CHEETAH: End-to-End Provisioned Network Testbed for eScience

Malathi Veeraraghavan, University of Virginia

Nagi Rao (raons@ornl.gov),
Bill Wing, Tony Mezzacappa, Oak Ridge National Laboratory

John Blondin, North Carolina State University

Ibrahim Habib, City University of New York

Sponsored by NSF EIN Program

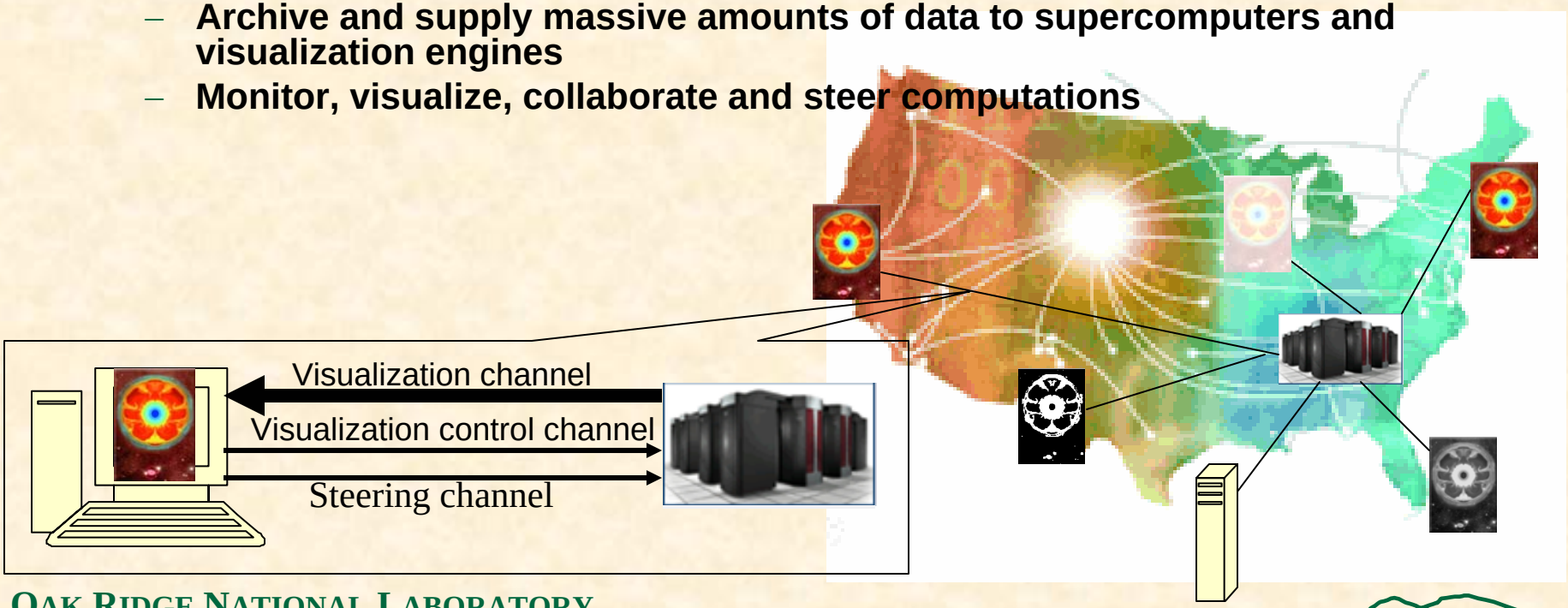
JET Roadmapping Workshop
April 13-15, 2004
Newport News, Virginia

Outline

- **Application Needs**
- **Project Details**
- **Preliminary Results**

Terascale Supernova Initiative (TSI)

- **Science Objective: Understand supernova evolutions**
  - **Department of Energy SciDAC Project: ORNL and 8 universities**
  - **Teams of field experts across the country collaborate on computations**
    - **Experts in hydrodynamics, fusion energy, high energy physics**
  - **Massive computational code**
    - **Terabyte in days are generated currently**
    - **Archived at nearby HPSS**
    - **Visualized locally on clusters – only archival data**
- **Desired network capabilities**
  - **Archive and supply massive amounts of data to supercomputers and visualization engines**
  - **Monitor, visualize, collaborate and steer computations**



Large-Scale eScience Needs (TSI and others)

- **Data Resources and Archives**
  - **Archive and retrieve massive data sets (Tera-Peta bytes)**
- **Visualizations**
  - **Stream, visualize and analyze massive datasets**
- **Computations on supercomputers and clusters**
  - **Archive and supply massive amounts of data**
  - **Monitor, visualize, collaborate and steer computations**
- **User Experimental Facilities**
  - **Setup, monitor and control experiments**
  - **Archive and supply experimental data**

They need revolutionary advances in network capabilities:

data transfers – petabytes at terabits/sec speeds

computational steering – real-time agile control

collaborative visualization – multiple synchronized streams with real-time control

instrument control – stabilized real-time control loops

Networking for Large-Scale eScience

A Promising Solution:

- **Provide application-level dedicated bandwidth channels to support**
  - **High bandwidth transfers**
    - **Simpler protocols – no congestion avoidance**
  - **Agile control operations**
    - **Easier stabilization and feedback control**
- **Challenges**
  - **Provisioning technologies**
    - **To be built using "gear" primarily designed for IP**
  - **Transport protocols**
    - **To be optimized for dedicated channel characteristics**
  - **Application immersion**
    - **Visualizations and computations must be tailored to and integrated with applications**

