

Multi-Core Based Parallel Computing Platforms at the Univ. of Oslo

Ole W. Saastad
Scientific Computing group
USIT, UiO

A balanced compute resource

- CPU registers L1 cache - ns access time
- L2 cache - 10 ns access time
- Memory - 0.1 μ s access time
- Interconnect - μ s access time
- IO and storage – ms access time
- Identify and solve bottlenecks
- Optimize code to utilize the cluster hardware

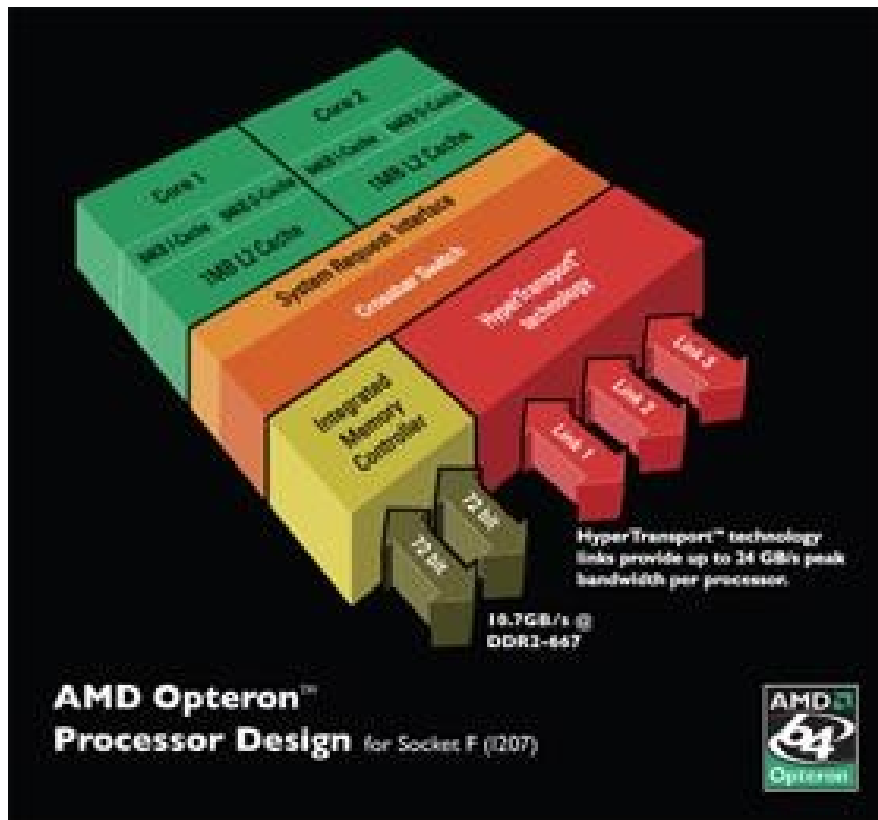
Multi core processors

- Dual core today
- Quad core in Q4
- Octa core in one year
- Hecto (or even more) cores in a few years
- 16 GBytes per node today, how much tomorrow ?
 - Larger problems can be challenged

Compute nodes

- AMD based dual core compute nodes
- Processor 2218
 - 2.6 GHz L2 cache 1M
 - 5.2 Gflops/s double prec.
- Each socket contains a processor with two CPUs / cores.

