenna from ‘behind’, i.e. from the direction opposite to the TEM cell feed point, effectively ‘hiding’ it by attaching it to the lower conducting plate with adhesive conducting tape.

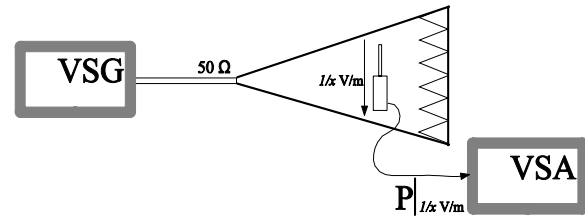


Figure 4: A block diagram view of the TEM cell antenna characterization. The TEM cell’s input is supplied by the VSG. The antenna, positioned such that it is exposed to a $1/x$ V/m field, has its output measured by a vector signal analyzer or similar instrument.

4. A Case Study

Our receiver-based measurement technique described above was utilized in conjunction with an independent study conducted by the Phoenix Fire Department [7]. Their aim was to compare the voice quality of radio transmissions at different frequencies and with different modulation schemes (e.g., analog, digital) in various typical building environments. The result is a subjective evaluation of the different schemes: individual communications between positions at which firefighters would typically be positioned are rated according to the criteria shown in Table 1. Our goal is to assign absolute field strength values to these subjective ratings.

Our aim was to attempt to match electric field strengths to the ratings (1-5) given in Table 1. The difficulty in comparing an objective quantity (the field strength) to a subjective rating will require additional study involving the firefighters who carried out the ratings. However as a

first cut, we can denote levels three, four, and five as “acceptable” communications, while levels zero, one and two will result in “unacceptable” communications. This division is borne out by the data shown in Fig. 5. These data, collected over a number of months by the Phoenix Fire Department, describes the perceived quality of voice transmissions over a wide spectrum of building types, at different frequencies and for different modulation schemes. There is a clear division between the ratings of 3-5, where a significant number of evaluations were made, and the ratings of 0-2, where very few were made.

Table 1: The criteria for the subjective evaluation of voice signal quality used by the Phoenix Fire Department in its study (from [7]).

Rating	Definition
0	No speech heard.
1	Unusable, speech present but unreadable.
2	Understandable with considerable effort. Frequent repetition due to noise or distortion.
3	Speech understandable with slight effort. Occasional repetition required due to noise or distortion.
4	Speech easily understood. Occasional noise or distortion.
5	Speech easily understood.

To investigate the link between these subjective ratings and absolute electric field strength, we first developed a map of signal strength in an eight-story building in Phoenix in which poor signal transmission quality had been observed in previous tests. A listening station—an implementation of the measurement set-up described in Fig. 1—was placed on the fifth floor of this building. Hand-held radios were set to transmit continually while being carried on a circuitous path through the building. At the same time, the radio bearers were regularly in voice communication with the listening station, allowing the quality of transmission to be judged and compared to the ratings from Table 1. Separate walks were done for transmissions at about 154, 765 and 867 MHz. Detailed notes were kept of the whereabouts of the transmitter as well as the signal quality, allowing the analysis shown below.

Since all results displayed the same trends, only the case for 867 MHz is discussed in detail here as a representative example. The measured electric field is shown in Fig. 6, plotted versus an ‘absolute time’ measured from the beginning of the walk. The

vertical marker lines and comments are based on the notes taken during the walk. Of particular interest are the comments regarding degraded signal quality: all occur when the measured electric field is below about 1×10^{-4} V/m, shown in Fig. 6 by the horizontal marker line. We assign this level to the “unacceptable” (levels 0-3) given in Table 1. This choice is somewhat arbitrary, since it is based on limited observations.

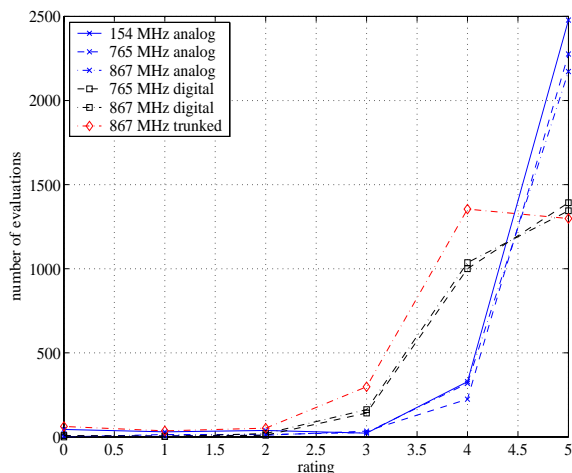


Figure 5: Sum of evaluations of radio transmission quality based on the rating descriptions in Table 1, reported for measurements of different modulation schemes at various frequencies in a wide range of buildings undertaken by the Phoenix Fire Department (based on data in [7]).

The route of the walk is roughly described by the comments along the top of the graph, e.g., ‘stair ascent’ refers to the transmitter being carried from the ground floor past the receiver position on the fifth floor to the roof. Note the correspondingly strong measured field just after the 300 second mark for this case. Access to the roof was not possible, so a brief circuit was walked on each floor before the sublevels were visited. As shown by the comments in Fig. 6, the measured field strengths were lowest when the transmitter was in the roof access hatch just above the eighth floor (at about the 600 second mark), and when it was in the parking garage sublevels (‘SL1’ and ‘SL2’). At these points the received audio quality was also worst, deserving ratings lower than 2 according to Table 1. A further aspect of the use of our receiver-based method for the detection of weak signals, indicated by the comments in Fig. 6, is that the carrier was visible even when the voice quality was very poor. This would appear to hold promise for alternative means of communication when voice transmission is difficult.

5. Conclusion

We have described a method, developed for the public safety sector, for detecting weak signals based on commercially available equipment. A focus of this paper was to calculate the absolute electric field strength from the measured signal. Use of electric field values allows the comparison of measurements made using different systems and makes the technique suitable for mapping signal propagation in complex environments. We applied the measurement technique to develop such a field-strength map in a large public building. Based on an assessment of audio quality, we assigned a field strength value below which communications were considered “unacceptable.”

