DOE UltraScience Net:
High-Performance Experimental Network Research Testbed

Presented by

Nagi Rao

Complex Systems
Computer Science and Mathematics Division

Research supported by the Department of Energy’s DOE Office of Science
Office of Advanced Scientific Computing High-Performance Networking Program

The need

- DOE large-scale science applications on supercomputers and experimental facilities require high-performance networking.
  - Moving petabyte data sets, collaborative visualization, and computational steering
- Application areas span the disciplinary spectrum: High-energy physics, climate, astrophysics, fusion energy, genomics, and others.

Promising solution

- High bandwidth and agile network capable of providing on-demand dedicated channels: multiple 10s Gb/s to 150 Mb/s
- Protocols are simpler for high throughput and control channels

Challenges

- In 2003, several technologies needed to be (fully) developed
- User-/application-driven agile control plane:
  - Dynamic scheduling and provisioning
  - Security—encryption, authentication, authorization
- Protocols, middleware, and applications optimized for dedicated channels

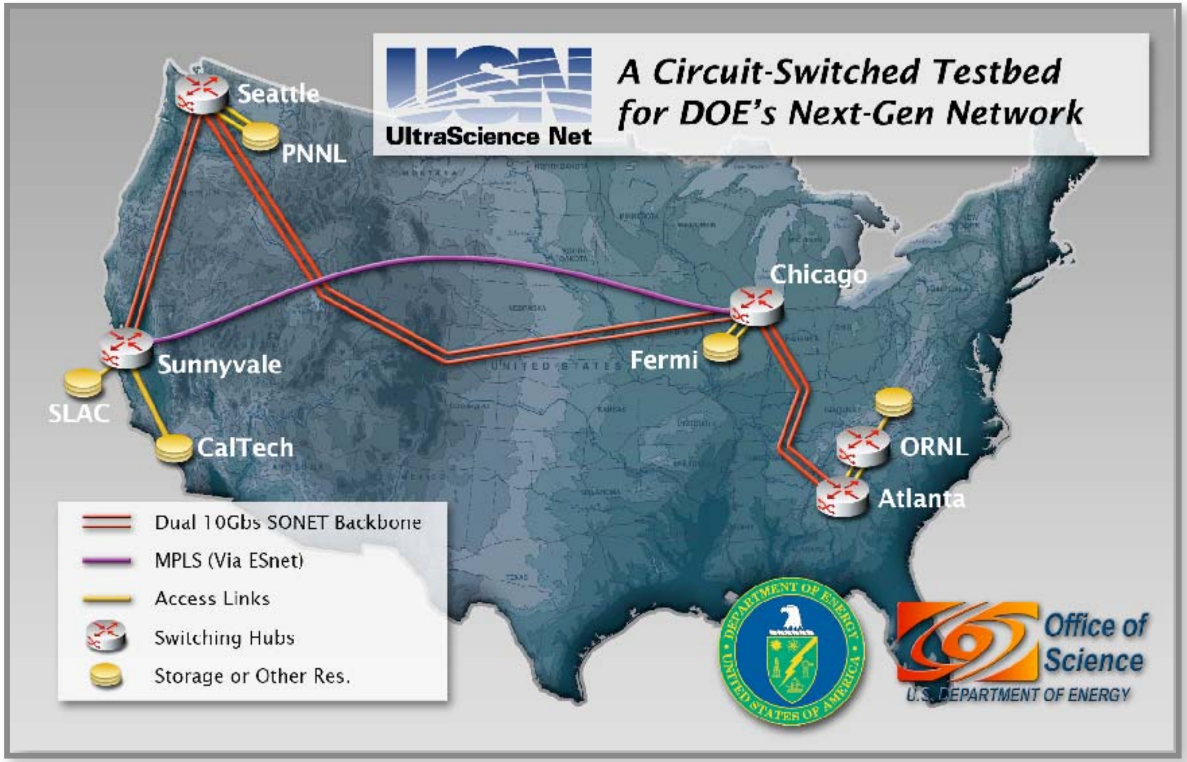
DOE UltraScience Net – In a nutshell

Experimental network research testbed

- To support advanced networking and related application technologies for DOE large-scale science projects

