ESnet

ENERGY SCIENCES NETWORK

Experiences with 40G Hosts

Brian Tierney, Eric Pouyoul: ESnet

Nathan Hanford: UC Davis

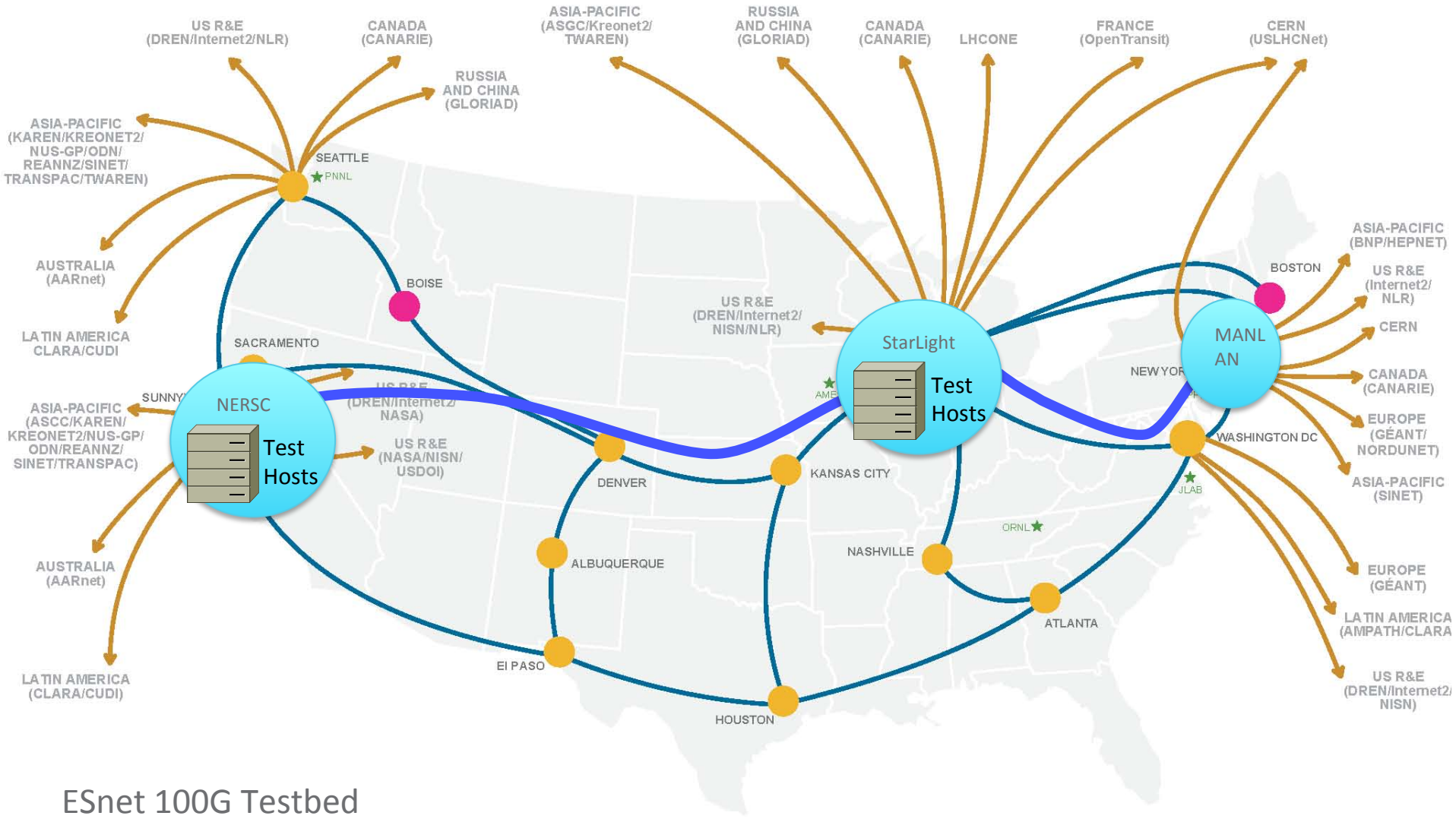
2014 Technology Exchange

October 29, 2014

U.S. DEPARTMENT OF ENERGY

Office of Science

BERKELEY LAB



ESnet 100G Testbed

ESnet
Energy Sciences Network

100G IP Hubs

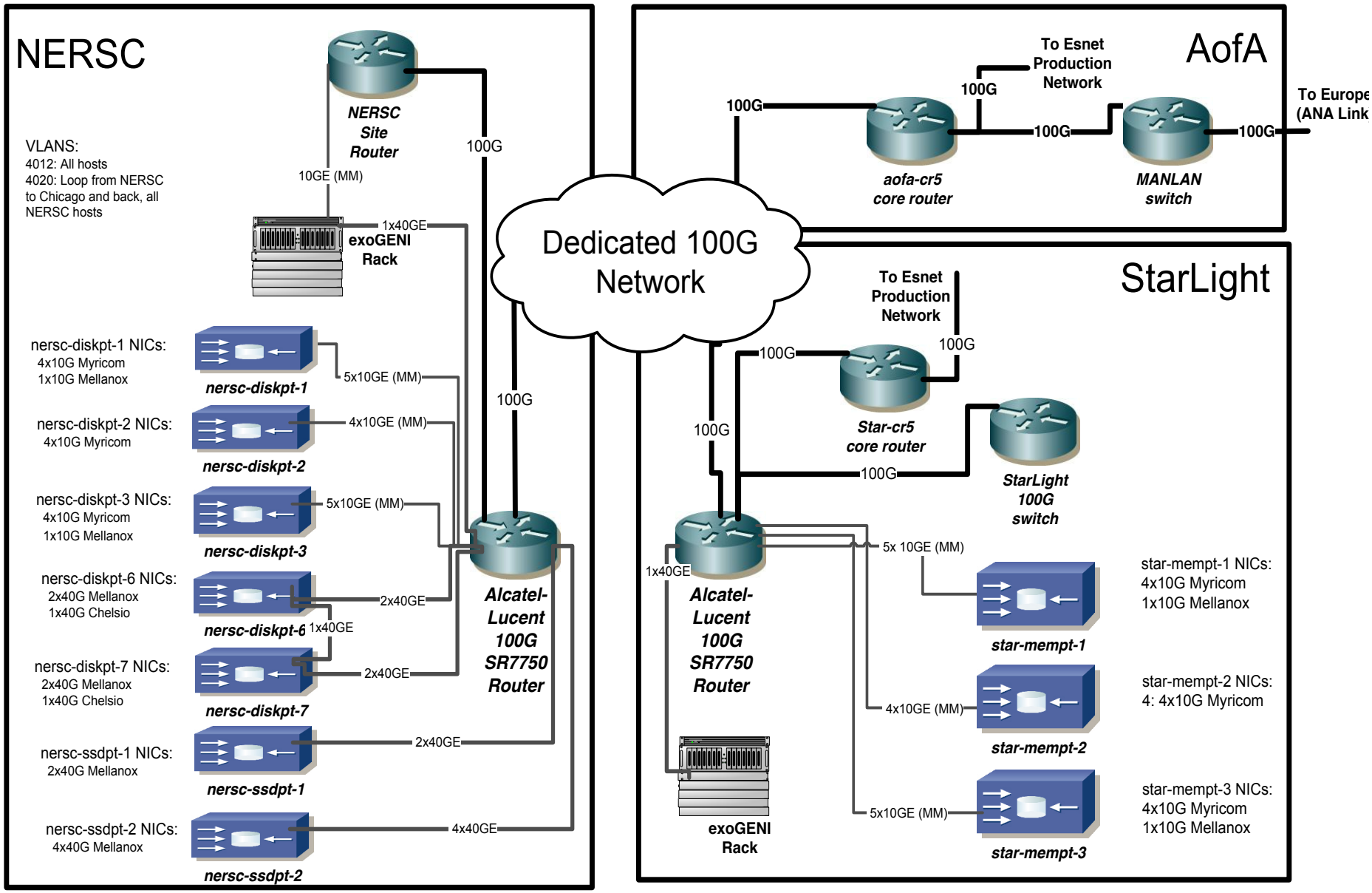
4x10G IP Hub

Major R&E
and International
peering connections

★ Office of Science National Labs

Ames	Ames Laboratory (Ames, IA)
ANL	Argonne National Laboratory (Argonne, IL)
BNL	Brookhaven National Laboratory (Upton, NY)
FNAL	Fermi National Accelerator Laboratory (Batavia, IL)
JLAB	Thomas Jefferson National Accelerator Facility (Newport News, VA)
LBNL	Lawrence Berkeley National Laboratory (Berkeley, CA)
ORNL	Oak Ridge National Laboratory (Oak Ridge, TN)
PNNL	Pacific Northwest National Laboratory (Richland, WA)
PPPL	Princeton Plasma Physics Laboratory (Princeton, NJ)
SLAC	Stanford Linear Accelerator Center (Menlo Park, CA)

ESnet 100G Testbed



Data Transfer Nodes (DTNs) Used in these tests

- Configuration:
  - Motherboard: SuperMicro X9DR3-F (PCIe Gen3)
  - Processors: 2 x Intel(R) Xeon(R) CPU E5-2667 0 @ 2.90GHz 6 Cores (Total 12 Cores)
  - Memory: 64GB DDR3-1600MHz RAM ECC/REG - (4x16GB)
  - HDD: (16) SAS Seagate Constellation 350GB
  - RAID controllers: 2 x 3ware Inc 9750 SAS2/SATA-II RAID PCIe (rev 05)
  - Ethernet: 40G RoCE Dual Port Mellanox MCX314A-BCBT, Chelsio 40GE Dual Port T580-LP-CR.