6. References

- [1] S. Qian, D. Chen, “Joint Time-Frequency Analysis,” *IEEE Sig. Proc. Mag.*, pp. 52-67, March 1999.
- [2] R. Johnk, G. Koepke, D. Novotny, K. Remley, C. Grosvenor, M. Rutschlin, D. Williams, “Propagation measurements using computer soundcard methods,” *NIST Technical Note*, in process.
- [3] J. Demarinis, “The Antenna Cable as a Source of Error in EMI Measurements,” in *Proc. 1988 IEEE Symp. on EMC*. IEEE, pp. 9–14, 1988.
- [4] C. Icheln, J. Ollikainen, and P. Vainikainen, “Reducing the Influence of Feed Cables on Small Antenna Measurements,” *Electron. Lett.*, Vol. 35, No. 15, pp. 1212–1214, July 1999.
- [5] M. L. Crawford, “Generation of Standard EM Fields Using TEM Transmission Cells,” *IEEE Trans. EMC*, Vol. 16, No. 4, pp. 189–195, Nov. 1974.
- [6] P. Hui, “Small Antenna measurements Using a GTEM Cell,” in *Antennas and Propagation Society International Symposium*, 2003. IEEE, Vol. 4, pp. 715–718, June 2003.
- [7] Mike Worrell and Andy MacFarlane, “Phoenix Fire Department Radio System Safety Project,” *Final Report*, Oct. 2004.

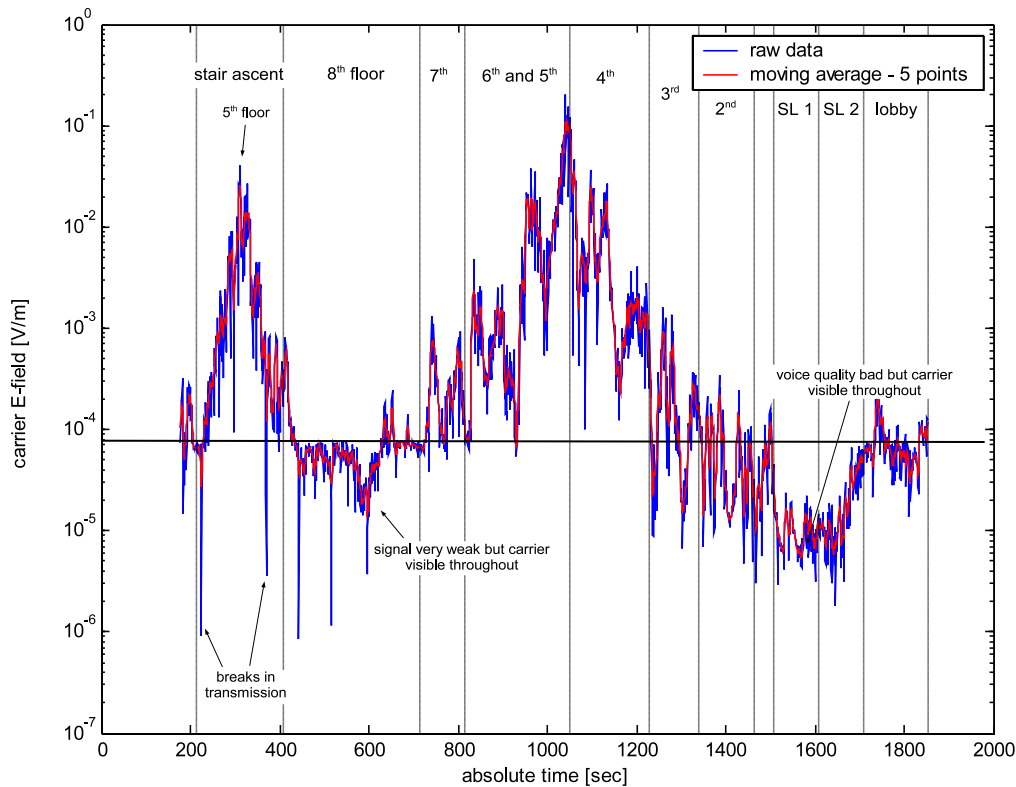


Figure 6: Electric field strengths measured during the 867 MHz building walk-through. The vertical marker lines represent boundaries in time between different sections of the walk. The horizontal marker at 1×10^{-4} V/m is the threshold below which poor signal quality, corresponding to ratings of 1 or 2 in Table 1, was observed. The lighter curve is a five-point moving average of the raw measured data shown by the darker curve.

Using a PCS Self-Interference Model to Evaluate the Effects of Cell Damage or Failure

Timothy J. Riley Teresa L. Rusyn
National Telecommunications and Information Administration
Institute for Telecommunication Sciences, Boulder, CO
303-497-5735 303-497-3411
triley@its.blrdoc.gov trusyn@its.blrdoc.gov

Using a self-interference model developed at the Institute for Telecommunication Sciences (ITS), the effects of system damage, load shifting, and increased traffic on personal communications service (PCS) systems can be studied, allowing emergency service providers to anticipate system availability and the need for supplemental emergency communications equipment. Currently implemented for systems following the ANSI/TIA/EIA 95B standard, this model characterizes the self-interference by producing a multiple-channel multiple-base station air interface signal with a variable number of base stations and channels per base station. The model also includes power control for individual signals, which is an important aspect of the interference level existing in the spectrum. The study of self-interference helps to develop more useful schemes to reduce the interference levels. The change in the aggregate air-interface spectrum is calculated and the resultant signal to interference (C/I) values can be used to determine the change in service quality and the probability of service availability. Any number of scenarios can be developed, depending on the configuration of the system under study and the level of detail desired. Since the model produces a cumulative baseband signal for both the forward and reverse directions, it can be implemented in a real-time hardware channel simulator, or as a component of higher-level software simulation and modeling. Predicted C/I values can be used in software-based network models to anticipate system limitations, traffic bottlenecks, and the probability of overall system failure. The baseband signal produced can be used to generate a simulated traffic signal for the testing and evaluation of commercial equipment. The model is particularly well-suited for independent PCS system evaluation by other Federal agencies, system manufacturers, and service providers.

