Building 96-processor Opteron Cluster at Florida International University (FIU)

January 5-10, 2004

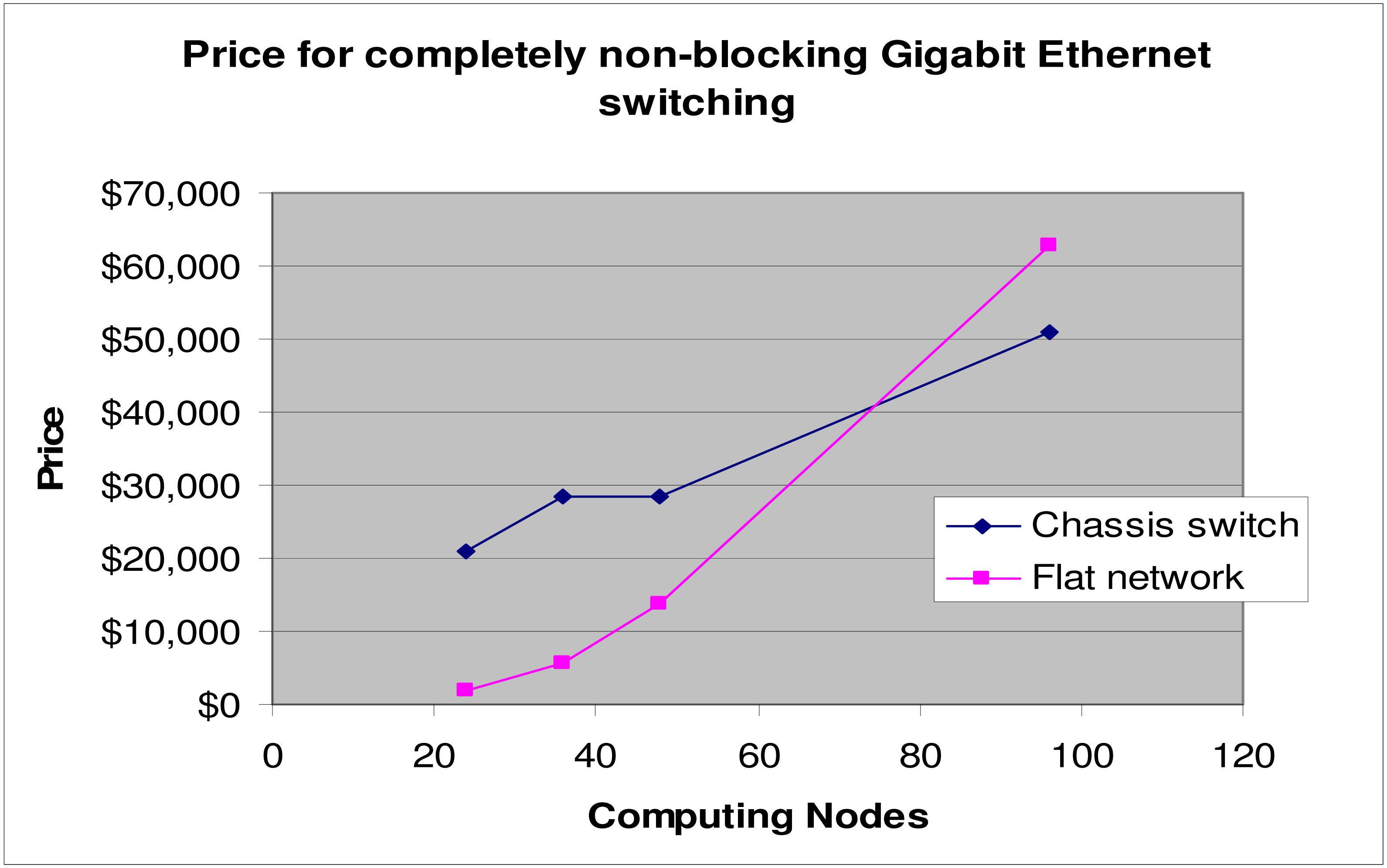
Brian Dennis, Ph.D.
Visiting Associate Professor
University of Tokyo

Designing the Cluster

- Goal: provide a cluster that performs well for a wide variety of engineering applications for less that $100,000
- Supports researchers in Mechanical Engineering department of FIU
- Applications: design optimization, molecular dynamics simulations, large-scale electromagnetics and fluid dynamics simulations, Lattice-Boltzman methods
- A vendor offered a discounted Xeon system with gigabit Ethernet networking for $155,000
- Analysis indicated that self-built system would result in a higher performance/price ratio

Gigabit Ethernet Options

- The cost of the gigabit network switching became the first factor limiting the size of the computer



Networking

- Analysis indicated that 48 dual processor machines networked by Fast Ethernet and Gigabit Ethernet would be the best solution for $100,000

- Gigabit Ethernet for more than 36 machines was determined to be too expensive or too complicated

