



# Evaluation of 40 Gigabit Ethernet technology for data servers

**Azher Mughal, Artur Barczyk**

**Caltech / USLHCNet**

*CHEP-2012, New York*

*http://supercomputing.caltech.edu*

# Agenda

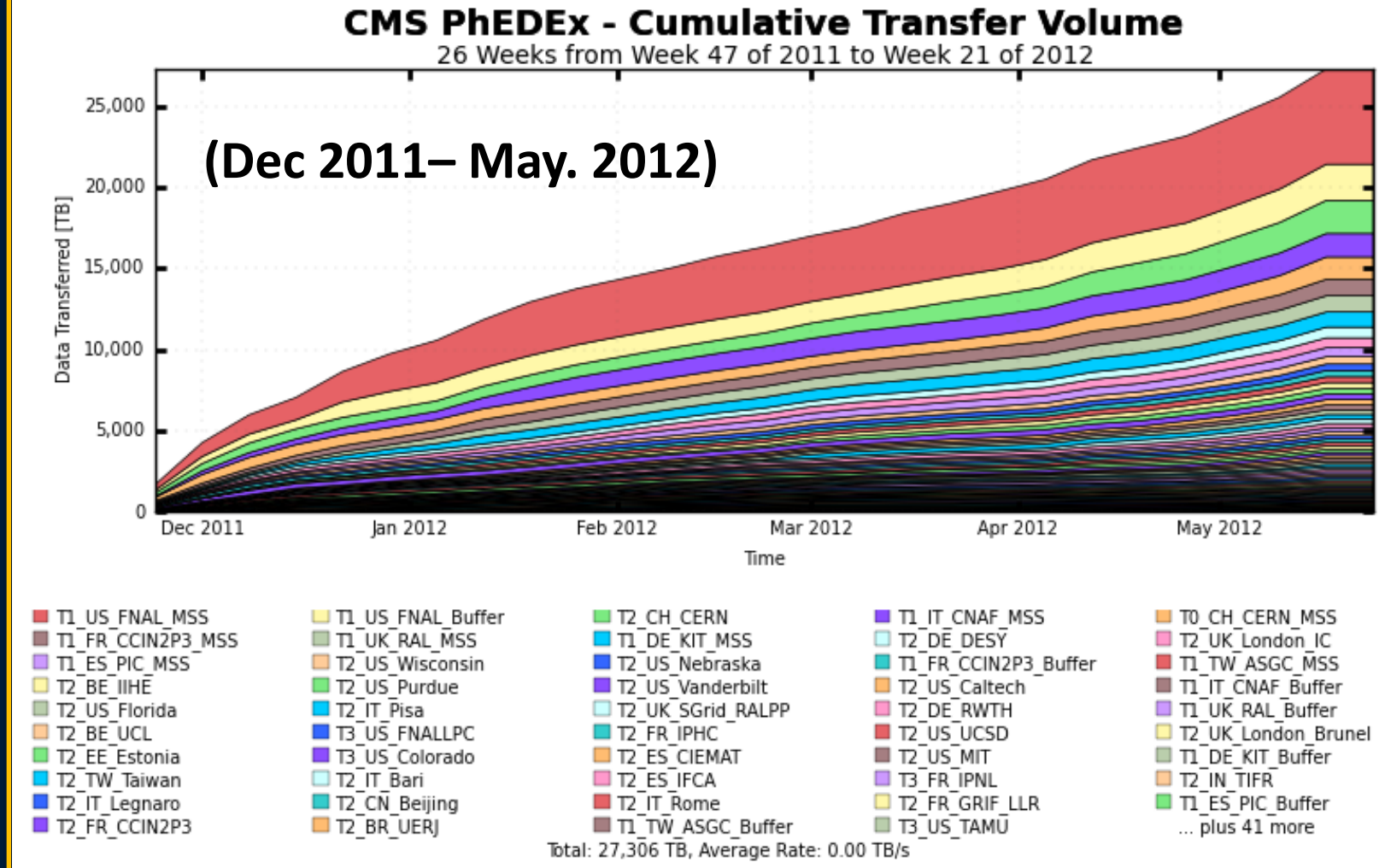
- Motivation behind 40GE in Data Servers
- Network & Servers Design
- Designing a Fast Data Transfer Kit
- PCIe Gen3 Server Performance
- 40G Network testing
- WAN Transfers
- Disk to Disk Transfers
- Questions ?

# The Motivation

- **The LHC experiments, with their distributed Computing Models and global program of LHC physics, have a renewed focus on networks, and correspondingly a renewed emphasis on "capacity" and "reliability" of the networks**
- **Networks have seen an exponential growth in capacity**
  - 10X in usage every 47 months in ESnet over 18 years
  - About 6M times capacity growth over 25 years across the Atlantic (LEP3Net in 1985 to USLHCNet as of today)
  - LHC experiments (CMS / ATLAS) are generating large data sets which need to be efficiently transferred to end sites, anywhere in the world
- **A sustained ability to use ever-larger continental and transoceanic networks effectively: high throughput transfers**
- **HEP as a driver of R&E and mission-oriented networks**
- **Testing latest innovations both in terms of software and hardware**

**Harvey Newman, Caltech**

# 27.3 PetaBytes Transferred Over 6 Months
# average transfer rate = 14 Gbps



# Target Features in 40GE Server

- Has at least one 40GE port connecting to LAN/WAN
- Able to read from Disks at near 40Gbps (4.9 GB/sec)
- Able to write on Disks at near 40Gbps (4.9 GB/sec)
- Line rate Network throughput with minimum CPU utilization (therefore more headroom for applications)