Features

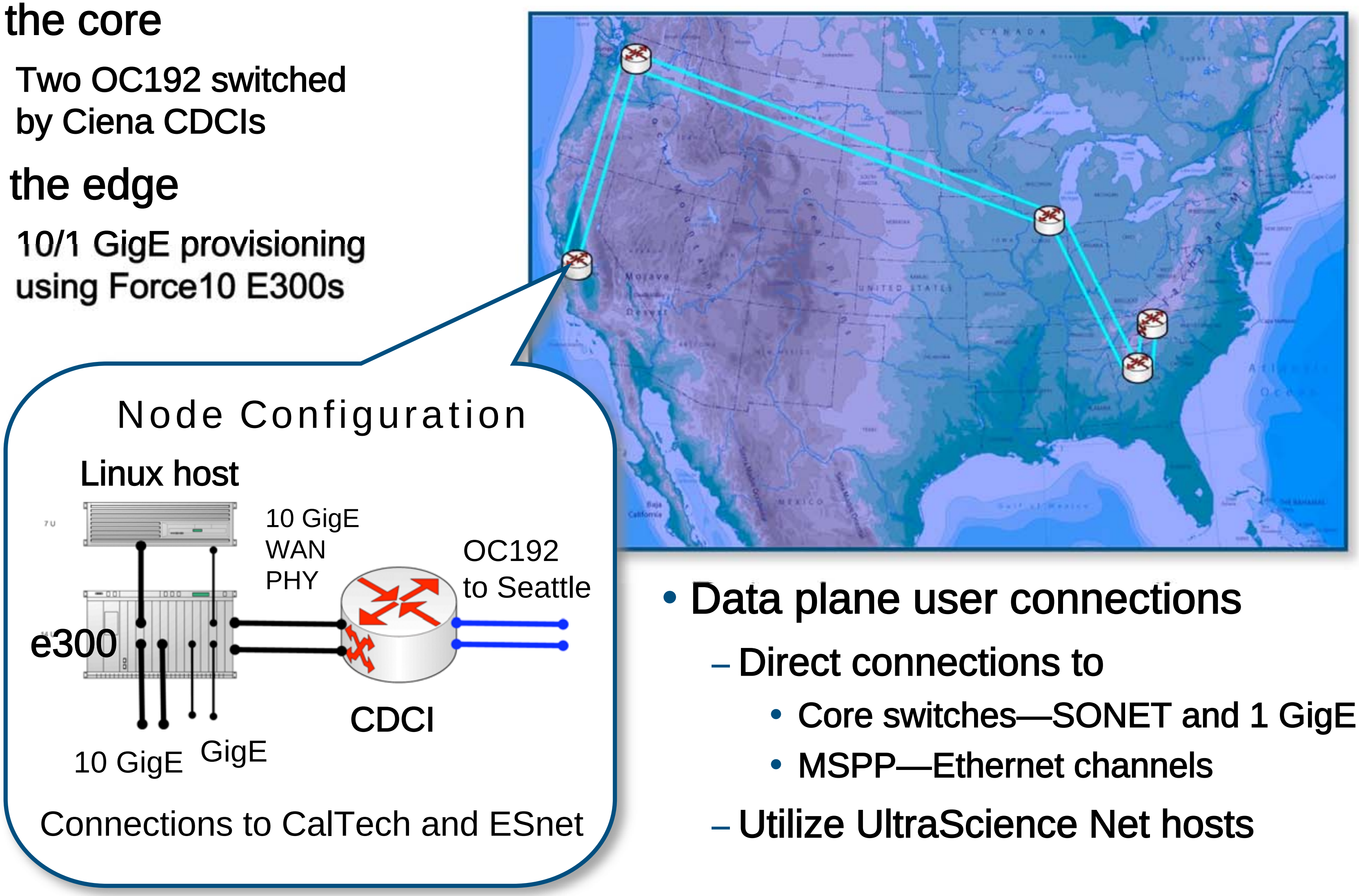
- End-to-end guaranteed bandwidth channels
- Dynamic, in-advance reservation and provisioning of fractional/full lambdas
- Secure control-plane for signaling
- Proximity to DOE sites: LCF, Fermi National Laboratory, National Energy Research Scientific Computing Center
- Peering with ESnet, National Science Foundation's CHEETAH, and other networks



OAK RIDGE National Laboratory

USN data plane: Node configuration

- In the core
  - Two OC192 switched by Ciena CDCIs
- At the edge
  - 10/1 GigE provisioning using Force10 E300s



