UltraScience Net Update: Network Research Experiments

Nagi Rao, Bill Wing, Susan Hicks, Paul Newman, Steven Carter
Oak Ridge National Laboratory
raons@ornl.gov
https://www.csm.ornl.gov/ultranet

February 12, 2007
JointTechs Meeting
Minneapolis, MN

Sponsored by
High-Performance Networking Program
Office of Science
U.S. Department of Energy
Outline

USN Architecture and Deployment
Testbed Network Research Activities
  Peering and Alignment – UVA, HOPI
  VLANS Testing and Analysis – HOPI, ESnet-OSCARS
  Infiniband over SONET – ORNL/NLCF
Microscope Control – PNL
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: NLCF, FNL, NERSC
• Peering with ESnet, NSF CHEETAH, HOPI and other networks
• $6M over 4 years
USN Architecture:
Separate Data-Plane and Control-Planes

No data plane continuity: can be partitioned into “islands”
- necessitated out-of-band control plane

Secure control-plane:
Encryption, authentication and authorization
On-demand and advanced provisioning

Dual OC192 backbone:
SONET-switched in the backbone
Ethernet-SONET conversion
USN Data-Plane: Node Configuration

In the Core:
- Two OC192 switched by Ciena CDCIs at OC3c resolution

At the Edge
- 10/1 GigE provisioning using Force10 E300s using VLANs

Data Plane User Connections:
Direct connections to:
- core switches –SONET & 1GigE
- MSPP – Ethernet channels (VLANs)
- Utilize USN hosts
Secure 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
Also used as management plane
USN Control Plane

- **Phase I (2004-2005)**
  - Centralized path computation for bandwidth optimization
  - TL1/CLI-based communication with CoreDirectors and E300s
  - User access via centralized web-based scheduler

- **Phase II (2006)**
  - Webservices interface
  - X509 authentication for web server and service

- **Phase II (current)**
  - GMPLS wrappers for TL1/CLI
  - Inter-domain “secured” GMPLS-based interface

Both use USN SSL Certificates for authorization

Webpage for manual bandwidth reservation

WSDL for webservice bandwidth reservation

SDL for webservice
Outline

USN Architecture and Deployment

Testbed Network Research Activities
  Peering and Alignment - UVA
  VLANS Testing and Analysis – ISI, TnTech
  Infiniband over SONET – ORNL/NLCF
  Microscope Control – PNL
Data-Plane Alignment Over Layer 1-3 Networks

Peering data-paths across networks that provide over Layer 1-3 connections:
- SONET layer-1 paths
- MPLS layer-3 paths

Broader Networking Question:
Layer-1 or layer-2 or layer-3 channels for dedicated bandwidth connections?

Feasibility and Performance of Composed SONET-MPLS VLANS:
Data-plane unification of dedicated paths over layer-1, layer-2 and layer-3 paths

Broader Question: Peering data-paths across networks that provide VLANs over Layer-1 or layer-2 or layer-3

Systematic analysis of application and IP level measurements:
Using USN, ESnet and CHEETAH, we collected iperf, ping and TCP measurements performed comparative performance analysis composed and tested VLANS over SONET and IP connections
VLAN – Unifying Data-Plane Technology for Peering Layer 1-2 and 3 Networks

- IP networks
  - VLANs Implemented in MPLS tunnels
- Circuit switched networks
  - VLANs Implemented on top of Ethernet or SONET channels
- Align IP and circuit connections at VLAN level

![Diagram of VLAN over MPLS and Circuit Switched networks](image)
USN–ESnet VLAN through L3-L2 paths
Collaborators: Chin Guok, Eli Dart (ESnet)

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

1Gbps channel over USN and ESnet - cross-connect on e300

UltraScience Net: Layer-2
E300 - CDCI - ... - CDCI - E300

OC21c 700, 2100, 3500, 4900 miles
1Gbps layer-2 connection Ethernet over SONET
USN– CHEETAH VLAN through L3-L2 paths
Collaborators: Malathi Veeraraghvan, U. Virginia