# History of 40GE NICs

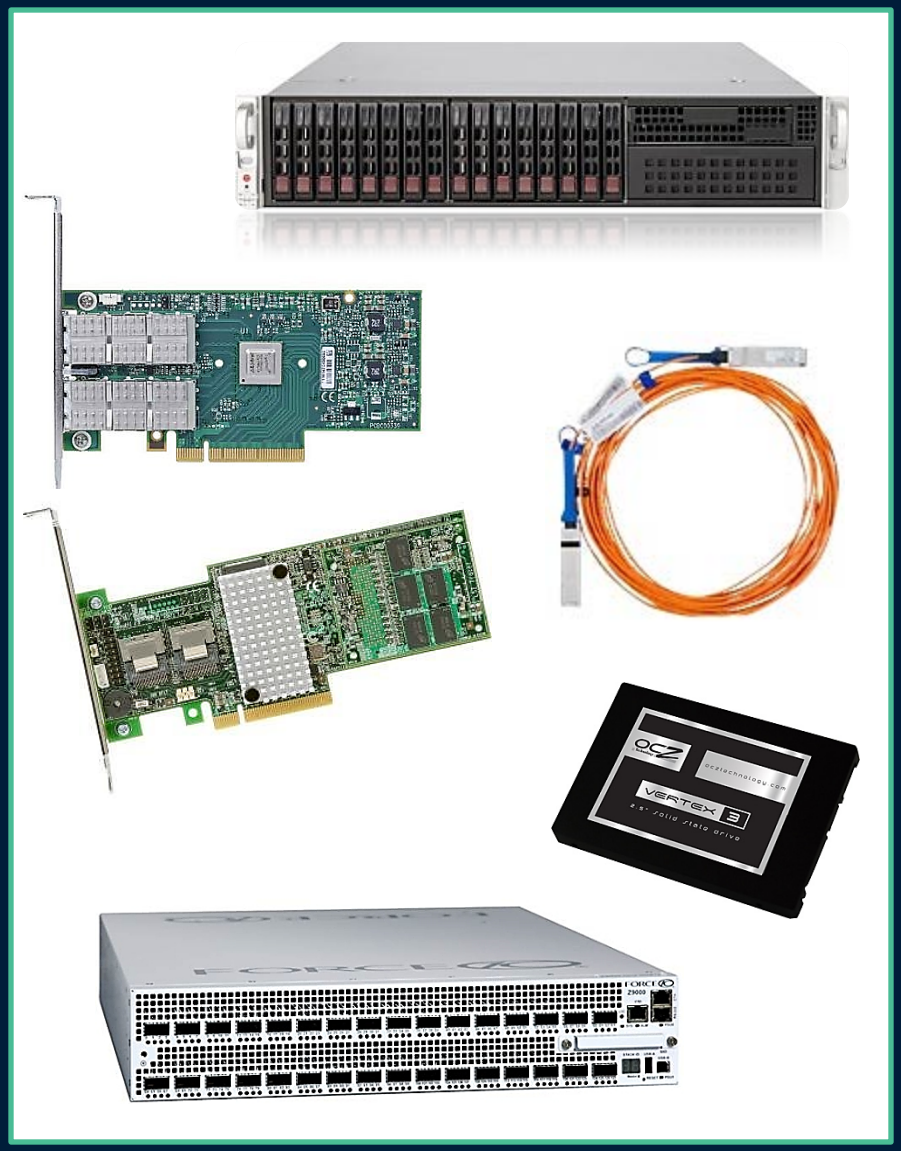
Mellanox is the only vendor offering 40GE NICs (since 2010).
Mainly Three variants:

- 40GE Gen2 NIC
  - ConnectX-2
  - PCIe Gen 2.0 x8 interface, 32Gbps FD
  - 8b/12b line encoding, 20% overhead, 25.6Gbps FD
- 40GE Gen 3 NIC
  - PCIe Gen 3.0 x8 interface, 1GB per lane or 64Gbps FD
  - More efficient 64/66 encoding
- 40/56Gbps Ethernet/VPI NIC
  - Faster Clock rate, can go upto 56Gbps using IB FDR mode

# 40GE Server Design Kit

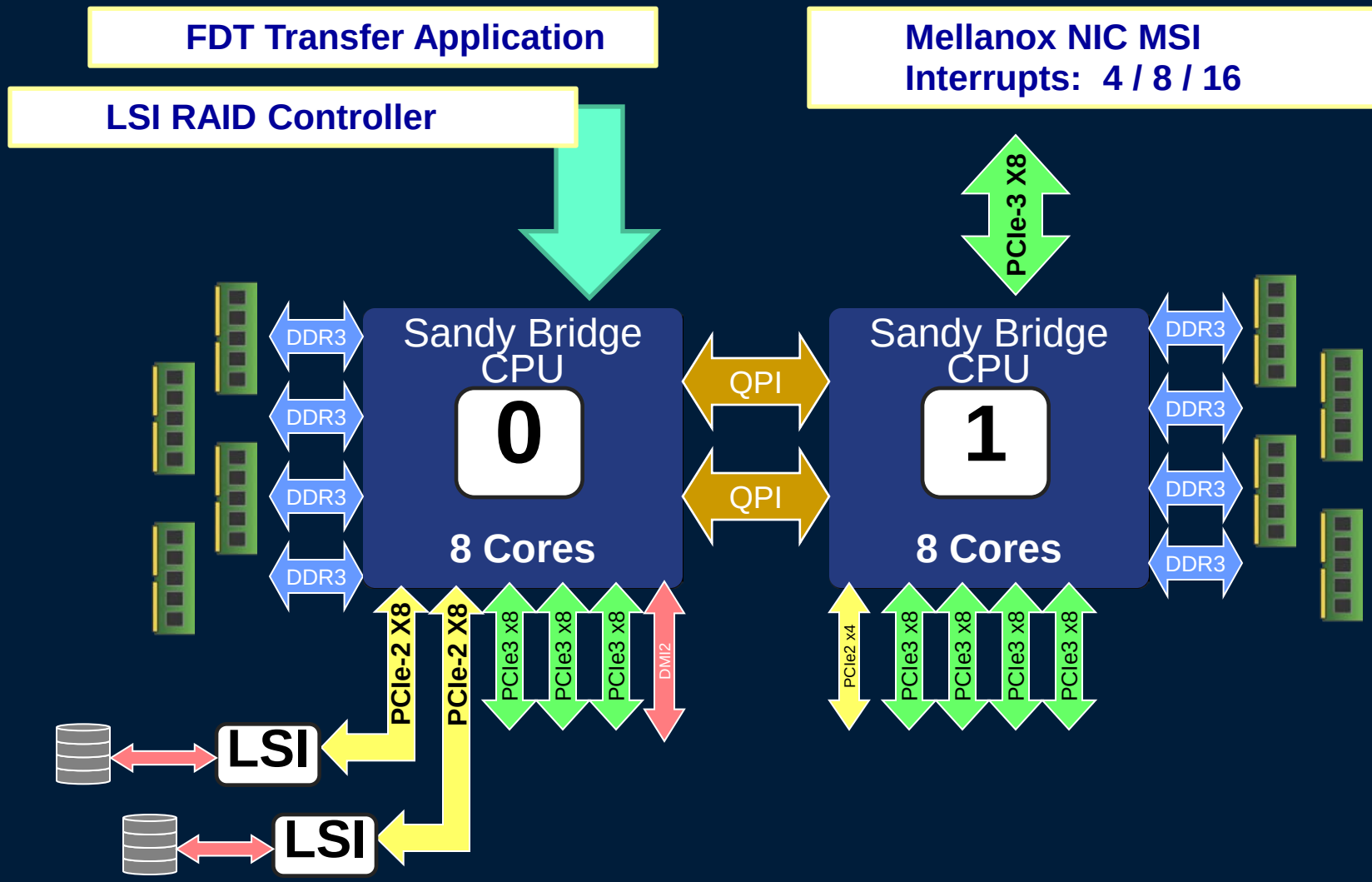
- ✓ **SandyBridge E5 Based Servers:**
  (SuperMicro X9DRi-F or Dell R720)
  Intel E5-2670 with C1 or C2 Stepping
  128GB of DDR3 1600MHz RAM
- ✓ **Mellanox VPI CX-3 PCIe Gen3 NIC**
- ✓ **Dell / Mellanox QSFP Active Fiber Cables**
- ✓ **LSI 9265-8i, 8 port SATA 6G RAID Controller**
- ✓ **OCZ Vertex 3 SSD, 6Gb/s**
  **(preferably enterprise disks like Deneva 2)**
- ✓ **Dell – Force10; Z9000 40GE Switch**

