CMSC 838T – Lecture 7

- **Parallel computer architectures**
  - Characteristics of interconnection architecture
  - Methods for interprocessor communication, synchronization

Cray X1

Latest Cray Supercomputer

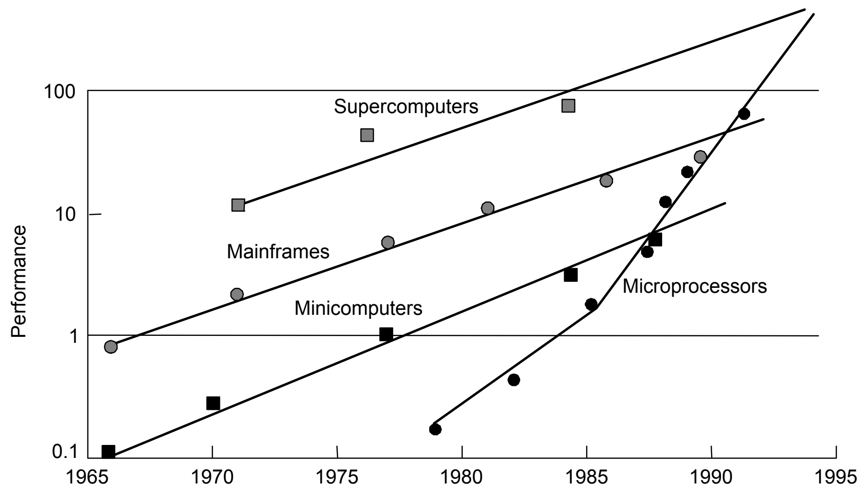
Up to 52.4 teraflops of peak computing power and 65.5 TB of memory

U.S. list pricing starts at about $2.5 million.

Why Parallel Architectures

- **Basic Motivation: Performance**
  - Only way to significantly increase speed
    - Orders of magnitude higher performance beyond uniprocessor
  - More cost effective than custom uniprocessor
    - Processor design / fabrication cost >> $100 million
    - Commercial microprocessor costs can be amortized
  - Single chip multiprocessors
    - Architects running out of uses for more transistors
    - Except for larger on-chip caches
    - Parallelism becomes effective use of chip density

Technology Trends



Commodity microprocessors caught up to and killed off most custom processors in 1990's

Parallel Architectures Today

- **Common parallel architectures**
  - **Desktop multiprocessor**
    - **Small number (2-4) of processors in single PC**
  - **Multiprocessor servers (SMP – Symmetric Multi-Processor)**
    - **Medium number (4-64) of processors in single system**
    - **HP SuperDome, Sun SunFire, IBM Regatta**
  - **Clusters of Workstations (COW)**
    - **Large collections (clusters) of processors / SMPs**
    - **HP AlphaServer, Beowulf Linux PCs**
- **Supercomputer architectures (less common)**
  - **Distributed memory multiprocessors (MPP)**
  - **Constellations – clusters of custom vector processors**

Parallel Architecture Overview

- **Motivation**
- **Forms of parallelism**
  - **Pipeline, SIMD, MIMD**
- **Types of parallel architectures**
  - **Shared memory multiprocessor (SMP)**
  - **Distributed memory multiprocessors (MPP)**
  - **Cluster**
  - **Constellation**
- **Summary**

Definition of Parallel Computer

- **A collection of processing elements that can communicate and cooperate to solve large problems fast**

32 processor Xeon Beowulf cluster

(COTS – Commercial Off-The-Shelf)

4096 processor Cray X1 constellation

(Customized Design)

A collection of processing elements …

- **How many**
  - A few, dozens, hundreds, thousands, many thousands or more
- **How powerful is each**
  - 1-bit processors, microprocessor, vector processors
- **How much memory**
  - Kilobytes to gigabytes

… that can communicate and coordinate …

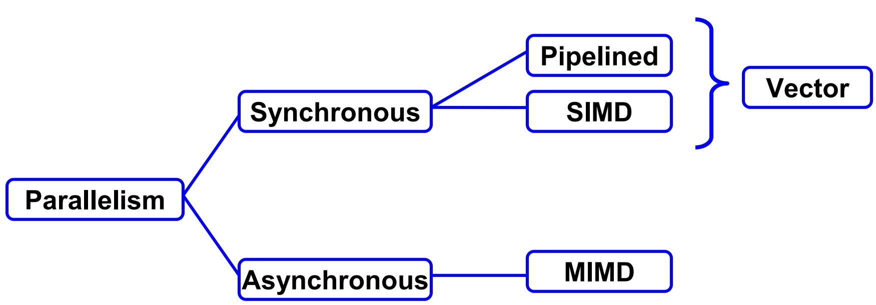
- **How do they exchange data**
  - Shared memory vs. message passing
- **What is the address space of programs**
  - Global vs. local address space
- **How are they interconnected**
  - Bus vs. switching network vs. loosely coupled network
- **How do they coordinate (synchronize) their execution**
  - Shared memory: locks, barriers, monitors
  - Message passing: blocking msgs. / collective communications
- **How frequently do they coordinate their execution**
  - Granularity: refers to the average size of subtasks separated by synchronization points, measured in instructions executed