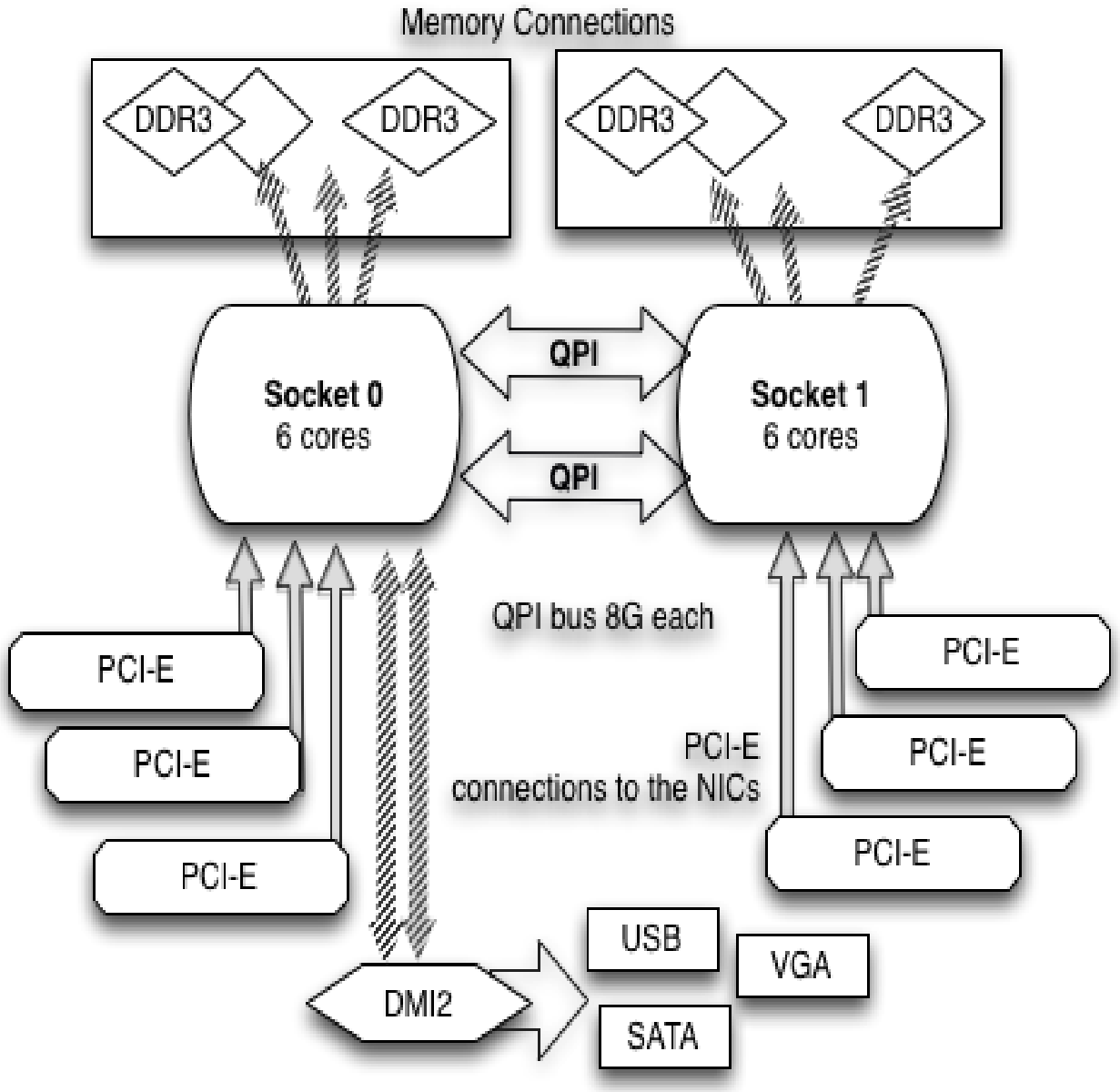
ESnet

40G Lessons Learned

- Single flow:
  - TCP: 39 Gbps (with the right core)
  - UDP: 22 Gbps, CPU limited
- Multiple Flows:
  - Easily fill 40G NIC
- Tuning for 40G is not just 4x Tuning for 10G
  - Some of the conventional wisdom for 10G Networking is not true at 40Gbps
