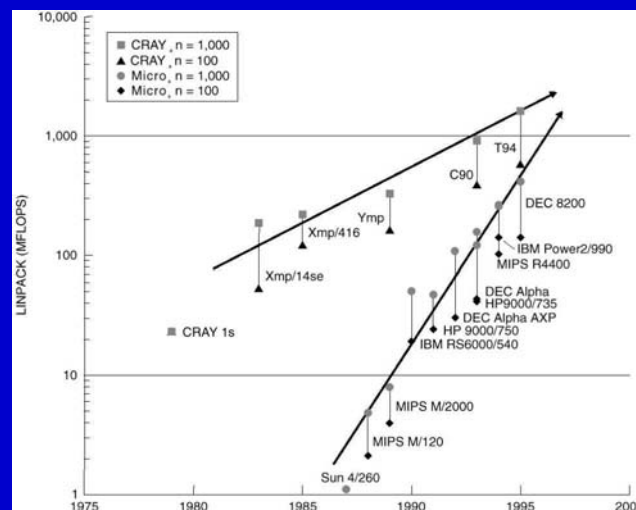


Node Hardware

- Improved microprocessor performance means availability of desktop PCs with performance of workstations (and of supercomputers of 10 years ago) at significantly lower cost
- Parallel supercomputers are now equipped with COTS components, especially microprocessors
- Increasing usage of SMP nodes with two to four processors
- The average number of transistors on a chip is growing by about 40% per annum
- The clock frequency growth rate is about 30% per annum

Performance Convergence



Three Basic Operations

- Instruction execution
 - Involves only CPU and registers
- Register loading
 - Load data from cache or memory into registers
 - Involves CPU, front-side bus, cache, memory
- Peripheral usage
 - Copying data through I/O bus from peripheral to memory
 - Involves peripheral, I/O bus, interface from I/O bus into peripheral and memory, memory



Fall 2008

Paul A. Farrell
Cluster Computing 3

Commodity cluster node

- Processor (CPU)
- On processor registers
- Cache – 10 times faster than memory
- Memory
- Motherboard
- Bus
- Power Supply
- Network Interface Controller (NIC)
- Disk controller
- Disks



Fall 2008

Paul A. Farrell
Cluster Computing 4

Processor

- Binary encoding determined by Instruction Set Architecture (ISA)
- Processors can share part of ISA but not have identical ISAs due to addition of features (instructions such as Streaming SIMD Extensions)
 - SSE, SSE2, SSE3 are numerical instructions for PIII and P4
- Processor clock rate in MHz or GHz is number of clock ticks per second (up to 3.8GHz in 2005 or 4.2GHz)
 - CPUs with different clock rates can perform equivalently
 - CPUs with same rate can perform differently
- Instructions per second / Floating point instructions per second (fps) depend also on ISA, and components on chip



Fall 2008

Paul A. Farrell
Cluster Computing 5

Processor

- Cache mitigates the effect of much slower memory
- CPUs can have cache from kilobytes to 4 to 8 megabytes
- Cache can be of different speeds (L1, L2, L3)
 - Faster is more expensive, therefore less used



Fall 2008

Paul A. Farrell
Cluster Computing 6

Processors

- Intel IA32 (x86) Processors
 - Pentium 3, Pentium4, *Pentium Pro* and Pentium Xeon
 - Athlon, *AMD x86*, *Cyrix x86*, etc.
- Digital Alpha 21364
 - Alpha 21364 processor integrates processing, memory controller, network interface into a single chip
- IBM PowerPC G5
- IA64 (Itanium, Athlon 64, Opteron)
- Sun SPARC
- SGI MIPS
- HP PA-RISC
- Berkeley Intelligent RAM (IRAM) integrates processor and DRAM onto a single chip



Fall 2008

Paul A. Farrell
Cluster Computing 7

IA32

- 32 bit instruction set
- Binary compatibility specification
 - Hardware may be very different but instruction set is the same
 - Pentium III, 4 and Athlon
- Additions to ISA include SSE, SSE2 and SSE3 (streaming SIMD extensions)
 - Can substantially increase performance
 - Important to consider
- Hyperthreading : multiple threads per CPU
 - Negatively impacts performance
 - Can be turned off
- Multicore : multiple cores per processor



Fall 2008

Paul A. Farrell
Cluster Computing 8

IA32

- Pentium III
 - Has SSE and L2 cache on chip
 - Can be used in 2 CPU SMPs
 - Xeon can be used in 4 CPU SMPs
- Pentium 4
 - Designed for higher clock cycles, but less computing power per cycle
 - Also has SSE2 (SSE3 since Prescott) and Hyperthreading
- Athlon
 - Processor architecture like PIII, bus like Compaq Alpha
 - Two 64KB L1 caches and one 256 KB L2 cache
 - Has SSE but not SSE2
 - Can be used in 2 CPU SMPs



Fall 2008

Paul A. Farrell
Cluster Computing 9

HP/Compaq/DEC Alpha 21264

- True 64 bit architecture
- RISC (Reduced Instruction Set Computer)
 - Simple instructions at high clock rate
- Fastest for a long time
- Used in Cray T3D and T3E
- Popular in early and large clusters due to superior fp performance e.g. Los Alamos NL ASCI Q



Fall 2008

Paul A. Farrell
Cluster Computing 10

Power PC G5

- IBM and Apple Mac
- 64 bit CPU running at over 2GHz (2003), 2.5GHz (2005)
- Up to 1.25GHz front-side bus
- Multiple functional units



Fall 2008

Paul A. Farrell
Cluster Computing 11

IA64 Itanium

- New IS, cache design, fp processor
- Clock rates up to 1.6 GHz (2004), multiway fp instruction issue
- Aimed at 1 to 2 Gflops performance
 - HP Server rx 4610, 800 Mhz Itanium SPecfp2000 of 701
 - HP rx2600, 1.5 GHz I2, SPecfp2000 of 2119
 - I2 is significantly faster
- Both need efficient compilers to exploit EPIC (Explicitly Parallel Instruction Computing)



Fall 2008

Paul A. Farrell
Cluster Computing 12

AMD Opteron

- Supports IA32 and IA64 ISA
- Can run legacy 32 bit codes
- Can access in excess of 4GB memory with new 64 bit instructions
- Integrated DDR memory controller
- Up to 3 high-performance “Hypertransport” interconnects with 6.4GB/sec bandwidth per CPU
- Early Opterons had SPECfp2000 of 1154
- Can have 2 CPU SMPs each with separate memory busses
- More popular than I2 for clusters (cost-performance)



