UltraScience Net:
High-Performance Network Research Test-Bed

Nagi Rao, Bill Wing,
Susan Hicks, Steve Poole, Frank Denap
Oak Ride National Laboratory
raons@ornl.gov

Steven Carter
Cisco Systems

Qishi Wu
University of Memphis

July 21, 2009
5th GENI Engineering Conference, Seattle, WA

Sponsored by
U.S. Department of Defense
U.S. Department of Energy

Outline

- **Motivation and Background**
- **USN infrastructure**
  - Architecture
  - Data-plane
  - Control-plane
  - Connection Suites
- **USN Networking Experiments**
  - Hybrid Network Connections
  - Infiniband over Wide-Area
  - Connections to Supercomputers
  - Transport Methods for Dedicated Channels
  - Wide-Area Application Accelerators
  - Encryption Devices

Motivation

UltraScience Net

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

OAK RIDGE
National Laboratory

Outline

- **Motivation and Background**
- **USN infrastructure**
  - Architecture
  - Data-plane
  - Control-plane
  - Connection Suites
- **USN Networking Experiments**
  - Hybrid Network Connections
  - Infiniband over Wide-Area
  - Connections to Supercomputers
  - Transport Methods for Dedicated Channels
  - Wide-Area Application Accelerators

UltraScience Net – In a nutshell

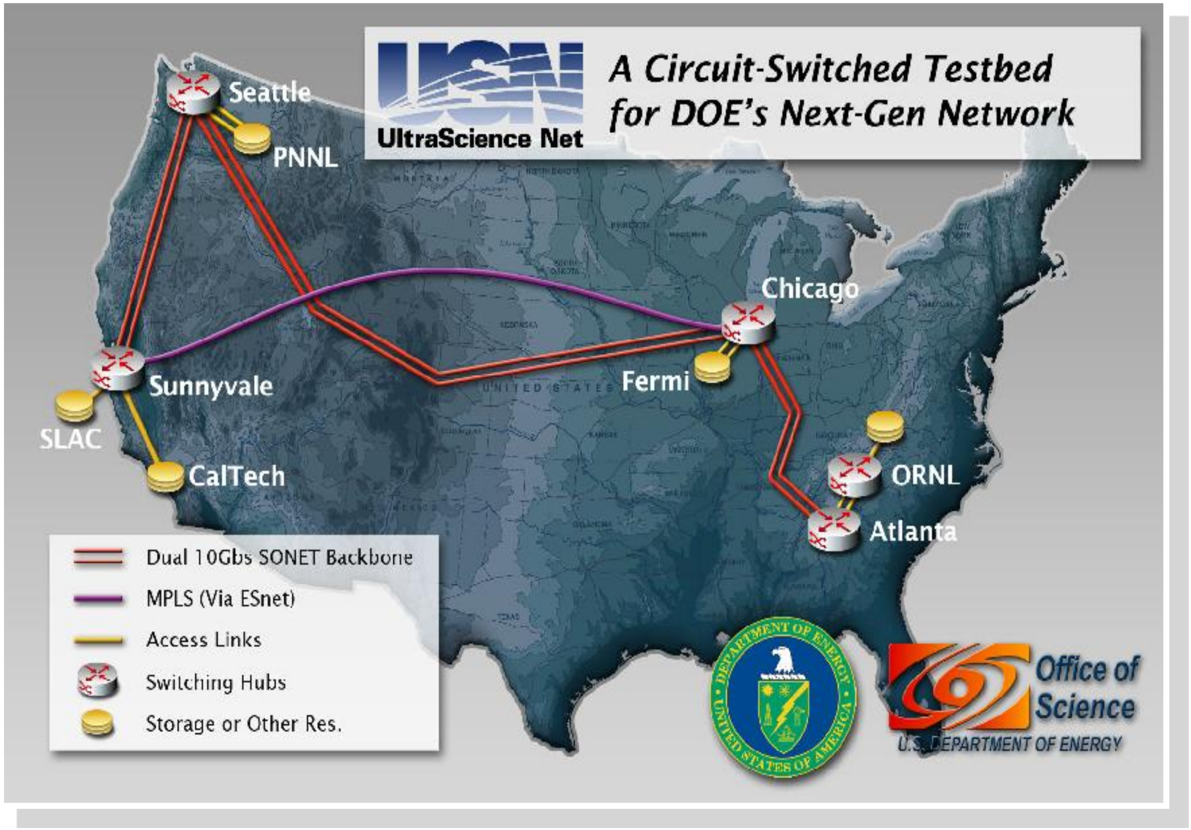
Experimental network research testbed:

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

Currently funded by Department of Defense; by Department of Energy (2004-2007)

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: National Leadership Computing Facility, Fermi National Laboratory, National Energy Research Scientific Computing
- Peering with ESnet, National Science Foundation CHEETAH, and other networks



USN Contributions

Network research testbed for high-performance networking

  - dedicated connections between limited number of sites – not for Internet

- Provides long haul production links for experimentation
  - 8000 mile 10Gbps and 70,000 mile 1Gbps connections
  - Automated scripts for testing over multiple connections

2004

- First advanced reservation and scheduling of dedicated connections
  - Showed the problem to be polynomial-time solvable
  - Deployed in USN control plane in 2005 – demonstrated at SC2005

2005

- Identified network throughput bottlenecks in dedicated connections supercomputers

2007

- Peering of layer-2 and layer-3 networks using VLANS:
  - coast-to-coast connections over USN, Esnet and CHEETAH
- Infiniband extensions to thousands of miles
  - IB-RDMA throughputs: local 7.6 Gbps: 8600 miles: 7.2 Gbps: SC2008

2008

- 10Gbps Crypto devices
  - TCP performance improved: higher throughput with less #streams

2009

Outline

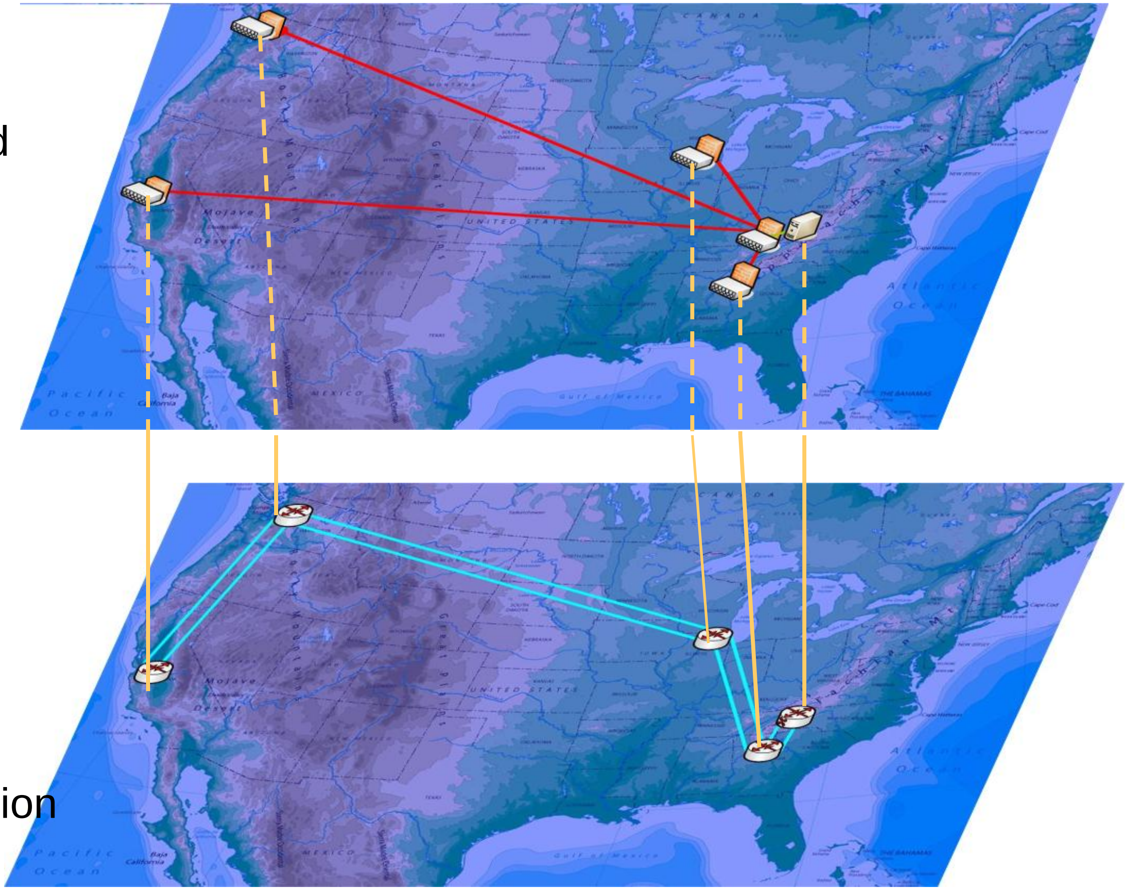
- **Motivation and Background**
- **USN infrastructure**
  - Architecture
  - Data-plane
  - Control-plane
  - Connection Suites
- **USN Networking Experiments**
  - Hybrid Network Connections
  - Infiniband over Wide-Area
  - Connections to Supercomputers
  - Transport Methods for Dedicated Channels
  - Wide-Area Application Accelerators
  - Encryption Devices

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 with:
- Encryption, authentication and authorization
- On-demand and advanced provisioning

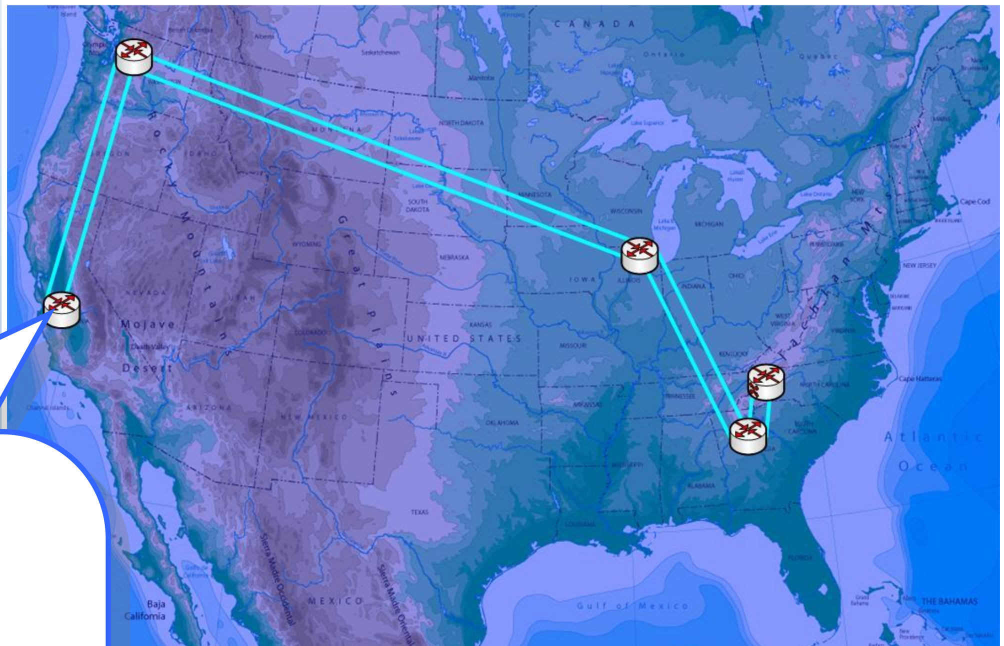
GMPLS in IP is not secure enough:
- Messages can be sniffed
- Control messages can be injected



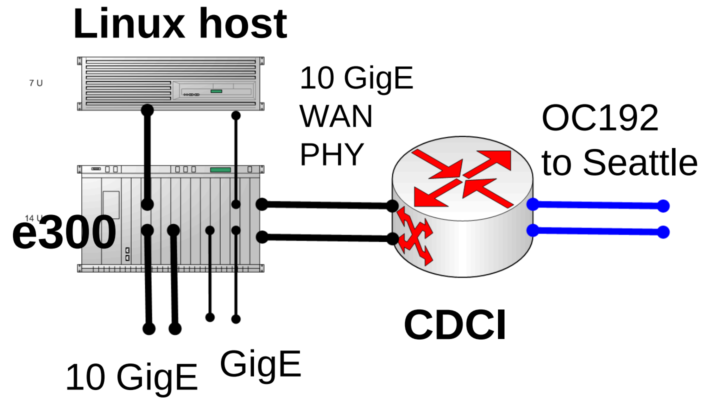
Dual OC192 backbone:
- SONET-switched in the backbone
- Ethernet-SONET conversion

USN data-plane: Node configuration

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



Node Configuration



Connections to CalTech and ESnet

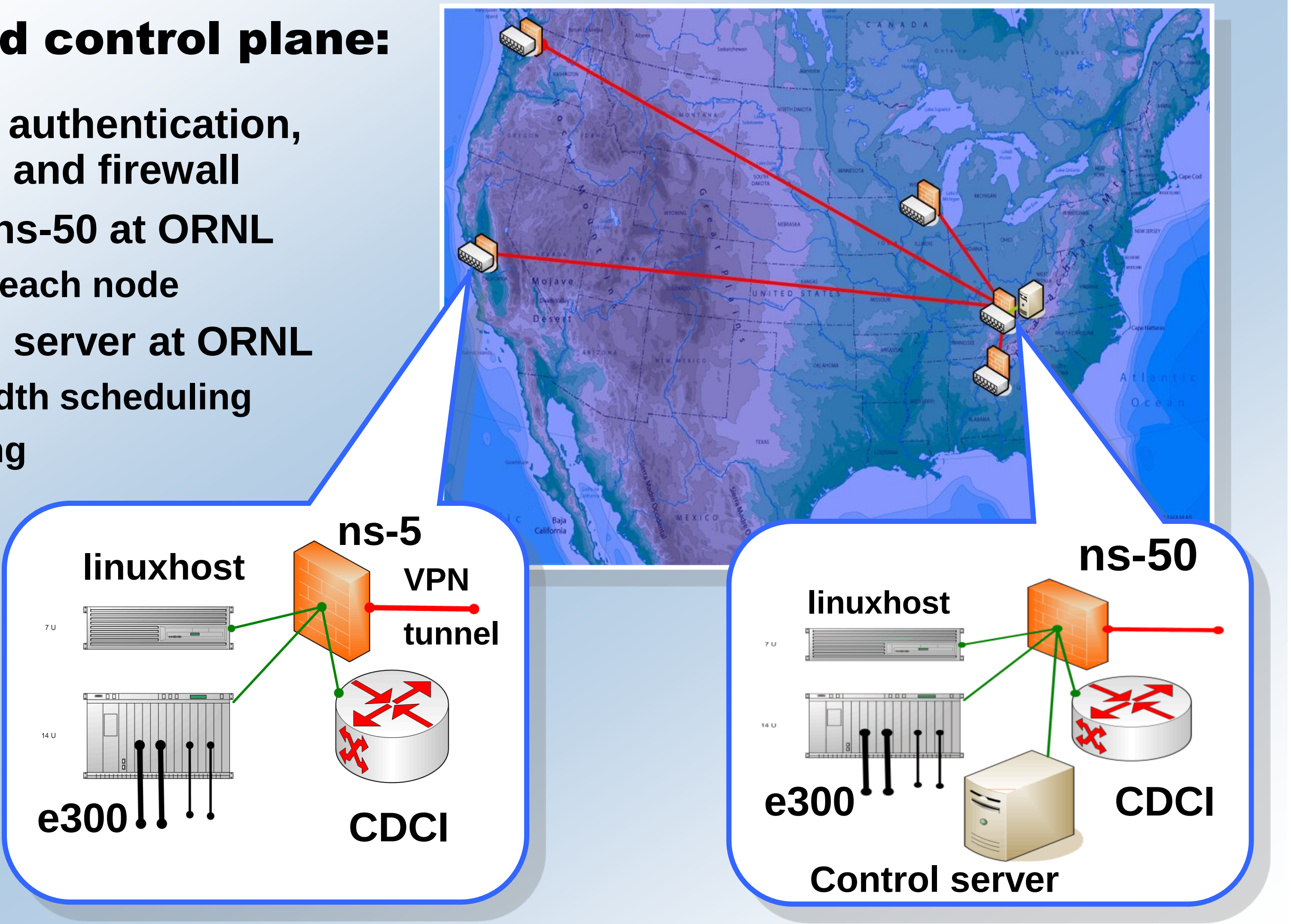
Data plane user connections:

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

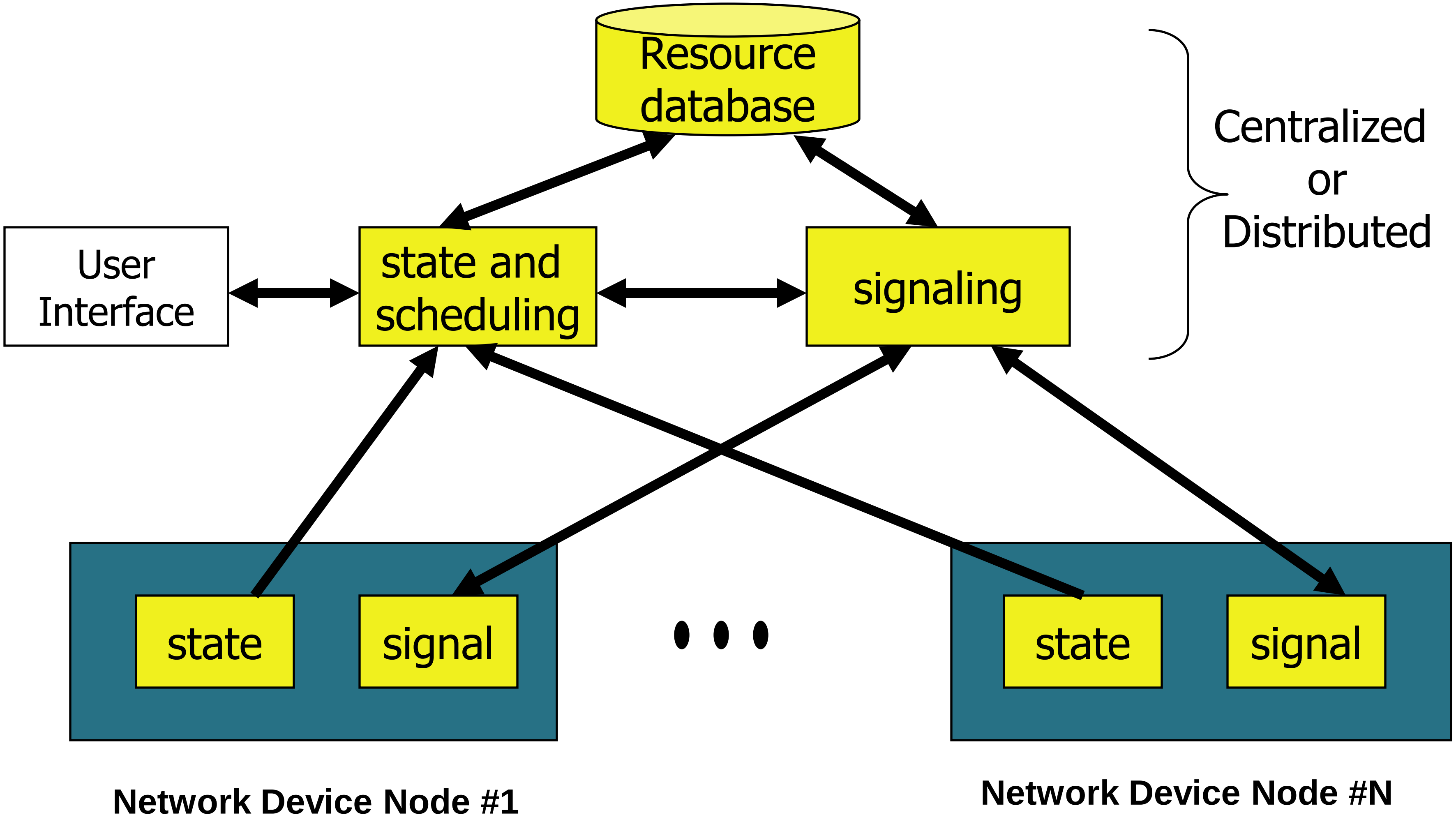
Secure control plane

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



A General Control-Plane Architecture



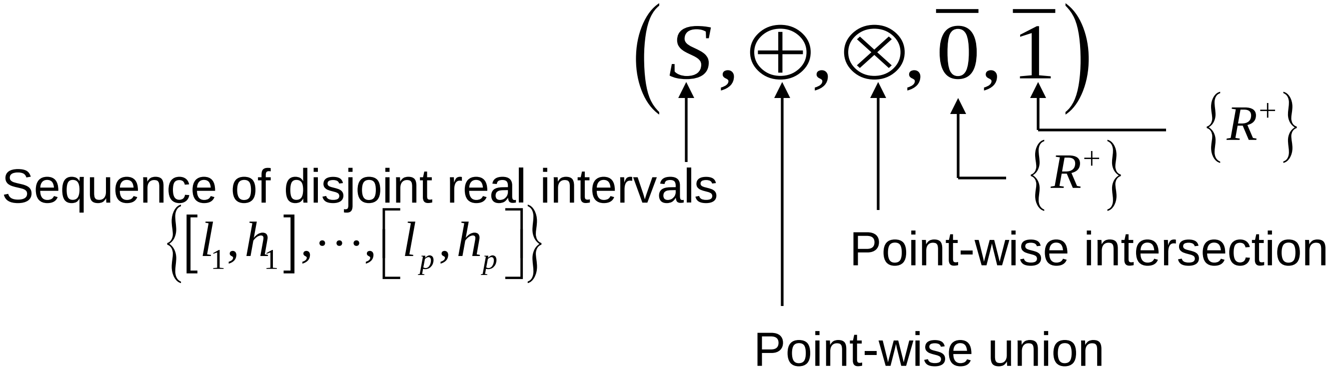
USN Path Computation – Bandwidth Optimization Collaboration with Sartaj Shani

Different paths may be computed: specify source and destination ports

- (i) A specified bandwidth in a specified time slot,
- (ii) Earliest available time with a specified bandwidth and duration,
- (iii) Highest available bandwidth in a specified time slot,
- (iv) All available time slots with a specified bandwidth and duration.

