Demanding Applications Networking

40G and beyond

Ronald van der Pol

<rvdp@sara.nl>

RoN Spring Meeting, 19 May 2010, Utrecht

Contributors

- Ronald van der Pol
- Freek Dijkstra
- Pieter de Boer
- Igor Idziejczak
- Mark Meijerink
- Hanno Pet
- Peter Tavenier

Outline

- Network bandwidth requirements
- Status 100GE deployment
- Scaling networking I/O
- Towards terabit networking
- Disk and network I/O measurements
- Conclusions
- Items for discussion

Outline

- Network bandwidth requirements
- Status 100GE deployment
- Scaling networking I/O
- Towards terabit networking
- Disk and network I/O measurements
- Conclusions
- Items for discussion

Data Transfer Challenge

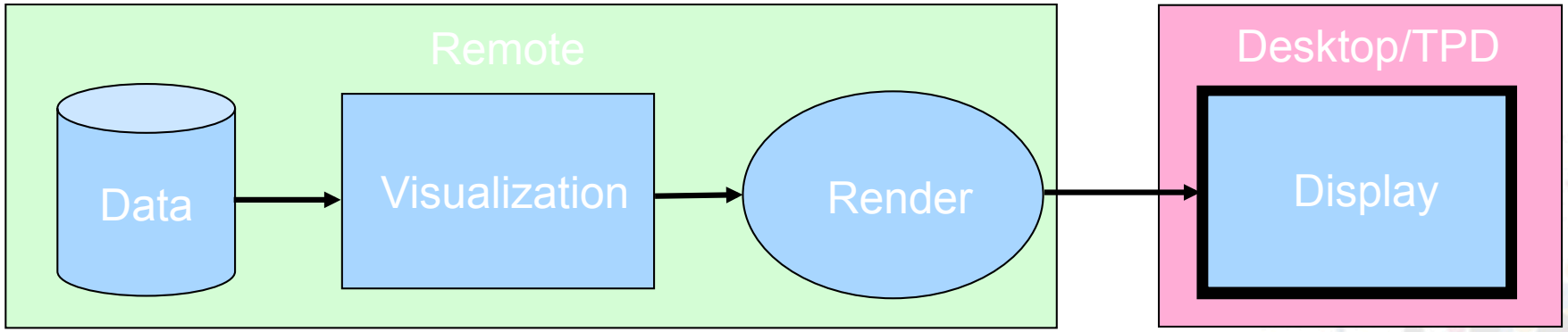
- Scientific data sets continue to grow
  - Petaflop computing
  - Petabyte storage challenge
  - Petabit/s network challenge
- Supercomputing 2009 bandwidth challenge:
  - *"How fast can you move a scientific petabyte data set over a high performance research network?"*
- Transfering 1 Petabyte of data takes:

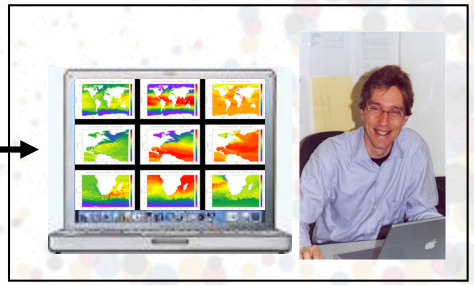
Bandwidth	Transfer Time
10 Gbit/s	9 days, 6:13 hours
40 Gbit/s	2 days, 7:33 hours
100 Gbit/s	22:13 hours
1 Tbit/s	2:13 hours
1 Pbit/s	8 seconds

Data Streaming Challenge

- Supercomputing grand challenges produce large data sets
- Results need to be visualised
- Transfering large data sets to the scientist is impractical
  - It takes too long to transfer the data
- Remote streaming of visualisation data is an option
  - Keep data at the supercomputer site
  - Visualise and render at the supercomputer site
  - Stream pixels over the network to the scientist
- But even this setup is a network challenge
- Streaming to a 60 Mpixel tiled panel @ 30 fps generates more than 40 Gbit/s of network traffic

Streaming Visualisation Data





Department of Energy
FY 2011 Congressional Budget Request

"ESnet will be upgraded seamlessly to meet the growing, complex needs of DOE and remains on a path to deliver 1 terabit per second connectivity in 2014."

Outline

- Network bandwidth requirements
- Status 100GE deployment
- Scaling networking I/O
- Towards terabit networking
- Disk and network I/O measurements
- Conclusions
- Items for discussion

Commercial 100G Deployment

PCWorld News Reviews How-To's Downloads Shop & Compare Forums Business Center

Let's protect the network by sealing the gateway.

PCWorld Business Center

Software & Services | Office Hardware | Security | Servers & Storage | Cell Phones & Mobile | Virtualization

CELL PHONES / VOIP December 14, 2009 3:00 PM

Verizon Business Lights up 100G Backbone in Europe

By Stephen Lawson, IDG News Service

Verizon Business has deployed 100Gb-per-second equipment on a fiber link between Paris and Frankfurt, activating the next generation of optical networking for commercial services to enterprises.

NETWORKWORLD

This story appeared on Network World at
http://www.networkworld.com/news/2009/090209-qwest-ethernet.html

Qwest upgrading backbone to 100Gbps

Alcatel-Lucent supplying Qwest with 100Gbps technology
By Brad Reed, Network World
September 02, 2009 11:27 AM ET