Server Cost = ~ $15k



*http://supercomputing.caltech.edu/40gekit.html*

# System Layout



# Hardware Setting/Tuning

- SuperMicro Motherboard – X9DRi/F
  - PCI-e slot needs to be manually set to Gen3, otherwise defaults are Gen2
  - Disable Hyper threading
  - Change PCI-e payload to the maximum (for Mellanox NICs)
- Mellanox – CX3 VPI
  - Use latest firmware and drivers
  - Use QSFP Active Fiber cables
- Dell-Force10 Switch - Z9000
  - Flow control needs to be turned on for server facing ports
  - Single Queue compared to 4 Queue model
  - MTU = 9000

# Software and Tuning

- Scientific Linux 6.2 Distribution, default kernel
- Fast Data Transfer (FDT) utility for moving data among the sites
  - Writing on the RAID-0 (SSD disk pool)
  - /dev/zero → /dev/null  memory test
- Kernel smp affinity:
  - Bind the Mellanox NIC driver queues to the processor cores where NIC's PCIe Lane is connected
  - Move LSI Driver IRQ to the second processor
- Using NUMA Control to bind FDT application to the second processor
- Change Kernel TCP/IP parameters as recommended by Mellanox

# System Tuning Details

- /etc/sysctl.conf (added during Mellanox driver installation)

```
## MLXNET tuning parameters ##
net.ipv4.tcp_timestamps = 0
net.ipv4.tcp_sack = 0
net.ipv4.tcp_low_latency = 1
net.core.netdev_max_backlog = 250000
net.core.rmem_max = 16777216
net.core.wmem_max = 16777216
net.core.rmem_default = 16777216
net.core.wmem_default = 16777216
net.core.optmem_max = 16777216
net.ipv4.tcp_rmem = 4096 87380 16777216
net.ipv4.tcp_wmem = 4096 65536 16777216
## END MLXNET ##
```

- Ethernet Interface

```
Ifconfig eth2 mtu 9000
ethtool  -G eth2 rx 8192
```

- Numactl (with local node memory binding)

```
numactl --physcpubind=1,2   --localalloc /usr/java/latest/bin/java -jar /root/fdt.jar &
```

