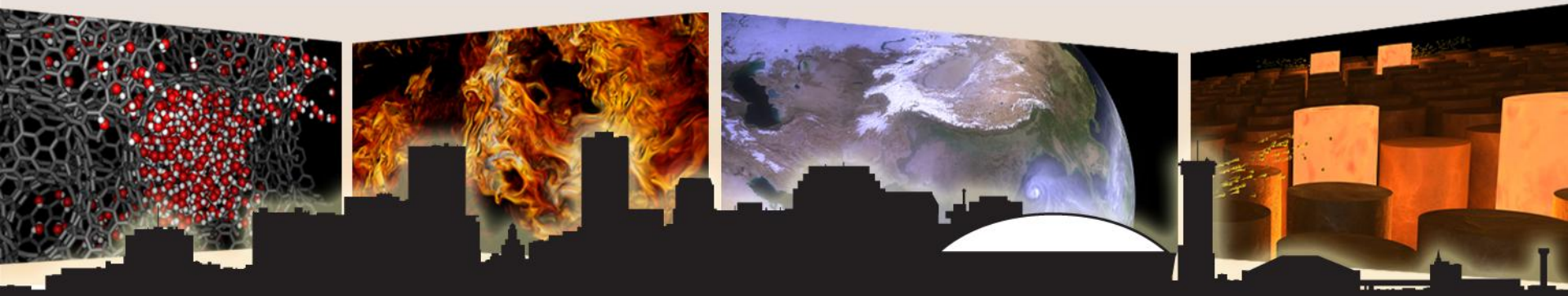
DOE UltraScience Net: High-Performance Experimental Network Research Testbed

Presented by

Nagi Rao
Steve Poole

Extreme Scale Systems Center
Computer Science and Mathematics Division

Research currently supported by the
Department of Defense and previously supported by the
Department of Energy's Office of Science
Office of Advanced Scientific Computing
High-Performance Networking Program



U.S. DEPARTMENT OF ENERGY

OAK RIDGE National Laboratory

The need

- Large-scale 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
• 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	• 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

UltraScience Net – In a nutshell

Experimental network research testbed

- **To support advanced networking and related application technologies for 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**

Peered with ESnet, National Science Foundation's CHEETAH, and other networks

USN Contributions

- Provided long haul production links for experimentation
  - 8000 mile 10 Gbps and 70,000 mile 1 Gbps 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

