CHEMNITZ UNIVERSITY OF TECHNOLOGY

Design and Evaluation of a 2048 Core Cluster System

Frank Mietke, Torsten Höfler, Torsten Mehlan and Wolfgang Rehm

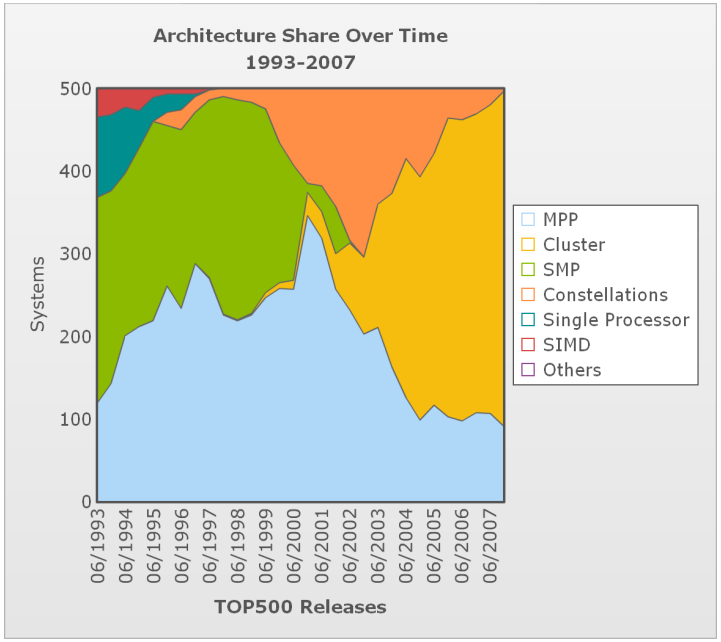
Computer Architecture Group
Department of Computer Science
Chemnitz University of Technology

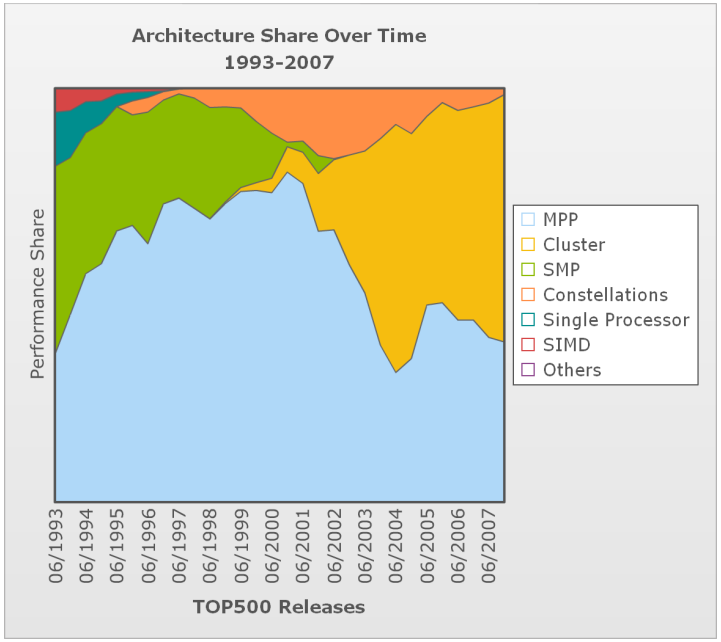
December 12, 2007

CHEMNITZ UNIVERSITY OF TECHNOLOGY

Outline

Supercomputing in General





- Clusters are dominant (81.2%)
- Power Consumption problematic (Green500)