1.0 Introduction

Historically, the U.S. Government has allowed its various agencies to develop their own communications systems independently and without oversight. In addition, there has been little effort to coordinate federal communication systems with state and local agency resources. The major philosophy in designing and acquiring communication systems was to closely tailor the system to the specific agency's needs and desires. The result has been a plethora of unique, one-of-a-kind, non-interoperable systems that are expensive and time-consuming to acquire, implement, maintain, and upgrade.

In the past, the problems that the lack of interoperability caused were apparent during responses to large-scale natural emergencies (tornados, floods, earthquakes, hurricanes, etc.). Initiatives to coordinate communications usually began following each emergency, but faded as the emergency receded in time. However, recent man-made disasters have strongly impressed the need for communication coordination across all levels (federal, state, and local) of emergency preparedness and response agencies and organizations. This follows attempts to wean agencies away from expensive, custom systems and towards less expensive, easily obtained, off-the-shelf, commercial equipment.

Federal, state and local emergency response agencies have begun coordinating efforts to develop standards and specifications for across-the-board communication systems to allow full interaction between all agencies under all circumstances. The support and usage of a common standard minimizes the development of incompatible equipment and eventually will result in lower costs, making such equipment available to small, less well-funded groups.

Until such standards are in place and compliant equipment is readily available and affordable to all, most smaller agencies (and to some extent, all agencies, regardless of size) are dependent on currently available, commercial communications equipment. Given the unpredictable nature of all disasters,

flexible and portable communications are required. In today's world, that points to cellular communications. Even when standardized emergency communication equipment is readily available, commercial communication services will still be of use to responding personnel due to their availability and flexibility.

With conventional wired communications, the *last mile*, the connection between the user and the system's main structure, is the most vulnerable to damage. Cellular systems are immune to that vulnerability and offer mobility to the user within the immediate service area, as well as across service boundaries. This is due to the design of the system, a network of small service areas, cells, centered around service providing equipment, the base station. The base stations of the various cells are inter-connected through various means (wire, fiber, terrestrial wireless, satellite) to the overall public switched telephone network (PSTN). The base stations and their connection to the PSTN are the most vulnerable to damage. However, two factors make cellular communications a more desirable form of emergency communications. The first is the relative ease of repair of damaged equipment and installation of new or temporary base stations offering cellular services. The second is the adaptability of cellular design to system outages and damage.

In the current economy, insufficient forethought is given to effective and efficient system design and implementation. A brute force approach is often taken when installing cellular services. What works in normal day-to-day situations may not work effectively, if at all, in an emergency situation, when traffic load volume and patterns change drastically. Add in random system damage, and large areas of an affected region may lose communication services altogether. Agencies that depend on commercial communication services may find themselves isolated and ineffective.

Prior planning, study, and evaluation of system design, capacity, reliability, and vulnerability is necessary to identify

necessary changes to the system to improve usability during emergency situations. To this end, the Institute for Telecommunication Sciences (ITS) is developing a self-interference model for cellular communication systems. While there are still some first-generation systems in use, they are steadily being replaced with second-generation (and beyond) technologies. The ITS model is aimed at second-generation systems and is being expanded to cover 2½ and third-generation technologies. For the purposes of this paper, we will concentrate on the version of the self-interference model based on the ANSI/TIA/EIA-95-B standard (hereafter referred to as 95-B).

2.0 Model Description

The model based on the 95-B standard [1] produces a representation of an instantaneous 95-B air interface signal. The signal can contain outputs of multiple base stations with variable numbers of channels for each base station and can assign relative power levels for each individual channel. Both forward and reverse link processes are included in the model, as shown in Figures 1 and 2.

The input for the model is a random data sequence, although there is no requirement that it be random. For forward link signals, the appropriate Walsh code and orthogonal I and Q short PN codes spread the input sequence. For reverse link signals, the model modulates the input sequence with Walsh codes and then spreads the sequence with long and short PN codes. The resulting I and Q data streams pass through a baseband filter and a quadrature phase shift keyed (QPSK) or an offset quadrature phase shift keyed (OQPSK) modulation scheme. The model calculates each channel signal contribution separately from all other channel signals and then adds the processed signal to the other signal contributions to form a composite output signal. The power level for a single channel (including an estimated power loss for the defined scenario) is included as a gain factor in the

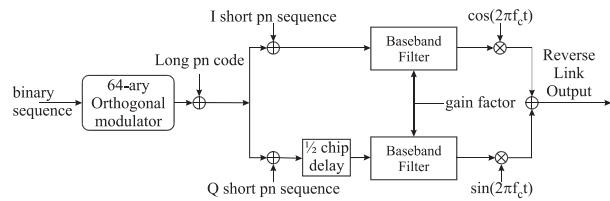


Figure 2. Reverse link model process flow diagram.

baseband filter which is set separately for each channel. All the Walsh and PN code definitions come from requirements in the 95-B standard [1]. The output of the model consists of a vector of numerical values representing a sampled QPSK or OQPSK signal.

There is no error correction added to the input sequence, only spreading codes and modulation processes are used. This model does not check for recovery information contained in the input. Its only purpose is to determine how well the system can transmit the bits of the input binary sequence.

Some characteristics of the 95-B air interface signal are:

- Peak amplitude does not change over a chip interval (½ chip for reverse link).
- The chip rate for both forward and reverse link signals is 1.2288 Mcps.
- The input symbol rate for the forward link is 19.2 ksps.
- The input symbol rate for the reverse link is 28.8 ksps.
- The limiting factor for system capacity is the interference level in the signal.
- Only four phase states, which are 90° apart, are valid.
- The signal can change to one of four phase states from chip to chip (one of three phase states for reverse link).