- It was decided to use a 48-way Fast Ethernet switch to connect all machines

- A 36-way flat network would be used to connect 36 machines by gigabit Ethernet

- For some applications, i.e. optimization, Fast Ethernet is good enough

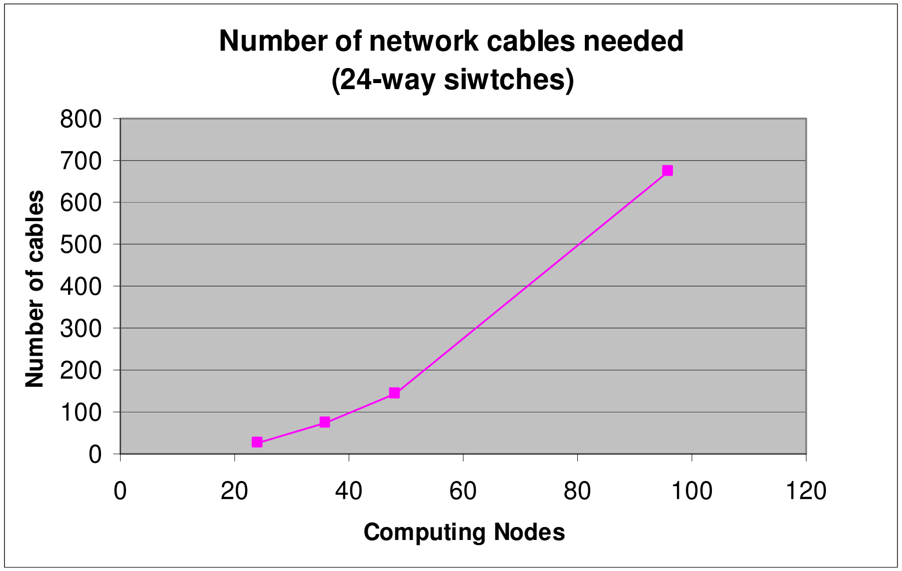
- With dual processor computers, a total of 96 processors on the fast Ethernet and 72 processors on gigabit Ethernet

Flat Networking

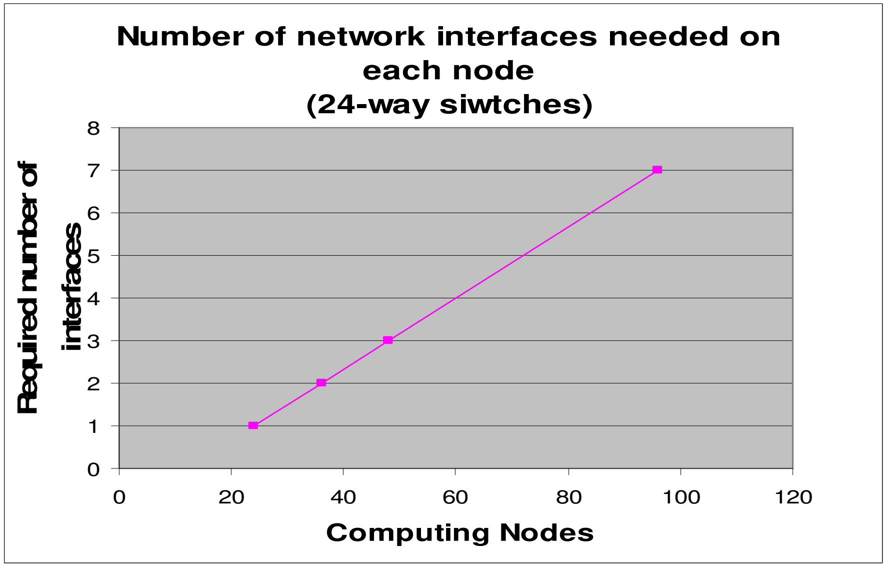
- Flat networking is a way to build a cheap non-blocking gigabit switching network using multiple small switches
- The cluster with world’s largest performance/price ratio uses flat networking
- Non-blocking with only one level of packet routing latency – better than trunking multiple switches into a tree
- Multiple network interfaces are needed in each machine
- Best to use as large a switch as is economically possible (currently 24-way switches) to minimize number of network interfaces
- Linux networking is easily configured to automatically route messages to the appropriate interface