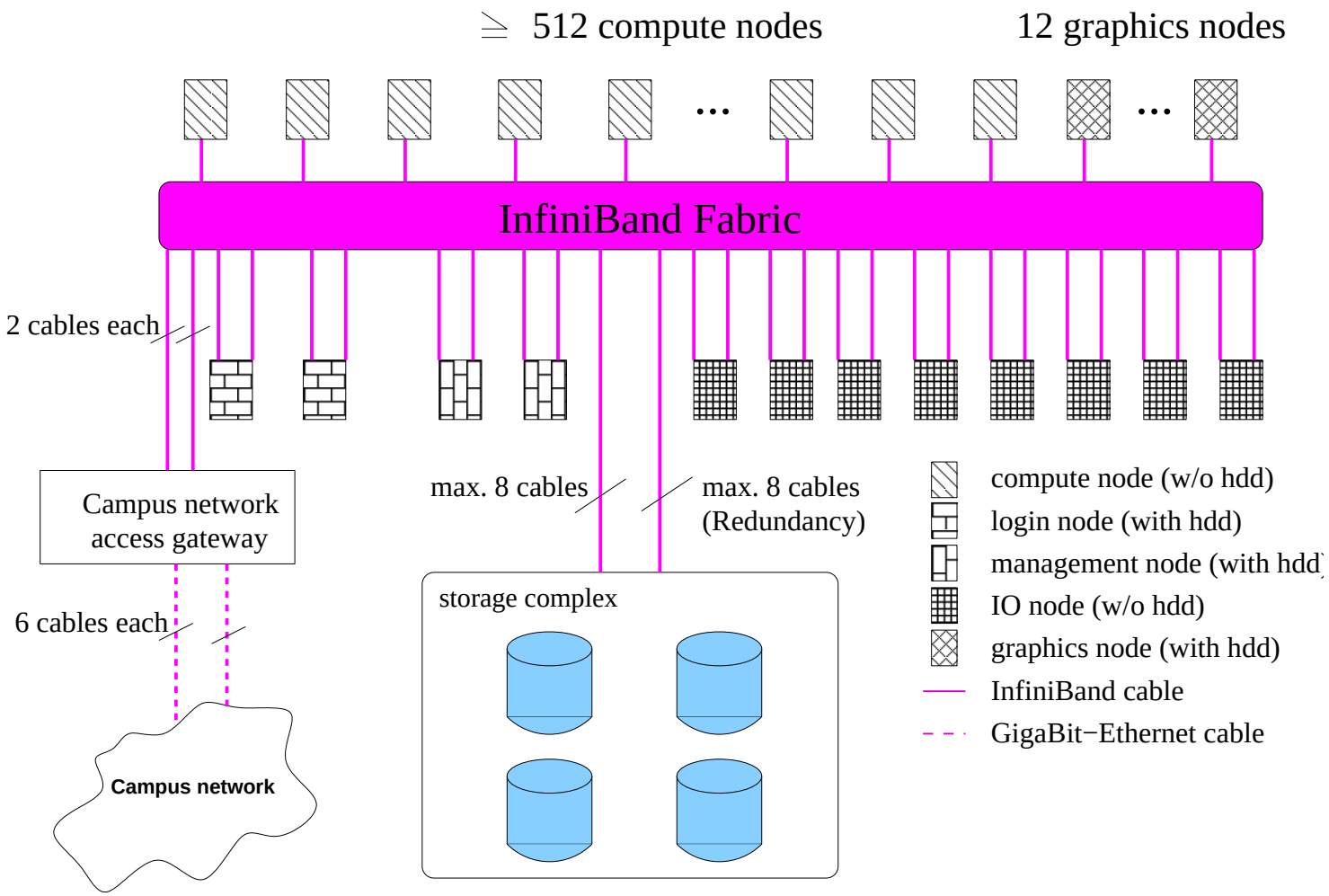
Supercomputing at Chemnitz

- Since 1994
- Growing User Community

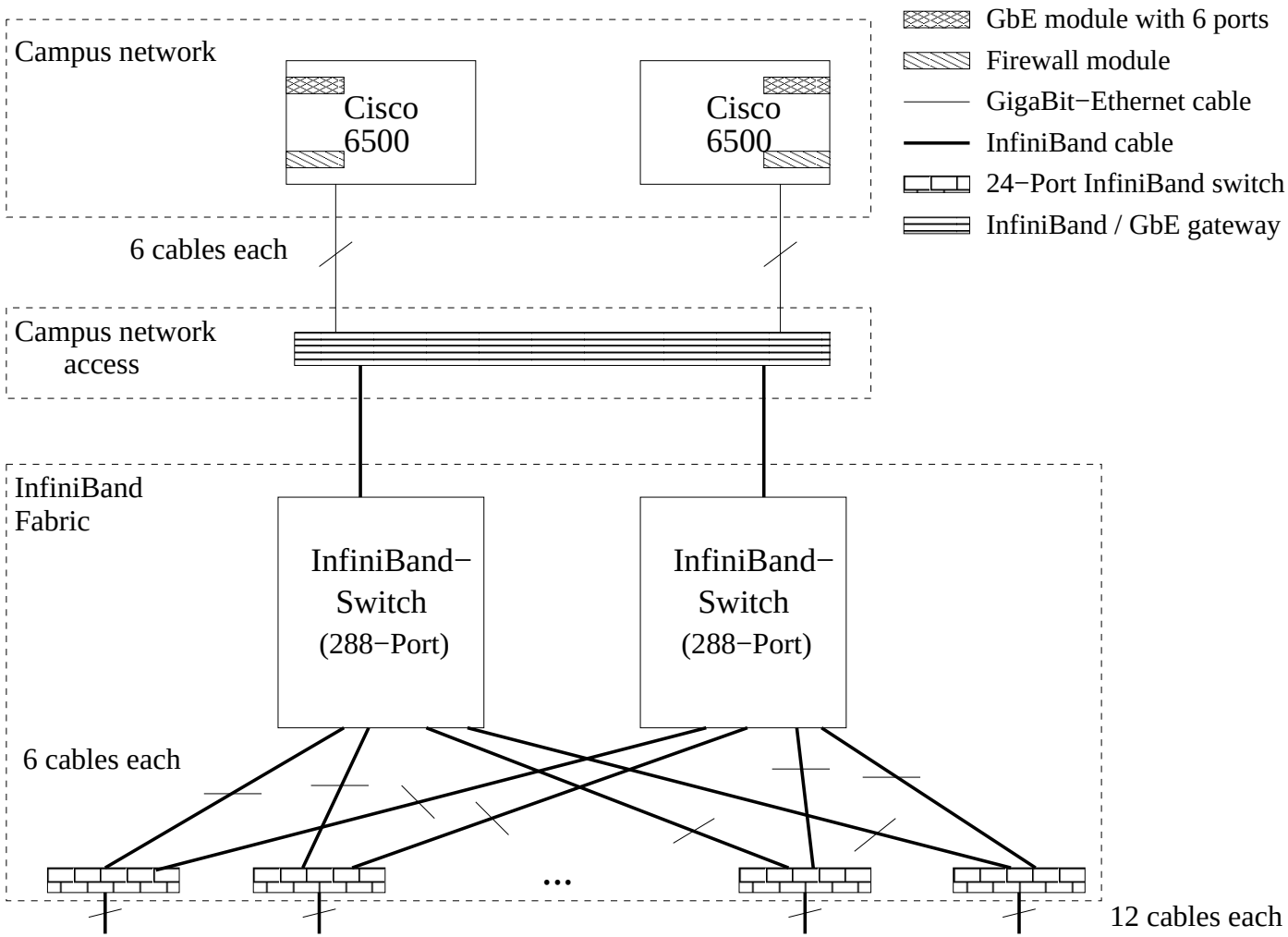
Parsytec – 20 GFlop/s

CLiC – 221.6 GFlop/s

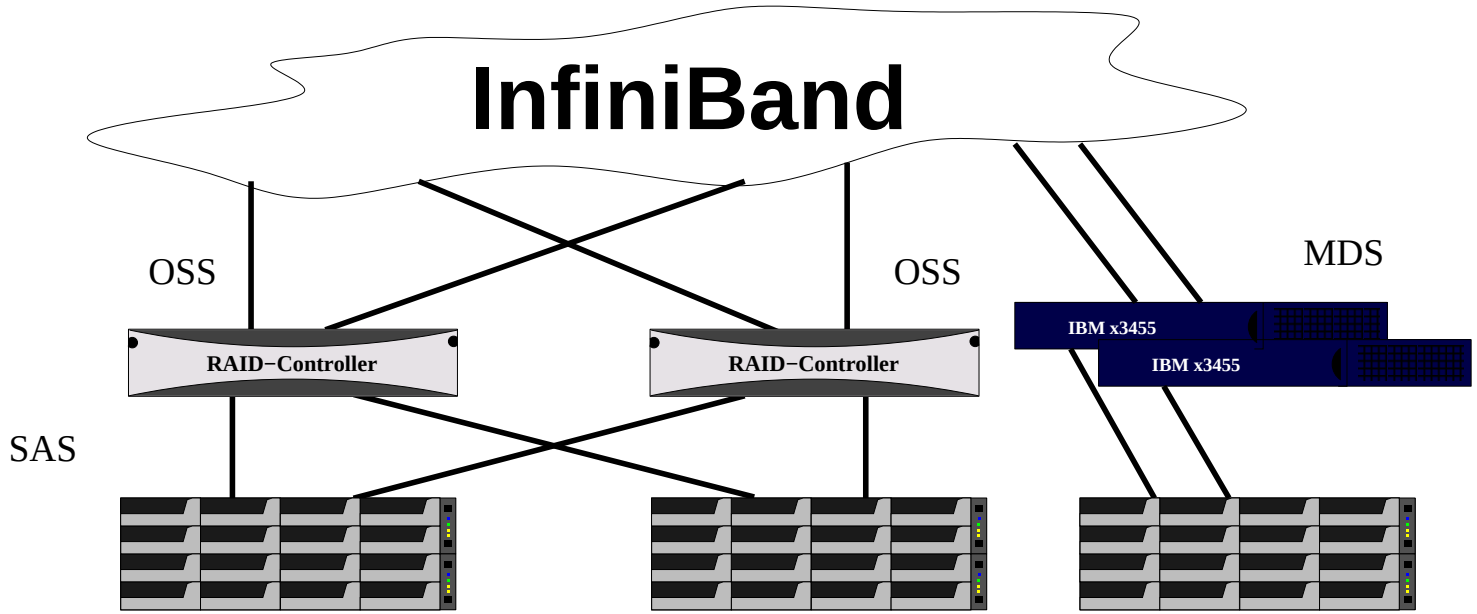
Cluster Design



Network Design



Storage Design



5x

The CHiC – Top500

- Rank 80 (Nov. 2006 - inofficial)
- Rank 117 (Jun. 2007)
- Rank 237 (Nov. 2007)

CHiC – 8.21 TFlop/s

But we provide more ...

- 12+ TFlops (Single Precision)
- www.gpgpu.org

Experiences

Hardware

- Very good Hardware Reliability (so far)
- IB-Eth-Gateway or Fabric Inconsistencies (Load Sit.)
  - Complex IB Fabric (3,5,7-stage CLOS)
- RAID-Controller in Storage Hardware (Config. Issues)

Software

- Lustre-1.6b7 and Lustre-1.6.3 (Bugs)
- OFED-1.1 and IPoIB Failover
- MPI Start-Up (Failed Processes and Scalability)
- TORQUE and `ulimit` Values