LIGHTWAVE

Trusted technical insights for optical communications p

HOME FTTX TEST & MEASUREMENT EQUIPMENT DES

Home > Networking > Networking News

Text Size RSS Feed email

Ericsson, Telefonica test 100 Gbps

January 18, 2010

TELECOMPAPER

Verizon, Juniper, NEC, Finisar complete 100G field trial

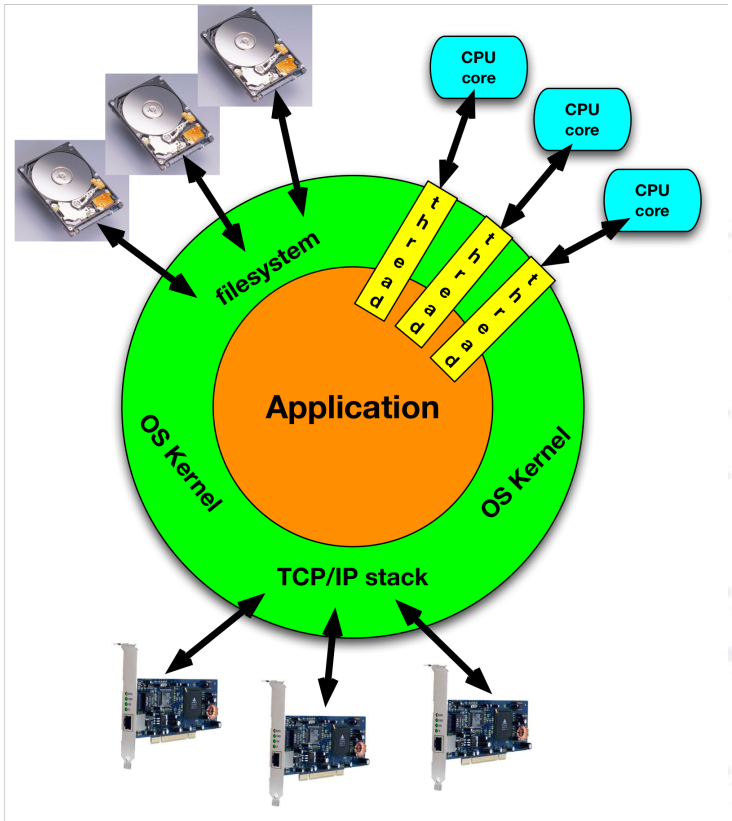
Published: Tuesday 9 March 2010 | 08:01 AM CET, Telecompaper

Outline

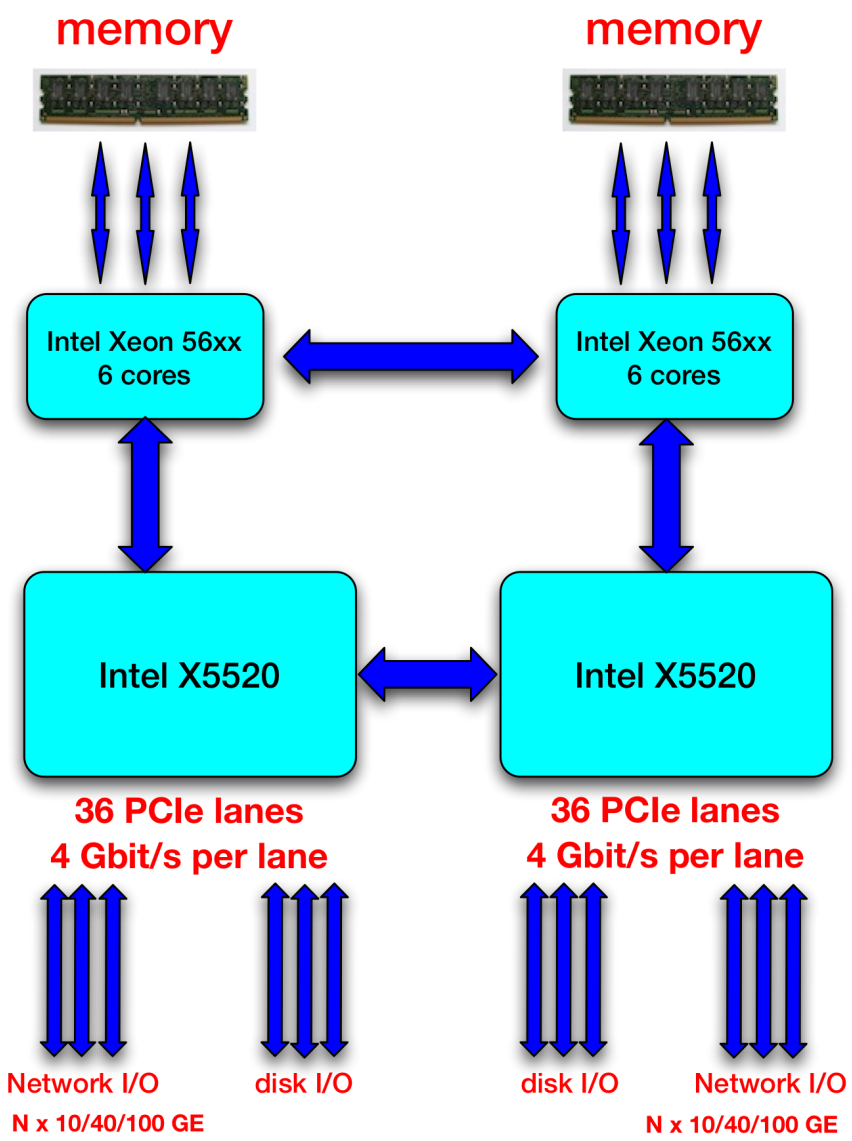
- Network bandwidth requirements
- Status 100GE deployment
- Scaling networking I/O
- Towards terabit networking
- Disk and network I/O measurements
- Conclusions
- Items for discussion

