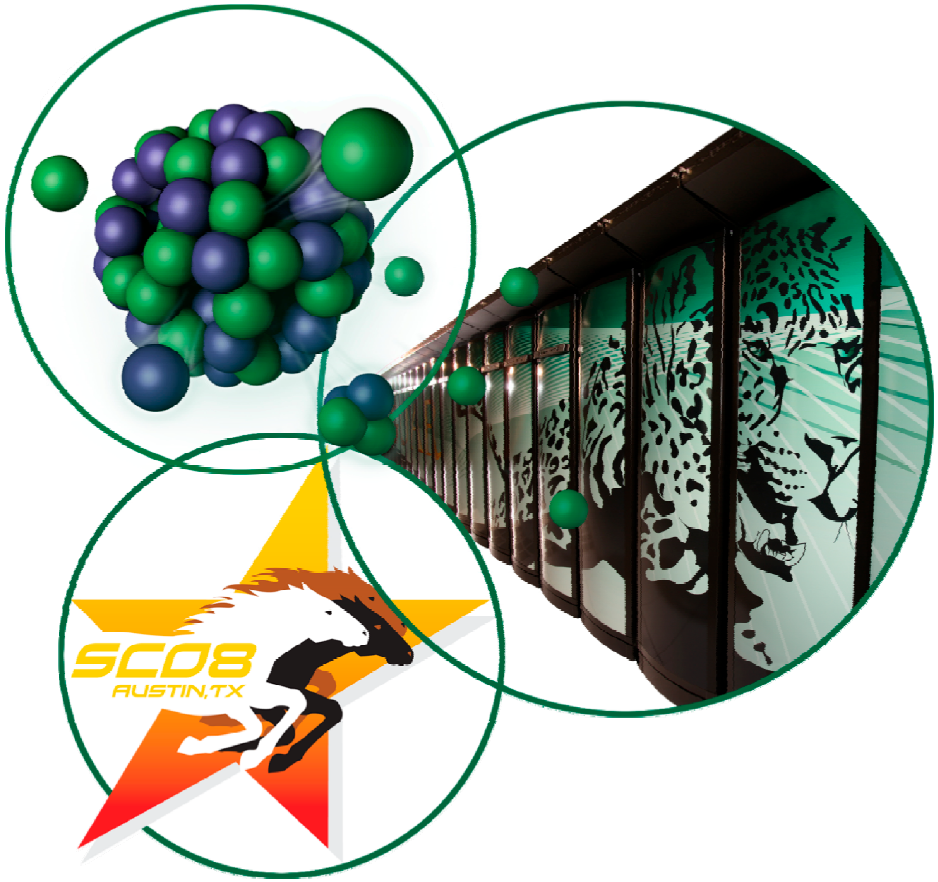
DOE UltraScience Net: High-Performance Experimental Network Research Testbed