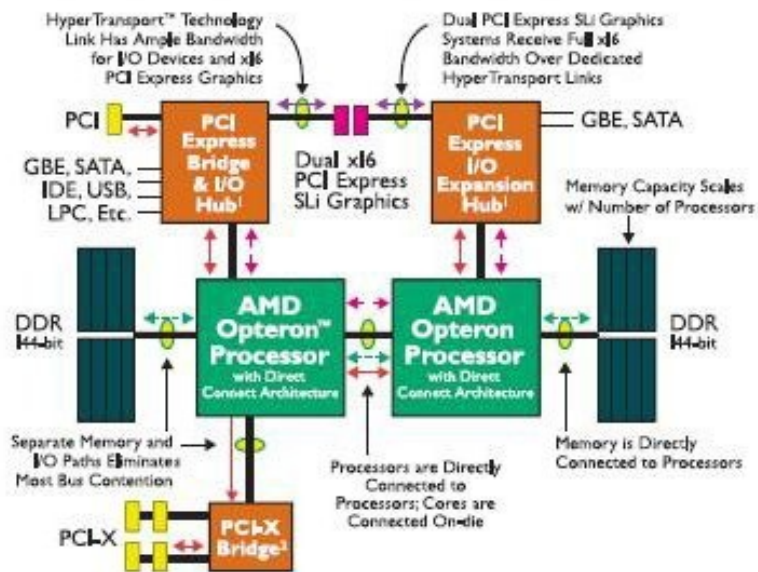
CPU = core



Compute nodes

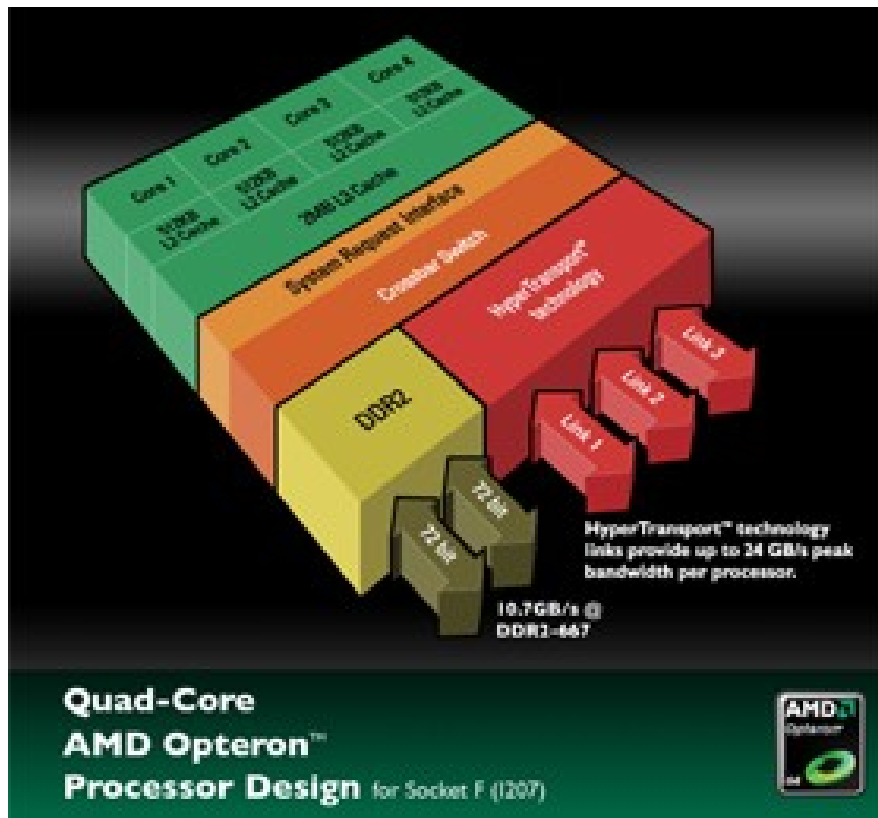
- Each node 16 GBytes main memory
- Local disk for OS
- 4 x Gigabit Ethernet
 - Only one is used
 - InfiniPath InfiniBand
 - Fraction of the nodes
- 21 Gflops/s per node

