Recent experience in buying and configuring a cluster

John Matrow, Director

High Performance Computing Center

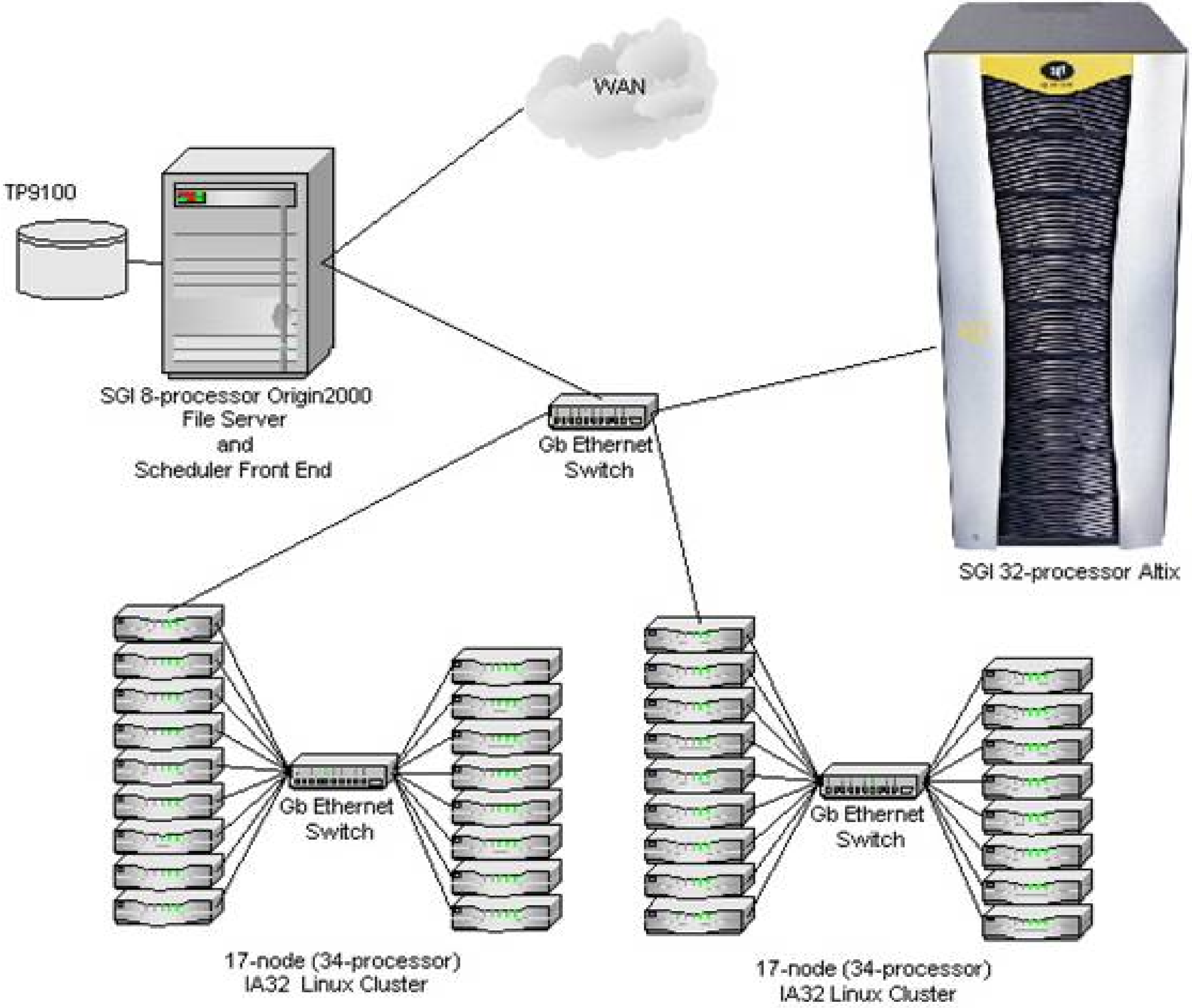
Wichita State University

OR

How I spent my summer vacation

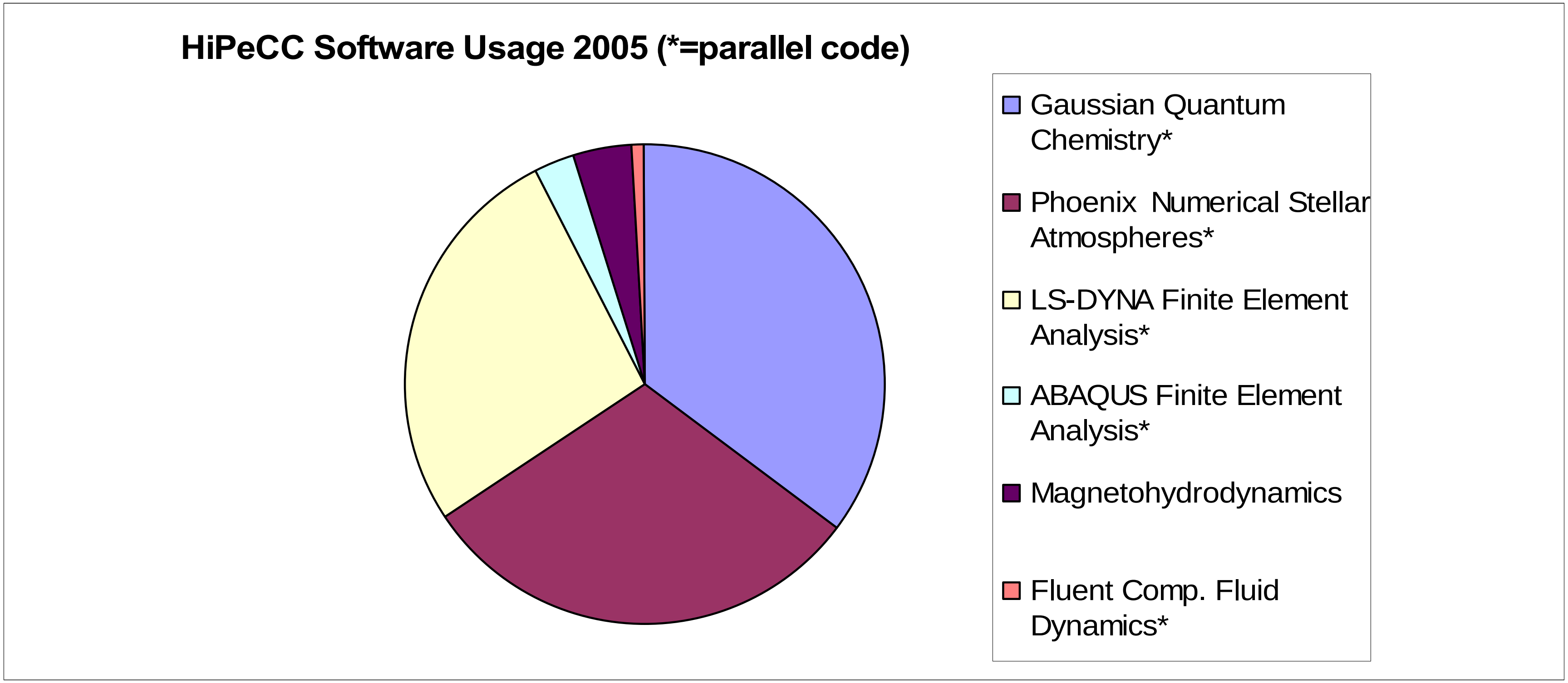
2003 New Configuration

- 32p SGI Altix (1.3GHz Itanium2)
- 34p Xeon cluster (2.666 Xeon)
- 34p Xeon cluster (2.666 Xeon)
- 100GB RAM
- Gigabit Ethernet
- SGI O2000 scheduler (PBSPro) and 1TB NFS file server (user directories)


WAN
TP9100
SGI 8-processor Origin2000
File Server
and
Scheduler Front End
Gb Ethernet
Switch
SGI 32-processor Altix
Gb Ethernet
Switch
Gb Ethernet
Switch
17-node (34-processor)
IA32 Linux Cluster
17-node (34-processor)
IA32 Linux Cluster