Presented by

Nagi Rao

Complex Systems
Computer Science and Mathematics Division

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

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

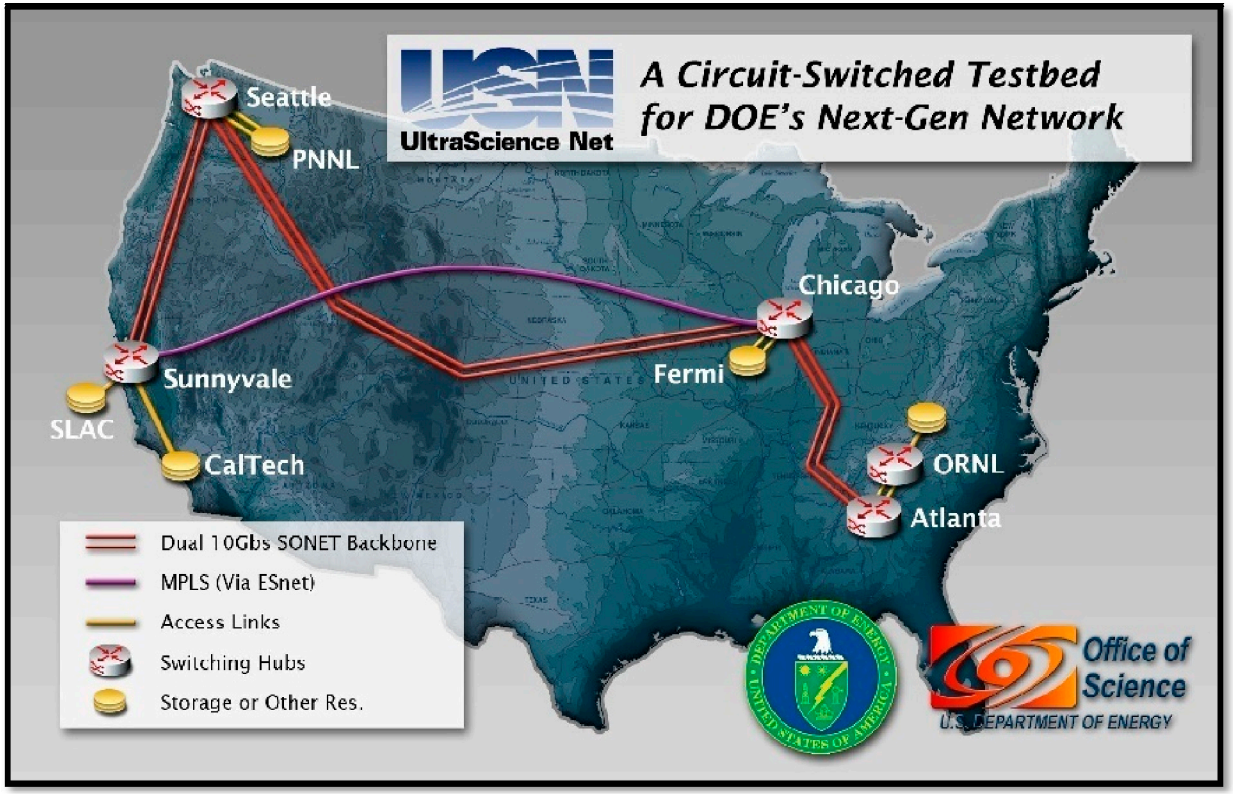
UltraScience Net – In a nutshell

Experimental network research testbed

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

Features

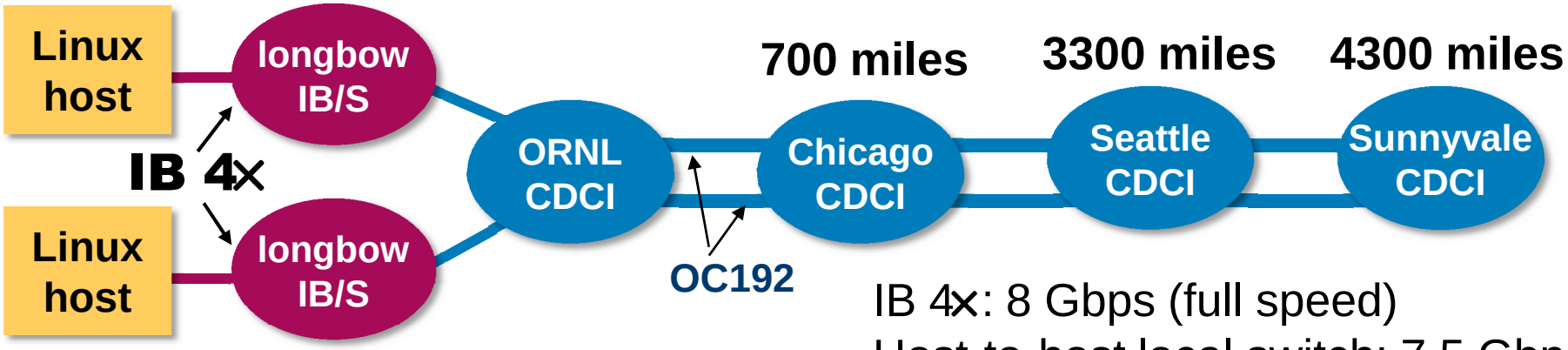
- End-to-end guaranteed bandwidth channels
- Dynamic, in-advance reservation and provisioning of fractional/full lambdas
- Secure control-plane for signaling
- Proximity to DOE sites: LCF, Fermi National Laboratory, National Energy Research Scientific Computing Center



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

Infiniband over SONET

Demonstrated that IB can scale to thousands of miles over OC192: 2% throughput reduction over 8000 miles