Flat Networking

- Large number of network cables required – complicated to assemble



- Flat networks with small switches don't scale well to large systems

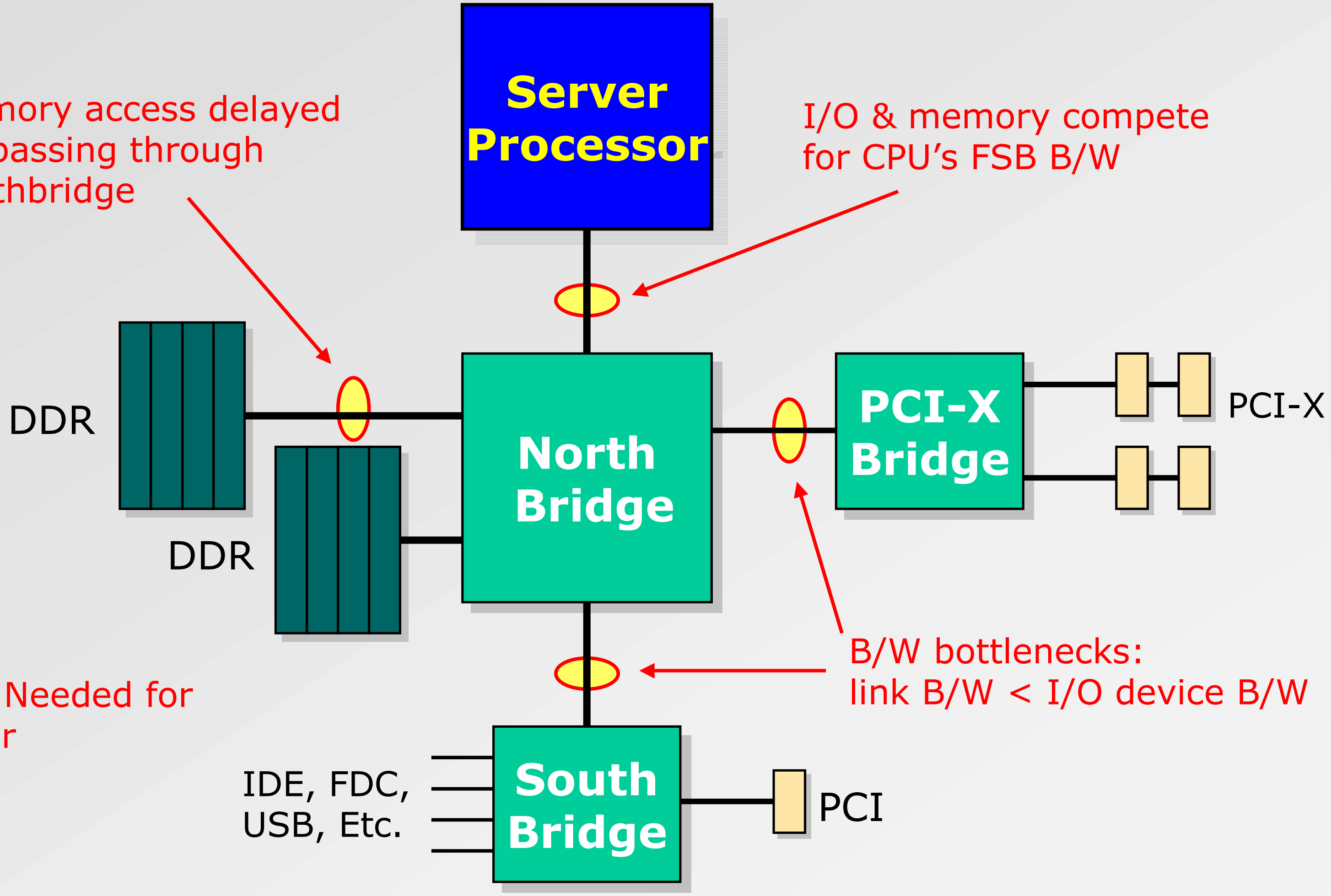


Opteron features

- 1 MB L2 cache
- 64-bit instruction set
  - Twice the memory bandwidth for 64-bit floats
  - Twice the registers (16 SSE vector registers vs. 8 on Intel)
  - Can access over 4 GB of memory
- NUMA architecture – memory bandwidth/latency scales with number of processors
- High bandwidth integrated memory controller with separate memory and I/O paths to eliminate contention
- Integrated memory controller runs at processor speed, not at FSB speed – larger memory bandwidth
- Can run existing 32-bit executables with no recompile

