E2E Performance Tools: Internet2 Performance Architecture and Technologies Update

Eric L. Boyd

Director of Performance Architecture and Technologies

Internet2

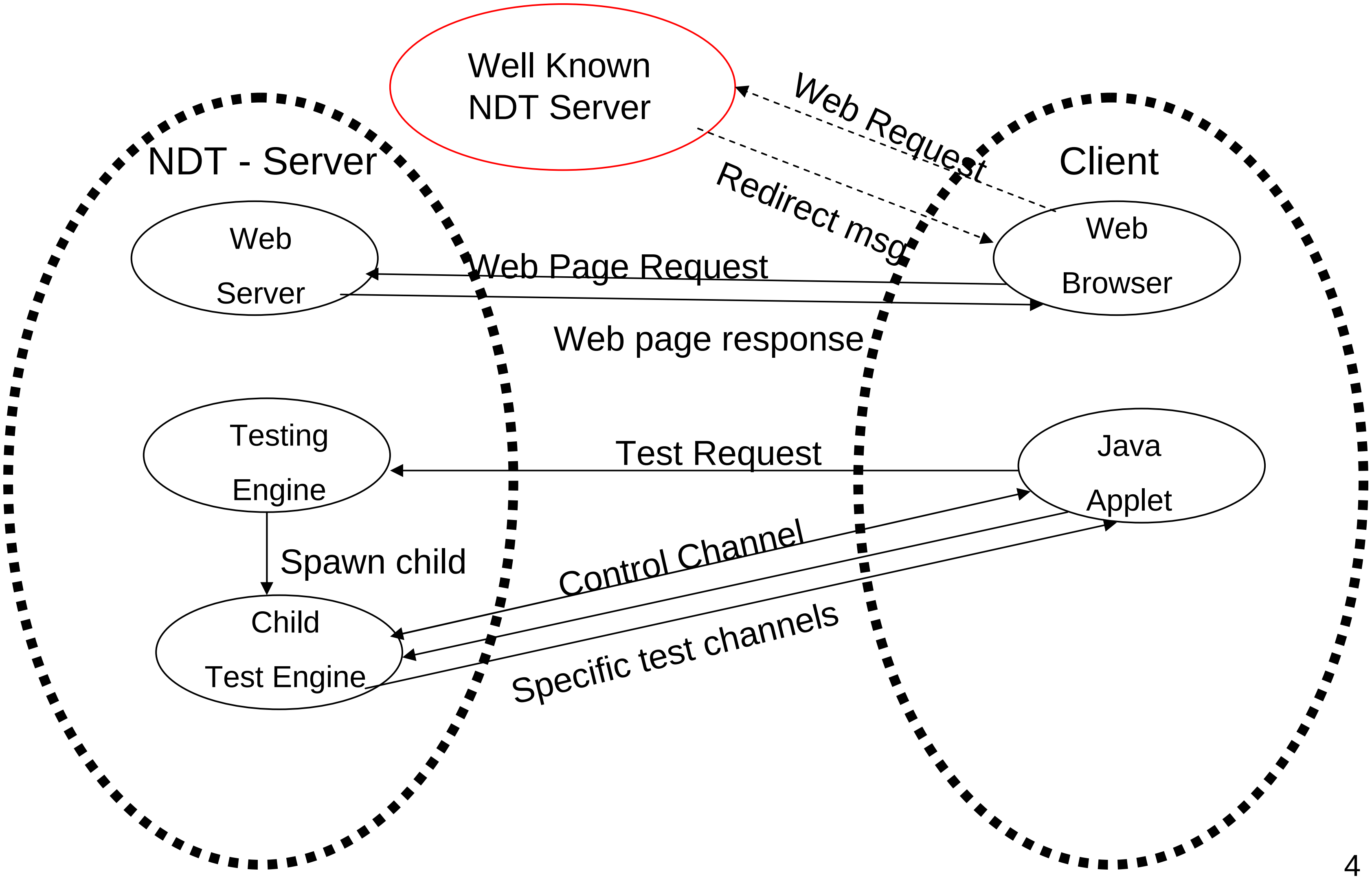
Current Projects

- Performance Tools
  - BWCTL
  - NDT
  - OWAMP
  - Thrulay
- Performance Measurement Framework
  - piPEs -> perfSONAR
  - GGF NMWG
- Member Outreach
  - Network Performance Measurement Workshops
  - Performance Tool Cookbooks
- Bulk Transport
  - Design Space
  - Prototype