All are computed by extending the shortest path algorithms using a closed semi-ring structure defined on sequences of real intervals

- (i)-(ii): Extended breadth-first search algorithm
- (iii)-(iv): Variation of Bellman-Ford algorithm;
  - previously solved using transitive-closure algorithm



$$\left(S, \oplus, \otimes, \overline{0}, \overline{1}\right)$$

Sequence of disjoint real intervals

$\left\{[l_1,h_1],\cdots,[l_p,h_p]\right\}$

Point-wise union

Point-wise intersection

$\left\{R^+\right\}$

$\left\{R^+\right\}$

All-Slots Algorithm

Given network with bandwidth allocations on all links

ALL-SLOTS returns all possible starting times for a connection with bandwidth *b* duration *t* between source node *s* and destination node *d*

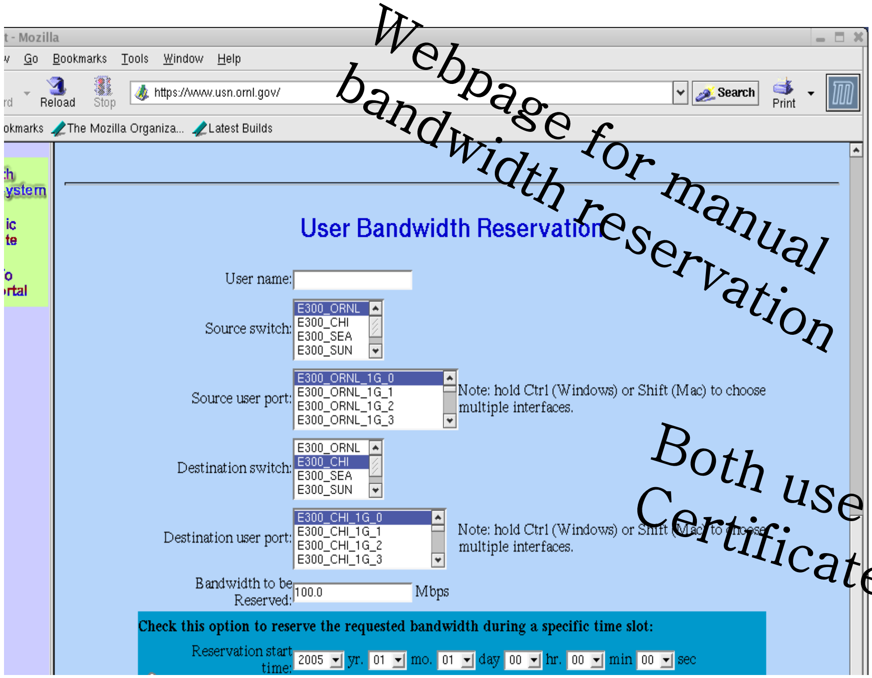
```
Algorithm ALL-SLOTS
$\tau(s) \leftarrow \{\Re\}$;
$\tau(v) \leftarrow \{\emptyset\}$ for all $v \neq s$;
for $k = 1, 2, \ldots, n-1$ do
    for each edge $e = (v, w)$ do
        $\tau(w) \leftarrow \tau(w) \oplus \{\tau(v) \otimes L_e\}$;
return $(\tau(d))$.
```

Modified Bell-Ford algorithm:
Time-complexity: $O(mn)$

More efficient than transitive-closure algorithm: $O\left(n^3\right)$

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 (2007-2009)
  - GMPLS wrappers for TL1/CLI
  - Inter-domain “secured” GMPLS-based interface



https://www.usn.ornl.gov/

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

Webpage for manual bandwidth reservation

http://ozy8/nusoap/server.php?wsdl

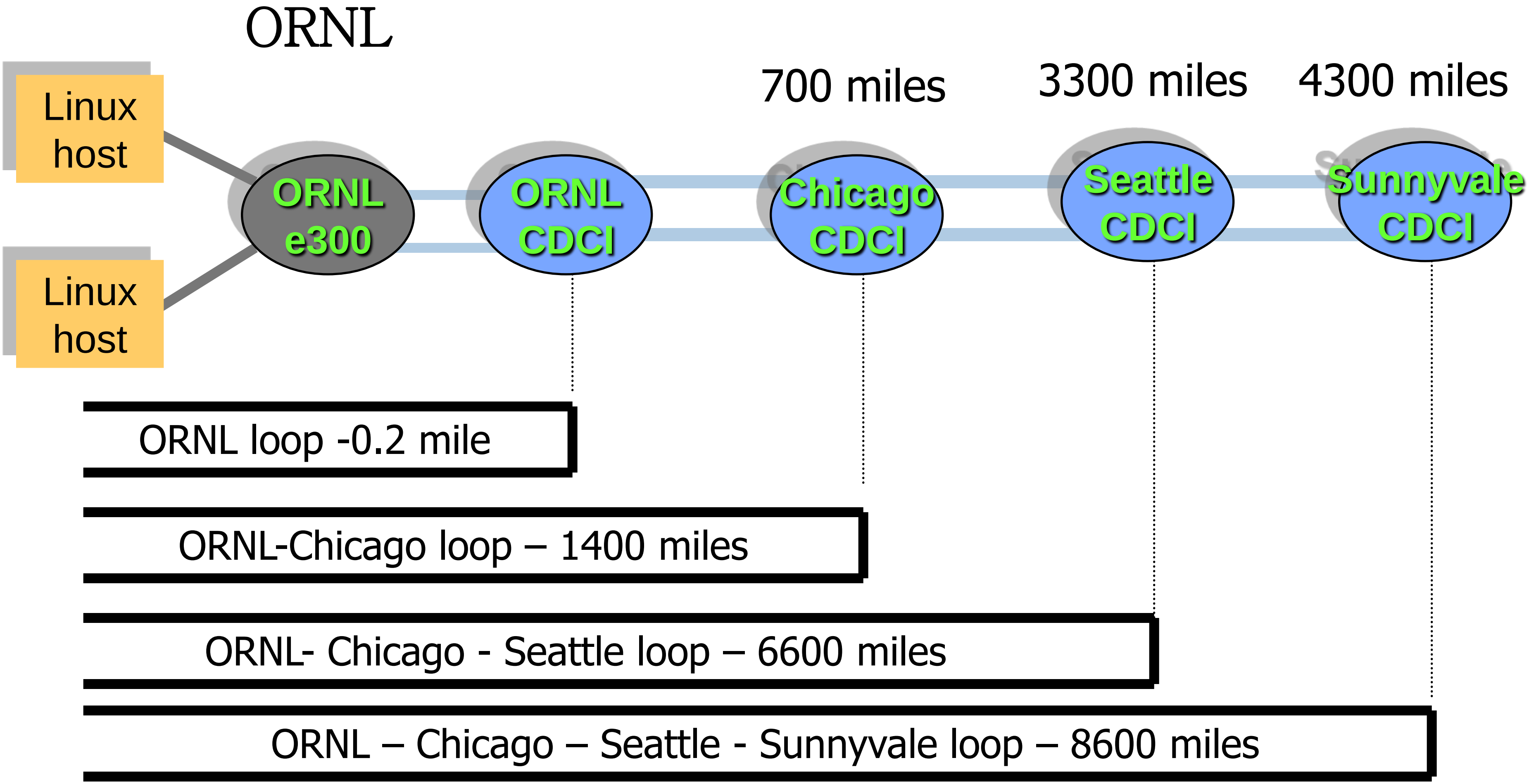
ot appear to have any style information associated with it. The document tree is shown below.

```
lamespace="urn:reservation">

rgetNamespace="urn:reservation">
namespace="http://schemas.xmlsoap.org/soap/encoding/"/>
namespace="http://schemas.xmlsoap.org/wsdl/"/>

createReservationRequest"></message>
createReservationResponse">
turn" type="xsd:string"/>

"reservationsd1PortType">
ie="createReservation">
ion>Make Bandwidth Reservation</documentation>
ge="tns:createReservationRequest"/>
age="tns:createReservationResponse"/>

eservationsd1Binding" type="tns:reservationsd1PortType">
style="rpc" transport="http://schemas.xmlsoap.org/soap/http"/>
ie="createReservation">
ion soapAction="urn:reservationwsd1#hello" style="rpc"/>

y use="encoded" namespace="urn:reservationwsd1" encodingStyle="http://schemas.xmlsoap.org/soap/encoding/"/>

y use="encoded" namespace="urn:reservationwsd1" encodingStyle="http://schemas.xmlsoap.org/soap/encoding/"/>

eservationsd1">
servationsd1Port" binding="tns:reservationsd1Binding">
s location="http://ozy8/nusoap/server.php"/>
```

WSDL for webservice Bandwidth reservation

Both use USN SSL Certificates for authorization

OC192 SONET Connections



ORNL

Linux host

Linux host

ORNL e300

ORNL CDCI

700 miles

Chicago CDCI

3300 miles

Seattle CDCI

4300 miles

Sunnyvale CDCI

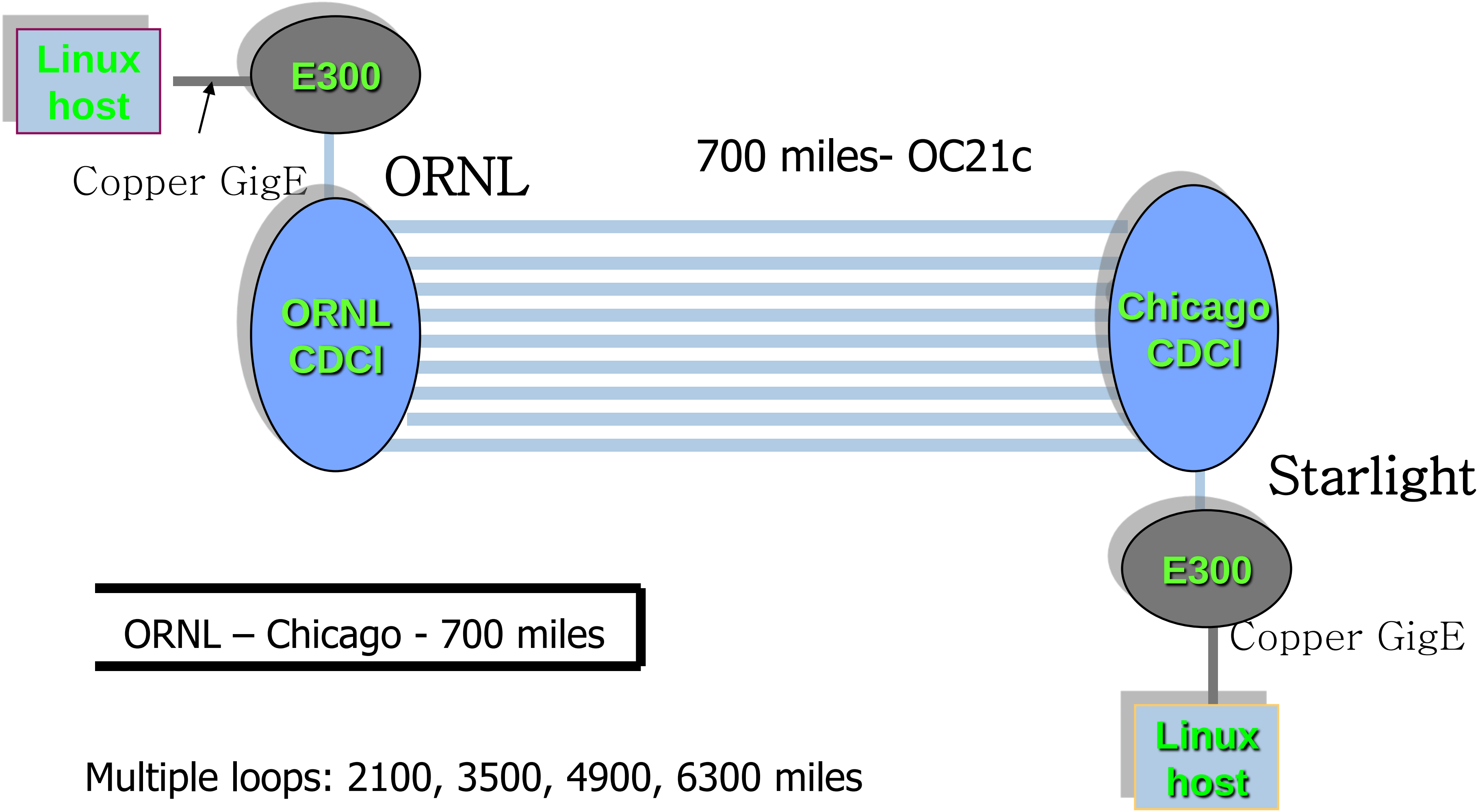
ORNL loop -0.2 mile

ORNL-Chicago loop – 1400 miles

ORNL- Chicago - Seattle loop – 6600 miles

ORNL – Chicago – Seattle - Sunnyvale loop – 8600 miles

OC21c SONET: USN test configurations



Linux host

E300

Copper GigE

ORNL

700 miles- OC21c

ORNL CDCI

Chicago CDCI

Starlight

E300

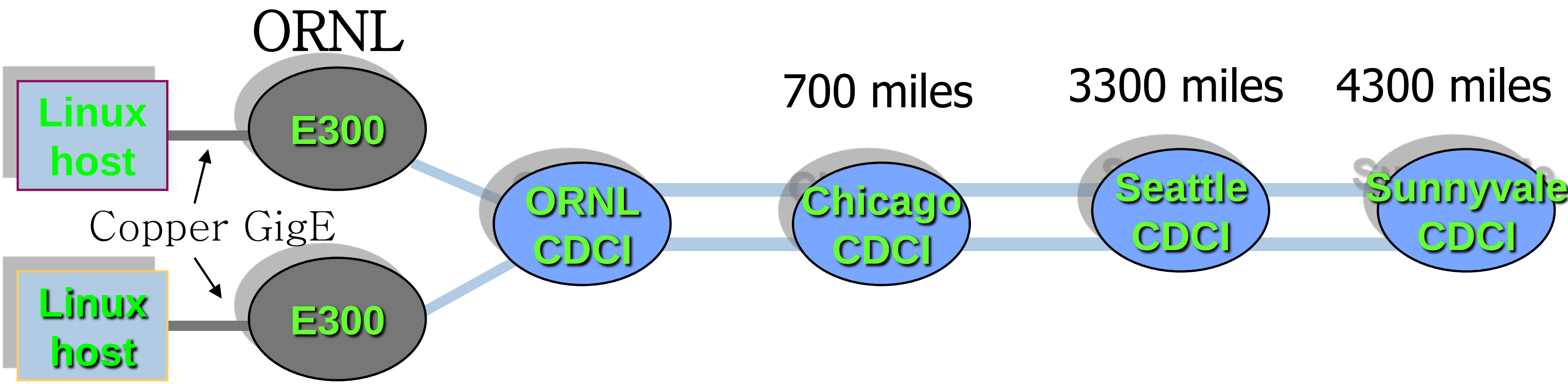
Copper GigE

Linux host

ORNL – Chicago - 700 miles

Multiple loops: 2100, 3500, 4900, 6300 miles

1GigE Over SONET: USN test configurations



ORNL

Linux host

Linux host

Copper GigE

E300

E300

ORNL CDCI

700 miles

Chicago CDCI

3300 miles

Seattle CDCI

4300 miles

Sunnyvale CDCI

ORNL – Chicago - loop – 1400 miles

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

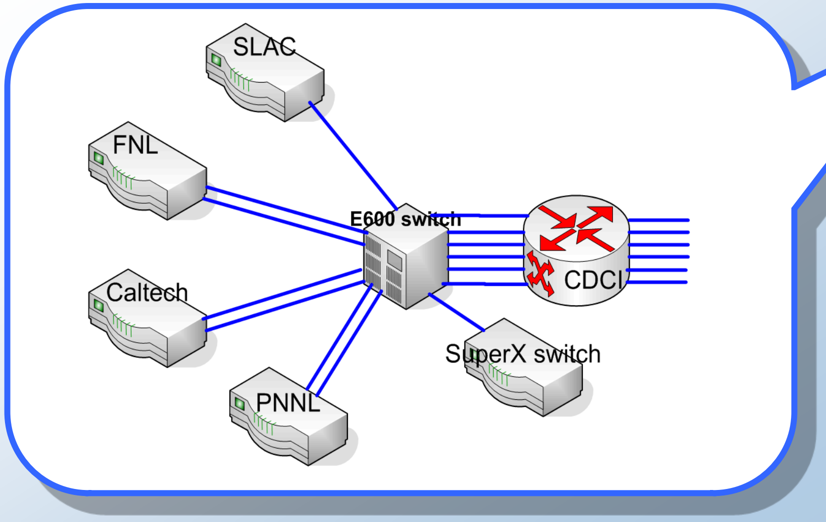
ORNL – Chicago – Seattle – Sunnyvale - loop – 8600 miles

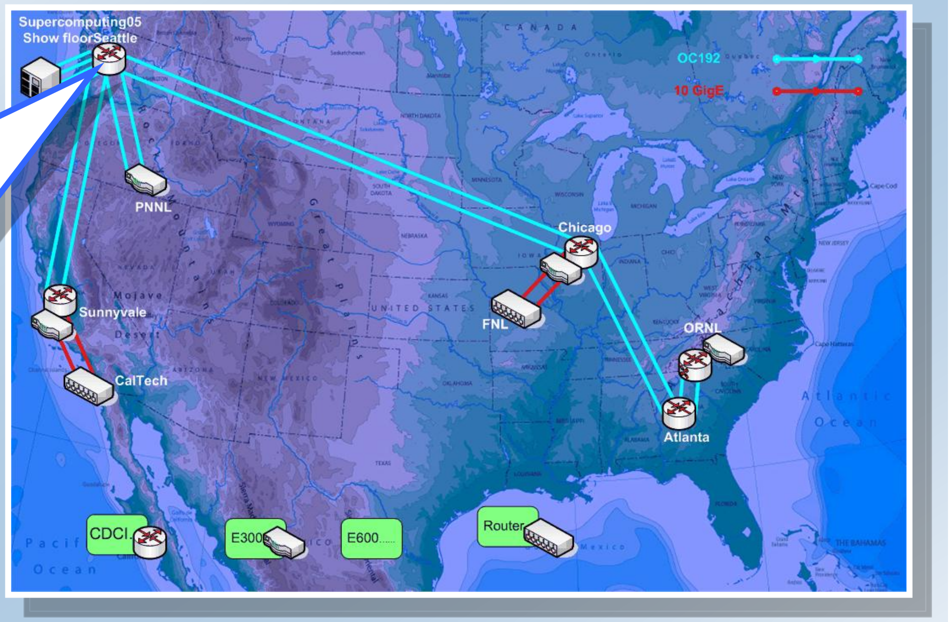
Multiple loops: 8600, 17200, 25800, 34400 miles

Around the earth once

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 re-created USN-Seattle node on
- Pacific Northwest National Laboratory, FNL, ORNL, Caltech, Stanford Linear Accelerator Center at various booths to support:
  - applications and bandwidth challenge

Helped Caltech team win Bandwidth Challenge:

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

USN
UltraScience Net

OAK RIDGE
National Laboratory

Outline

- Motivation and Background
- USN infrastructure
  - Architecture
  - Data-plane
  - Control-plane
  - Connection Suites
- USN Networking Experiments
  - Hybrid Network Connections
  - Infiniband over Wide-Area
  - Connections to Supercomputers
  - Transport Methods for Dedicated Channels
  - Wide-Area Application Accelerators
  - Encryption Devices

Interoperability data-planes of different networks

Another way of providing dedicated connections (layer 3):

- Multiple Protocol Label Switching (MPLS) tunnels over IP routers

Important question:

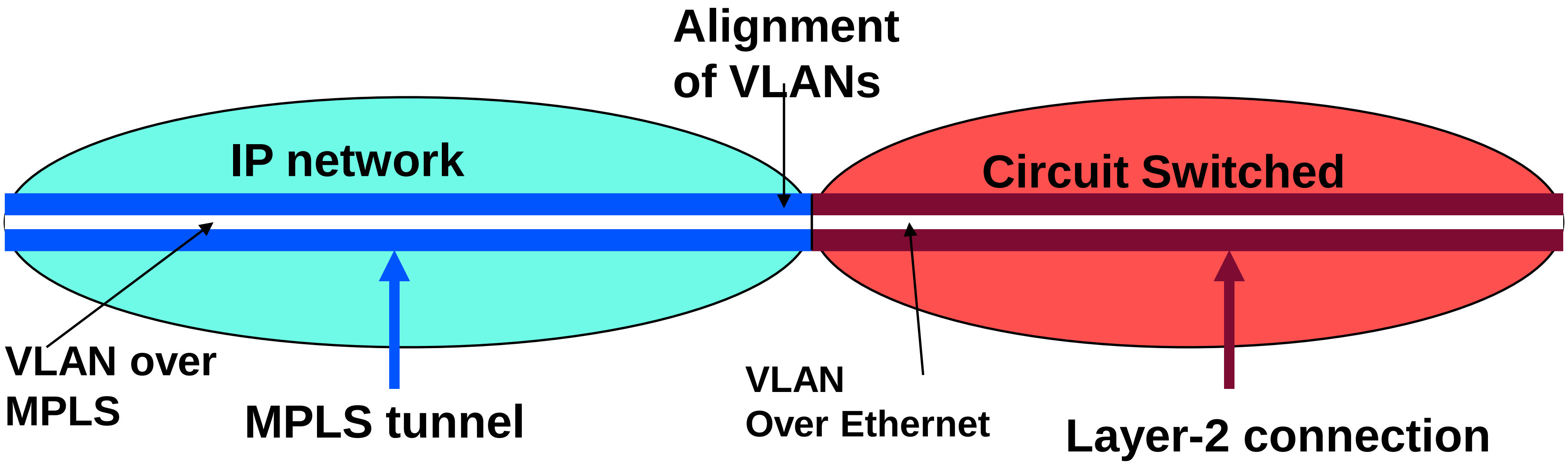
- Peering of dedicated paths provisioned at layers 1 through 3?

Virtual Local Area Network (VLAN) technologies provide a solution

- VLANs are typically native to layer-2: other layers need to be moved up/down to implement VLANs:
  - SONET connections (layer1): VLANs are provisioned using edge switches (E300 in our case)
  - Layer-2 connections – VLANs are provisioned natively
  - IP networks (layer 3) – VLANs are provisioned over MPLS tunnels using IEEE 802.1q – router implementations differ

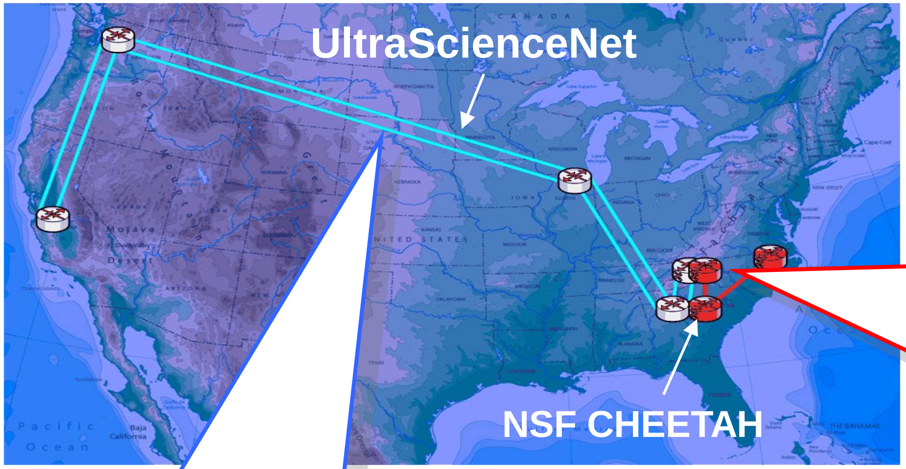
VLAN – Unifying Data-Plane Technology for Peering Layer 1-2 and 3 Networks

- 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

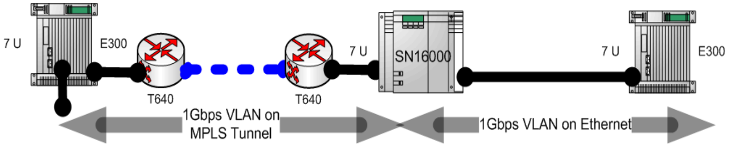


Demonstrated peering circuit-packet switched networks:

USN–CHEETAH VLAN through L3-L2 paths

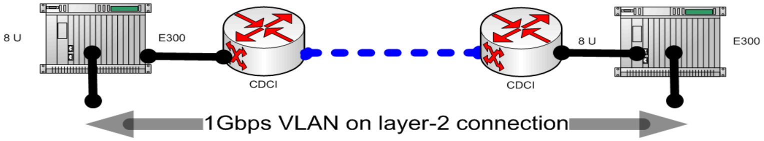


CHEETAH VLAN



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

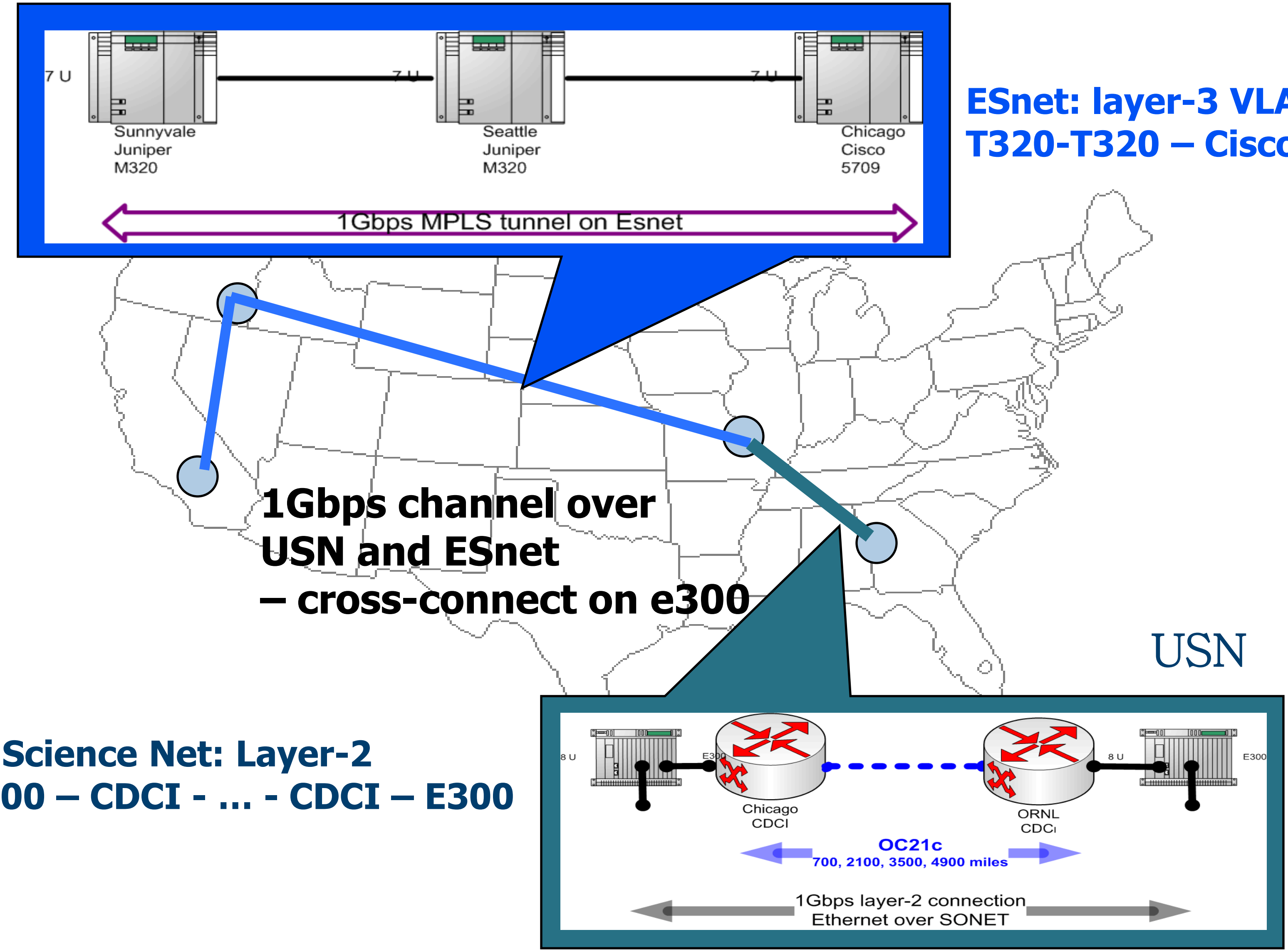
- DOE UltraScience Net: Layer-2
  - VLAN: E300–CDCI - … -CDCI–E300



USN VLAN

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

USN–ESnet Peering of L2 and L3 paths



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

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

USN

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

Performance of Dedicated Channels

Relative performance of VLANs provisioned over:

SONET: layer-1 – Ethernet: layer-2 – MPLS: layer-3

Building networks to provide dedicated channels:

Which layer to build? layer-1, 2, 3 or mixed?

