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 Department of Energy's Office of Science Office of Advanced Scientific Computing High-Performance Networking Program

The need

- Large-scale applications on supercomputers and experimental facilities require high-performance networking
  - Moving exascale 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: 150 Mbps to multiple 10/40/100Gbps
- Protocols are simpler for dedicated high throughput and control channels with limited/known traffic floes

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 and multi-core hosts

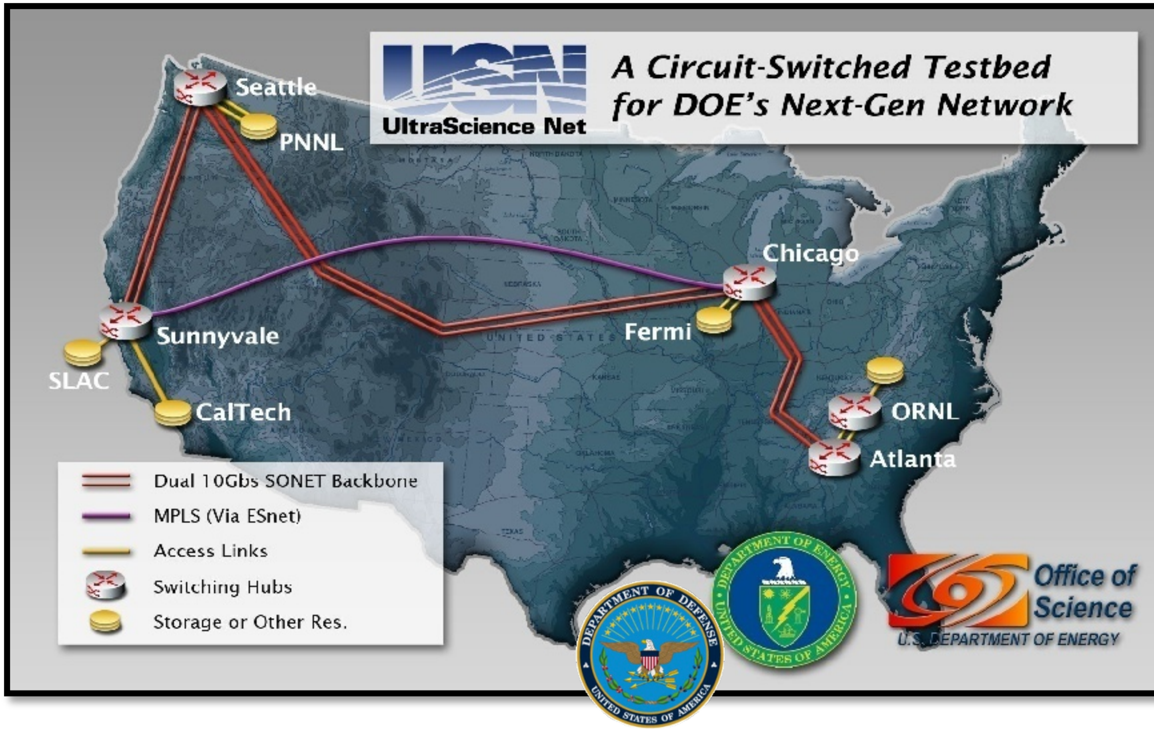
UltraScience Net – In a nutshell

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

ORNL-Atlanta connections upgraded to 40 Gbps

- 10Gbps Infrastructure emulated using ANUE devices

USN Contributions

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

2004

- **First advanced reservation and scheduling of dedicated connections**
  - 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**

2008

  - IB-RDMA throughputs: local 7.6 Gbps: 8600 miles: 7.2 Gbps: SC2008
- **10Gbps Crypto devices**
  - TCP performance improved: higher throughput with less #streams