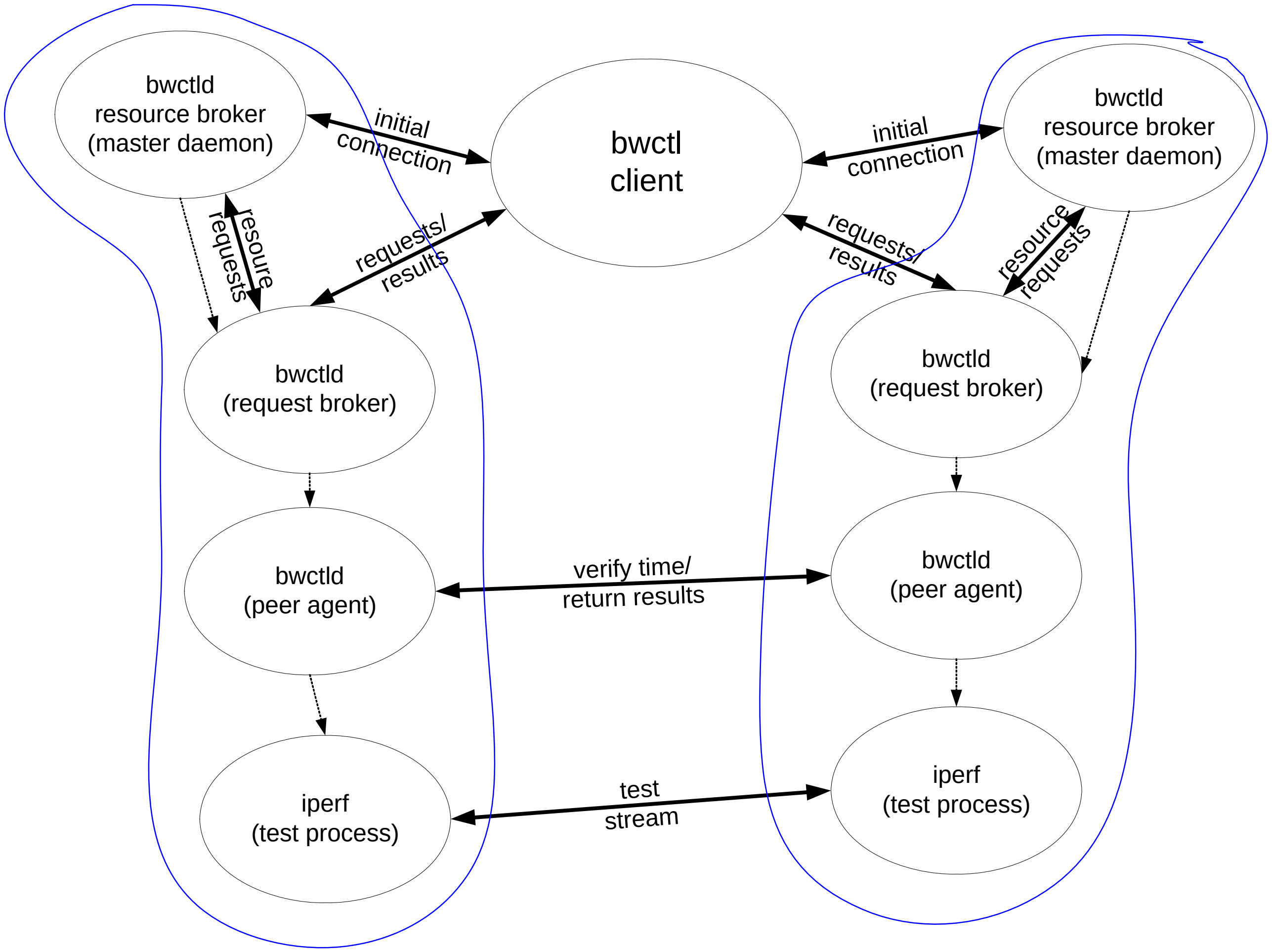
NDT: Network Diagnostic Tool

- Web100 enhanced server handles testing and diagnostic services
- Java based and command line clients allows testing from any client (local or remote)
- Performance and configuration faults reported back to client
- Drill-down functions provide more details & error reporting capabilities
- Grant from NIH/NLM to explore duplex mismatch detection

NDT Flow Diagram



BWCTL: 3-Party Flow Diagram