AMD Opteron™ Processor-based 2P Workstation

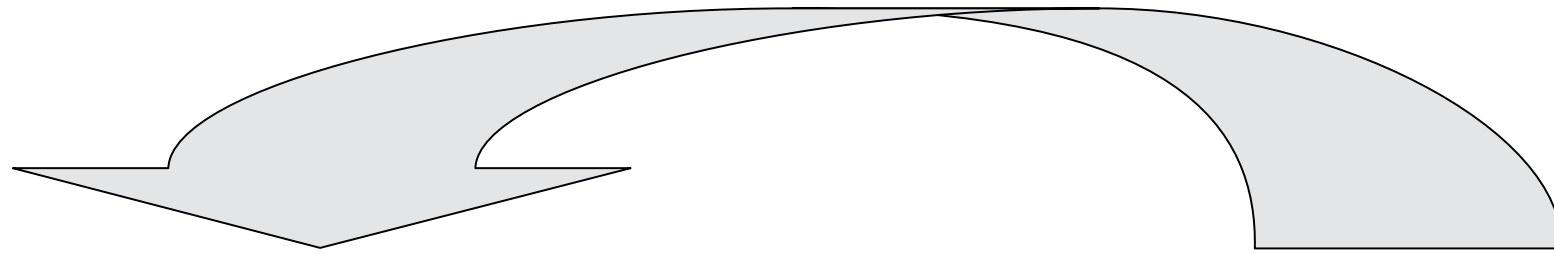


Compute nodes upgrade

- Upgrade to AMD quad core in near future
- Double # of cores/CPU's
- 2x Floating Point performance per CPU
 - 128-bit floating-point pipeline enhancements

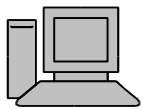


Clusters are complex



Perceived Complexity

Real Complexity



- CPUs
- I/O Bridges
- Interconnect
- OS
- Libraries
- Compilers
- MPI
- Management

PCs

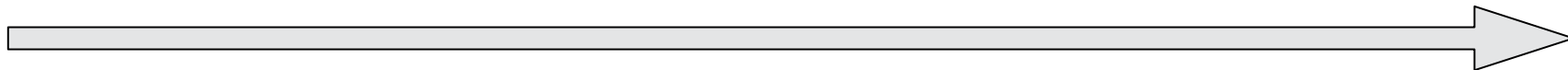
Clusters

Supercomputers

Clusters

Commodity

Complexity



Cluster of compute nodes

- 450 SUN Fire 2200m2 Dual socket AMD nodes
- Theoretical peak performance 9.36 Tflops/s (37.4 upg)
- 7.2 TBytes main memory, RAM
- Cisco 2960 Rack switches 8 links to core switch
- Cisco 6509 Core switch
- InfiniPath InfiniBand PCIe SDR cards
- Silverstorm 96 port DDR InfiniBand switch

Cluster of compute nodes

- Rocks cluster software
 - Centos operating system
- Ganglia cluster admin and monitoring software
- Intel, Pathscale, Portland and GNU compilers
- Sun Grid Engine batch queue system
- Scali MPI, MPICH2 and OpenMPI
- AMD Core Math Library (ACML), Goto BLAS

Cluster of compute nodes

Ganglia Cluster Toolkit: Cluster Report

<http://titan.uio.no/ganglia/?c=TITAN%20II&m=&r=hour&s=descending&hc=4&jr=-3600&js=1193...>