UltraScience Net: Layer-2
VLAN: E300 – CDCI – … - CDCI – E300
CHEETAH: layer-3 + layer 2
VLAN: T640-T640 - SN16000 - Cisco 3750

Coast-to-cost 1Gbps channel demonstrated over in 2006
USN and CHEETAH – simple cross-connect on e300
Layer-3 and Layer-2 Connections: iperf TCP Throughput Measurements

No. streams 1-10 repeated 100 times

**Comparison**

**ESnet**
Chicago-Sunnyvale

Layer-3:
- MPLS tunnel
- Ping: 67ms
- ~3000 miles

Layer-2:
- 906 Mbps
Layer-3:
- 852 Mbps

**USN**
ORNL-Chicago...

Layer 2 over OC21c
- Ethernet over SONET
- Ping: 66ms
- ~3500 miles

---

OAK RIDGE NATIONAL LABORATORY
U. S. DEPARTMENT OF ENERGY

UT-BATTENELL
Layer 3 and Layer 2 Connections: iperf UDP Throughput Measurements
Target bw: 100, 200, ..., 1100 repeated 100 times

Comparison
Profiles are almost identical
Layer 2: 958 Mbps
Layer 3: 953 Mbps

ESnet
Chicago-Sunnyvale
Layer-3: MPLS tunnel
Ping: 67ms
~3000 miles

USN
ORNL-Chicago...
- ORNL-Chicago
Layer 2 over OC21c
Ethernet over SONET
Ping: 66ms
~3500 miles

throughput

throughput

通过put

target b/w
repetition

throughput

target b/w
repetition
1GigE Over SONET: USN test configurations

- ORNL - Chicago - 700 miles
- Multiple loops: 2100, 3500, 4900, 6300 miles
USN Connections:
iperf TCP Measurements

Throughput profile “lowers” with distance:
Linux-host – e300 – CDCI …
   … CDCI – e300 – linux host
   2100 miles
   3500 miles
   4900 miles
   6300 miles
Throughput profile “lowers” with distance:

ESnet: T320 – T320 – Cisco6509
USN: e300 – CDCI … CDCI – e300

700+3000 miles
2100 + 3000 miles
3500 + 3000 miles
4900 + 3000 miles
6300 + 3000 miles
Connection Profile: Window-based UDP transport
Collaboration with Qishi Wu, University of Memphis

Layer-3:
MPLS tunnel
Ping: 67ms
~3000 miles

Layers 1-3:
Hybrid connection
Ping: 67ms
~3500 miles

Layer 2 over OC21c
Ethernet over SONET
Ping: 134ms
~6500 miles
Peak Link Utilization Protocol
Collaboration with Qishi Wu, University of Memphis

ESnet
Chicago-Sunnyvale

ESnet-USN
ORNL-Chicago-Sunnyvale

USN
ORNL-Chicago-Sunnyvale
1GigE Over SONET: USN test configurations
Collaborators: Tom Lehman, HOPI; Chin Guok, OSCARS; Nasir Ghani, Tn Tech

ORNL

Linux host

E300

Copper GigE

Linux host

E300

ORNL

700 miles

CDCI

ORNL

3300 miles

CDCI

Chicago

4300 miles

CDCI

Seattle

Sunnyvale

E300

E300

ORNL – Chicago - loop – 1400 miles

Multiple loops: 1400, 2800, 4200, 5600, 7000, 8400, 9800, 11200, 12600 miles

ORNL – Chicago – Seattle – Sunnyvale - loop – 8600 miles

Multiple loops: 8600, 17200, 25800, 34400 miles
USN test configurations: Ping RTT