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

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

ORNL loop—0.2 mile: 7.48 Gb/s

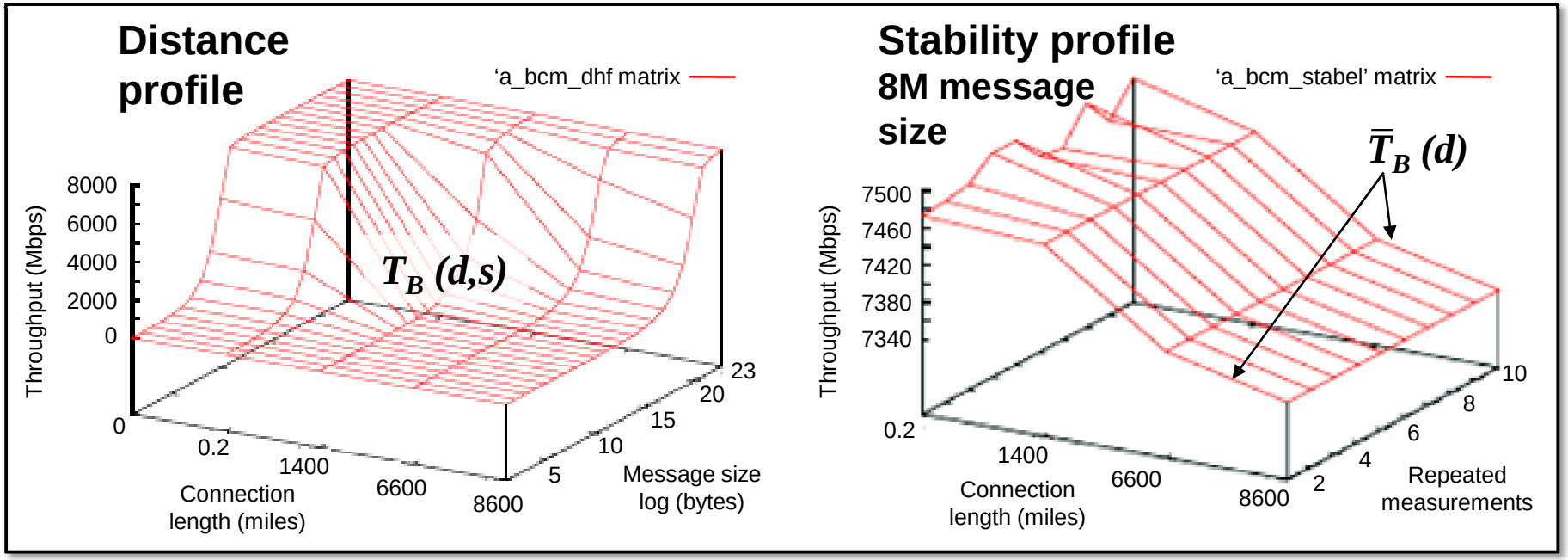
ORNL Chicago loop—1400 miles: 7.47 Gb/s