Cluster Report for Tue, 23 Oct 2007 10:10:00 +0200

Get Fresh Data



Metric Last Sorted

Physical View

Grid > TITAN II >

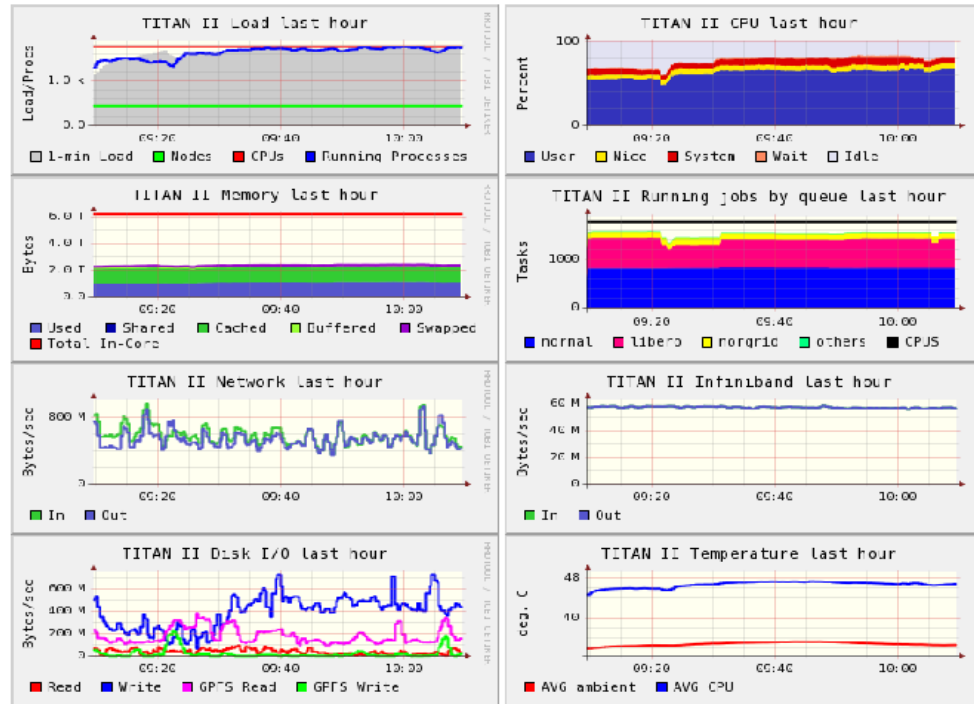
Overview of TITAN II

CPUs Total: 1796
 Hosts up: 442
 Hosts down: 7
 Max CPU temp: 69 °C
 Max Amb temp: 46 °C

Avg Load (15, 5, 1m):
 95%, 97%, 97%
 Localtime:
 2007-10-23 10:09

Rocks Tools:
[Job Queue](#) | [Cluster Top](#) | [Gmetrics](#)

