Parallel and Distributed Computing
Chapter 2: Parallel Programming Platforms

Jun Zhang

Laboratory for High Performance Computing & Computer Simulation
Department of Computer Science
University of Kentucky
Lexington, KY 40506

2.1a: Flynn's Classical Taxonomy

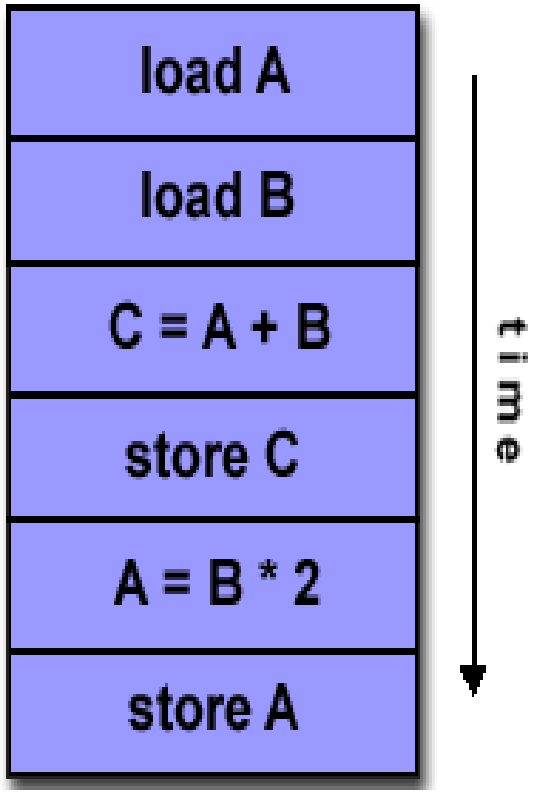
One of the more widely used parallel computer classifications, since 1966, is called Flynn's Taxonomy

It distinguishes multiprocessor computers according to the dimensions of **Instruction** and **Data**

- **SISD**: Single instruction stream, Single data stream
- **SIMD**: Single instruction stream, Multiple data streams
- **MISD**: Multiple instruction streams, Single data stream
- **MIMD**: Multiple instruction streams, Multiple data streams

2.1b: SISD Machines

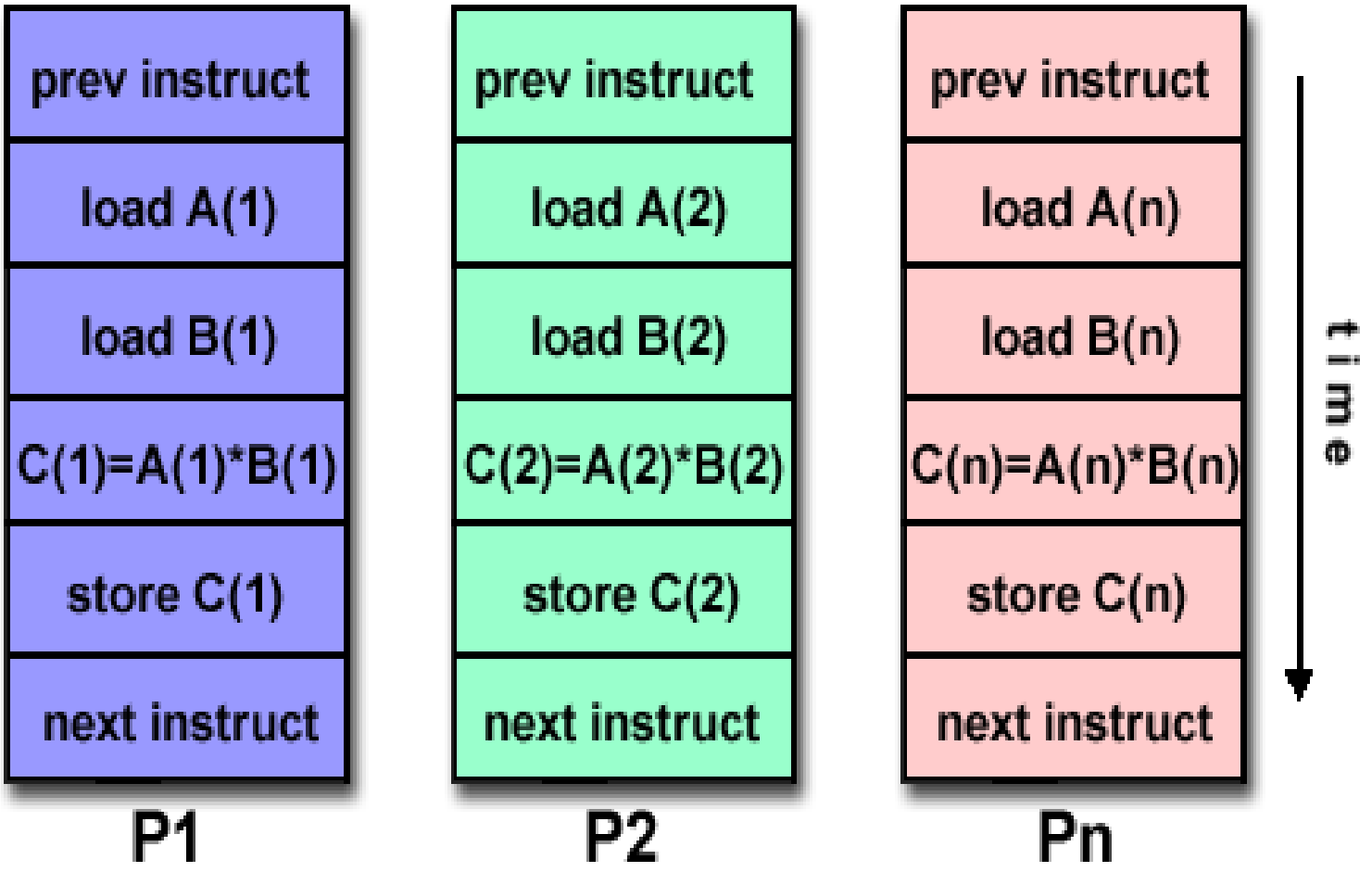
- A serial (non-parallel) computer
- Single instruction: Only one instruction stream is acted on by CPU during any one clock cycle
- Single data: Only one data stream is used as input during any one clock cycle
- Deterministic execution
- Oldest and most prevalent form computer
- Examples: Most PCs, single CPU workstations and mainframes



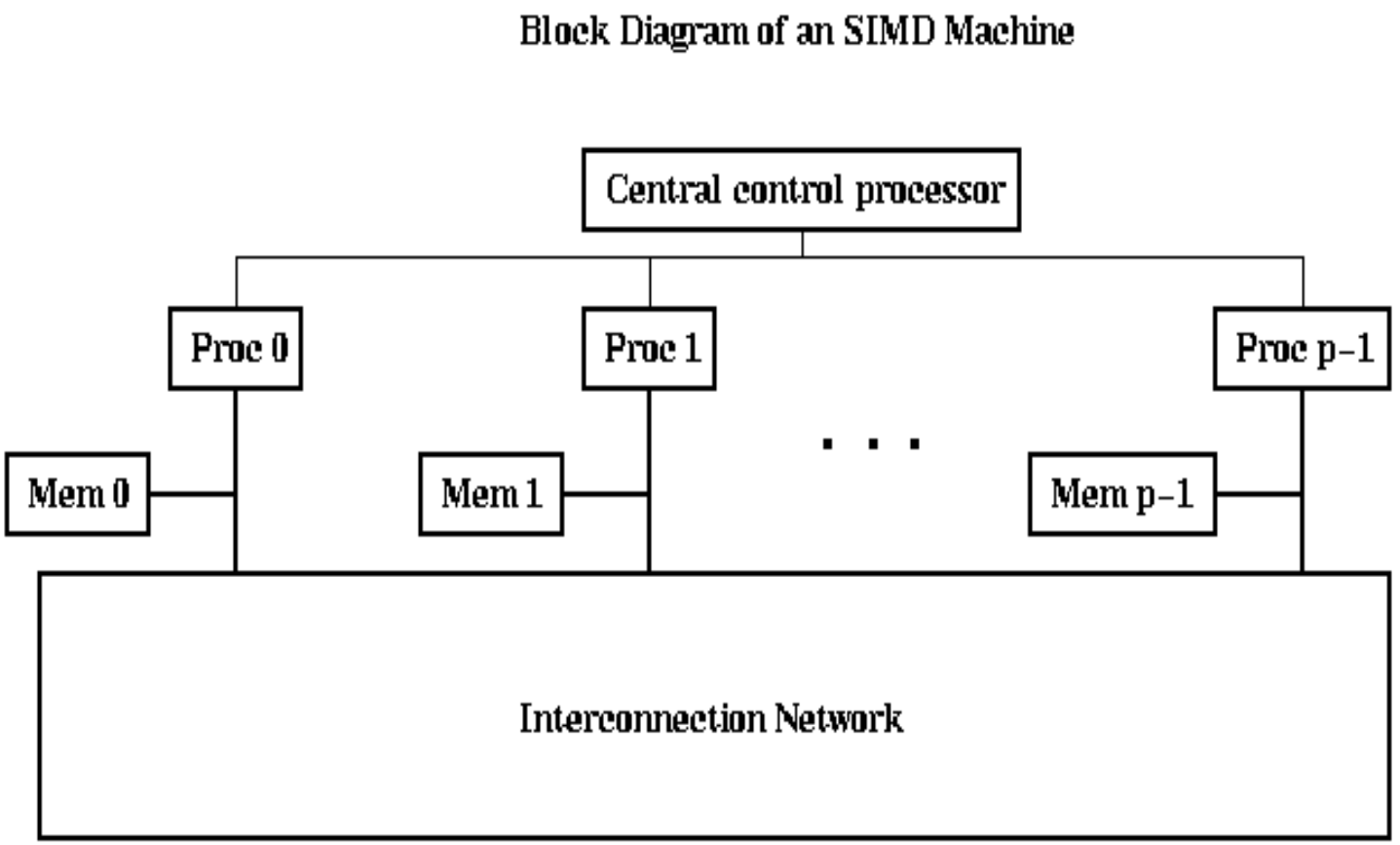
2.2a: SIMD Machines (I)

- A type of parallel computers
- Single instruction: All processor units execute the same instruction at any give clock cycle
- Multiple data: Each processing unit can operate on a different data element
- It typically has an instruction dispatcher, a very high-bandwidth internal network, and a very large array of very small-capacity instruction units
- Best suitable for specialized problems characterized by a high degree of regularity, e.g., image processing
- Two varieties: Processor Arrays and Vector Pipelines
- Examples: Connection Machines, MasPar-1, MasPar-2;
- IBM 9000, Cray C90, Fujitsu VP, etc

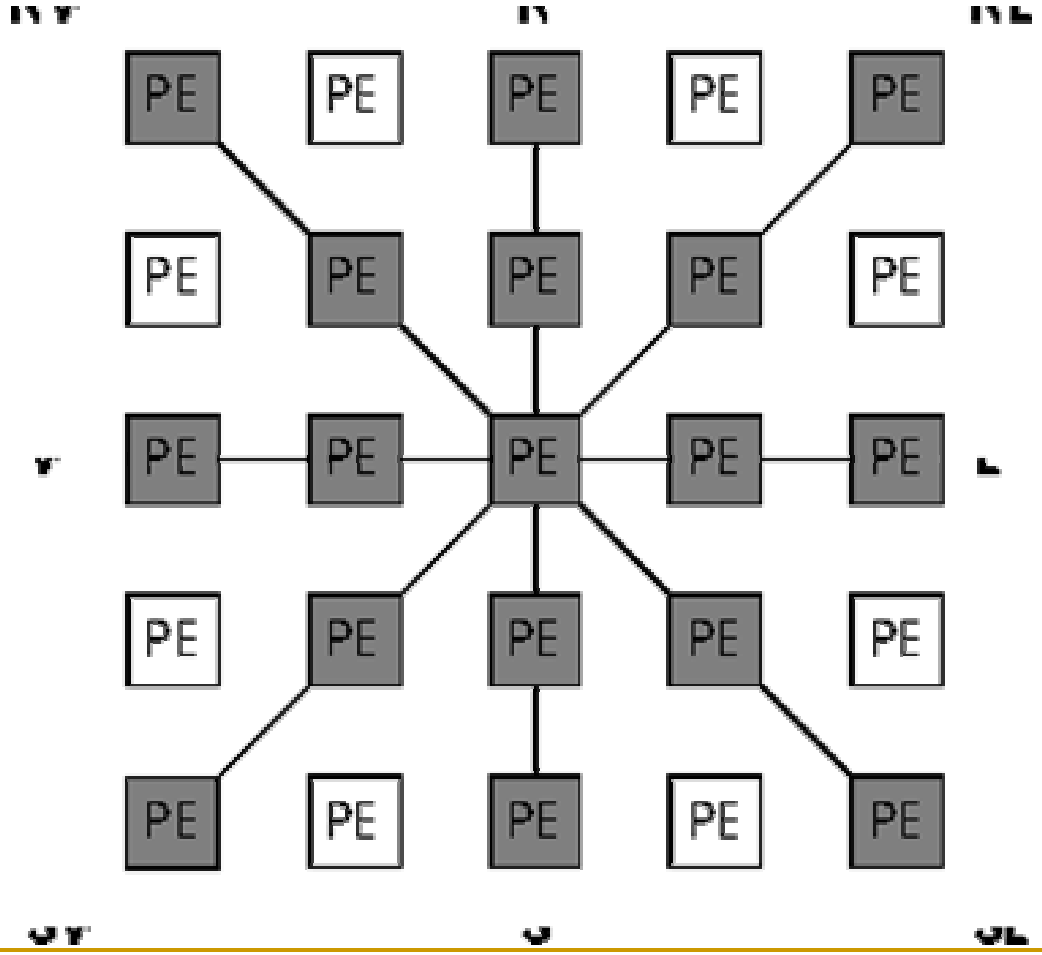
2.2b: SIMD Machines (II)