  - e.g.: Parallel streams more likely to hurt than help
- "Sandy Bridge" Architectures require extra tuning
- Details at http://fasterdata.es.net/science-dmz/DTN/tuning/

ESnet

Intel Sandy/Ivy Bridge



ESnet

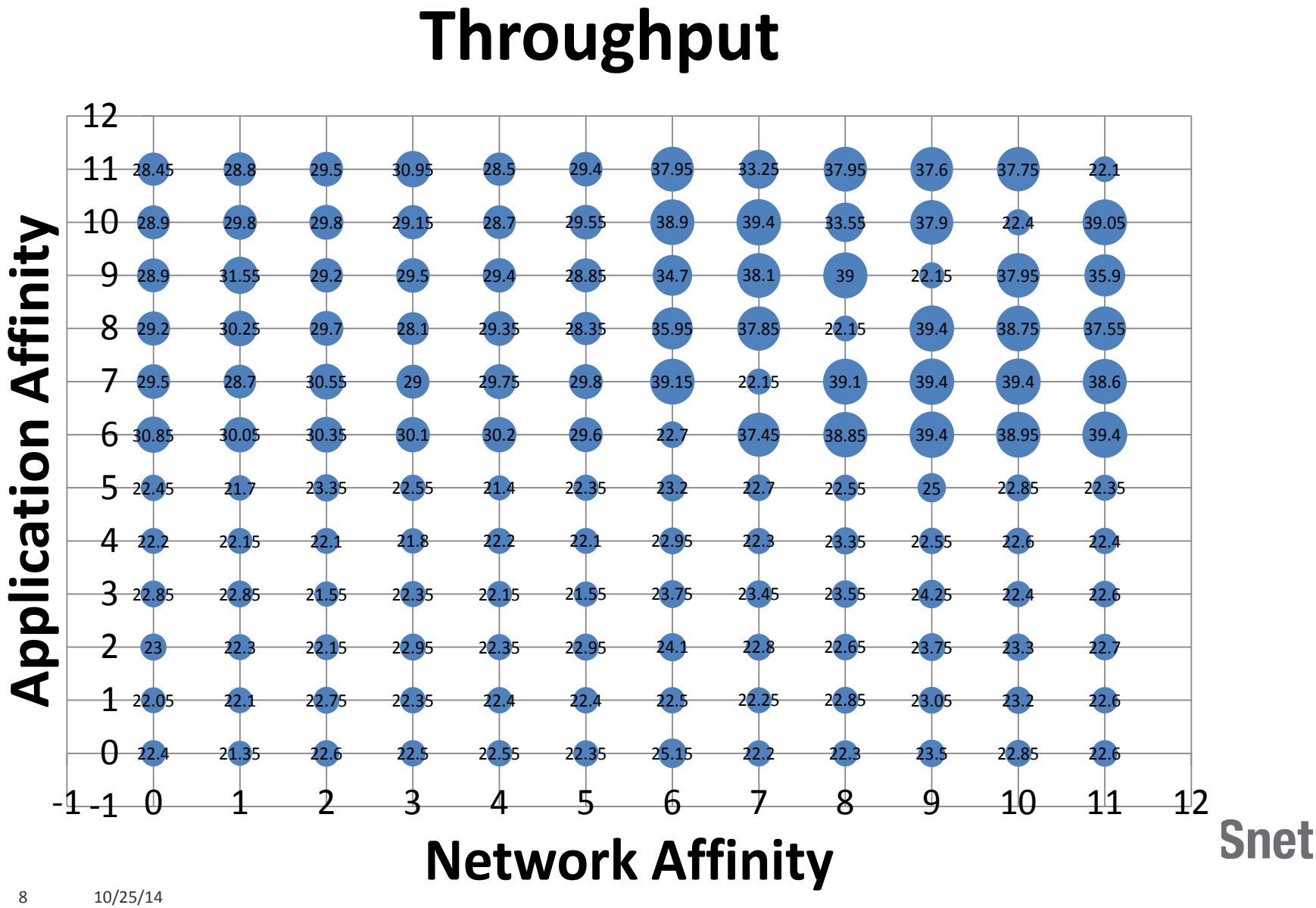
Core selection matters! (iperf3 –A N,N option)