Cluster Load Percentages



Show Hosts: yes no | TITAN II load_one last hour sorted descending | Columns

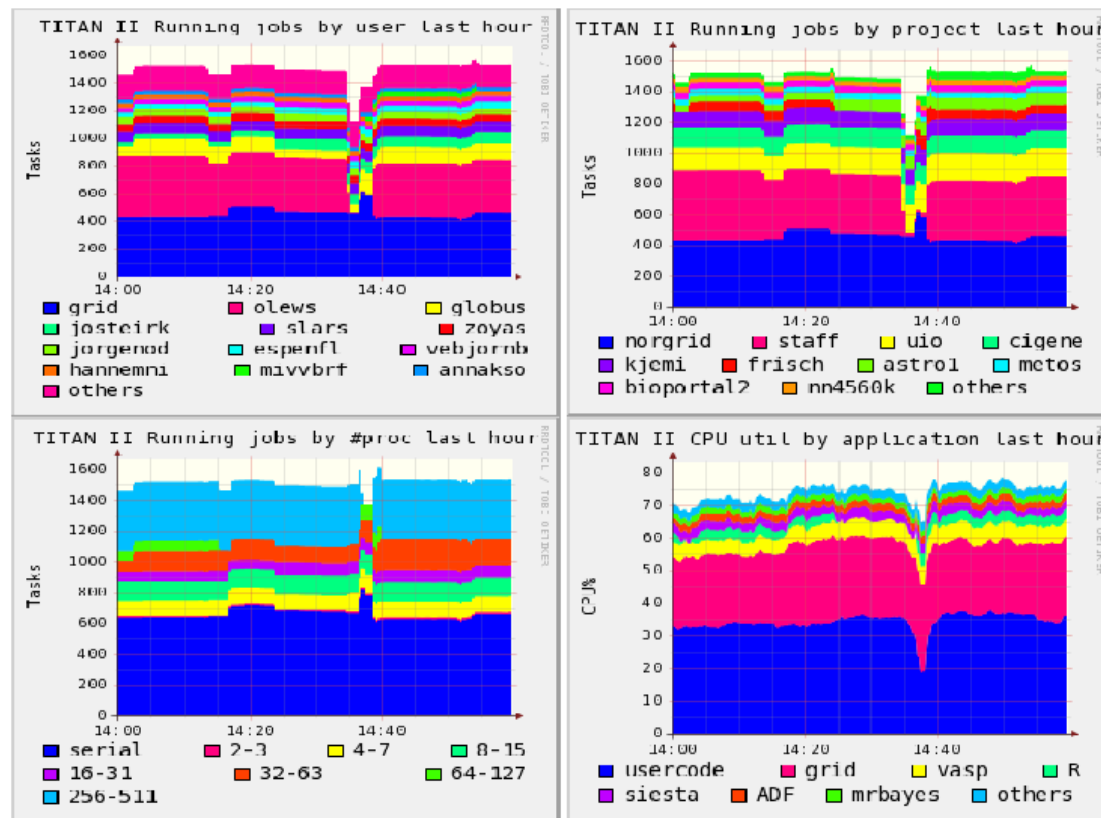
Cluster queue system

Job Queue

<http://titan.uio.no/ganglia/addons/rocks/queue.php?c=TITAN%20II&r=1>

Job Queue

Tue, 23 Oct 2007 14:59:57 +0200 Physical Job Assignments



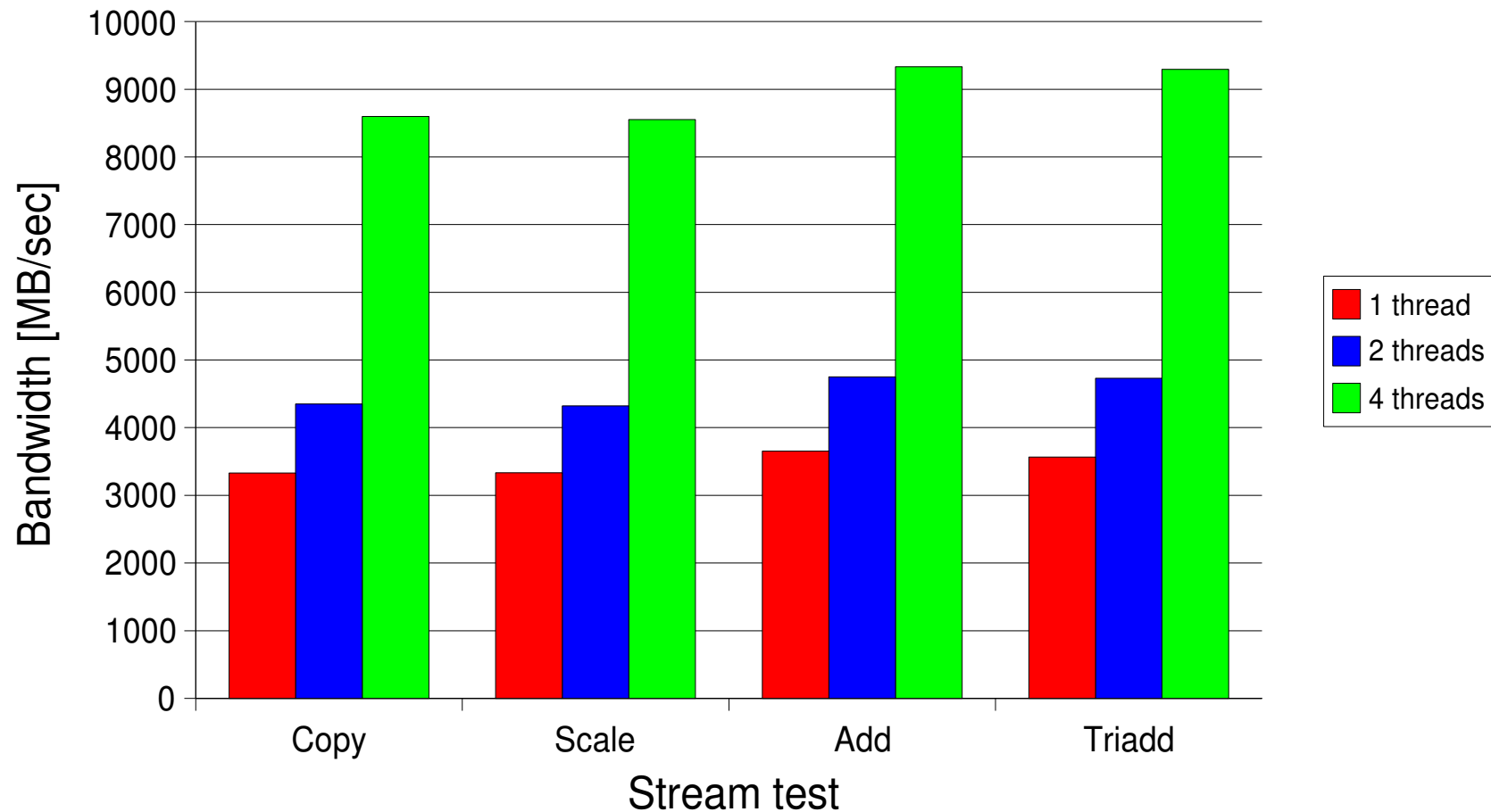
Performance – single CPU

```
=====
T/V                N    NB    P    Q                Time                Gflops
-----
WR10R4C4          10000  360    1    1                141.39                4.716e+00
-----
||Ax-b||_oo / ( eps * ||A||_1 * N ) =          0.1036849 ..... PASSED
||Ax-b||_oo / ( eps * ||A||_1 * ||x||_1 ) =       0.0245349 ..... PASSED
||Ax-b||_oo / ( eps * ||A||_oo * ||x||_oo ) =       0.0054841 ..... PASSED
=====
```