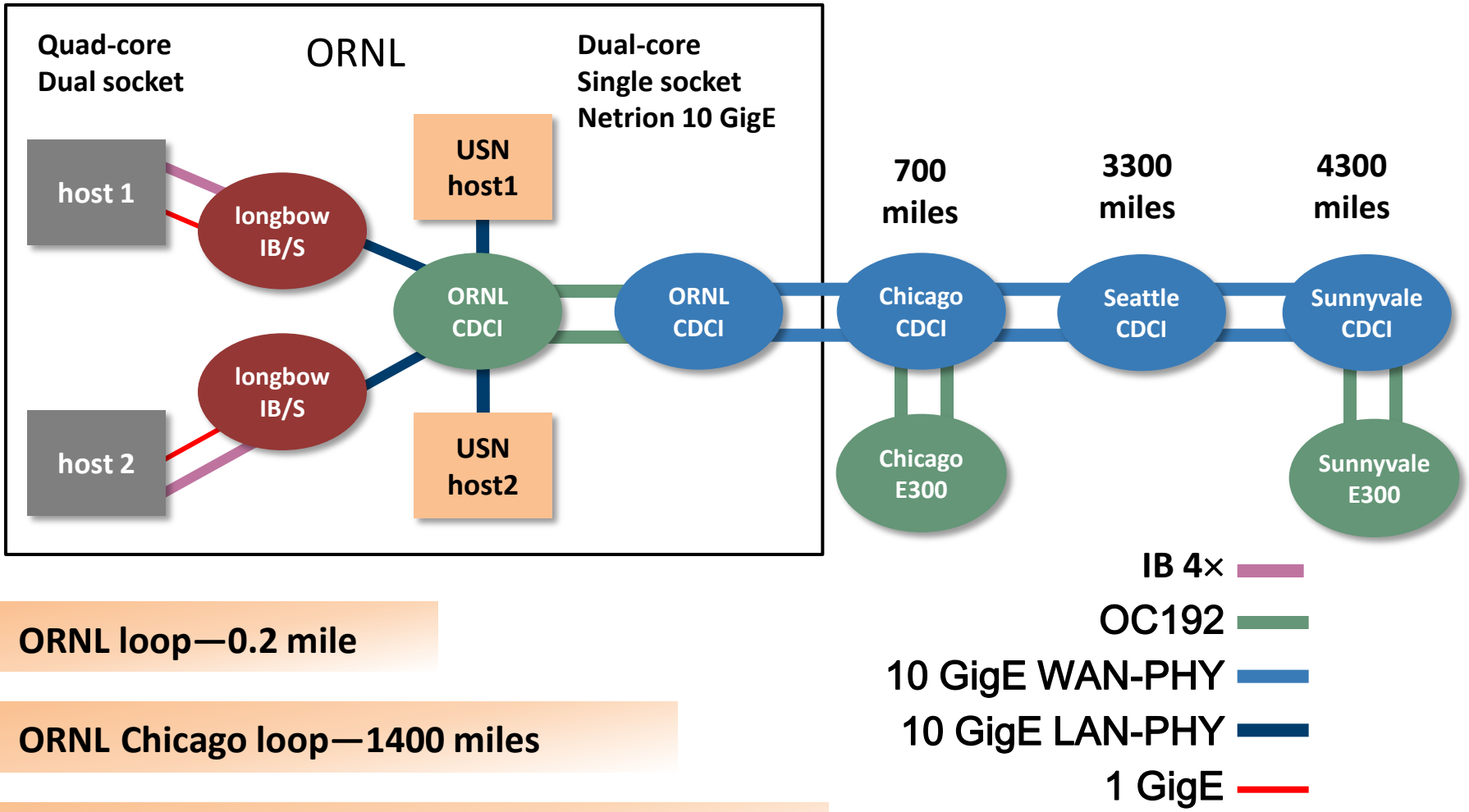
- 10 Gbps Crypto devices
  - TCP performance improved: higher throughput with less #streams

2009

- Cross-calibration of simulations, emulations, and testbed connections
  - Regmented regression to extend measurements to other modalities