ORNL - Chicago - Seattle - Sunnyvale - loop - 8600 miles

<table>
<thead>
<tr>
<th>miles</th>
<th>rtt (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8,600</td>
<td>163</td>
</tr>
<tr>
<td>17,200</td>
<td>327</td>
</tr>
<tr>
<td>25,800</td>
<td>490</td>
</tr>
<tr>
<td>34,400</td>
<td>653</td>
</tr>
</tbody>
</table>

ORNL - Chicago - loop - 1400 miles

<table>
<thead>
<tr>
<th>miles</th>
<th>1,400</th>
<th>2,800</th>
<th>4,200</th>
<th>5,600</th>
<th>7,000</th>
<th>8,400</th>
<th>9,800</th>
<th>11,200</th>
<th>12,600</th>
</tr>
</thead>
<tbody>
<tr>
<td>rtt (ms)</td>
<td>26.79</td>
<td>53.4</td>
<td>79.90</td>
<td>106</td>
<td>132</td>
<td>159</td>
<td>185</td>
<td>212</td>
<td>238</td>
</tr>
</tbody>
</table>
Jitter Measurements Suite

1. TCP client-server: client sends a message and server echoes back
2. Tcpmon: client sends a message size and server sends the message
3. Ping

5600 miles 1GigE VLAN
Four 1400 mile loops
USN: ORNL-Chicago OC192
Jitter Regression Band

Concept:
Regression Fit – delays vs connection length $\bar{D}(l)$

Decompose measurements into above and below regression
Separate Upper and Lower regression fit
$\bar{D}_U(l)$ $\bar{D}_L(l)$

Regression band if difference between upper and lower regression functions
$\bar{B}_{UL}(l) = \bar{U}(l) - \bar{L}(l)$
Measurements Summary – 12,000 miles

VLAN on SONET OC-21
Multiple USN ORNL-Chicago loops
Each loop is 1400 miles

<table>
<thead>
<tr>
<th>miles</th>
<th>1,400</th>
<th>2,800</th>
<th>4,200</th>
<th>5,600</th>
<th>7,000</th>
<th>8,400</th>
<th>9,800</th>
<th>11,200</th>
<th>12,600</th>
</tr>
</thead>
<tbody>
<tr>
<td>rtt (ms)</td>
<td>26.79</td>
<td>53.4</td>
<td>79.90</td>
<td>106</td>
<td>132</td>
<td>159</td>
<td>185</td>
<td>212</td>
<td>238</td>
</tr>
</tbody>
</table>
Jitter regression bands for VLANs on SONET paths: [5600, 12,600] miles

Jitter regression band is narrow: weighted with frequency

Ping measurements are constant -1000 times

\[ \overline{B}_{UL}(l) = \overline{U}(l) - \overline{L}(l) \]
Normalization Framework

**Basic Question:** Measurements are collected on two connections of different lengths and types. How do we objectively compare them?

**Example:** Ping measurements on 1000 mile SONET-VLAN and 300 mile MPLS-VLAN, can we objectively conclude about jitter on such VLANs?

\[
\begin{align*}
M_T(d) & \quad \text{Measurements on path of type } T \text{ of distance } d \\
\hat{M}_T(d) & \quad \text{Estimates of measurements on path of type } T \text{ of distance } d \\
P \odot \hat{M}_T(d) & \quad \text{Parameters computed using measurements}
\end{align*}
\]

\[
\begin{align*}
M_{T_1}(d_1) & \quad \Theta_{T_1} \quad \text{Interpolation based on regression} \\
\hat{M}_{T_1}(d) & \quad \Theta_{T_2} \\
P \odot \hat{M}_{T_1}(d) & \quad \mathcal{N}_P \\
P \odot \hat{M}_{T_2}(d)
\end{align*}
\]
Jitter Comparison on SONET-MPLS VLANs

- USN ORNL-Chicago 1Gig VLAN on SONET – 1400 miles
  - E300- CDCI – CDCI – E300
- ORNL ATL sox 1Gig production IP connection – 300 miles
  - T640 – T640

\[ M_{\text{SONET}}(1400) \]