4.716 Gflops/s is 90% of theoretical peak performance (Goto BLAS library).

Node Performance - memory

Stream memory Bandwidth



Interconnect performance

- InfiniBand latency 1.68 microseconds !

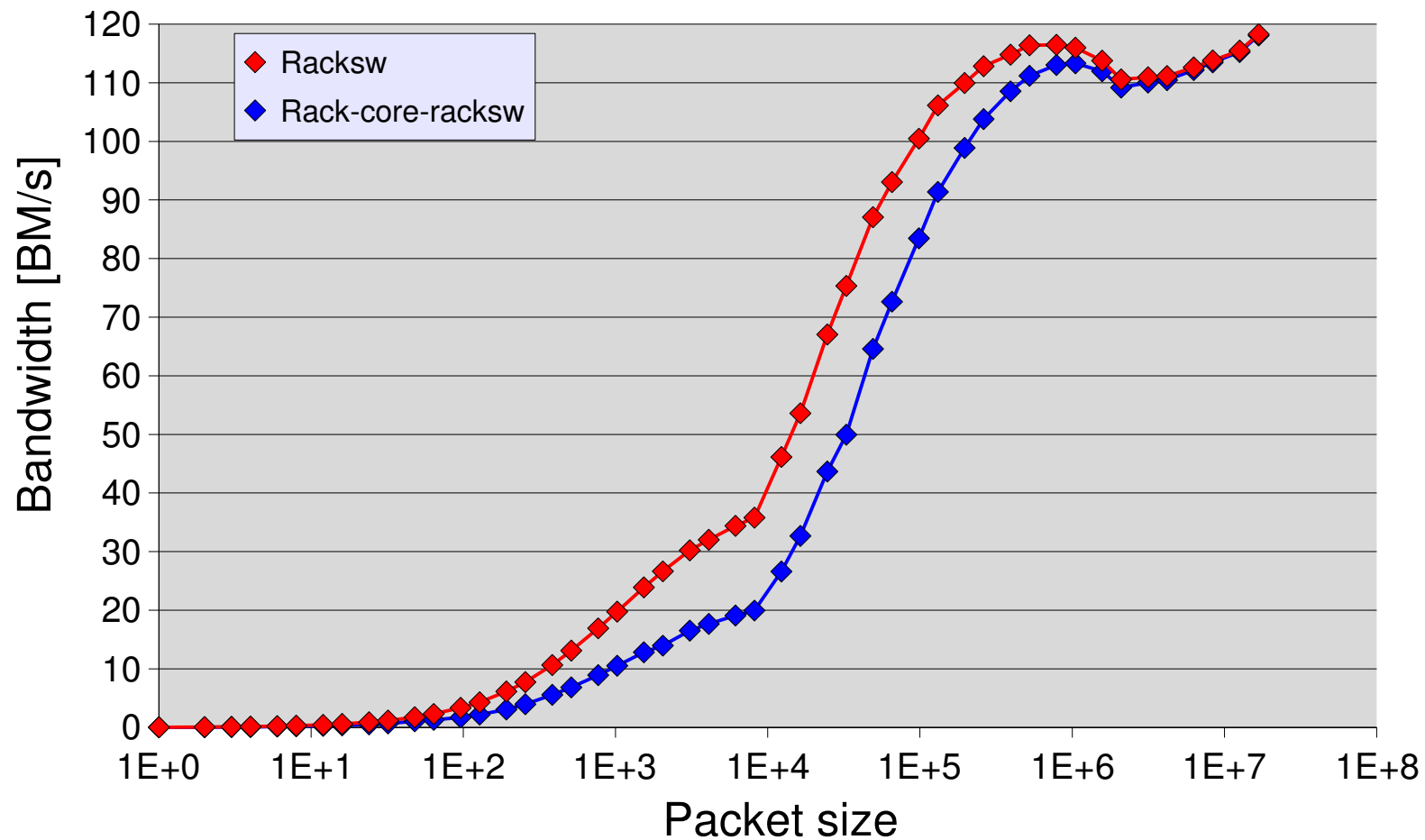
Benchmark ping-pong

=====

length (bytes)	iterations (count)	elapsed time (seconds)	transfer rate (Mbytes/s)	latency (usec)
0	29579	0.100	0.000	1.682
1	30801	0.103	0.599	1.669
2	30801	0.104	1.189	1.682
3	30452	0.102	1.787	1.679
4	30731	0.103	2.390	1.674
6	30452	0.102	3.574	1.679
8	30731	0.103	4.781	1.673

Ethernet performance

Bandwidth benchmark



Top500 Linpack perf.

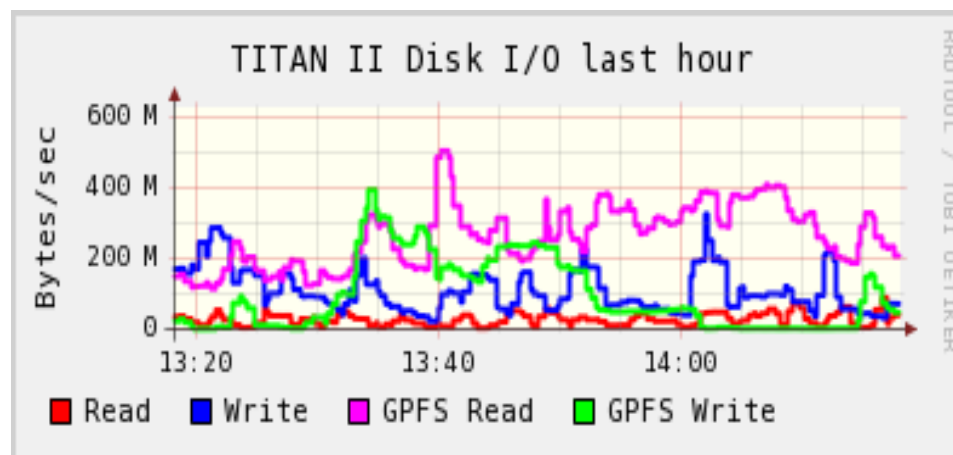
High perf. linpack results:

Params	size	block	nxm	time	Gflops	%peak
WR10R4C4	330000	360	8x32	24913.45	961.700	70.6
WR10R4C4	280000	360	6x32	19067.15	767.500	75.1
WR10R4C4	280000	360	12x16	19265.99	759.600	74.4

Preliminary results, only part of cluster (64 nodes) has been used and only Gigabit Ethernt.

IO subsystem and storage

- General Parallel File System
- 8 x IO nodes HP DL385
- 2 x HP Enterprise Virtual Array 8000 w. 240 Fibre ATA disks in each



IO subsystem and storage

- Performance
 - IOZone show performance close to wire speed