2010

InfiniBand over 10 GigE: cross-traffic



ORNL loop—0.2 mile

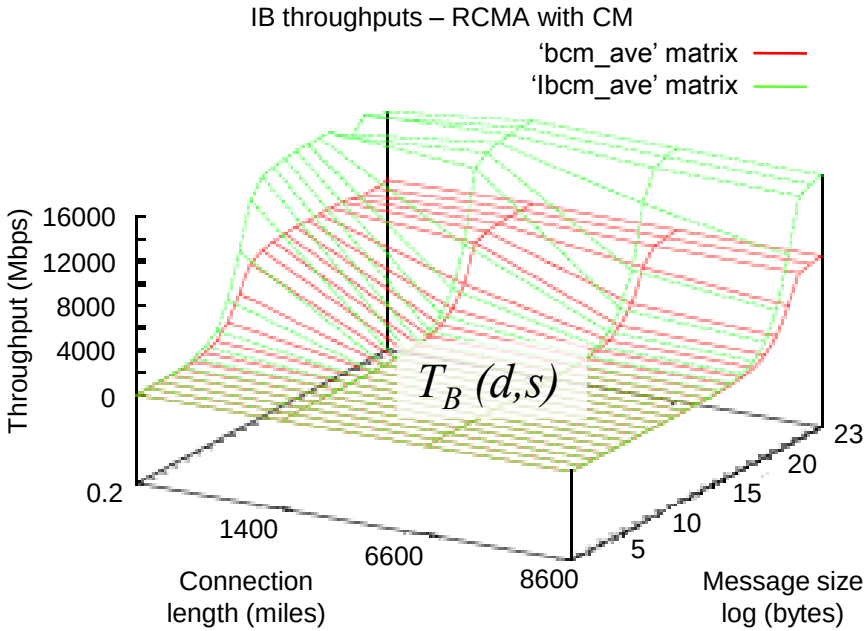
ORNL Chicago loop—1400 miles

ORNL Chicago Seattle loop—6600 miles

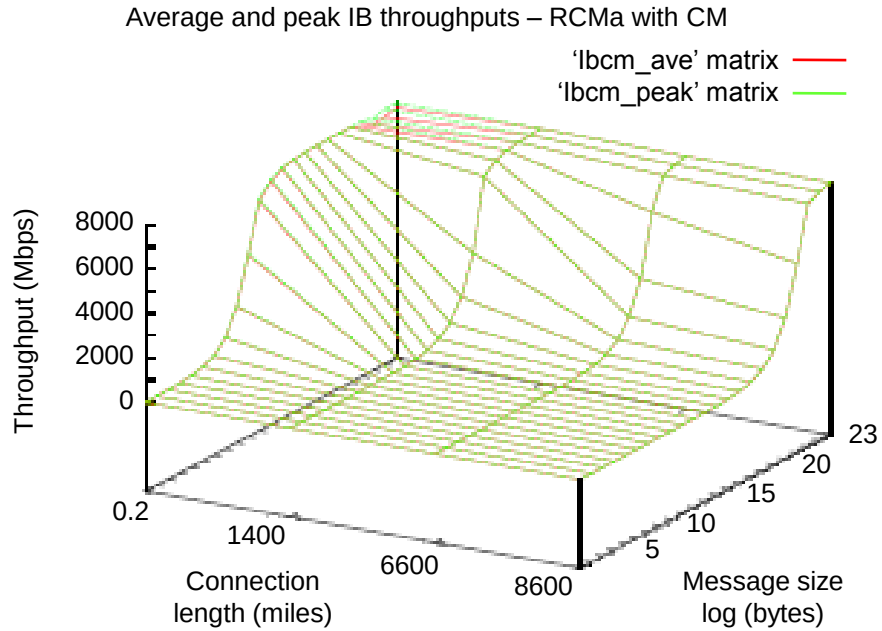
ORNL Chicago Seattle Sunnyvale loop—8600 miles

Performance profiles of IB over 10 GigE

Distance profile



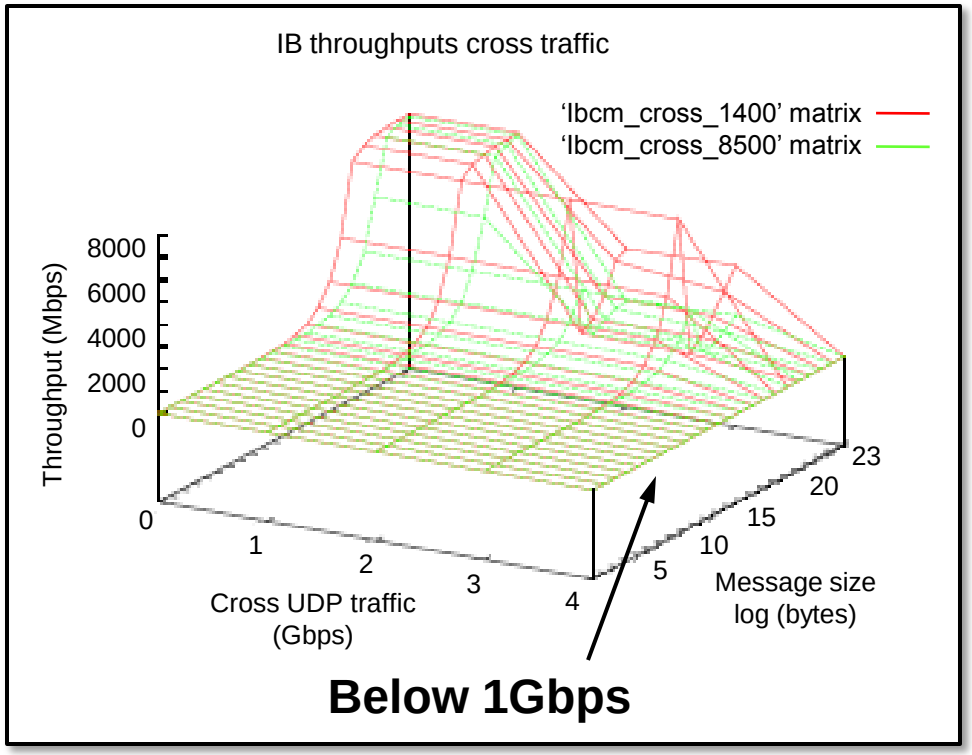
Peak distance profile
Average distance profile