Project Details

- **Objective: Develop the infrastructure and networking technologies to support a broad class of eScience projects and specifically the** Terascale Supernova Initiative
  - **Optical network testbed**
  - **Transport protocols**
  - **Middleware and applications**
- **Sponsor: National Science Foundation**

  NSF EIN: Experimental Infrastructure Network

  Title: CHEETAH: Circuit-switched High-Speed End-to-End Transport ArcHitecture

  Project: Jan. 2004 – Dec. 2007

  Award: $3.5M

  Institutions: University of Virginia, North Carolina State University, Oak Ridge National Laboratory, City University of New York

Project Team

- **Astrophysics computations**
  - **Mezzacappa (ORNL) and Blondin (NCSU)**
- **Provisioning technologies**
  - **Habib (CUNY), Veeraraghavan (UVA), Wing (ORNL)**
- **Transport technologies**
  - **Veeraraghavan (UVA) and Rao (ORNL)**
- **Visualization support**
  - **Rao (ORNL) and Blondin (NCSU)**
- **Application immersion**
  - **Rao (ORNL), Blondin (NCSU) and Mezzacappa (ORNL)**

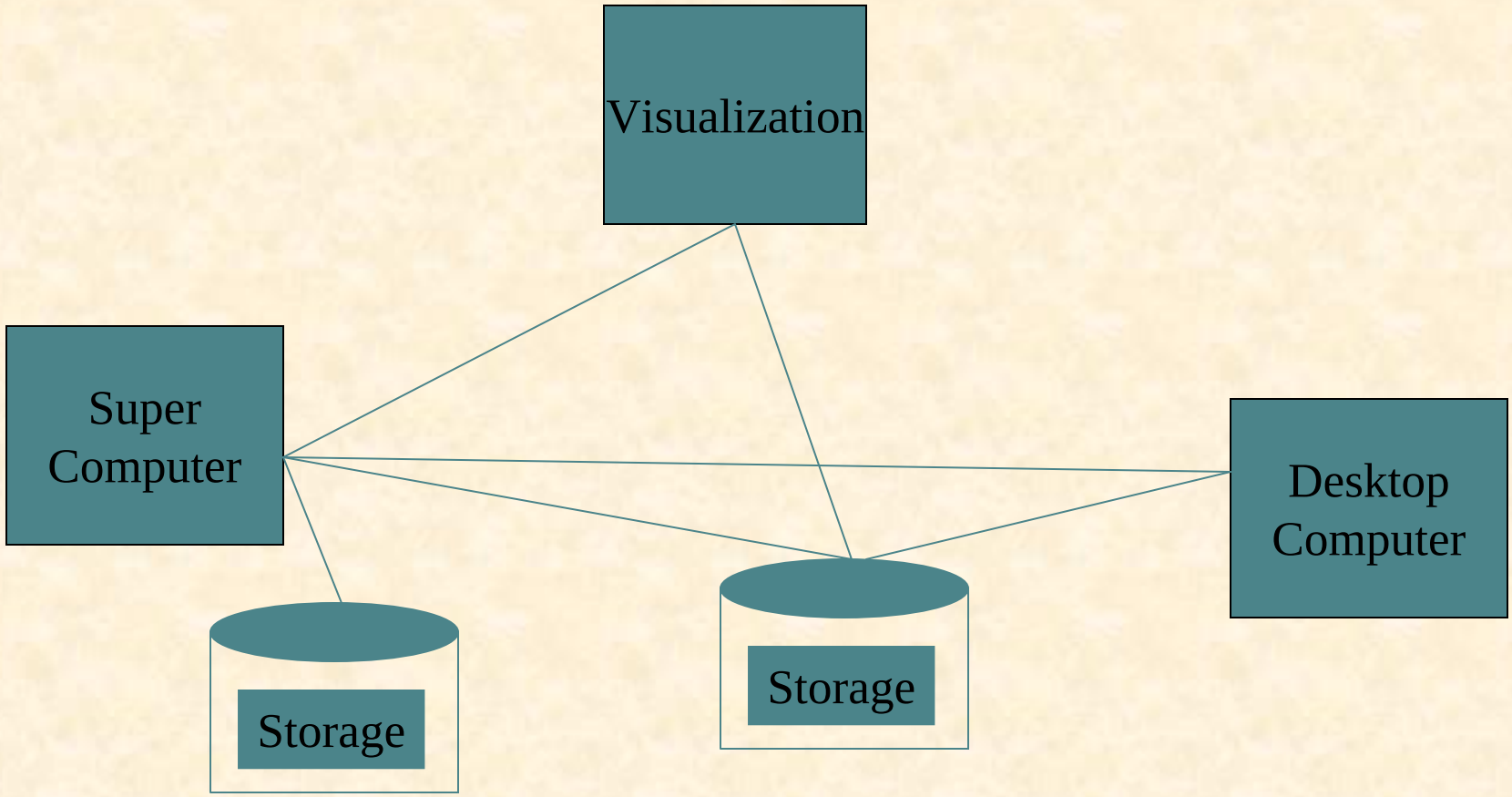
Project Conceived at 2nd DOE workshop

1. High-Performance Networks for High-Impact Science, Aug 13-15, 2002.

2. Network Provisioning and Protocols for High-Impact Science, April 10-11, 2003.
   report: www.csm.ornl.gov/ghpn/wk2003.html

3. DOE Science Networking Challenge: Roadmap to 2008, June 3-5, 2003

Provide dedicated channels to applications



Long way to go!