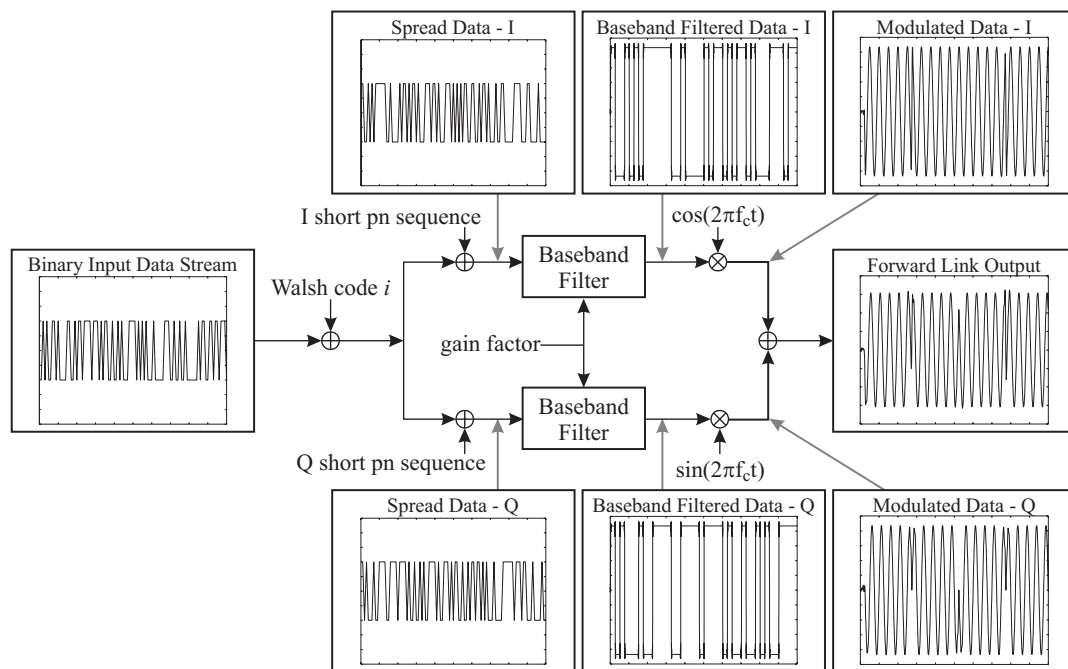


Figure 1. Forward link model process flow diagram.

3.0 Simplifications and Assumptions

The basic geometric framework for the model is based on earlier work performed at ITS [2, 3]. As shown in Figure 3, all cells are circular, cover equal areas, and are arranged in a hexagonal pattern with some overlap. Base stations are located at the centers of their respective cells. All mobile stations are stationary and have an arbitrary position in the cell. Positions of base stations and mobile stations are determined by a Cartesian coordinate system centered at the target cell's (cell 1) base station.

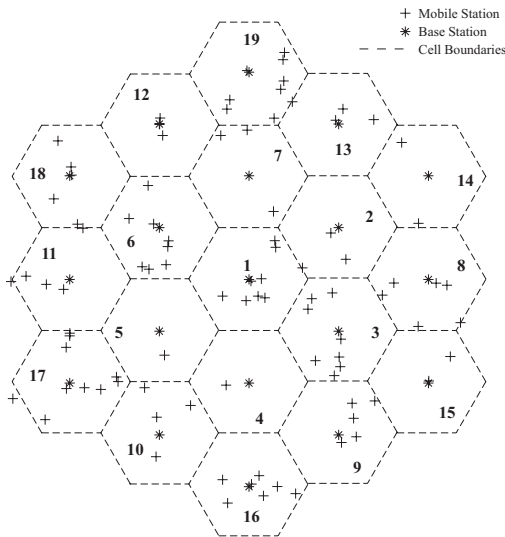


Figure 3. Fully-populated test scenario geometry.

All models of complex systems are simplified to some degree since including all the details of a system is often computationally infeasible. The ITS interference model has excluded many details of the 95-B system that have little or no effect on self-interference. Two simplifications already mentioned are the lack of any error correcting code, and the lack of any processing of the information in a data stream. Additional limitations for the model are [4, 5]:

- Terrain is not a factor (i.e., obstructions and multipaths are not considered in the channel model).
- Path loss can utilize any propagation model. For the sake of simplicity, the calculations presented in this paper will use a free space model with a factor of $r^{-2.5}$.
- All antennas are considered omnidirectional so that the patterns have constant gain and phase characteristics throughout the required field of view.

4.0 Software Validation

The ITS model is developed in software and much of the validation of the model is supported through software-based methods. Plots of the output of the model show physical characteristics of the signal such as the amplitude envelope and the phase changes. The signal's power spectrum shows components of the signal's bandwidth. Power spectra also can reveal approximate values for signal to interference ratio (SIR).

A software program was developed to demodulate and despread the output of the model. This program can recover

the original input binary sequence for both forward and reverse link outputs. The recovered signal has no errors for single channel and multiple channel (for low numbers of channels) signals. Multiple base station signals also can produce error-free data streams. As expected, increasing the number of code channels in the composite signal does increase the possibility of errors in a recovered signal.

Figure 4 shows a portion of the output of the model consisting of a single forward link base station signal containing channels for 5 mobile stations. The gain factor for the pilot channel is 1.76 dB and the gain factors for the traffic channels are all 0 dB. In this example, the carrier frequency is only 10 times the chip rate (12.5 MHz) for illustration purposes. The amplitude envelope for the signal is constant, after a transition period, over each chip interval and the phase changes occur at chip edges. A plot for a reverse link signal (not shown here) would have the same characteristics except that the amplitude would be constant over a $\frac{1}{2}$ chip interval and the phase would not include 180° changes as is possible in a QPSK signal. Figure 4 shows that the output signal of the model has the expected physical characteristics of a 95-B modulated signal.