STREAM – Triad

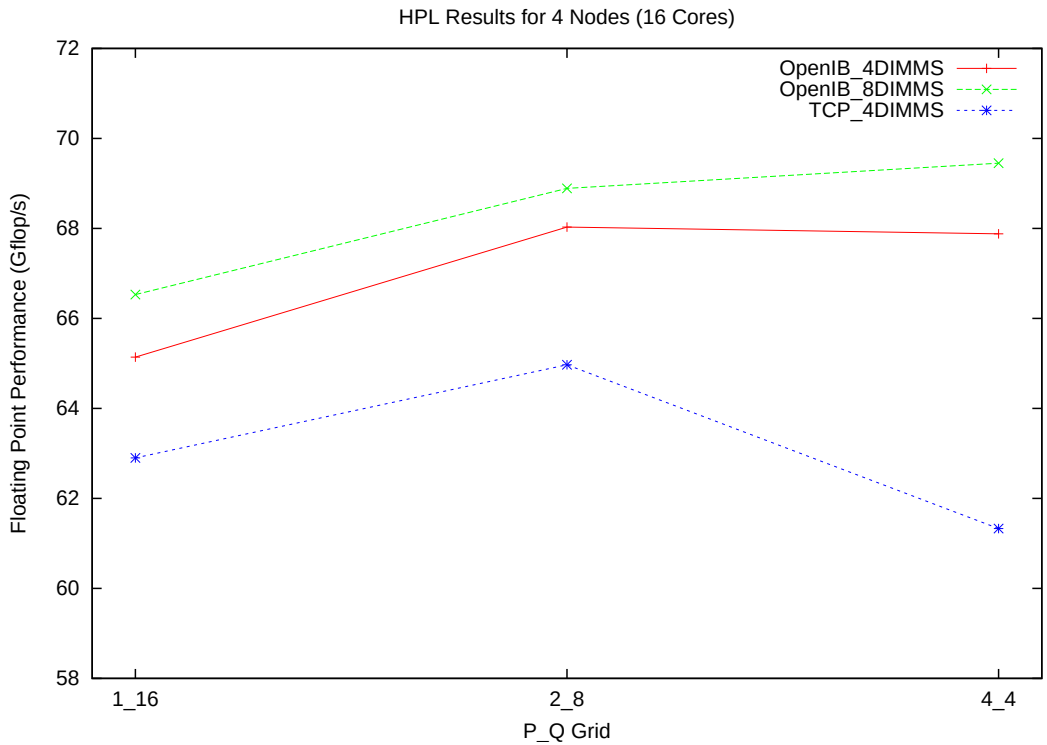
$$a[i] = b[i] + q \cdot c[i]$$

$$balance = \frac{peak\ floating\ ops/s}{sustained\ memory\ ops/s}$$

		pathscale-3.0			
		Opteron		Woodcrest	
		BW (MB/s)	Balance	BW (MB/s)	Balance
1 T	2 Ds	5655.7	7.3	3672.8	17.4
	4 Ds	5572.9	7.4	3896.4	16.4
	8 Ds	5769.8	7.2	3959.6	16.2
2 Ts	2 Ds	6056.0	13.7	3967.9	32.2
	4 Ds	6114.7	13.6	5061.7	25.3
	8 Ds	6520.9	12.7	5876.6	21.8
4 Ts	2 Ds	5025.1	33.1	3949.3	64.8
	4 Ds	11527.4	14.4	5111.2	50.1
	8 Ds	12796.4	13.0	5653.6	45.3

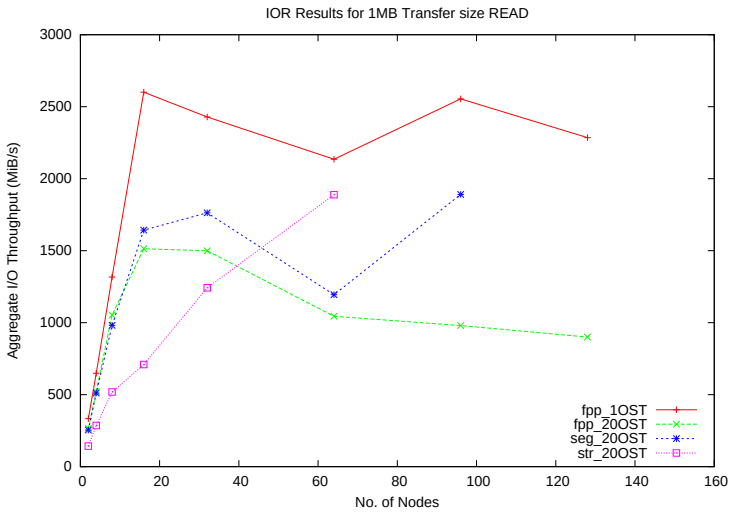
HPL

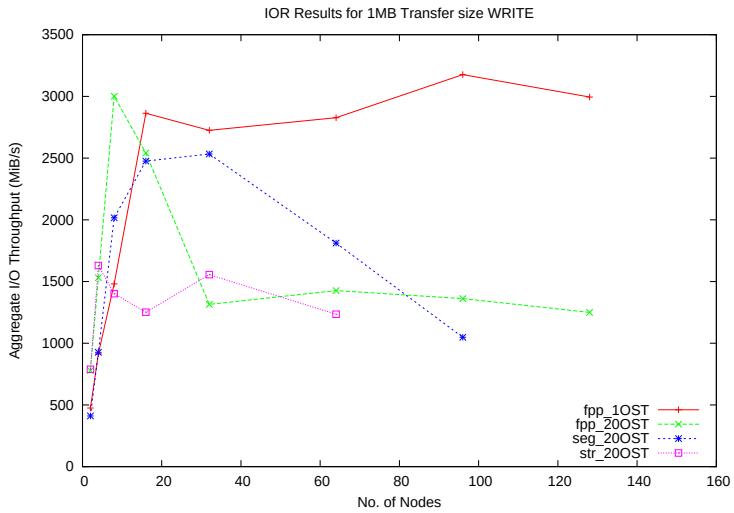
- 8.21 TFlop/s (76%) measured (2080 Cores)



IOR

- 20 Object Storage Targets (RAID-5 a 8 HDDs)
- 3.2 GiB/s Write Performance
- 2.6 GiB/s Read Performance