Node Configuration

Linux host

10 GigE WAN PHY

OC192 to Seattle

e300

CDCI

10 GigE

GigE

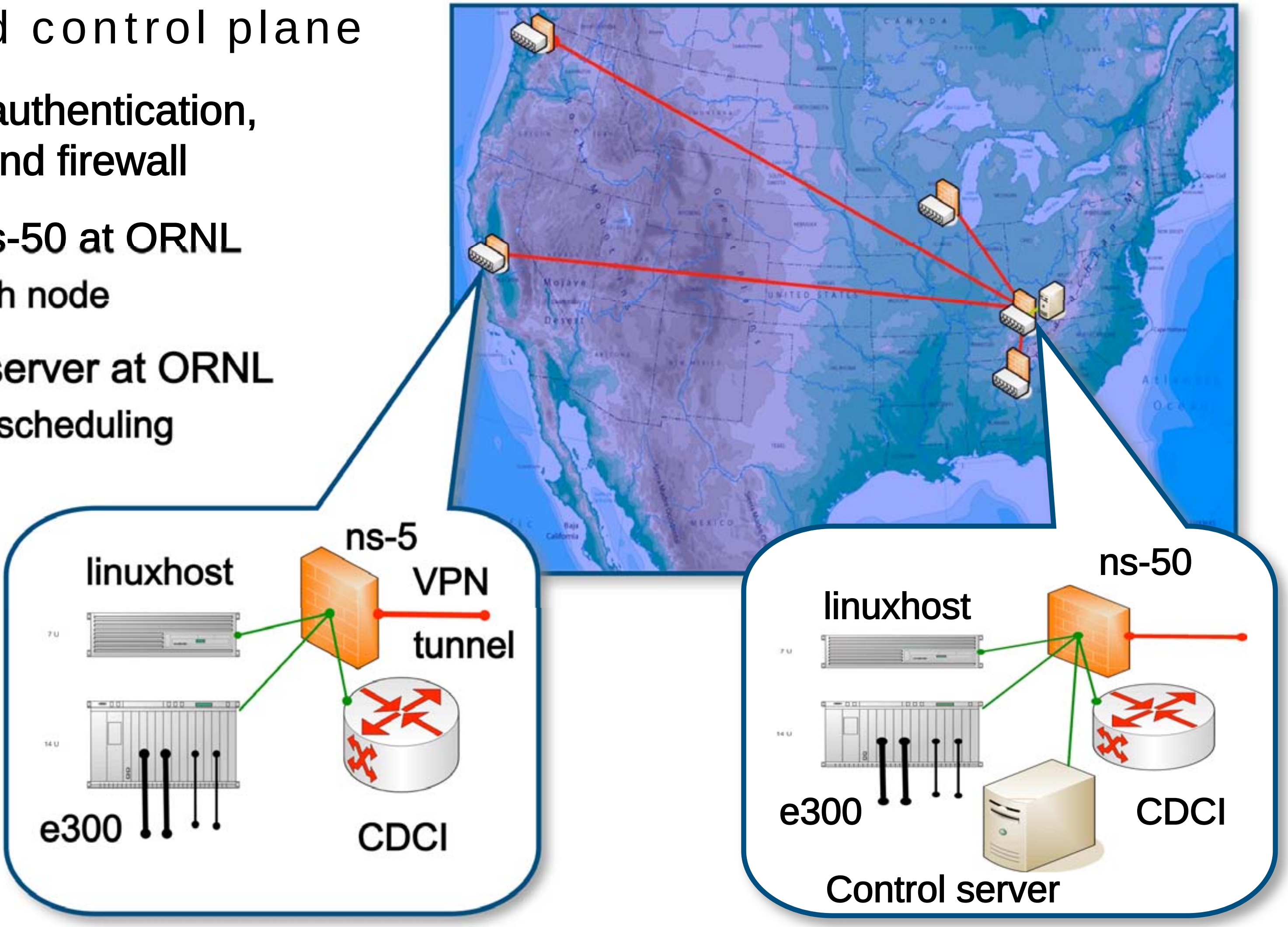
Connections to CalTech and ESnet

- Data plane user connections
  - Direct connections to
    - Core switches—SONET and 1 GigE
    - MSPP—Ethernet channels
  - Utilize UltraScience Net hosts

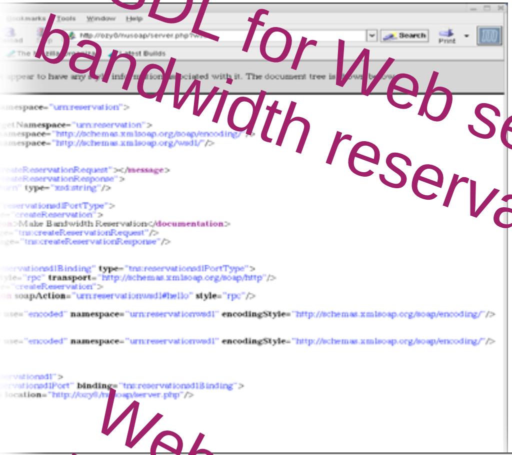
Secure control plane

Out-of-band control plane

- VPN-based authentication, encryption, and firewall
- Netscreen ns-50 at ORNL
  - ns-5 at each node
- Centralized server at ORNL
  - Bandwidth scheduling
  - Signaling



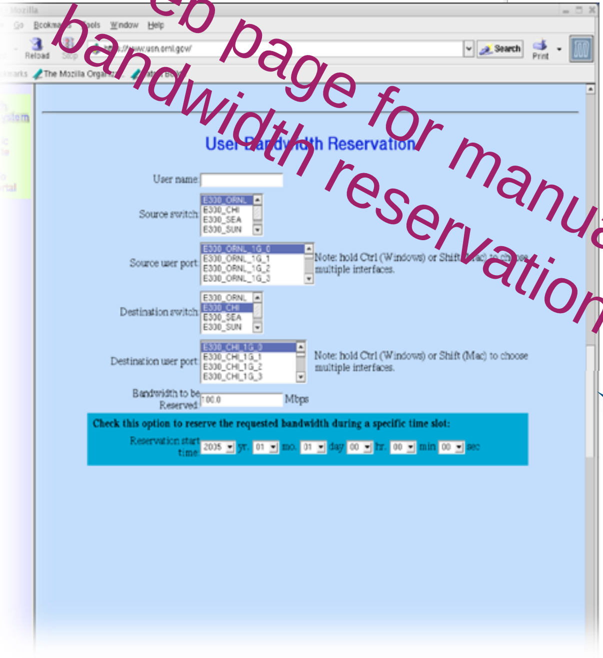
USN control plane



WSDL for Web service bandwidth reservation

appear to have any style information associated with it. The document tree is shown below.

```
amespace="urn:reservation">
getNamespace="urn:reservation">
amespace="http://schemas.xmlsoap.org/soap/encoding/"/>
amespace="http://schemas.xmlsoap.org/wsdl/"/>

reateReservationRequest"></message>
reateReservationResponse">
turn" type="xsd:string"/>

reservationwsdlPortType">
="createReservation">
ion>Make Bandwidth Reservation</documentation>
ge="tns:createReservationRequest"/>
age="tns:createReservationResponse"/>

servationwsdlBinding" type="tns:reservationwsdlPortType">
tyle="rpc" transport="http://schemas.xmlsoap.org/soap/http"/>
="createReservation">
ion soapAction="urn:reservationwsdl#hello" style="rpc"/>
 use="encoded" namespace="urn:reservationwsdl" encodingStyle="http://schemas.xmlsoap.org/soap/encoding/"/>

 use="encoded" namespace="urn:reservationwsdl" encodingStyle="http://schemas.xmlsoap.org/soap/encoding/"/>

ervationwsdl">
ervationwsdlPort" binding="tns:reservationwsdlBinding">
 location="http://ozy0/nusoap/server.php"/>
```