I/O Scalability

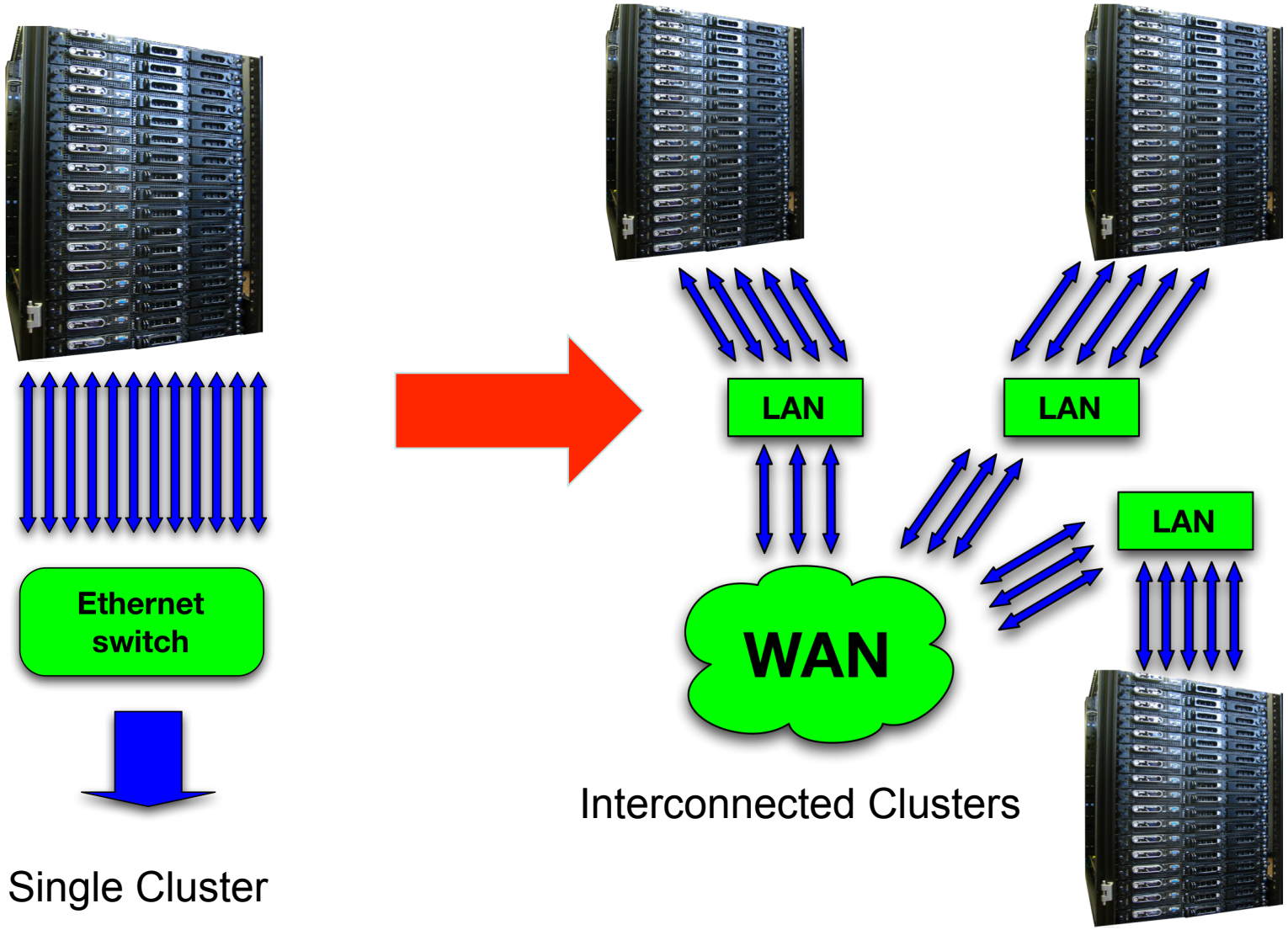
- Storage I/O speedup with multiple disks (RAID-1/RAID-Z)
- Compute speedup with multi-core systems
- Network I/O speedup with multiple NICs



Single Server Architecture



Scaling to Clusters



Single Cluster

Interconnected Clusters

CosmoGrid

- Dutch computing challenge project
- Cosmological N-body simulation with 8,589,934,592 particles
- Distributed application using several European and Japanese supercomputers

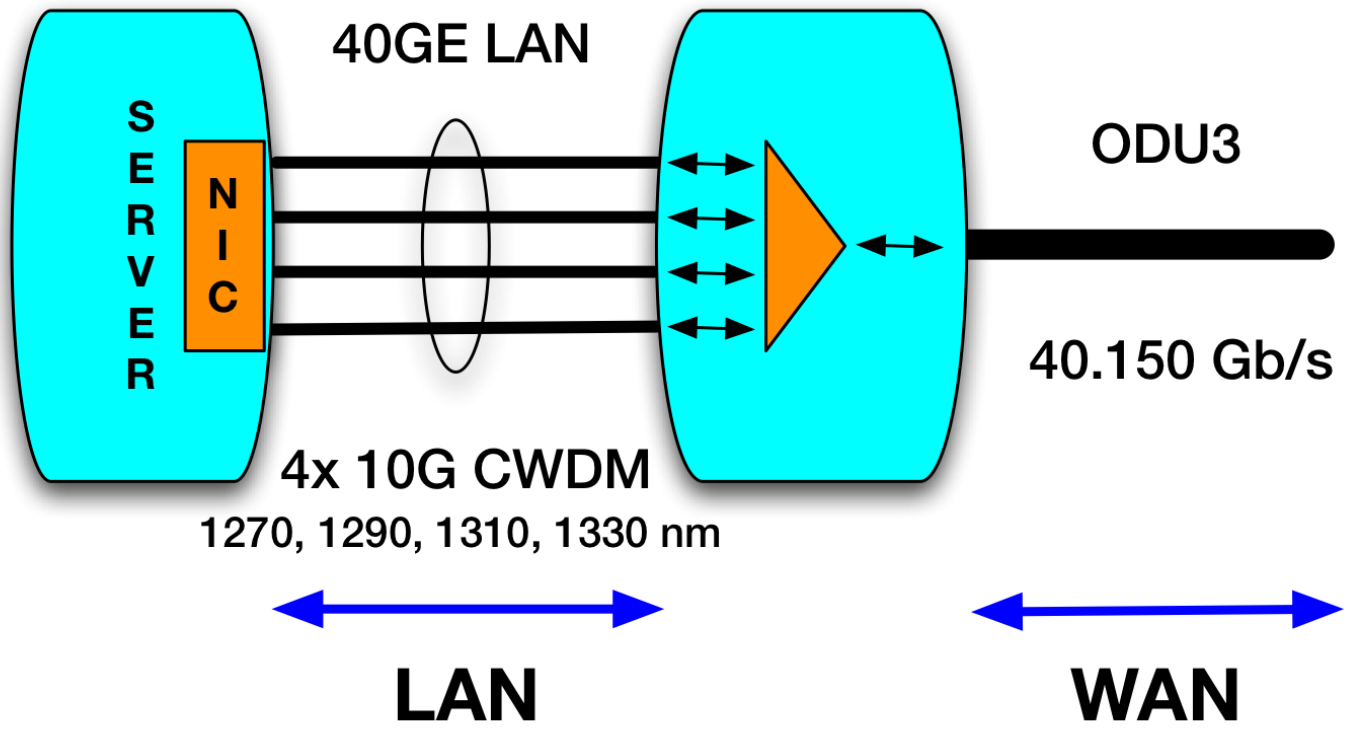
Outline

- Network bandwidth requirements
- Status 100GE deployment
- Scaling networking I/O
- Towards terabit networking
- Disk and network I/O measurements
- Conclusions
- Items for discussion