ORNL Chicago Seattle loop—6600 miles: 7.37 Gb/s

ORNL Chicago Seattle Sunnyvale loop—8600 miles: 7.34 Gb/s

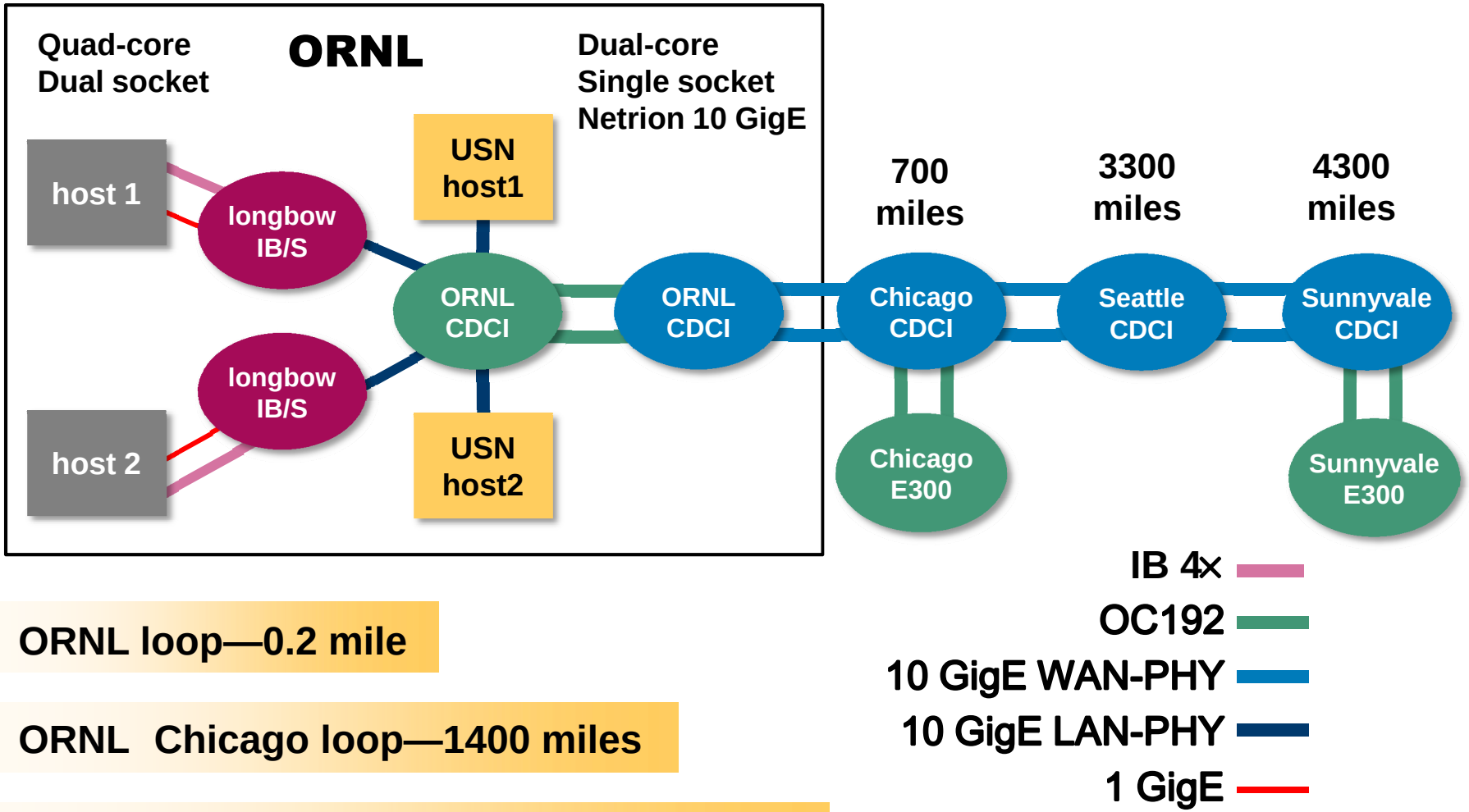
Distance and stability profiles of IB over SONET