Fall 2008

Paul A. Farrell
Cluster Computing 13

Athlon 64

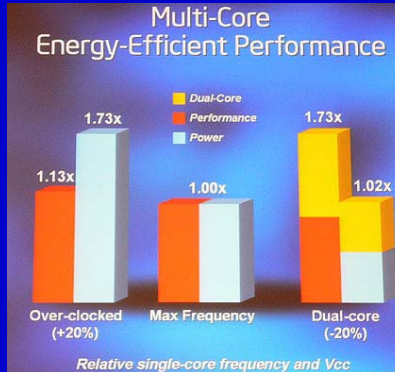
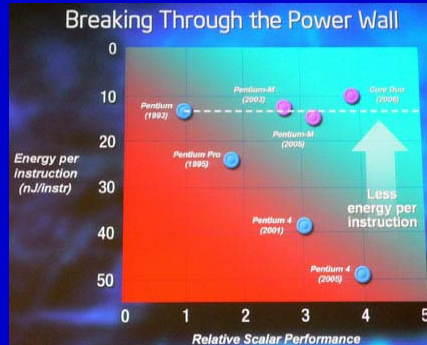
- HyperTransport™ technology with up to 14.4 GB/sec total processor-to-system bandwidth
- integrated DDR memory controller
- Supports SSE2
- Later versions (since Venice Stepping E3 and San Diego Stepping E4) support SSE3



Fall 2008

Paul A. Farrell
Cluster Computing 14

Energy Issue



Evolution of Multi-Core Technology

Single-threading:

Only one task processes at one time.

Multitasking and Multithreading:

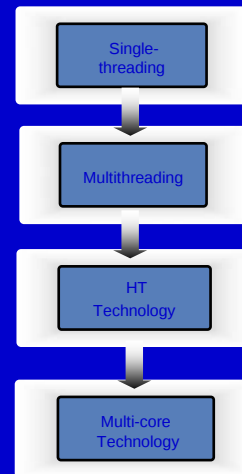
Two or more tasks execute at one time by using content switching (Functionality).

HT Technology:

Two single threads execute simultaneously on the same processor core.

Multi-core Technology:

Computational work of an application is divided and spread over multiple execution cores (Performance).



Evolution of Multi-Core Technology

Formal Definition of Threads, Processors, and Hyper-Threading in Terms of Resources

- Thread
 - Basic unit of CPU utilization
 - Program counter, CPU state information, stack
- Processor
 - Architecture state: general purpose CPU registers, interrupt controller registers, caches, buses, execution units, branch prediction logic
- Logical Processor
 - Duplicates the architecture space of processor
 - Execution can be shared among different processors
- HyperThreading
 - Intel version of Simultaneous Multi-Threading (SMT)
 - Makes single physical processor appear as multiple logical processors
 - OS can schedule to logical processors



Fall 2008

Paul A. Farrell
Cluster Computing 17

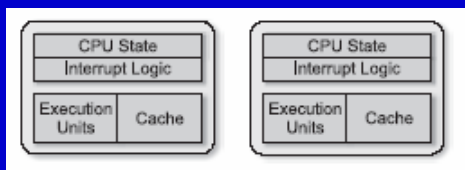
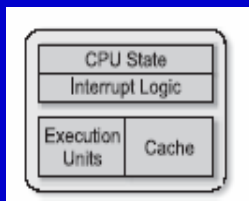
HyperThreading and Multicore

- HyperThreading
 - Instructions from logical processors are persistent and execute on shared execution resources
 - Up to microarchitecture how and when to interleave the instructions
 - When one thread stalls, another can proceed
 - This includes cache misses, and branch mispredictions
- Multicore
 - Chip multiprocessing
 - 2 or more execution cores in a single processor
 - 2 or more processors on a single die
 - Can share an on-chip cache
 - Can be combined with SMT



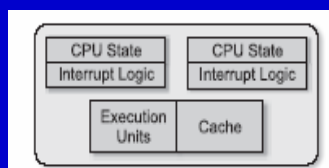
Fall 2008

Paul A. Farrell
Cluster Computing 18

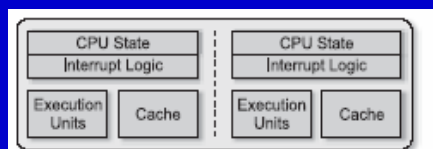


Multiprocessor

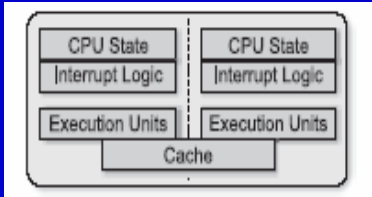
- Single Processor



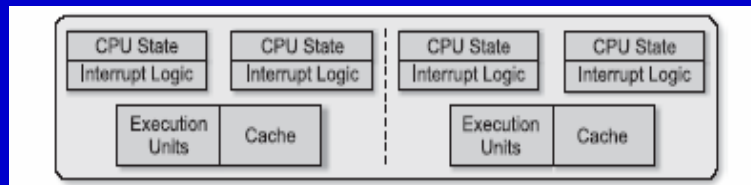
- Hyper-Threading



- Multi-core Processor



- Multicore Shared Cache (Intel)



- Multicore with HyperThreading Technology (AMD)



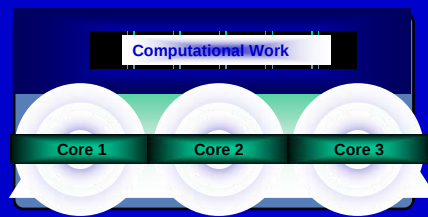
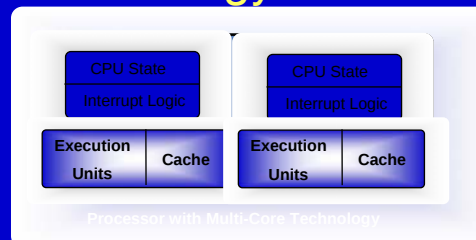
Fall 2008

Paul A. Farrell
Cluster Computing 21

Defining Multi-Core Technology

With Multi-core technology:

- Two or more complete computational engines are placed in a single processor.
- Computers split the computational work of a threaded application and spread it over multiple execution cores.
- More tasks get completed in less time increasing the performance and responsiveness of the system.