BWCTL (Bandwidth Controller)

- What is it?

  A resource allocation and scheduling daemon for arbitration of iperf tests

- Typical Solution
  - Run “iperf” or similar tool on two endpoints and hosts on intermediate paths
- Typical road blocks
  - Need permissions on all systems involved
  - Need to coordinate testing with others
  - Need to run software on both sides with specified test parameters

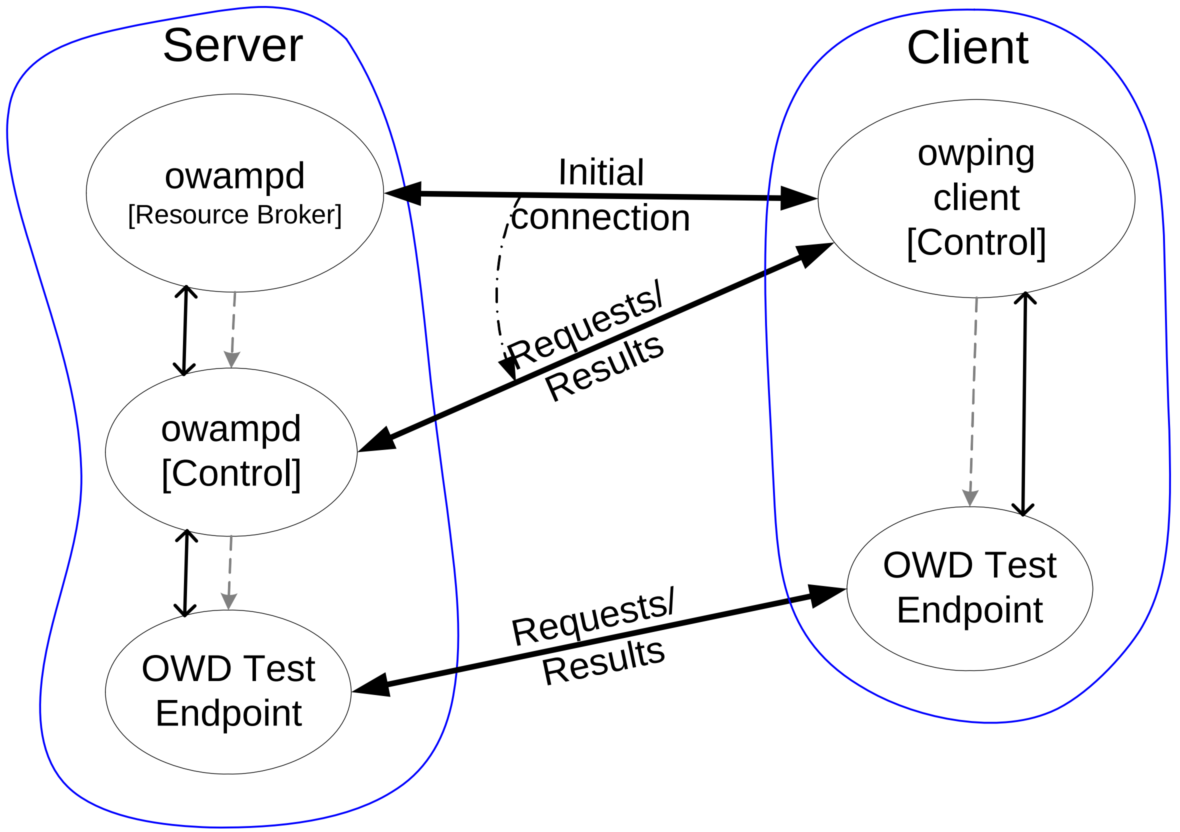
BWCTL (Bandwidth Test Controller)