2.2c: SIMD Machine (III)

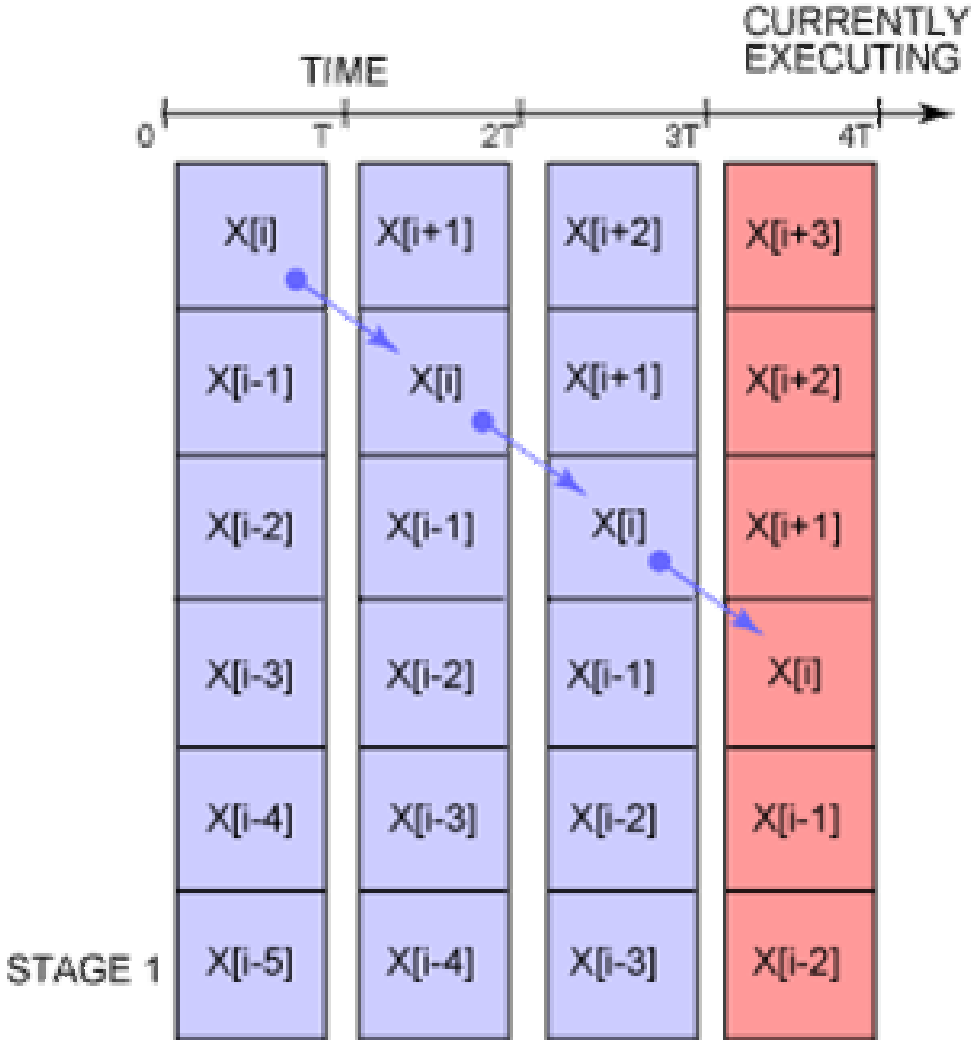


2.2d: Processing Array



2.2e: MasPar Machine

2.2f: Pipelined Processing



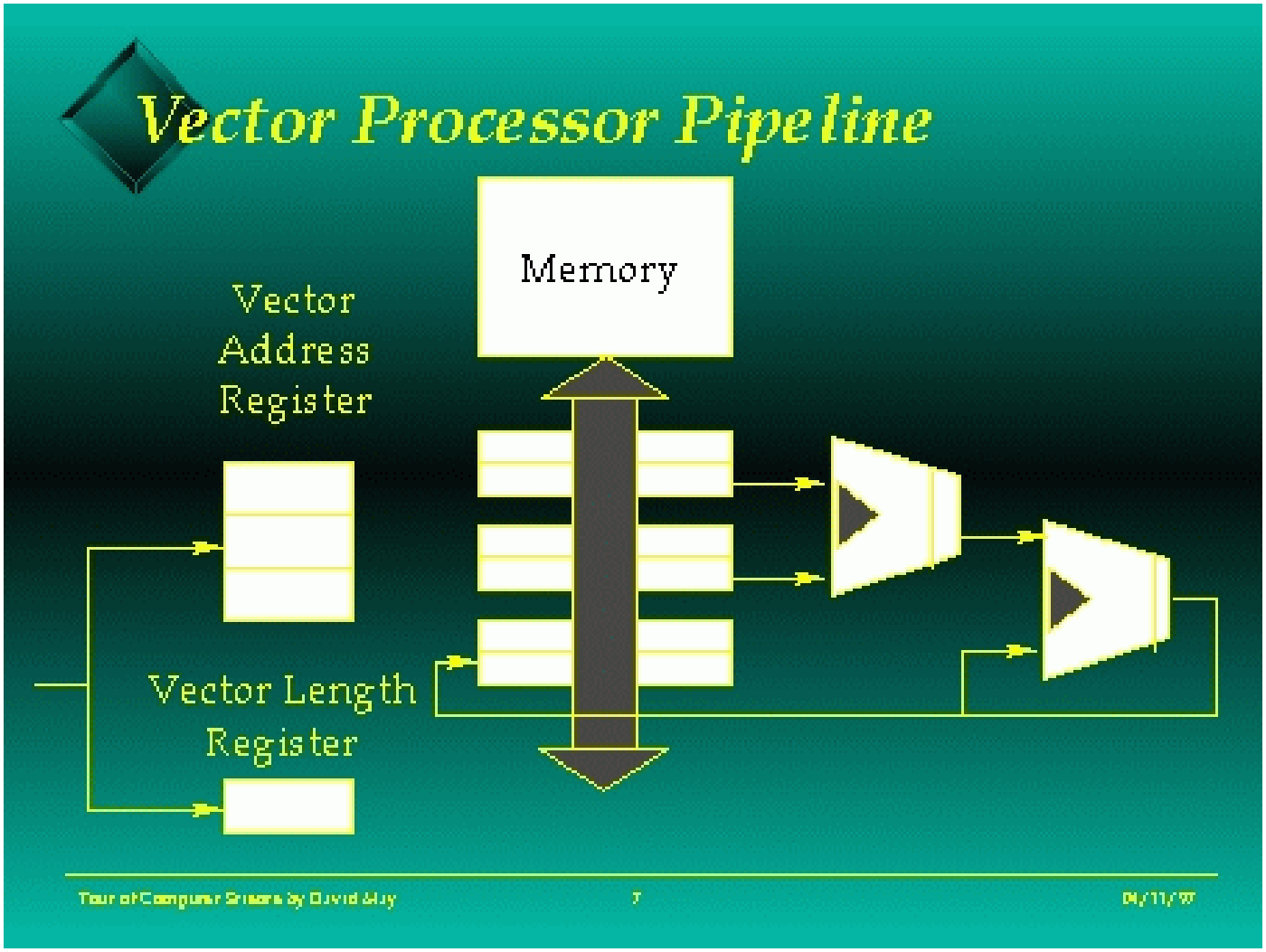
A six step pipeline with IEEE arithmetic hardware

Parallelization happens behind the scene

Not true parallel computers

2.2g: Assembly Line

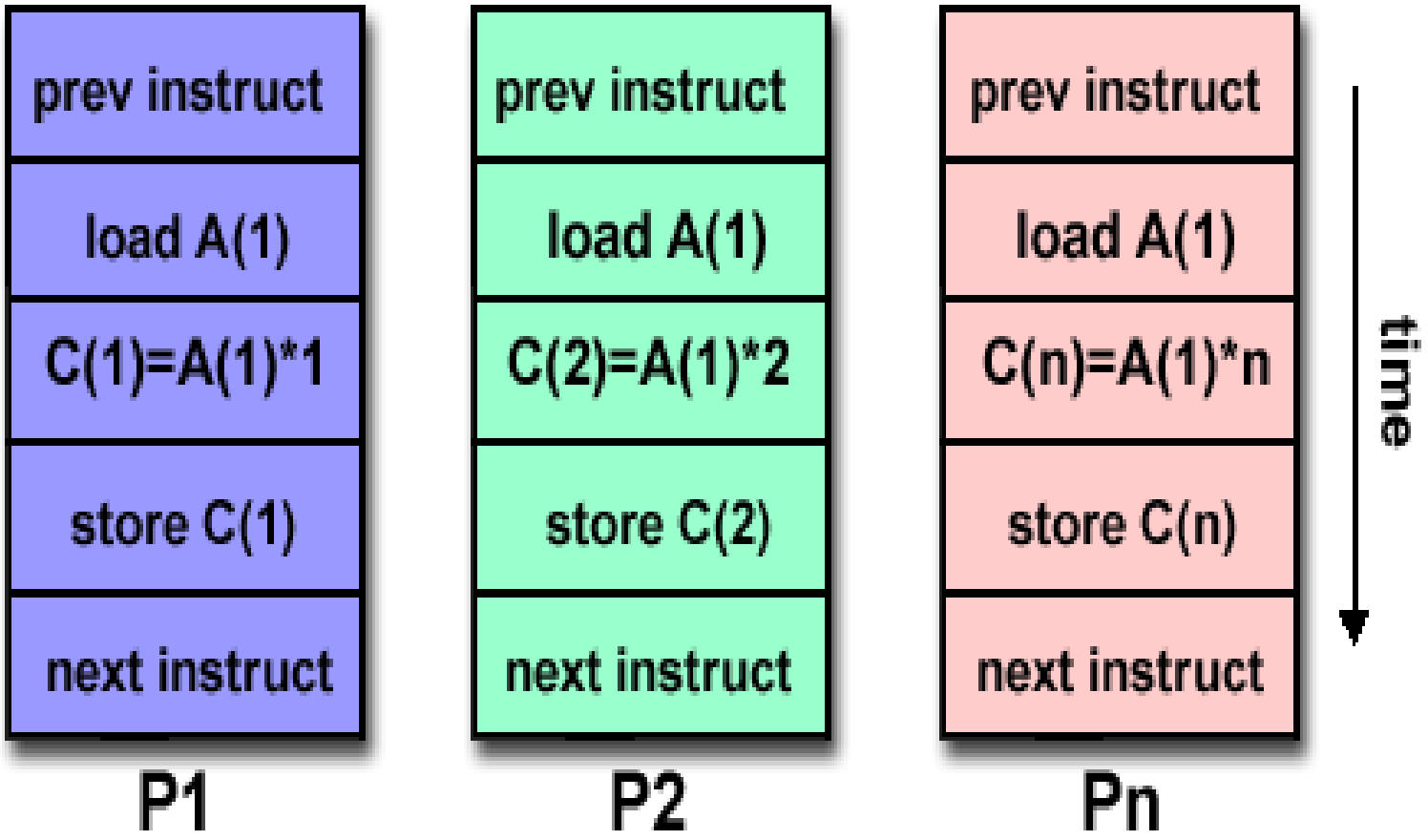
2.2h: Vector Processor Pipeline



2.3a: MISD Machines (I)

- A single data stream is fed into multiple processing units
- Each processing unit operates on the data independently via independent instruction streams
- Very few actual machines: CMU's C.mmp computer (1971)
- Possible use: multiple frequency filters operating on a single signal stream