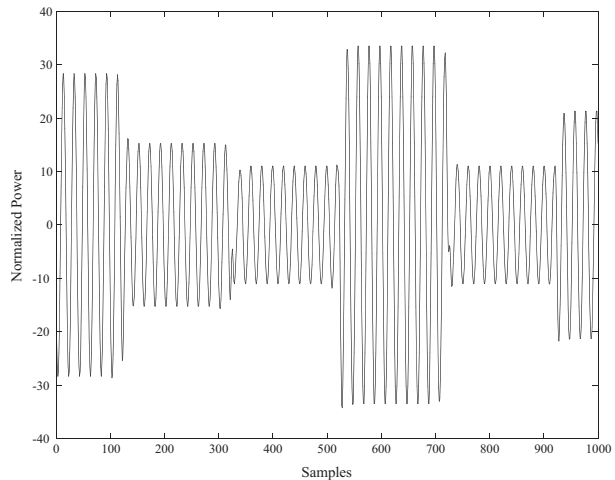


Figure 4. Multiple channel 95-B air interface signal – forward link.

Power spectra can show several characteristics of a 95-B signal. The modulated signal's spectrum will show both the carrier frequency and the bandwidth of the spread data stream. Spectra of the data stream, both with and without spreading, show the bandwidths of the respective data streams. The relative amplitude can be compared for an empirical evaluation of the processing gain of the system. The carrier frequency dominates the power spectrum of the modulated signal. The next dominant feature in a forward link signal is the 1.25 MHz bandwidth of the unmodulated spread data stream. The power spectrum for the reverse link would show the bandwidth for a half-chip interval. Figure 5 shows the power spectrum produced by the ITS model of a forward link signal with the carrier frequency set to 12.5 MHz. The peak of the signal occurs at the carrier frequency and a 1.25 MHz bandwidth signal is clearly visible.

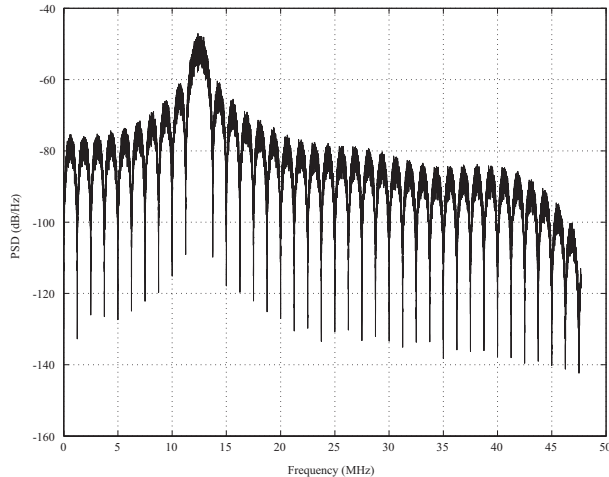


Figure 5. Power spectral density of a 95-B reverse link air interface-type signal.

Figure 6 shows the power spectral density (PSD) of a single channel data stream before and after spreading. There is no modulation included. The ~ 18 dB difference between the two plots for the first 10 kHz of the frequency range ($\frac{1}{2}$ of the bandwidth of the unspread signal) represents the empirical value for the SIR of this signal. This SIR value corresponds to an amplitude difference of 64 which is the processing gain expected for a 95-B signal, due to a Walsh code length of 64 bits.

The major limiting factor for 95-B system capacity is self-interference [4, 5]. Therefore, the SIR of a particular system is an indicator of how much capacity the system has. A software simulation which uses the output of the 95-B model compares SIR values to study the effects of interference levels on transmission performance. Figure 7 uses SIR values to illustrate how the interference power level changes as the number of channels in an air interface signal changes. The figure shows the SIR values for 1 base station with an increasing number of code channels in the composite signal. The code channel for mobile station #12 has a power

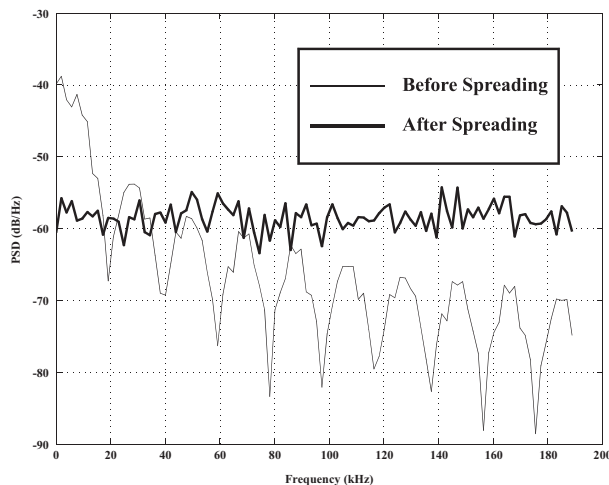


Figure 6. PSD of 95-B data stream before and after spreading.

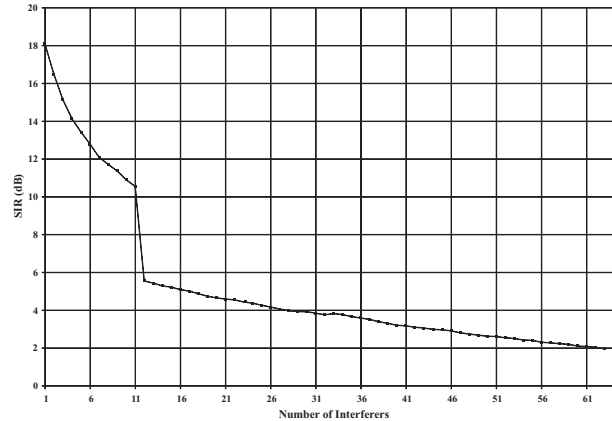


Figure 7. SIR (1 base station – mobile #12 with high power).

level which is five times that of the other mobile channels. Since the signal strength of the desired data stream does not change, the decrease in SIR shows how the interference level rises as channels are added to the composite signal [3].