Results are almost the same as in SONET case

Connection length (miles) d_i	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) $D_B(d_i)$	0	0.012	0.017	0.016

Cross-traffic effect of IB over 10 GigE WANPHY



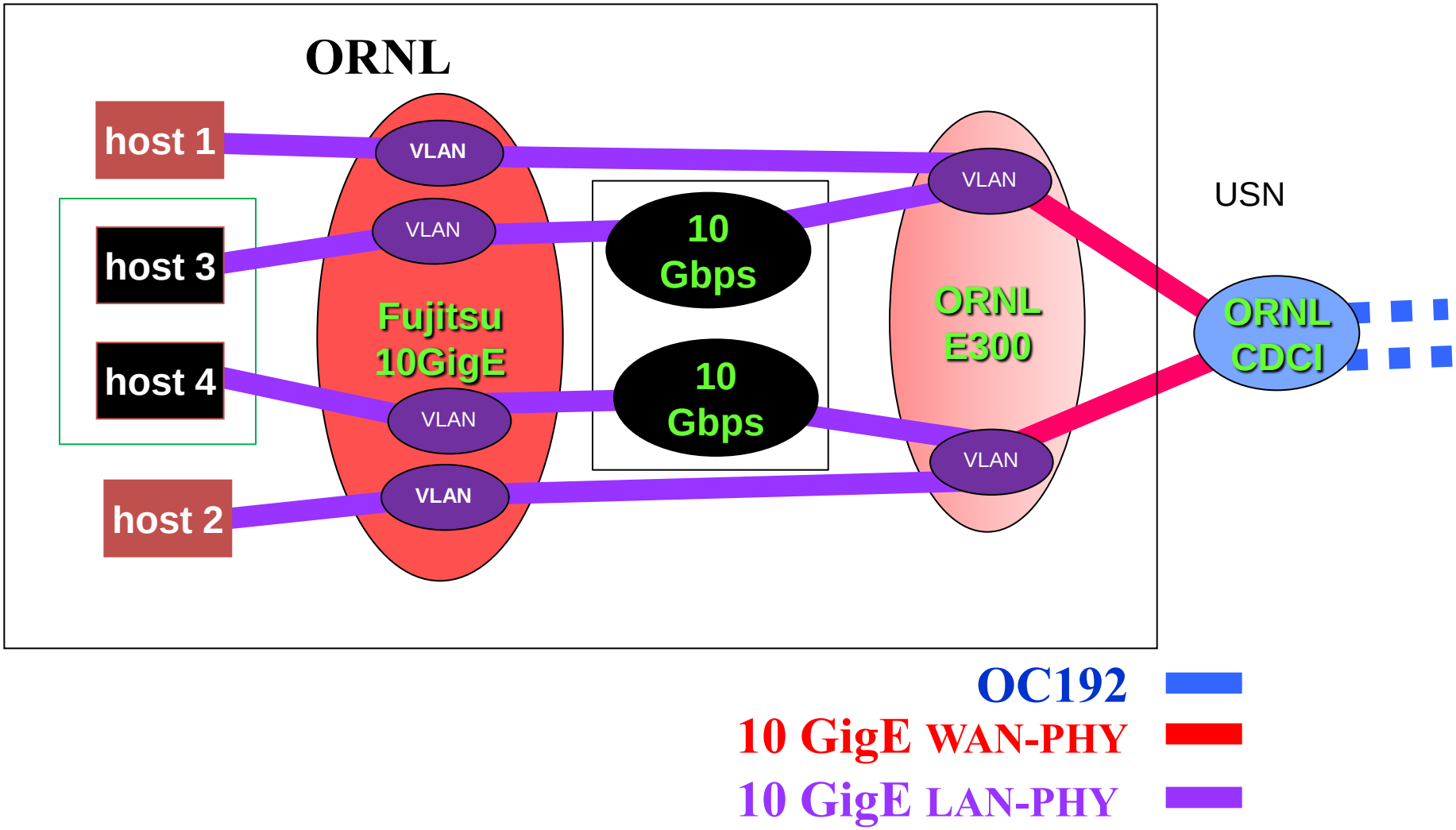
Average throughput for 8M			
Cross-traffic	1400 Miles	6600 Miles	8600 Miles
0 G	7.49	7.39	7.36
1 G	7.49	7.39	7.36
2 G	3.13	1.38	0.74
3 G	3.25	1.97	1.02
4 G	2.91	1.82	0.96