- New version 1.2a
- Mostly bug fixes
- NTP requirement removed
  - Still best to use it
- Improved error reporting
- Solaris port
- OS X port

OWAMP: One-Way Active Measurement Protocol

- What is it?
  - Measures one-way latency: 1-way ping
  - Control connection used to broker test request based upon policy restrictions and available resources. (Bandwidth/disk limits)
- Specification
  - http://ietfreport.isoc.org/ids/draft-ietf-ippm-owdp-10.txt

OWAMP Flow Diagram



OWAMP

- LOTS of new deployments (Network Performance Workshop Attendees)
- New “developers” release to support latest version (14) of owdp spec
  - TTL (hop count)
  - Early terminated sessions handled more gracefully
  - Sender will skip sending “late” records and shares that information with receiver
- Public release this summer
  - Solaris
  - Incremental summary data from powstream (better database support)
- In the IESG, waiting for Security Review and IANA port number

Thrulay Overview

- Network capacity tester
- Same class of tools as iperf, netperf, nettest, nuttcp, ttcp, etc.
- Unique features not found in other tools:
  - measures round-trip delay along with goodput
  - output easy to parse by machine (gnuplot input format)
  - can send extremely precise Poisson stream of UDP packets

Thrulay Update

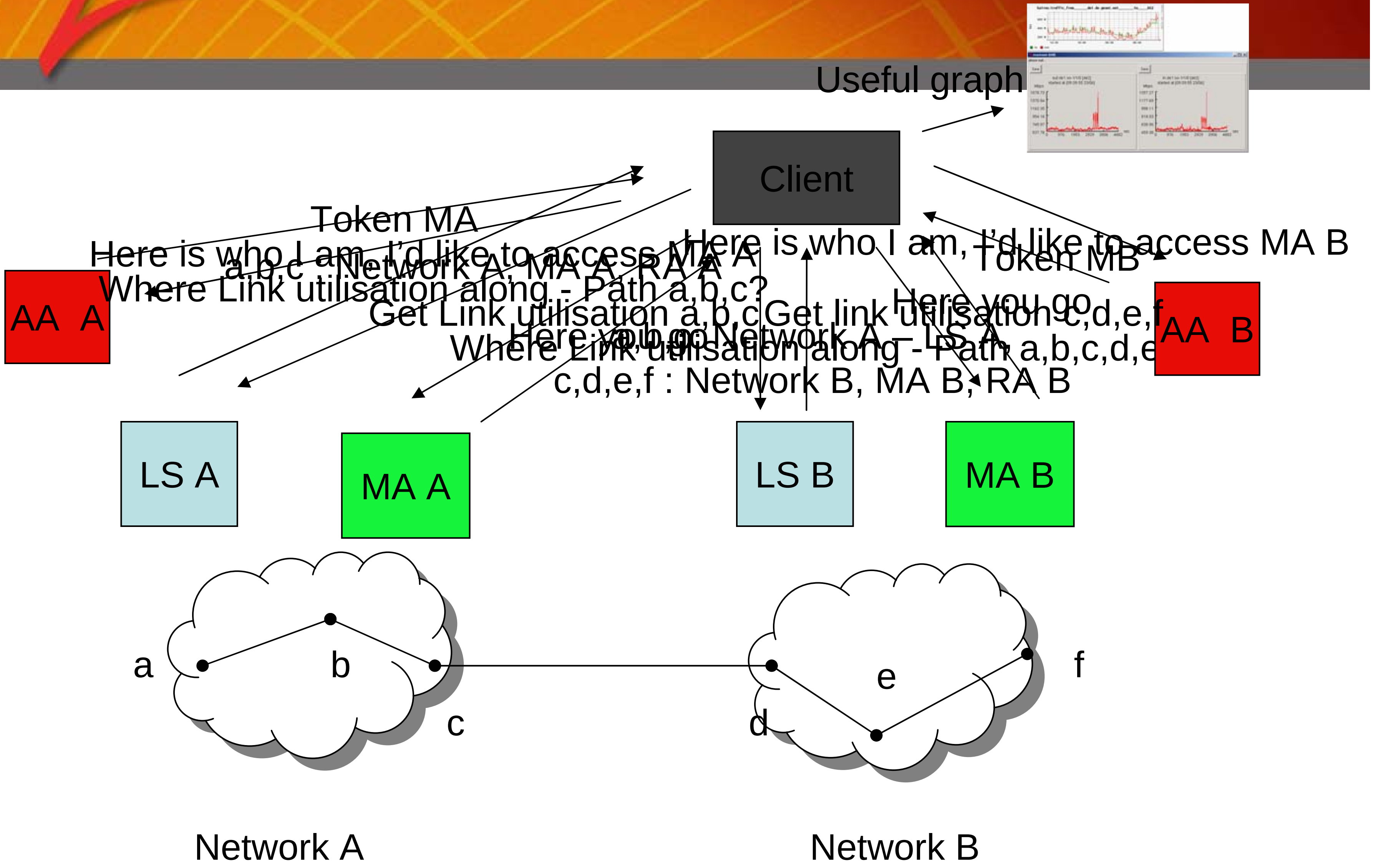
- New release v0.8
- Tests with multiple TCP streams
- Set DSCP (a.k.a. first 6 bits of the TOS byte)
- Report MTU and/or MSS (whichever the OS makes available)
- More UDP statistics: duplication, reordering, quantiles of delay
- SPARC/Solaris support
- Mac OS X support
- IPv6 support
- Non-busy-waiting UDP mode (less precise, but can run more concurrent tests)
- Documentation: manual pages have been added
- Basic client authorization based on IP address
- Integration of TSC timekeeping projects for faster and more precise timestamping

perfSONAR: Overview

- Joint effort of ESnet, GÉANT2 JRA1 and Internet2
- Herding cats or babysitting rottweilers?
- Webservices network performance framework
    - Network measurement tools
    - Network measurement archives
    - Distributed scheduling/authorization
    - Multi-domain policy

perfSONAR: Services (1)

- Measurement Point Service (MP)
- Measurement Archive Service (MA)
- Look-up Service (LS)
- Authentication Service (AS)
- Transformation Service (TS)
- Topology Service (ToS)
- Resource Protector Service (RP)



Useful graph

Client

Token MA

Here is who I am, I'd like to access MA A

a,b,c : Network A, MA A, RA A

Where Link utilisation along - Path a,b,c?