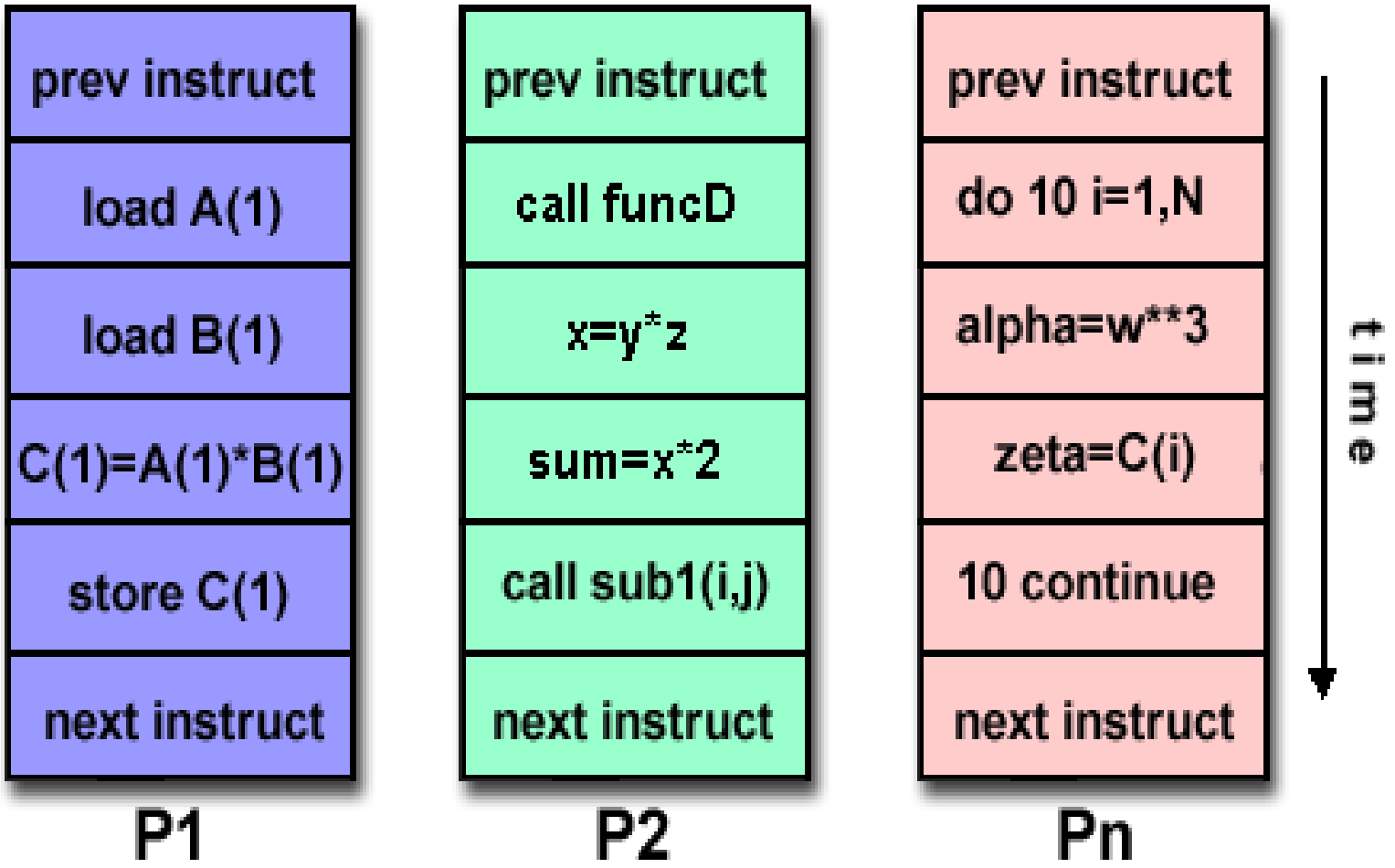
2.3b: MISD Machines (II)



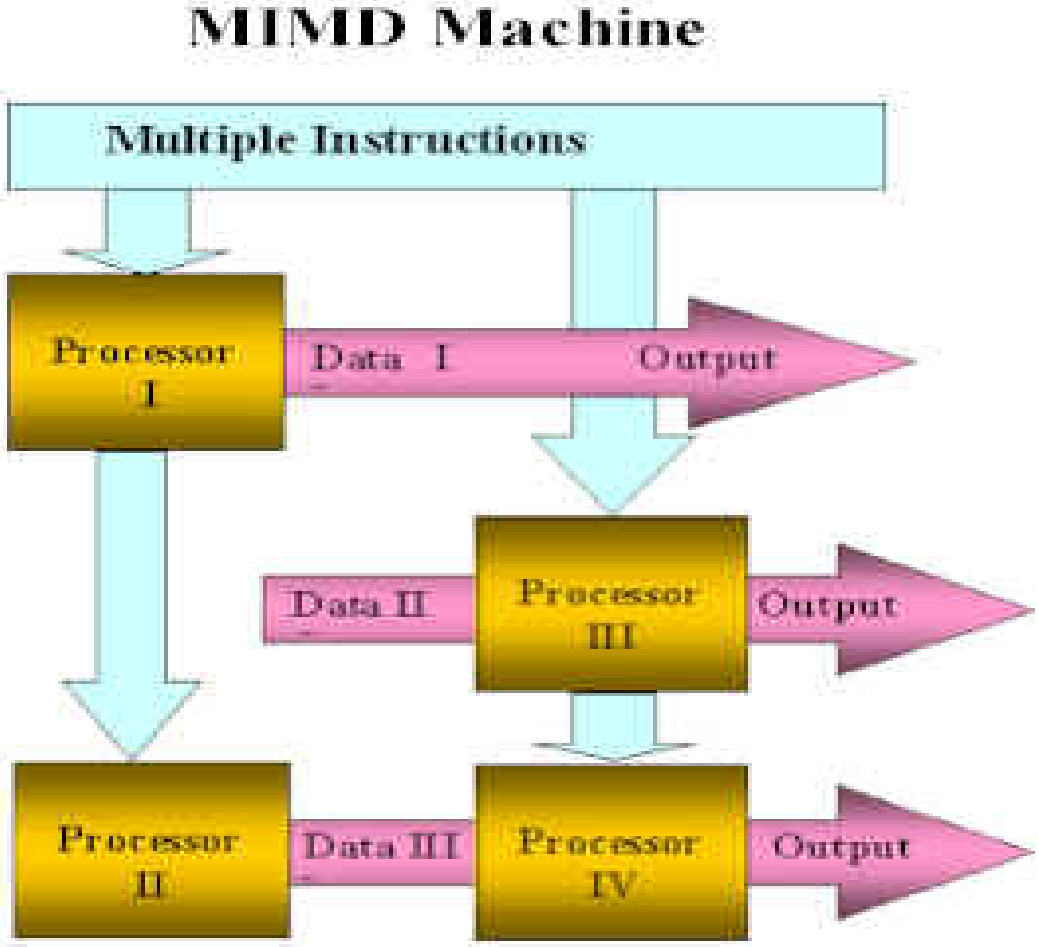
2.4a: MIMD Machines (I)

- Multiple instruction: Every processor may execute a different instruction stream
- Multiple data: Every processor may work with a different data stream
- Execution can be synchronous or asynchronous, deterministic or non-deterministic
- Examples: most current supercomputers, grids, networked parallel computers, multiprocessor SMP computer

2.4b: MIMD Machines (II)



2.4c: MIMD Machines (III)



2.4d: MIMD Machines (T3E-Cray)

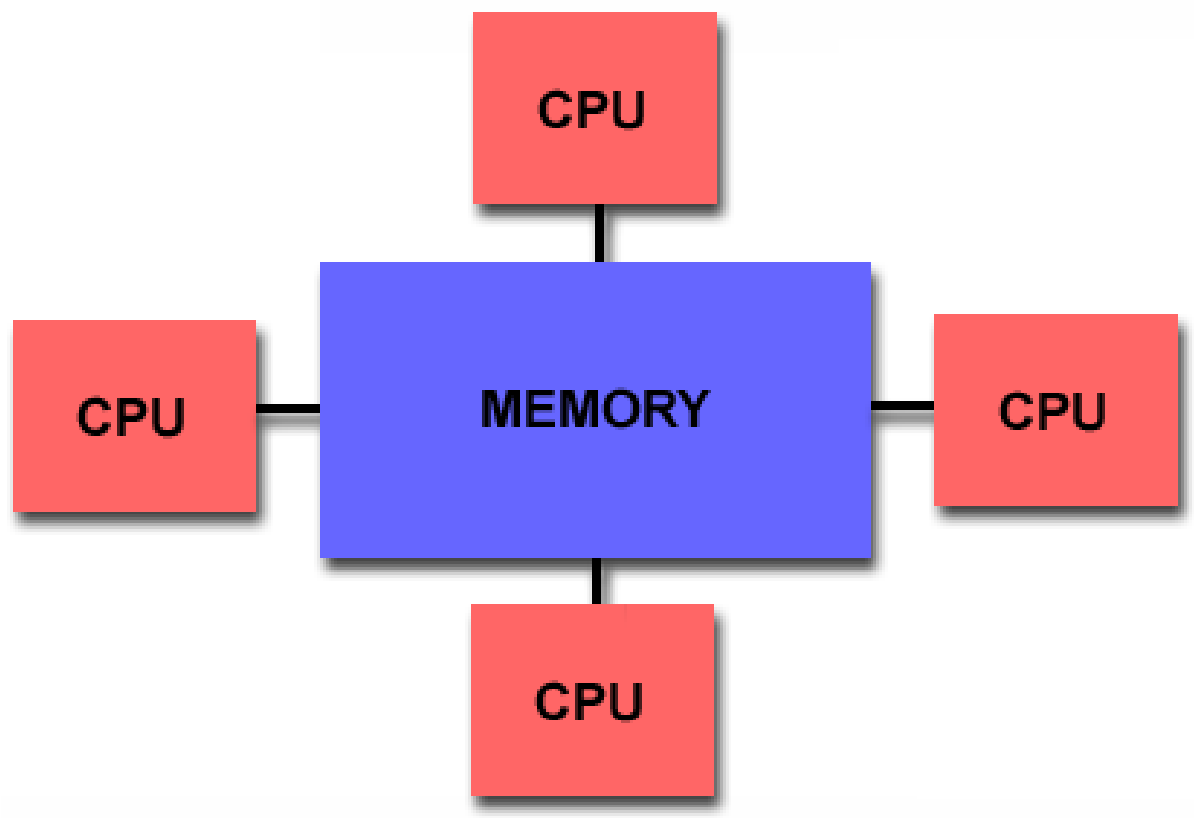
2.5: Shared Memory and Message Passing

- Parallel computers can also be classified according to memory access
- **Shared memory computers**
- **Message-passing (distributed memory) computers**
- Multi-processor computers
- Multi-computers
- Clusters
- Grids

2.6a: Shared Memory Computers

- All processors have access to all memory as a global address space
- Multiple processors can operate independently, but share the same memory resources
- Changes in a memory location effected by one processor are visible to all other processors
- Two classes of shared memory machines: UMA and NUMA, (and COMA)

2.6b: Shared Memory Architecture

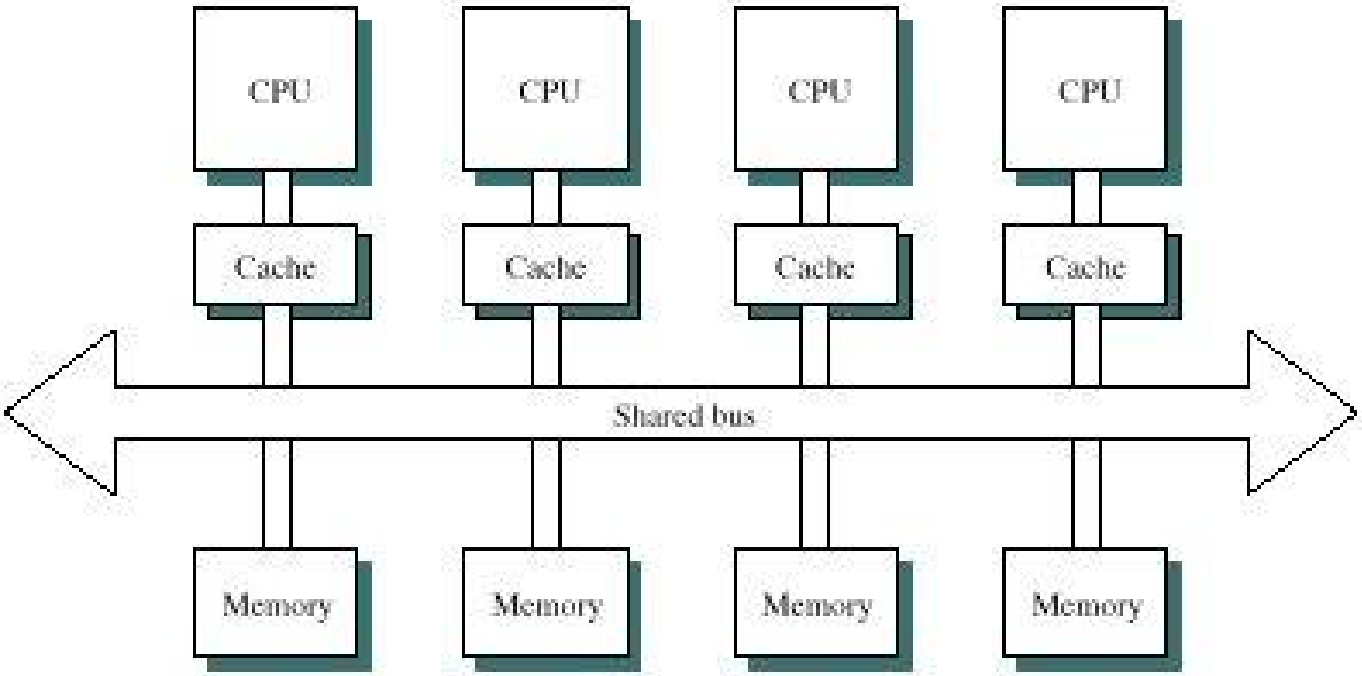


2.6c: Uniform Memory Access (UMA)