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 affected at cross-traffic level of 4 Gbps
- Effect of cross-traffic is more on large message sizes

Testing of 10 Gbps encryption devices:

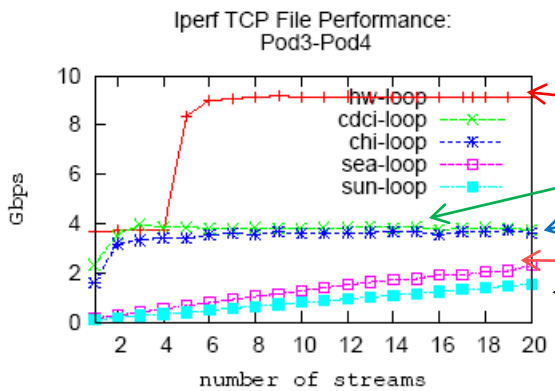
host1-host2 – plain connections
host3-host4 – encrypted connections



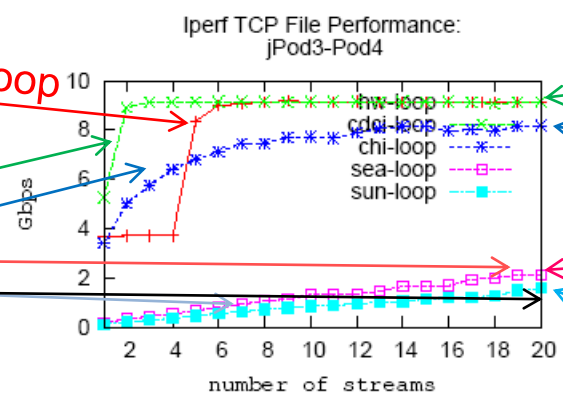
TCP profiles: Before and after MTU alignment host3-4 encrypted connection: File transfer