Computational Work Distributed Between Multiple Cores



Fall 2008

Paul A. Farrell
Cluster Computing 22

Evolution to Dual Core



Basic Processor
Pipelined Execution
1 Instr. / Cycle



Fall 2008

Paul A. Farrell
Cluster Computing 23

Evolution to Dual Core



Superscalar
Parallel Functional Units



Fall 2008

Paul A. Farrell
Cluster Computing 24

Evolution to Dual Core



Double the transistors

Scope



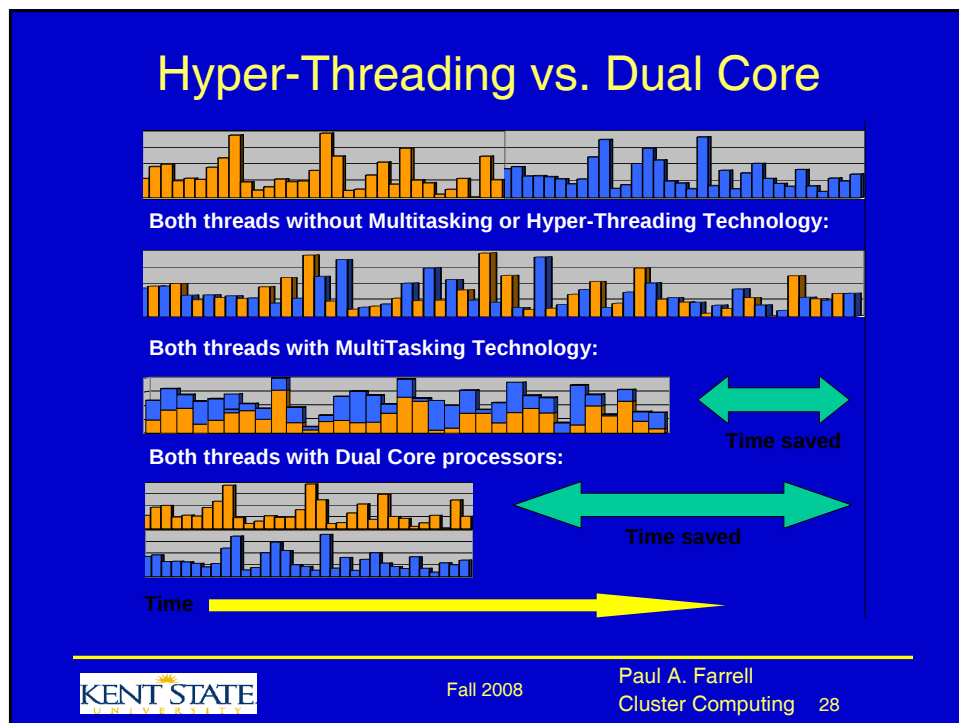
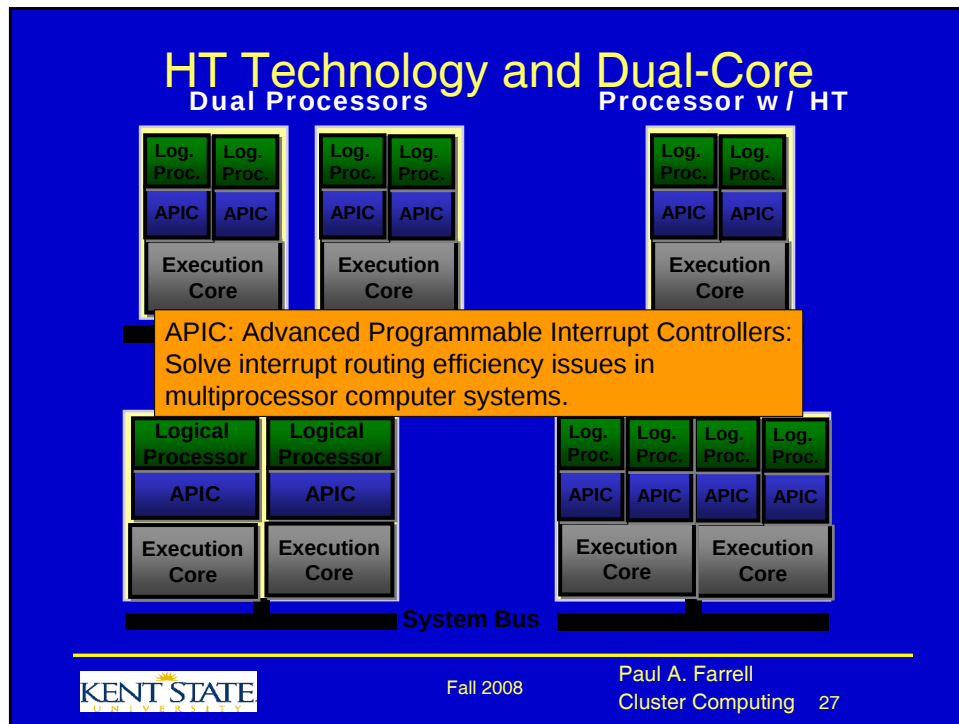
Instruction Level – encoding/pipelined execution – increase throughput



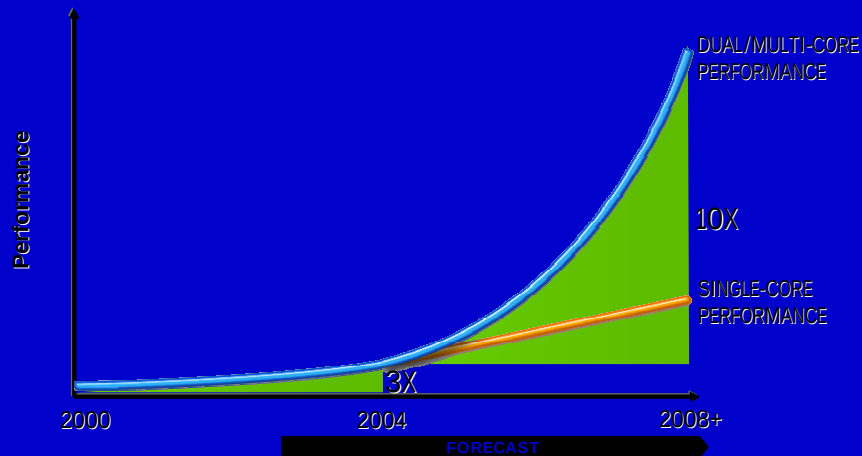
Instruction Level – scheduling, superscalar



Outermost Loops, whole program – data/functional

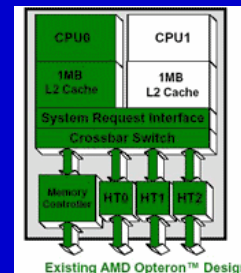


Driving Greater Parallelism



Dual Core Processors

- Opteron
 - for servers by mid-2005
 - for clients late-2005
 - In laptops in 2006
- One die with 2 CPU cores, each core has its own 1MB L2 cache
- Drops into existing AMD Opteron 940-pin sockets
- 2 CPU cores share the same memory and HyperTransport™ technology resources found in single core AMD Opteron processors
- Followed by P4 EE (4/05), Athlon 64, laptops



Intel Dual Core Roadmap April 2005

The Move to IA Multi-core

Platform	Current	2005	2006+	Future
Itanium® processor MP	Itanium® 2 Processor	Montecito	Montvale	Tukwila
Itanium® processor DP	Itanium® 2 Processor - 3M (Fanwood)	Millington	DP Montvale	Dimona
MP Server	Intel® Xeon™ Processor MP	64-bit Intel® Xeon™ processor MP	Paxville	Tulsa
DP Server / WS	64-bit Intel® Xeon™ Processor w/ 2MB cache		Dempsey	Future
Desktop Client	Pentium® 4 processor	Pentium® Processor Extreme Edition Smithfield	Presler	Future
		Pentium® 4 processor	Cedar Mill	
Mobile Client	Pentium® M processor		Yonah	Future
			Yonah	

All products and dates are preliminary and subject to change without notice.