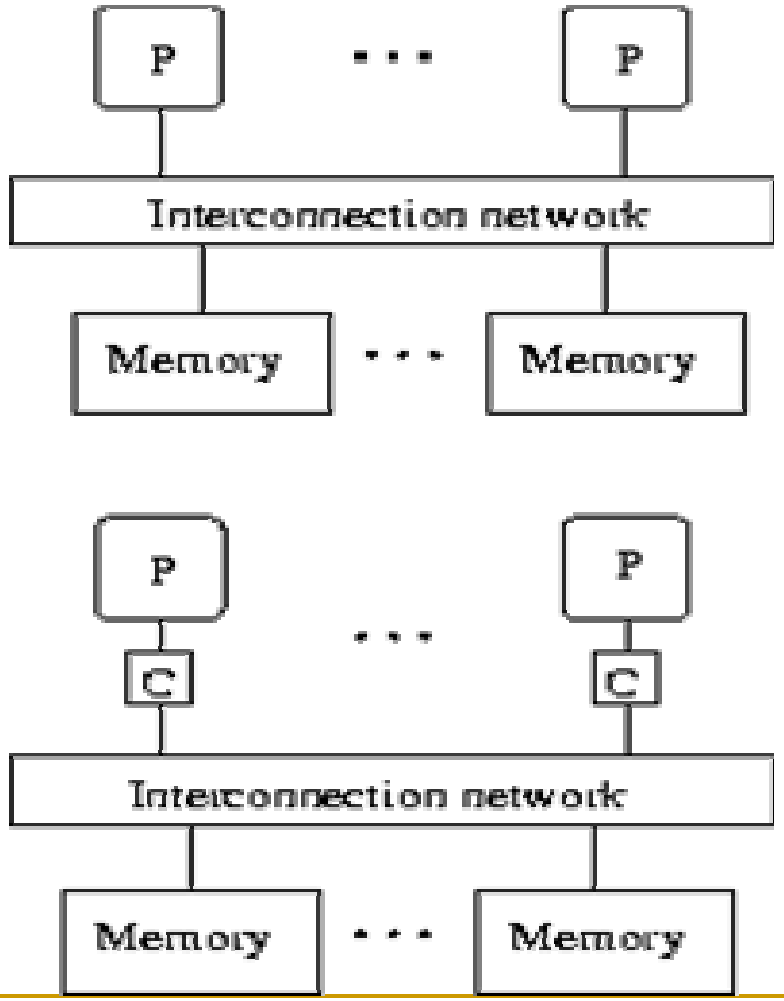
- Most commonly represented today by Symmetric Multiprocessor (SMP) machines
- Identical processors
- Equal access and access time to memory
- Sometimes called CC-UMA – Cache Coherent UMA
- **Cache Coherence:**

  If one processor updates a location in shared memory, all the other processors know about the update. Cache coherence is accomplished at the hardware level

2.6d: Symmetric Multiprocessor (SMP)



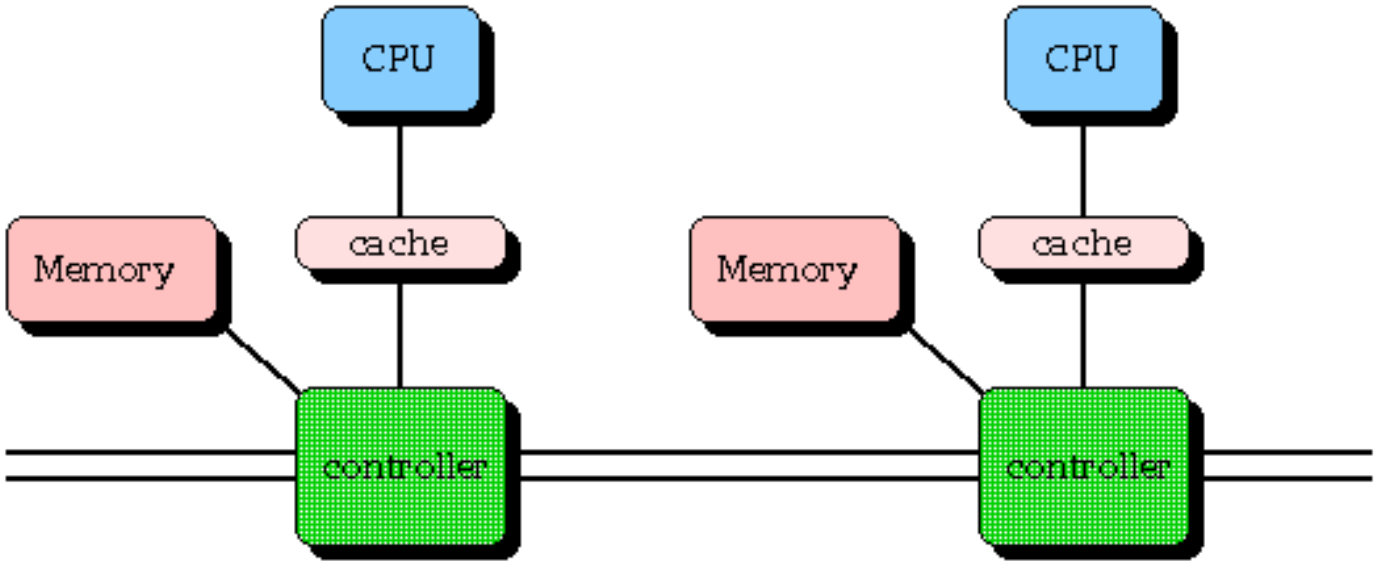
2.6e: UMA –with and without caches



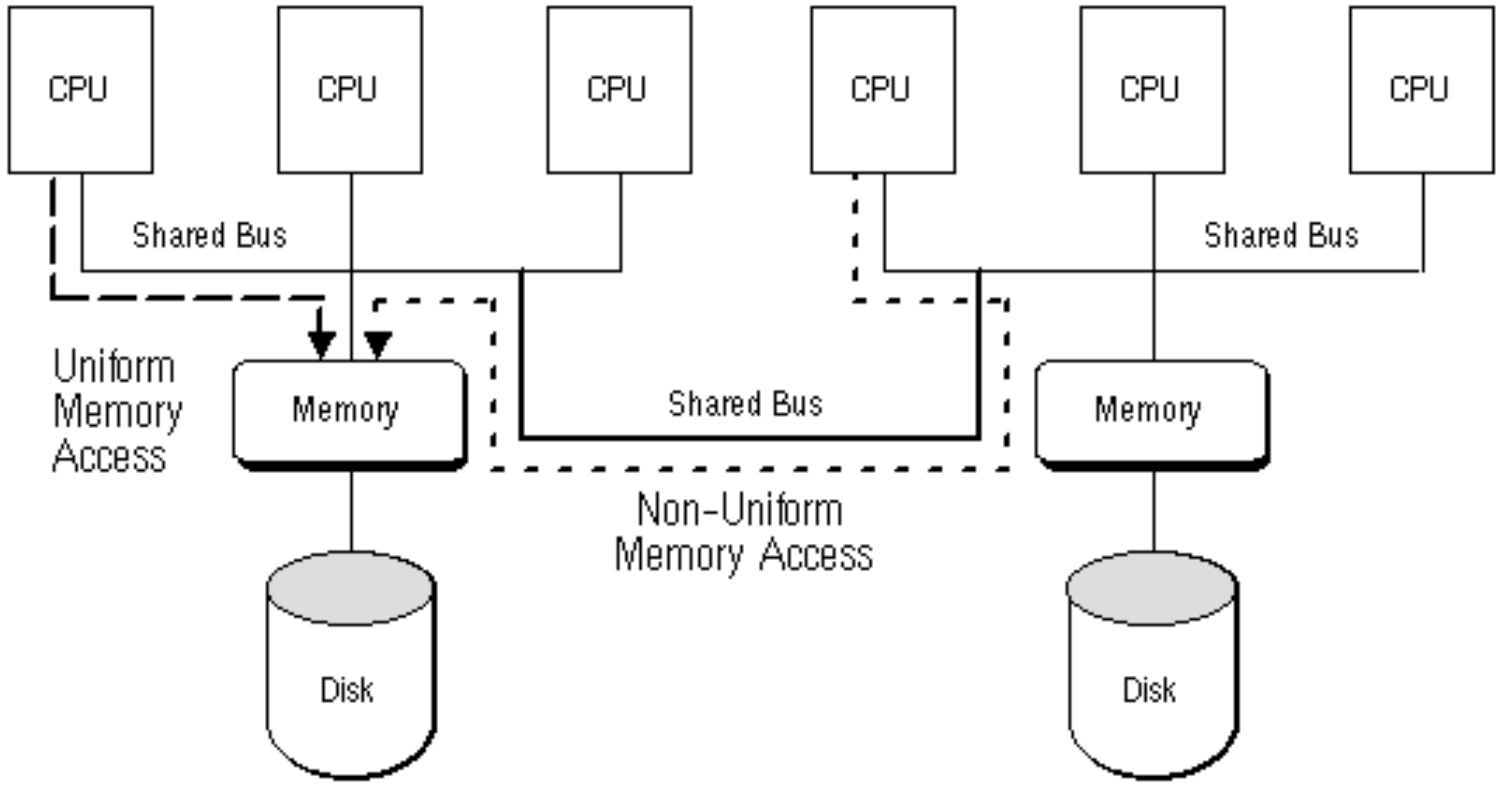
2.6f: Nonuniform Memory Access (NUMA)

- Often made by physically linked two or more SMPs
- One SMP can directly access memory of another SMP
- Not all processors have equal access time to all memories
- Memory access across link is slower
- If cache coherence is maintained, they may also be called CC-NUMA

2.6g: NUMA



2.6h: Hybrid Machine (UMA & NUMA)



2.6i: Advantages of Shared Memory Machines

- Global address space provides a user-friendly programming environment to memory access
- Data sharing between tasks is both fast and uniform due to proximity of memory to CPUs

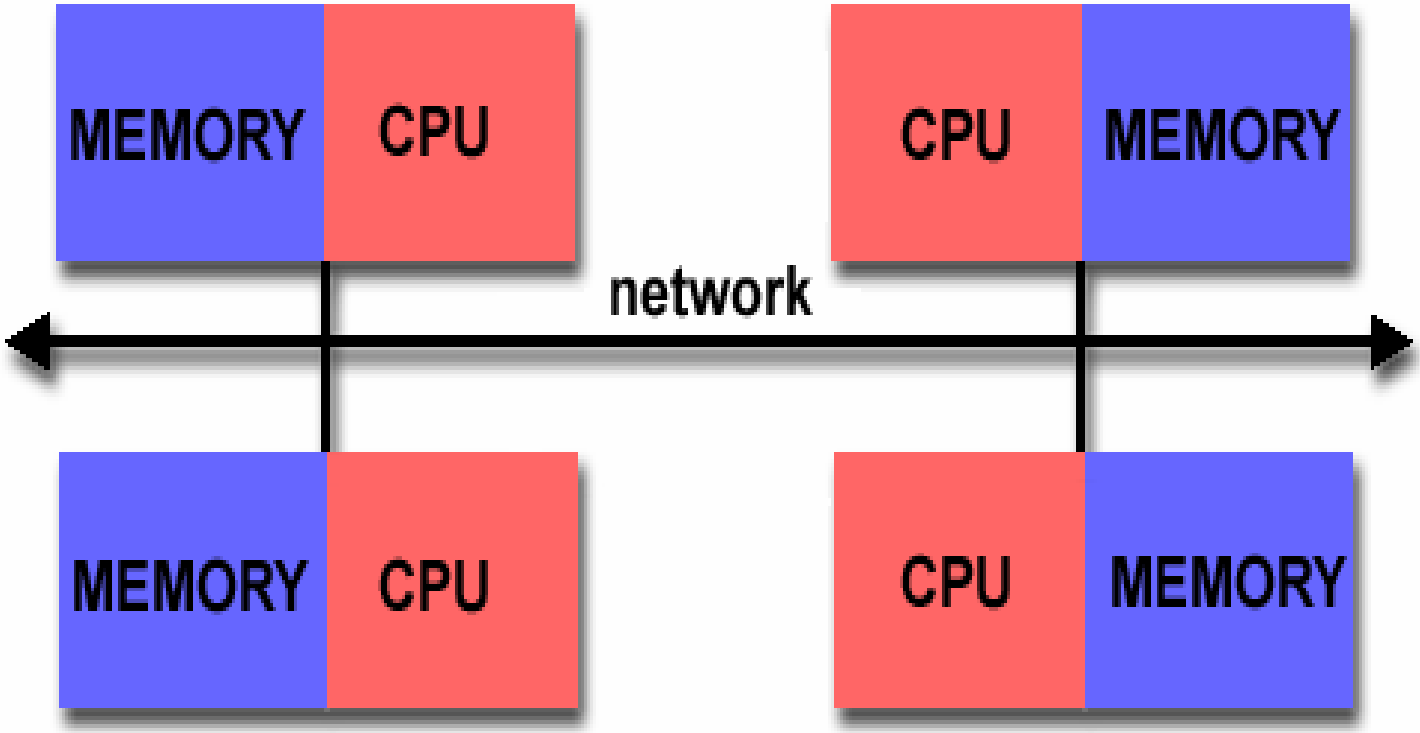
2.6j: Disadvantages of Shared Memory Machines