Layer-1: Most "separated" and flexible

Layer-2: Cheapest to build from scratch

Layer-3: Cheapest if IP infrastructure already exists

Performance of Composed SONET-MPLS VLANS:

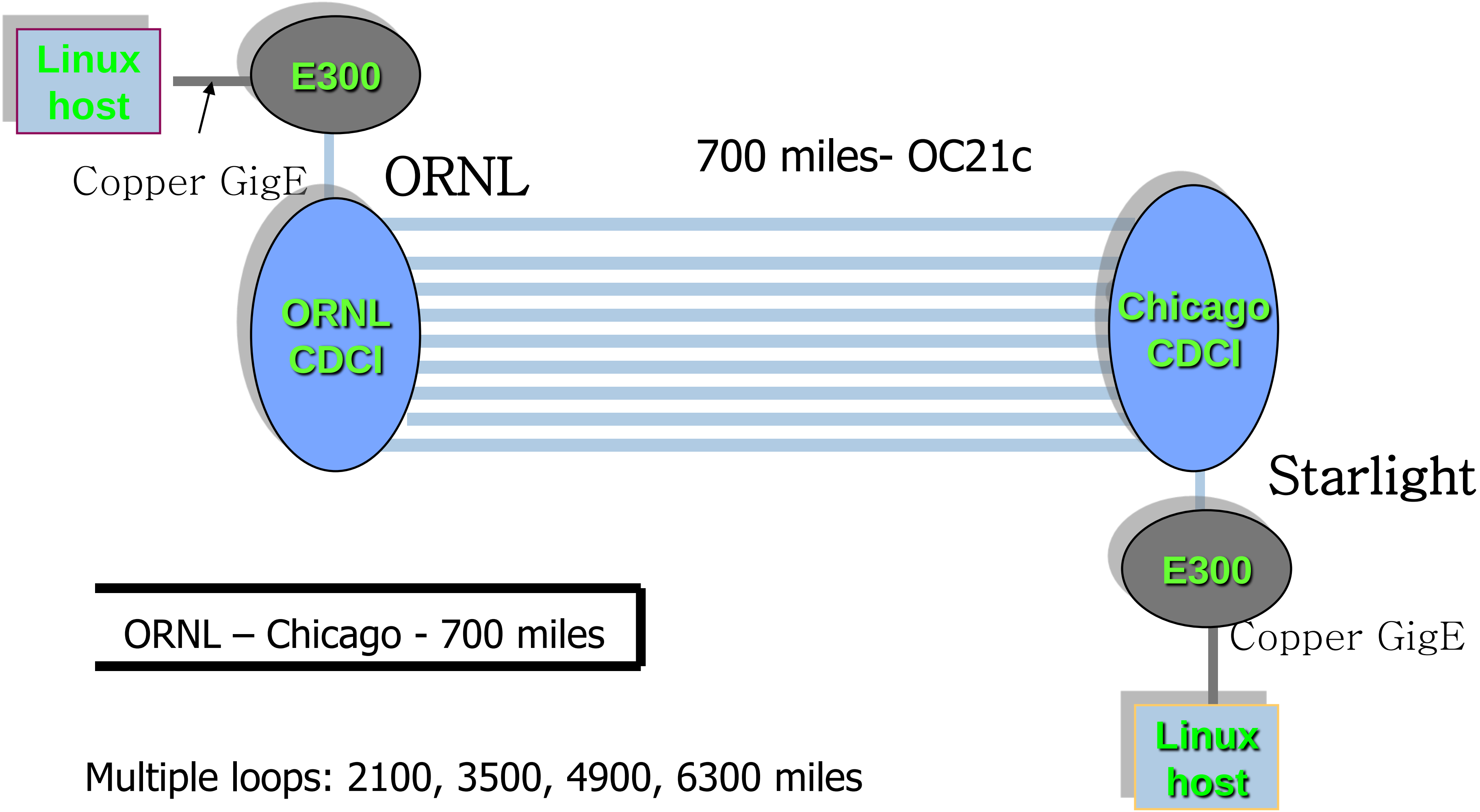
Data-plane unification of dedicated paths over layer-1, layer-2 and layer-3 paths

Need systematic analysis of application and IP level measurements:

Using USN, CHEETAH and Esnet, we

- collected ping, iperf andTCP measurements
- performed comparative performance analysis
- composed and tested VLANS over SONET and IP connections

1GigE Over SONET: USN test configurations



ORNL – Chicago - 700 miles

Multiple loops: 2100, 3500, 4900, 6300 miles

Channel Throughput profile

Plot of receiving rate as a function of sending rate

Its precise interpretation depends on:

- Sending and receiving mechanisms
- Definition of rates

For protocol optimizations, it is important to use its own sending mechanism to generate the profile

Window-based sending process for UDP datagrams:

Send $W_c(t)$ datagrams in a one step – *window size*

Wait for $T_S(t)$ time called *idle-time* or *wait-time*

Sending rate at time resolution $T_S(t)$:

$$r_s(t) = \frac{W_c(t)}{T_s(t) + T_c(t)}$$

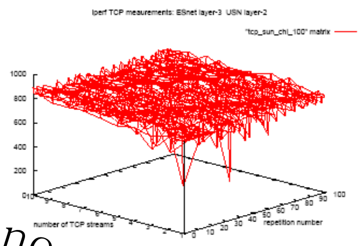
Layer 3 and Layer 1 Connections: iperf TCP Throughput Measurements

No. streams 1-10 repeated 100 times

ESnet

Chicago-Sunnyvale

Layer-3:
MPLS tunnel
Ping: 67ms
~3600 miles



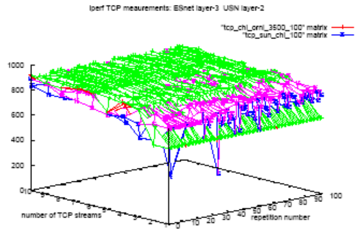
no. streams

repetitions

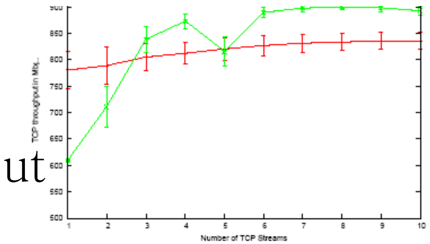
Comparison

On layer-2 connection higher throughput is achieved with more streams

USN: 906 Mbps
ESnet: 852 Mbps



throughput

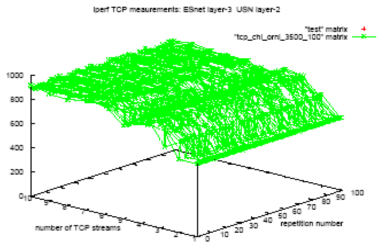


no. of streams

USN

ORNL-Chicago-..- ORNL-Chicago

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



TCP peak rates: 7–8 streams

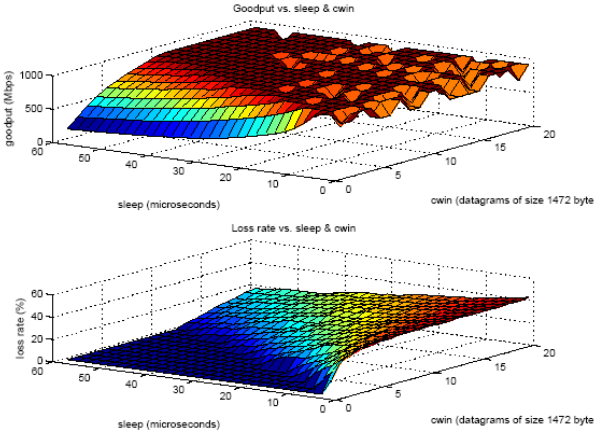
SONET: 906Mbps
MPLS: 852 Mbps
Hybrid: 852 Mbps

Connection Profile: Window-based UDP transport

Collaboration with Qishi Wu, University of Memphis

ESnet

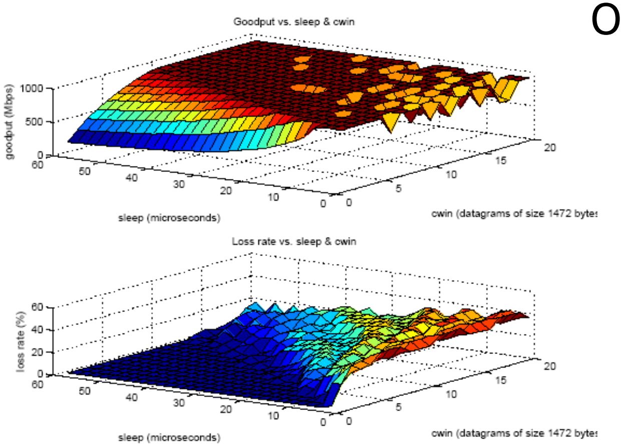
Chicago-Sunnyvale



Layer-3:
MPLS tunnel
Ping: 67.5ms
~3600 miles

ESnet-USN

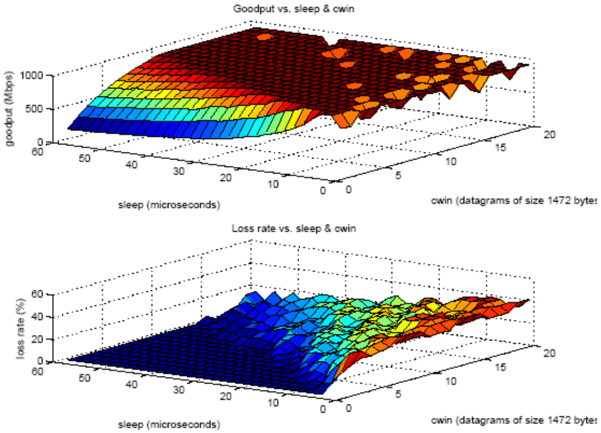
ORNL-Chicago-Sunnyvale



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

USN

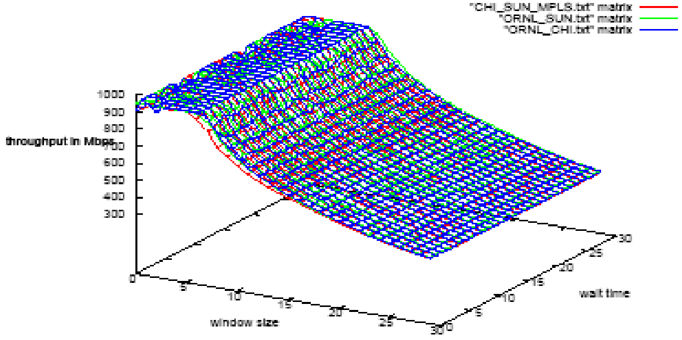
ORNL-Chicago-..- ORNL-Chicago



Layer 2 over OC21c
Ethernet over SONET
Ping: 134ms
~7100 miles

Throughput comparisons: Summary

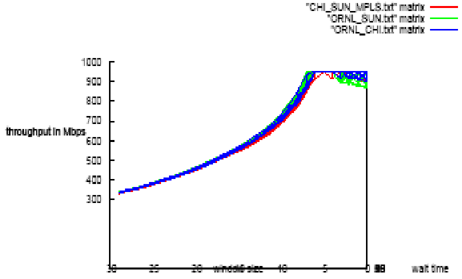
	PLUT	UDP peak	TCP peak	PLUT−TCP diff
MPLS:	952 Mbps	953	840	112
SONET:	955 Mbps	957	900	55
Hybrid:	952 Mbps	953	840	112
Difference	3Mbps	5Mbps	60Mbps	



USN
ORNL-Chicago-..- ORNL-Chicago

ESnet
Chicago-Sunnyvale

ESnet-USN
ORNL-Chicago-Sunnyvale



Special purpose UDP-PLUT transport achieved higher throughput than multi-stream TCP

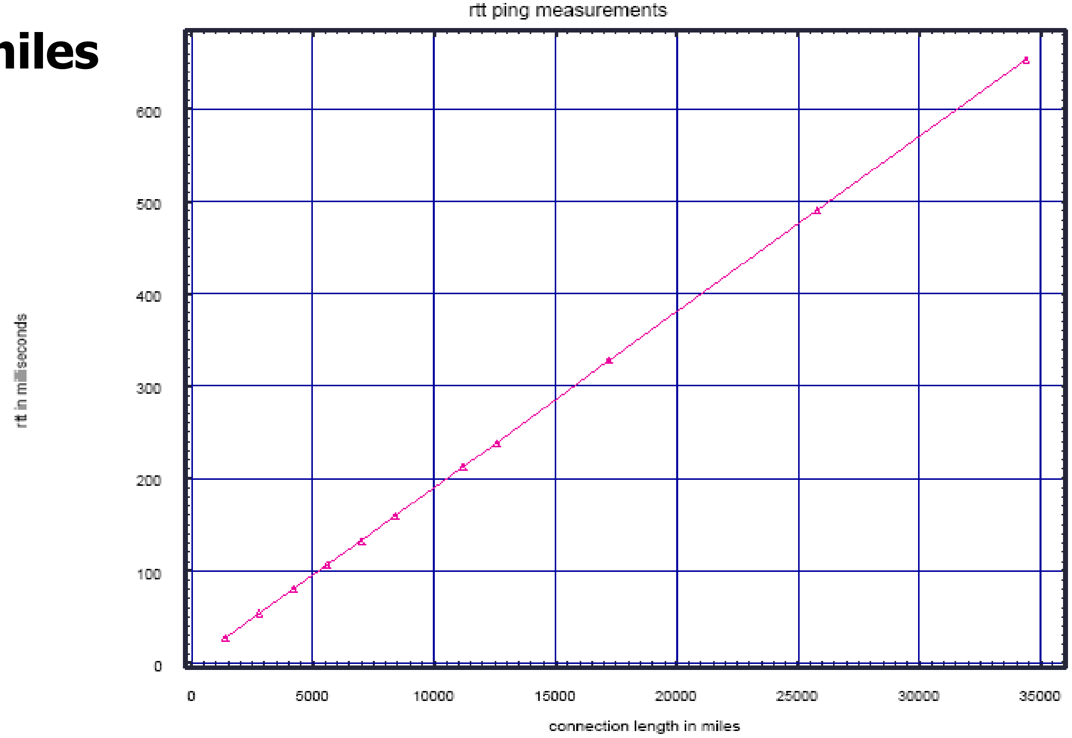
Outline

- **Motivation and Background**
- **USN infrastructure**
  - **Architecture**
  - **Data-plane**
  - **Control-plane**
  - **Connection Suites**
- **USN Networking Experiments**
  - **Hybrid Network Connections - jitter**
  - **Infiniband over Wide-Area**
  - **Connections to Supercomputers**
  - **Transport Methods for Dedicated Channels**
  - **Wide-Area Application Accelerators**
  - **Encryption Devices**

USN test configurations: Ping RTT

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

miles	rtt(ms)
8,600	163
17,200	327
25,800	490
34,400	653

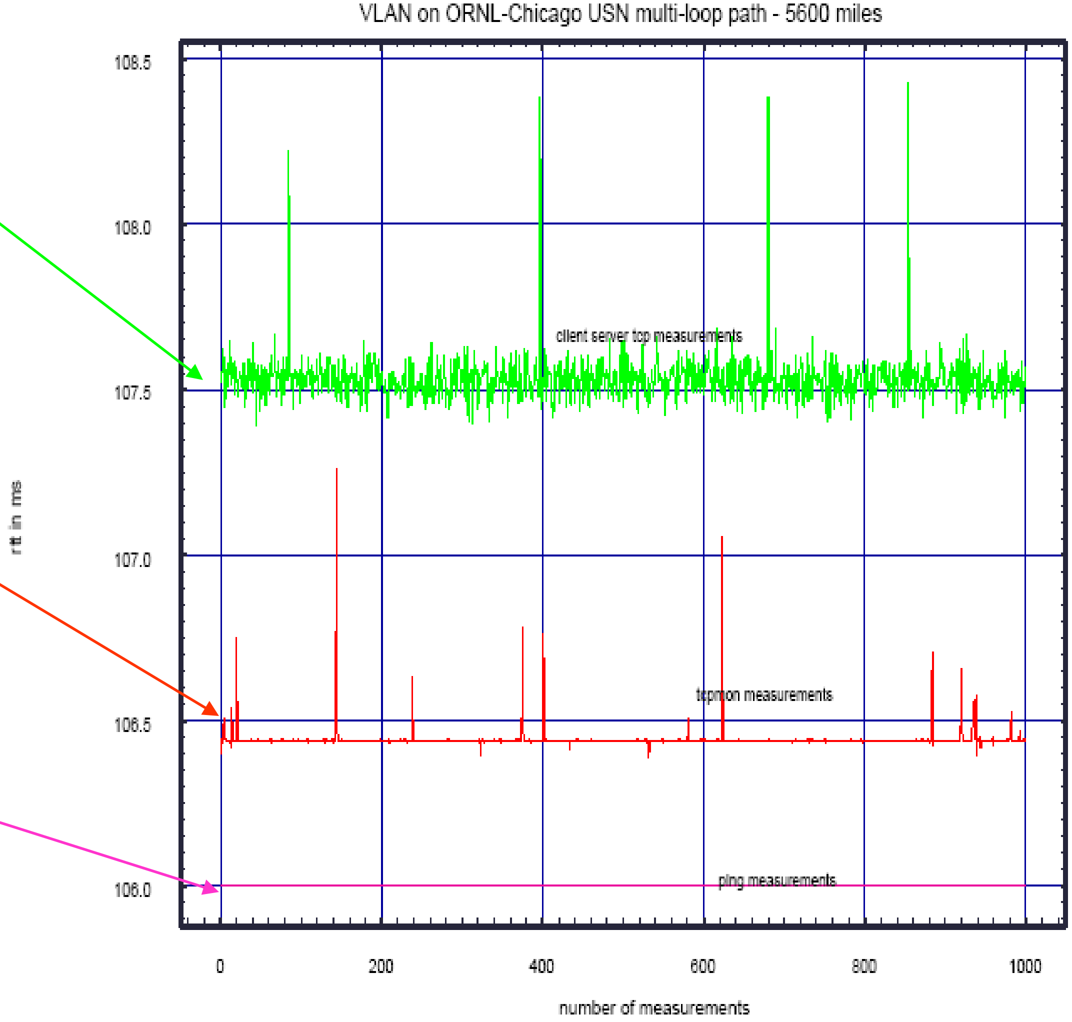


ORNL – Chicago - loop – 1400 miles

miles	1,400	2,800	4,200	5,600	7,000	8,400	9,800	11,200	12,600
rtt (ms)	26.79	53.4	79.90	106	132	159	185	212	238

Jitter Measurements Suite

1. **TCP client-server: client sends a message and server echo 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

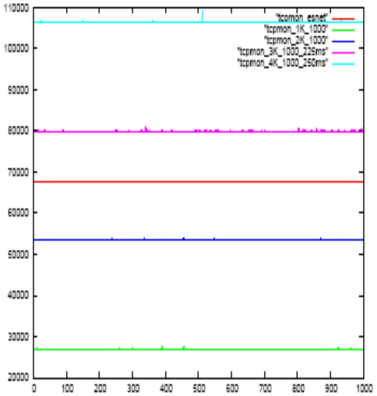
TCP Client-Server Measurements MPLS tunnel and Ethernet over SONET

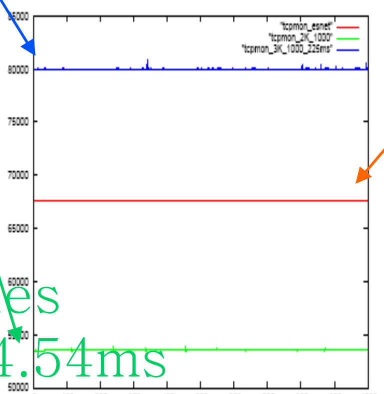
MPLS tunnel measurements seem comparable

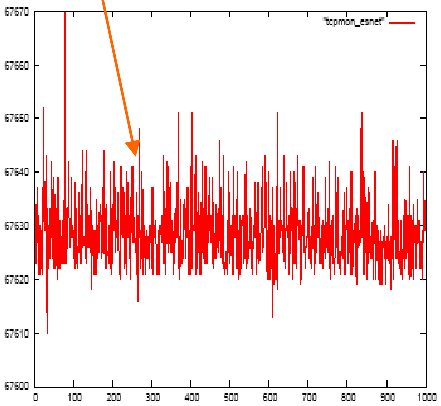
USN
ORNL-Chicago-..- ORNL-Chicago

4200 miles
Mean: 81.03ms
Range: 0.29%
Std dev: 0.05%

ESnet
MPLS tunnel
Chicago-Sunnyvale
Mean: 68.71ms
Range: 0.29%
Std dev: 0.07%







2800 miles
mean: 54.54ms
Range: 0.43%
Std dev: 0.097%

More detailed analysis
is needed to quantify
the relative performance

Objective Comparison of Measurements

Basic Problem

Measurements are collected for two types of connections at different connection lengths d_1 and d_2

Question: how do we objectively compare them?

Considerations:

Ideally, we may replace all the devices on one type of connection with the other and repeat the measurements – this is not a feasible solution

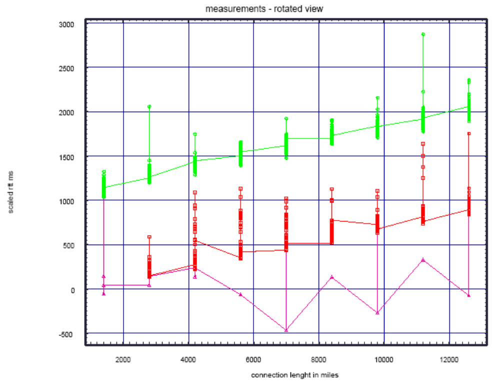
Computing mean and variances at non-commensurate lengths is not very instructive

Particular version of regression

- Small number of connection lengths
- Several measurements at each length

Characteristically different from the usual

scatter-plot regression



Normalization Framework

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

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

$M_T(d)$	Measurements on path of type *T* of distance *d*
$\hat{M}_T(d)$	Estimates of measurements on path of type T of distance d
$P \odot \hat{M}_T(d)$	Parameters computed using measurements

$$\begin{tikzcd}
M_{T_1}(d_1) \arrow[d, "\Theta_{T_1}"] & & M_{T_2}(d_2) \arrow[d, "\Theta_{T_2}"'] \\
\hat{M}_{T_1}(d) \arrow[d] & & \hat{M}_{T_2}(d) \arrow[d] \\
P \odot \hat{M}_{T_1}(d) & \aleph_P \arrow[l] \arrow[r] & P \odot \hat{M}_{T_2}(d)
\end{tikzcd}$$

Interpolation based on regression

Regression Method

Basic Problem

Parameters are measured or estimated for a particular connection-type at different connection lengths $d_1, d_2, \cdots, d_n$

Question: Estimate the parameters at distance d

Two solutions: Measurements at distance $M_1(d_i), M_2(d_i), \cdots, M_{n_i}(d_i)$

Linear regression: L_{-1} computes

$$\min\left[\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(L(d_i) - M_j(d_i)\right)^2\right]$$

over all lines – it does no achieve 0 MSE and too-sensitive to point variations

Fully-segmented regression L_n is linear interpolation of points

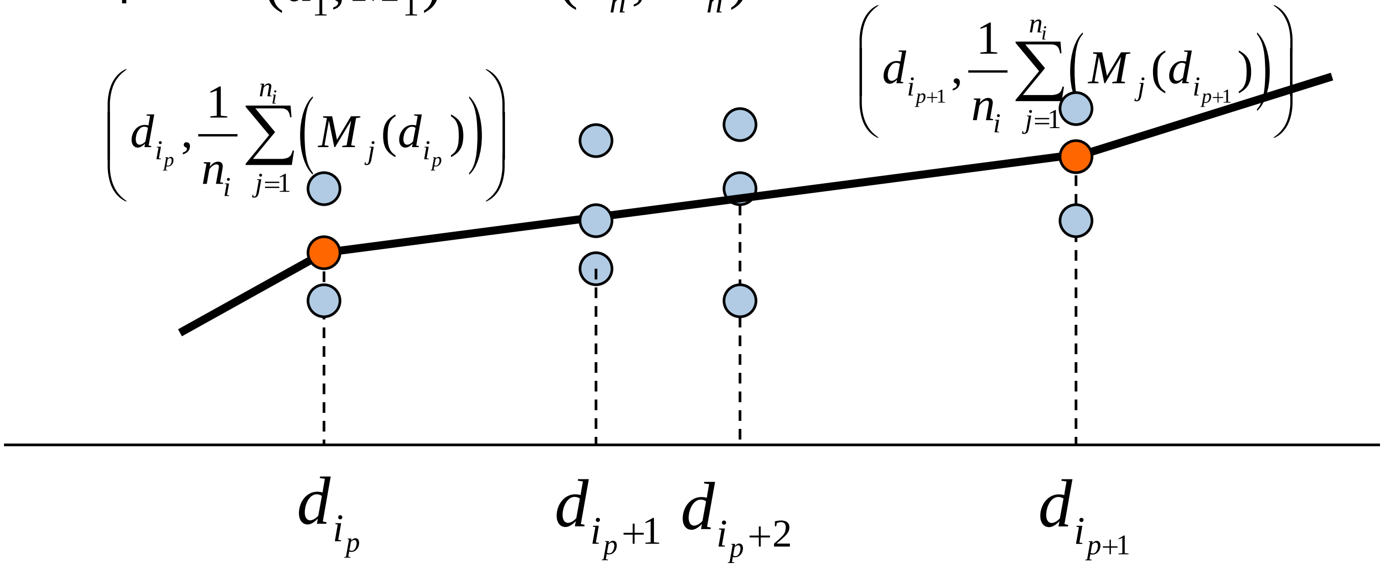
$$(d_i, \bar{M}_i) = \left(d_i, \frac{1}{n_i}\sum_{j=1}^{n_i}\left(M_j(d_i)\right)\right)$$

It achieves 0 MSE but has lower predictive quality – higher Vapnik and Chervonenkis dimension of 2(n-1)