Latest IOZone Results

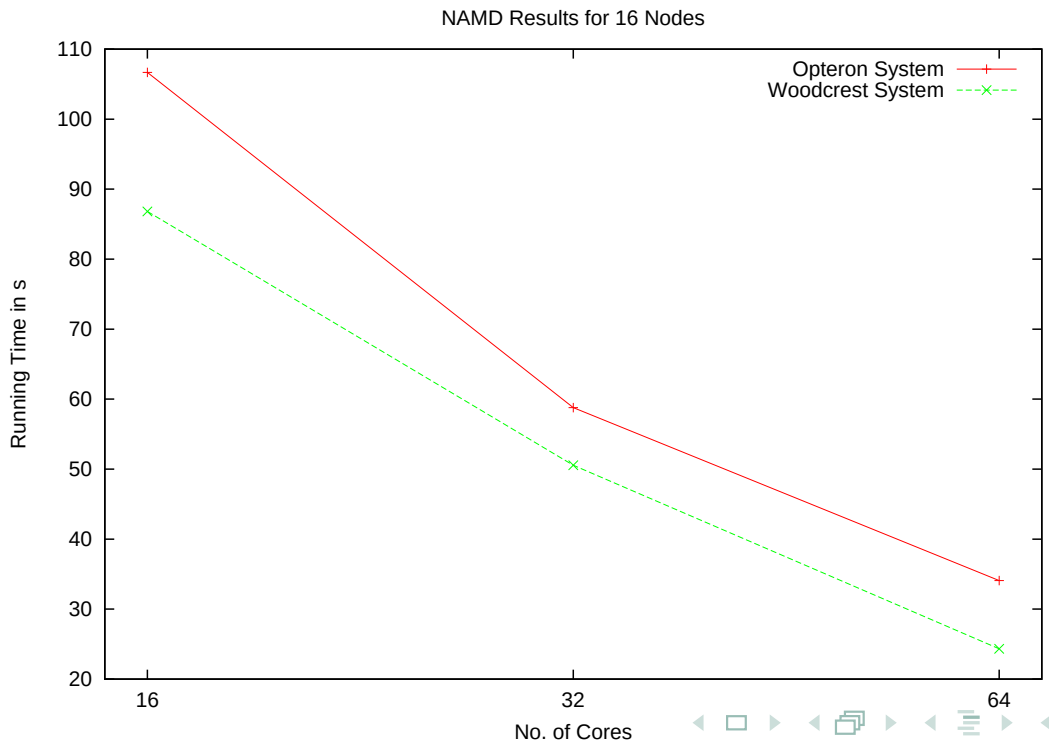
- 18 Object Storage Targets (RAID-6)
  - 9 RAID-6 with 10 HDDs
  - 9 RAID-6 with 6 HDDs
- 120 Clients (Lustre-1.6.3)
- 5GB Data File each
- **3.7GiB/s Read Performance**
- **3.2GiB/s Write Performance**

Application Benchmarks

ABINIT:

	AMD Cluster	Intel Cluster
Time in s	1,384.6	1,454.2

NAMD:



CHEMNITZ UNIVERSITY OF TECHNOLOGY

Summary

- Extremely Good Price-Performance Ratio Achieved
- Ambitious Project Deadlines (Compromises)
- Self-Design vs. Self-Made
- Performance Numbers of Intel/AMD Processor (Memory Bandwidth more important for us)
- Lustre Failover Configuration Expensive (Backup Strategy)

CHEMNITZ UNIVERSITY OF TECHNOLOGY

Thank You!

Any Questions?

CHEMNITZ UNIVERSITY OF TECHNOLOGY

Backup Slides

Software–Environment

- Scientific Linux 4.4 / 5.0
- Open Fabcris Enterprise Ed. 1.2
- Lustre 1.6.3 –> Lustre 1.6.4
- Open MPI 1.2.4, MVAPICH-1.0beta and MVAPICH2-1.0.1
- GNU Compiler 3.4.6 and 4.2.2, and EKOPath Compiler 3.1
- TORQUE 2.1.8 and Maui 3.2.6p13
- Nagios 2.9
- xCAT 1.2.0 and Warewulf 2.6

Cluster Installation

- 1 Month Deployment
- 21,6 Tons Material (Racks + Components)
- 4200 Nuts and 4600 Skrews necessary
- 4900 Cables with 9800 Connectors (8km Length)
- 300 Man-Days Effort