2009

- **Cross-Calibration of emulations and testbed connections**

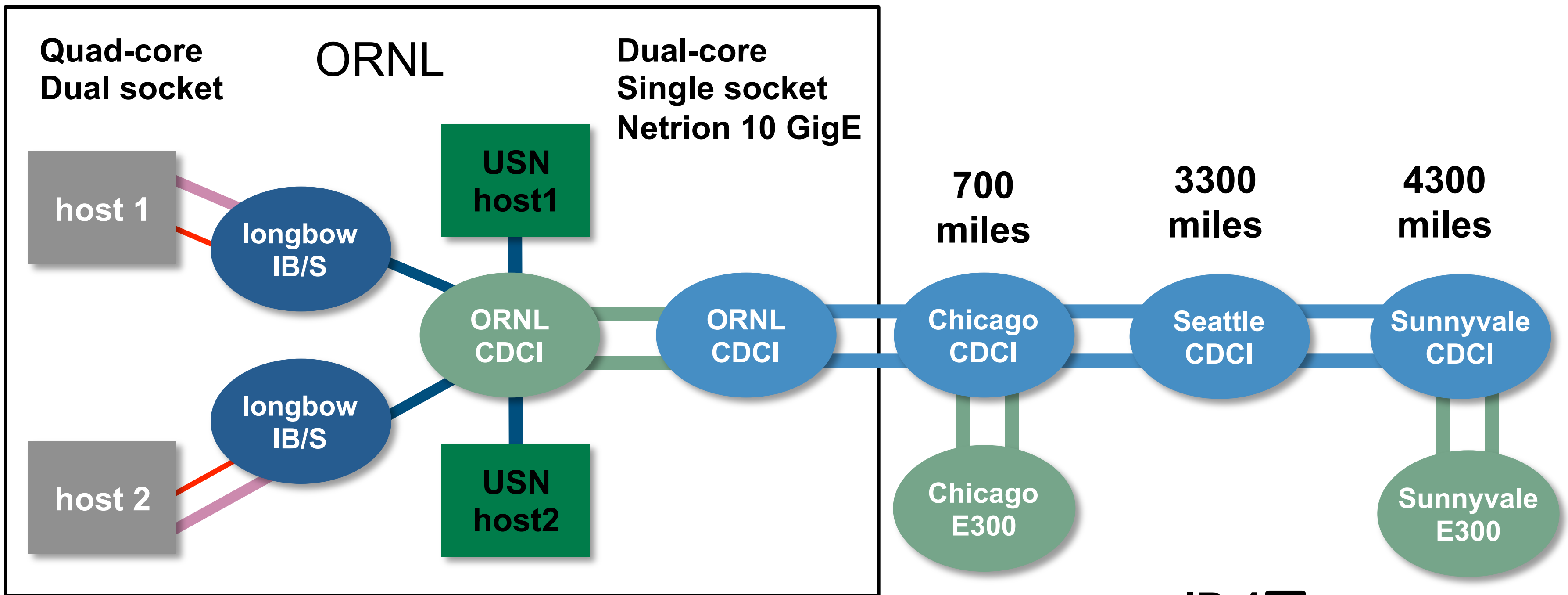
2010

  - Segmented regression to extend measurements to other modalities
- **40 Gbps upgrade to ORNL-Atlanta infrastructure**