Legacy Northbridge Architecture

Typical System



Memory access delayed by passing through northbridge

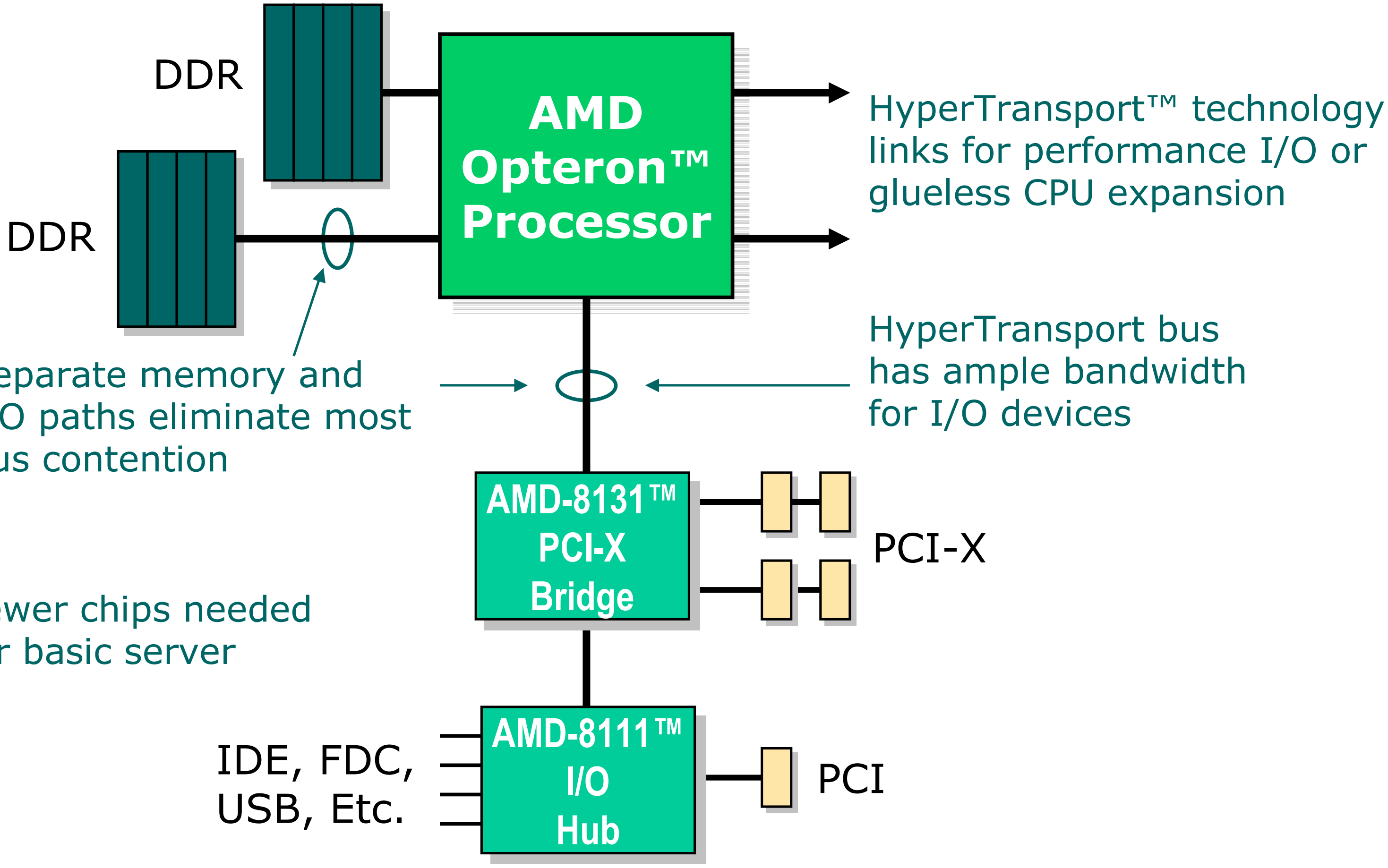
I/O & memory compete for CPU’s FSB B/W

B/W bottlenecks: link B/W < I/O device B/W