Web page for manual bandwidth reservation

User Bandwidth Reservation

User name:

Source switch: E300_ORNL / E300_CHI / E300_SEA / E300_SUN

Source user port: E300_ORNL_1G_0 / E300_ORNL_1G_1 / E300_ORNL_1G_2 / E300_ORNL_1G_3 — Note: hold Ctrl (Windows) or Shift (Mac) to choose multiple interfaces.

Destination switch: E300_ORNL / E300_CHI / E300_SEA / E300_SUN

Destination user port: E300_CHI_1G_0 / E300_CHI_1G_1 / E300_CHI_1G_2 / E300_CHI_1G_3 — Note: hold Ctrl (Windows) or Shift (Mac) to choose multiple interfaces.

Bandwidth to be Reserved: 100.0 Mbps

Check this option to reserve the requested bandwidth during a specific time slot:

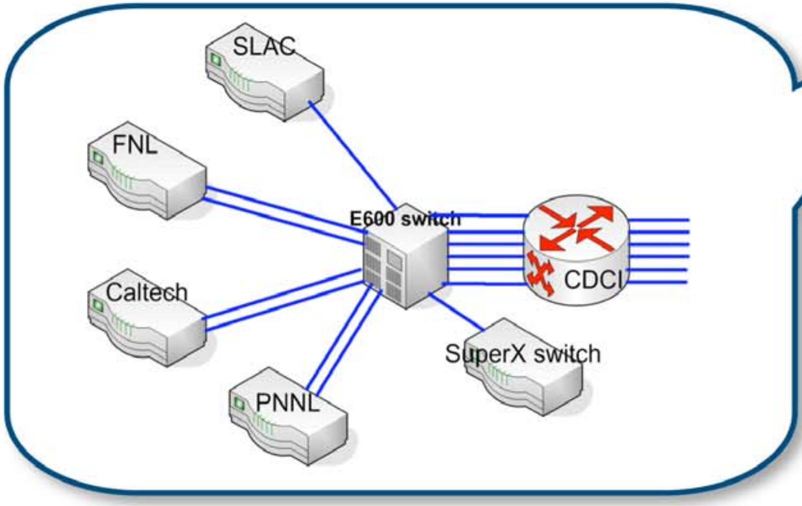
Reservation start time: 2005 yr. 01 mo. 01 day 00 hr. 00 min 00 sec

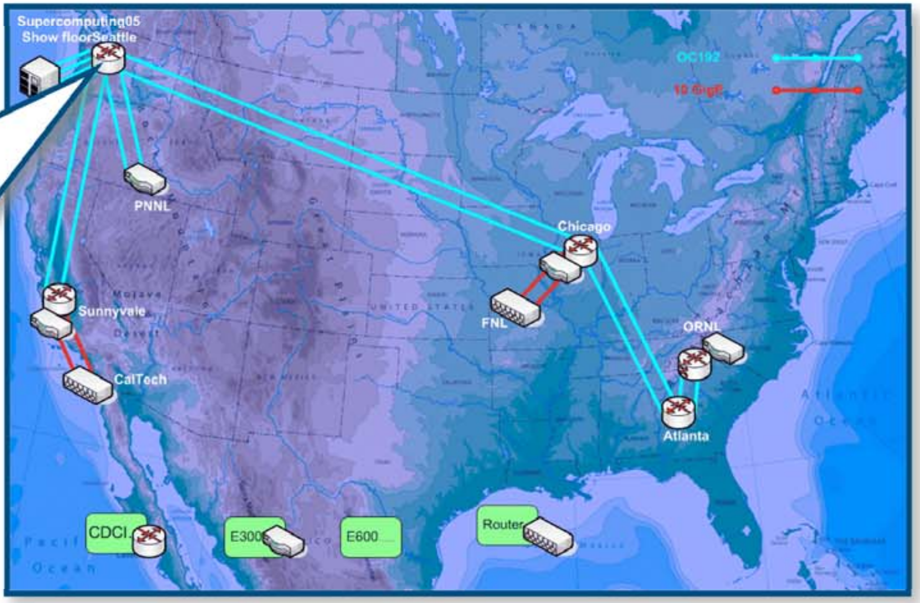
- Phase I (completed)
  - Centralized path computation for bandwidth optimization
  - TL1/CLI-based communication with Core Directors and E300s
  - User access via centralized Web-based scheduler
- Phase II (completed)
  - Web services interface
  - X509 authentication for Web server and service
- Phase II (current)
  - Generalized Multiprotocol Label Switching (GMLS) wrappers for TL1/CLI
  - Inter-domain “secured” GMPLS-based interface

Both use USN SSL certificates for authorization.