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



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

IB over 10 GigE: cross-traffic



ORNL loop—0.2 mile

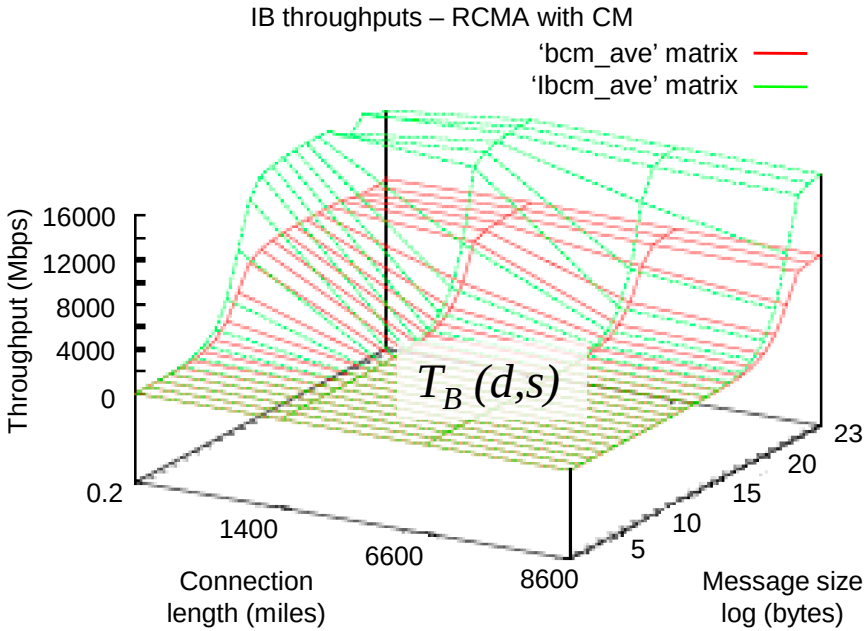
ORNL Chicago loop—1400 miles

ORNL Chicago Seattle loop—6600 miles

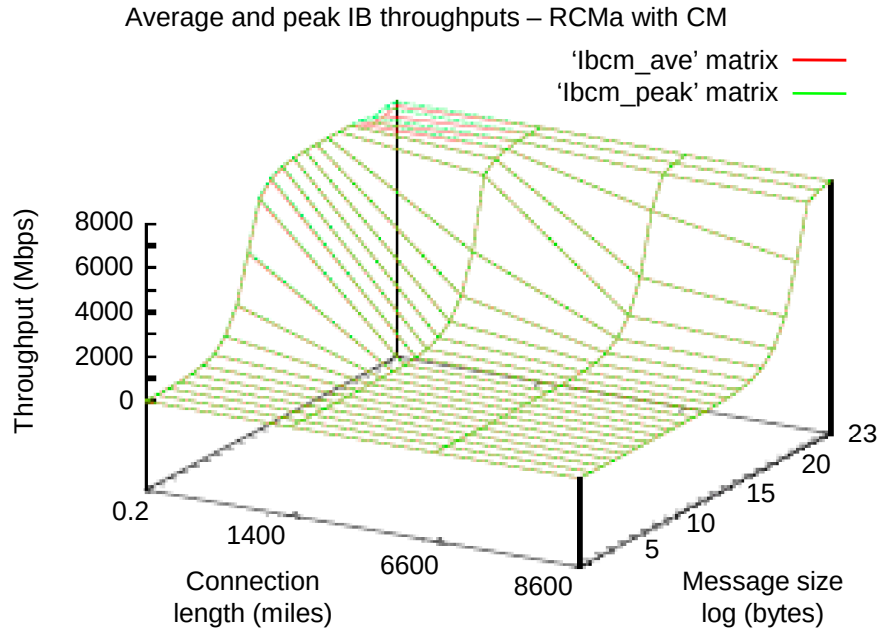
ORNL Chicago Seattle Sunnyvale loop—8600 miles

Performance profiles of IB over 10 GigE

Distance profile



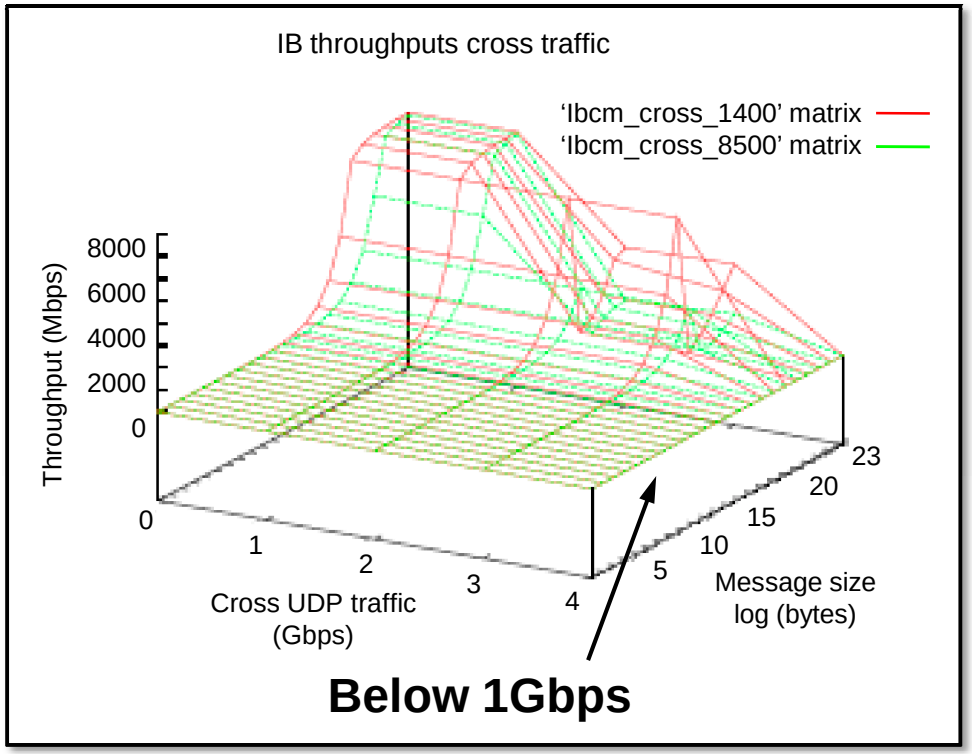
Peak distance profile
Average distance profile