- Lack of scalability between memory and CPUs: Adding processors can geometrically increase traffic on the shared memory-CPU path and for cache coherence management
- Programmer's responsibility for synchronization constructs (correct access to memory)
- Expensive to design shared memory computers with increasing numbers of processors

2.7a: Distributed Memory Computers

- Processors have their own local memory
- It requires a communication network to connect inter-processor memory
- Memory addresses in one processor do not map to another processor – no concept of global address space across all processors
- The concept of cache coherence does not apply
- Data are exchanged explicitly through message-passing
- Synchronization between tasks is the programmer's responsibility

2.7b: Distributed Memory Architecture



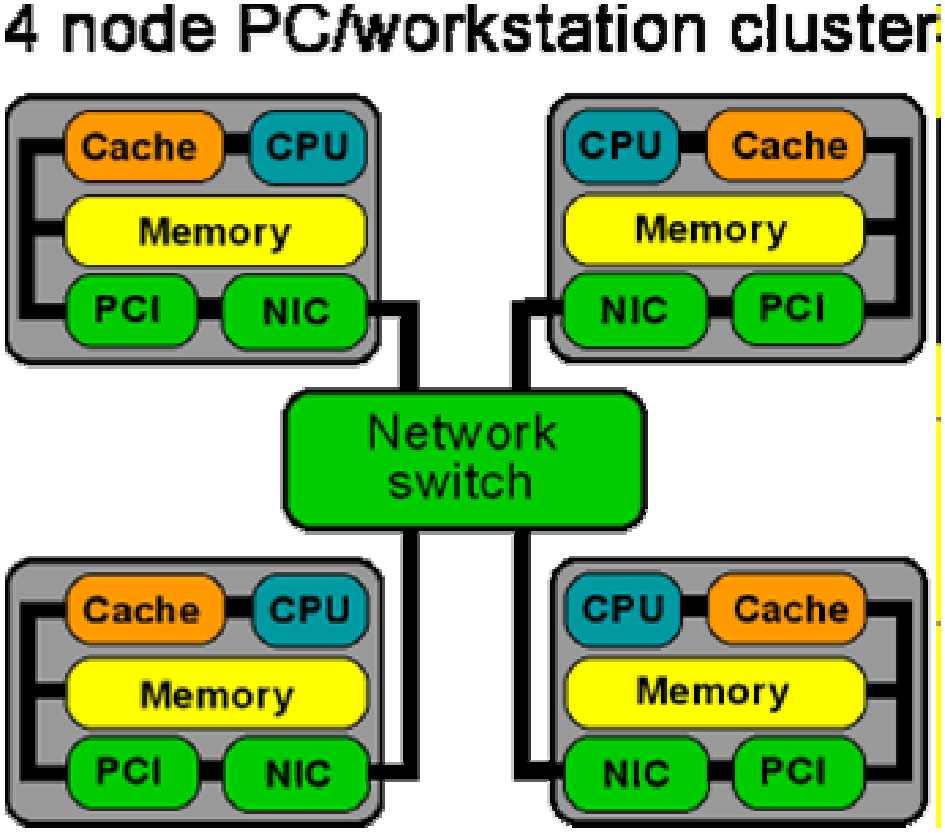
2.7c: Advantages of Distributed Memory Machines

- Memory is scalable with the number of processors
- Increase the number of processors, the size of memory increases proportionally
- Each processor can rapidly access its own memory without interference and without the overhead incurred with trying to maintain cache coherence
- Cost effectiveness: can use commodity, off-the-shelf processors and networking

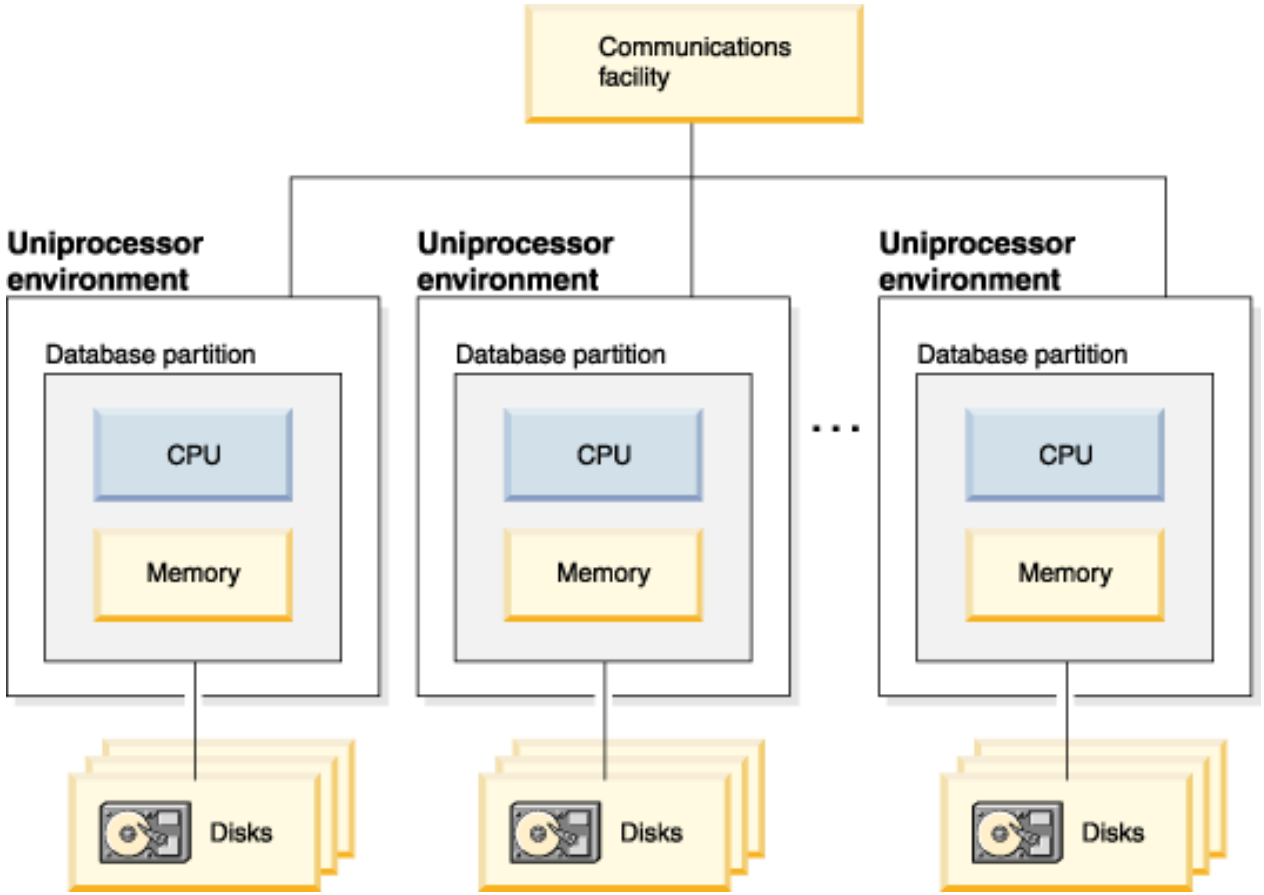
2.7d: Disadvantages of Distributed Memory Machines

- Difficult to program: Programmer has to handle data communication between processors
- Nonuniform memory access (NUMA) times
- It may be difficult to map existing data structures, based on global memory, to distributed memory organization

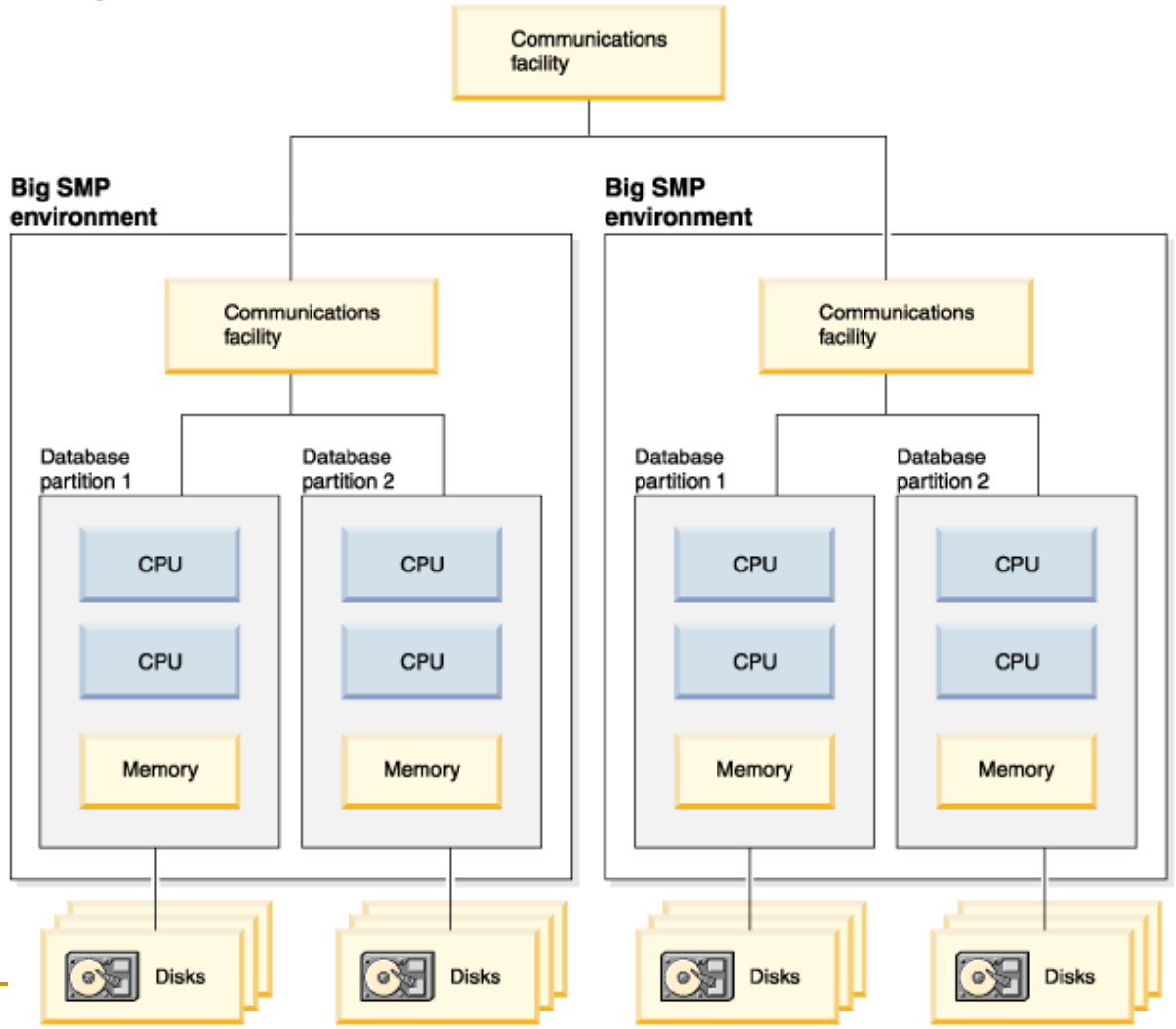
2.7e: Networked Cluster of Workstatsions (PCs)