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

1400 byte MTU



jumbogram



hw-loop

cdci-loop

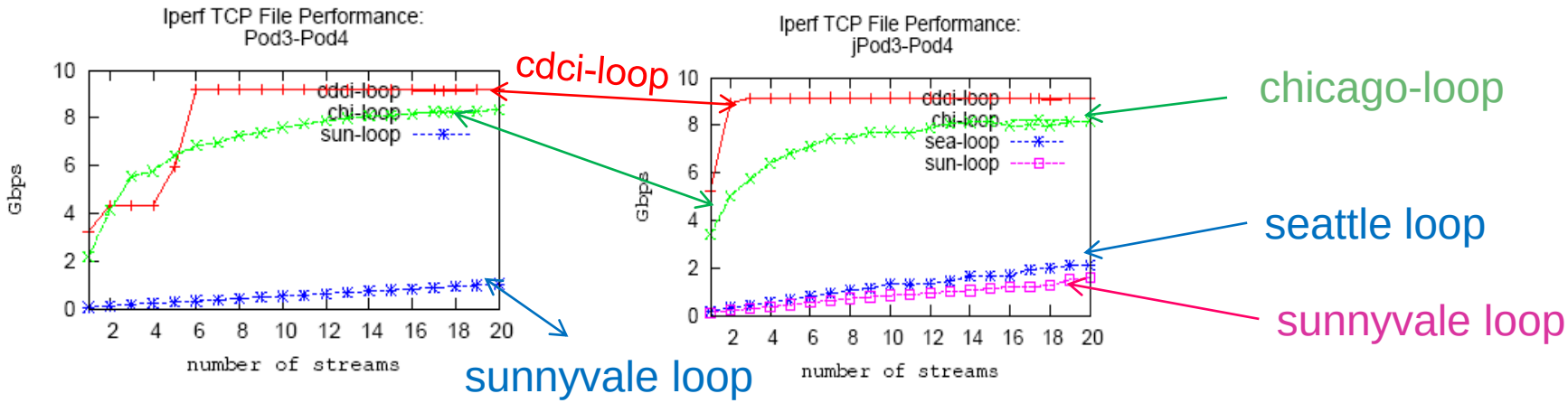
chicago loop

seattle loop

sunnyvale loop

TCP profiles comparison: Better throughput with 10 Gbps devices

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



Compared to plain connections, for encrypted connections we observe that

- **High throughput is achieved with fewer streams**
- **Higher throughput is achieved at longer distances**

Simulation or emulation or realization WAN connections

1. **Network simulation: s/w**
   - **Tools: OPnet, OMNET, NS-2**
2. **Network emulation: s/w + h/w**
   - **Routers and switches with link delays emulated using ANUE**
3. **Fiber loop realizations or USN connections: h/w**
   - **Switches and links are realized as fiber spools or wide-area circuits**

Increased complexity

Closer to real network

	Implementation	Strengths	Limitations
Analytical modeling	Mathematical models and software	Rigorous analysis and design	Challenge to achieve right abstraction
Simulations OPNET and OMNINET	Software on workstations	Broad what-if capability	Limited reflection of networks
ANUE emulations	Laboratory hardware	Closer to actual network	High cost; not mobile
Fiber loop or USN connections	Laboratory hardware, fiber spools	Closest to actual network	Highest cost; not mobile

Differential regression method for cross-calibration

Basic Question: **Predict performance on connection length d not realizable on USN**

Example: IB-RDMA or HTCP throughput on 900 mile connection

$M_S(d)$ Measurements on OPNET simulated path of distance *d*

$M_E(d)$ Measurements on ANUE emulated path of distance *d*

$M_U(d_i)$ Measurements on USN path distance d_i

Measurement Regression: for $A \in \{S, E, U\}$

$\bar{M}_A(.)$ Regression of measurements on

Differential Regression: for $A \in \{S, E, U\}, B \in \{S, E, U\}$

$$\Delta\bar{M}_{A,B}(.) = \bar{M}_A(.) - \bar{M}_B(.)$$

Approach: **Under active development**

1. **Collect simulation or emulation measurement for** d
2. **Apply differential regression to obtain the estimate** $C \in \{S, E\}$

$$\hat{M}_U(d) = M_C(d) - \Delta\bar{M}_{C,U}(d)$$

simulated/emulated measurements

point regression estimate

Conclusions

- USN infrastructure
  - Its architecture has been adopted by LHCnet and Internet2
  - It has provided experimental connections to supercomputers
  - It has enabled testing: VLAN performance, peering of packet circuit switched networks, control plane with advanced reservation, Infiniband and crypto systems over wide area
- USN continues to play a research role in advanced networking capabilities
  - Networking technologies for high-performance computers
    - Connectivity to supercomputers
  - Testing 10 Gbps data transport
    - TCP variants and Inifiniband
    - Encryption devices
  - Cross-calibration with ANUE emulations and OPNET simulations
  - Plans to upgrade USN to 40/100 Gbps in 2010

Contact

Nageswara (Nagi) S. Rao

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

Stephen (Steve) W. Poole

Computer Science and Mathematics Division
(865) 574-9008
spoole@ornl.gov

Acknowledgements

This work is supported by the United States Department of Defense & used resources of the Extreme Scale Systems Center at Oak Ridge National Laboratory.