Segmented Regression Method

K-Segmented Regression: L_k Utilizes k distances $d_{i_1}, d_{i_2}, \cdots, d_{i_k}$ as anchors, and uses linear interpolation between them $k = 0, 1, \cdots, n-2$

with end points $(d_1, \bar{M}_1)$ and $(d_n, \bar{M}_n)$



$$\left(d_{i_p}, \frac{1}{n_i}\sum_{j=1}^{n_i}\left(M_j(d_{i_p})\right)\right)$$

$$\left(d_{i_{p+1}}, \frac{1}{n_i}\sum_{j=1}^{n_i}\left(M_j(d_{i_{p+1}})\right)\right)$$

d_{i_p} d_{i_p+1} d_{i_p+2} $d_{i_{p+1}}$

Optimal L_k can be computed using dynamic programming for fixed

Optimal ***k*** is computed using Vapnik-Chervonenkis bound equations

Best in Class Estimator

<u>Prediction Error:</u> $f : \Re \rightarrow \Re$ corresponding to unknown distribution $P_{M,d}$

Error corresponding to measure measurement (M, d)

$$E(f) = \int_{M,d} \left(f(d) - M\right)^2 P_{M,d} \qquad E(f^*) = \min_{f \in \mathbb{F}} E(f)$$

Empirical Error

$$\hat{E}(f) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left(f(d_i) - M_j(d_i)\right)^2 \qquad \hat{E}(\hat{f}) = \min_{f \in \mathbb{F}} \hat{E}(f)$$

Vapnik and Chervenenkis Theory: For function class $\mathbb{F}$

$$E(\hat{f}) \leq \hat{E}(\hat{f}) + \frac{B \in (l)}{2}\left(1 + \sqrt{1 + \frac{\hat{E}(\hat{f})}{B \in (l)}}\right)$$

$$\in (l) = 4\left(\frac{1}{l}\left(h\left(\ln(2l/h) + 1\right) - \ln(\eta/4)\right)\right)$$

;

$h = VC\dim(\mathbb{F})$ $\quad \left(f(d) - M\right)^2 \leq B$ and $l = \sum_{i=1}^{n} n_i$

Best Segmented Regression Estimator

VC-Dimension estimates: L_k

Linear regression class: $VC\dim(\mathbf{L}_{-1}) = 2$

Segmented regression class of $VC\dim(\mathbf{L}_k) = 2(k+1)$

$k = 0,1,\cdots,n-1$

For delay estimates, regresssion could be monotonic: VCdim=2

Choose estimator to minimize the prediction error bound:

for $k = -1,0,1\cdots,n-1$

$$E(L_k) \le \hat{E}(L_k) + \frac{B\in(l)}{2}\left(1+\sqrt{1+\frac{\hat{E}(L_k)}{B\in(l)}}\right)$$

$$\in(l) = 4\left(\frac{1}{l}\left(VC\dim(\mathbf{L}_k)\left[\ln(2l/VC\dim(\mathbf{L}_k))+1\right]-\ln(\eta]4)\right)\right)$$

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**

$$\begin{tikzcd}
M_{SONET}(1400) \arrow[d, "\Theta_{T_1}"] & & M_{MPLS}(300) \arrow[d, "\Theta_{T_2}"'] \\
\hat{M}_{SONET}(300) \arrow[d] & & \hat{M}_{MPLS}(300) \arrow[d] \\
P \odot \hat{M}_{SONET}(300) & & P \odot \hat{M}_{MPLS}(300) \\
& \aleph_P \arrow[ul] \arrow[ur] &
\end{tikzcd}$$

Interpolation based on linear regression

identity

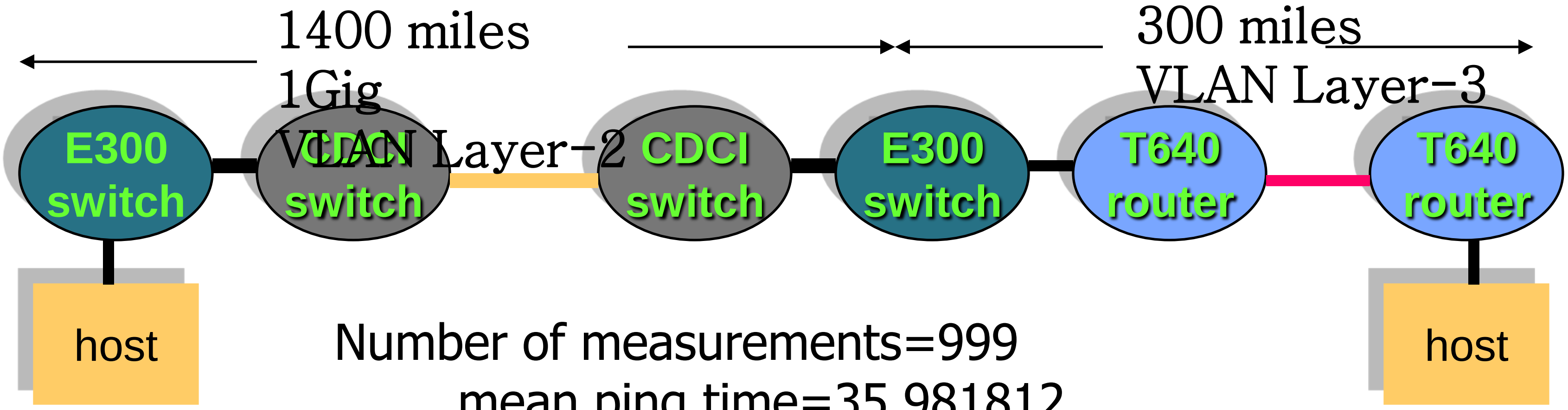
Align jitter regression band

Another Method

P – FFT

$\aleph_P$ – Identity

Composed VLAN:
SONET and Layer-3 Channels - Gig 1300 miles



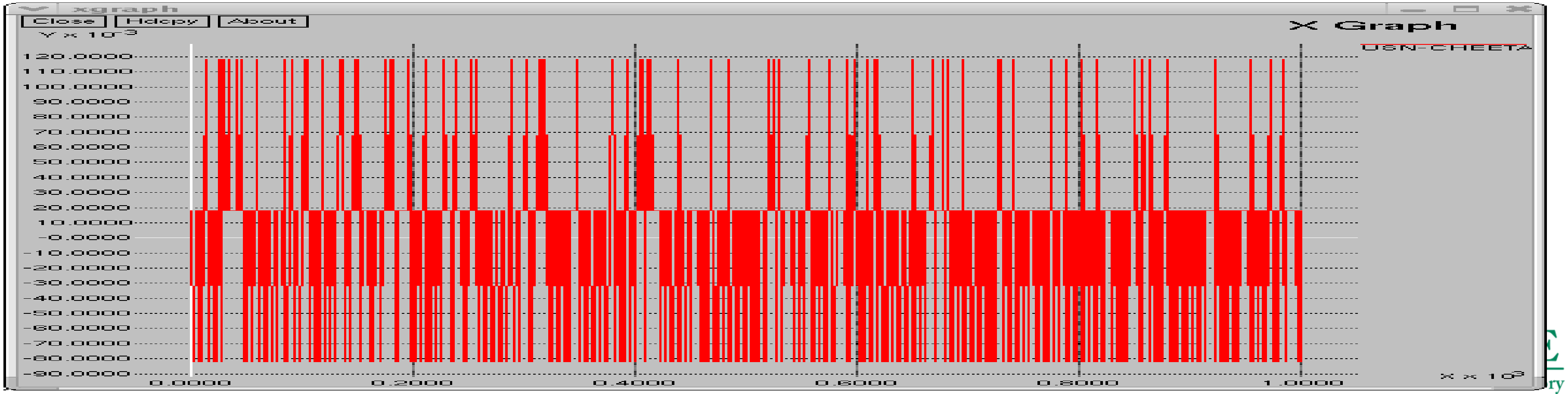
Number of measurements=999

mean ping time=35.981812

percent range: [99.772635,100.328463]

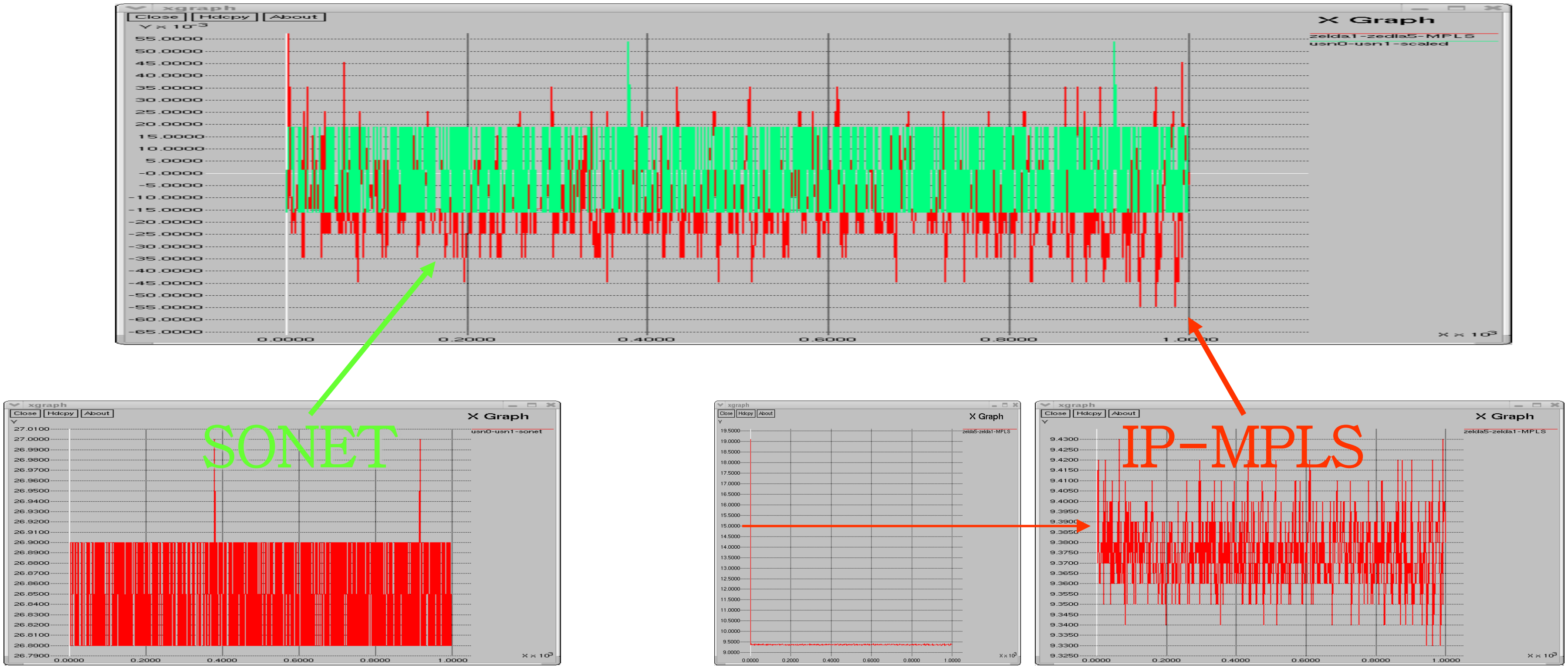
range: [35.900002,36.099998]: 0.199997

std_deviation (percent)= 0.151493



Comparison of VLANs: SONET vs. MPLS tunnels

Measurements are normalized for comparison:



mean time=26.845877ms
percent range: [99.8,100.6]
std_dev (%)= 0.187035

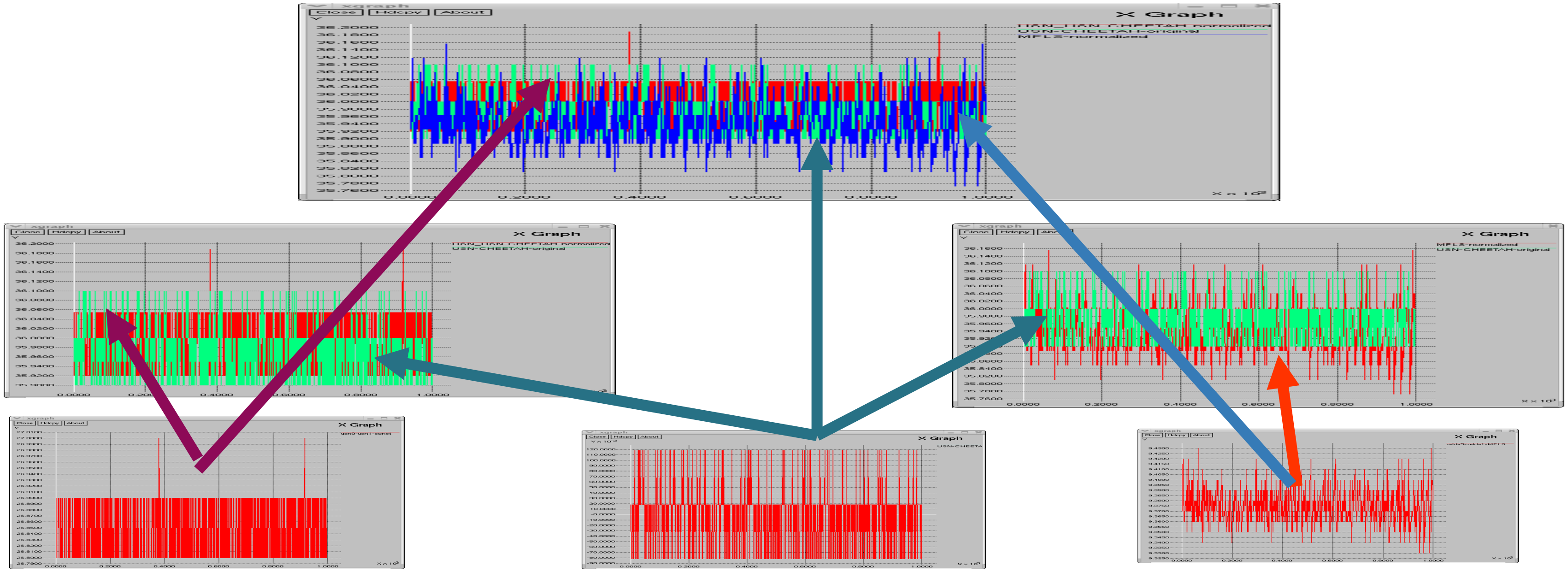
mean time=9.384557m
percent range:[99.4,20
std_dev (%)= 3.281692

Conclusion

VLANs over SONET have smaller jitter levels

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

Measurements are normalized for comparison:



SONET	**SONET-MPLS composite**	**L2MPLS**
mean time = 26.845877 ms std_dev (%) = 0.187035	mean time = 35.981812 ms std_dev (%) = 0.151493	mean time = 9.384557 ms std_dev (%) = 3.281692

SONET channels have smaller jitter levels

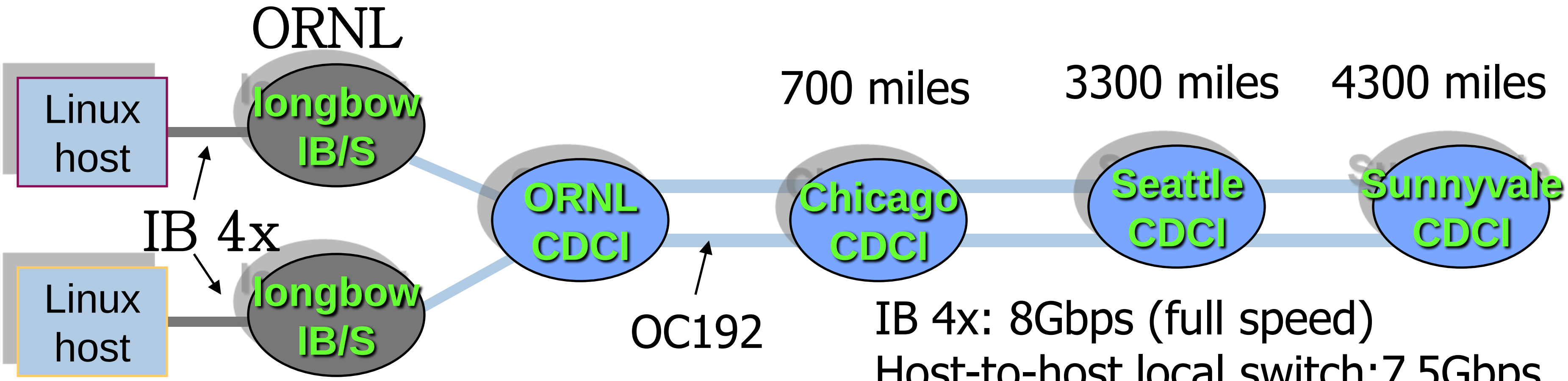
Outline

- **Motivation and Background**
- **USN infrastructure**
  - Architecture
  - Data-plane
  - Control-plane
  - Connection Suites
- **USN Networking Experiments**
  - Hybrid Network Connections
  - Infiniband over Wide-Area
  - Connections to Supercomputers
  - Transport Methods for Dedicated Channels
  - Wide-Area Application Accelerators
  - Encryption Devices

But, Supercomputers do much faster local transfers …

- Infiniband at 4X routines achieves 7.6Gbps
  - Is it very effective data transport protocol for storage networks (few miles)?
  - Question: Can we natively support IB over wide-area?
- Related Comments:
  - Additional Benefit: data and file systems can be “transparently” access – remote mount a file system
  - TCP is not easily extended and not optimal for such data transfers

Infiniband Over SONET: Obsidian Longbows RDMA throughput measurements over USN



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

Hosts:
dual-socket quad-core 2GHz AMD Opteron, 4GB memory
8-lane PCI-Express slot
Dual-port Voltaire 4x SDR HCA.

ORNL loop -0.2 mile: **7.48**Gbps

ORNL-Chicago loop – 1400 miles: **7.47**Gbps

ORNL- Chicago - Seattle loop – 6600 miles: **7.37**Gbps

ORNL – Chicago – Seattle - Sunnyvale loop – 8600 miles: **7.34**Gbps

Performance Profiles – IB RDMA Throughputs

- Throughput Distance Profile
  - Plot throughput as a function connection length and message size
  - B=SONET, WAN-PHY

$$T_B(d,s)$$

- Throughput Stability Profile
  - Plot throughput as function of connection length and repetition number for fixed message size

$$T_B(d,s) \quad - - - \quad T_B(d,s)$$

  - Average throughput over 10 iterations with 8M message size

$$\bar{T}_B(d)$$

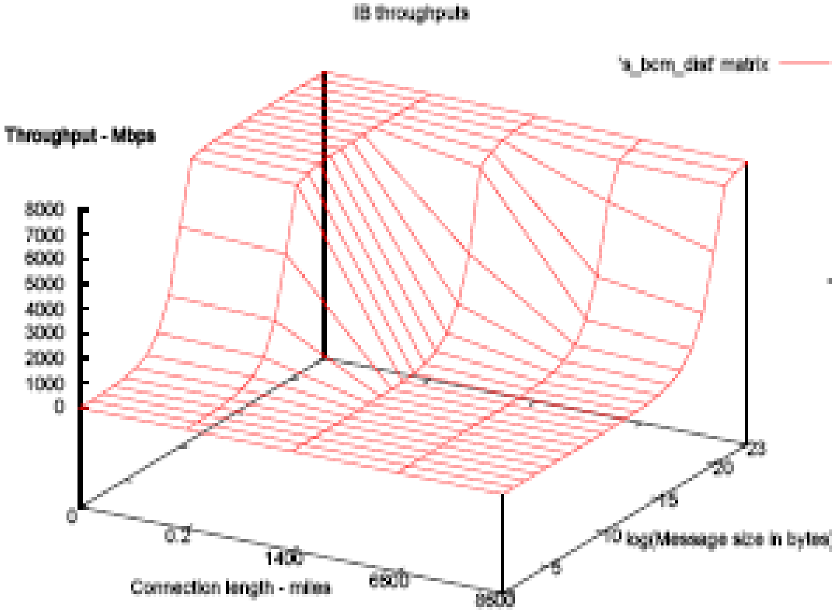
- Throughput Decrease Per Mile

$$D_B(d_i) = \frac{\bar{T}_B(d_0) - \bar{T}_B(d_i)}{d_i - d_0}$$

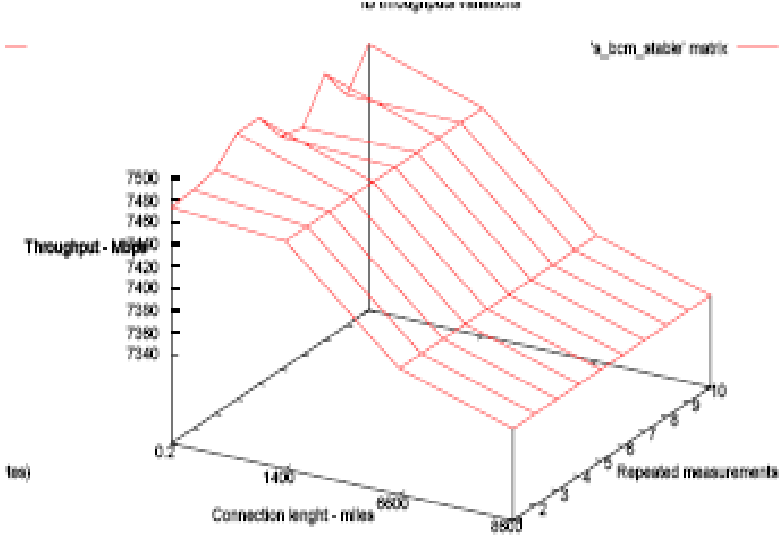
Distance and Stability Profiles of IB over SONET

Measurements using ib_rdma-bw – c
It uses IB CM for connection setup and management

distance profile

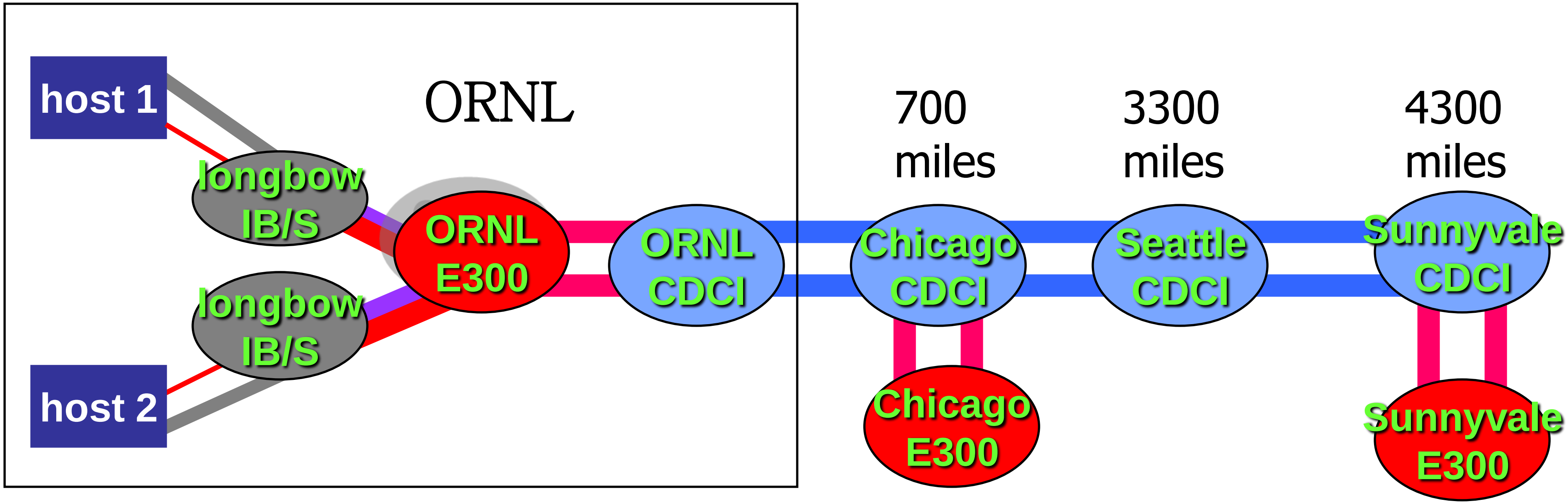


stability profile
8M message size



Connection length (miles)	0.2	1400	6600	8600
Throughput (Gbps) – 8M msg	7.48	7.47	7.37	7.34
Std-dev (Mbps)	45.27	0.07	0.09	0.07
DPM (Mbps)	0	0.012	0.017	0.016

IB over 10GigE LAN-PHY and WAN-PHY



IB 4x ▬
OC192 ▬
10 GigE LAN-PHY ▬
10 GigE LAN-PHY ▬
1GigE —

ORNL loop -0.2 mile

ORNL-Chicago loop – 1400 miles

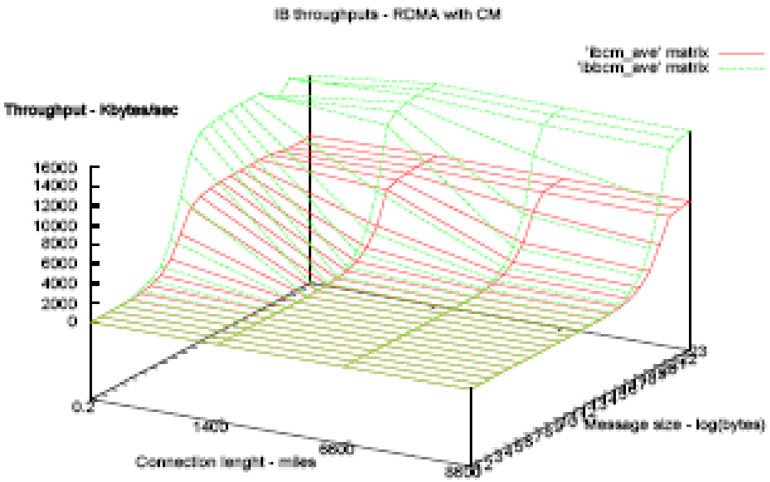
ORNL- Chicago - Seattle loop – 6600 miles