1. **Network switches are available off-the-shelf with capability to provide bandwidth-on-demand**
   - **It is not sufficient to just buy these and hook them together?**
2. **Implement a "socket" to enable applications to request bandwidth channels on demand and release when done**
3. **Need "application immersion" into dedicated channels**
   - **Scientists need enabled applications (toolkit) to use the dedicated bandwidth**
4. **Test the applications with the integrated BW-on-demand "socket" on a lab BW-on-demand network testbed**
5. **Finally, "take" the network wide area**

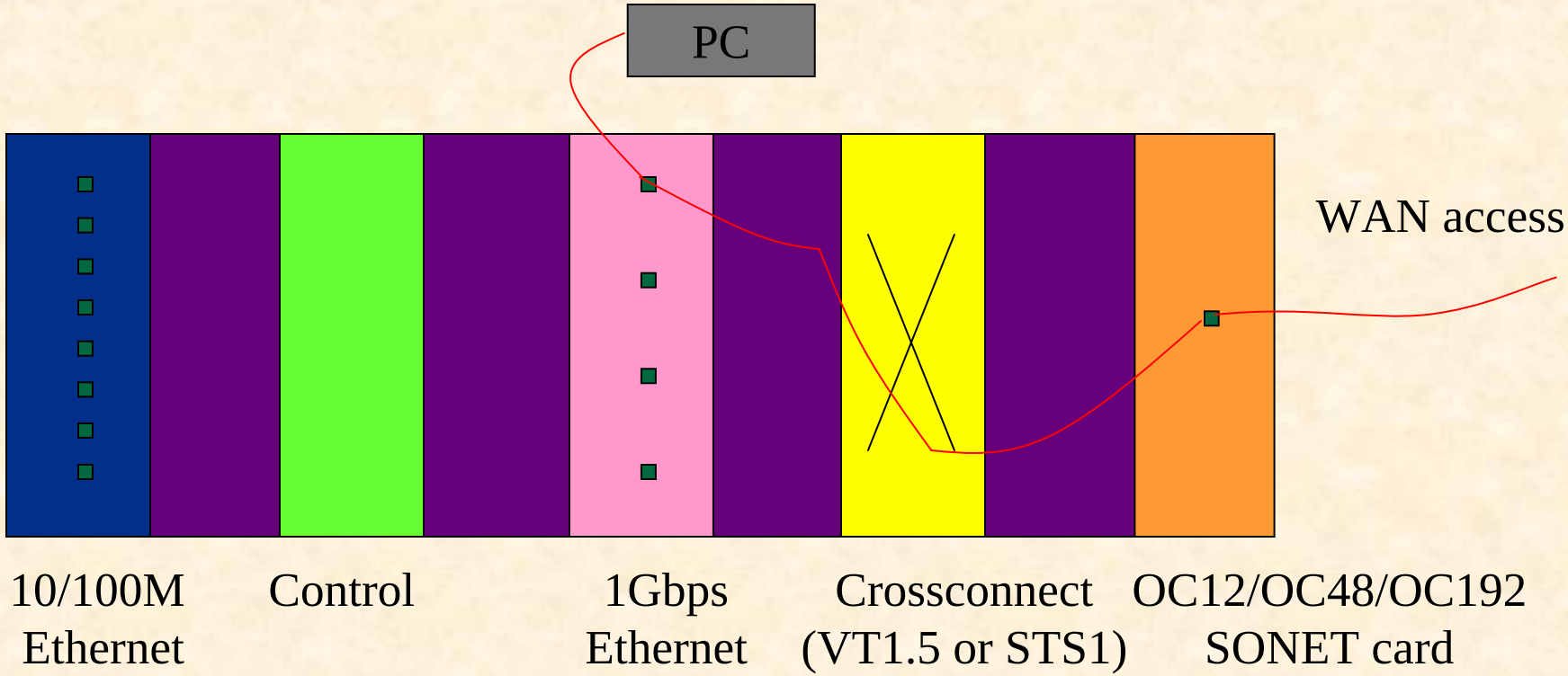
Project concept

- **Network:**
  - **CHEETAH: Circuit-switched High-Speed End-to-End Transport ArcHitecture**
  - **Create a network that on-demand offers end-to-end dedicated bandwidth channels to applications**
  - **Operate a PARALLE network to existing high-speed IP networks – NOT AN ALTERNATIVE!**
- **Transport protocols:**
  - **Design to take advantage of dual end-to-end paths**
    - **IP path and dedicated channel**
- **TSI applications:**
  - **High-throughput file transfers**
  - **Interactive remote visualization**
  - **Remote computational steering**
  - **Multipoint collaborative computation**