Protocol	40G NIC	"good" core	"bad" core
TCP	Mellanox	37 Gbps	26 Gbps
	Chelsio	39.5 Gbps	5 Gbps
UDP	Mellanox	21 Gbps	16 Gbps
	Chelsio	22 Gbps	16 Gbps

IRQ binding bootscripts
Mellanox: /usr/sbin/set_irq_affinity_bynode.sh 1 ethN
Chelsio: /sbin/t4_perftune.sh

ESnet

Plot of Application core vs NIC interrupt core (slide from Nate Hanford, UC Davis)



Snet

Sample results: TCP Single vs Parallel Streams: 40G to 40G

- **1 stream:** iperf3 -c 192.168.102.9

```
[ ID] Interval           Transfer     Bandwidth       Retransmits
[  4]   0.00-1.00   sec  3.19 GBytes  27.4 Gbits/sec    0
[  4]   1.00-2.00   sec  3.35 GBytes  28.8 Gbits/sec    0
[  4]   2.00-3.00   sec  3.35 GBytes  28.8 Gbits/sec    0
[  4]   3.00-4.00   sec  3.35 GBytes  28.8 Gbits/sec    0
[  4]   4.00-5.00   sec  3.35 GBytes  28.8 Gbits/sec    0
```

- **2 streams:** iperf3 -c 192.168.102.9 -P2

```
[ ID] Interval           Transfer     Bandwidth       Retransmits
[  4]   0.00-1.00   sec  1.37 GBytes  11.8 Gbits/sec    7
[  6]   0.00-1.00   sec  1.38 GBytes  11.8 Gbits/sec   11
[SUM]   0.00-1.00   sec  2.75 GBytes  23.6 Gbits/sec   18
……
- - - - - - - - - - - - - - - - - - - - - - - - -
[  4]   9.00-10.00  sec  1.43 GBytes  12.3 Gbits/sec    4
[  6]   9.00-10.00  sec  1.43 GBytes  12.3 Gbits/sec    6
[SUM]   9.00-10.00  sec  2.86 GBytes  24.6 Gbits/sec   10
- - - - - - - - - - - - - - - - - - - - - - - - -
[ ID] Interval           Transfer     Bandwidth       Retransmits
[  4]   0.00-10.00  sec  13.8 GBytes  11.9 Gbits/sec   78
[  6]   0.00-10.00  sec  13.8 GBytes  11.9 Gbits/sec   95
[SUM]   0.00-10.00  sec  27.6 GBytes  23.7 Gbits/sec  173
```