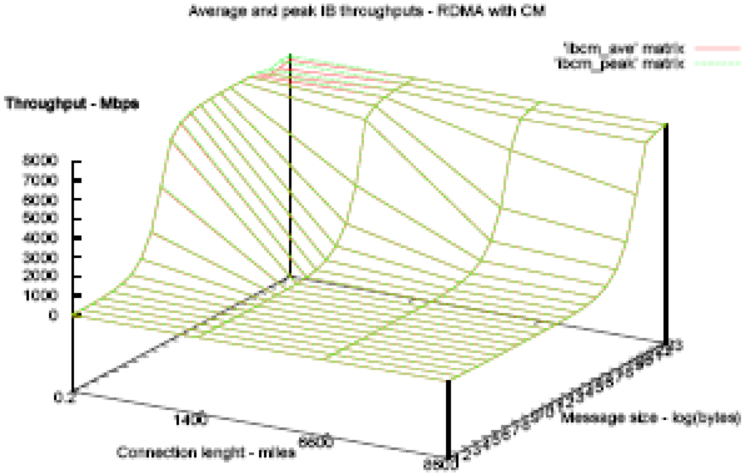
ORNL – Chicago – Seattle - Sunnyvale loop – 8600 miles

Performance Profiles of IB Over 10GigE WAN-PHY

distance profile

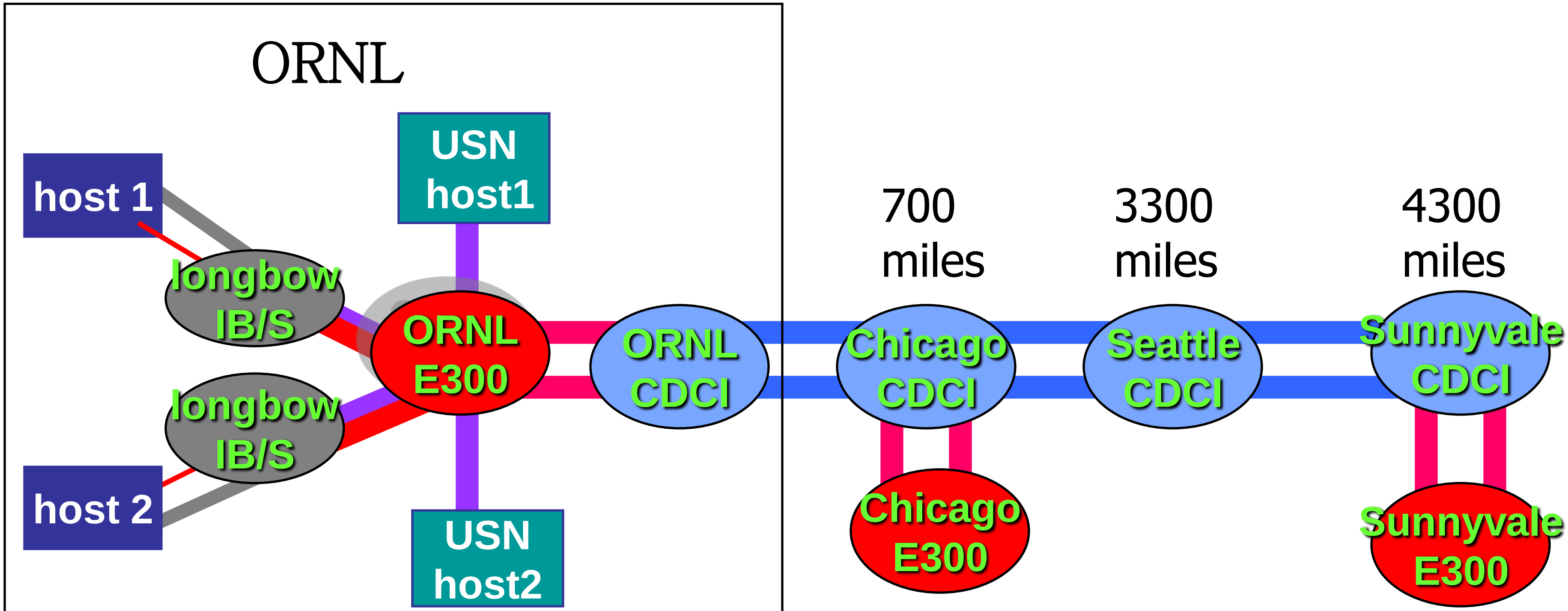
peak distance profile
average distance profile





Connection length (miles)	0.2	1400	6600	8600
Throughput (Gbps) – 8M msg	7.5	7.49	7.39	7.36
Std-dev (Mbps)	0.07	0.69	0.00	0.20
DPM (Mbps)	0	0.012	0.017	0.016

Cross-Traffic Generation



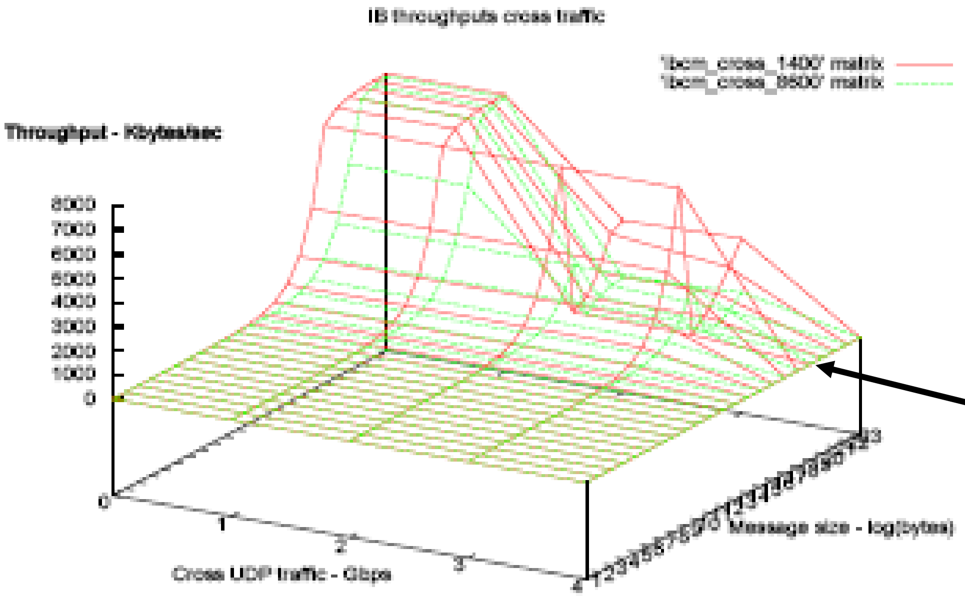
ORNL loop -0.2 mile

ORNL-Chicago loop – 1400 miles

ORNL- Chicago - Seattle loop – 6600 miles

ORNL – Chicago – Seattle - Sunnyvale loop – 8600 miles

Cross-Traffic Effect of IB over 10GigE WANPHY

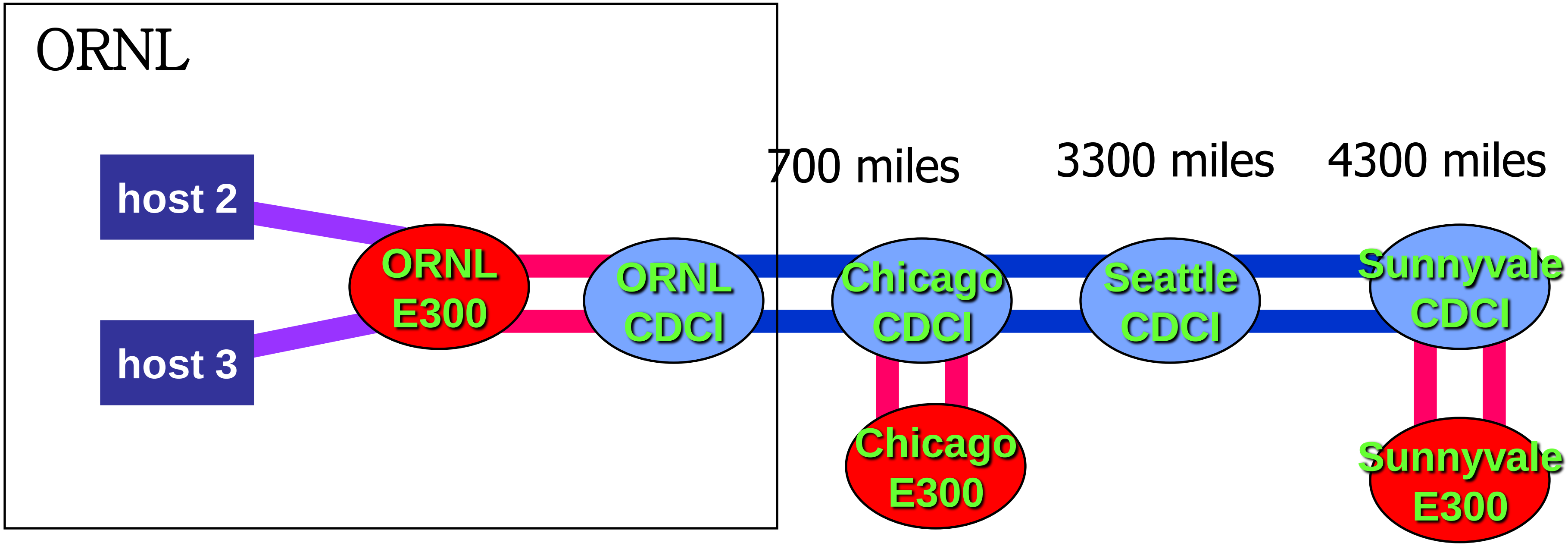


Below 1Gbps

Competing traffic: UDP streams on WAN at 1,2,3,4 Gbps

- Distance profiles are unaffected for cross-traffic levels of up to 1Gbps
- IB throughput was drastically effected at cross-traffic level of 4 Gbps
- Effect of cross-traffic is more on large message sizes

10GigE Connections



ORNL loop -0.2 mile

ORNL-Chicago loop – 1400 miles

ORNL- Chicago - Seattle loop – 6600 miles

ORNL – Chicago – Seattle - Sunnyvale loop – 8600 miles

10 GigE WAN−PHY 🟥
10 GigE LAN−PHY 🟪
OC192 🟦

Performance Profiles – TCP Throughputs

BIC and Hamilton TCP – pluggable Linux modules

- Throughput Distance Profile
  - Plot throughput as a function connection length and number of streams
  - A=BIC,HTCP

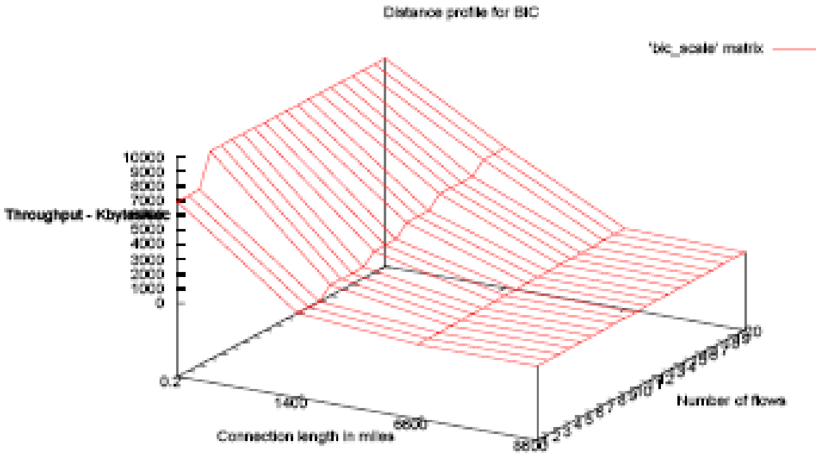
$$T_A(d,n)$$

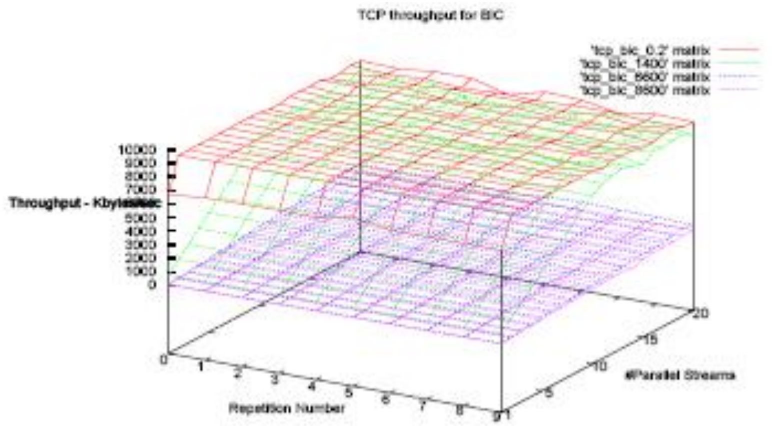
- Throughput Stability Profile
  - Plot throughput as function of connection length and repetition number of streams
  - Average throughput over repetitions and range of number of streams 15-20 $\bar{T}_B(d)$

- Throughput Decrease Per Mile

$$D_A(d_i)=\frac{\bar{T}_A(d_0)-\bar{T}_A(d_i)}{d_i-d_0}$$

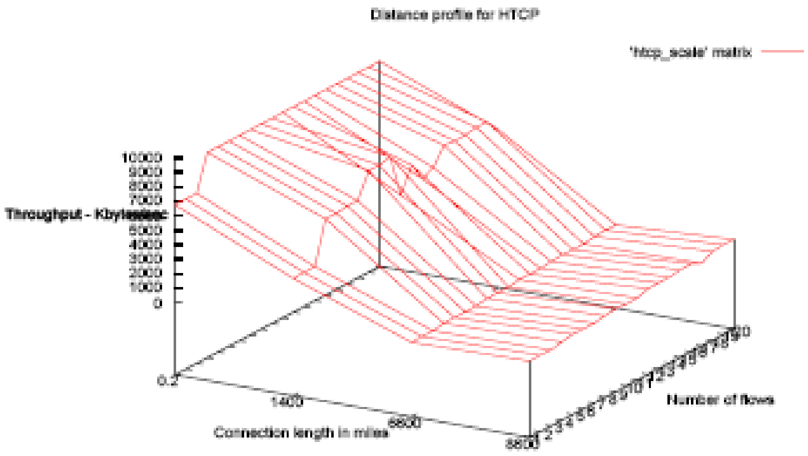
Performance of TCP over 10GigE
BIC with Linux auto-tuning

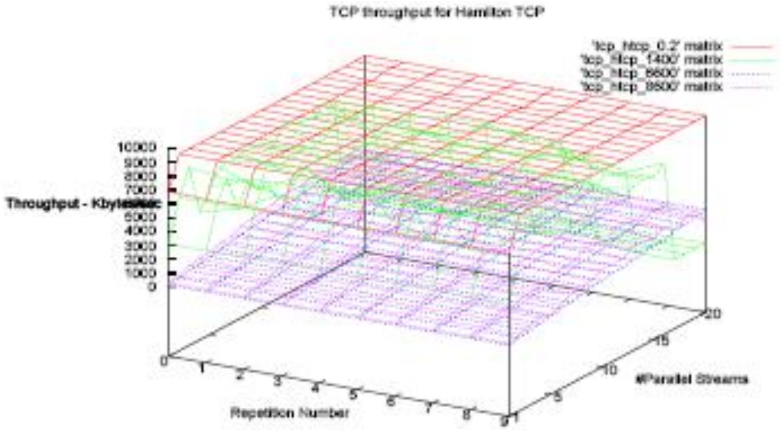




Connection length (miles)	0.2	1400	6600	8600
Throughput (Gbps) – 8M msg	9.12	6.69	0.76	0.50
Std-dev (Mbps)	64.11	70.08	24.96	21.08
DPM (Mbps)	0	1.74	1.27	1.00

Performance of TCP over 10GigE
Hamilton TCP with Linux auto-tuning



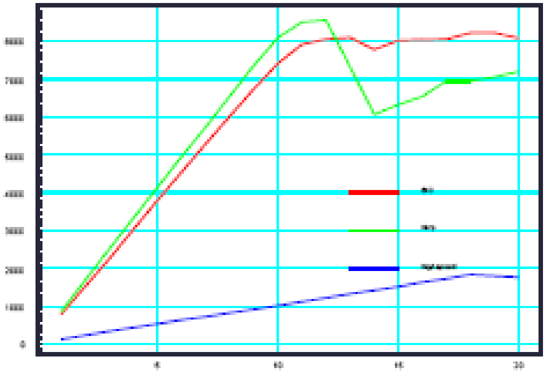


Connection length (miles)	0.2	1400	6600	8600
Throughput (Gbps) – 8M msg	9.21	6.71	1.22	1.79
Std-dev (Mbps)	12.25	37.42	18.96	128.15
DPM (Mbps)	0	1.79	1.21	0.87

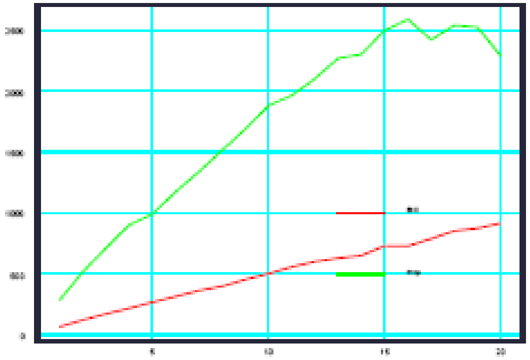
Comparative Performance of BIC and Hamilton TCP

BIC

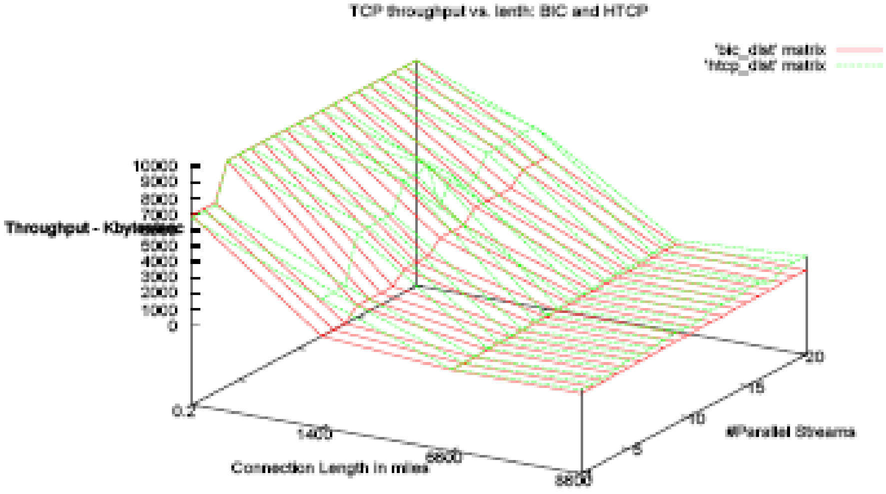
1400 miles



8600 miles



HTCP



Outline

- **Motivation and Background**
- **USN infrastructure**
  - Architecture
  - Data-plane
  - Control-plane
  - Connection Suites
- **USN Networking Experiments**
  - Hybrid Network Connections
  - Infiniband over Wide-Area
  - Connections to Supercomputers
  - Transport Methods for Dedicated Channels
  - Wide-Area Application Accelerators
  - Encryption Devices

Connecting Supercomputers: Complex Problem Space

- Requires knowledge in networking and supercomputer architectures – no single answer
- Just adding 10GigE NICs is not sufficient
- Internal data paths must be carefully configured
  - Cray X1 SPC-FC-Ethernet
- Execution paths are just as important
  - Network stack is implemented as thread migration to OS nodes
- Cross-Connects must match the impedances
- High-Performance wide-area storage and file systems need further development

Experimental Results:
Production 1GigE Connection Cray X1 to NCSU

- Tuned/ported existing bbcp protocol (unicos OS):
  - optimized to achieve 250-400Mbps from Cray X1 to NCSU;
    - actual throughput varies as a function of Internet traffic
    - tuned TCP achieves ~50 Mbps.

  currently used in production mode by John Blondin
- developed new protocol called Hurricane
  - achieves *stable* 400Mbps using a single stream from Cray X1 to NCSU;

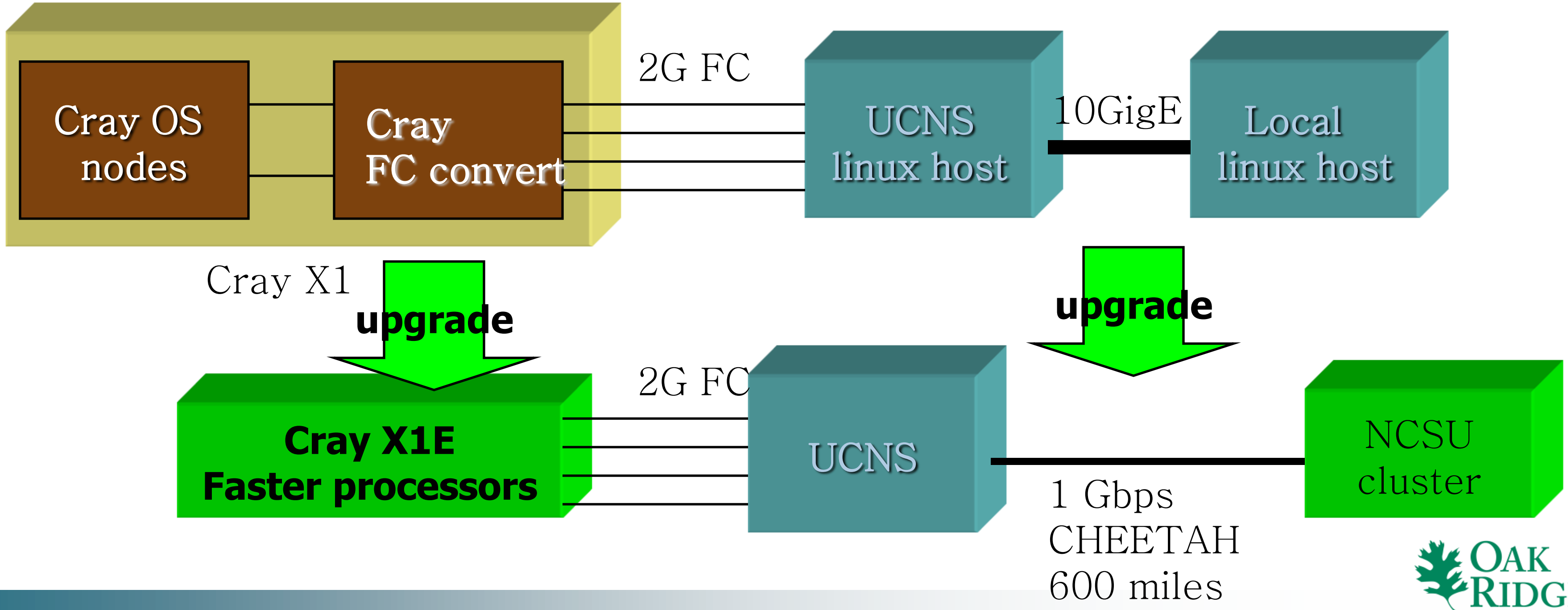
These throughput levels are the highest achieved (2005) between ORNL Cray X1 and a remote site located several hundred miles away.