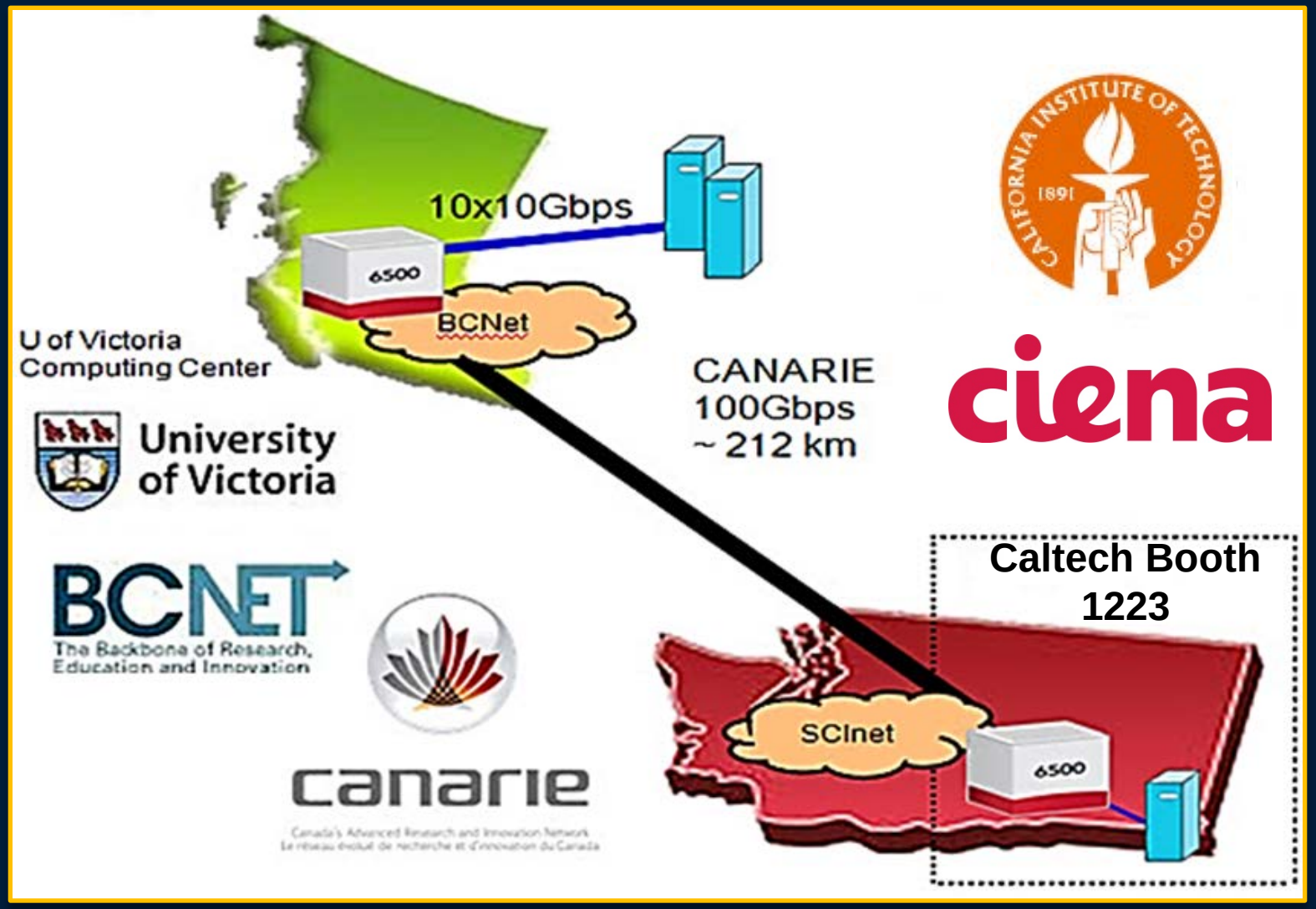
- Smp Affinity (Mellanox NIC)

```
set_irq_affinity_bynode.sh 1 eth2
```

- Smp Affinity (LSI RAID Controller)

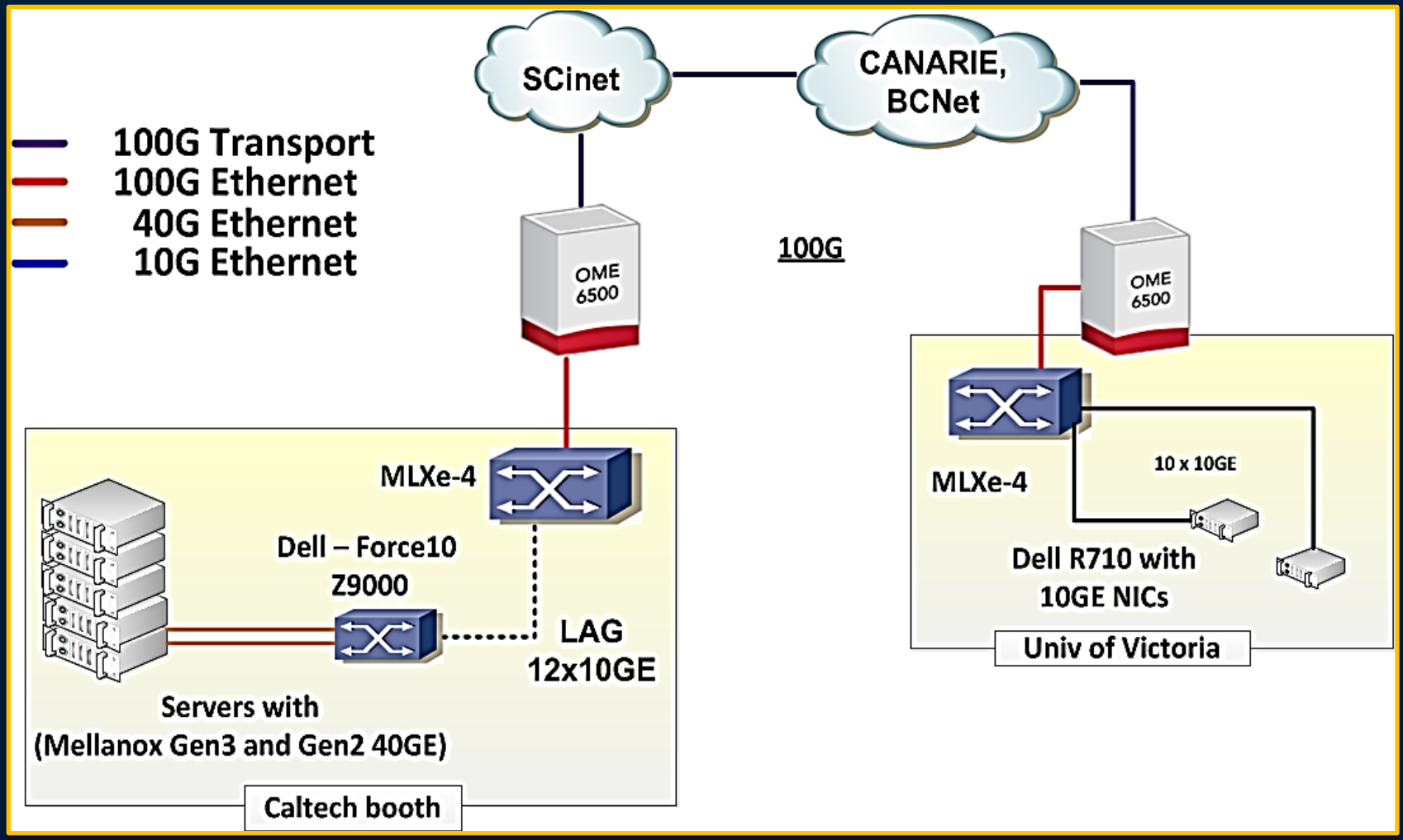
```
echo 20 > /proc/irq/73/smp_affinity
```