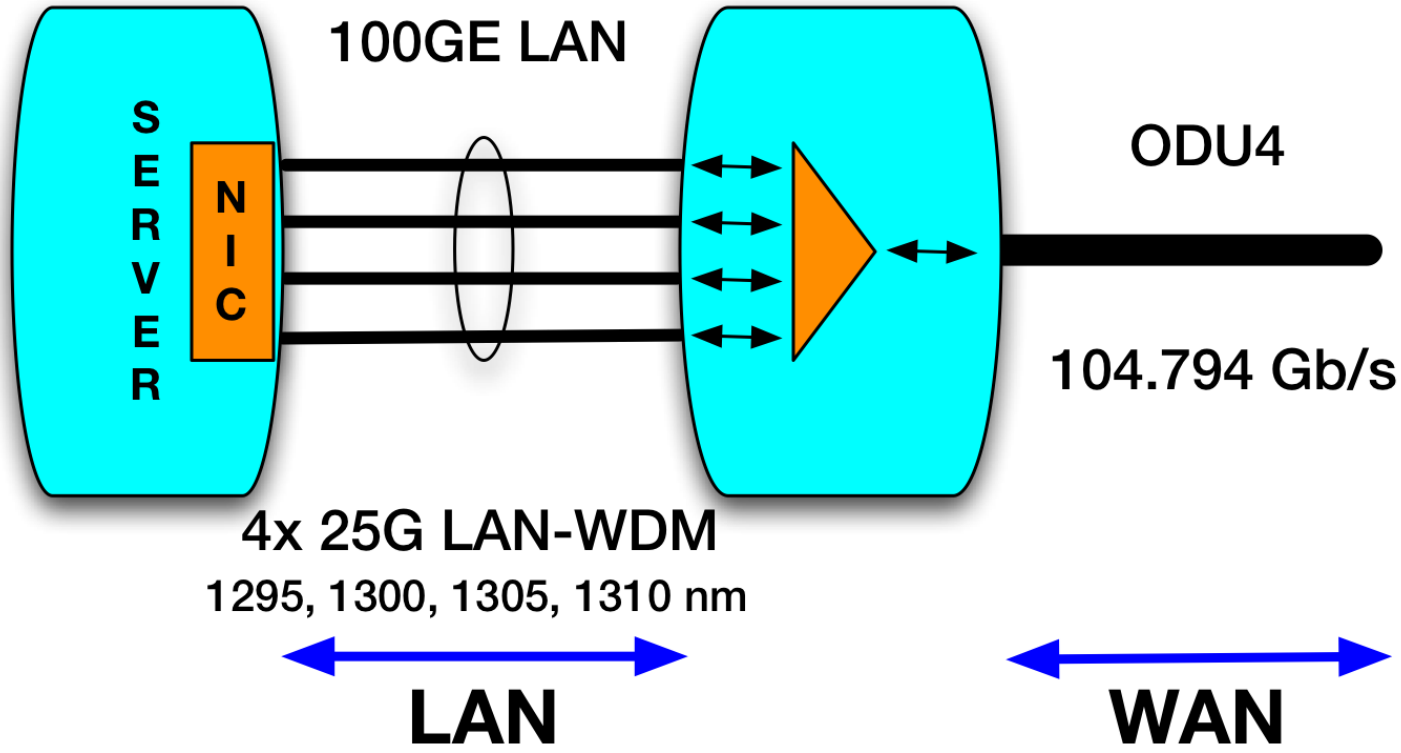
Optical Transport Network

OTU	rate	ODU	rate	client
OTU2	255/237 x 9.953 = 10.709 Gb/s	ODU2	239/237 x 9.953 = 10.037 Gb/s	STS-192
OTU2e	255/237 x 10.3125 = 11.096 Gb/s	ODU2e	239/237 x 10.3125 = 10.400 Gb/s	10GE LAN
OTU3	255/236 x 39.813 = 43.018 Gb/s	ODU3	239/236 x 39.813 = 40.319 Gb/s	STS-768
OTU3e1	255/236 x 10.3125 = 44.571 Gb/s	ODU3e1	239/236 x 10.3125 = 41.774 Gb/s	40GE
OTU4	255/237 x 99.5328 = 111.81 Gb/s	ODU4	239/227 x 99.5328 = 104.79 Gb/s	100GE