Experimental Results Cray X1: Dedicated Connection

Dedicated Channel

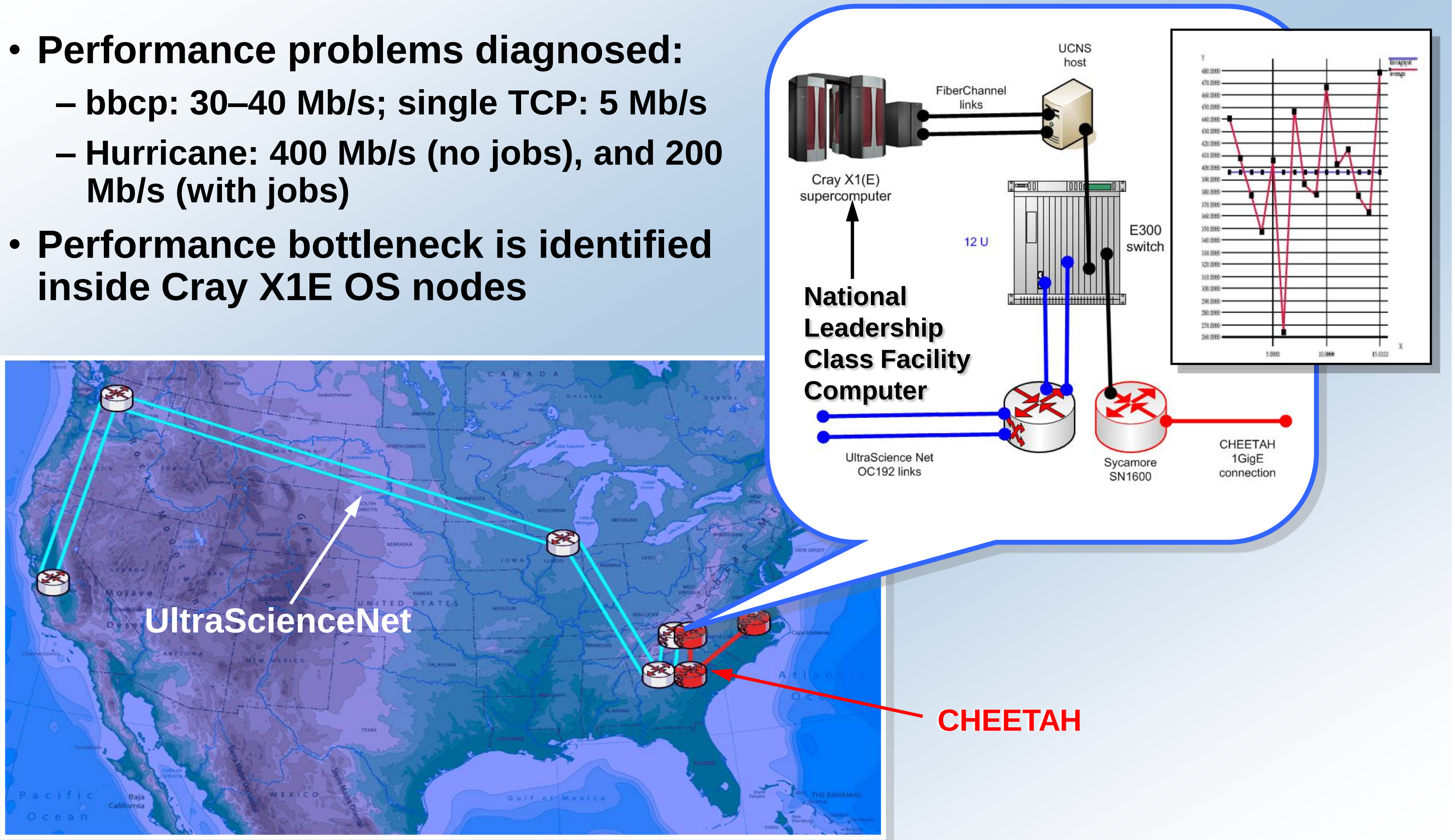
- UCNS connected to Cray X1 via four 2Gbps FC connections.
- UCNS is connected to another linux host via 10 GigE connection
- Transfer results:
  - 1.4Gbps using single flow using Hurricane protocol

highest file transfer rates achieved over Ethernet connections from ORNL Cray X1 to an external (albeit local) host



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



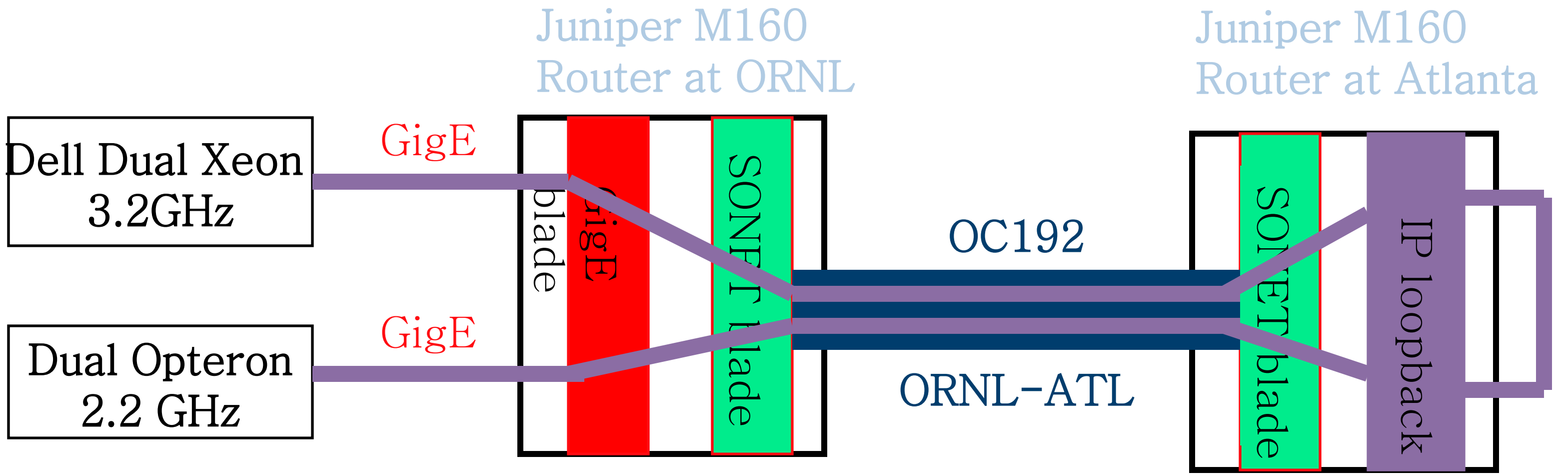
Outline

- **Motivation and Background**
- **USN infrastructure**
  - Architecture
  - Data-plane
  - Control-plane
  - Connection Suites
- **USN Networking Experiments**
  - Hybrid Network Connections
  - Infiniband over Wide-Area
  - Connections to Supercomputers
  - Transport Methods for Dedicated Channels
  - Wide-Area Application Accelerators
  - Encryption Devices

Transport Methods for Dedicated Channels

- Needed both research and development
  - TCP is sub-optimal:
    - Even multiple stream TCP can be analytically shown to under-utilize some bandwidth (6-12)
    - Congestion control takes processing time on hosts and absolutely not needed – does lower throughput
  - Hurricane Protocol
    - Optimized goodput and no congestion control
    - Needed detailed connection profile analysis
      - Typically achieved 99% of profile BW on 1Gbps 500 mile link
      - Light-weight flow control - NACK

1Gbps ORNL-ATL-ORNL Dedicated IP Channel



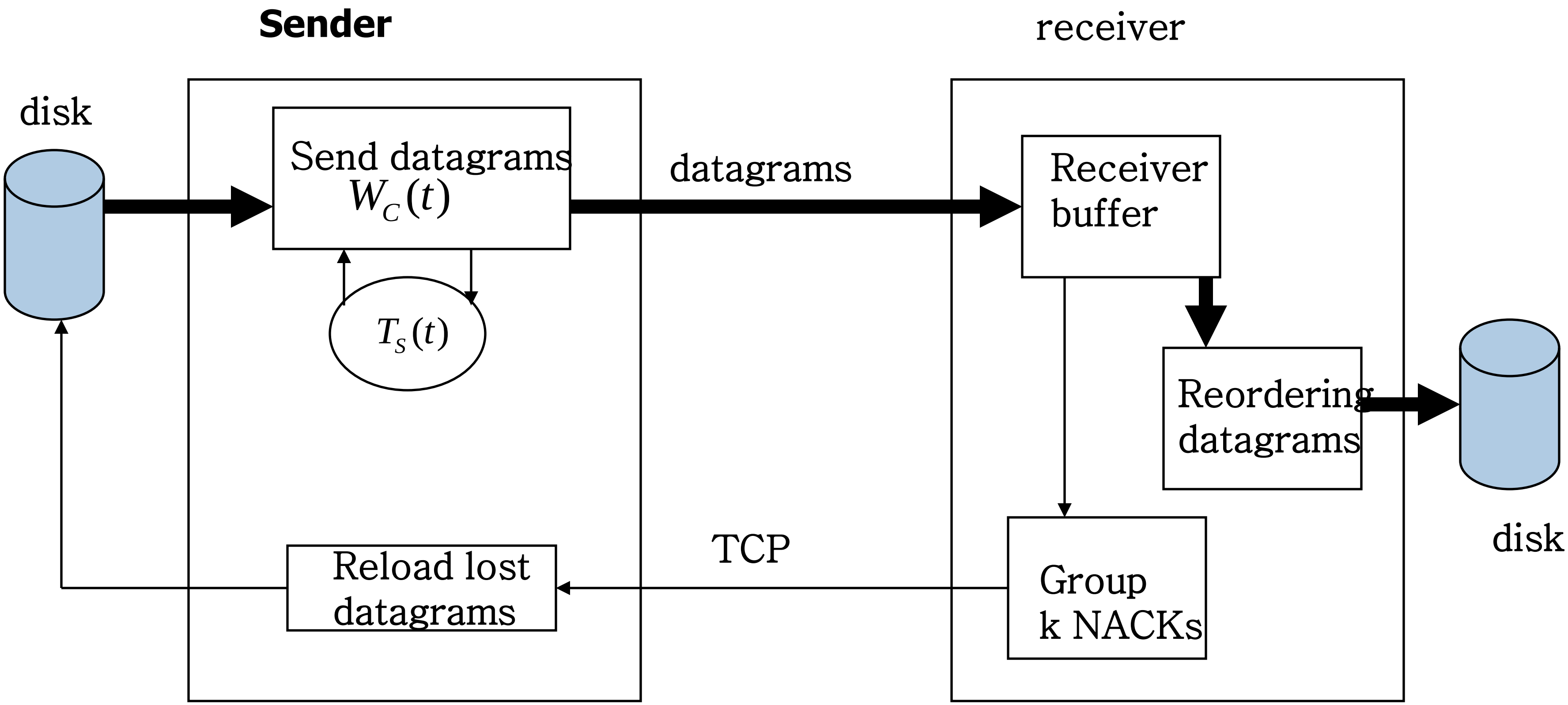
- **Non-Uniform Physical Channel:**
  - GigE – SONET – GigE
  - ~500 network miles
- **End-to-End IP Path**
  - Both GigE links are dedicated to the channel
  - Other host traffic is handled through second NIC
- **Routers, OC192 and hosts are lightly loaded**
- **IP-based Applications and Protocols are readily executed**

Hurricane Protocol

Collaboration with Qishi Wu, University of Memphis

- **Composed based on principles and experiences with UDT and SABUL**
  - **was not easy for us to figure out all tweaks for pushing peak performance**
- **UDP window-base flow-control**
  - **Nothing fundamentally new but needed for fine tuning**
  - **990 Mbps on dedicated 1Gbps connection disk-to-disk**
  - **No attempt for congestion control**

Hurricane Control Structure



Different subtasks are handled by threads, which are woken up on demand
Thread invocations are reduced by clustered NCKs instead of individual ACKS

Transport Modules Needed Careful Analysis

Disk-to-Disk Transfers (unet2 to unet1)

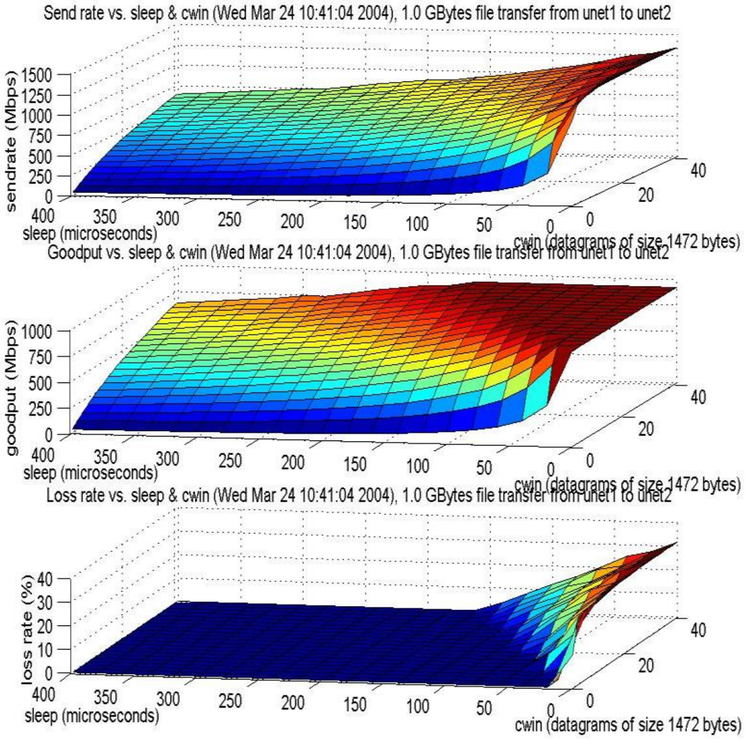
Protocol	**goodput**
tsunami	919 Mbps
UDT	890 Mbps
FOBS	708 Mbps
Hurricane	**990 Mbps**

Memory-to-Memory Transfers

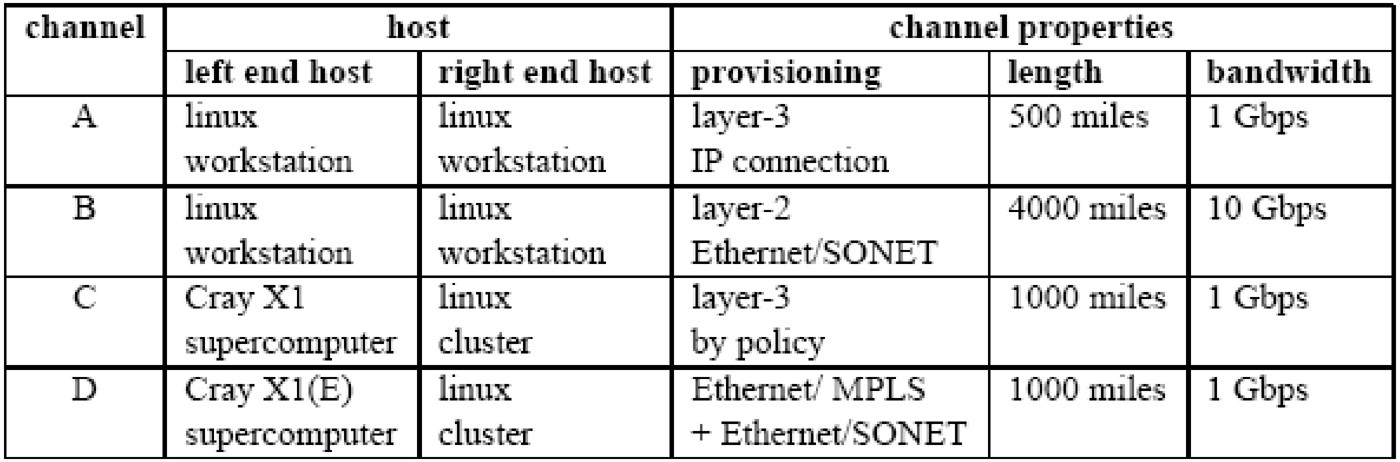
UDT: 958Mbps

Both Iperf and throughput profiles indicated 990 Mbps levels

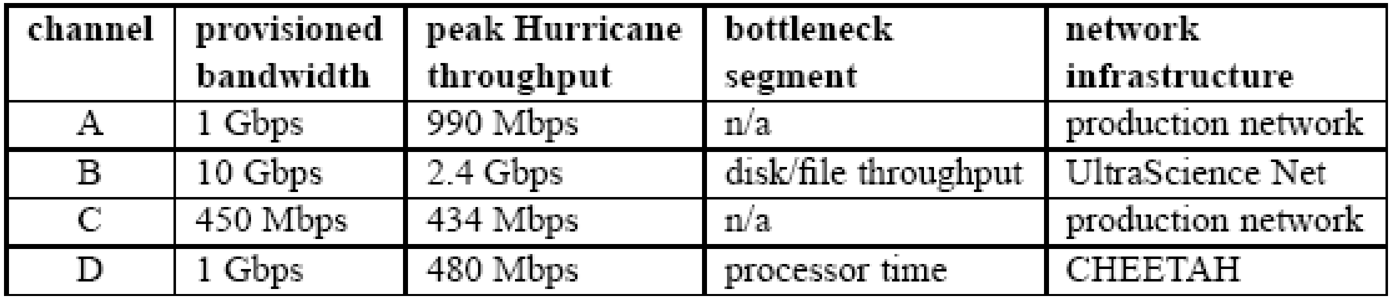
Potentially such rates are achievable if disk access and protocol parameters are tuned



Summary of Hurricane Protocol Performance



channel	host		channel properties		
	left end host	right end host	provisioning	length	bandwidth
A	linux workstation	linux workstation	layer-3 IP connection	500 miles	1 Gbps
B	linux workstation	linux workstation	layer-2 Ethernet/SONET	4000 miles	10 Gbps
C	Cray X1 supercomputer	linux cluster	layer-3 by policy	1000 miles	1 Gbps
D	Cray X1(E) supercomputer	linux cluster	Ethernet/ MPLS + Ethernet/SONET	1000 miles	1 Gbps



channel	provisioned bandwidth	peak Hurricane throughput	bottleneck segment	network infrastructure
A	1 Gbps	990 Mbps	n/a	production network
B	10 Gbps	2.4 Gbps	disk/file throughput	UltraScience Net
C	450 Mbps	434 Mbps	n/a	production network
D	1 Gbps	480 Mbps	processor time	CHEETAH

Adhoc Optimizations

- Manual tuning of parameters
    - Wait-time parameter: $T_s(t)$
        - Initial value chosen from throughput profile
        - Empirically, goodput is "unimodel" in $T_s(t)$: pairwise measurements for binary search
    - Group size for *k* for NACKs
        - empirically, goodput is unimodel in *k* and is tuned
- Disk-specific details
    - Reads done in batch – no input buffer
    - NAKs are handled using fseek – attached to the next batch

- This tuning is not likely to be transferable to other configurations and different host loads
    - More work needed: automatic tuning and systematic analysis

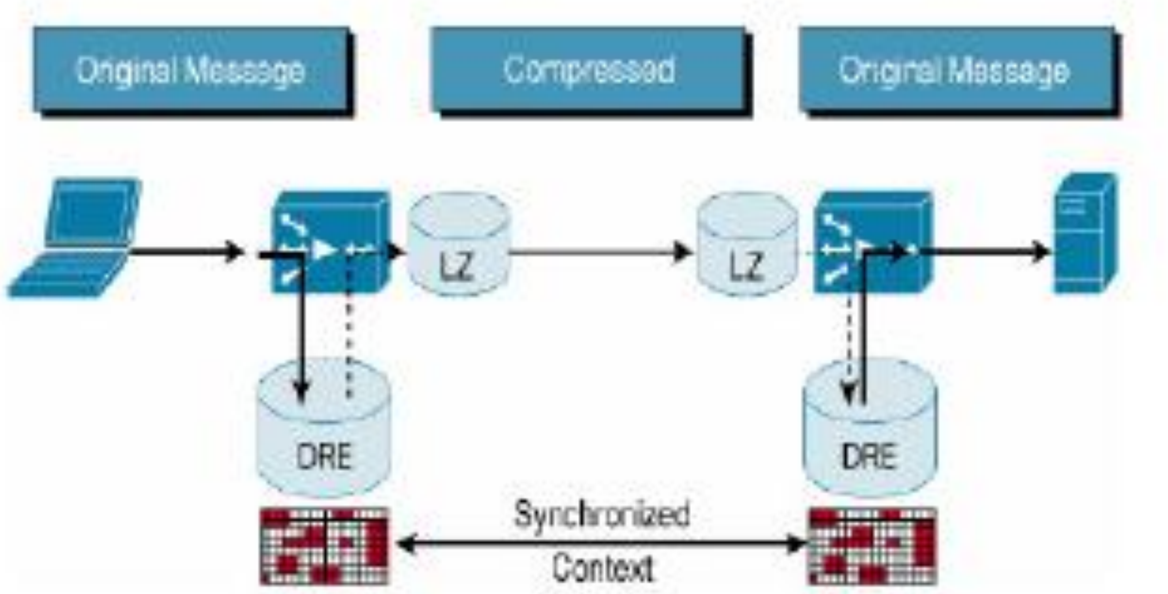
Outline

- **Motivation and Background**
- **USN infrastructure**
  - Architecture
  - Data-plane
  - Control-plane
  - Connection Suites
- **USN Networking Experiments**
  - Hybrid Network Connections
  - Infiniband over Wide-Area
  - Connections to Supercomputers
  - Transport Methods for Dedicated Channels
  - Wide-Area Application Accelerators
  - Encryption Devices

Transport Improvements Based on Data Contents

Examines payload contents to improve network throughputs:
- Can achieve data transfer rates higher than connection capacities

Three separate optimization methods implemented by Cisco WAE devices:
TFO – TCP Flow Optimization
DRE - Data Redundancy Elimination for aggregate flows
LZ – Limple-Ziv Data Compression on per flow basis



Experiments Overview

Detailed experimental analysis of effects of:

TFO – TCP Flow Optimization
DRE - Data Redundancy Elimination
LZ – Limple-Ziv Data Compression
All options

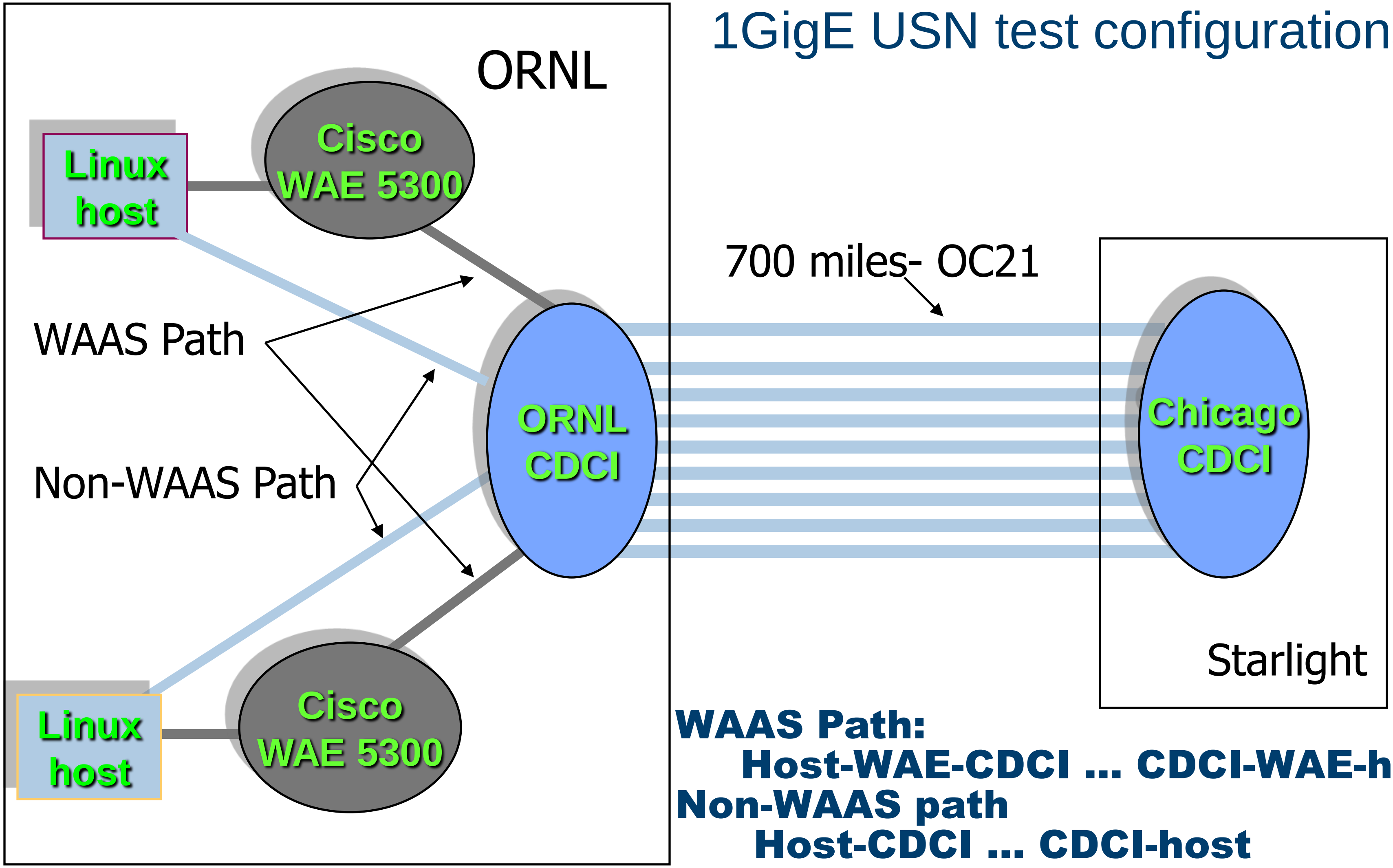
Performance affects on file transfers:

- Duplicated contents
- Uniformly random contents - baseline for non-compressible data
- Gziped uniformly random contents
- Terascale supernova files – HDF format – used extensively in scientific applications
- Gziped Terascale supernova files

Compression ratios using gzip on complete files

Duplicated contents - gziped file is 1030 times compressed
Uniformly random contents – gziped version is slightly larger (0.01%)
HDF supernova datasets – gziped version is 0.6831 times original size

1GigE USN test configurations



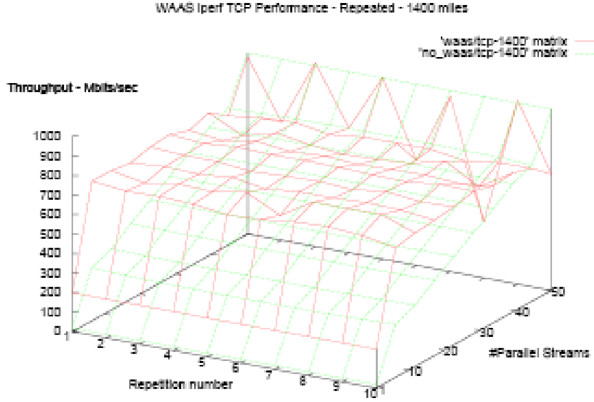
WAAS Path:
Host-WAE-CDCI … CDCI-WAE-host
Non-WAAS path
Host-CDCI … CDCI-host