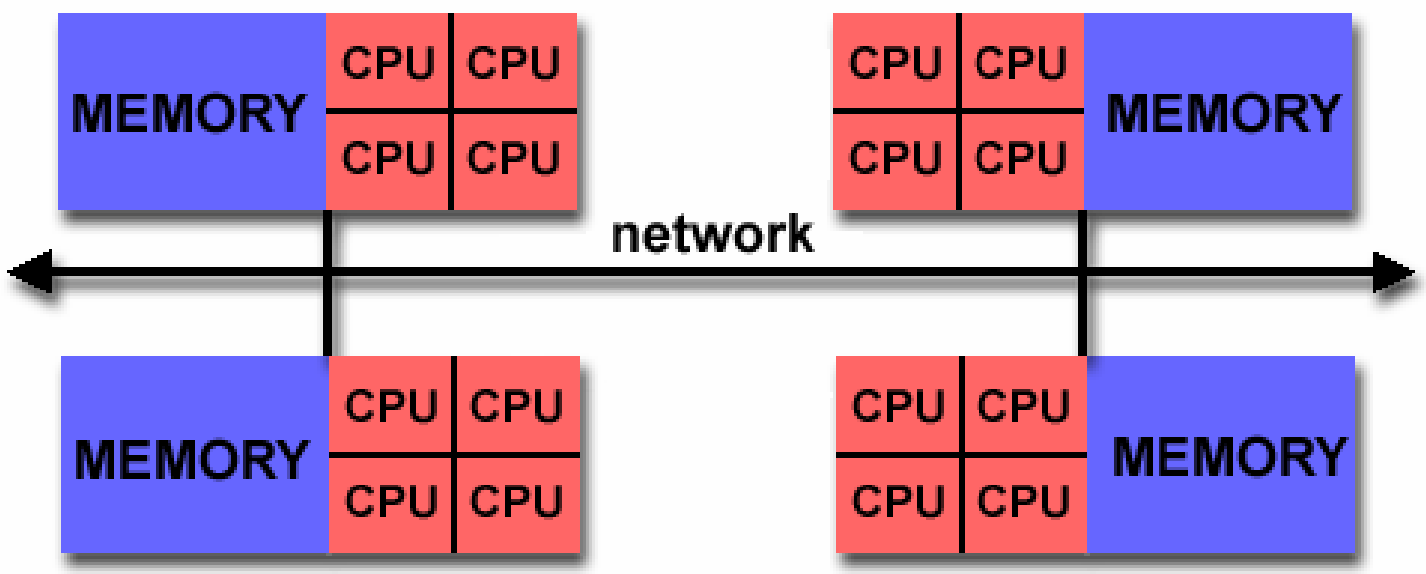
2.7f: Massively Parallel Processing (MPP) Environment



2.7g: Distributed Shared Memory (DSM) Computer



2.7h: Hybrid and Combinations (Very Large Parallel Computing System)



2.8: Architectural Design Tradeoffs

	Shared memory	Distributed memory
Programmability	easier	harder
Scalability	harder	easier

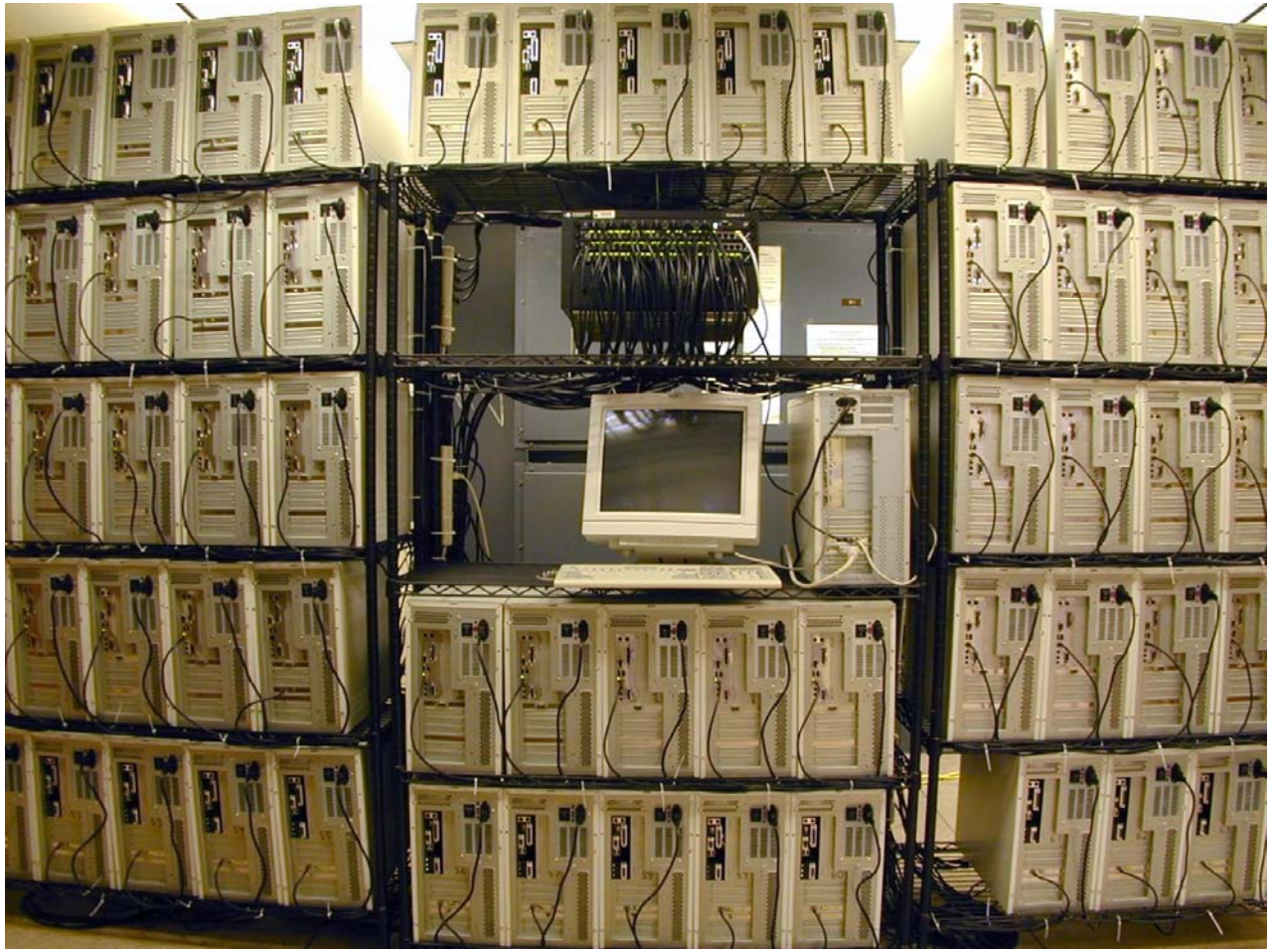
2.9:Parallel Architectural Issues

- Control mechanism: SIMD vs MIMD
- Operation: synchronous vs asynchronous
- Memory organization: private vs shared
- Address space: local vs global
- Memory access: uniform vs nonuniform
- Granularity: power of individual processors
  coarse grained system vs fine grained system
- Interconnection network topology

2.10a: Beowulf Cluster System

- A cluster of tightly coupled PC's for distributed parallel computation
- Moderate size: normally 16 to 32 PC's
- Promise of good price/performance ratio
- Use of commodity-of-the-self (COTS) components (PCs, Linux, MPI)
- Initiated at NASA (Center of Excellence in Space Data and Information Sciences) in 1994 using 16 DX4 processors
- http://www.beowulf.org

2.10b: NASA 128-Processor Beowulf Cluster



2.11a: Sun Enterprise 10000 Server (I)

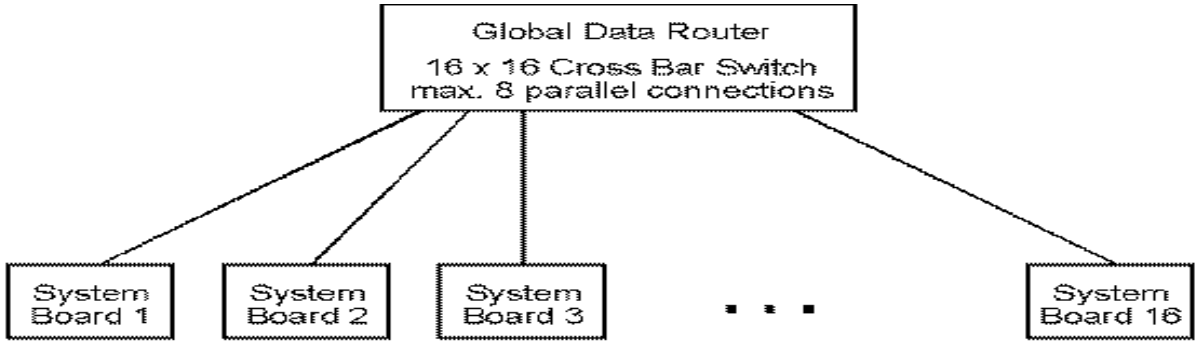
- Starfire is a shared-address space machine
- 4 to 64 processors, each is an UltraSPARC II

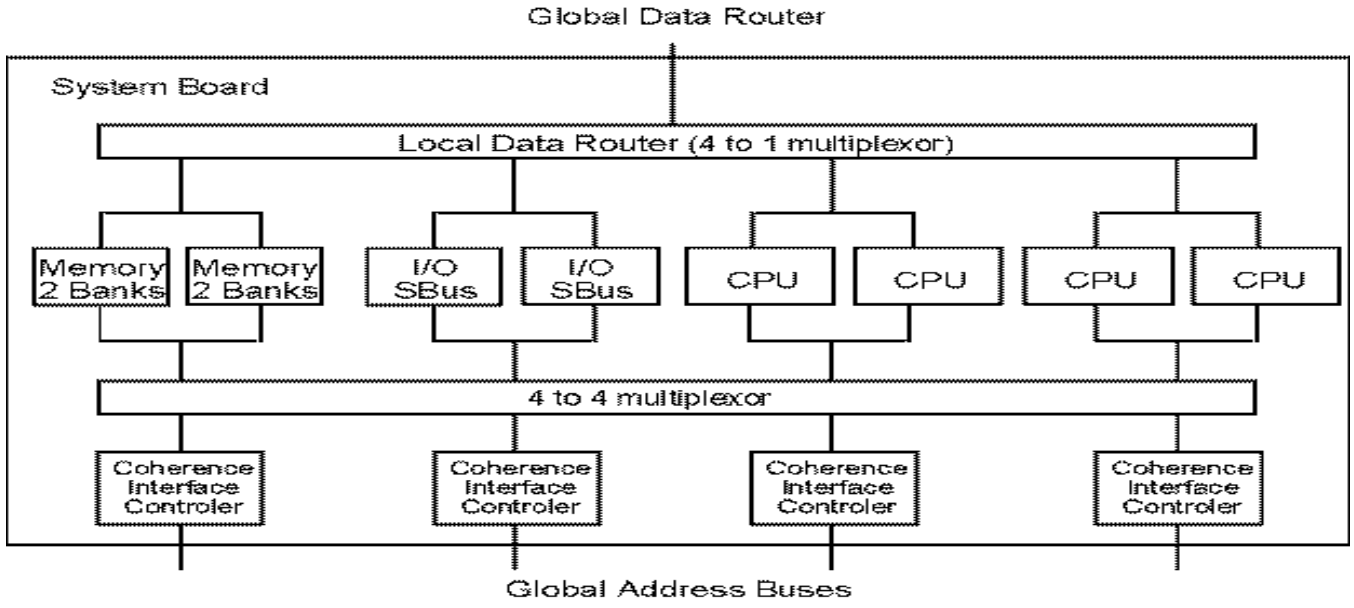
  running at 400 MHz with a peak rate 800 MHz
- System board has up to 4 CPUs and 4GB RAM local memory
- Up to 64 GB memory in total
- 16 such boards can be connected via a Gigaplane-XB interconnect

2.11b: Sun Enterprise 10000 Server (II)

- Intra- and inter-board connectivity uses different interconnection networks
- Intra-board connectivity is by a split transaction system bus (Gigaplane bus)
- Global data router is a 16X16 non-blocking crossbar for the data buses and four global address buses
- http://www.sun.com/servers/highend/e10000/

2.11c: Sun Enterprise 10000 Server (III)





2.11d: Sun Enterprise 10000 Server (IV)

2.12a: SGI Origin Servers (I)

- SGI Origin 3000 Servers support up to 512 MPIS 14000 processors in a NUMA cache coherent shared-address-space configuration
- Each processor operates at 500 MHz with a peak rate of 1.0 gigaflops
- Modular framework with 4 CPUs and 8GB of RAM (C-Brick)
- Interconnection by crossbar between C-Bricks and R-Bricks (routers)

2.12b: SGI Origin Servers (II)