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

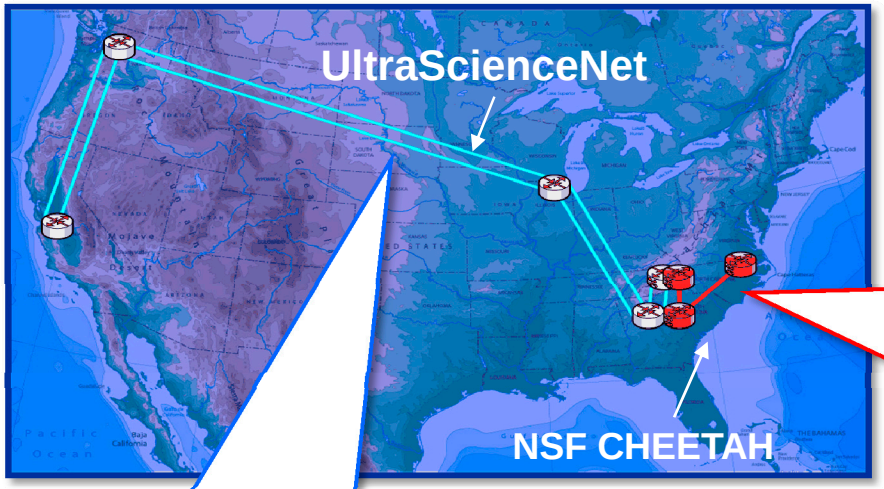


Average throughput for 8M			
Cross-traffic	1400 Miles	6600 Miles	8600 Miles
0 G	7.49	7.39	7.36
1 G	7.49	7.39	7.36
2 G	3.13	1.38	0.74
3 G	3.25	1.97	1.02
4 G	2.91	1.82	0.96

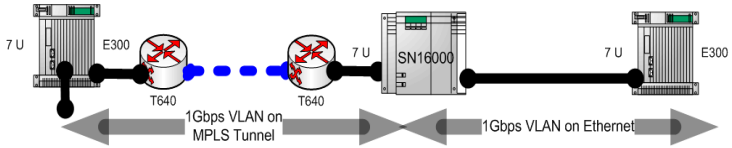
Competing traffic: UDP streams on WAN at 1,2,3,4 Gbps