ORNL–Chicago loop: 1400 miles

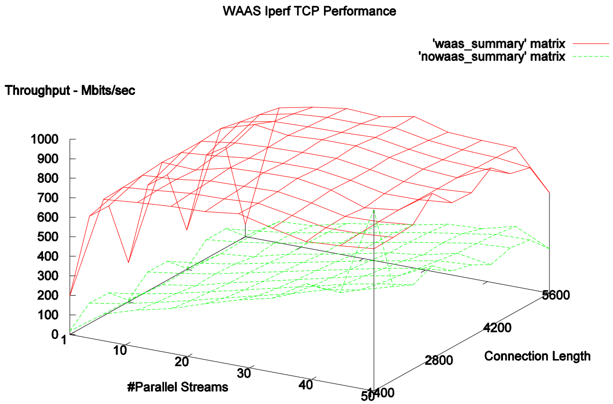
Multiple loops: 2800, 4200, 5600 miles

Throughput Performance Profile Examples
To Capture Overall Qualitative Behavior

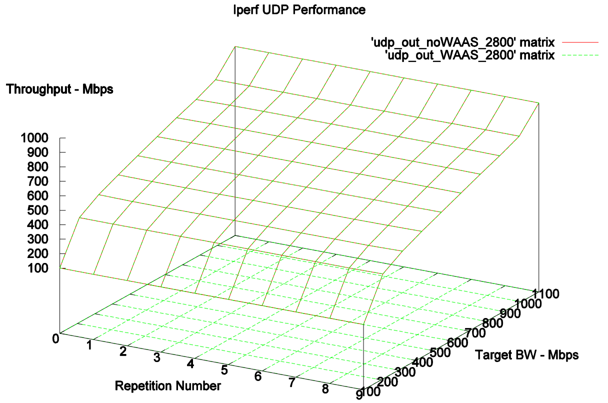
TCP throughput:
Repetition and #streams



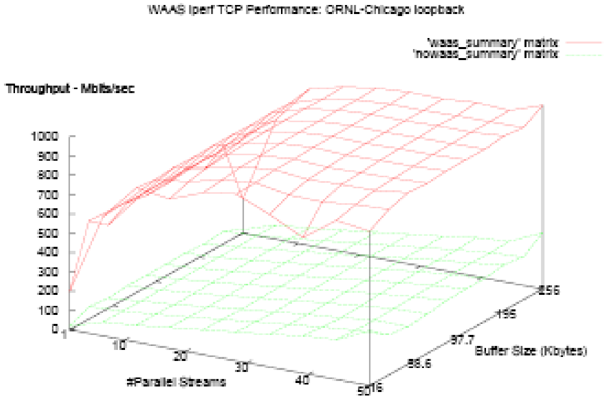
TCP throughput:
#streams and connection length



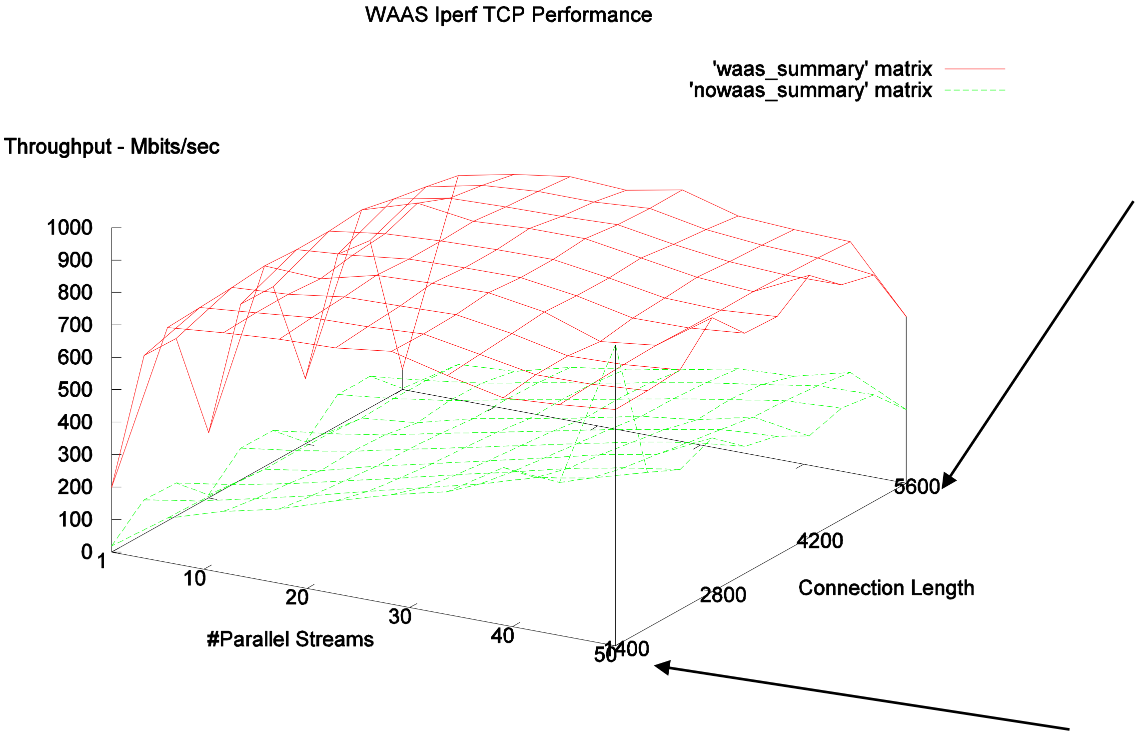
UDPP throughput:
Repetition and target rate

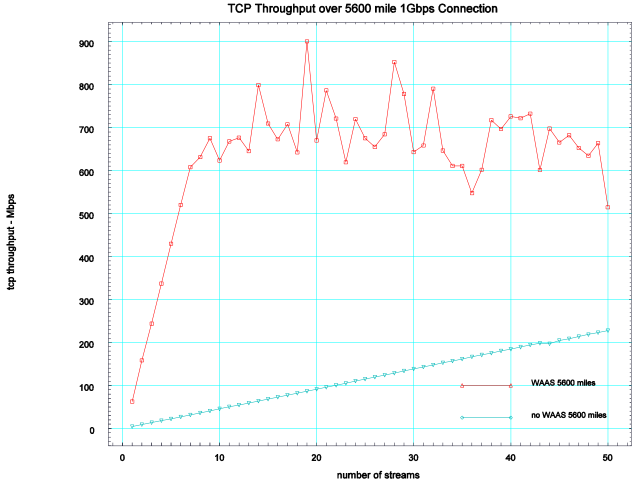


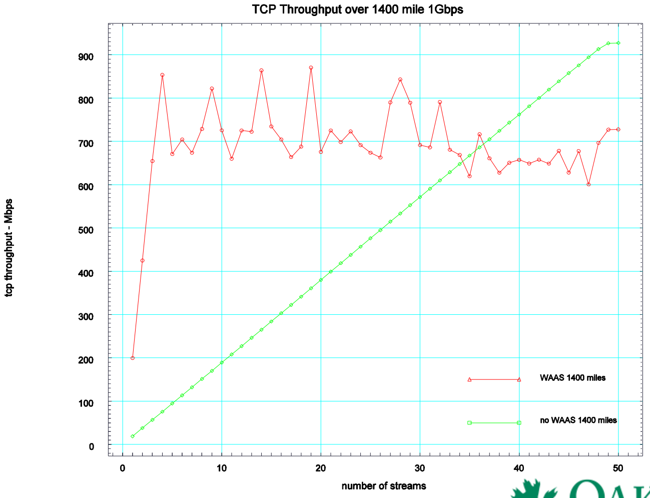
TCP throughput:
#streams and buffersize



Average TCP iperf Throughput – Distance Scalability

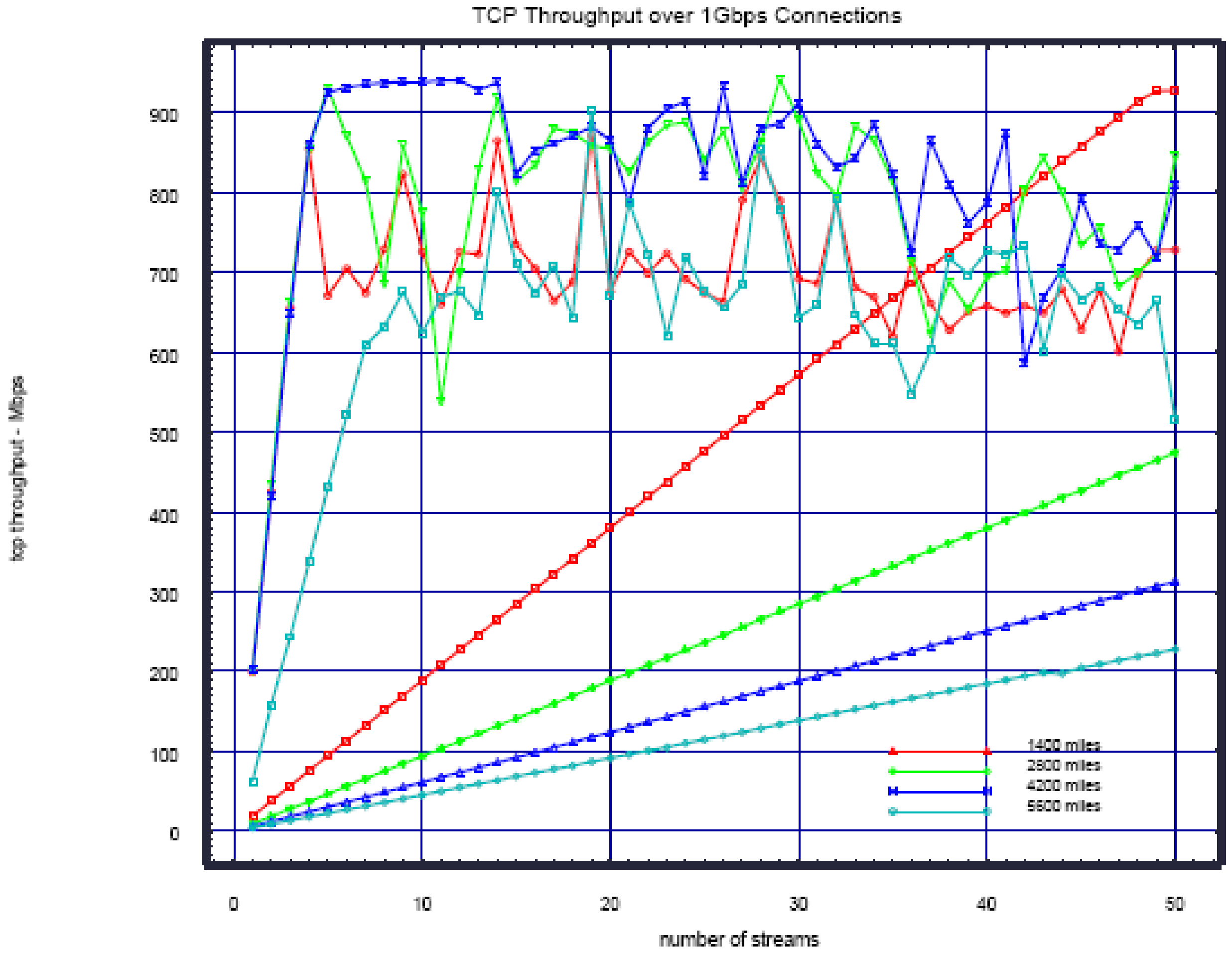






WAAS scales well with distance
Peak performance is reached with <10 streams

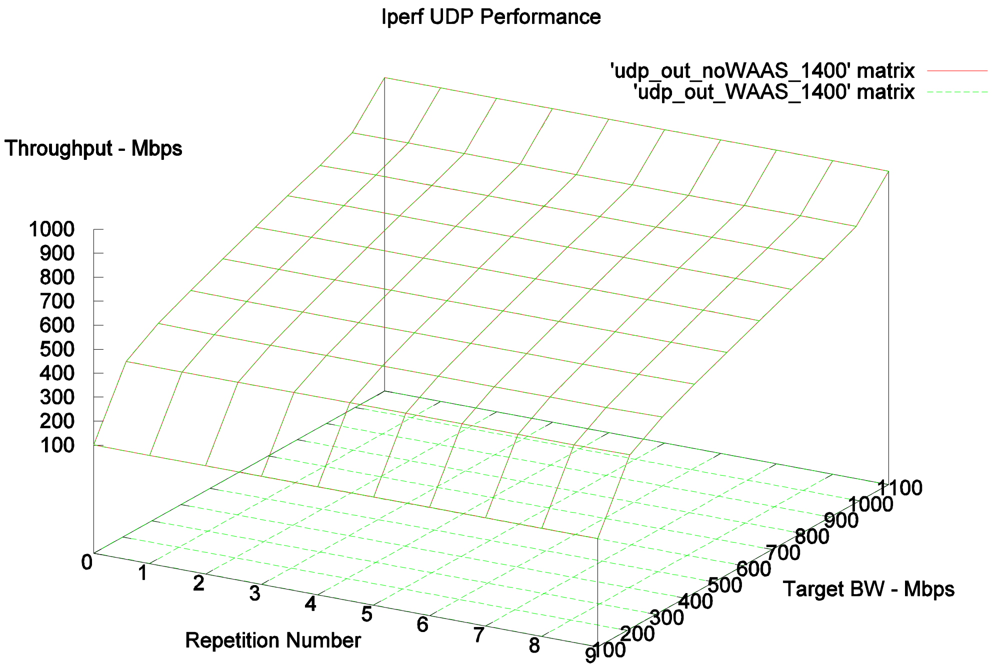
Typical Performance of Parallel-TCP iperf

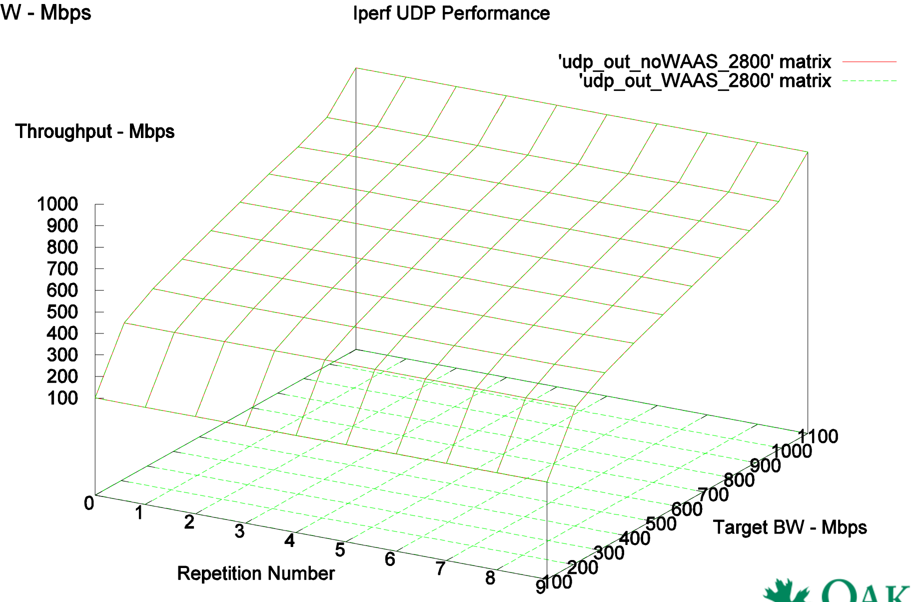

TCP Throughput over 1Gbps Connections

WAAS performance scales well with distance

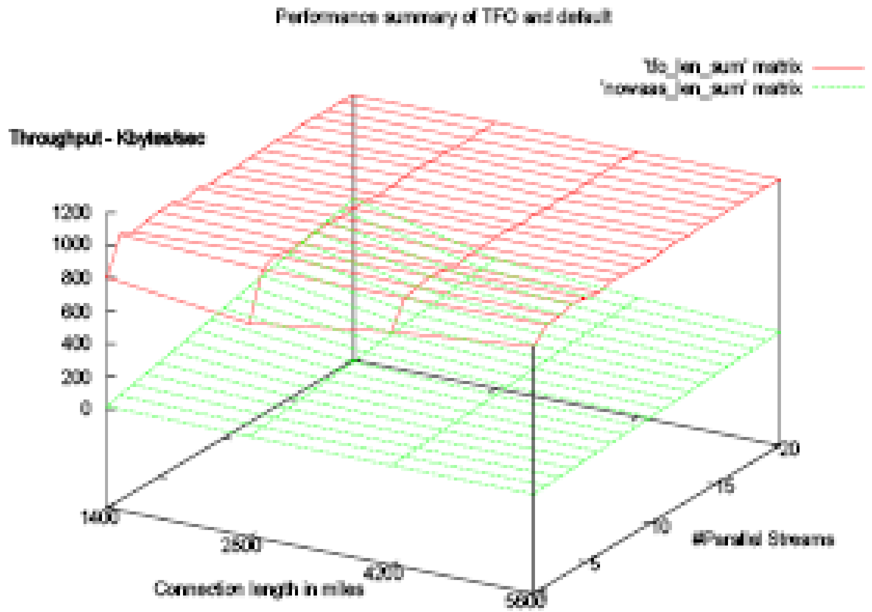
Non-Monotonic with respect to number of streams

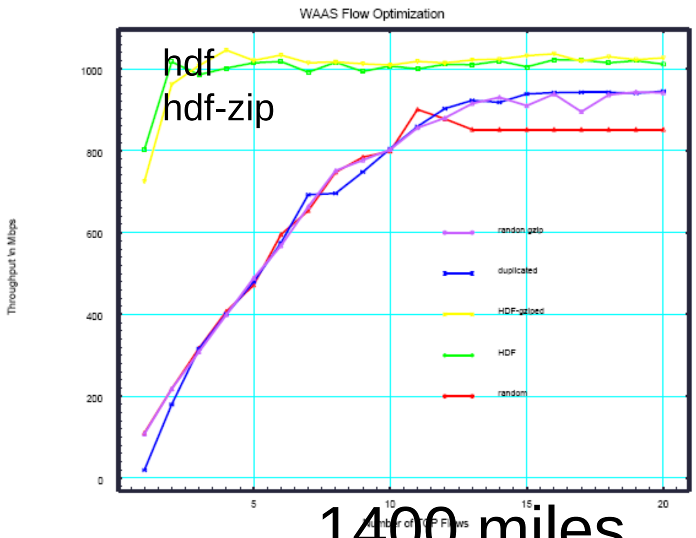
UDP iperf Performance is unaffected





TCP Flow Optimization





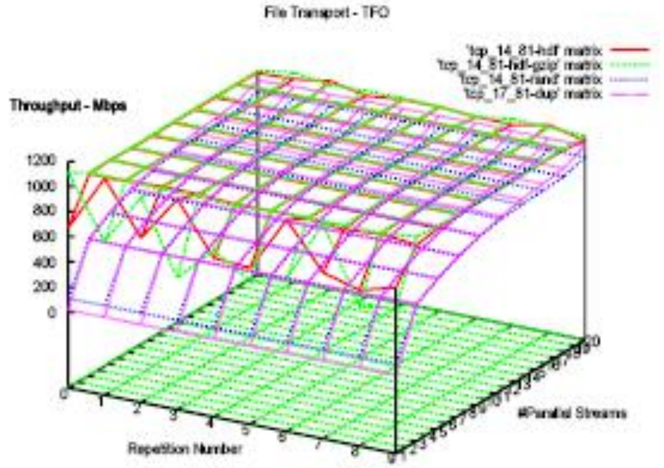
hdf
hdf-zip

1400 miles

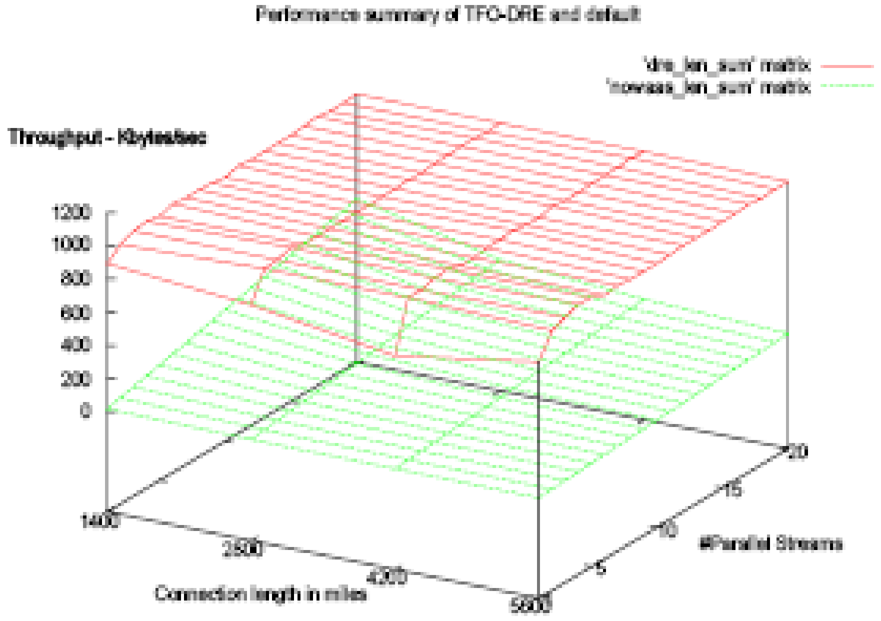
HDF files have good performance
-Gzip did not make much difference

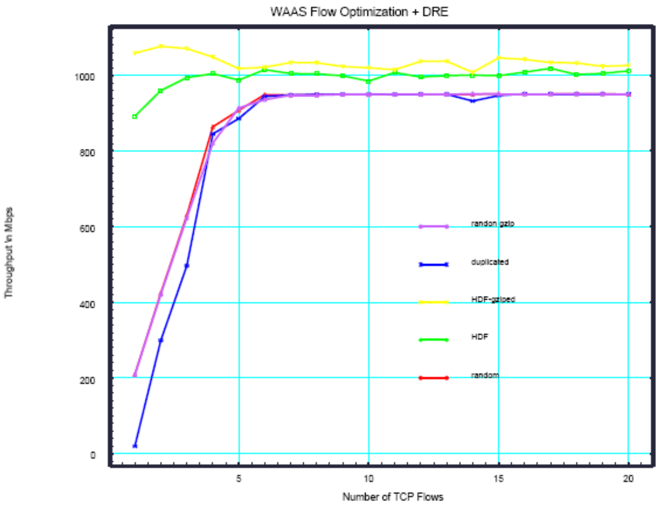
-Uniform random contents are most challenging
-Gzip again did not make much difference

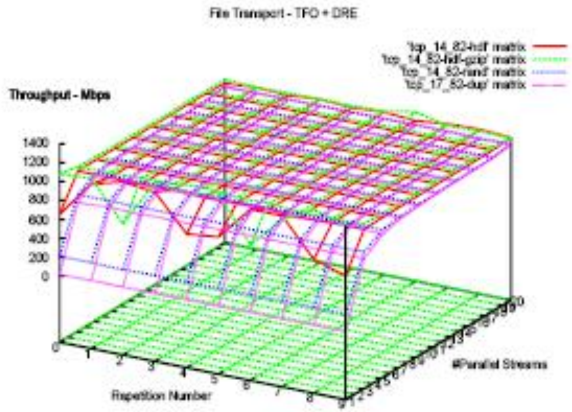
-Duplicated contents performed same as random



TCP Flow Optimization + Data Redundancy Elimination



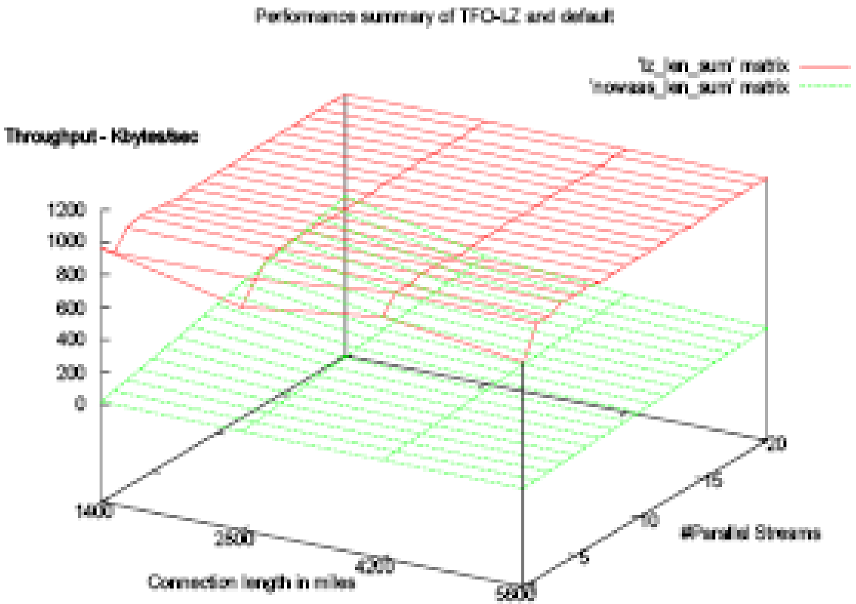


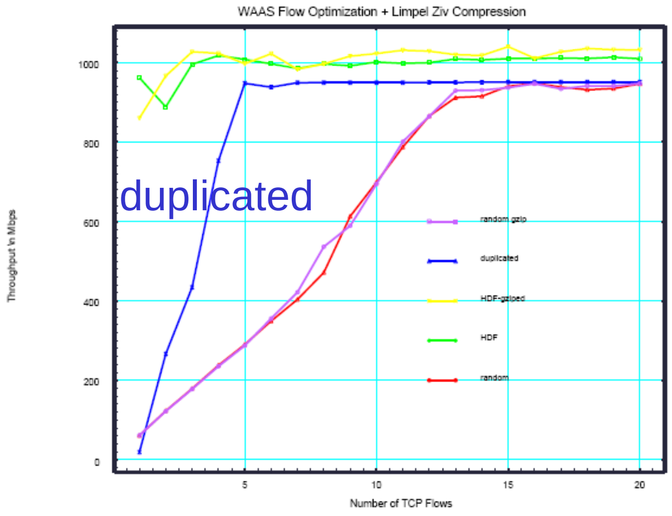


DRE improved all cases, but relative behaviors is same as TFO
  HDF files have good performance
  Gzip did not make much difference

  Uniform random contents are most challenging
  Gzip again did not make much difference
  Duplicated contents performed same as random