5.0 Hardware Validation

The 95-B system uses a QPSK modulation scheme for the forward link and an OQPSK modulation scheme for the reverse link. After filtering and modulation in hardware, the model signal should have QPSK characteristics similar to those seen in the output of the model for software-based simulations. The envelope should be constant over a chip or $\frac{1}{2}$ chip period and the phase changes should be on chip or $\frac{1}{2}$ chip boundaries.

The output of the model is fed into an arbitrary waveform generator (AWG), filtered, and modulated. The two channels of the AWG are fed into a vector signal analyzer (VSA). The VSA demodulates the QPSK or OQPSK signal and plots the constellation diagram of the signal showing the amplitude-phase characteristics of the signal.

As shown in Figure 8, the forward link model signal has four phase states with transition lines connecting each phase state to the other three. The amplitude of the phase states is constant and the separation between the states is 90° . Figure 9 shows the constellation diagram for a reverse link signal. The signal has four phase states with constant amplitude and 90° separation. The major difference between Figures 8 and 9 is that there are no transitions through the center of the plot in Figure 9, a characteristic of an OQPSK signal as opposed to a QPSK signal. Both figures show some outlying nodes due to wraparound from the AWG or from discontinuities at chip edges when the phase changes.

6.0 Test Scenario

Figure 10 displays an example scenario showing the effects of system failure on the self-interference experienced by mobile stations in a cellular system (Figure 3). It consists of a target cell, surrounded by six immediate, secondary, cells (2-7). Beyond these are two layers of tertiary cells (8-13 and 14-19). All cells in the system are populated by a single, centrally located base station and a variable number of mobile stations which are randomly positioned within the

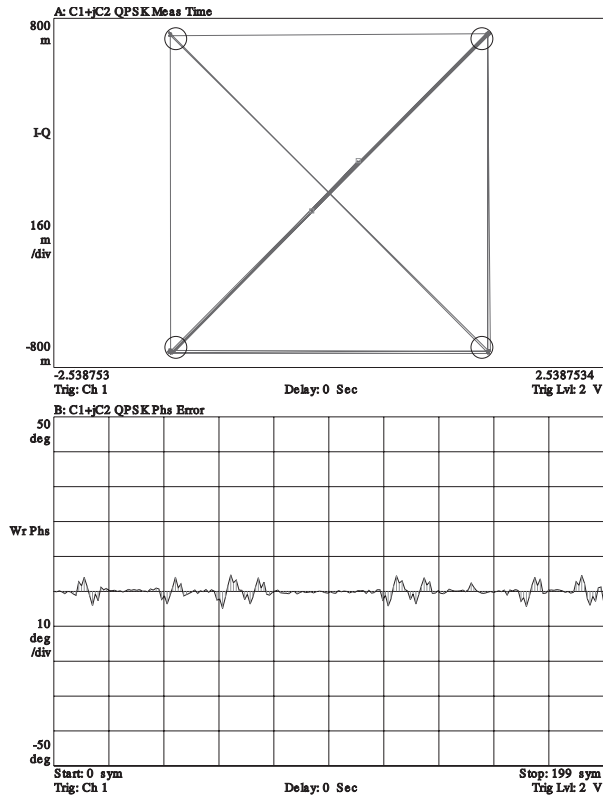


Figure 8. QPSK constellation and phase error plots.

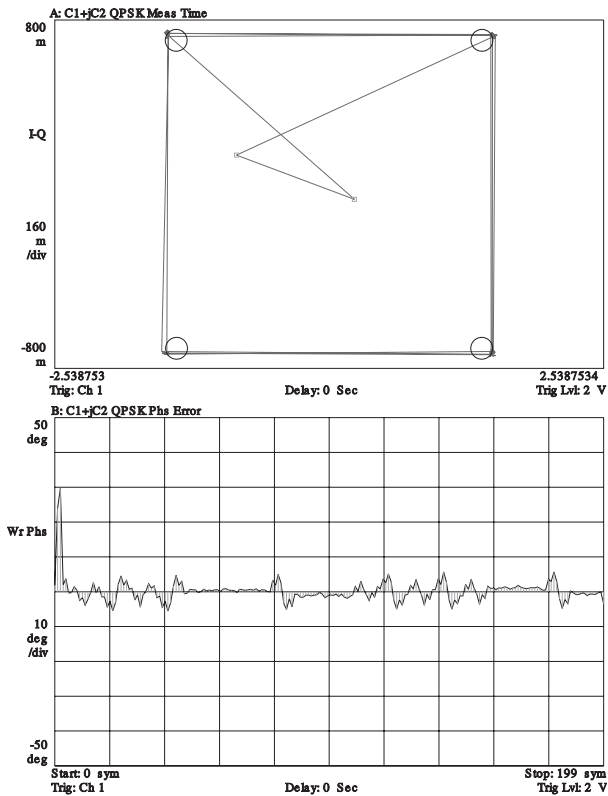


Figure 9. OQPSK constellation and phase error plots.