USN at Supercomputing 2005

Supercomputing 2005 exhibit floor





- Extended USN to exhibit floor
  - Eight dynamic 10-Gb/s long-haul connections over time
- Moved and recreated USN-Seattle node on various booths
  - Pacific Northwest National Laboratory, FNL, ORNL, Caltech, Stanford Linear Accelerator Center at various booths
- Supported applications and bandwidth challenge

Helped Caltech team win Bandwidth Challenge

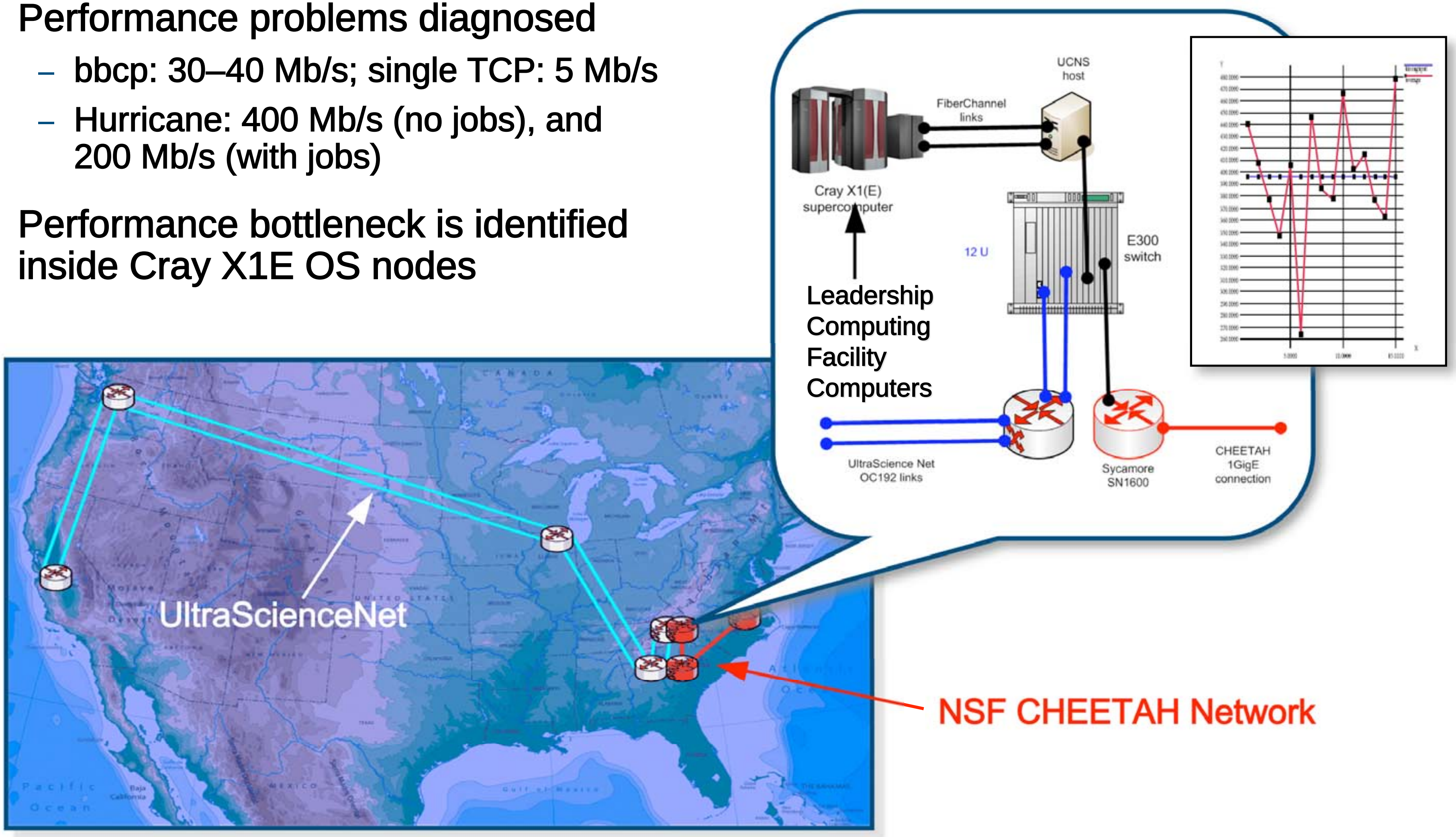
- 40 Gb/s aggregate bandwidth
- 164 terabytes transported in a day

UltraScience Net

Dedicated connections to supercomputers:

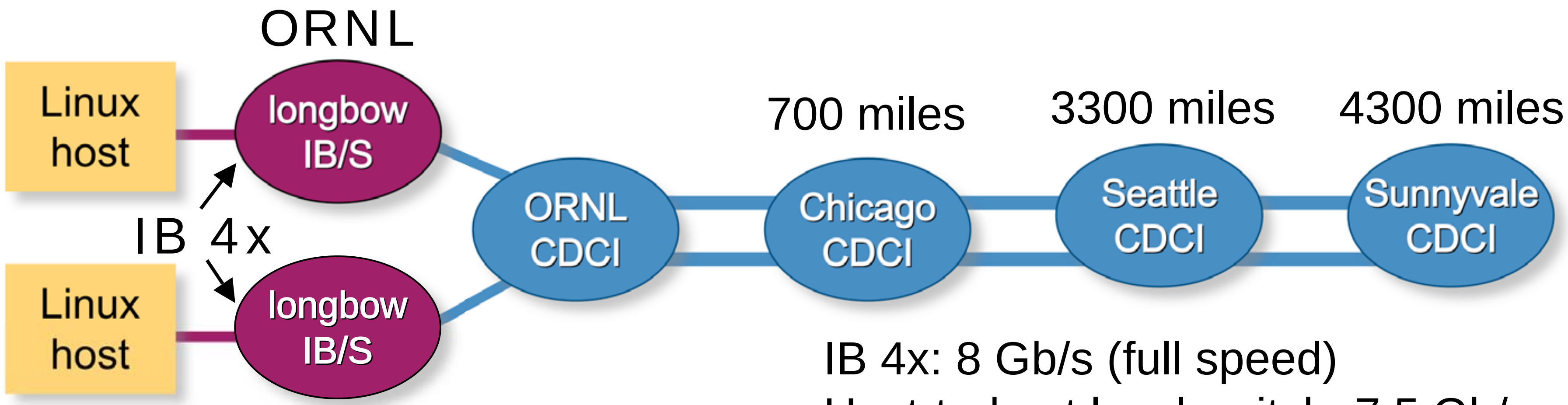
1 Gb/s dedicated connection: Cray X1E—NSCU Cluster

- Performance problems diagnosed
  - bbcp: 30–40 Mb/s; single TCP: 5 Mb/s
  - Hurricane: 400 Mb/s (no jobs), and 200 Mb/s (with jobs)
- Performance bottleneck is identified inside Cray X1E OS nodes



Infiniband over SONET

Demonstrated that IB can scale to thousands of miles over SONET: 5% throughput reduction over 8000 miles



IB 4x: 8 Gb/s (full speed)
Host-to-host local switch: 7.5 Gb/s

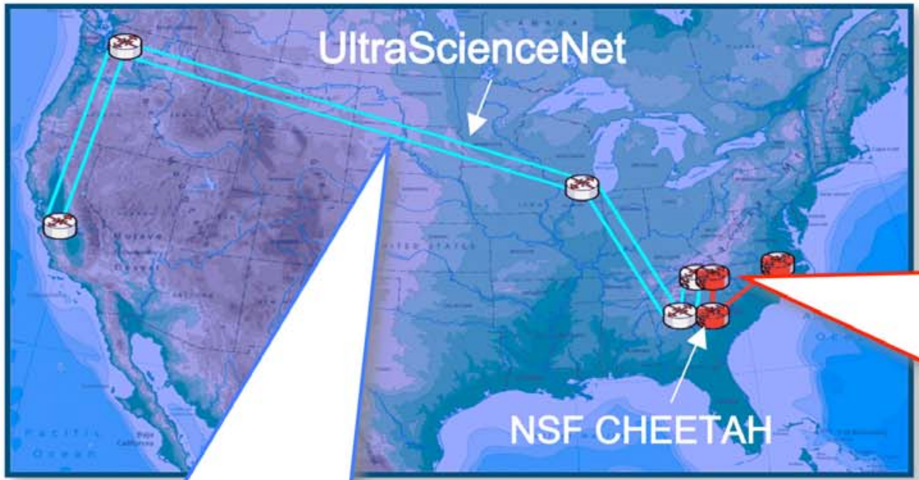
ORNL loop – 0.2 mile: **7.5** Gb/s

ORNL–Chicago loop – 1400 miles: **7.46** Gb/s

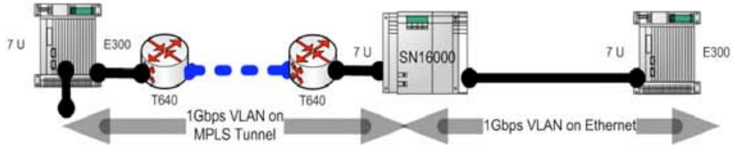
ORNL–Chicago–Seattle loop – 6600 miles: **7.23** Gb/s

ORNL–Chicago–Seattle–Sunnyvale loop – 8600 miles: **7.20** Gb/s

Demonstrated peering circuit-packet switched networks: USN–CHEETAH VLAN through L3-L2 paths