Get Link utilisation a,b,c

Get link utilisation c,d,e,f

Here you go

Here you go

a,b,c : Network A – LS A

Where Link utilisation along - Path a,b,c,d,e

c,d,e,f : Network B, MA B, RA B

Here is who I am, I'd like to access MA B

Token MB

perfSONAR: Services (2)

- Lookup Service
  - Allows the client to discover the existing services and other LS services.
  - Dynamic: services registration themselves to the LS and mention their capabilities, they can also leave or be removed if a service gets down.
- Authentication Service
  - Internet2 MAT, GN2-JRA5
  - Authentication functionality for the framework
  - Users can have several roles, the authorisation is done based on the user role.
  - Trust relationship between networks

perfSONAR Services (3)

- Transformation Service
  - Transform the data (aggregation, concatenation, correlation, translation, etc).
- Topology Service
  - Make the network topology information available to the framework.
  - Find the closest MP, provide topology information for visualisation tools
- Resource protector
  - Arbitrate the consumption of limited resources.

perfSONAR: Prototype

- Phase 0
  - Simplistic client which requests data to a MA (RRD filesystem) using web-services (we stand here)
- Phase 1
  - Include simplistic LS web-services (Static list)
  - Trivial AA – always say yes (need interface)
  - Visualisation
- Phase 2
  - Request additional data (OWD, packet drops)
  - Dynamic registration to LS
- Phase 3
  - AA handle attributes for other services
  - Distributed LS data across several domains
  - MP get's on-demand capability
  - Make use of the attributes to offer different functionalities to the users

perfSONAR: Deployment Plans

- Several networks have mention they would deploy the prototype phase1 (link utilisation and link capacity)
  - Abilene
  - ESnet
  - GARR
  - GEANT
  - GRNet
  - Hungarnet
  - RedIris
  - Uninett

perfSONAR: Accomplishments

- Architecture document (Fall ’04)
- Detailed Design document (Spring ’05)
- Workshops in Brussels (09/04), Zurich (04/05), Ann Arbor (05/05), and Poznan (08/05)
- Development Environment (05/05)
- Communications:
  - E2EMON submission (03/05)
  - TNC05 paper / presentation (06/05)
  - ICSOC05 paper (12/05)

perfSONAR: Progress

- Work up to early this year focused on a very detailed functional specification
  - Document deliverable for the EU
- This spring we worked on converting that to a more concrete design specification
  - XML schema defined for message communication
  - Java/Tomcat selected for prototype development
- This summer/fall we coded and coded …
- Prediction for winter: More coding!

perfSONAR: Current Work

- Current work is focused on developing a prototype that will allow interface utilization data to be shared.
- Relatively simple use case, but will demonstrate the feasibility of sharing data across multiple administrative domains

perfSONAR: Current Status

- Prototype is “done”, but …
- Code base still undergoing rapid change
- Current status:
  - Using Java/Axis/Tomcat/rrdjtool for rrd access

perfSONAR: Demos

- perfSONAR Demo in the demo room
  - Jason Zurawski, University of Delaware
- GGF and Supercomputing demos (10-11/05)
- Support EGEE demos (10/05)
- Participants who have deployed infrastructure over RRD files:
  - Abilene
  - ESnet
  - Geant
  - Other NRENs (PSNC, GRnet)
  - University of Delaware

What’s Next?

- Current Status:
  - Regular discussions
  - Development is underway
- We are at a key moment of the collaboration:
  - Distributed development process emerging
  - Making compromises between the vision and the technology
  - Where should we cut corners on the prototype?
  - Does the prototype form the basis of the deployed system?

Timing of Deliverables

- Prototype
  - Link Utilization (Abilene, ESnet, GÉANT, various European NRENs)
  - Generic service and interface
- Licensing and naming
  - Working name: perfSONAR
  - Working license: modified Berkeley
- Main services: MP, MA, LS, TS
- AA model to follow and policies
- Multi-domain AA integration

GGF NMWG

- Version 1 of the schema “all but done”
  - Employed by piPEs, Advisor, AMP, MonALISA, and SLAC
- Version 2 of the schema continues to evolve
  - perfSONAR work benefits from and informs this project

Network Performance Measurement Workshops

- Grow installed base of BWCTL/Iperf, OWAMP, and NDT at GigaPoP and regional campuses.
  - http://e2epi.internet2.edu/pipes/pmp/pmp-dir.html
- Begin integration into IT support processes.
- Create an installed base for perfSONAR deployment.
- Give each participant tool-specific cookbooks.