# SuperComputing 2011 Collaborators

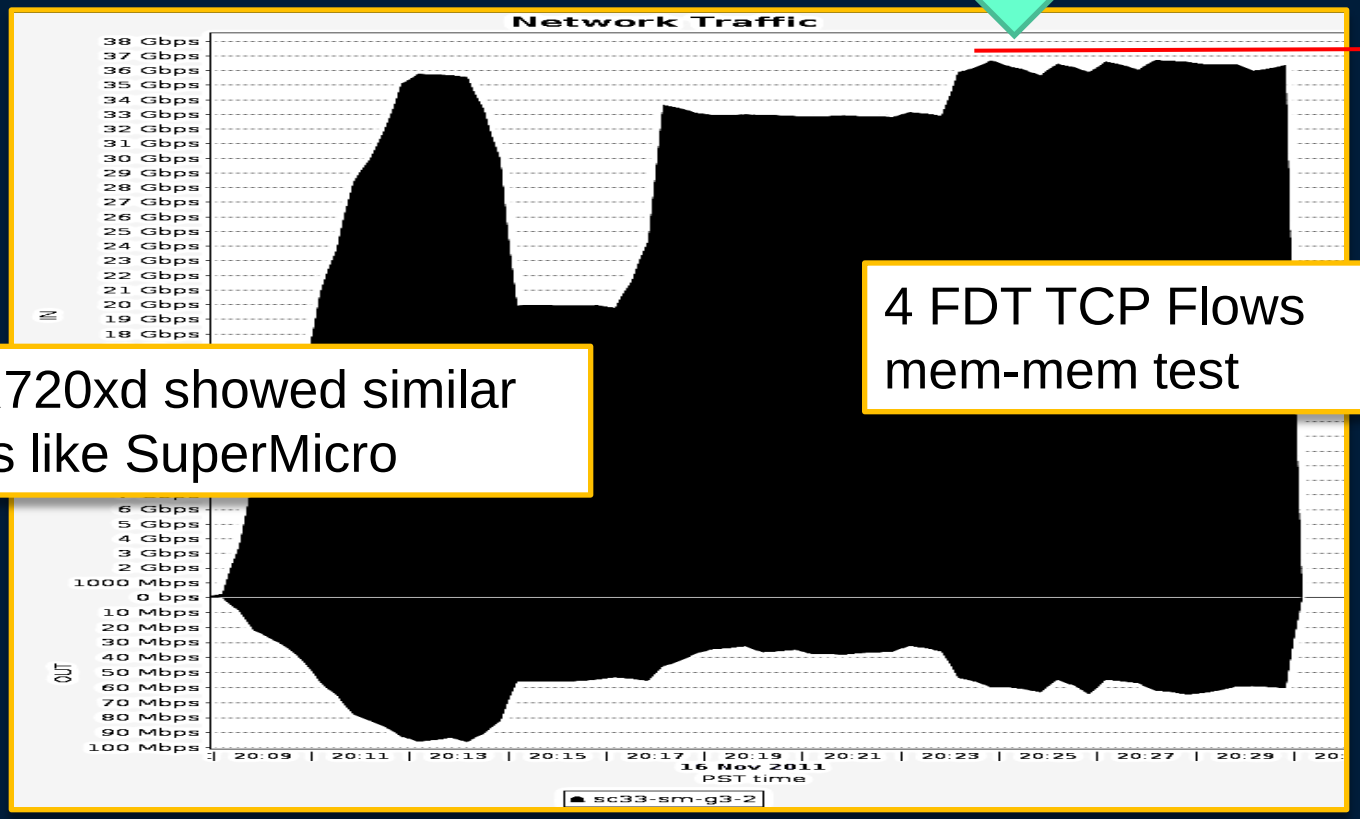



*Courtesy of Ciena*

# SC11 - WAN Design for 100G



# SC11 - PCIe Gen3 performance: 36.8 Gbps