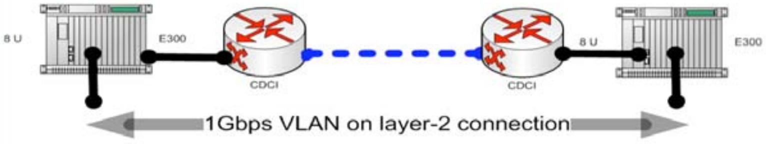


CHEETAH VLAN



- NSF CHEETAH: layer 3 + layer
  - VLAN: T640-T640–SN1600–Cisco 3750

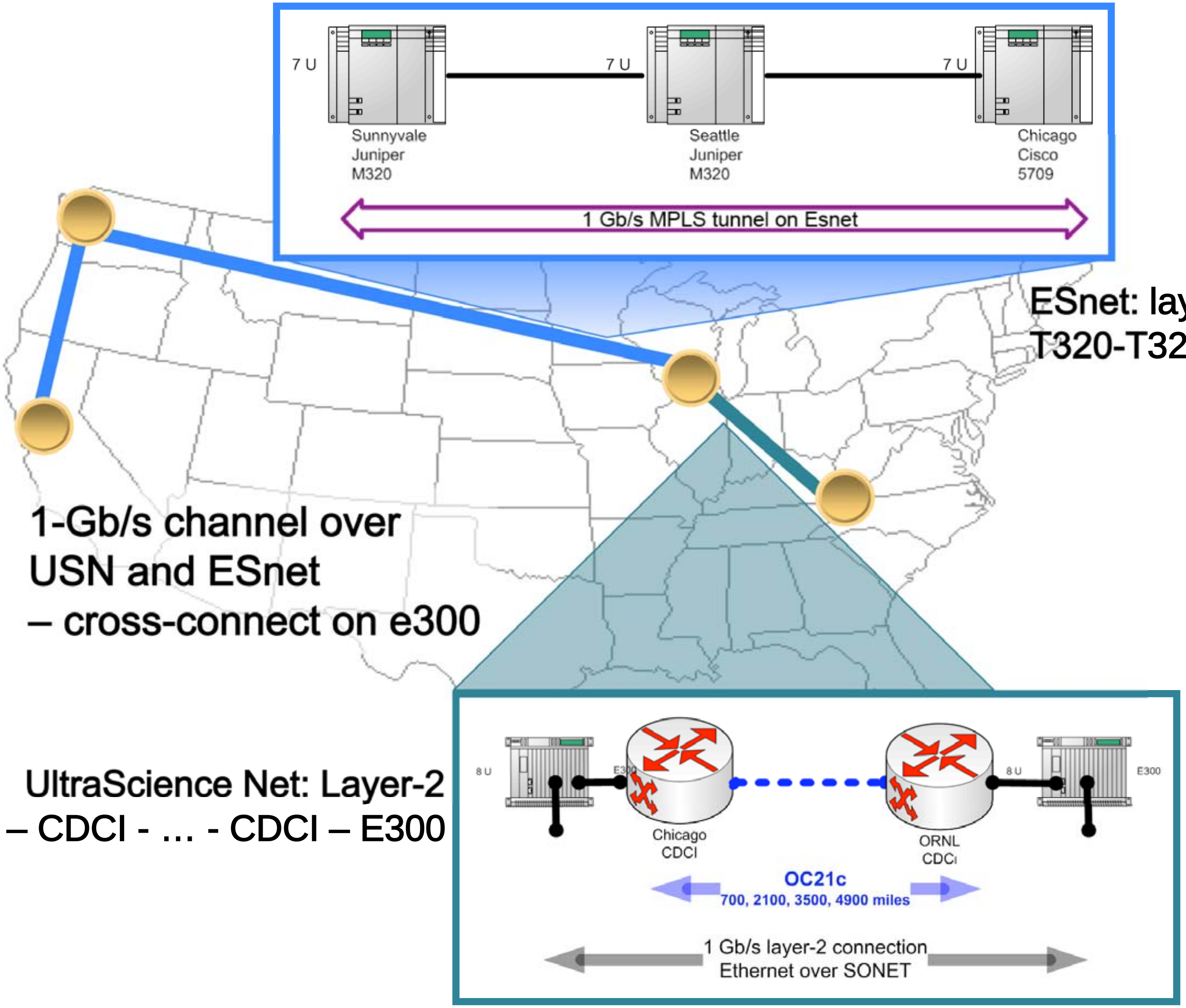
- DOE UltraScience Net: Layer-2
  - VLAN: E300–CDCI - … -CDCI–E300



USN VLAN

Coast-to-cost 1-Gb/s channel demonstrated over USN and CHEETAH—simple cross-connect on e300.

USN–ESnet Peering of L2 and L3 paths



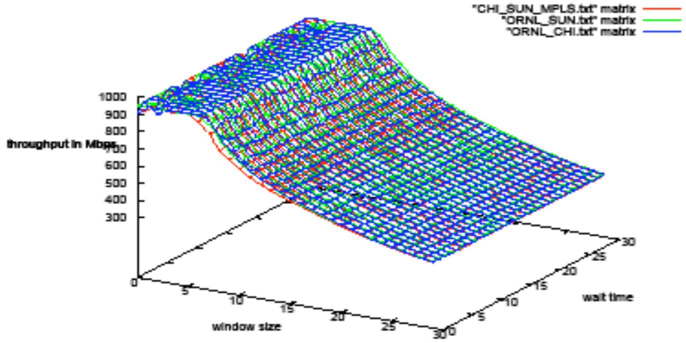
ESnet: layer-3 VLAN:
T320-T320 – Cisco 6509

1-Gb/s channel over
USN and ESnet
– cross-connect on e300

UltraScience Net: Layer-2
E300 – CDCI - … - CDCI – E300

Throughput comparisons: Summary

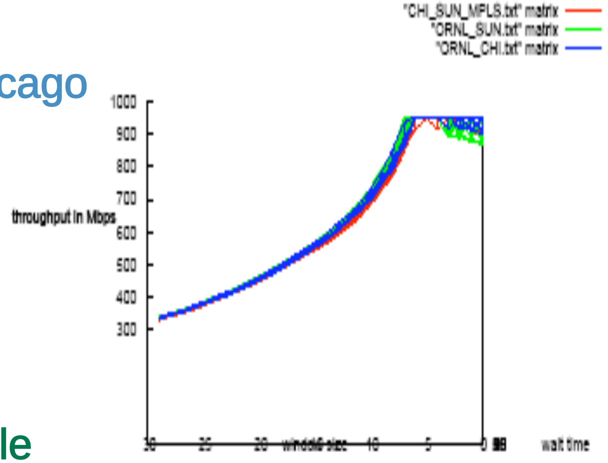
	PLUT	UDP peak	TCP peak	PLUT-TCP diff
MPLS	952 Mb/s	953	840	112
SONET	955 Mb/s	957	900	55
Hybrid	952 Mb/s	953	840	112
Difference	3 Mb/s	5 Mb/s	60 Mb/s	