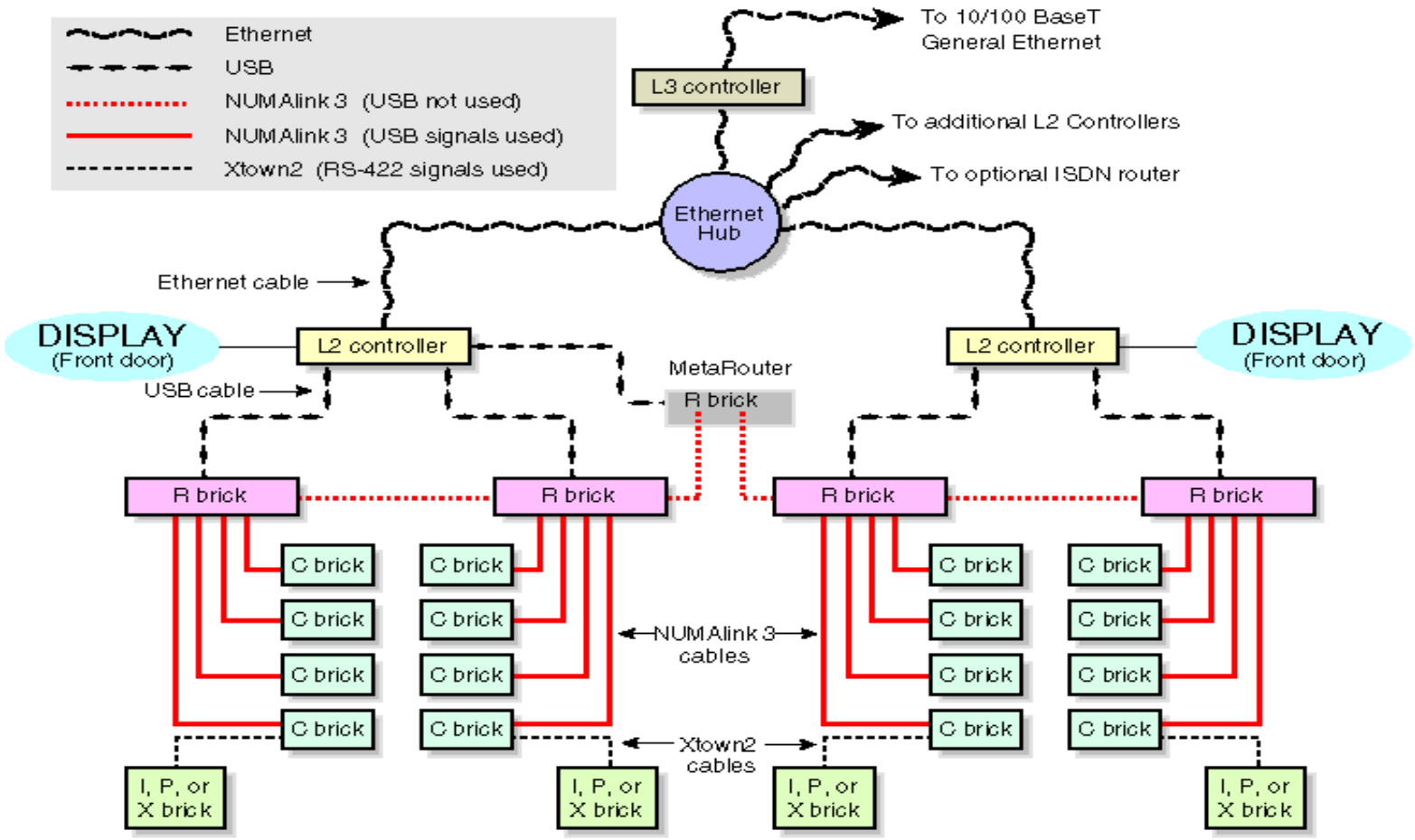
- Larger configuration are built by connecting multiple C-Bricks via R-Bricks using 6-port or 8-port crossbar switches and metarouters with full-duplex links operating at 1.6GB/s
- http://www.sgi.com/products/servers/origin/3000/overview.html

2.12c: C-Brick and R-Brick

C Brick

R Brick

2.12d: SGI Origin Server

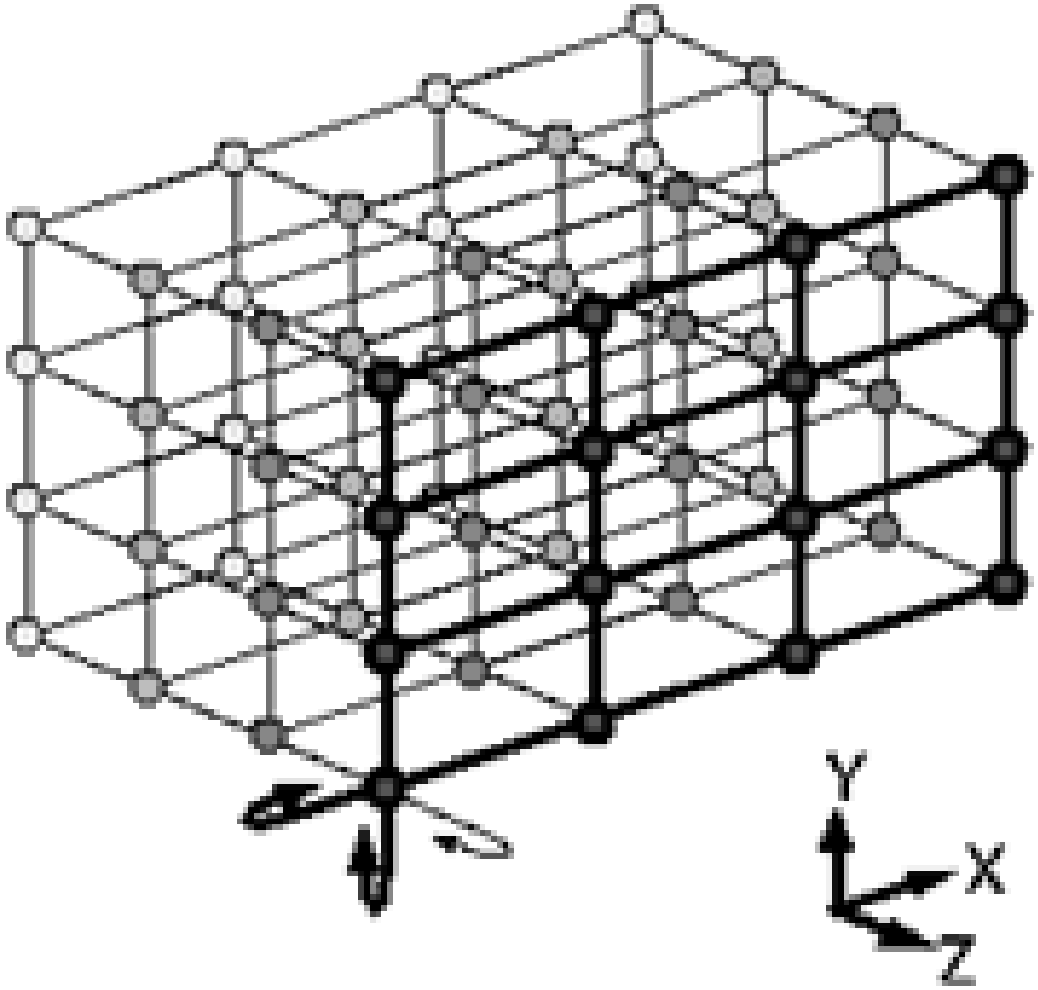


2.12e: SGI Origin Server Machine

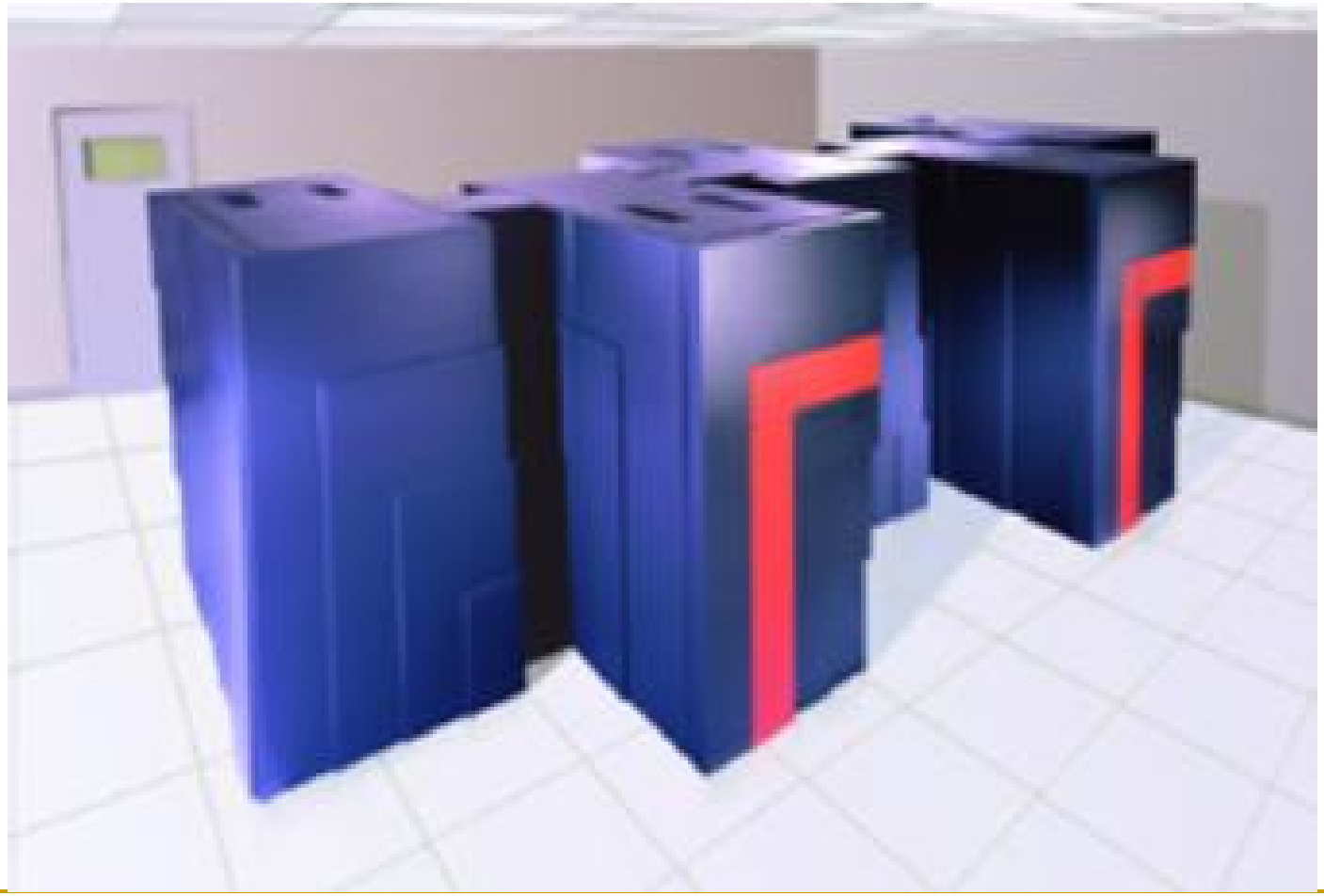
2.13a: Cray T3E/1350

- Use Alpha 21164A processor with 4-way superscalar architecture, 2 floating point instruction per cycle
- CPU clock 675 MHz, with peak rating 1.35 Gigaflops, 512 MB local memory
- Parallel systems with 40 to 2176 processors (with modules of 8 CPUs each)
- 3D torus interconnect with a single processor per node
- Each node contains a router and has a processor interface and six full-duplex link (one for each direction of the cube)

2.13b: Cray T3E Topology



2.13c: Cray T3E Machine



2.14a: UK HP Superdome Cluster

- 4 HP superdomes
- 256 total processors (64 processors per host/node)
- Itanium-2 processor at 1.6 TF peak rate
- 2 GB memory per processor
- 7 Terabytes of total disk space
- High speed, low latency Infiniband internal interconnect