Network specifics

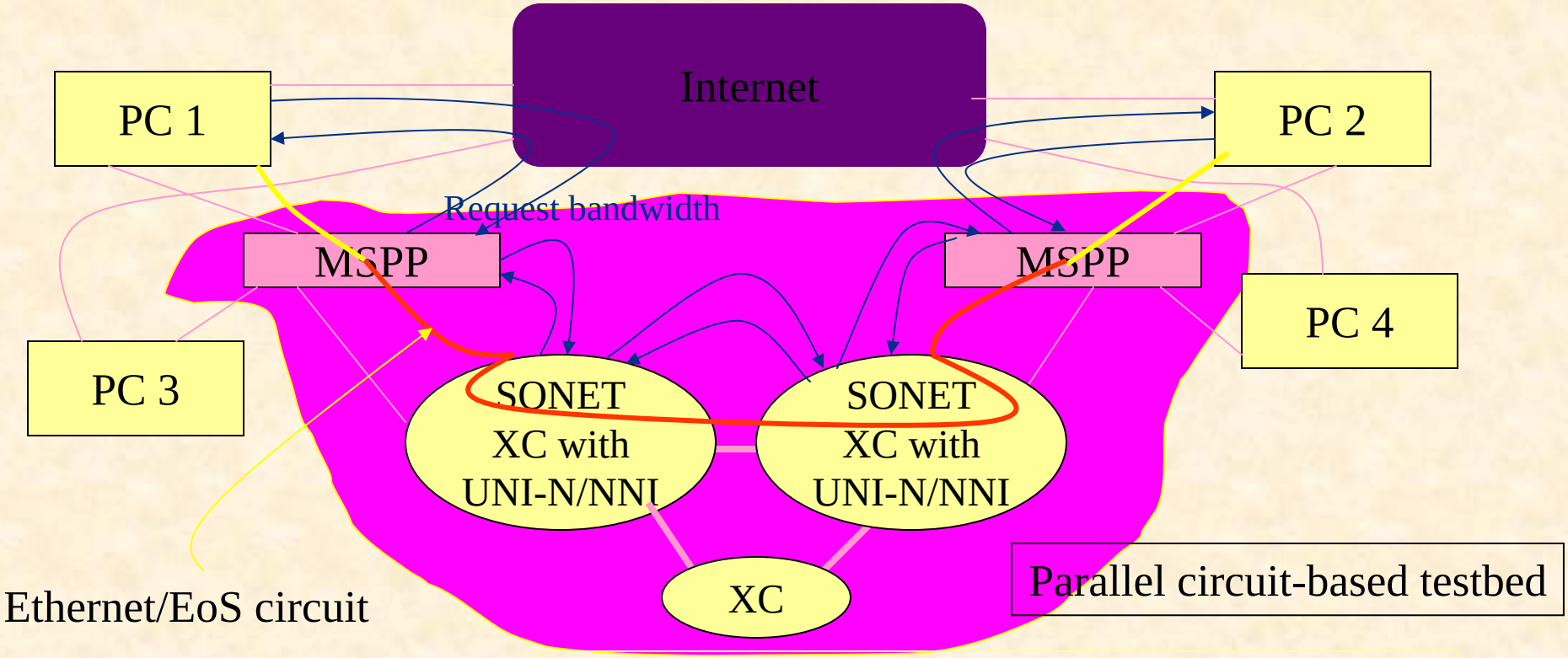
- **Dedicated channel:**
  - **High-speed Ethernet mapped to Ethernet-over-SONET circuit**
- **Leverage existing technologies:**
  - **100Mbps/1Gbps Ethernet in LANs**
  - **SONET in MANs/WANs**
  - **Availability of Multi-Service Provisioning Platforms (MSPP) class devices**
    - **Can map Ethernet to Ethernet-over-SONET**
    - **Can cross-connect dynamically**
    - **Can rate-control Ethernet ports**

Multi-Service Provisioning Platform (MSPP)



- **MSPPs already deployed within enterprises**
- **Some routers can achieve similar capability with filters – somewhat less dynamically**
- **Combination of traditional routers and VLAN-enabled Ethernet switches can work**