USN
ORNL–Chicago-..-ORNL–Chicago

ESnet
Chicago–Sunnyvale

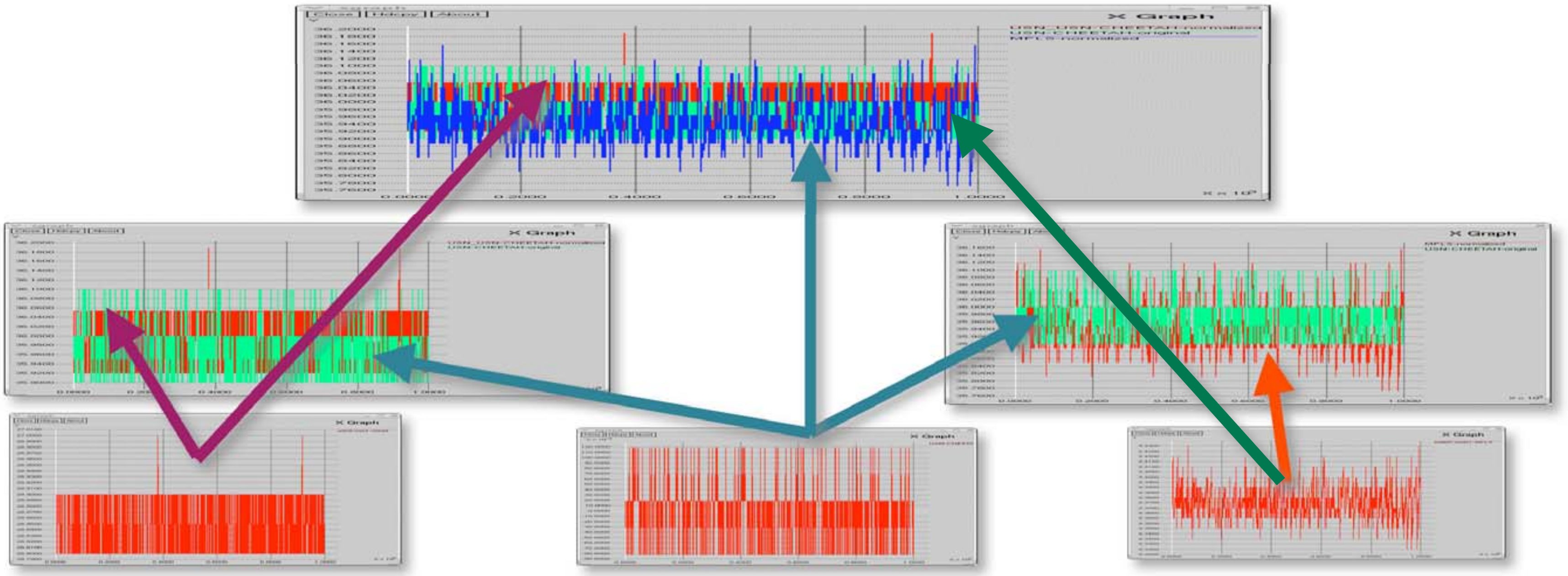
ESnet
ORNL–Chicago–Sunnyvale



Special purpose UDP-PLUT transport achieved higher throughput than multistream TCP.

USN-enabled comparison of VLANs:
SONET–SONET–MPLS composed–L2MPLS

Measurements are normalized for comparison



SONET

mean time = 26.845877 ms
std_dev (%) = 0.187035

SONET-MPLS composite

mean time = 35.981812 ms
std_dev (%) = 0.151493

L2MPLS

mean time = 9.384557 ms
std_dev (%) = 3.281692

SONET channels have smaller jitter levels.

Conclusions

- USN infrastructure development is close to completion:
  - Its architecture has been adopted by LHCnet and Internet2.
  - It has provided special connections to supercomputers.
  - It has enabled testing: VLAN performance, peering of packet-circuit switched networks, control plane with advanced reservation, Lustre and Infiniband over wide-area.
- USN continues to play a research role in advanced networking capabilities:
  - Networking technologies for LCFs
    - Connectivity to supercomputers
    - Testing of file systems: Lustre over TCP/IP and Inifiniband/SONET
  - Integrated multidomain interoperation: USN-ESnet-CHEETAH-HOPI
    - On-going efforts with OSCARS and HOPI
  - Hybrid optical packet and switching technologies
    - VLAN testing and analysis over L1-2 and MPLS connections (this presentation)
    - Configuration and testing of hybrid connections

Contacts

Nageswara (Nagi) S. Rao

Complex Systems
Computer Science and Mathematics Division
(865) 574-7517
raons@ornl.gov

William (Bill) R. Wing

Computer Science Research Group
Computer Science and Mathematics Division
(865) 574-8839
wingwr@ornl.gov