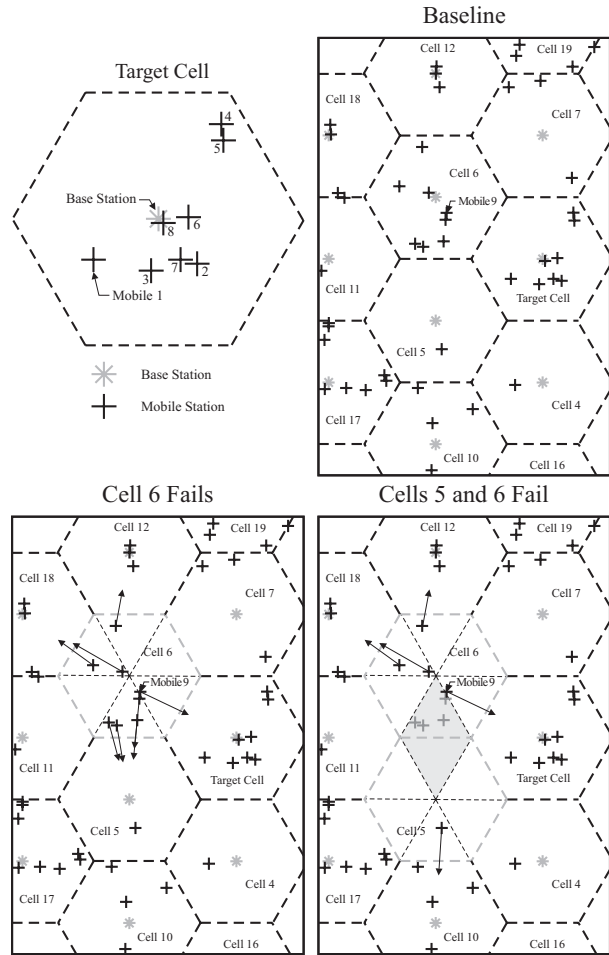


Figure 10. Example scenario.

cell. The target cell contains eight mobiles which are numbered as shown. Cell 6 contains eight mobiles, while cell 5 contains a single mobile.

For the purposes of clarity, each cell is drawn with a hexagonal shape. In reality, each cell's coverage is circular, given the assumption of lack of terrain (and structure) effects on coverage. Therefore, mobile stations located near, or on, the hexagonal boundaries between two cells, fall under the coverage of the base station of either cell. In more detailed scenarios, rules will need to be established regarding the determination of base-to-mobile relations.

Using the simplest scenario possible, all of the mobile stations are assumed to be stationary and no dynamics will be considered save one: the loss of base stations in immediately adjacent (secondary) cells.

In the baseline state, all base stations are operating and are servicing their respective mobile stations. When the base station of cell 6 is removed, the mobiles of cell 6 are picked up by their nearest base stations. In this situation, the target cell's base station picks up a single mobile (designated as mobile 9), cell 5's base station picks up four mobiles, while the remainder are picked up by other surrounding cells.

Continuing the scenario, cell 5's base station is removed. When this occurs, an assumption is made that the four mobiles cell 5 picked up from cell 6 (those located in the shaded, diamond-shaped area) are no longer in position to receive service from any adjacent cell's base station. They are removed from service and no longer affect the interference levels of nearby mobiles. Cell 5's single mobile is picked up by its adjacent cell, and the interference levels for the nine mobiles being observed are recalculated.

Several things happen when a base station ceases to function. First, the signal the base station produced ceases to interfere with the target base station. Then, communications between cell 6's mobiles and base station ceases, further reducing interference.

Due to the nature of the system's design, once a mobile loses contact with one base station, it tries to establish a connection with another base station in an adjacent cell. Under normal circumstances, this would occur when a mobile travels beyond the range of one base and enters the range of another. In this case, the mobile is stationary and it is the base station that has changed. In the first step of this scenario, it is assumed that all mobiles of an affected cell will be picked up by the surviving base station closest to it. In reality, not all mobiles will find themselves in a position that is covered by an adjacent base station, that is, the loss of a base station will result in a hole in the coverage pattern. In addition, the nearest base station, although physically able to communicate with the mobile, may have reached capacity with its existing mobiles and is unable to offer service to another.

If a mobile connects to a base station closer to the target cell's base station, the interference to the target cell's mobiles is likely to increase. If the new base station is further away, the interference is likely to decrease. If both the original base station and the new base station are roughly the same distance from the target cell or are both more than one or two cell sites away from the target cell, the interference levels are likely to be relatively unaffected.

Now that the scenario is established, it is time to decide what information is useful to determine the system's status or availability. Spectral usage, especially in the case of a spread-spectrum system, can tell little except in a gross sense. Figure 11 shows two spectra. The grey trace is the signal generated by the target cell's base station alone. The black trace is the cumulative spectra generated by all base and mobile stations in the area defined by the scenario. In a static system such as the one being considered, the spectra will not vary over time; in reality, they will vary significantly. We can consider the spectra in Figure 11 to be an instantaneous snapshot of spectral usage for the purpose of the current discussion. The spectra of the base station is the desired signal, while the difference between that signal and the cumulative signal is the interference the base station experiences. If this interference exceeds the level with which the base station's design can contend, then the base station's performance is negatively affected.

This format makes it difficult to evaluate the operating margin that the base station has to work with. A better way to present the data is to calculate the signal-to-interference ratio (SIR) under various conditions. A sample plot of SIR versus an

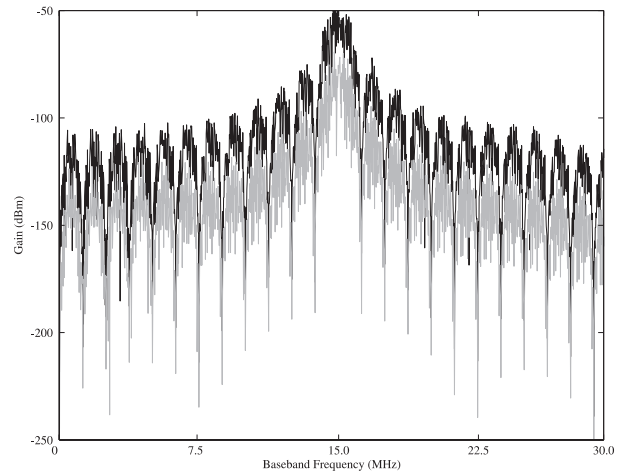


Figure 11. Comparison of target spectrum to fully-populated geometry spectrum.

increasing number of interferers is shown in Figure 12 (any station other than those of a base-mobile station pair is considered an interferer, even those mobiles in the immediate cell). The solid line shows the SIR for a fully populated system (19 base stations), while the dotted line shows the SIR when the base station in cell 6 stops functioning. The y-axis scale is rather narrow, indicating that the difference in SIR values is rather small. In some instances, there will be no effect on system performance. But if the system's level of operation is nearing its effective limit, the change in SIR could cause a detrimental effect on its availability and performance.