Dynamic circuit sharing

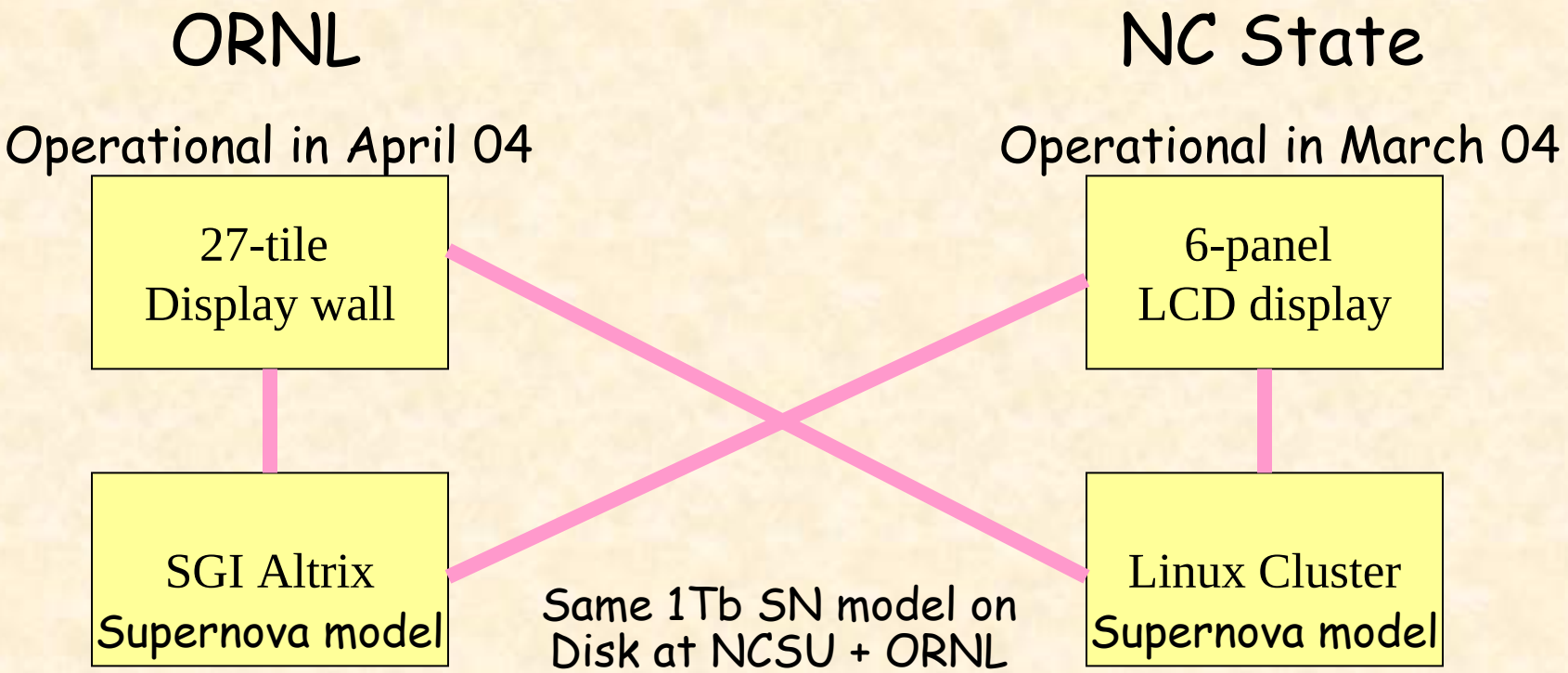


- **Steps:**
  - **Route lookup**
  - **Resource availability checking and allocation**
  - **Program switch fabric for the crossconnection**

TSI Application Activities –NCSU-ORNL

- **Construct local visualization environment**
  - **Added 6 cluster nodes, expanded RAID to 1.7TB**
  - **Installed dedicated server for network monitoring**
  - **Began constructing visualization cluster**
  - **Wrote software to distribute data on cluster**
- **Supernova Science**
  - **Generated TB data set on Cray X1 @ ORNL**
  - **Tested ORNL/NCSU collaborative visualization session**

LAN and WAN testing



Currently testing viz on Altrix + cluster using single-screen graphics

OAK RIDGE NATIONAL LABORATORY
U. S. DEPARTMENT OF ENERGY

John M. Blondin
John_Blondin@ncsu.edu

UT-BATTELLE

Applications Enabled for TSI project

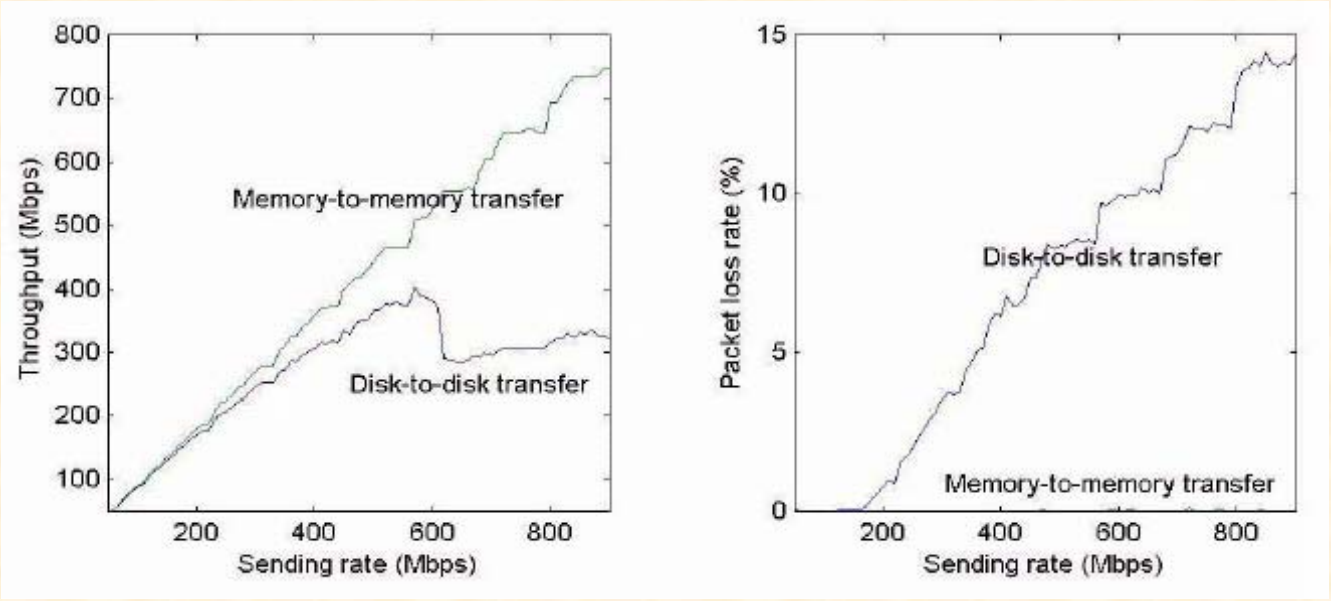
- **To provide scientists dedicated channels on CHEETAH network**
  - **File transfer tools**
  - **Visualization tools:**
    - **Ensight or Aspect/Paraview**
    - **Custom OpenGL codes**
  - **Computational steering tools**

Transport protocols – UVA, ORNL

- **File transfers**
  - **Tested various rate-based transport solutions**
    - **SABUL, UDT, Tsunami, RBUDP, hurricane**
    - **UVA – Local Connection: Two Dell 2.4Ghz PCs with 100Mhz 64-bit PCI buses**
    - **ORNL - Atlanta Testbed: Dell dual 2.4Ghz and dual optron hosts with PCIX buses, dedicated 1Gbps channel**
  - **Why rate based protocols:**
    - **No congestion control after the channel is setup**
    - **instead for flow control**
- **Control Channels**
  - **Channel stabilization under random losses and jitter**
  - **Typically only a small portion of channel bandwidth**
  - **Stochastic approximation method**

Rate-based flow control (UVA)