Fall 2008

Paul A. Farrell
Cluster Computing 31

Dual Core Pentium

- Dec 2005 Intel Pentium Extreme Edition 840 at 3.2GHz
 - Multiple cores in single processor
 - Can use hyperthreading to give 4 threads
 - Aimed at reducing power consumption
 - 65nm technology
 - 2MB L2 cache



Fall 2008

Paul A. Farrell
Cluster Computing 32

Intel Core MicroArchitecture

- July 2006 Core 2 Duo
 - New Micro Architecture
 - <http://www.intel.com/technology/architecture/coremicro/>
 - Unified (shared) L2 cache (2MB or 4MB)
 - **Wide Dynamic Execution**
 - **Advanced Smart Cache**
 - **Smart Memory Access**
 - **Advanced Prefetch**
 - **Memory Disambiguation**
 - **Advanced Digital Media Boost**
 - **Intelligent Power Capability**
 - **Presentation:**
 - <http://www.intel.com/technology/architecture/coremicro/demo/demo.htm>

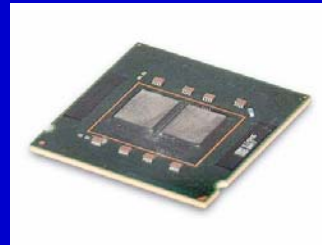


Fall 2008

Paul A. Farrell
Cluster Computing 33

Intel Quad Core

- Nov 2006 : Core 2 Quad Core - Kentsfield
 - two Core 2 Duo E6700 cores in one socket package
- Dec 2006: Xeon Quad Core Clovertown
 - two dual-core Woodcrest-type Xeons in a single processor package
 - 2x 4 MB L2 cache and FSB1066 or FSB1333



Fall 2008

Paul A. Farrell
Cluster Computing 34

AMD and Intel Differences

- Shared L2 cache v separate caches
- Shared memory v separate memory
- AMD processors : built-in DDR or DDR2 memory controller
- Intel : memory controller, which is part of the platform core logic
 - data path is still shared by all available cores

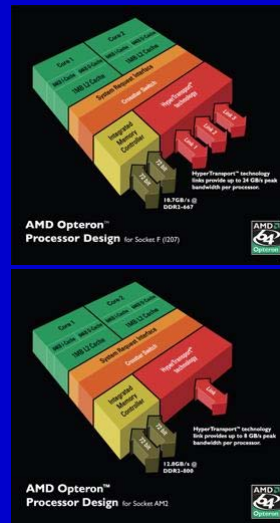


Fall 2008

Paul A. Farrell
Cluster Computing 35

Second Generation Opteron

- Direct Connect Architecture
 - HyperTransport™ technology provides a scalable bandwidth interconnect between processors, I/O subsystems, and other chipsets, with up to three coherent HyperTransport technology links providing up to 24.0 GB/s peak bandwidth per processor
- AMD Virtualization™ (AMD-V™)
- energy efficient DDR2 memory
 - Integrated Memory Controller on-die
- Up to 8 processor (16 core) systems
- quad-core upgradeability in 2007



Fall 2008

Paul A. Farrell
Cluster Computing 36

AMD Barcelona plan

- Quad core Opteron

	AMD CPUs	Barcelona
SSE execution width	64 bits wide	128 bits wide
Instruction fetch bandwidth	16 bytes/cycle	32 bytes/cycle
Data cache bandwidth	2 x 64 bit loads/cycle	2 x 128 bits loads/cycle
L2 cache/memory controller bandwidth	64 bits/cycle	128 bits/cycle
Floating-point scheduler depth	36 dedicated x 64-bit ops	36 dedicated x 128-bit ops



Fall 2008

Paul A. Farrell
Cluster Computing 37

Instructions per Cycle in AMD's Barcelona

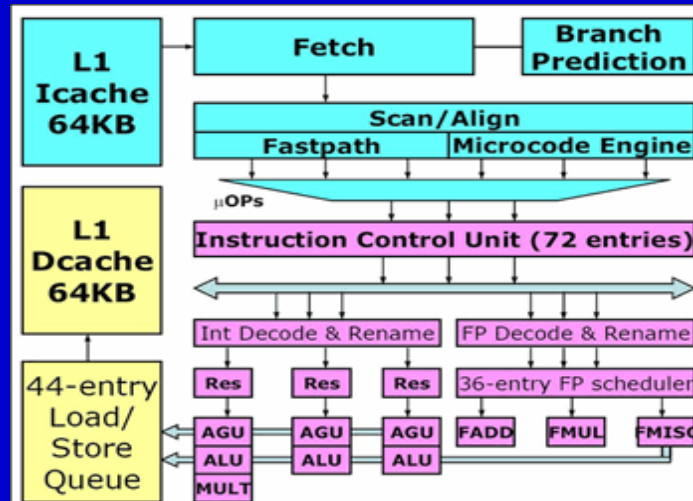
- Enhancements
- **Advanced branch prediction.** AMD's architects have doubled the return stack size, added more branch history bits, and built in a 512-entry indirect branch predictor.
- **32-byte instruction fetch.** Increases efficiency by reducing split-fetch instruction cases.
- **Sideband stack optimizer.** Adjustments to the stack don't take up functional unit bandwidth.
- **Out-of-order load execution.** Load instructions can actually bypass other loads in some cases, as well as stores that are not dependent on the load in question. This minimizes the effect of L2 cache latency



Fall 2008

Paul A. Farrell
Cluster Computing 38

Instructions per Cycle in AMD's Barcelona

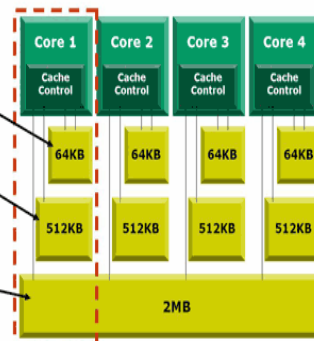


Memory Bandwidth

- Independent memory controllers
- full 48-bit hardware addressing
- 1GB memory page size in addition to the common 4KB and 2MB page sizes.
- L1 cache is 64KB, the L2 cache is 512KB and the L3 cache is 2MB
- virtualized address translation, instead of the current shadow paging