More Chips Needed for Basic Server

AMD Opteron Processor

System Architecture



HyperTransport™ technology links for performance I/O or glueless CPU expansion

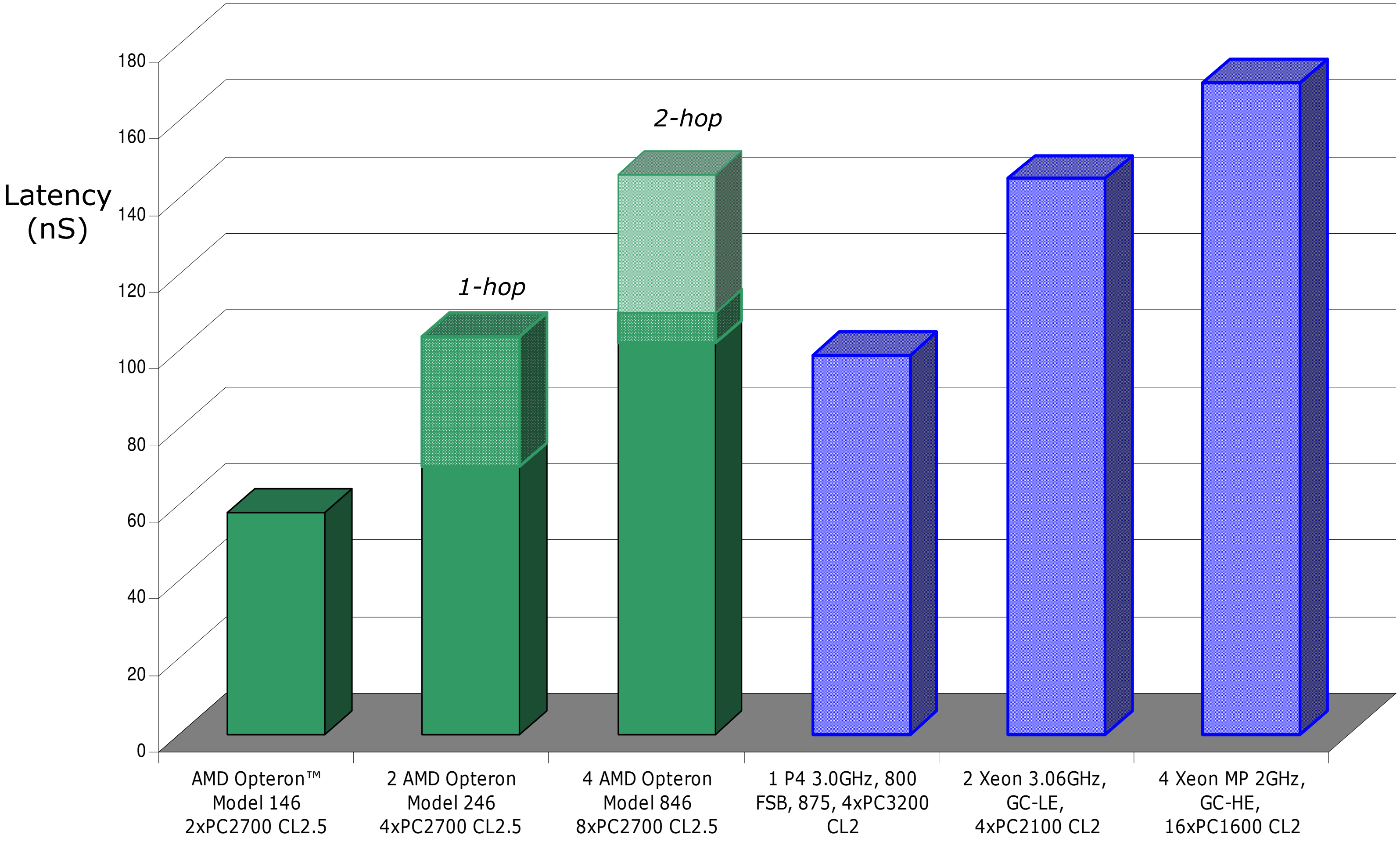
Separate memory and I/O paths eliminate most bus contention