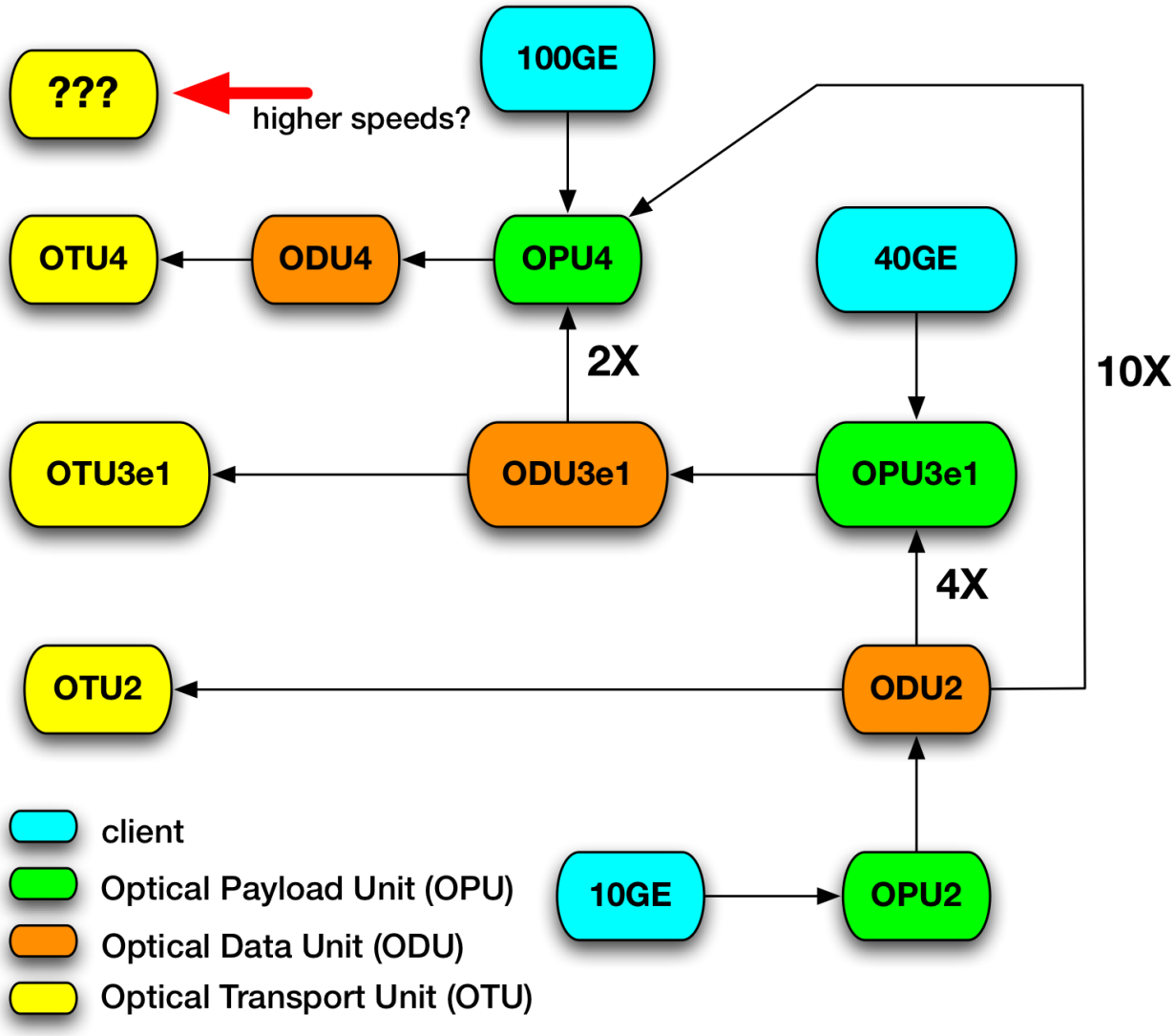
40 Gigabit Ethernet



100 Gigabit Ethernet



OTN Hierarchy



Outline

- Network bandwidth requirements
- Status 100GE deployment
- Scaling networking I/O
- Towards terabit networking
- Disk and network I/O measurements
- Conclusions
- Items for discussion

Single Server Measurements

- What can we squeeze out of a single server?
- Network I/O must be balanced with compute power and storage capacity
- Tuning and optimizing needed to get terabit I/O
- What disk I/O can we get with many parallel disks?
  - Do we get linear speedup with additional disks?
- What network I/O can we get with several NICs?
  - Can we get good load balancing with a single flow?
  - Do we get linear speedup with additional NICs?

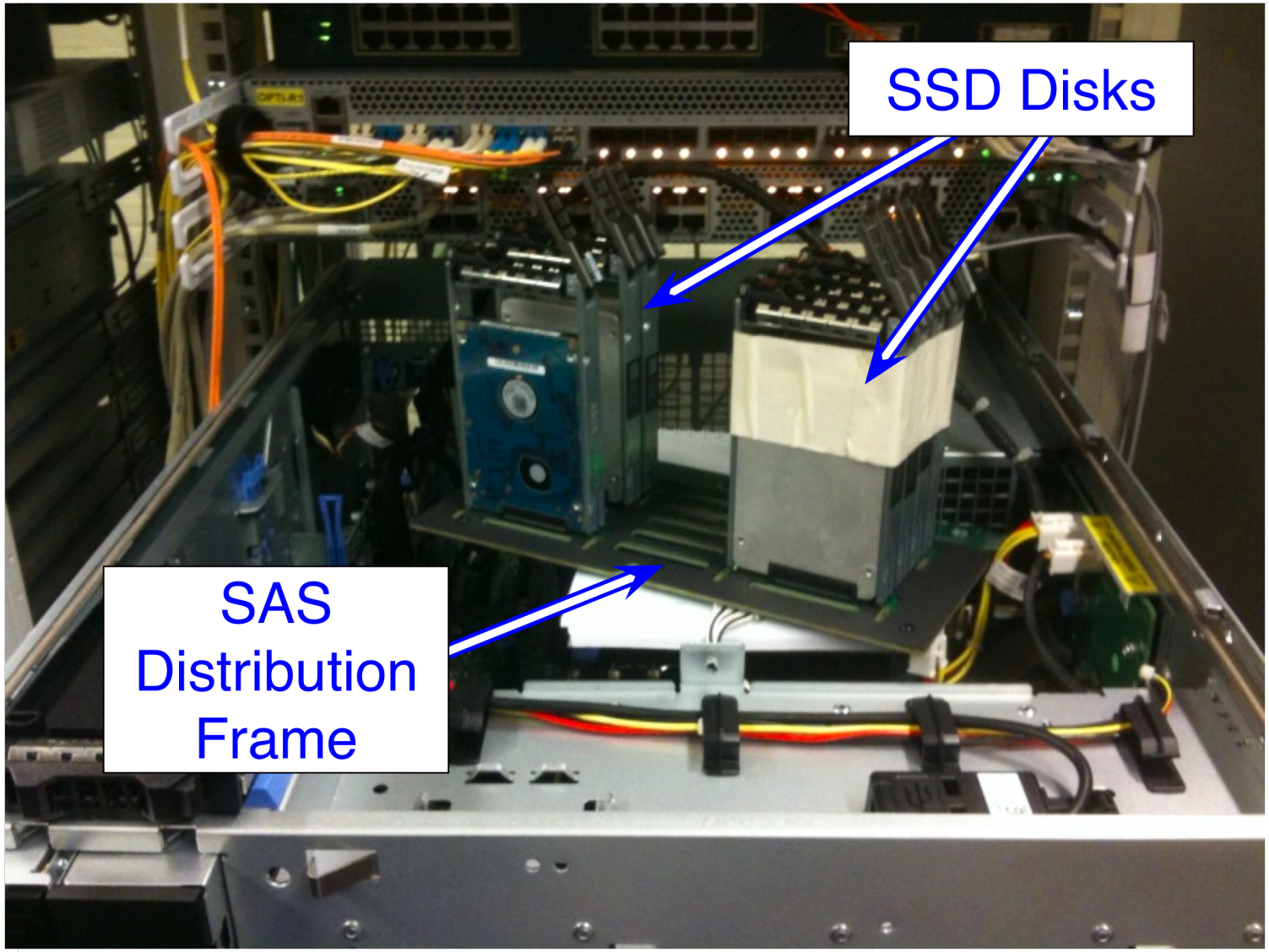
Disk I/O Scalability

- 9 Intel X-25 Solid State Disks
- Dell Perc6/i RAID controller
- Intel Xeon 5550 @ 2.66 GHz
- 6 GB DDR3 @ 1333 MHz