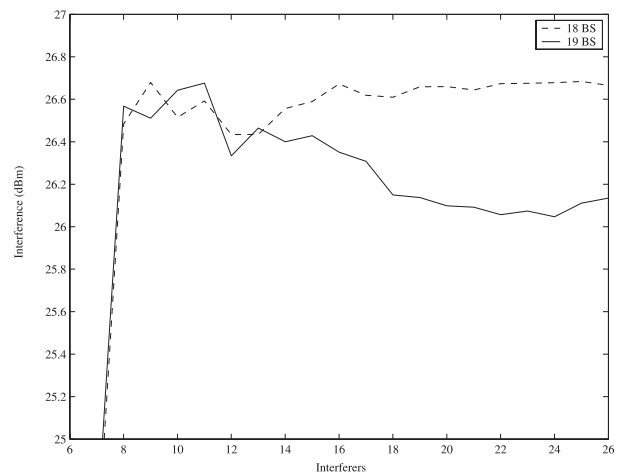


Figure 12. Change in SIR experienced by the target cell's base station due to the loss of a nearby base station.

The levels of interference experienced by the mobiles in the target cell also can be examined, compared and used to modify the scenario. Figure 13 is a plot of interference calculated at each of the eight mobiles in the target cell, along with the single mobile of cell 6 that is picked up by the target base station. Following the loss of cell 6's base station, the plot shows the increase in self-interference experienced by the mobiles. When cell 5's base station is removed, the interference is reduced due to the removal of the four

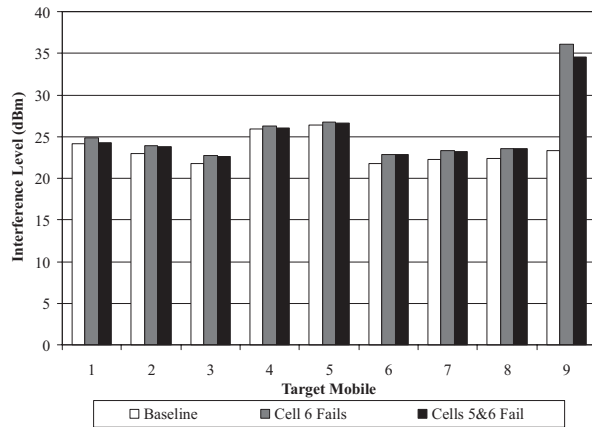


Figure 13. Interference experienced by the target cell's mobiles.

mobiles that lost their service. Mobiles 6 and 8 show almost no change in experienced interference due to their proximity to the target cell's base station.

The decision to remove the four mobiles in the second step of the scenario was made arbitrarily. A more realistic scenario would utilize SIR thresholds; a mobile would be removed from service once the SIR experienced by that mobile drops below a given threshold. Similarly, the capacity of a base station could be limited due to the level of SIR it experiences. The complexity of a scenario is limited by the amount of detail the user wishes to include and the computational power and time available.

7.0 Summary

The ITS self-interference model produces a sampled output representing an air interface signal. The output of the model can feed either a software- or a hardware-based simulation, and is able to support predictions of simple and complex channel scenarios. The degree of detail is limited only by the needs and intent of the user and the time available to produce the results, since more complex scenarios are more computationally intense. One level of detail that this paper purposely avoided was geometric dynamics. While many mobile stations are, in fact, stationary (either in actuality, or in terms of scale with respect to the overall cellular structure), those that are mobile will affect the performance of the system, especially in a situation where the system is not operating at an optimum level. In addition, mobiles that remain within a single cell's service area have less effect than those that cross cell boundaries and alter the load on the related base stations.

The basic geometry is highly stylized and generic. There is no reason that the geometry cannot be modified to reflect actual system implementations. While the basic propagation model used is the free-space model, the user can implement more realistic urban and rural propagation models such as Okumura and COST-231/Walfish/Ikegami. However, each modification introduces a new level of complexity and an additional requirement for information about the system being analyzed. This results in greater amounts of time needed to compute and evaluate the results.

Software shows that the output of the model has QPSK or OQPSK characteristics, including a constant amplitude envelope and phase changes. The original data stream can be recovered from the output of the model, although the SIR of the output signal decreases when the number of channels increases. Power spectra show the carrier frequency of the signal, the bandwidth of the signal, and the relative SIR values between signals.

Hardware phase measurements of the output signal show that the signals have the phase characteristics of QPSK and OQPSK signals. There are four phase states 90° apart. There are no 180° transitions in OQPSK. Phase error is less than 5° for the examples run in the validation process.

The benefit of the model lies in a more accurate understanding of the system's capabilities and limitations, and can lead to better design decisions in current system improvements, future system installations, and more accurate estimations in the need for emergency and backup communication equipment.

8.0 References

- [1] "Mobile Station-Base Station Compatibility Standard for Wideband Spread Spectrum Cellular Systems," ANSI/TIA/EIA-95B-1999, Telecommunications Industry Association, Arlington, VA, 1999.
- [2] J. G. Ferranto, "Interference simulation for personal communications services testing, evaluation, and modeling," NTIA Report 97-338, 1997.
- [3] T. L. Rusyn, "Co-channel Interference Modeling of the ANSI/TIA/EIA-95-B Code Division Multiple Access Cellular System," IEEE 2002 EMC Conference Proceedings, 2002.
- [4] S. C. Yang, *CDMA RF System Engineering*, Norwood, MA: Artech House, 1998.
- [5] T. S. Rappaport, *Wireless Communications Principles and Practice*, Upper Saddle River, NJ: Prentice Hall, 1996.