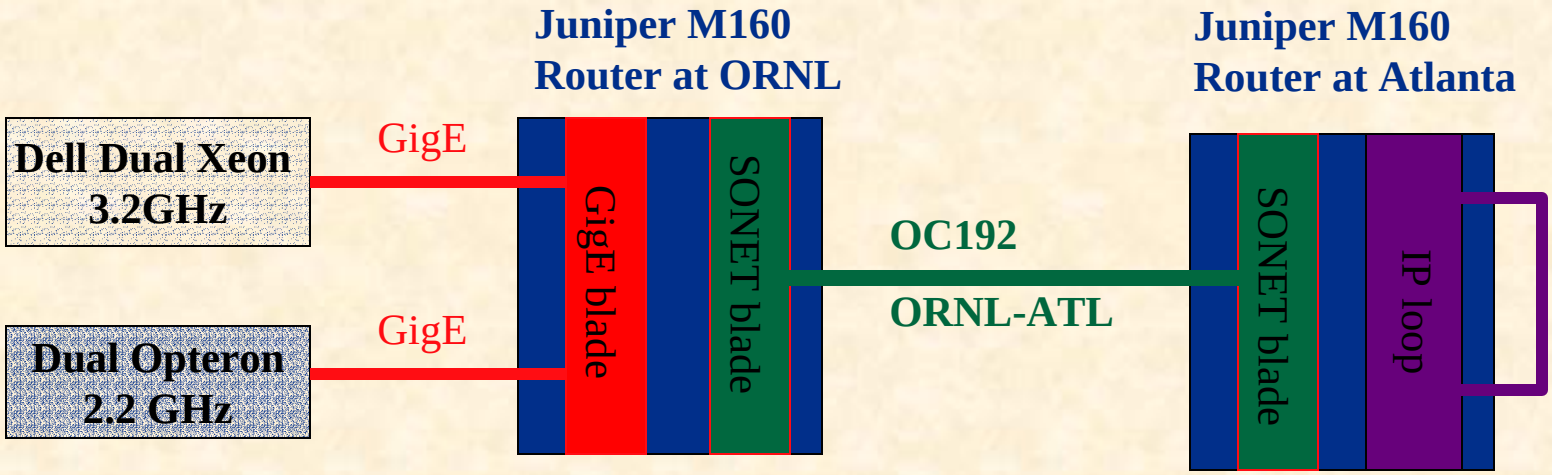
- **Receive-buffer overflows: a necessary evil**
  - **Play it safe and set a low rate: avoid/eliminate receive-buffer losses**
  - **Or send data at higher rates but have to recover from losses**



(MTU=1500B, UDP buffer size=256KB, SABUL data block size=7.34MB)

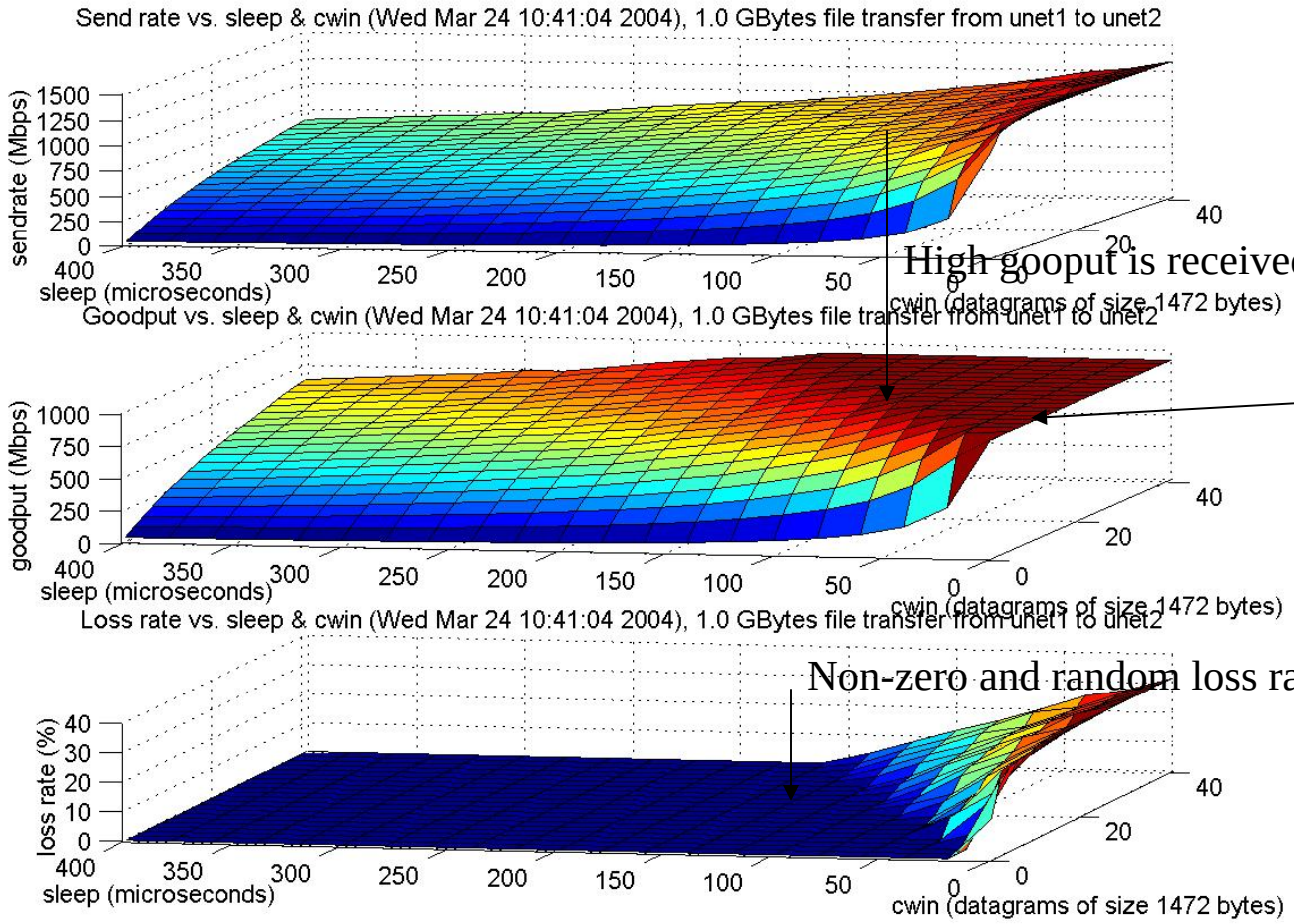
- **Two Dell 2.4Ghz PCs with 100Mhz 64-bit PCI buses**
  - **Connected directly to each other via a GbE link**
    - **Emulates a dedicated GbE-EoS-GbE link**
  - **Disk bottleneck: IDE 7200 rpm disks**