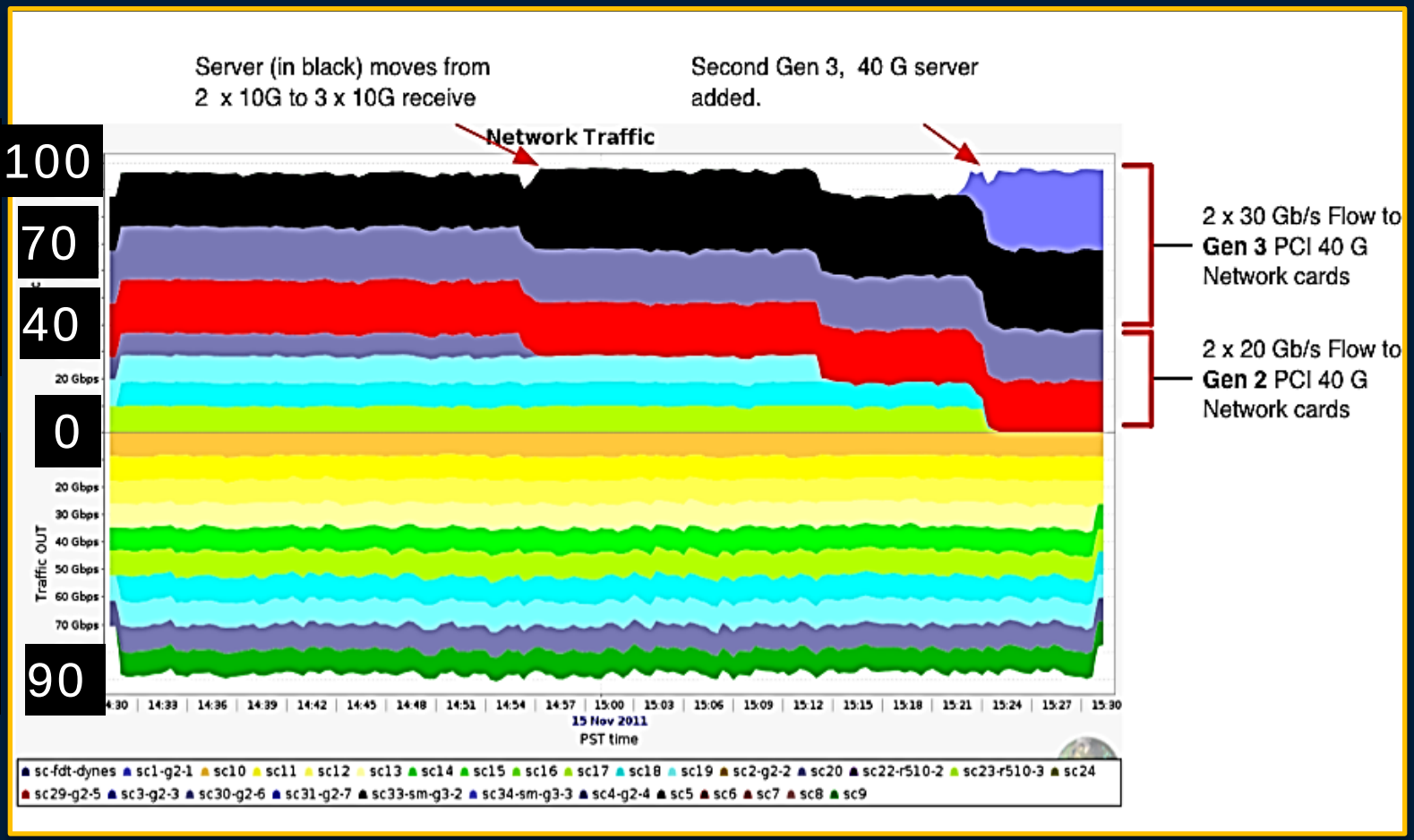



37Gbps

4 FDT TCP Flows
mem-mem test

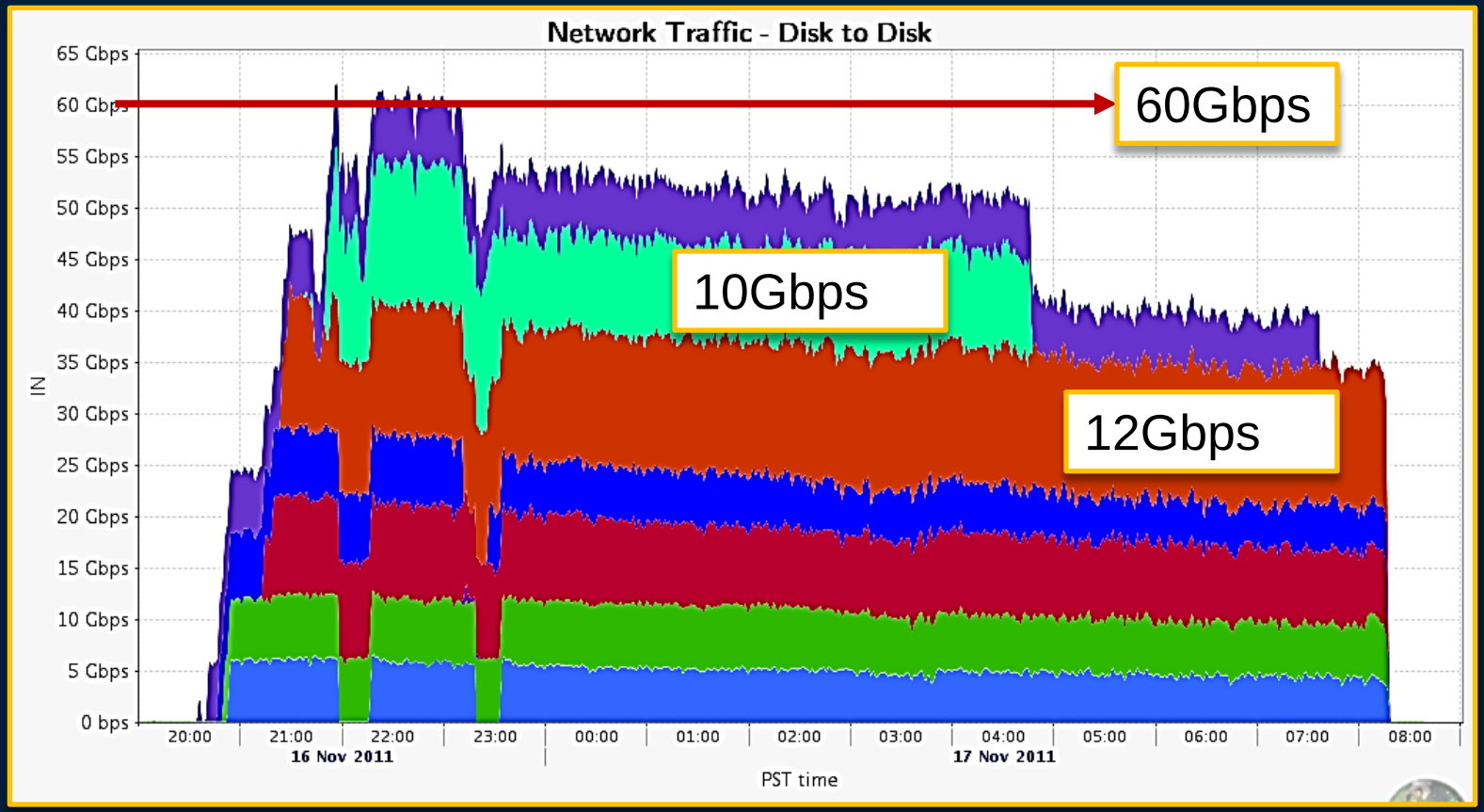
Dell R720xd showed similar results like SuperMicro