Balanced, Highly Efficient Cache Structure

- Dedicated L1**
- Locality keeps most critical data in the L1 cache
 - Lowest latency
 - 2 loads per cycle
- Dedicated L2**
- Sized to accommodate the majority of working sets today
 - Dedicated to eliminate conflicts common in shared caches
 - Better for Virtualization
- Shared L3 - NEW**
- Victim-cache architecture maximizes efficiency of cache hierarchy
 - Fills from L3 leave likely shared lines in the L3
 - Sharing-aware replacement policy
 - Ready for expansion at the right time for customers



Power Management in AMD's Barcelona

- AMD's 65nm SOI process, which allows for lower voltage and thermal output while increasing circuit densities
- separate CPU core and memory controller power
- enhanced version of AMD's PowerNow
 - will allow individual core frequencies to lower while other cores may be running full bore
- two-socket system, plus memory, plus the chipset, is estimated at about 240W (not counting graphics and storage.)
- shipping by mid-2007



Fall 2008

Paul A. Farrell
Cluster Computing 41

References

- **Product Brief: Quad-Core AMD Opteron™ Processor**
 - http://www.amd.com/us-en/Processors/ProductInformation/0,,30_118_8796_15223_0_0.html
- **Intel® Xeon® Processor 5000 Sequence**
 - <http://www.intel.com/products/processor/xeon5000/>
 - http://download.intel.com/products/processor/xeon/dc54kplat_brief.pdf



Fall 2008

Paul A. Farrell
Cluster Computing 42

Memory (RAM)

- Standard Industry Memory Module (SIMM) – RDRAM and SDRAM
- Access to RAM is extremely slow compared to the speed of the processor
 - Memory busses (front side busses FSB) run at 100MHz to 800MHz
 - Memory speed metrics
 - Peak memory bandwidth: burst rate from RAM to CPU
 - Currently 1 to 4 GB/secs
 - FSB must be fast enough for this
 - Latency: now under 6 nanosecs (2003)
- Extended Data Out (EDO)
 - Allow next access to begin while the previous data is still being read
- Fast page
 - Allow multiple adjacent accesses to be made more efficiently



Fall 2008

Paul A. Farrell
Cluster Computing 43

RAM size

- RAM size determines size of problem that can be run at reasonable speed
- Alternatives:
 - Out-of-core calculations
 - Virtual memory
- Old rule of thumb
 - 1B RAM per 1 flop (gross approximation)

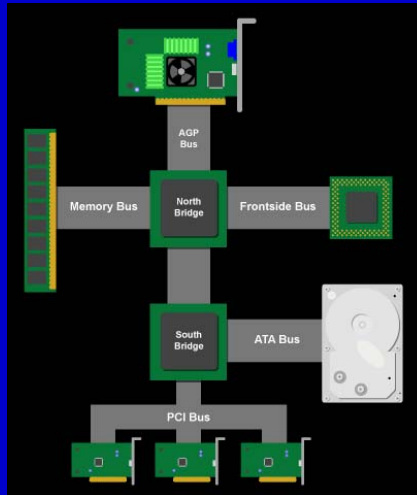


Fall 2008

Paul A. Farrell
Cluster Computing 44

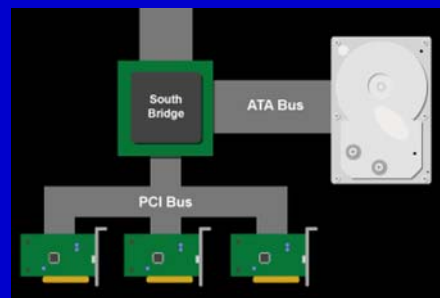
Chipsets

- North Bridge: FSB connects CPU, memory bus, AGP
- South Bridge: I/O bus bridge, disk & USB controllers



I/O Channels

- Bus from peripherals to main memory
- Connected by a bridge (PCI chipset) to memory



I/O Channels

- PCI bus (1994)
 - 32 bit/33MHz : 133MB/s peak, 125MB/s attained
 - 64 bit/66MHz : 500MB/s peak, 400-500M/s in practice
- PCI-X
 - 64bit/133MHz : 900MB/s - 1GB/s peak
- PCI-X 2
 - 64bit/PCI-X 266 and PCI-X 533, offering up to 4.3 gigabytes per second of bandwidth
- PCI-X requires a controller for every slot - just too expensive for desktop deployment



Fall 2008

Paul A. Farrell
Cluster Computing 47

I/O Channels

- AGP (not really a bus)
 - High speed graphics adapters
 - Better peak than PCI and PCI-X
 - Not bus
 - Not symmetric – slower from graphics card to memory
 - Directly addresses main memory – can only support one device
 - AGP 2.0 (4x) peak 1GB/s to main memory, AGP 3.0 (8x) is 2.1 GB/s
- Legacy Busses (Slow)
 - ISA bus (AT bus)
 - Clocked at 5MHz and 8 bits wide
 - Clocked at 13MHz and 16 bits wide
 - VESA bus
 - 24/32 bits bus matched system's clock speed



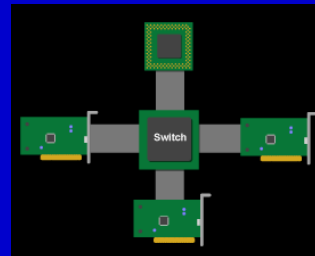
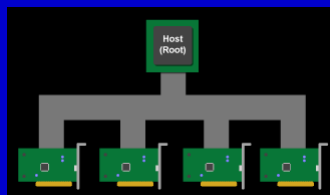
Fall 2008

Paul A. Farrell
Cluster Computing 48

PCI-Express

- High-bandwidth, low pin count, serial, interconnect technology
<http://developer.intel.com/technology/pciexpress/devnet/desktop.htm>
- An assembly of serial, point-to-point wired, individually clocked 'lanes' each consisting of two data lines of 2.5Gb/s, one upstream and one downstream
- x1 : 2.5Gb/s (250MB/s) bidirectional for Gigabit Ethernet, TV Tuners, 1394a/b controllers, and general purpose I/O.
- x16 : 4GB/s bidirectional for video cards (double AGPx8)
- Express Card (successor to PCMCIA for laptops)
 - Supports x1 PCI-Express and Fast USB

PCI Bus v PCI Express Switch



Switch avoids bus contention/arbitration

Motherboard

- PCB (Printed Circuit Board)
- Next to CPU most important component for performance
- Sockets/connectors include:
 - CPU, Memory, PCI/PCI-X, AGP, Floppy disk
 - ATA and/or SCSI
 - Power
 - LEDs, speakers, switches, etc
 - External I/O
- Chips
 - System bus to memory
 - Peripheral bus to system bus
 - PROM with BIOS software



Fall 2008

Paul A. Farrell
Cluster Computing 51

Motherboard

- Choice restricts
 - CPU
 - Clock speed
 - # of CPUs
 - Memory capacity, type
 - Disk interfaces
 - Number and types of I/O busses