HyperTransport bus has ample bandwidth for I/O devices

Fewer chips needed for basic server

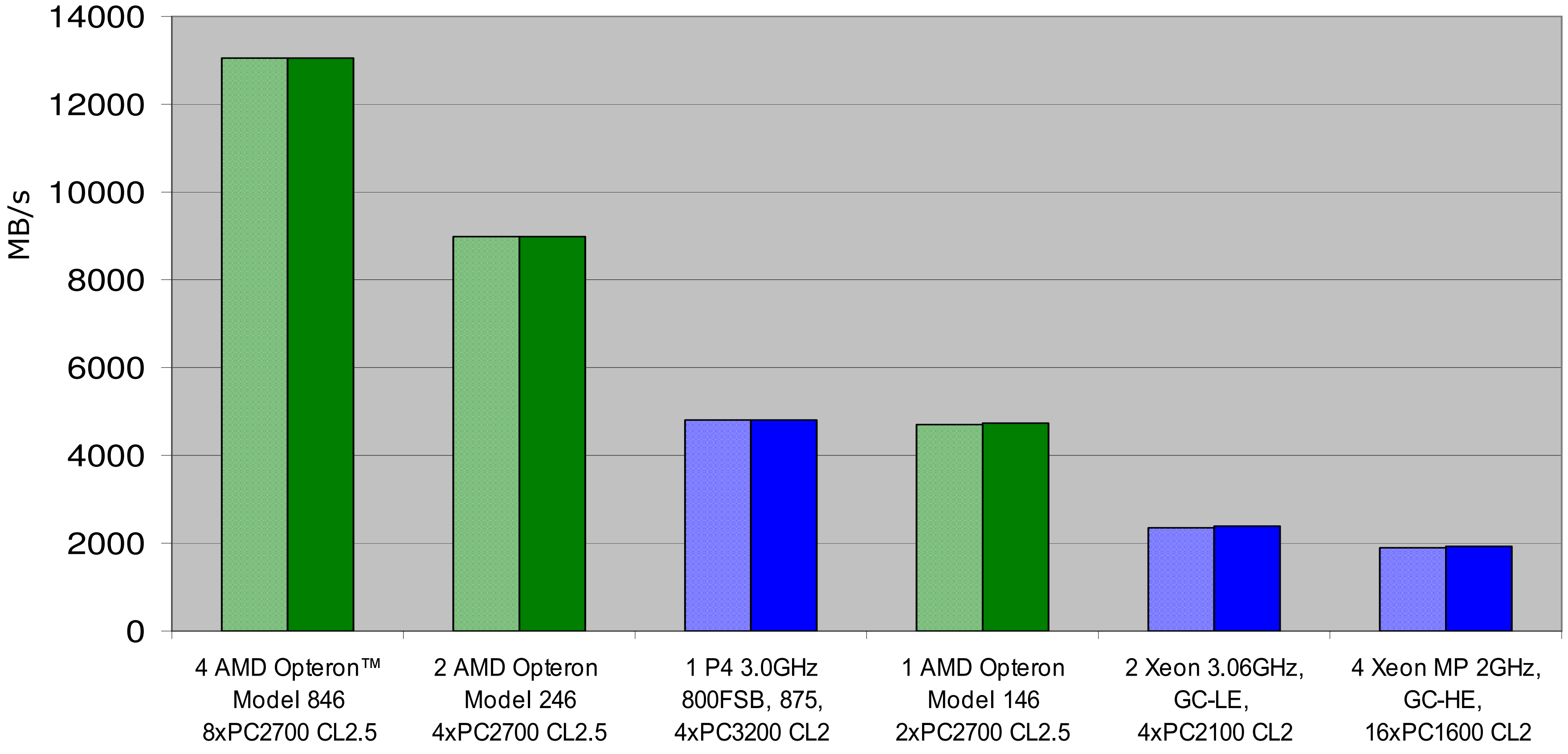
Low Memory Latency

ScienceMark 2.0 Beta, 512-Byte Stride



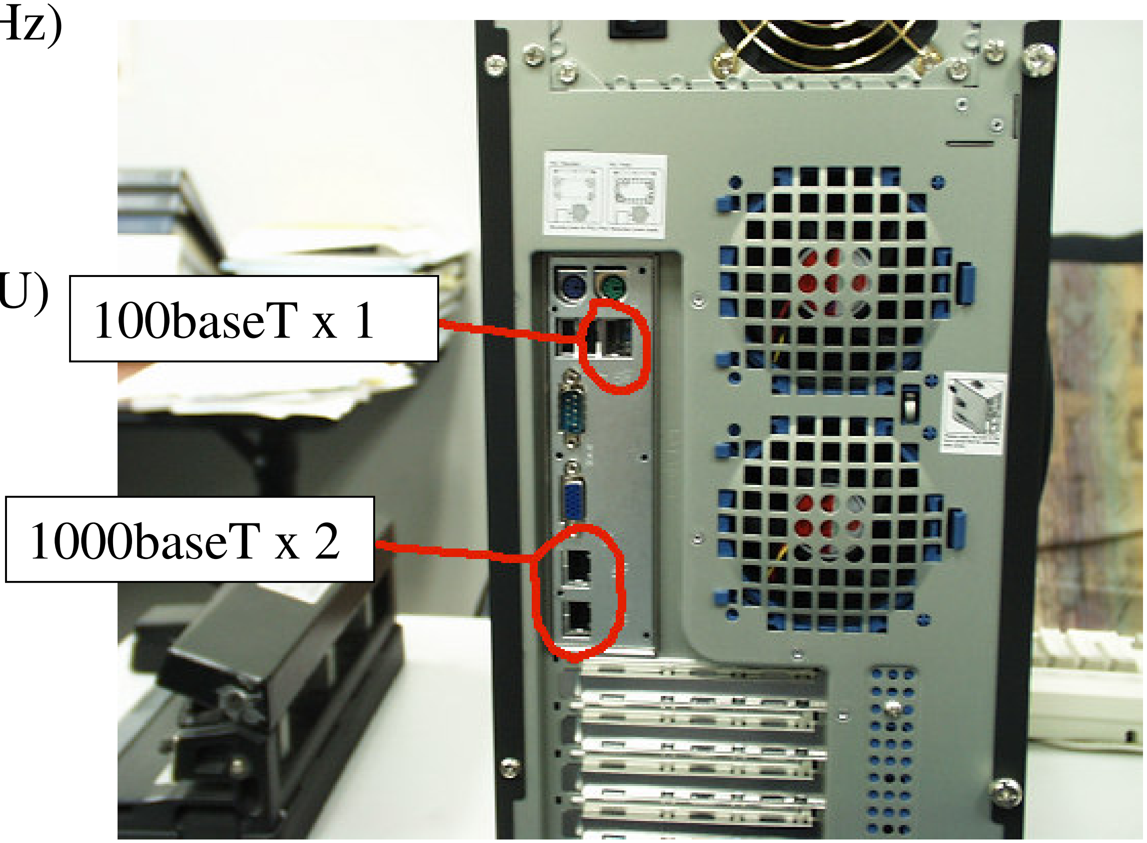
Scalable Memory Bandwidth

Sisoft Sandra Standard 2003



Computing Node Hardware

- Dual Opteron computers were chosen for the computing nodes
- Tyan Dual - Opteron S2882 Server Board was selected
  - -Dual Broadcom 1000baseT adapter on PCI-X bus
  - -Single Intel(R) Ether PRO/100 100baseT adapter
- AMD Opteron 242 CPU (1.6GHz)
- 40GB EIDE Hard Drive
- 1GB Registered ECC DDR PC-2700 (512 MB for each CPU)