# SC11 - Servers Testing, reaching 100G




Sustained 186 Gbps; Enough to transfer 100,000 Blue-ray per day

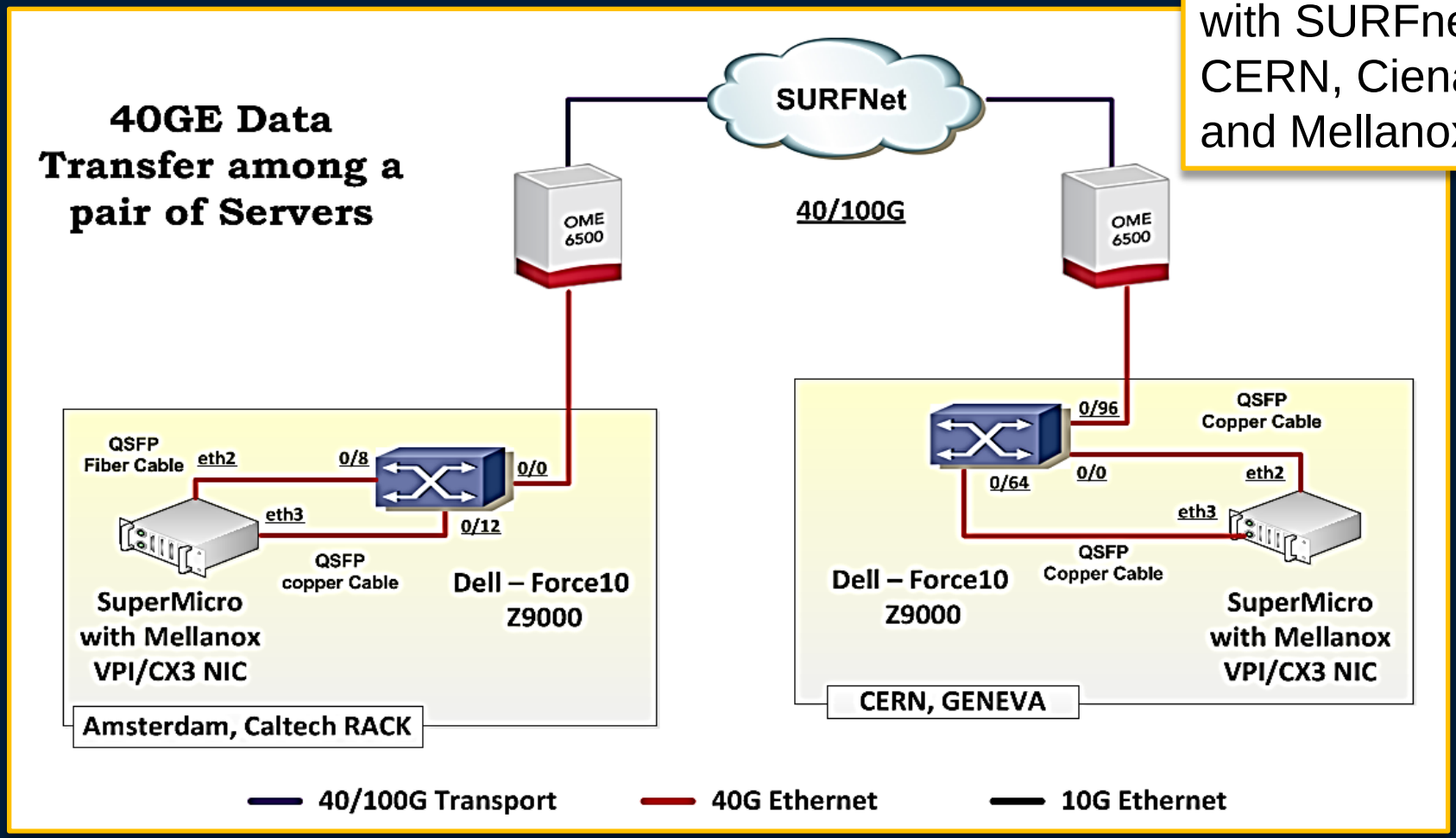
# SC11 - Disk to Disk Results; Peaks of 60Gbps




Disk write on 7 Supermicro and Dell servers with a mix of 40GE and 10GE Servers.

# 40GE Data Transfer Demo between Amsterdam and Geneva

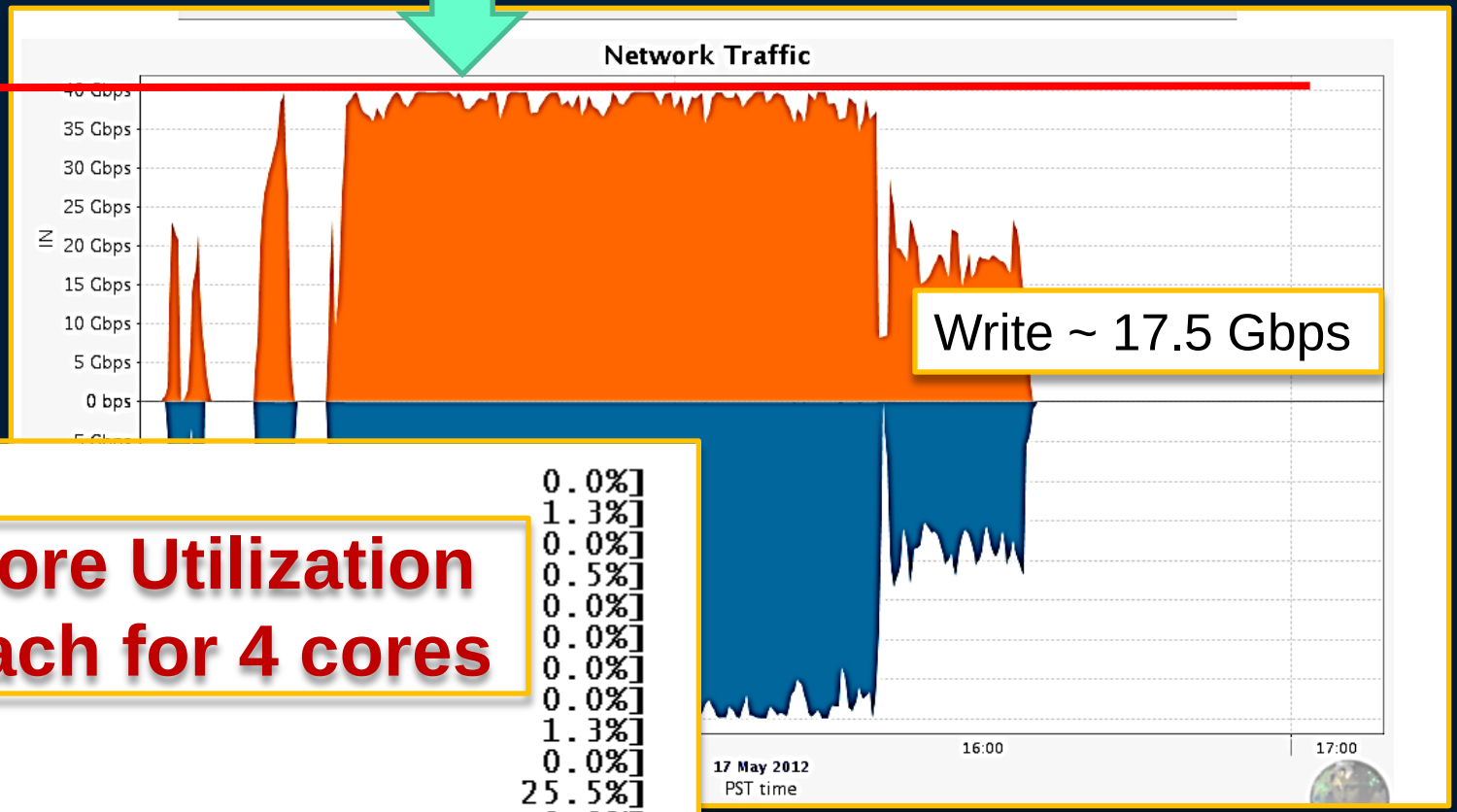
In Collaboration with SURFnet, CERN, Ciena, and Mellanox