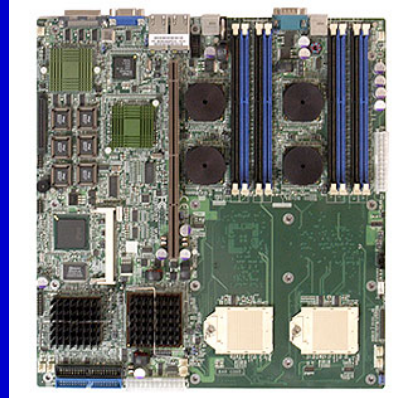


Fall 2008

Paul A. Farrell
Cluster Computing 52

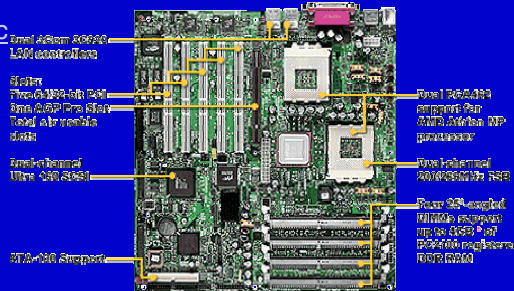
Motherboards – I2

- Supermicro i2DML-8G2
- Dual I2
- Intel E8870 chipset
- **400 MHz** FSB
- 64-bit 133/100 MHz PCI-X
- <http://www.supermicro.com/PRODUCT/MotherBoards/E8870/i2DML-8G2.htm>



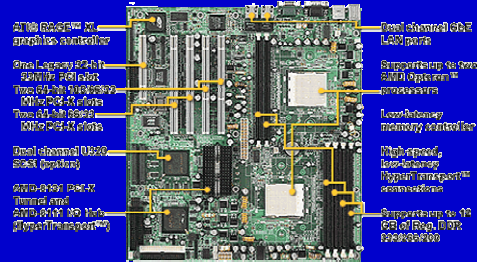
Motherboards - Opteron

- Tyan Thunder K7
- <http://www.tyan.com/products/html/thunderk7.html>
- Dual Opteron
- AMD's new 760MP chipset
- , DDR memory support,
- 64-bit PCI slots,
- AGP Pro slot (and integrated VGA),
- dual LAN controllers,
- dual-channel Ultra160 SCSI,



Motherboards - Opteron

- Tyan Thunder K8S
- Dual Opteron
- Two 128-bit DDR memory buses
- Two independent PCI-X buses
- Two 64-bit 66/33 MHz (3.3-volt) PCI-X slots - from PCI-X bridge A
- Two 64-bit 133/100/66/33 MHz (3.3-volt) PCI-X slots - from PCI-X bridge B (closer to CPUs)
- One Legacy 32-bit 33MHz (5-volt) PCI slot 64-bit PCI slots
- dual LAN controllers,
- dual-channel Ultra160 SCSI,

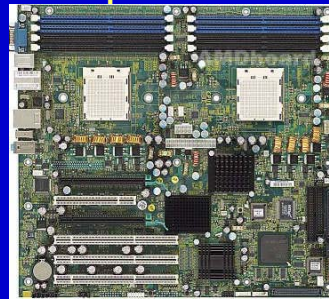


Fall 2008

Paul A. Farrell
Cluster Computing 55

Motherboards - Opteron

- Tyan Thunder K8WE
 - Dual Opteron
 - Three [HyperTransport](#) links support up to 6.4GB/s data transfer rate each link
 - 144-bit [DDR](#) interface (128-bit data + 16 bit ECC)
 - Supports up to 16GB of [Registered DDR](#) 400/333/266 meory
 - dual LAN controllers,
 - dual-channel Ultra160 SCSI,
- Two x16 [PCI Express](#) FULL SPEED slots
 - Slot1 PCI-E x16 from nForce Prof. 2200
 - Slot3 PCI-E x16 from nForce Prof. 2050
 - Two independent 64-bit PCI-X buses
 - Slot 4 & 5: PCI-X 100 MHz max. (Bridge B)
 - Slot 6: PCI-X 133 MHz max. (Bridge A)
 - One 32-bit 33MHz PCI v2.3 (Slot 2)



Fall 2008

Paul A. Farrell
Cluster Computing 56

Motherboard – quad opteron

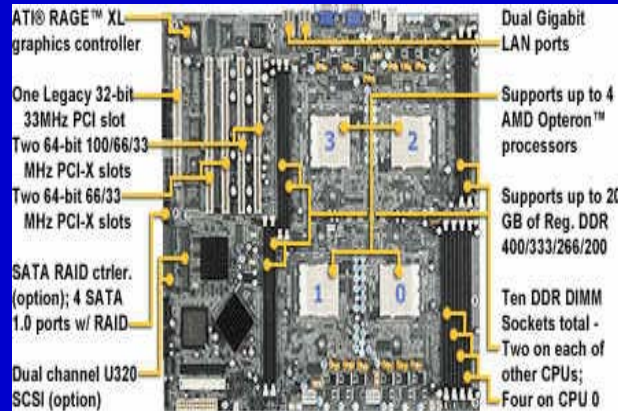
Thunder K8QS - S4880

128-bit dual channel memory bus

- Total ten DDR DIMM sockets (4 on CPU 0 and 2 each on CPU1, CPU2, CPU3 respectively)
- Supports up to 20 GB Registered DDR
- Supports [PC1600](#), [PC2100](#), & [PC2700](#) DDR

Two independent PCI-X buses

- Two 64-bit 66/33 MHz (3.3-Volt) PCI-X slots from PCI-X bridge A



- Two 64-bit 100/66/33 MHz (3.3-Volt) PCI-X slots from PCI-X bridge B
- One legacy 32-bit 33MHz (5-Volt) PCI slot



Fall 2008

Paul A. Farrell
Cluster Computing 57

8-way Opteron

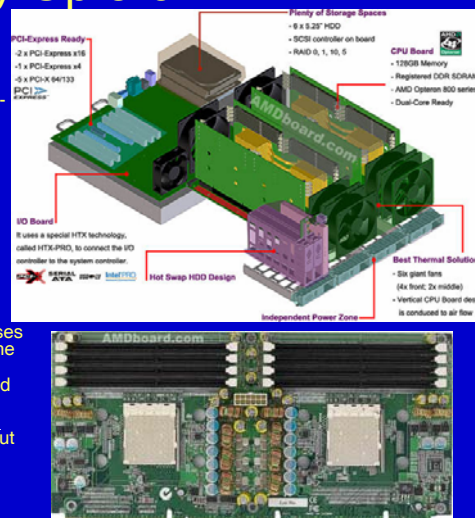
- **IWILL H8502**
- **8-way computing power**
Linpack benchmark raised 80% from 4-way to 8-way system. Dual-core ready
- **128G Memory**

PCI-Express Ready

IWILL Octet-way server adopts nForce Professional 2200 plus 2050. It could provide up to 40 lanes [PCI-Express](#) bandwidth.

