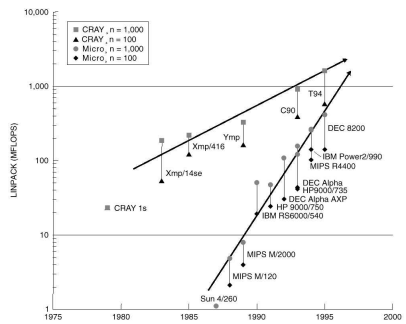
Node Hardware

- Improved microprocessor performance means availability of desktop PCs with performance of workstations (and of supercomputers of 10 years ago) at significanty 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

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

Processor

- Binary encoding determined by Instruction Set Architecture (ISA)
- Processors can share ISA but not have identical ISAs due to addition of features (instructions)
  - SSE and SSE2 are numerical instructions for PIII and P4
- Processor clock rate in MHz or GHz is number of clock ticks per second (up to 3GHz in 2003)
  - 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

Processor

- Cache mitigates the effect of much slower memory
- CPUs can have cache kilobytes to 4 to 8 gigabytes

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
- Opteron
- Sun SPARC
- SGI MIPS
- HP PA-RISC
- Berkeley Intelligent RAM (IRAM) integrates processor and DRAM onto a single chip

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 and SSE2 (streaming SIMD extensions)
  - Can substantially increase performance
  - Important to consider
- Hyperthreading : multiple threads per CPU
  - Negatively impacts performance
  - Can be turned off

IA32

- Pentium 4
  - Designed for higher clock cycles, but less computing power per cycle
  - Also has SSE2 and Hyperthreading
- Pentium III
  - Has SSE and L2 cache on chip
  - Can be used in 2 CPU SMPs
  - Xeon can be used in 4 CPU SMPs
- 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

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

Power PC G5

- IBM and Apple Mac
- 64 bit CPU running at over 2GHz (2003)
- 1GHz front-side bus
- Multiple functional units

IA64 Itanium

- New IS, cache design, fp processor
- Clock rates 1GHz plus, multiway fp instruction issue
- Aimed at 1 to 2 Gflops performance
  - HPServer 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)

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

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

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)

I/O Channels

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

I/O Channels

- AGP (not really a bus)
  - High speed graphics adapters
  - Better peak than PCI and PCI-X
  - Not bus
  - Directly addresses main memory – can only support one device
  - AGP 2.0 peak 1GB/s to main memory, AGP 3.0 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

PCI-Express

- High-bandwidth, low pin count, serial, interconnect technology http://developer.intel.com/technology/pciexpress/devnet/desktop.htm
- x1 : 2.5GB/s for Gigabit Ethernet, TV Tuners, 1394a/b controllers, and general purpose I/O.
- X4 : 16GB/s for video cards (double AGPx8)
- Express Card (successor to PCMCIA for laptops)
  - Supports x1 PCI-Express and Fast USB

Motherboard

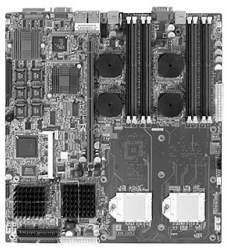
- PCB (Printed Circuit Board)
- Next to CPU most important component for performance
- Sockets/connectors include:
  - CPU, Memory, PCI/PCI-X, AGP, Floppy disk
  - ATA and/or SCSI
  - Power
  - LEDs, speakers, switches, etc
  - Externel I/O
- Chips
  - System bus to memory
  - Peripheral bus to system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

Motherboard

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

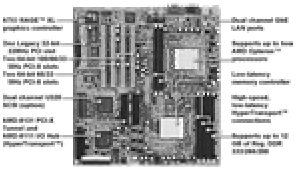
Motherboards – I2

- Supermicroi2DML-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



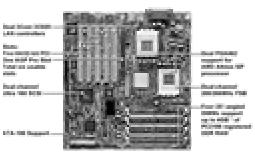
Motherboards - Opteron

- Tyan Thunder 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,



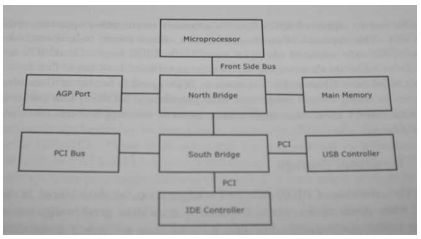
Motherboards - Opteron

- 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,



Chipsets

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



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)

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

Local Hard Disks

- Overall improvement in disk access time has been less than 10% per year
- Amdahl's law
  - Speed-up obtained by 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

RAID

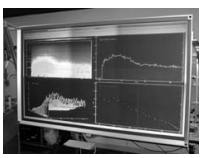
- Redundant Array of Inexpensive Disks
- Disk aggregate appear as single disk
- Adv: larger data, faster, redundancy
- Software (possibly high CPU utilization) or hardware
- RAID versions
  - RAID0: striping across multiple disks, faster reads & writes
  - RAID1: mirroring, 2 copies of data, faster read, 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

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 dedicatednet and protocol
- Network file systems : NFS & PVFS
  - Data transmitted with filesystem semantics

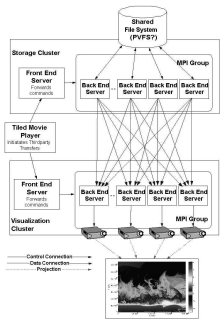
Video

- 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



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