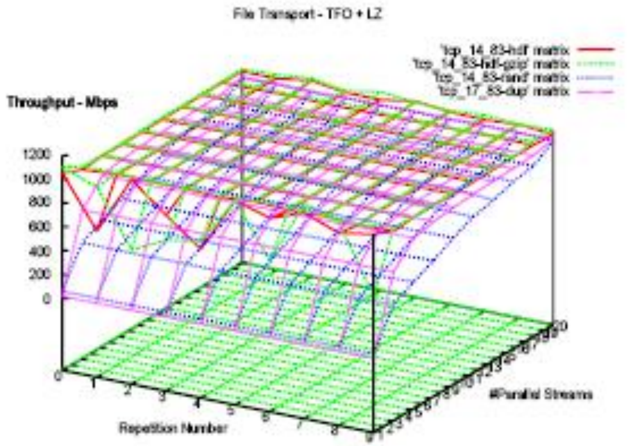
TCP Flow Optimization + Limpel-Ziv Compression





HDF files have good performance
-Gzip did not make much difference

-Uniform random contents are most challenging
-Gzip again did not make much difference

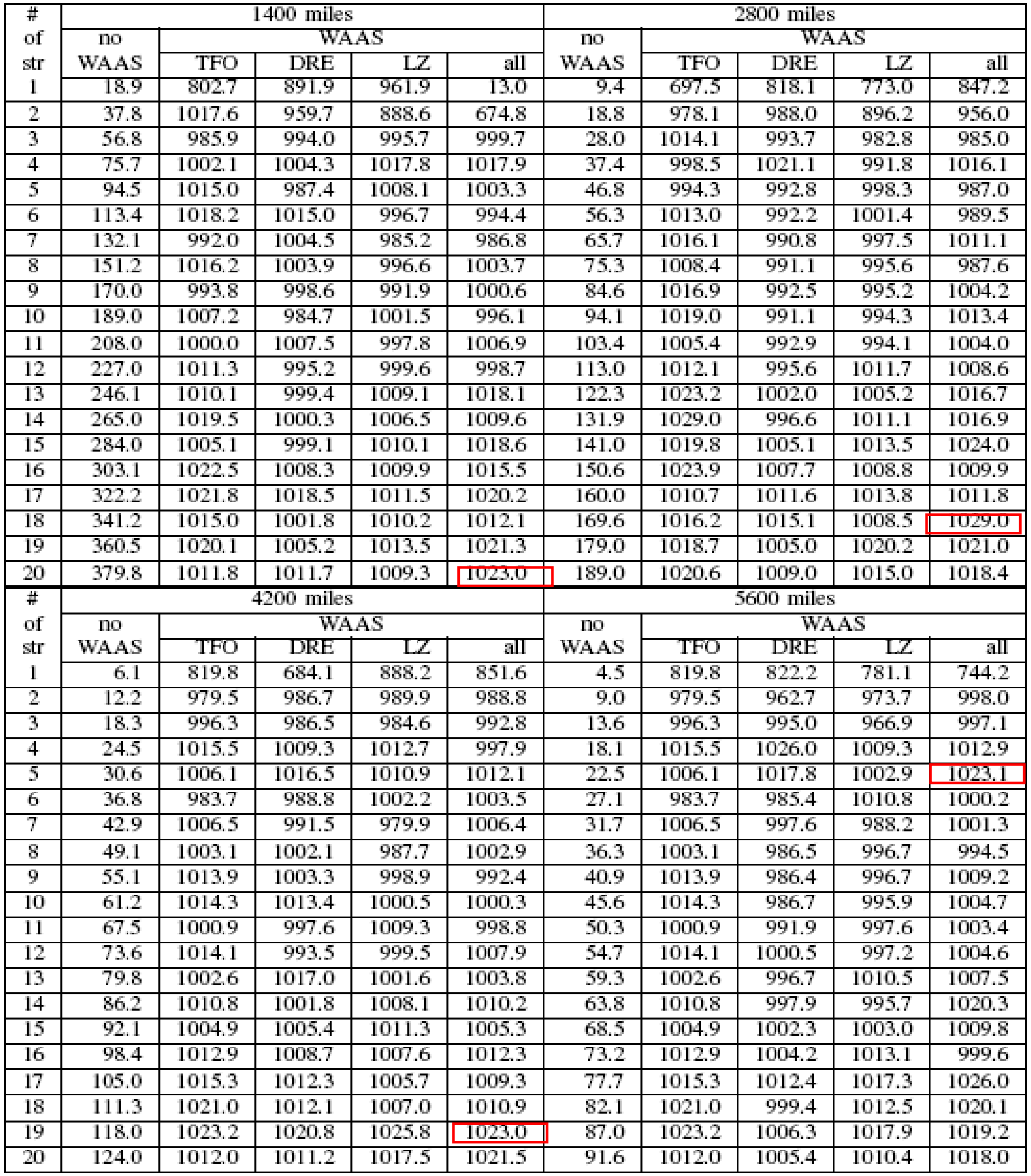


-Duplicated contents performed much better than random

Measurements for hdf files

- Most-effective on hdf files:
    - 1.02Gbps on 1GigE connection
- Scalability up to 5600 miles with essentially no decrease
    - 1.023 Gbps
- Non-monotonic throughput with increased number of streams
- Needed multiple streams to reach highest throughput
    - 20 at 1400 miles
    - 18 at 2800 miles
    - 19 at 4200 miles
    - 5 at 5600 miles

- Least-effective on files with uniform random contents
- Gzipping the files did not make much difference



# of str	1400 miles					2800 miles				
	no WAAS	WAAS				no WAAS	WAAS			
		TFO	DRE	LZ	all		TFO	DRE	LZ	all
1	18.9	802.7	891.9	961.9	13.0	9.4	697.5	818.1	773.0	847.2
2	37.8	1017.6	959.7	888.6	674.8	18.8	978.1	988.0	896.2	956.0
3	56.8	985.9	994.0	995.7	999.7	28.0	1014.1	993.7	982.8	985.0
4	75.7	1002.1	1004.3	1017.8	1017.9	37.4	998.5	1021.1	991.8	1016.1
5	94.5	1015.0	987.4	1008.1	1003.3	46.8	994.3	992.8	998.3	987.0
6	113.4	1018.2	1015.0	996.7	994.4	56.3	1013.0	992.2	1001.4	989.5
7	132.1	992.0	1004.5	985.2	986.8	65.7	1016.1	990.8	997.5	1011.1
8	151.2	1016.2	1003.9	996.6	1003.7	75.3	1008.4	991.1	995.6	987.6
9	170.0	993.8	998.6	991.9	1000.6	84.6	1016.9	992.5	995.2	1004.2
10	189.0	1007.2	984.7	1001.5	996.1	94.1	1019.0	991.1	994.3	1013.4
11	208.0	1000.0	1007.5	997.8	1006.9	103.4	1005.4	992.9	994.1	1004.0
12	227.0	1011.3	995.2	999.6	998.7	113.0	1012.1	995.6	1011.7	1008.6
13	246.1	1010.1	999.4	1009.1	1018.1	122.3	1023.2	1002.0	1005.2	1016.7
14	265.0	1019.5	1000.3	1006.5	1009.6	131.9	1029.0	996.6	1011.1	1016.9
15	284.0	1005.1	999.1	1010.1	1018.6	141.0	1019.8	1005.1	1013.5	1024.0
16	303.1	1022.5	1008.3	1009.9	1015.5	150.6	1023.9	1007.7	1008.8	1009.9
17	322.2	1021.8	1018.5	1011.5	1020.2	160.0	1010.7	1011.6	1013.8	1011.8
18	341.2	1015.0	1001.8	1010.2	1012.1	169.6	1016.2	1015.1	1008.5	1029.0
19	360.5	1020.1	1005.2	1013.5	1021.3	179.0	1018.7	1005.0	1020.2	1021.0
20	379.8	1011.8	1011.7	1009.3	1023.0	189.0	1020.6	1009.0	1015.0	1018.4

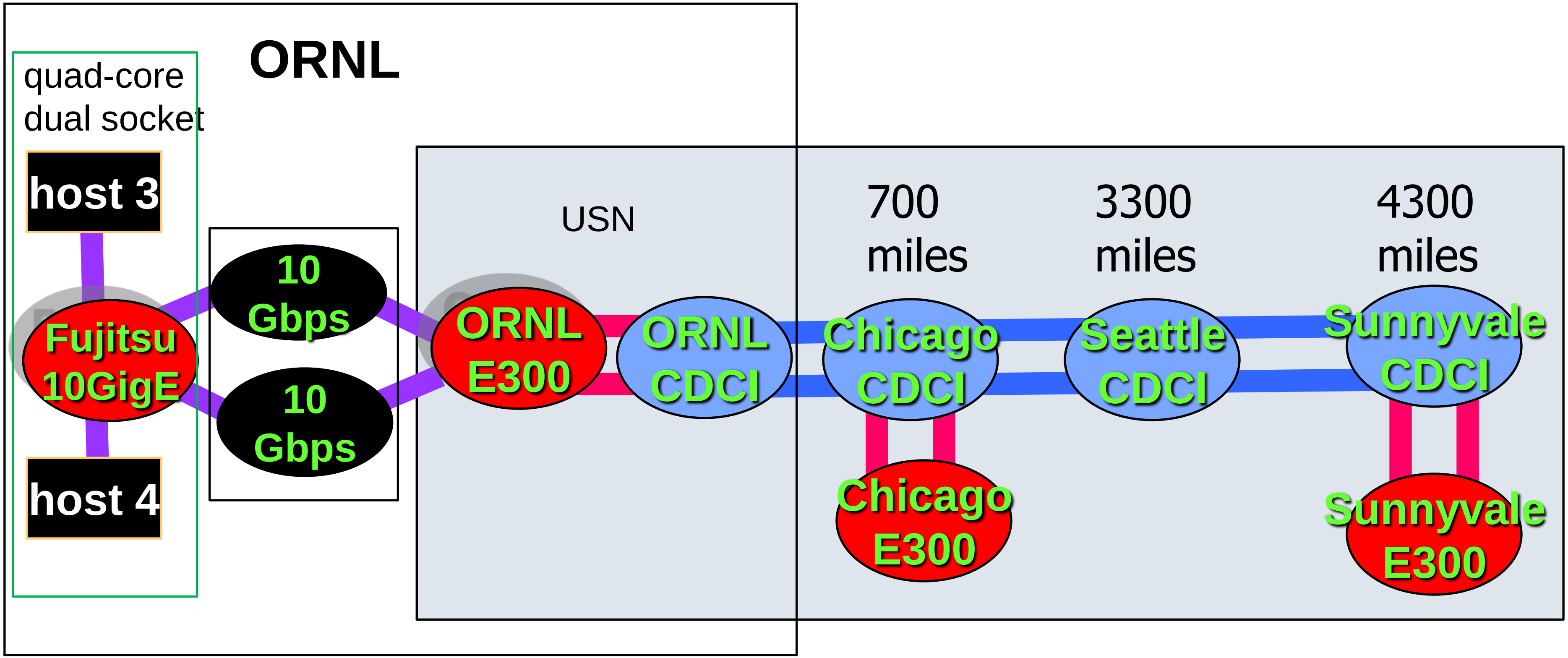
# of str	4200 miles					5600 miles				
	no WAAS	WAAS				no WAAS	WAAS			
		TFO	DRE	LZ	all		TFO	DRE	LZ	all
1	6.1	819.8	684.1	888.2	851.6	4.5	819.8	822.2	781.1	744.2
2	12.2	979.5	986.7	989.9	988.8	9.0	979.5	962.7	973.7	998.0
3	18.3	996.3	986.5	984.6	992.8	13.6	996.3	995.0	966.9	997.1
4	24.5	1015.5	1009.3	1012.7	997.9	18.1	1015.5	1026.0	1009.3	1012.9
5	30.6	1006.1	1016.5	1010.9	1012.1	22.5	1006.1	1017.8	1002.9	1023.1
6	36.8	983.7	988.8	1002.2	1003.5	27.1	983.7	985.4	1010.8	1000.2
7	42.9	1006.5	991.5	979.9	1006.4	31.7	1006.5	997.6	988.2	1001.3
8	49.1	1003.1	1002.1	987.7	1002.9	36.3	1003.1	986.5	996.7	994.5
9	55.1	1013.9	1003.3	998.9	992.4	40.9	1013.9	986.4	996.7	1009.2
10	61.2	1014.3	1013.4	1000.5	1000.3	45.6	1014.3	986.7	995.9	1004.7
11	67.5	1000.9	997.6	1009.3	998.8	50.3	1000.9	991.9	997.6	1003.4
12	73.6	1014.1	993.5	999.5	1007.9	54.7	1014.1	1000.5	997.2	1004.6
13	79.8	1002.6	1017.0	1001.6	1003.8	59.3	1002.6	996.7	1010.5	1007.5
14	86.2	1010.8	1001.8	1008.1	1010.2	63.8	1010.8	997.9	995.7	1020.3
15	92.1	1004.9	1005.4	1011.3	1005.3	68.5	1004.9	1002.3	1003.0	1009.8
16	98.4	1012.9	1008.7	1007.6	1012.3	73.2	1012.9	1004.2	1013.1	999.6
17	105.0	1015.3	1012.3	1005.7	1009.3	77.7	1015.3	1012.4	1017.3	1026.0
18	111.3	1021.0	1012.1	1007.0	1010.9	82.1	1021.0	999.4	1012.5	1020.1
19	118.0	1023.2	1020.8	1025.8	1023.0	87.0	1023.2	1006.3	1017.9	1019.2
20	124.0	1012.0	1011.2	1017.5	1021.5	91.6	1012.0	1005.4	1010.4	1018.0

TABLE II
AVERAGE OF THROUGHPUTS FOR HDF FILES OVER 10 REPETITIONS.

Outline

- **Motivation and Background**
- **USN infrastructure**
  - Architecture
  - Data-plane
  - Control-plane
  - Connection Suites
- **USN Networking Experiments**
  - Hybrid Network Connections
  - Infiniband over Wide-Area
  - Connections to Supercomputers
  - Transport Methods for Dedicated Channels
  - Wide-Area Application Accelerators
  - Encryption Devices

Test Configuration



ORNL loop -0.2 mile

ORNL-Chicago loop – 1400 miles

ORNL- Chicago - Seattle loop – 6600 miles

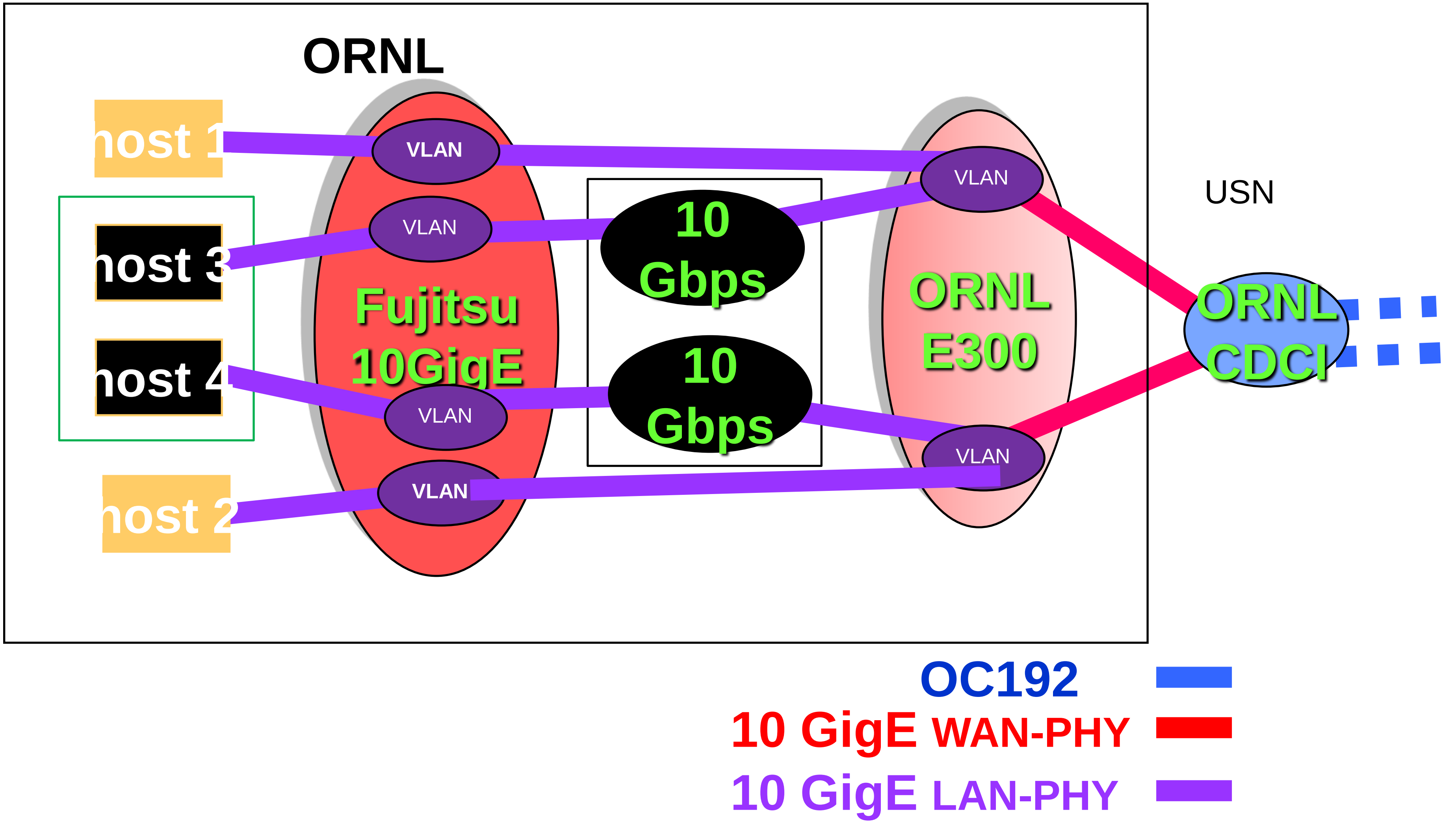
ORNL – Chicago – Seattle - Sunnyvale loop – 8600 miles

OC192 🟦

10 GigE WAN-PHY 🟥

10 GigE LAN-PHY 🟪

host1-host2 Connections
host3-host4 Connections through 10Gbps Devices

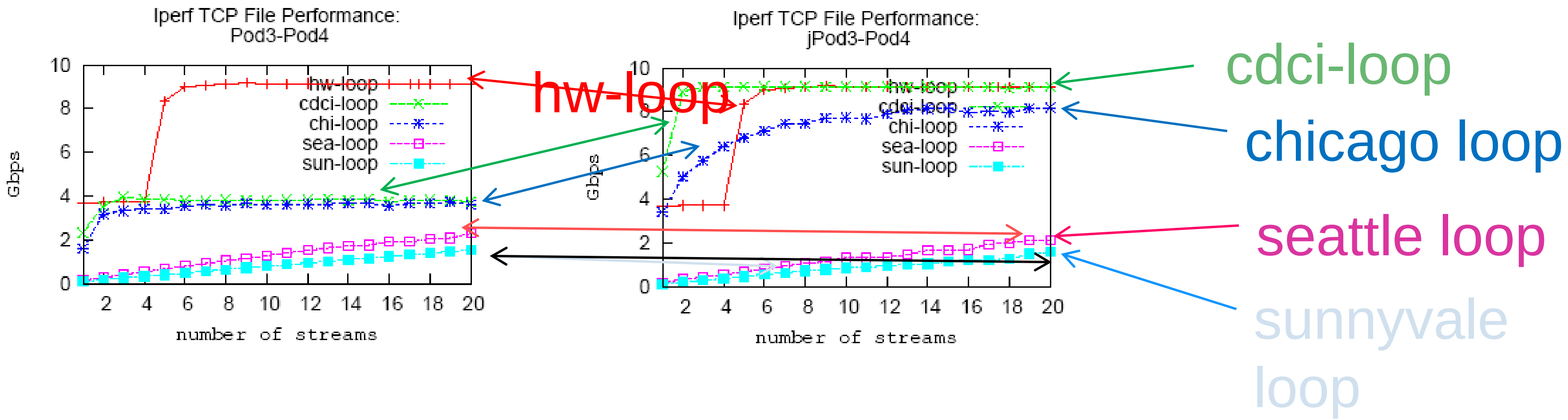


TCP Profiles: Before and after MTU Alignment host3-4 Encrypted Connection: File transfer

- Fiber loop between 10Gbps devices : 9 Gbps TCP throughput
- When connected to E300: 9Gbps throughput locally
  - MTU size is modified on E300
  - IP segment/datagram size set to 8950

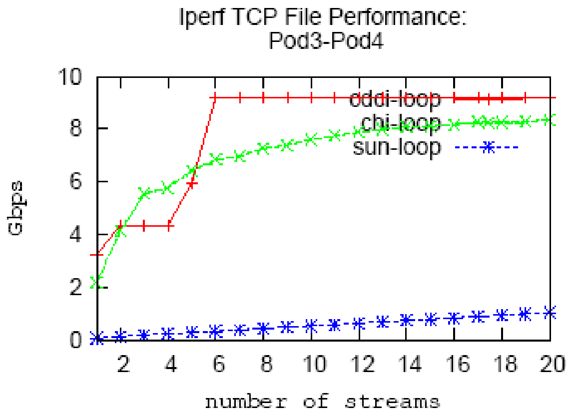
1400 byte MTU

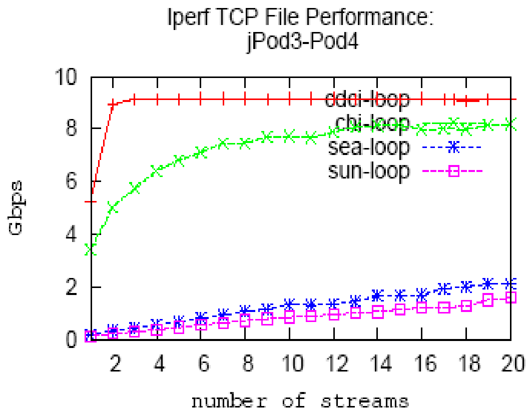
jumbogram



TCP Profiles Comparison:
Better Throughput with 10Gbps devices
host1-2 Plain and host3-4 Encrypted Connections

Fiber loop between 10Gbps devices : 9 Gbps TCP throughput
Chicago loop: host3-4 connection achieved 8Gbps
Sunnyvale loop: host3-4 connection 1.5 time higher throughput





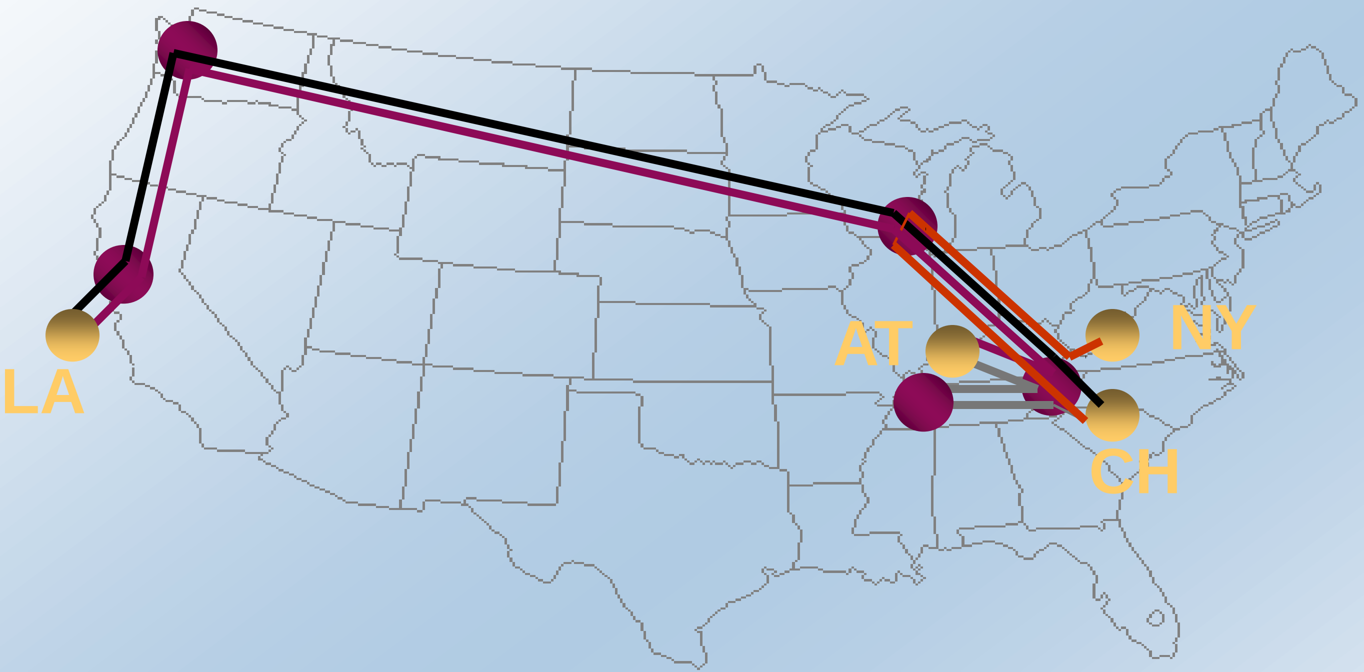
Observations: Compared to plain connections, for encrypted connections:

- High throughput is achieved with less number of streams
- Higher throughput is achieved at longer distances

Realizations on Extended USN
Specified target national-wide network

● Target location for third-party switch

Realization of Target Network on Proposed Extended USN (E-USN) with new node in Memphis



● Third party switch – Actual locations on E-USN

One at Sunnyvale – three at ORNL

● E-USN switches

Summary: USN Project

- **USN infrastructure**
  - 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, Infiniband over wide-area.
- **USN's research role in advanced networking capabilities**
  - Networking technologies
    - Connectivity to supercomputers
    - Testing of file systems: Lustre over TCP/IP and Inifiniband/SONET
  - Hybrid optical packet and switching technologies
    - VLAN testing and analysis over L1-2 and MPLS connections
    - Configuration and testing of hybrid connections