High-performance I/O

IWILL Octet-way server architecture uses a special HTX technology to connect the I/O controller to the system controller. The Serial ATA controller, USB 2.0, and serial ports are all integrated through a bidirectional 1000MHz HyperTransport interconnects for a maximum throughput of 1.6GBps.



Fall 2008

Paul A. Farrell
Cluster Computing 58

BIOS

- Software that initializes system so can boot, does POST (power on self test) including memory test, SCSI and IDE bus initialization
- BIOS is motherboard specific
- Various BIOSes
 - PXE (Pre-execution environment) allows boot from network config and boot images
 - Uses DHCP and tftp
 - Can be in BIOS or ethernet card initialization code
 - LinuxBIOS streamlined but does not support all OSes
 - Linux and Windows 2000
 - Adv: source available, faster (<5 sec v 10 to 90 secs)



Fall 2008

Paul A. Farrell
Cluster Computing 59

Local Hard Disks

- Disk busses: SCSI, IDE (EIDE or ATA), SATA (serial ATA)
- IDE controllers on motherboard support 2 busses of 2 devices each. Higher CPU utilization v SCSI.
 - Fastest UDMA133: 133 MB/s
- SCSI used in servers.
 - Faster (up to 320 MB/s), more devices, more expensive
- SATA: serial as opposed to parallel (ATA, SCSI)
 - 150 MB/s, smaller cables, 2 devices per bus, hot pluggable
 - Easier to increase bus speeds
- Disk platter speeds: 5400, 7200, 10000, 15000rpm



Fall 2008

Paul A. Farrell
Cluster Computing 60

Local Hard Disks

- Overall improvement in disk access time has been less than 10% per year
- Amdahl's law
 - Speed-up obtained from faster processors is limited by the slowest system component
- Parallel I/O
 - Carry out I/O operations in parallel, supported by parallel file system based on hardware or software RAID



Fall 2008

Paul A. Farrell
Cluster Computing 61

RAID

- Redundant Array of Inexpensive Disks
- Disk aggregate appear as single disk
- Adv: larger data, faster, redundancy
- Software (possibly high CPU utilization) or hardware
- RAID versions
 - RAID0: striping across multiple disks, faster reads & writes
 - RAID1: mirroring, 2 copies of data, faster read, slower write
 - RAID5: one disk for parity info, can recover data from disk failure, read faster, writes require checksum computation
- RAID used on cluster storage nodes



Fall 2008

Paul A. Farrell
Cluster Computing 62

Nonlocal Storage

- Storage device bus traffic transferred over network
 - Net may be dedicated or shared
- iSCSI: SCSI encapsulated in IP
 - Possible bottleneck
 - FibreChannel similar but dedicated net and protocol
- Network file systems : NFS & PVFS
 - Data transmitted with filesystem semantics

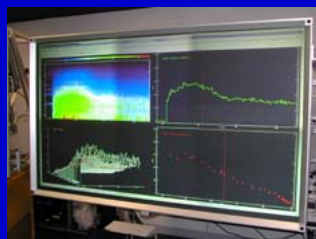


Fall 2008

Paul A. Farrell
Cluster Computing 63

Video

- Usually only to debug hardware & update BIOS
- Advanced not needed unless cluster used for visualization e.g. tiled displays
 - Used to show regions of 3D visualizations
- AGP or PCI
 - Nvidia GeForce, ATI Radeon, Matrox

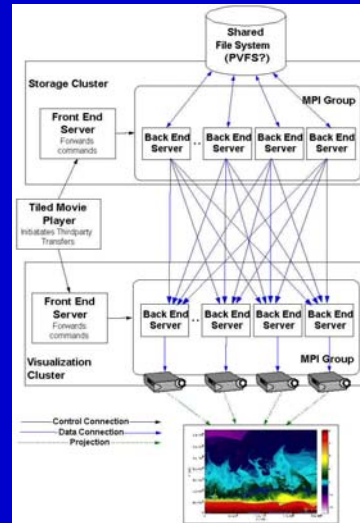


Fall 2008

Paul A. Farrell
Cluster Computing 64

Tiled Display

- Series of cluster nodes outputting to projector
- Usually back projection
- Synchronization issues
 - Software synch
 - genlock



Peripherals

- Other peripherals not usually used in clusters
 - USB (1.1, 2.0), Firewire
 - USB might be used for keyboard/mice
- Legacy interfaces
 - Keyboard, mice, serial (RS232), parallel

Packaging – Cases

- Early clusters used desktop or tower cases
- Cooling not a problem
- Density low



Fall 2008

Paul A. Farrell
Cluster Computing 67

Packaging – Rack Mount

- Low profile cases (1U, 2U, 3U, 4U)
- 1U = 1.75in high
- Designed for rack mount
- High density, good servability
- Need to be careful of cooling needs



Appro 1U Clusters



Fall 2008

Paul A. Farrell
Cluster Computing 68

Packaging – Blade Servers

- Machines packed as tightly as possible
- Very high density
- May have shared components e.g. power
- Blades specialized and not usually expandable
- Usually most expensive per unit option



Cluster Hardware Configuration

- Having chosen the hardware configuration, need to decide vendor
- Cluster vendors – integrated solutions, support
 - Large system vendors
 - Specialist vendors
- White boxes – use custom system vendors without cluster experience – hardware but not software maintenance
 - Specify exact parts
- DIY – specify and buy parts, and assemble yourself

Cluster Vendors

http://dmoz.org/Computers/Parallel_Computing/Beowulf/Vendors/

- Alinka
- Armari
- Aspen Systems
- Atipa
- GraphStream
- Linux NetworX
- Microway
- Penguin Computing
- PSSC Labs
- RackSaver
- RocketCalc



KENT STATE UNIVERSITY

Fall 2008

Paul A. Farrell
Cluster Computing 71

Large System Vendors



Cluster Computing 72

Clusters in Computer Science

- **Fianna 32 dual PIII-450 nodes with Gigabit & 100Mbps Ethernet**
 - 256MB memory per node
 - Osoleted Summer 2008
- **Cluster Ohio cluster : 8 nodes each with four 550 Megahertz Intel Pentium III Xeon with Myrinet**
 - Two Gigabytes of RAM
 - Two Myrinet interfaces
 - One 100Base-T Ethernet interface



Fall 2008

Paul A. Farrell
Cluster Computing 73

Clusters in Computer Science

- **Two RocketCalc clusters**
 - 4 boards each with dual Intel Pentium Xeon 2.4 GHz processors with Gigabit Ethernet
- **Rocketcalc Opteron Cluster with 4 Panel Display**
 - Four Tyan Thunder K8W motherboards with onboard Broadcom Gigabit NetXtreme BCM5703 64bit network interface each with
 - Two AMD Opteron 1.8 GHz model 244 processors
 - 2 GB Registered 333 MHz ECC DDR SDRAM per processor