Network Performance Measurement Workshop Dates / Locations

- Completed
  - SOX / GaTech (03/05)
  - CENIC / UCLA (06/05)
  - JT – Vancouver (07/05)
  - OARNet / OSU (09/05)
  - MAGPI / FMM (09/05)
- Planned
  - MAX / College Park (12/05)
  - APAN (01/06)
  - JT - Albuquerque (02/06)
- Under Consideration
  - MERIT, Wisconsin, Alaska, …

www.internet2.edu

perfSONAR Open issues: AA (1)

- Authorization is based on role in group.
  - 4 "classes" of users: root, super, regular, untrusted
  - Default class is: regular - everyone that can authenticate gets this unless we specify something else.
  - As part of bilateral agreements, we may learn about projects at other institutions and specifically map users with those "project" attributes to another "class".
  - Likewise, we may map individuals who are part of projects "locally" to another "class".
- How do we deal with attributes?
  - Each network has it’s own attributes, how can we make things more common globally to minimize the complexity of bilateral agreements?

perfSONAR Open Issues: AA (2)

- AA system between now and full solution
  - Does Internet2 Middleware or GÉANT JRA5 have a central AA system (with the AA interface) that we could use and administer, so we don’t have to build it from scratch?
  - Does Shibboleth v1.3 (which implements SAML v2.0) meet our needs?
- Do the perfSONAR web services have the same “look and feel” as the AA interface?

perfSONAR Open issues: Deployment (1)

- How do we create a deployed base?
  - Critical mass deployment of tools (Underway)
    - Should Network Performance Measurement workshops be rolled out in Europe, FedNets?
    - What tool mix is appropriate in each administrative domain?
    - Target: GigaPoPs / NRENs? What about jointly tackling international application communities?
  - Critical mass deployment of measurement framework (TBD)
    - Should we jointly develop an Advanced Network Performance Measurement workshop to roll out perfSONAR?

perfSONAR Open Issues: Deployment (2)

- AA: What are our dependencies on deployment of AA infrastructure?
- Next Gen: Lightpath monitoring requirements?
  - DEISA will be using a lightpath, DANTE would like to provide them a monitoring infrastructure
  - Others?
- Security: How do we avoid creating “missile launchers”?

perfSONAR Open Issues: Bilateral Agreements

- Bilateral agreement: between any two entities (e.g. university, GigaPoP, NREN, backbone network)
- What should a bilateral agreement look like?
  - Agree on roles
  - Agree on what to measure
  - Agree on frequency of measurement
  - Agree on response to results
  - Can we “batch” agreements? (Can a measurement agreement between Internet2 and GÉANT make a bilateral agreement that covers an American university and a European university?)

Bulk Transport Tool (1)

- Killer App for High Performance Networks (i.e. why else do we need fat pipes)
- Remedies for TCP’s maladies
  - Tuning: buffers, window scaling, timestamps, SACK
  - Use multiple streams
  - Something Else
  - Replace the kernel and use different congestion control
  - Replace all the routers and kernels

Bulk Transport Tool (2)

- Many alternative TCP/IP congestion control algorithms
- Modified kernels are incompatible with regular kernel security patches
- Get the benefits of kernel-level modifications to TCP/IP congestion control algorithms in a user level tool, avoiding security issue with alternate kernels

Bulk Transport Tool (3)

- Design Space Document
- Early Stage Prototype

Google Summer of Code

- Google is "sponsoring" many students to work on open-source projects this summer.
  - Internet2 is mentoring 10 students. http://transport.internet2.edu/student-projects.html.
- Current Projects:
  - Timekeeping using TSC register - timestamp fetching without a context switch and relating the TSC value to UTC.
  - Noise calibration - data analysis of noise in delays for packet measurements and development of filtering algorithms.
  - Thrulay enhancements
  - Bulk Transport API over UDT
  - Rich Presence Project