100baseT x 1

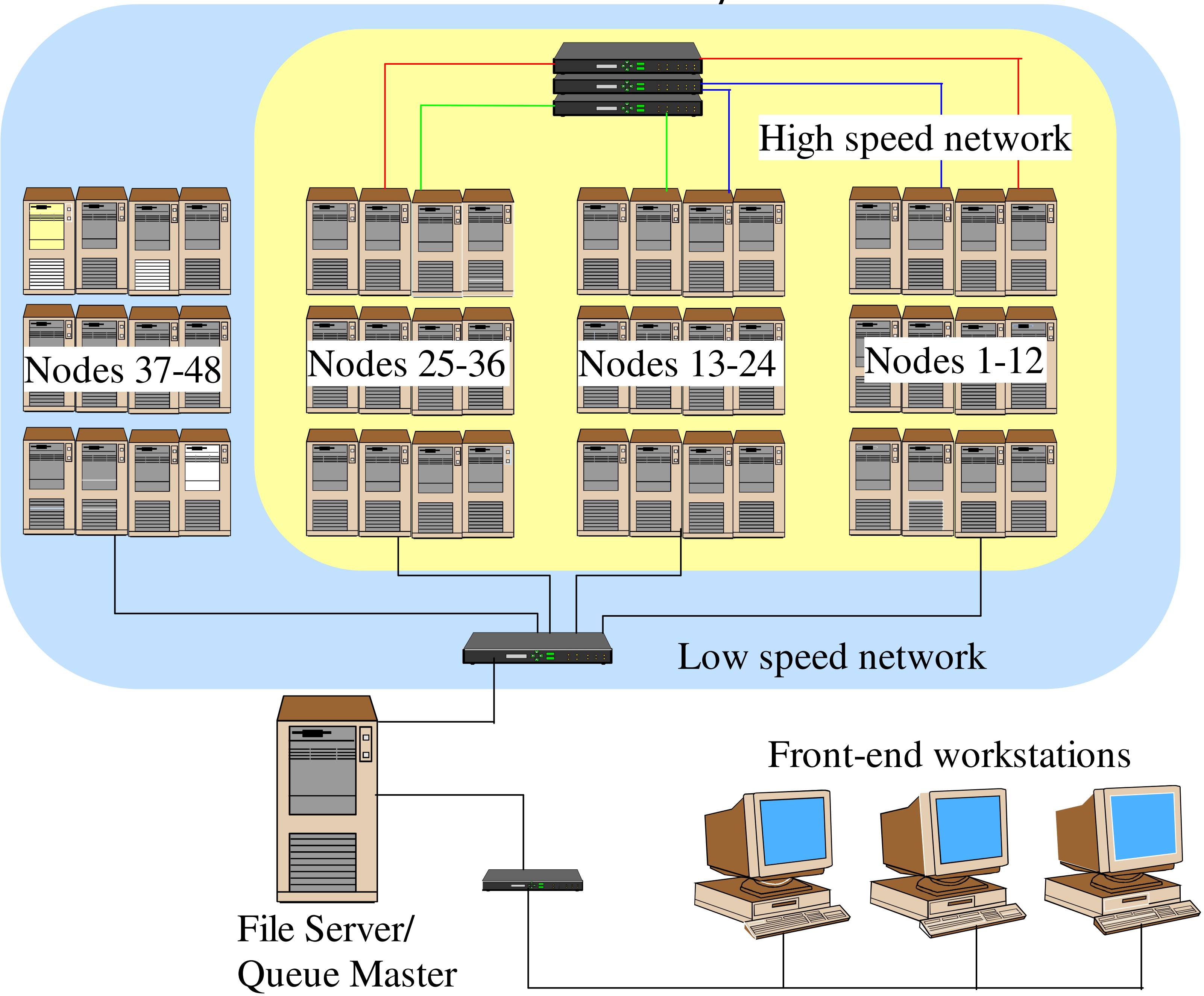
1000baseT x 2

Final Price List

Type	Quantity	Detailed Description	Unit price
Computing node	48	Dual Opteron	$1750
File server	1	Dual Xeon SCSI	$3200.00
Front-end	1	Opteron workstation with SUSE Linux	$2584.00
15’ cables	120	15 foot Cat 5e non-booted network patch cables	$2.0
Keyboard, Monitor, mouse	2	PS/2 keyboard, 3-button mouse, monitor for each front-end and server	$0
Compiler	1	PGI Workstation 64-bit/32-bit Node-locked single-user Academic license	$699
switch	1	48 100base-T with 2 1000base-T RJ-45 ports (13.6 gigabits/sec capacity)	$903.70
switch	3	24 1000base-T RJ-45 ports (48.0 gigabits/sec capacity)	$1901.35