2011

  - 39.5 Gbps throughputs between multi-core hosts

InfiniBand over 10 GigE: cross-traffic



IB 4X
OC192
10 GigE WAN-PHY
10 GigE LAN-PHY
1 GigE

ORNL loop—0.2 mile

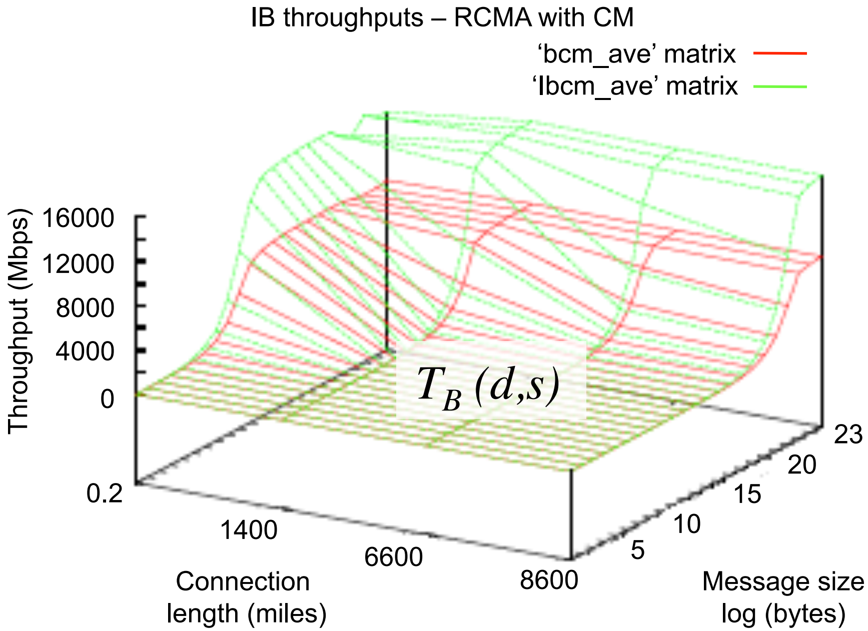
ORNL-Chicago loop—1400 miles

ORNL-Chicago-Seattle loop—6600 miles

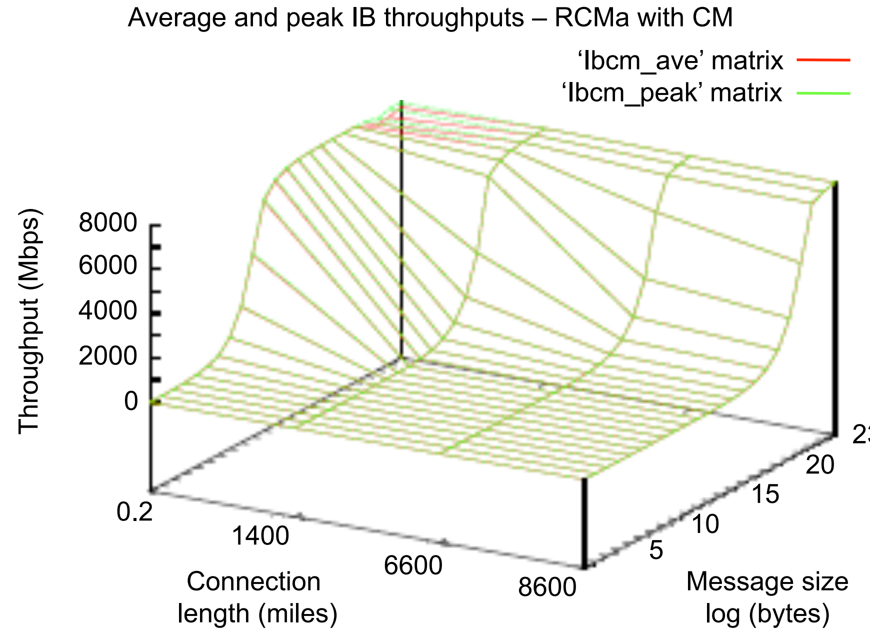
ORNL-Chicago-Seattle-Sunnyvale loop—8600 miles