sgi

Primary Applications



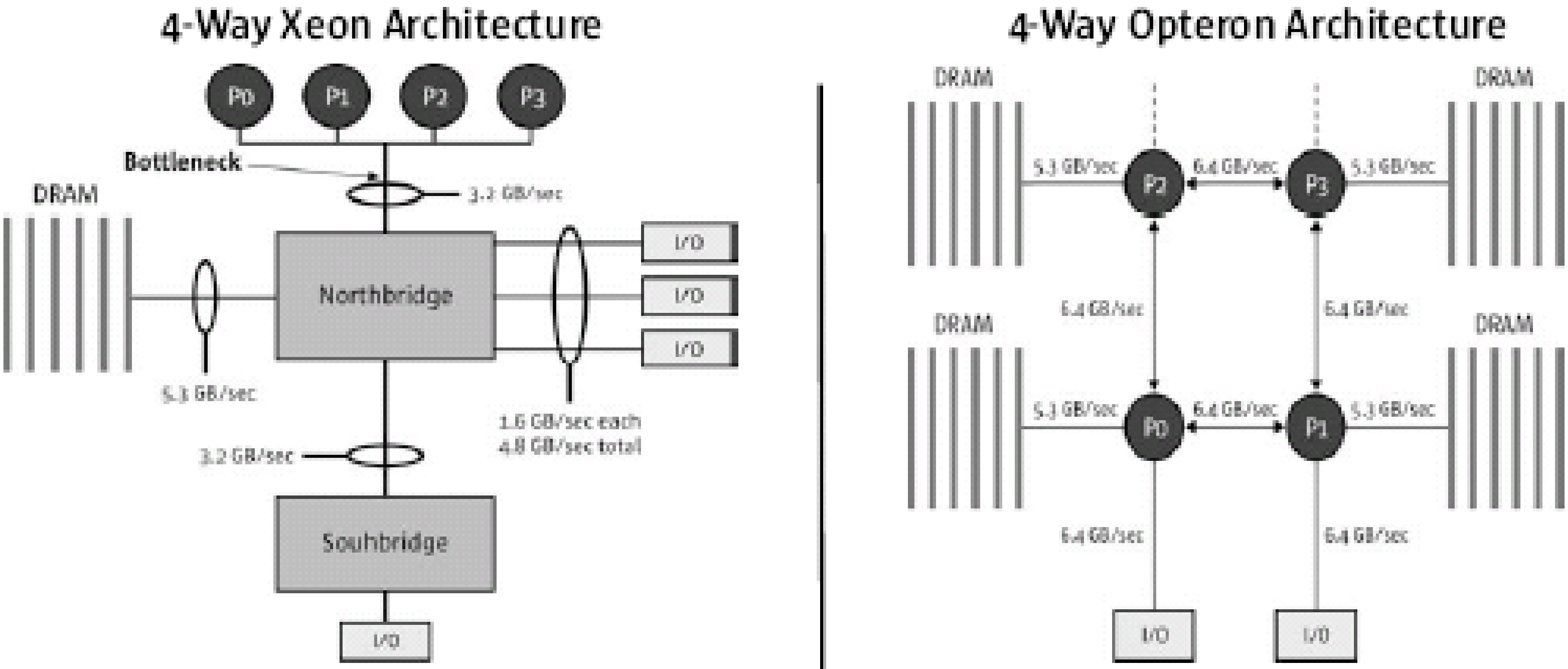
HiPeCC Software Usage 2005 (*=parallel code)

Professor X

- $125K grant for cluster for one application
- Went for bid with minimal specification
- Bidding cancelled and bids never opened
- Preferred state contracts to bids
- Feb. 2005: First quote of 64p (3.2) for $136K
- Best Opteron (2.4): 58p for near $100K
- Best Xeon EM64T (3.6): 92p for near $100K

Intel® Extended Memory 64 TechnologyΦ (Intel® EM64T)

- address more than 4 GB of both virtual and physical memory
- 64-bit flat virtual address space
- 64-bit pointers
- 64-bit wide general purpose registers
- 64-bit integer support
- Up to 1 terabyte (TB) of platform address space


4-Way Xeon Architecture

4-Way Opteron Architecture

Figure 2-3. 4-way Opteron architecture vs. 4-way Xeon architecture (also known as Northbridge/Southbridge architecture)

Texas Advanced Computing Center

The Intel Xeon memory architecture provides good memory bandwidth to a **single processor** from all levels of the memory hierarchy.

The higher CPU clock speed and slightly better bandwidth from cache to the registers relative to the AMD Opteron provide a performance edge to **computationally demanding** applications.

However, the limitations of the shared bus memory architecture become apparent when executing two memory intensive processes in parallel on a **dual-processor** node.

Because bandwidth is shared, per-processor throughput will **decrease** for memory bound applications with synchronous memory access needs.

The effect of memory contention, and the resulting degradation in performance, is **especially bad** for random or strided memory access.

Texas Advanced Computing Center

The AMD Opteron processor has a slower clock rate than the Intel Xeon, and consequently will **not** perform as well for **compute-bound** applications.

Because of the memory architecture and (resulting) excellent scalability of the memory subsystem, the AMD Opteron node is best suited for **memory intensive applications** that are not cache-friendly, and/or have random or stride access patterns.

Bottom Line: SW Vendor

“Ran some 1p and 2p jobs (on one node), and indeed the Opteron is 2x faster on 2 processors. Traditionally we have only gotten something like 60-80% out of the second processor on Intel’s. So even if the per cpu speed is close, the Opteron should outperform the Intel by 25% ish (2/1.6). So a 29p Opteron could perform as well as a 36p Xeon.”