ESnet

40G Sender to 10G receiver; Parallel streams decrease throughput even more

- **Single Stream:**

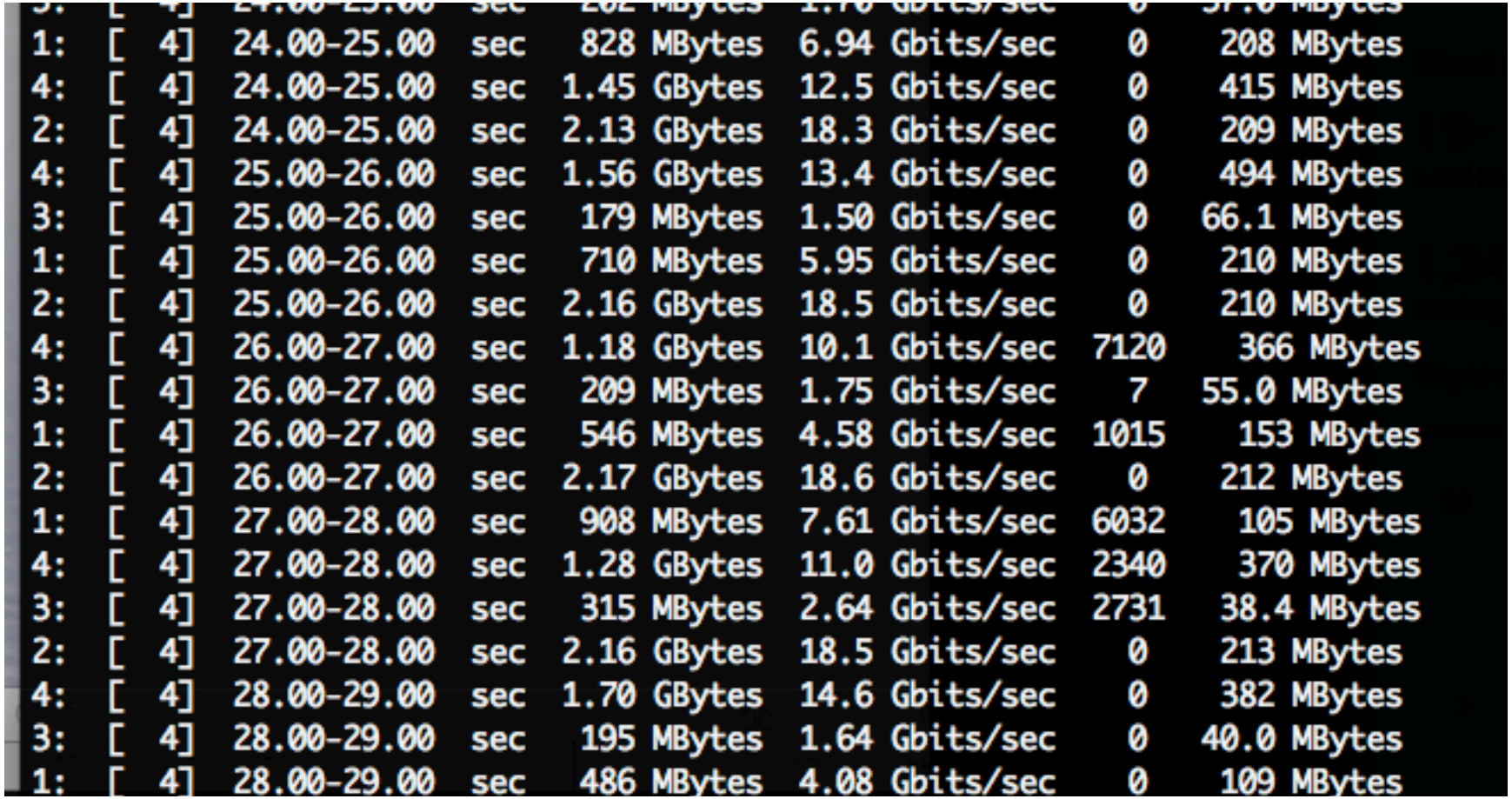
```
iperf3 -c 10.12.1.128 -w 128M
[ ID] Interval           Transfer     Bandwidth       Retr
[  4]   0.00-1.05   sec   238 MBytes  1.90 Gbits/sec    0
[  4]   1.05-2.05   sec   404 MBytes  3.38 Gbits/sec    0
[  4]   2.05-3.05   sec  1.06 GBytes  9.10 Gbits/sec  404
[  4]   3.05-4.05   sec  1.08 GBytes  9.30 Gbits/sec    0
[  4]   4.05-5.05   sec  1.14 GBytes  9.78 Gbits/sec    0
[  4]   5.05-6.05   sec  1.15 GBytes  9.89 Gbits/sec    0
```