Interpolation based on linear regression

\[ M_{\text{MPLS}}(300) \]

Align jitter regression band

\[ \Theta_{T_1} \]

\[ \Theta_{T_2} \]

Identity

\[ \hat{M}_{\text{SONET}}(300) \]

\[ \hat{M}_{\text{MPLS}}(300) \]

Another Method

\[ P \rightarrow \text{FFT} \]

\[ \hat{M}_{\text{SONET}}(300) \]

\[ \hat{M}_{\text{MPLS}}(300) \]

\[ P \circ \hat{M}_{\text{SONET}}(300) \]

\[ P \circ \hat{M}_{\text{MPLS}}(300) \]

\[ \Lambda_P \]

\[ \Lambda_P \rightarrow \text{Identity} \]
Composed VLAN:
SONET and Layer-3 Channels - Gig 1300 miles

Number of measurements = 999
mean ping time = 35.981812
percent range: [99.772635, 100.328463]
range: [35.900002, 36.099998]: 0.199997
std_deviation (percent) = 0.151493
Comparison of VLANs: SONET vs. MPLS tunnels

Measurements are normalized for comparison:

- **SONET**
  - Mean time: 26.845877ms
  - Percent range: [99.8, 100.6]
  - Std. dev (%): 0.187035

- **IP-MPLS**
  - Mean time: 9.384557ms
  - Percent range: [99.4, 203.5]
  - Std. dev (%): 3.281692

**Conclusion**

VLANs over SONET have smaller jitter levels.
Normalized Comparison of VLANs:
SONET - SONET-MPLS composed – L2MPLS

Measurements are normalized for comparison:

- **SONET channels** have smaller jitter levels

  - **SONET**
    - mean time = 26.845877ms
    - std_dev (%) = 0.187035

  - **SONET-MPLS composite**
    - mean time = 35.981812ms
    - std_dev (%) = 0.151493

  - **L2MPLS**
    - mean time = 9.384557ms
    - std_dev (%) = 3.281692
Infiniband Over SONET – Joint with ORNL/NLCF

collaborators: Makia Minich, Feiyi Wang, Steven Carter

• Infiniband is effective data transport protocol for storage networks (few miles):
• TCP is not easily extended or not optimal for such data transfers

Question: Is IB effective over wide-area? - Yes

IB 4x: 8Gbps (full speed)
Host-to-host local switch: 7.5Gbps
Infiniband Over SONET

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

ORNL loop -0.2 mile: 7.5Gbps

ORNL-Chicago loop – 1400 miles: 7.46Gbps

ORNL- Chicago - Seattle loop – 6600 miles: 7.23Gbps

ORNL – Chicago – Seattle - Sunnyvale loop – 8600 miles: 7.20Gbps
Remote Microscope Control Over USN
PNL Project: Tom McKenna

ORNL

 OC192 700 miles

Chicago CDCI

3300 miles

Seattle CDCI

PNL

E300

E300

10GigE Windows host

Windows sever

Demonstrated control of confocal microscopes between ORNL and PNL
Viewed subject online
Controlled microscope view

10GigE

3300 miles

10GigE

Demonstrated control of confocal microscopes between ORNL and PNL
Viewed subject online
Controlled microscope view
Conclusions – UltraScience Net

• USN Infrastructure development is close to completion
• USN continues its research role in advanced networking projects:
  – Networking Technologies for Leadership Class Facilities
    • Connectivity to supercomputers
    • Testing of storage/file systems: Infiniband, Lustre
  – Integrated Multi-Domain Interoperation System for USN-ESnet-CHEETAH-HOPI
    • on-going efforts with OSCARS and HOPI
  – Control of confocal microscopes at PNL
    • Configuration and testing of hybrid connections

USN website:  www.csm.ornl.gov/ultranet

Email: Nagi Rao (raons@ornl.gov) or Bill Wing (wrw@ornl.gov)