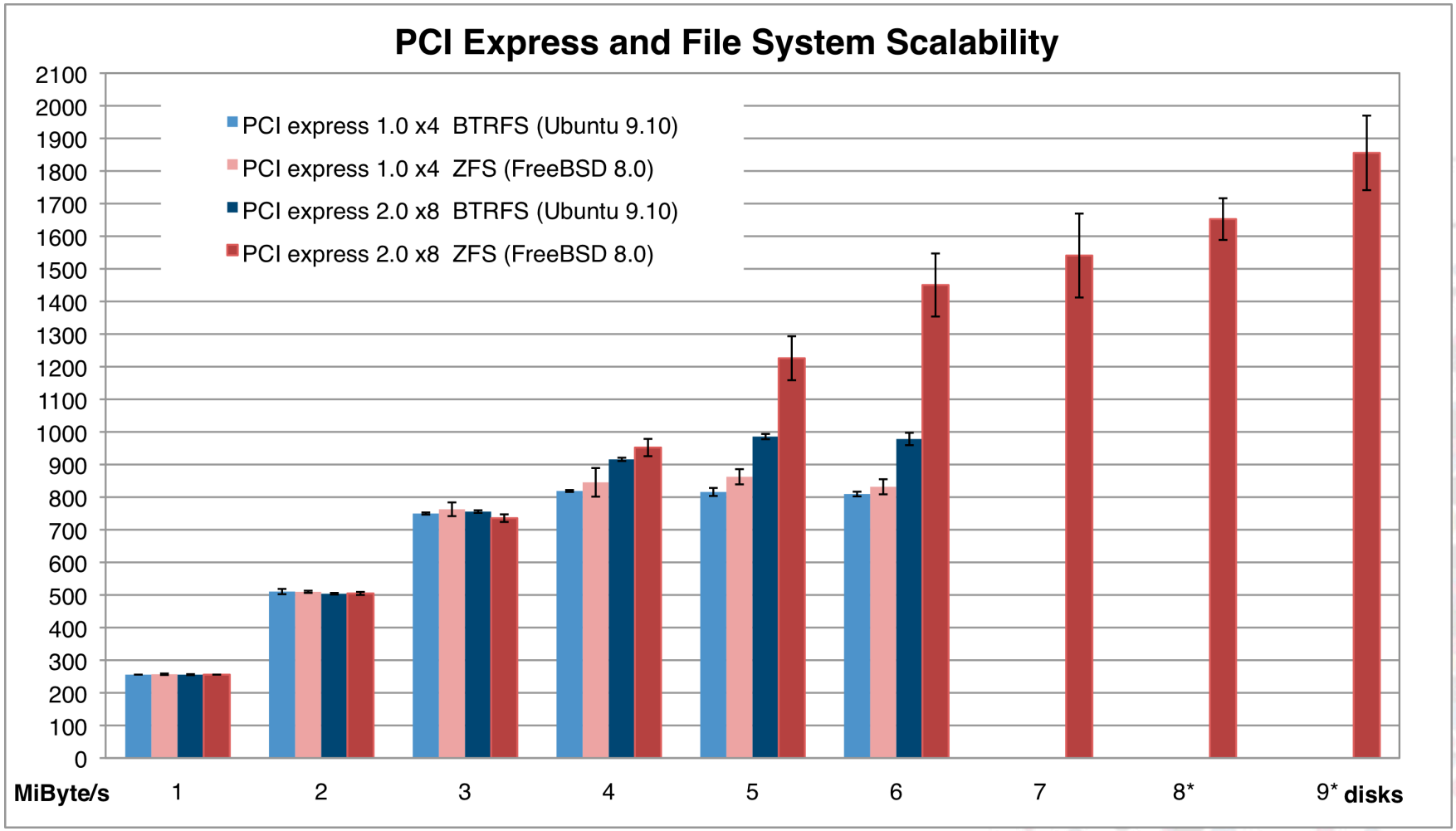
- Dedicated PERC6/i slot turned out to be PCIe 1.0 x4
  - PCIe 1.0 x4 I/O limited to 8 Gbit/s
- Moving PERC6/i card to PCIe 2.0 x8 required disassembling the complete disk subsystem
  - PCIe 2.0 x8 throughput is 16 Gbit/s