- **2 Parallel Streams**: iperf3 -c 10.12.1.128 -P2

```
[ ID] Interval           Transfer     Bandwidth       Retr
[  4]   0.00-10.05  sec  2.69 GBytes  2.30 Gbits/sec  311
[  6]   0.00-10.05  sec  2.69 GBytes  2.30 Gbits/sec  613
[SUM]   0.00-10.05  sec  5.38 GBytes  4.60 Gbits/sec  924
```

ESnet

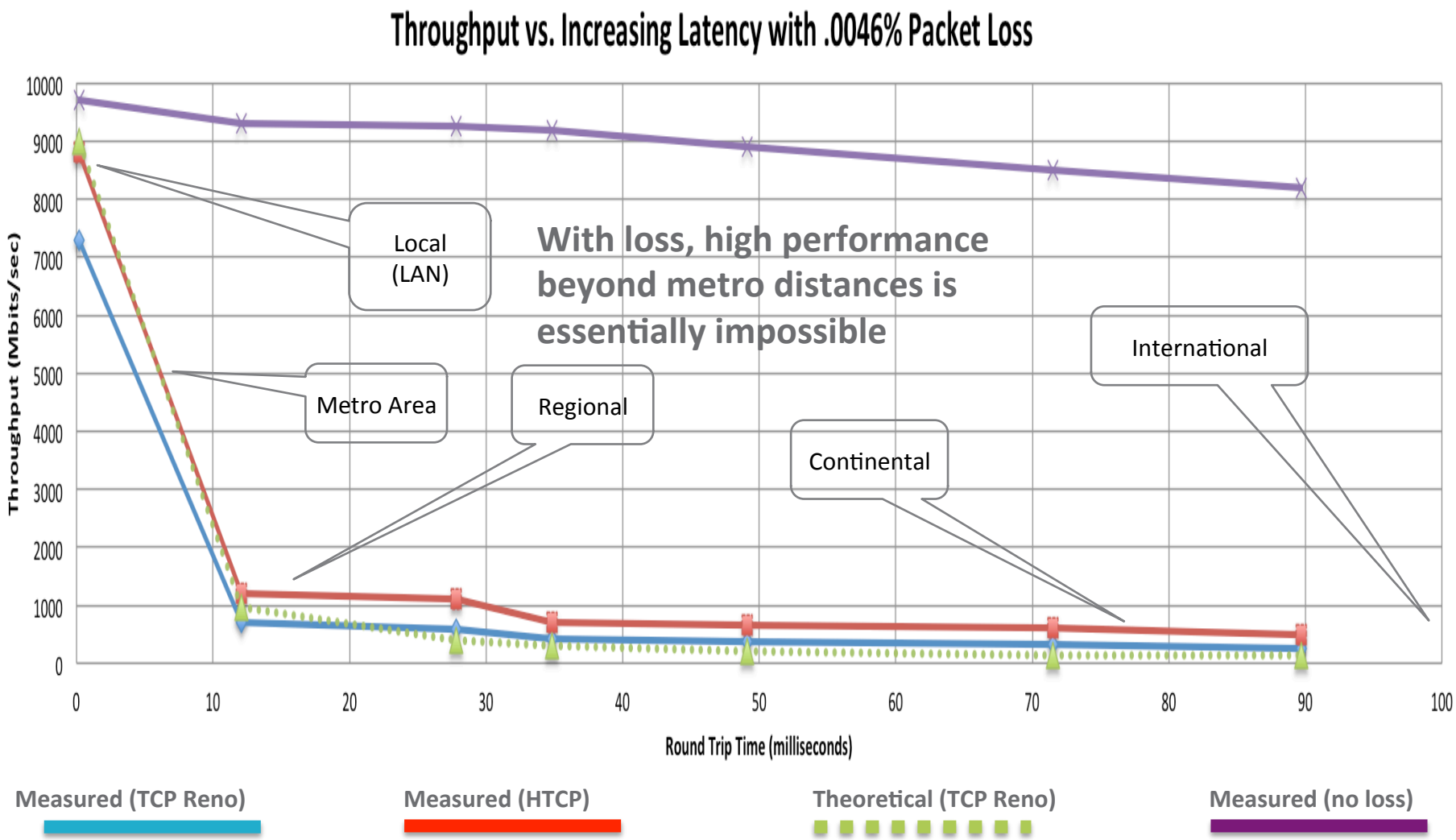
Parallel streams on different cores still not stable: wide variance in per stream throughput