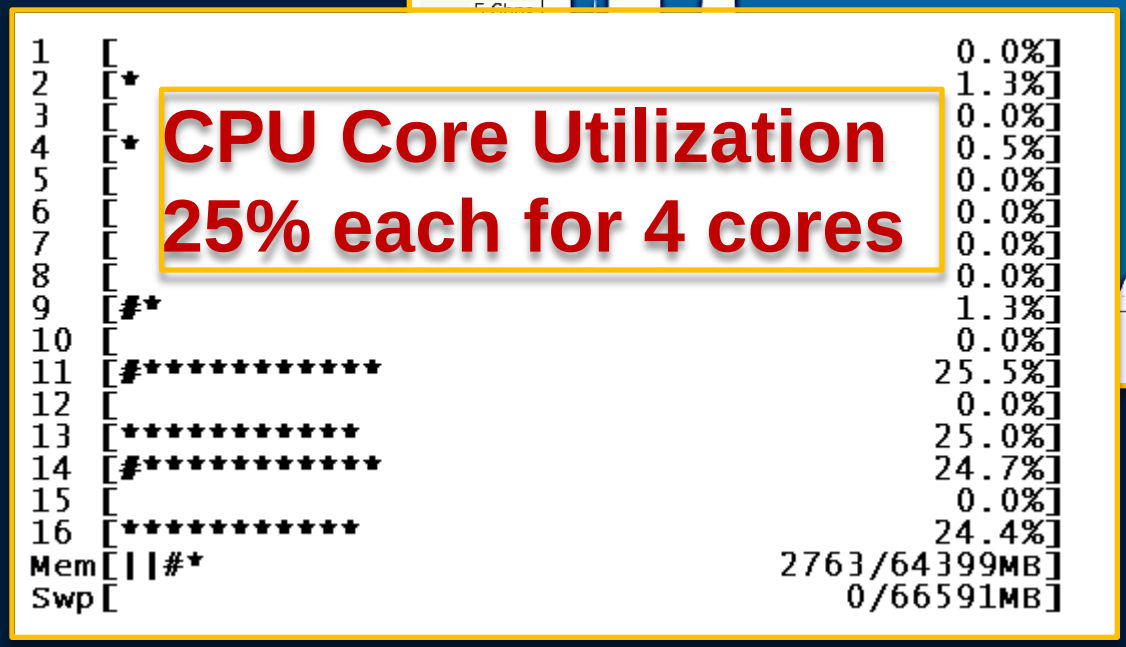



Distance = 1650km, RTT=16ms

# 40GE line rate at 39.6Gbs

40Gbps




**CPU Core Utilization**
**25% each for 4 cores**

# Key Challenges encountered

- SuperMicro Servers, Mellanox CX3 NIC and drivers were all in BETA stage.
- First hand experience with PCIe Gen3 servers using sample E5 Sandy Bridge processors, Not many vendors were available for testing.
- What do we know on the BIOS settings for Gen3 (Slots, processor performance mode)
- Mellanox NIC randomly throwing interface errors.
- QSFP Passive Copper cable has issues at line rate (39.6Gbps).
  - Use Fiber Cables
- LSI drivers, single threaded, utilizing a single core to maximum.
- Will FDT be able to go close to the line rate of Mellanox Network Cards, 39.6Gbps (theoretical peak)
- End to End 100G and 40G testing, any transport issues ?

# Future Directions

- Investigate bottlenecks in the LSI Raid Card driver, New driver supporting MSI-x Vectors is available (many configurable queues)
- Optimizing Linux kernel, SSD tuning as compared to mechanical disks, kernel timers, other unknowns
- Investigate performance for PCIe based SSD drives from vendors like Intel, FusionIO, OCZ
- Ways to lower CPU Utilization, investigate RoCE
- Understand/overcome the SSD wearing out problems over a time

# Summary

- The 40/100Gbps network technology has shown the potential possibilities to transfer peta-scale physics datasets in a matter of hours around the world.
- Three highly tuned servers can easily reach the 100GE line rate, effectively utilizing the PCIe Gen3 technology.
- Individual Server tests using E5 processors and PCIe Gen-3 based Network Cards have shown stable network performance reaching line rate at 39.6Gbps.
- During SC11, Fast Data Transfer (FDT) application achieved an aggregate disk write of 60Gbps.
- MonALISA intelligent monitoring software, effectively recorded and displayed the traffic at 40/100G and the other 10GE links.

# Questions ?