- Total cost: $98,000
- Networking is 6% of total hardware cost

Final Layout

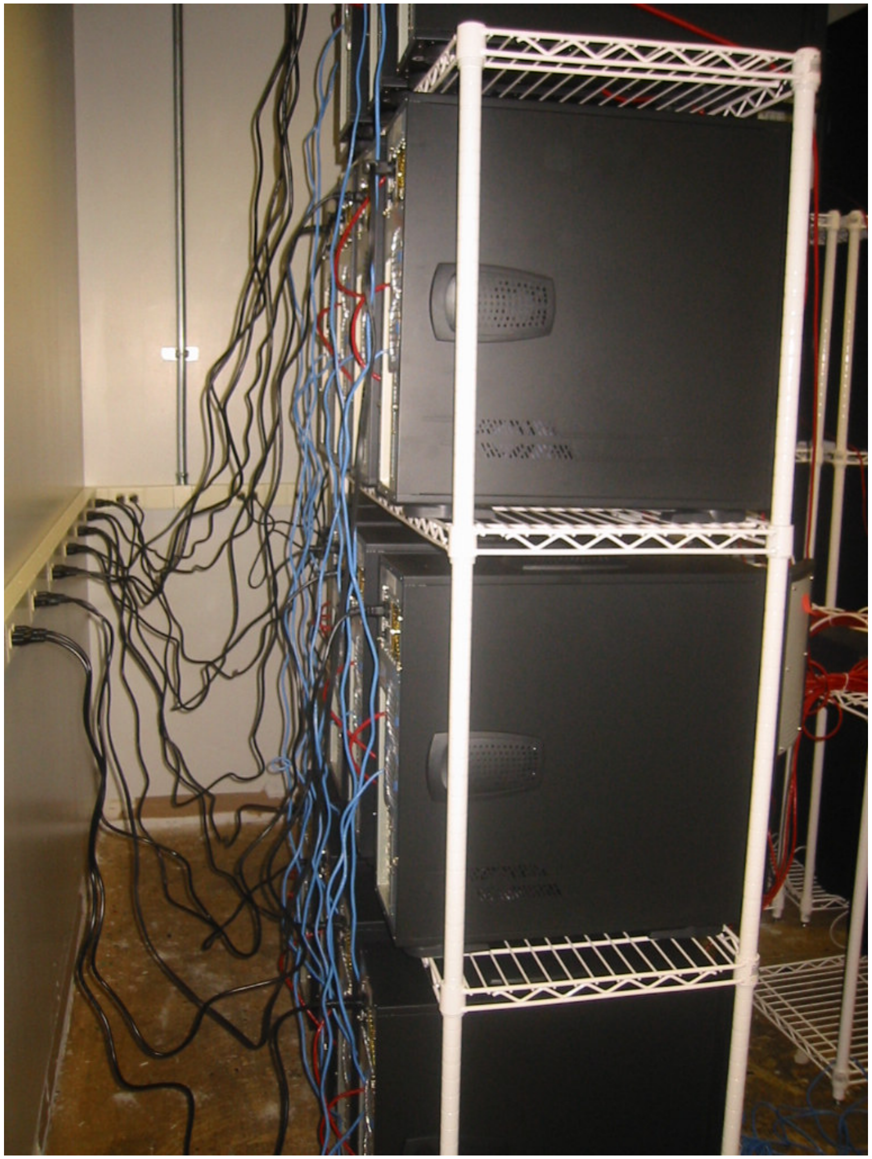


Assembly

CS-10182-BA
SUPER ATX TOWER
CS-1018
ENERMAX
HEAVY DUTY BEST VENT!
SUSE
CS-1018
SUPER ATX TOWER
CS-10182-BA
CS-10182-BA
ENERMAX
CS-1018
SUPER ATX TOWER

Assembly

Assembly



Assembly

Operating System

- 64-bit Linux is needed to get full benefits of Opteron processors
- In winter 2003 there were no *free* non-beta 64-bit distributions
- Suse Enterprise edition chosen for front-end machine
- Redhat “Taroon” beta distribution for computing nodes
- Redhat 9.0 for Xeon-based file server
- Operating system was copied from disk to disk using ‘dd’ command. Entire process took 10 hours.

```
“dd if=/dev/hdb of=/dev/hdd bs=512MB”
```

TIP: using a block size of 512MB reduced a single disk copy from 2 hours to 20 minutes.

Infrastructure Problems

- Room needs new air conditioning unit
  - Each machine dissipates 100-200 Watts
  - Computer room is small
  - Temperature sensors will be used to monitor thermal performance
- Special power supply needed
  - Each machine requires nearly 1.5 Amperes at full load
  - Several strange problems with some nodes traced to low quality power

Infrastructure Problems

- construction not completed on time. Machine had to be assembled then disassembled so construction could be completed
- Shelves are needed due to small space. Computers are heavy (30 kg) so not easy to move around.

Other Problems

- Trouble with 64-bit Linux
  - Not all packages are available for Redhat Taroon Linux
    - rpc server (rlogin, rsh, rcp) not available
  - GCC cannot produce 32-bit binaries
  - Must use commercial Linux on front-end machines
  - Kickstart with PXE doesn’t work
- Time
  - I had limited time for assembly (Jan. 5-Jan. 10 – 5 days)
  - Cabling for flat network is very complex.

    TIP: label each network cable on both ends before attaching