1:	[4]	24.00-25.00	sec	828 MBytes	6.94 Gbits/sec	0	208 MBytes
4:	[4]	24.00-25.00	sec	1.45 GBytes	12.5 Gbits/sec	0	415 MBytes
2:	[4]	24.00-25.00	sec	2.13 GBytes	18.3 Gbits/sec	0	209 MBytes
4:	[4]	25.00-26.00	sec	1.56 GBytes	13.4 Gbits/sec	0	494 MBytes
3:	[4]	25.00-26.00	sec	179 MBytes	1.50 Gbits/sec	0	66.1 MBytes
1:	[4]	25.00-26.00	sec	710 MBytes	5.95 Gbits/sec	0	210 MBytes
2:	[4]	25.00-26.00	sec	2.16 GBytes	18.5 Gbits/sec	0	210 MBytes
4:	[4]	26.00-27.00	sec	1.18 GBytes	10.1 Gbits/sec	7120	366 MBytes
3:	[4]	26.00-27.00	sec	209 MBytes	1.75 Gbits/sec	7	55.0 MBytes
1:	[4]	26.00-27.00	sec	546 MBytes	4.58 Gbits/sec	1015	153 MBytes
2:	[4]	26.00-27.00	sec	2.17 GBytes	18.6 Gbits/sec	0	212 MBytes
1:	[4]	27.00-28.00	sec	908 MBytes	7.61 Gbits/sec	6032	105 MBytes
4:	[4]	27.00-28.00	sec	1.28 GBytes	11.0 Gbits/sec	2340	370 MBytes
3:	[4]	27.00-28.00	sec	315 MBytes	2.64 Gbits/sec	2731	38.4 MBytes
2:	[4]	27.00-28.00	sec	2.16 GBytes	18.5 Gbits/sec	0	213 MBytes
4:	[4]	28.00-29.00	sec	1.70 GBytes	14.6 Gbits/sec	0	382 MBytes
3:	[4]	28.00-29.00	sec	195 MBytes	1.64 Gbits/sec	0	40.0 MBytes
1:	[4]	28.00-29.00	sec	486 MBytes	4.08 Gbits/sec	0	109 MBytes

ESnet

A small amount of packet loss makes a huge difference in TCP performance



Conclusions

- Conventional wisdom for tuning 10G does not necessarily work at 40G
- If you want to maximize flow throughput, get the fastest cores you can afford
  - Especially for the receive host
- Benchmarking 40G system requires understanding of these issues.
  - And these issues likely to be even worse with 100G.

ESnet

More Information

http://www.es.net/testbed/

email: BLTierney@es.net

"Hi, I'd like to hear a TCP joke."
"Hello, would you like to hear a TCP joke?"
"Yes, I'd like to hear a TCP joke."
"OK, I'll tell you a TCP joke."
"Ok, I will hear a TCP joke."
"Are you ready to hear a TCP joke?"
"Yes, I am ready to hear a TCP joke."
"Ok, I am about to send the TCP joke. It will last 10 seconds, it has two characters, it does not have a setting, it ends with a punchline."
"Ok, I am ready to get your TCP joke that will last 10 seconds, has two characters, does not have an explicit setting, and ends with a punchline."
"I'm sorry, your connection has timed out.
...Hello, would you like to hear a TCP joke?"