2.14b: HP Superdome Cluster

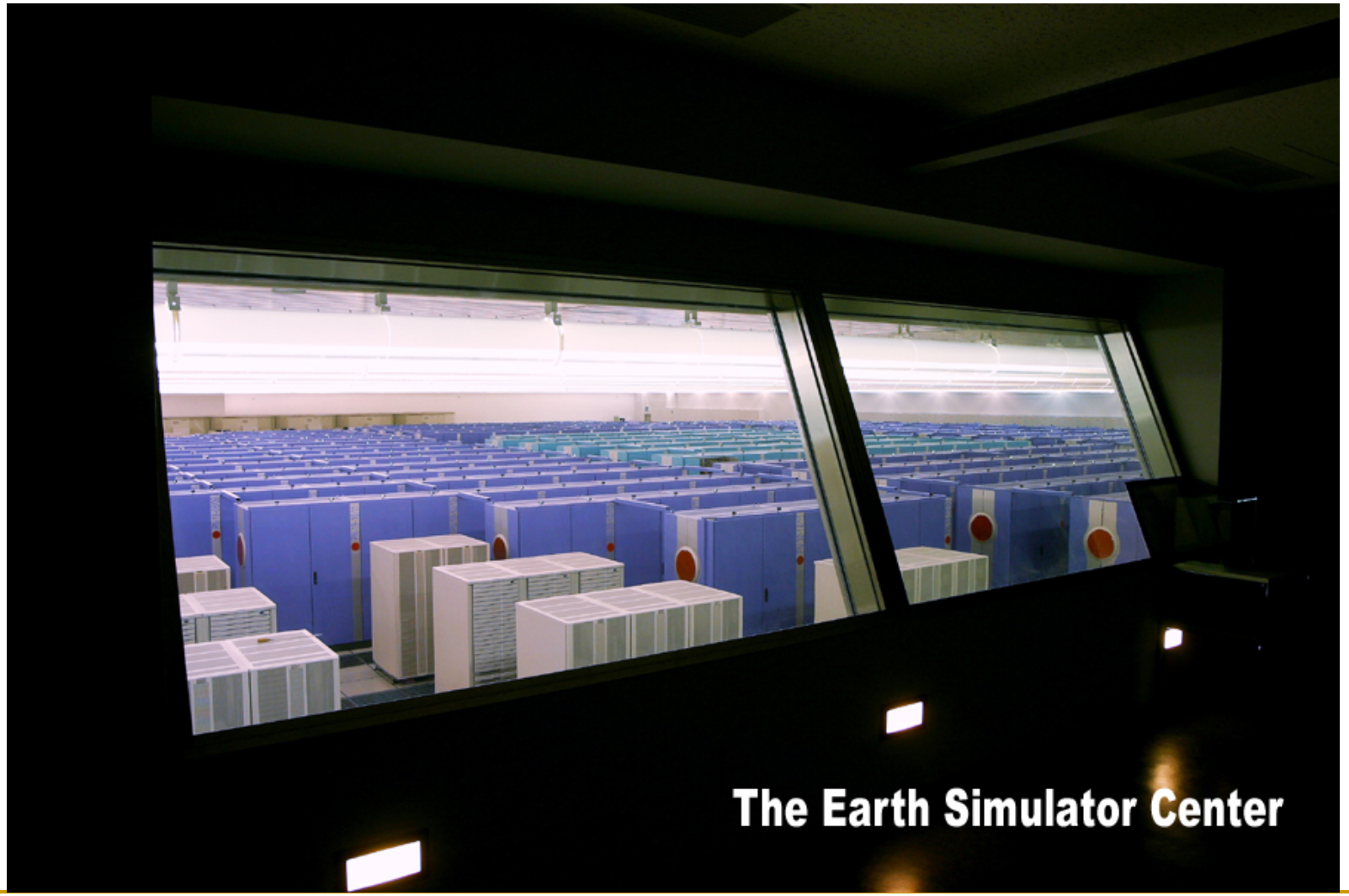
2.15a: The Earth Simulator (I)

- Each processor of the Earth Simulator runs at 2ns per cycle, with 16GB shared memory
- Total number of processors is 5,120. The aggregated peak performance is 40 TFlops with 10 TB memory
- It has a single stage crossbar (1800 miles of cable), 83,000 copper cables, 16 GB/s cross station bandwidth
- 700 TB disk space and 1.6PB mass storage

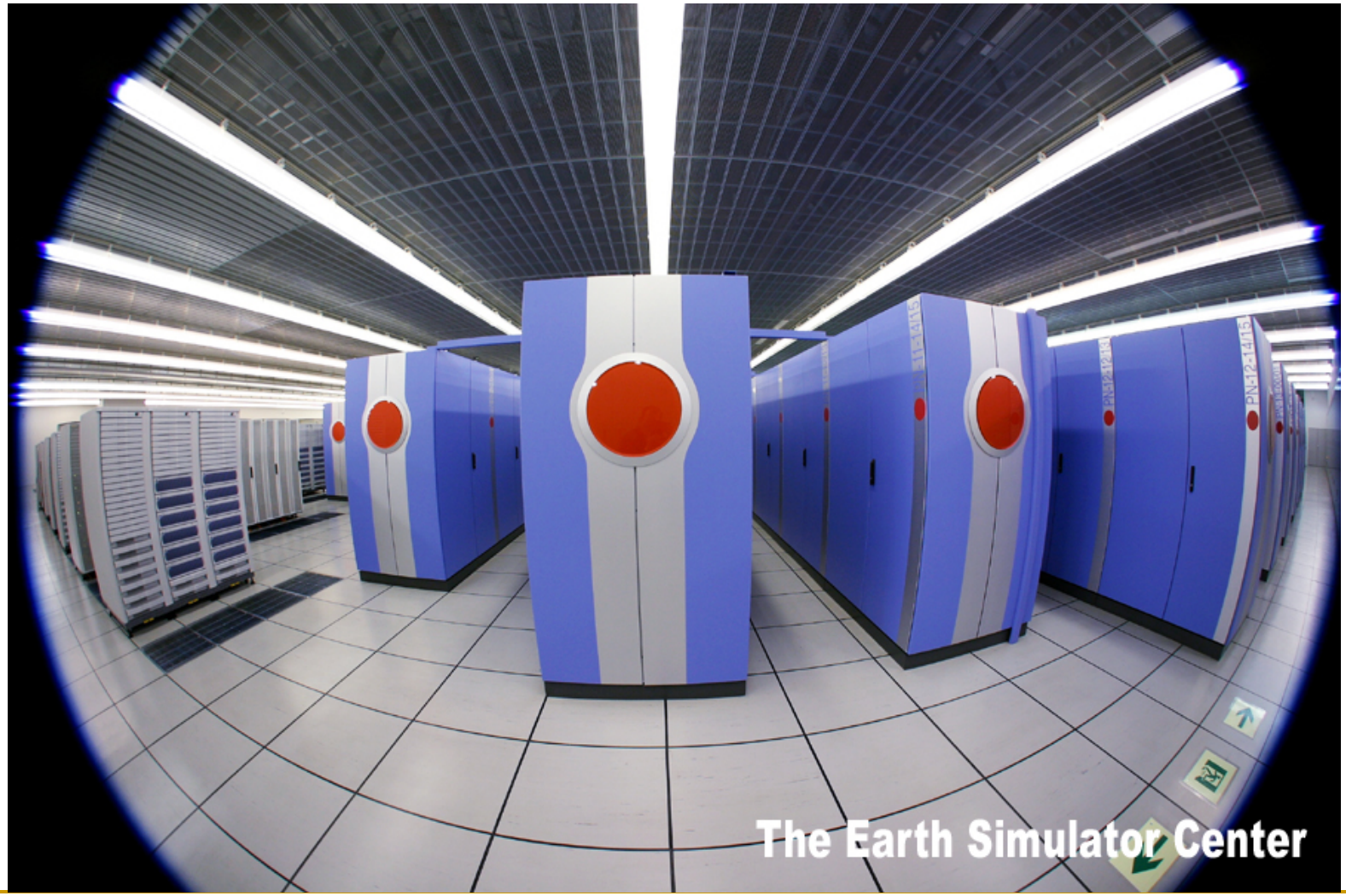
2.15b: The Earth Simulator (II)

- One node = 8 vector processors with a peak performance of 8G Flops (640X8 = 5,120)
- Area of the computer = 4 tennis courts, 3 floors
- Sum of all the US Department supercomputers = 24 TFlops/s in 2002
- Number 1 in 2002 top 500 list (five consecutive lists)
- http://www.es.jamstec.go.jp/esc/eng/

2.15c: The Earth Simulator Machine (I)



2.15d: The Earth Simulator Machine (II)



2.16a: IBM Blue Gene (I)

- In December 1999, IBM Research announced a 5-year $100M project, named Blue Gene, to develop a petaflop computer for research in computational biology
- Proposed architecture: SMASH (simple, multiple, self-healing)
- Collaboration between IBM Thomas J. Watson Research Center, Lawrence Livermore National Laboratory, US DOE and academia

2.16b: IBM Blue Gene (II)

- 1 million CPUs of 1 gigflop each = 1 petaflop
- 32 CPUs on a single chip, 16 MB memory
- 64 chips are placed on a board of 20 inches
- 8 boards form a tower, and 64 towers are connected into a 8X8 grid to form Blue Gene
- Blue Gene/L, Blue Gene/C, and Blue Gene/P for different applications
- # 1 and # 2 on top 500 list in December 2005
- http://domino.research.ibm.com/comm/research_projects.nsf/pages/bluegene.index.html

2.17a: Current Number 1 (Blue Gene/L)

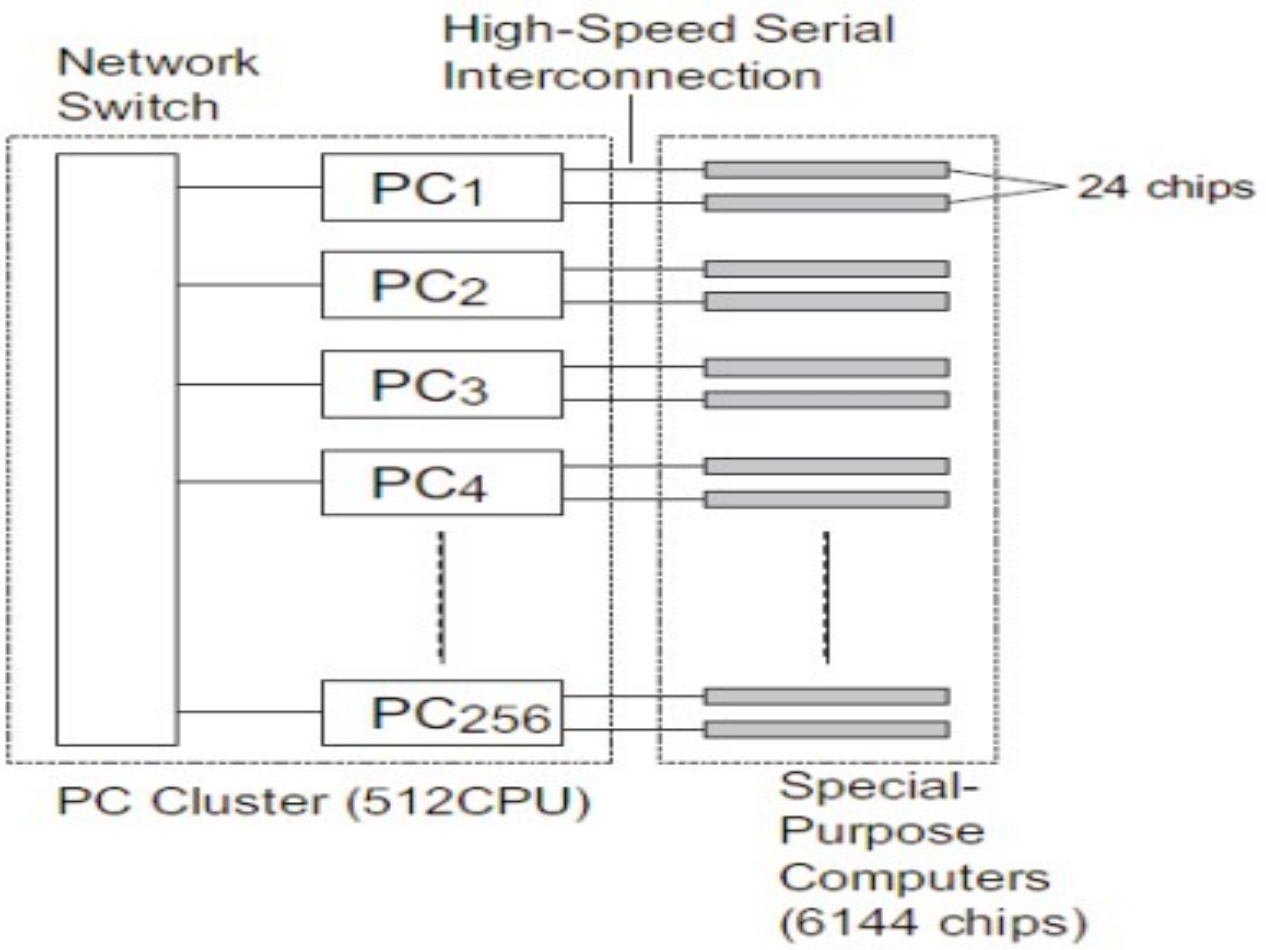
- eServer Blue Gene Solution (Blue Gene/L)
- 32,768 GB memory
- Installed in 2005 DOE/NNSA/LLNL
- 131,072 processors
- Peak performance 367.8 TFlops, 280.6 Tflops/s LINPACK benchmark
- See more at http://www.top500.org
- The Earth Simulator is currently # 10
- Next round announcement: November 2006, Tampa, Florida at Supercomputing 2006

2.17b: Blue Gene/L Picture

2.18a: The Really Fastest One

- On June 26, 2006, MD-Grape-3 at Riken in Japan, clocked at 1.0 petaflop. But it does not run LINPACK and not qualify for top 500 list
- Special purpose system for molecular dynamics – 3 times faster than Blue Gene/L
- 201 units of 24 custom MDGrape-3 chips (4808 total), plus 64 servers each with 256 Dual-Core Intel Xeon processor, and 37 servers each containing 74 Intel 3.2 GHz Xeon processors
- http://www.primidi.com/2004/09/01.html

2.18b: MDGrape-3 System



2.18c: One Board with 12 MDGrape-3 Chips