Fall 2008

Paul A. Farrell
Cluster Computing 74

Clusters in Computer Science

- Opteron Cluster powering 12 20" LCD panel display wall in 4x3 configuration
- Node configuration
 - Dual AMD Opteron 250 at 2.4GHz
 - 4GB memory
 - Nvidia GeForce 7900 GTX graphics cards



Fall 2008

Paul A. Farrell
Cluster Computing 75

New Cluster July 2008 - Gandalf



- 24 nodes – 48 processors, 144 cores
 - 12 nodes dual 2.8GHz, 2220, dual-core Opteron, socket F, 2MB cache
 - 12 nodes dual 2.0GHz, 2350, Quad-core Opteron, socket F, 4MB cache
- 16GB 667, DDR2 memory each
- SilverStorm, 9000, Dual port Infiniband NIC each
- SilverStorm 24 port Infiniband DDR Switch
- NVIDIA, 768MB, 8800GTX, PCI-E x16 in some nodes

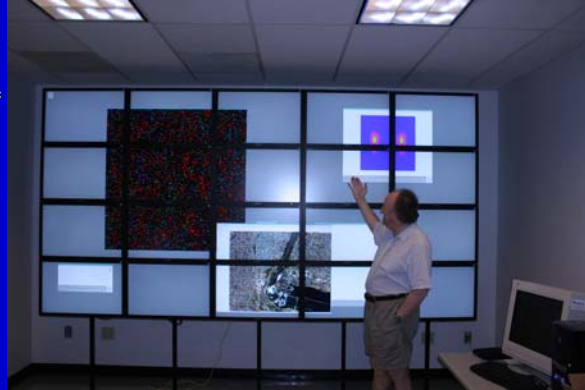


Fall 2008

Paul A. Farrell
Cluster Computing 76

New Display Wall

- 5x4 30in Samsung monitors
- Connected 10 nodes of to gandalf each with two Nvidia 8800GTX



Fall 2008

Paul A. Farrell
Cluster Computing 77

Clusters at OSC 2004

- **OSC IA32 Cluster**
 - 128 compute and 16 I/O nodes each with
 - Two gigabytes of RAM
 - Two, 1.4 GHz AMD Athlon MPs, each with 256kB of secondary cache
- **OSC Itanium 2 Cluster (includes SGI Altix 3000)**
 - 32 processor SGI Altix 3000 for SMP and large memory applications
 - 128 parallel compute nodes
 - 900 Megahertz Intel Itanium 2 processors with 1.5MB of tertiary cache, 4GB RAM
 - 20 serial compute nodes with 20GB RAM



Fall 2008

Paul A. Farrell
Cluster Computing 78

Clusters at OSC 2005-2007

- **OSC Pentium 4 Cluster**
 - 252 compute and 16 I/O nodes each with
 - Four gigabytes of RAM
 - Two 2.4 GHz Pentium Xeon, each with 512kB of secondary cache
 - 100Base-T and 1000Base-T Ethernet
 - 112 with Infiniband 10Gb interface
- **OSC Itanium 2 Cluster (includes SGI Altix 3000)**
 - 32 processor SGI Altix 3000 for SMP and large memory applications 64GB
 - 128 parallel compute nodes
 - 900 Megahertz Intel Itanium 2 processors with 1.5MB of tertiary cache, 4GB RAM, Myrinet 2000, Gigabit Ethernet
 - 20 serial compute nodes with 12GB RAM



Cluster



Fall 2008

Paul A. Farrell
Cluster Computing 79

Additions in 2005

- Itanium Cluster
- Two 16 processor SGI Altix 350 systems for SMP and large memory applications configured with:
 - 32 GB of memory
 - 16 1.4 Gigahertz Intel Itanium 2 processors
 - 4 Gigabit Ethernet interfaces
 - 2 Gigabit Fibre Channel interfaces
 - approximately 250 GB of temporary disk (/tmp)
- 110 compute nodes for parallel jobs configured with:
 - Four Gigabytes of RAM
 - Two, 1.3 Gigahertz Intel Itanium 2 processors
 - 80 Gigabytes, ultra-wide SCSI hard drive
 - One Myrinet 2000 interface card
 - One Gigabit Ethernet interface



Fall 2008

Paul A. Farrell
Cluster Computing 80

Glenn, the Ohio Supercomputer Center IBM Cluster 1350

- 877 System x3455 compute nodes
 - Dual socket, dual core 2.6 GHz Opterons
 - 8 GB RAM
 - 48 GB local disk space in /tmp
- 88 System x3755 compute nodes
 - Quad socket, dual core 2.6 GHz Opterons
 - 64 GB (2 nodes), 32 GB (16 nodes), or 16 GB (70 nodes) RAM
 - 1.8 TB (10 nodes) or 218 GB (76 nodes) local disk space in /tmp
- 1 e1350 Blade Center
 - 4 Dual **Cell Broadband Engine based** QS20 blades
 - Voltaire 10 Gbps PCI Express adapter
- All connected together by 10 Gbps Infiniband



Fall 2008

Paul A. Farrell
Cluster Computing 81

OSC – BALE Cluster

- 50 node dual purpose system
- Used to run classroom and when not in use to run graphics rendering jobs
 - Dual processor AMD system, Tyan Tiger MPX 760MPX Motherboard
 - 1.53GHz AMD Athlon 1800MP CPUs
 - Myrinet 2000 interface
 - Quadro4 900 XGL graphics board



Fall 2008

Paul A. Farrell
Cluster Computing 82

Apple G5 Cluster

- 1 front-end node configured with:
 - 4 Gigabytes of RAM
 - 2 - 2.0 Gigahertz PowerPC G5 processors
 - 2 Gigabit Fibre Channel interfaces
 - Approximately 750 GB of local disk
 - Approximately 12TB of Fibre Channel attached storage
- 32 compute nodes configured with:
 - Four Gigabytes of RAM
 - Two, 2.0 Gigahertz PowerPC G5 processors
 - Approximately 80 GB of local disk space (/tmp)
 - Dual Gigabit Ethernet interfaces



Fall 2008

Paul A. Farrell
Cluster Computing 83

Pitfalls

- Need to ensure adequate power and cooling
 - Otherwise unpredictable behavioral problems
- Use a console solution
 - Crash cart or KVM switch to see BIOS errors at boot
- Profile the target applications
 - Benchmarks a guide but no substitute
- Remember if you have 64 nodes you may have to do it 64 times
 - So saving small amounts by buying parts may not be cost effective



Fall 2008

Paul A. Farrell
Cluster Computing 84