- 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

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

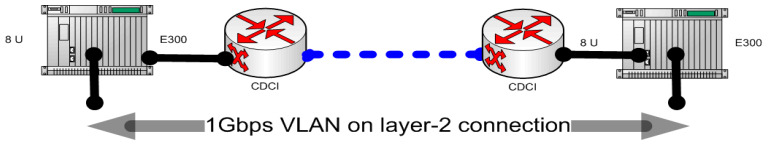


CHEETAH VLAN



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

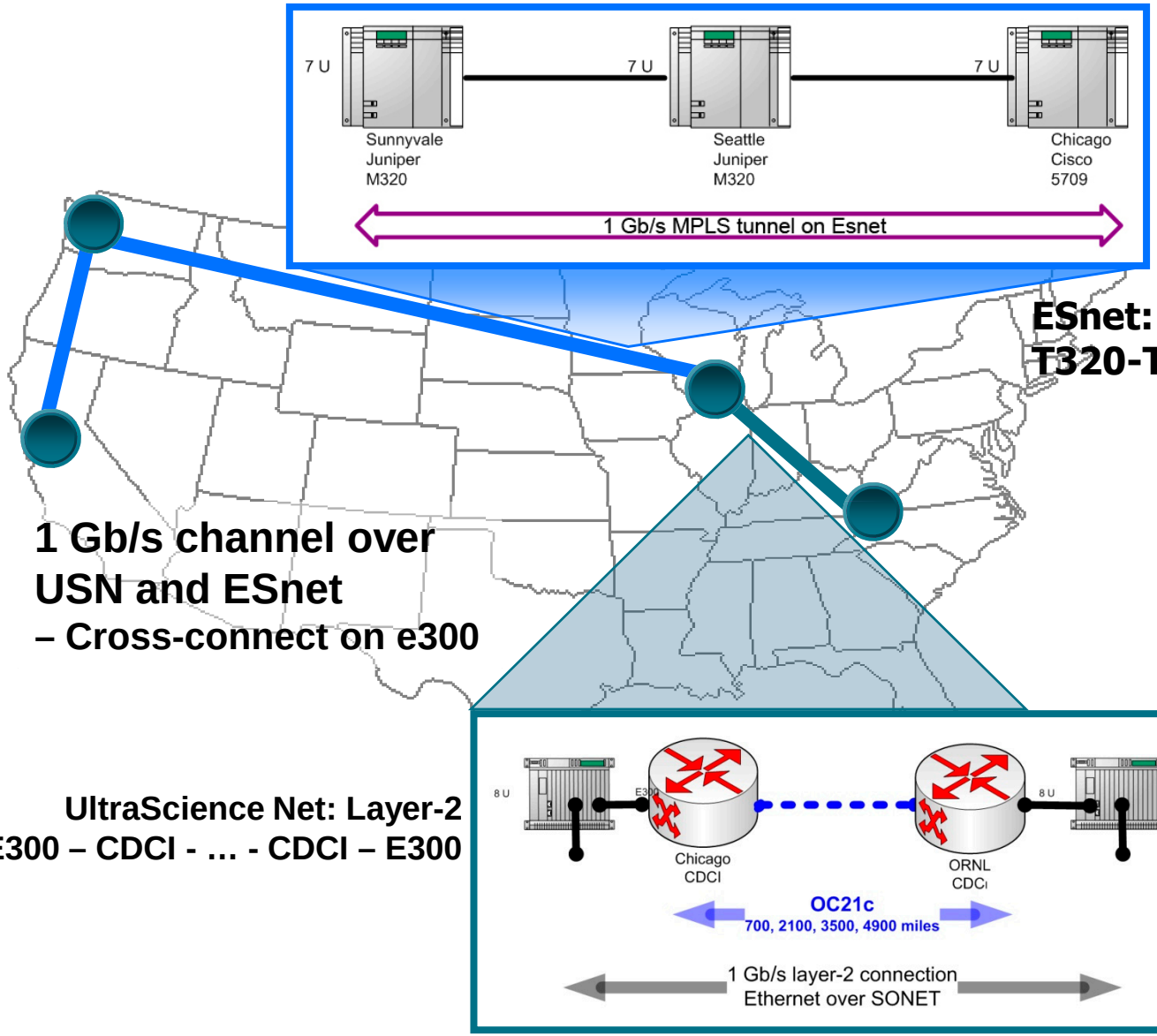
- **DOE UltraScience Net: Layer-2**
  - **VLAN: E300–CDCI - … -CDCI– E300**



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



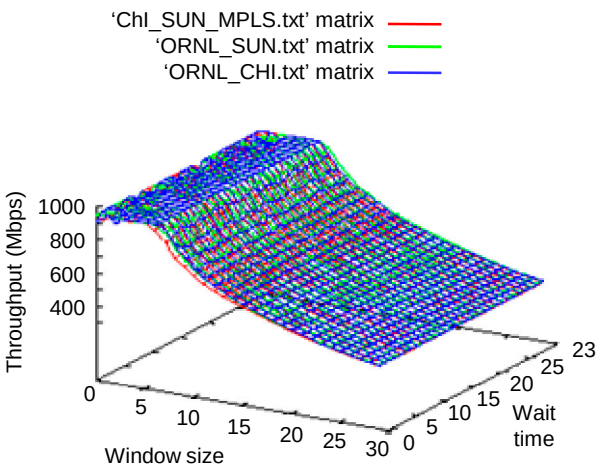
ESnet: layer-3 VLAN:
T320-T320 – Cisco 6509