Performance profiles of IB over 10 GigE

Distance profile



Peak distance profile
Average distance profile

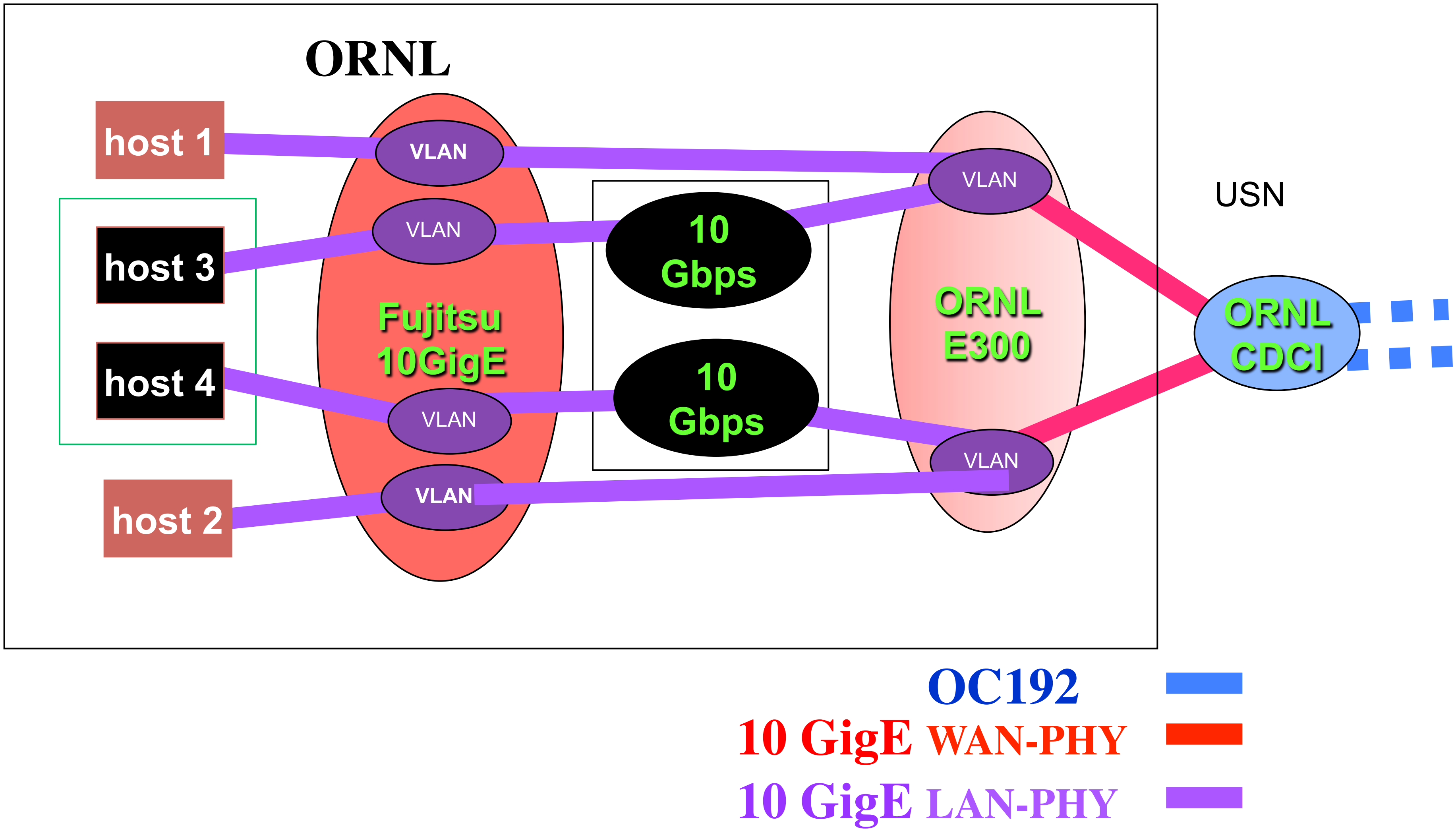


Results are almost the same as in SONET case

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

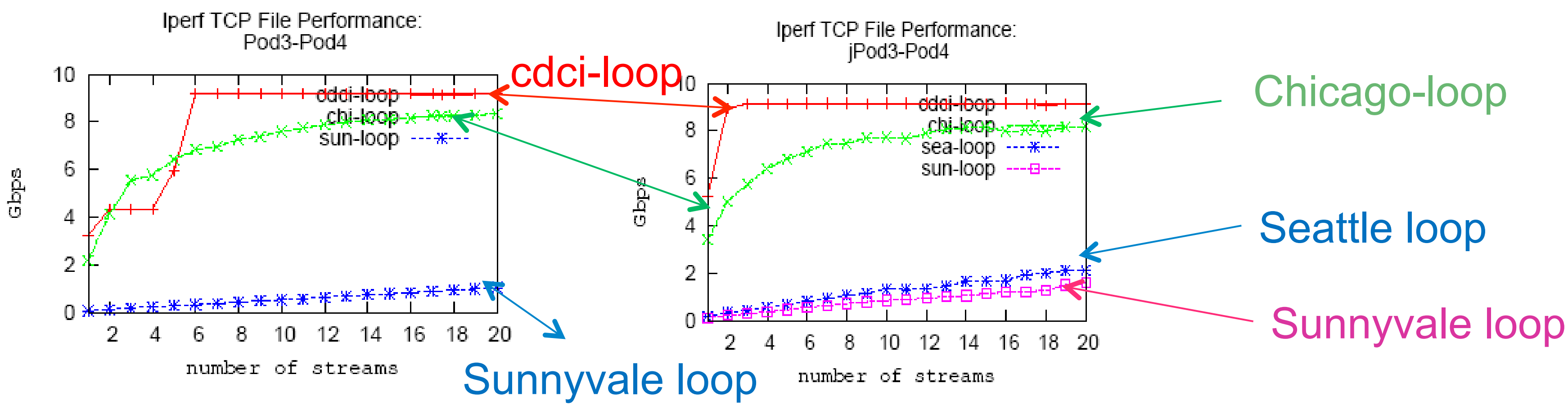
Testing of 10Gbps Encryption Devices:

host1-host2 Plain Connections
host3-host4 Encrypted Connections



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:

- **higher throughput is achieved with less number of streams**
- **higher throughput is achieved at longer distances**

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

Managed by UT-Battelle
for the U.S. Department of Energy

Analysis of iperf and XDD measurements - joint work with I/O Team

- **Estimated differential regressions:**

$T_M^{\varepsilon}(x)$ **:memory transfer throughput - emulated connection of length *x***

$T_D^{\varepsilon}(x)$ **: single disk transfer throughput - emulated connection**

$T_{DD}^{\varepsilon}(x)$ **: dual disk transfer throughput - emulated connection**

$f_M^{P\otimes\varepsilon}(x)$ **: differential regression for memory transfer throughput - between physical and emulated connections of length *x***

x (miles)	0.2	1400	6600	8600
$f_M^{P\otimes\varepsilon}(x)/T_M^{\varepsilon}(x)$	5.14%	5.80%	10.99%	14.98%
$T_M^{\varepsilon}(x)$ - Gbps	9.73	9.65	8.83	8.81
$f_D^{P\otimes\varepsilon}(x)/T_D^{\varepsilon}(x)$	28.03%	-2.82%	-2.26%	-0.91%
$T_D^{\varepsilon}(x)$ - MB/s	829.47	644.59	670.57	640.98
$f_{DD}^{P\otimes\varepsilon}(x)/T_{DD}^{\varepsilon}(x)$	7.37%	-2.19%	0.56%	-0.99%
$T_{DD}^{\varepsilon}(x)$ - MB/s	1233.98	899.91	715.89	684.06

Estimated memory transfer throughput: $\hat{Y}_{P;M} = Y_{\in;M} + f_M^{P\otimes\varepsilon}(x)$

measured memory transfer throughput: length *x*

Analysis of iperf and XDD measurements - joint work with I/O Team

- Measurements collected on USN connections and ANUE-emulated connections: Compared with measurements
  - Iperf memory transfers
  - XDD file transfers
  - Segmented Regression Method
    - Interpolation for 6600 mile connection
    - USN and ANUE measurements used to interpolate for
      - ANUE using ANUE
      - USN using USN
      - USN using ANUE + differential regression

memory transfer throughputs - Mbps			
	predicted	measured	percent error
anue	8981.09	9060.00	-0.87
usn	7920.56	7710.00	2.73
anue-usn	7999.47	7710.00	3.75
disk transfer throughputs - MB/s - single hosts			
	predicted	measured	percent error
anue	768.06	816.76	-5.96
usn	801.13	820.03	-2.31
anue-usn	849.82	820.03	3.63
disk transfer throughputs - MB/s - two hosts			
	predicted	measured	percent error
anue	925.02	890.65	3.86
usn	934.92	906.17	3.17
anue-usn	900.56	906.17	-0.62

- Summary:
  - For 10Gbps ANUE network emulators can closely match USN measurements – somewhat larger margins than IB measurements
  - Continue 10Gbps testing after USN de-commissioning

Analysis of iperf and XDD measurements - joint work with I/O Team

- Measurements collected ANUE-emulated USN connections used for interpolation/ extrapolation – compared with emulated connections

- Interpolation/extrapolation:
  - Apply differential regression

    to obtain USN predictions
    - Interpolation: 100 and 150ms
      - Not feasible on USN
        - In-between lengths
  - Extrapolation: 200 ms
    - Not feasible on USN
      - Too long

memory transfer throughputs - Mbps			
rtt	predicted	measured	percent error
100 ms	9235.70	9250.00	-0.15
150 ms	8912.75	8980.00	-0.75
200 ms	8525.25	8580.00	-0.64
disk transfer throughputs - MB/s - single hosts			
rtt	predicted	measured	percent error
100 ms	662.84	609.22	8.80
150 ms	656.51	598.83	9.63
200 ms	619.53	594.40	4.23
disk transfer throughputs - MB/s - two hosts			
rtt	predicted	measured	percent error
100ms	954.66	906.59	5.30
150ms	877.35	869.27	0.93
200ms	842.35	835.83	0.78

- Interpolation and extrapolation:
  - For 10Gbps ANUE network emulators can provide measurements for connection lengths not feasible (too long or in-between) on USN
  - Enable us to continue 10Gbps testing after 10Gbps USN de-commissioning