ORNL-Atlanta 1Gbps Channel



- **Disks: RAID 1 dual disks (140GB SCSI) with XFS file systems under linux**
  - **Peak disk data rate is ~1.2Gbps – disk is not a bottleneck**
- **Hurricane (ORNL) protocol achieved ~954Mbps for file transfers**
  - **Fastest for file transfers**
  - **UDT achieved 890Mbps**
- **Memory transfers**
  - **UDT – 958Mbps**

ORNL-ATL-ORNL 1Gbps channel UDP goodput and loss profile



High gooput is received at non-trivial loss

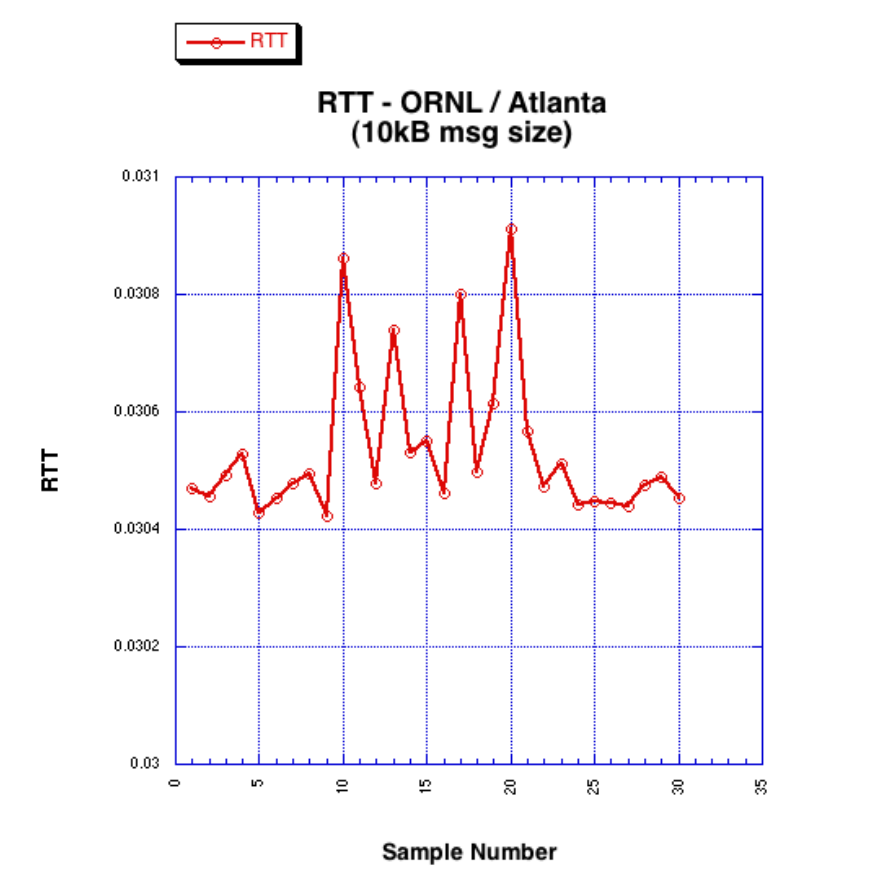
Gooput plateau ~970Mbps

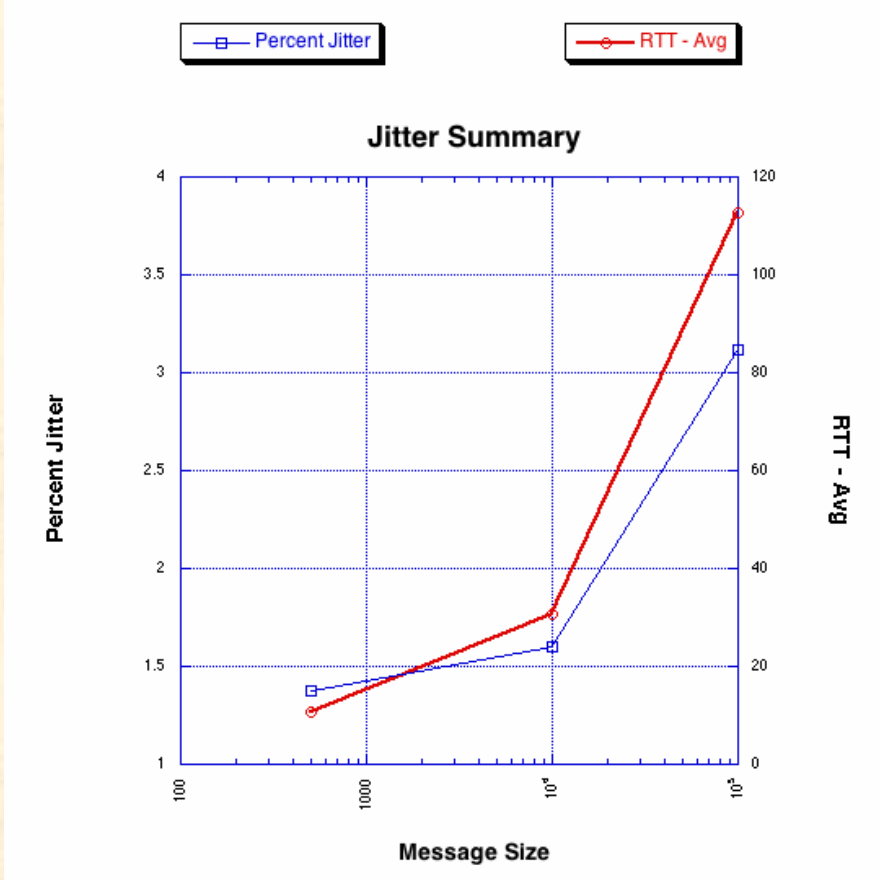
Non-zero and random loss rate

Point in horizontal plane – sending rate (waiting time, window size)

For Hurricane protocol we manually stabilized at plateau with low loss

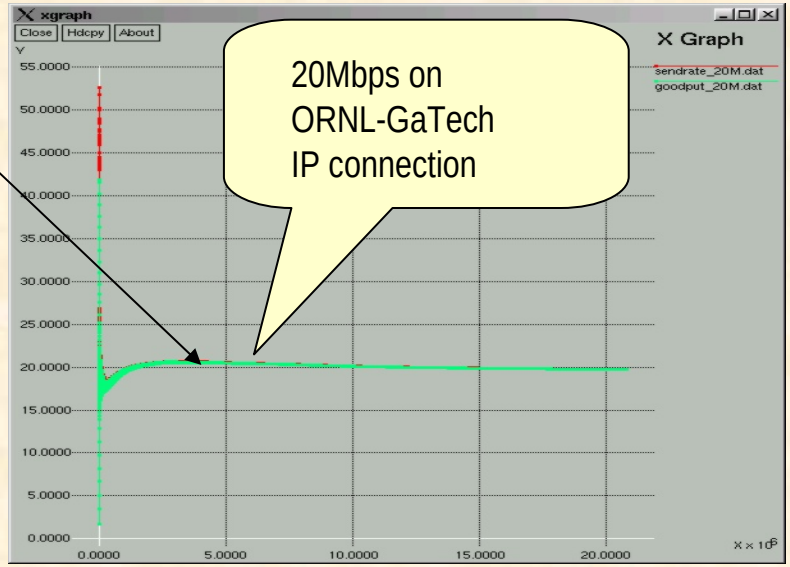
Jitter - ORNL / Atlanta

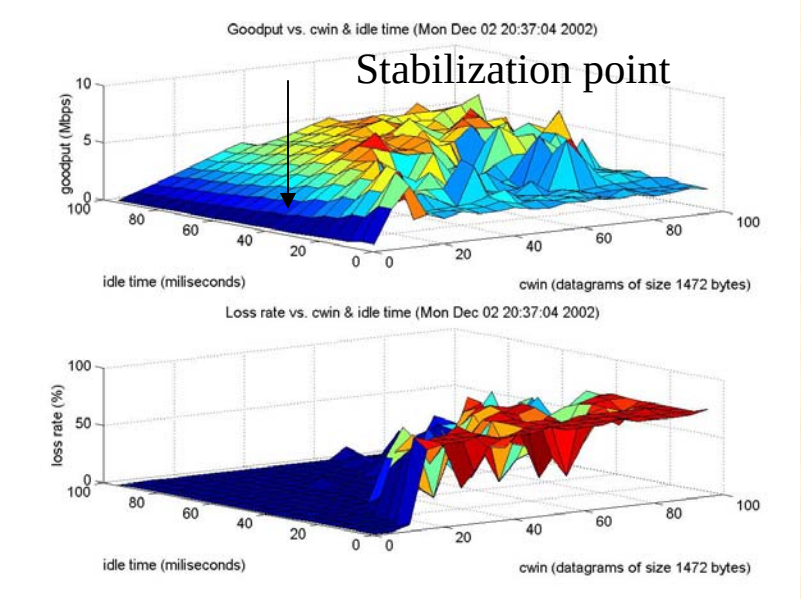




Control Channels Over Shared Connections

- **Stabilization protocols for visualization control streams**
  - **stochastic approximation methods for stable application-to-application streams**
- **Modularization and channel separation framework for visualization**
  - **decompose visualization pipeline into modules, measuring effective bandwidths and mapping them onto network**





Taking it wide-area

- **Three possible approaches**
  - **Collocate high-speed circuit switches at POPs and lease circuits from commercial service provider or NLR**
  - **Use MPLS tunnels through Abilene**
  - **Collocate switches at Abilene POPs and share router links – after thorough testing**

Summary

- **Implement building blocks for TSI scientists to take advantage of dedicated channels**
- **Demonstrate applications on dynamically shared high-speed circuit-switched network**
- **Take it to the wide area production quality deployment**

Thank You