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

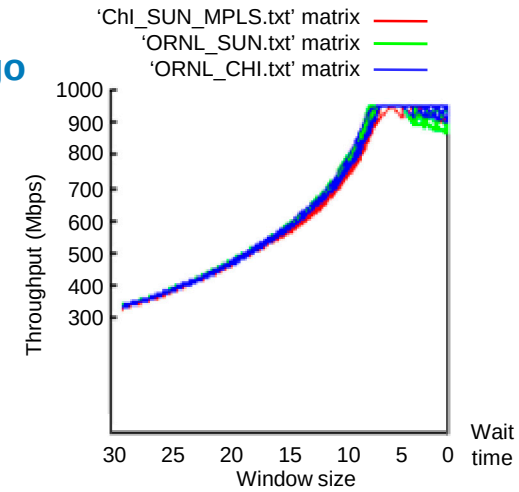
ORNL–Chicago-..-ORNL–Chicago

ESnet

Chicago–Sunnyvale

ESnet

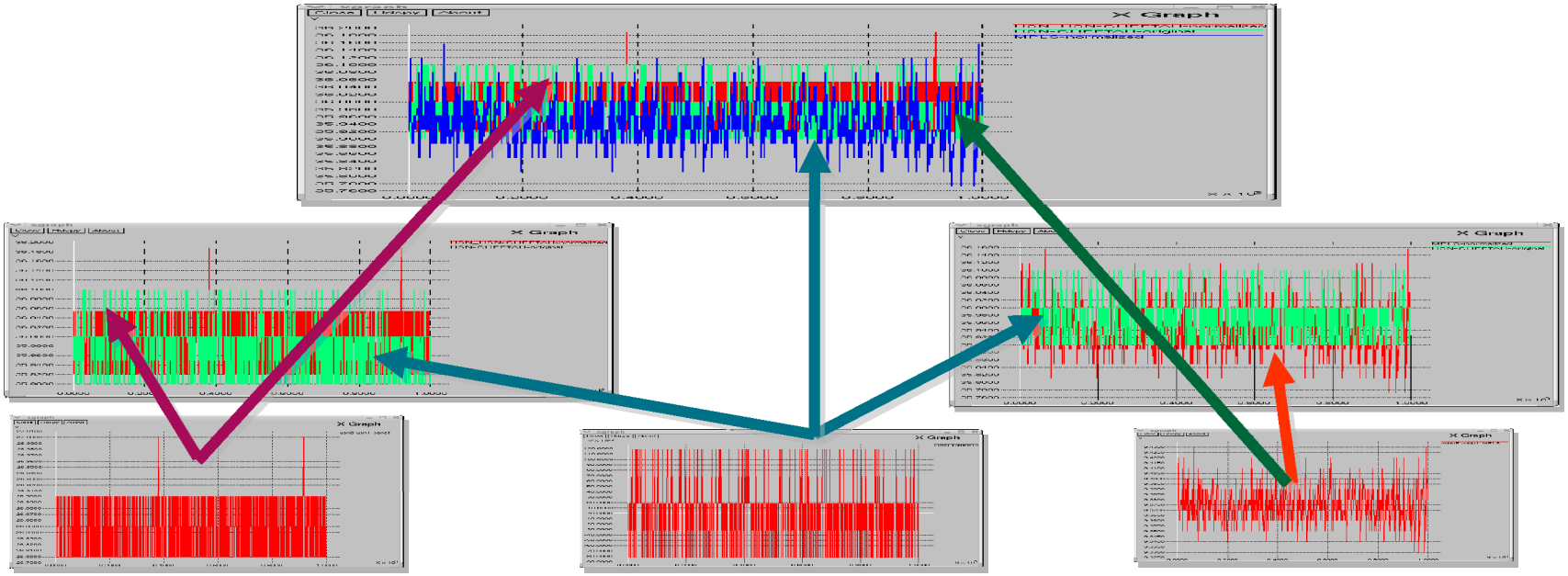
ORNL–Chicago–Sunnyvale



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
  - 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, and Infiniband over wide-area
- USN continues to play a **research role** in advanced networking capabilities
  - Networking technologies for LCFs
    - Connectivity to supercomputers
  - Testing 10 Gbps data transport
    - TCP variants and Inifiniband
  - 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