BUT, ...

SW Vendor

“A lot of the logic behind Opteron being better than Irwindale is extrapolation from slower/older processors. We've never run on more than 1 Opteron node, but we have run on 2040 Irwindale processors as of last week. The Intel option may not be the best price/performance wise, but I’ll bet it’s close, and very safe.”

Specification

- 3 year warranty
- Pull-out KVM
- 4GB RAM/node
- Redundancy in head node
- Installation/Training
- Console Server
- GANGLIA

Other costs

- Disks for NFS Server: $4330
- PBSPro license: $50/p
- CFD Software: $1000 + (#p-24)*$200
- Gridgen Software: $6000
- Power cables + installation: $1000

Order/Arrival

- PO faxed April 26
- Power cable/receptacles ordered
- BTU requirements calculated
- First nodes arrive May 17
- Counsel reviews Statement of Work
- By June 2: 65 boxes
- Power cable/receptacles arrive June 22. Ready for rack assembly.
- Conference call July 12
- Dell begins assembly July 18

Assembly

- Side panels missing
- 20’ Cat 5E cables but GigE likes Cat 6E
- 48 port switch had year old firmware
- Backup PS for switch bad
- No console server
- Excellent assembler person
- Responsive vendor

Leftover

- Management node
- 48 port switch
- 90 CAT5 cables
- 4 Outlet boxes and 61 power cords
- Pair of rack side panels
- 45 unopened copies of RHEL 3

ROCKS

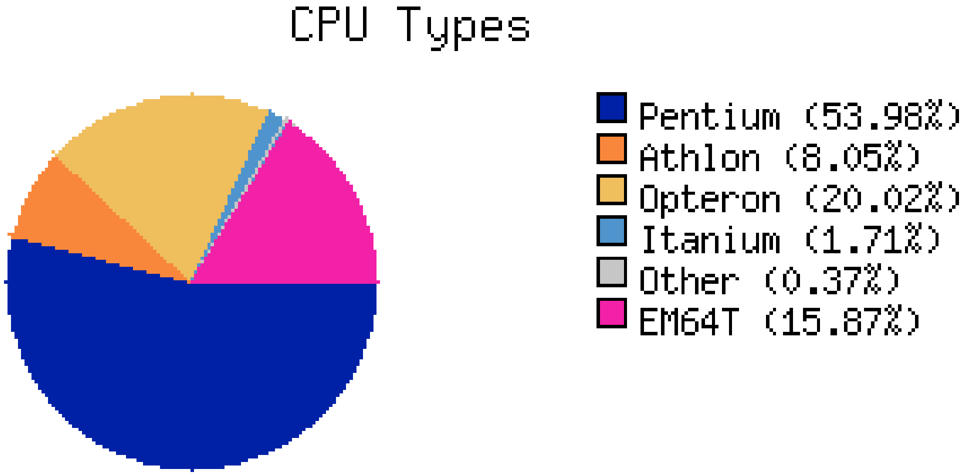
Introduction to ROCKS

While earlier clustering toolkits expend a great deal of effort (i.e., software) to compare configurations of nodes, Rocks makes complete Operating System (OS) installation on a node *the basic* management tool.

With attention to complete automation of this process, it becomes faster to reinstall all nodes to a known configuration than it is to determine if nodes were out of synchronization in the first place.

Unlike a user's desktop, the OS on a cluster node is considered to be *soft state* that can be changed and/or updated rapidly.

Rocks Cluster Register



525 systems registered

31602 cpu’s (average 60 cpu’s each)

NCSA system rated at 7488 GFLOPS (1040p) [#47 on TOP500]

Commercial ROCKS

- Based on V3.3.0
- Uses PXE Boot for bootstart
- Includes GANGLIA, PBS, MPI, Lava, LSF, …
- Existing PBSPro uses RSH. Missing svr rpm
- Configured OK for NFS mounts, /temp dir
- Aug. 16: nodes stuck on **anaconda bug**
- Aug. 23: have to delete partitions to reinstall
- Move 5 KVM cables to sets of 5 nodes
- Created custom config script

Log

- Aug. 26: primary app fails
- Aug. 29: compute node dead
- Sept. 9: installed two other apps
- Sept. 14: rsh fails. SSH works
- Sept. 19: pressing the POWER button to shut down causes node to boot in REINSTALL ROCKS mode
- Sept 26: all nodes up-to-date

Future

- Get primary app to work
- Add other apps
- Configure PBSPro for group priority
- Get console server

Computing Center
State University
sgi
ORIGIN
ICHIT