	X4	x8
PCIe 1.0	8 Gbit/s	16 Gbit/s
PCIe 2.0	16 Gbit/s	32 Gbit/s

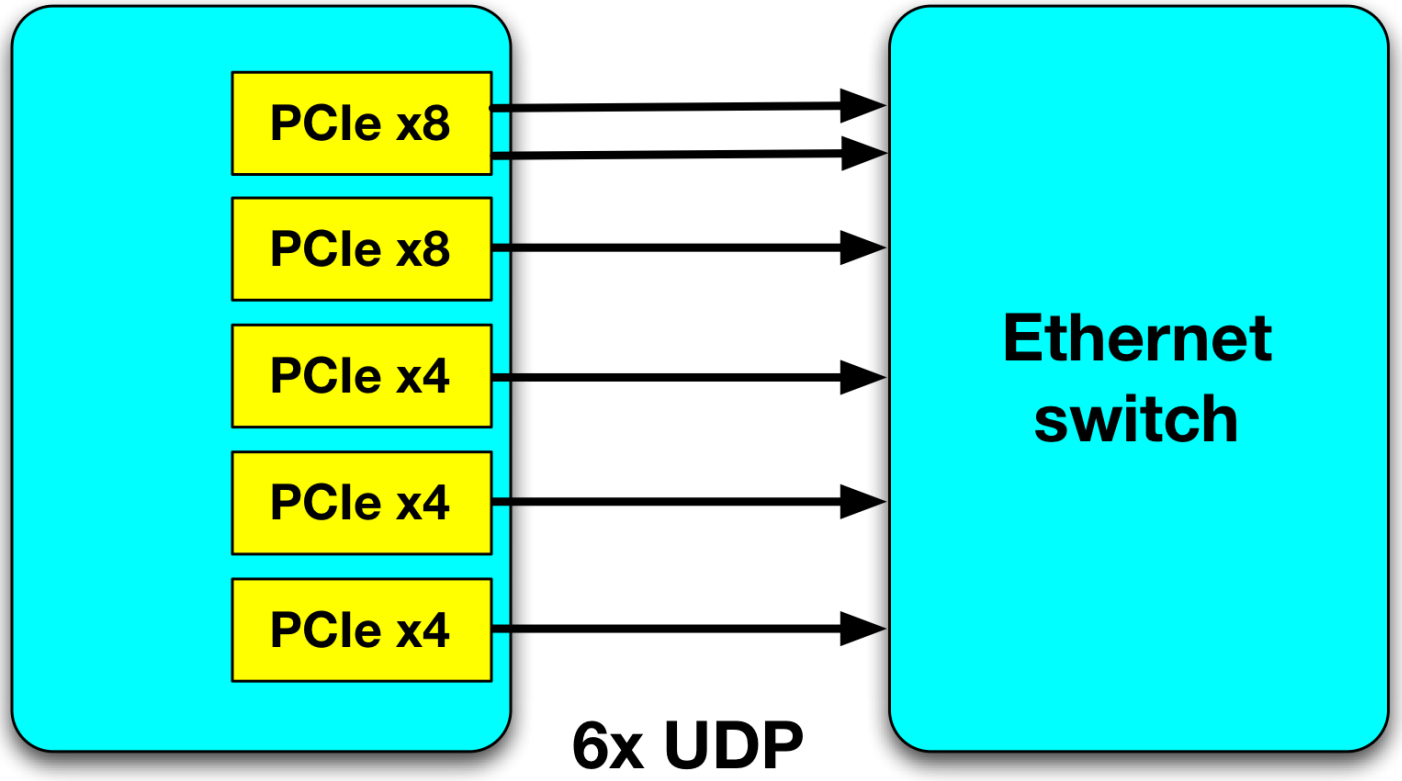
Disassembled Disk Subsystem



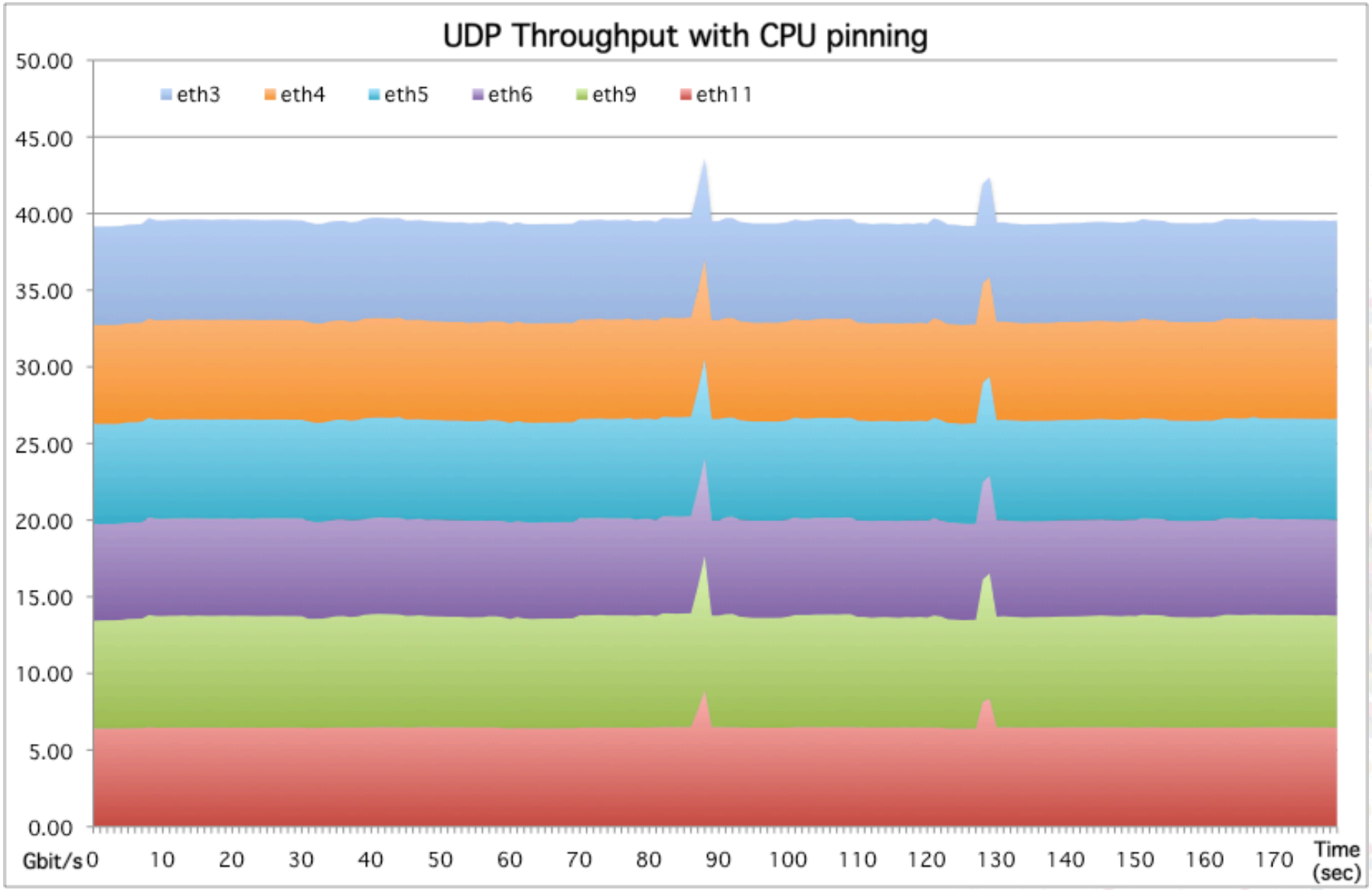
16.12 Gbit/s SSD Read Speed



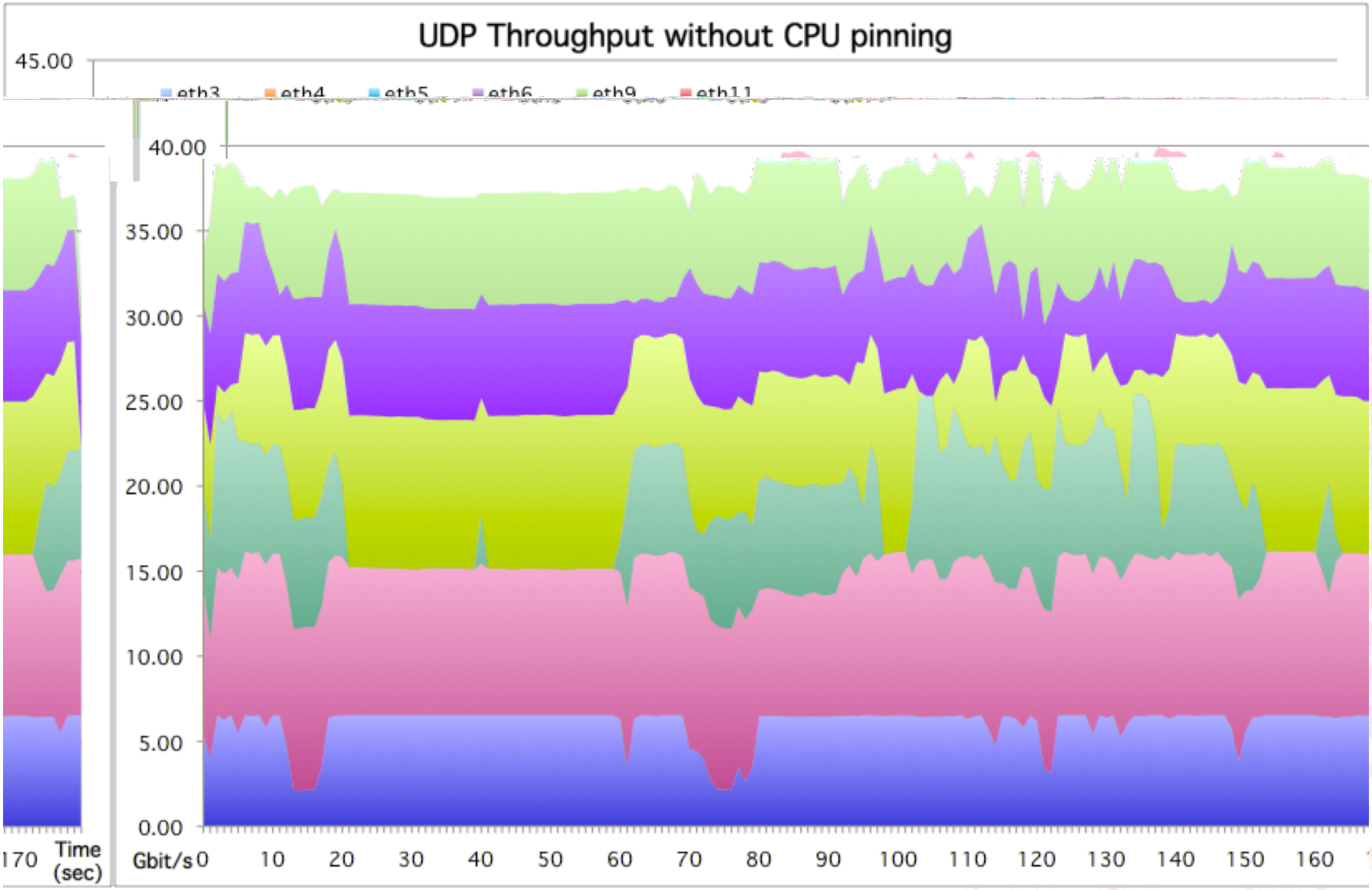
UDP Streaming Setup



40 Gbit/s Network Throughput

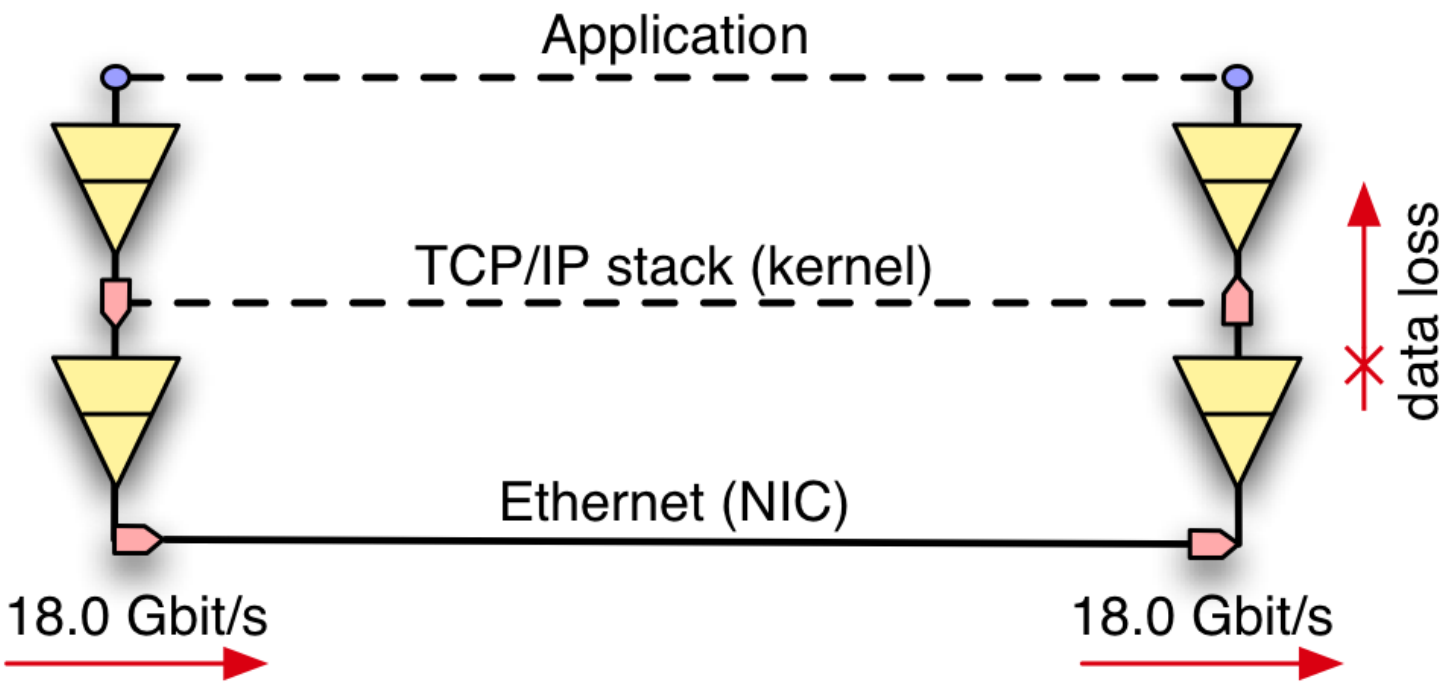


40 Gbit/s Network Throughput



20GE Ethernet Channel

- Receiving server could not process 18 Gbit/s data stream
- Time to buy a higher performance server ☺



Outline

- Network bandwidth requirements
- Status 100GE deployment
- Scaling networking I/O
- Towards terabit networking
- Disk and network I/O measurements
- Conclusions
- Items for discussion

Conclusions

- Linear disk I/O scaling with ZFS and SSD disks to 16 Gbit/s
- 40 Gbit/s UDP streaming with six 10GE NICs
- Additional measurements hindered by limited hardware
- New server should have enough PCIe 2.0 lanes
- CPU does not seem to be a bottleneck
- Memory I/O does not seem to be a bottleneck
- Measurements and analysis show that 40 Gbit/s from disk to network should be possible

Outline

- Network bandwidth requirements
- Status 100GE deployment
- Scaling networking I/O
- Towards terabit networking
- Disk and network I/O measurements
- Conclusions
- Items for discussion

Items for Discussion

- Topology for the 40G/100G demo at GLIF
- Topology for the 40G demo at SC10
- Towards terabit networking in SURFnet7/SURFnet8

Additional Information

- http://nrg.sara.nl/
- http://nrg.sara.nl/publications/RoN2010-D1.1.pdf
- http://nrg.sara.nl/publications/40G-Applications.pdf
- Email: nrg@sara.nl

Thank you

Ronald van der Pol
rvdp@sara.nl