… to solve large problems fast

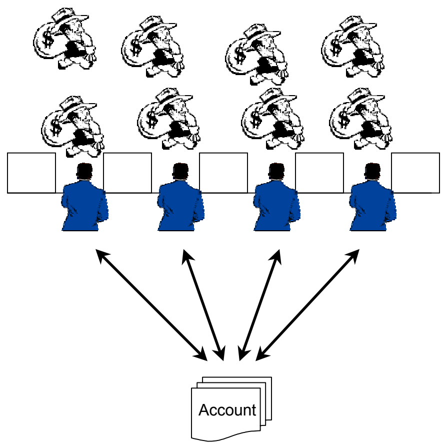
- **How large**
  - **Sufficiently beyond capability of single processor to make parallelism worthwhile**
- **How fast is fast**
  - **Fast enough for solution to be of interest**

Forms of Parallelism

- **Independence among computations**

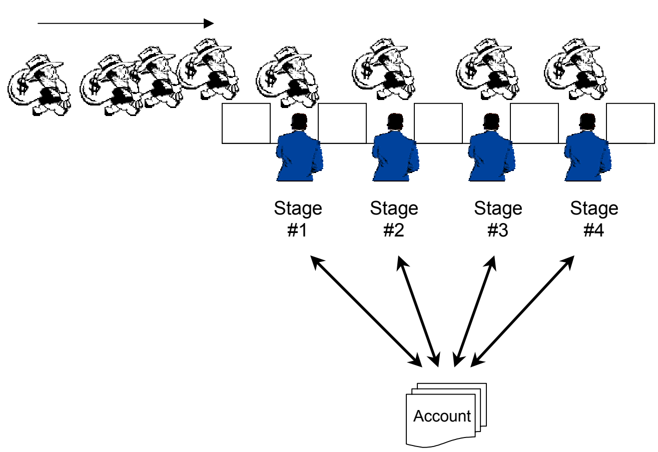


A Banking Analogy



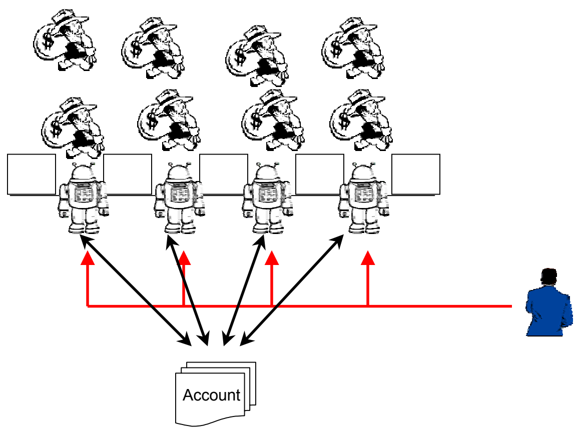
Pipeline Parallelism

- Tellers work as an assembly line of workers



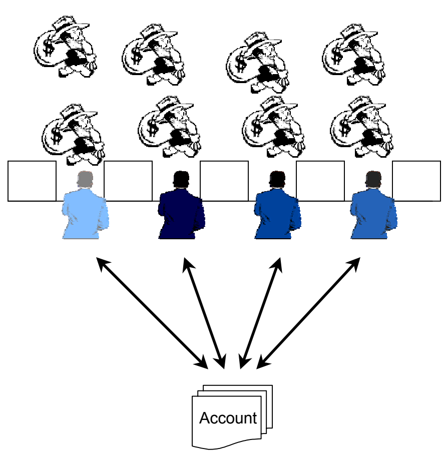
SIMD Parallelism

- **Single Instruction Multiple Data (SIMD)**
  - **All tellers do the same thing (or remain idle) at the same time**



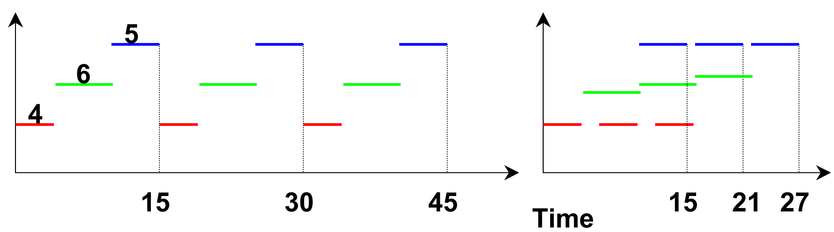
MIMD Parallelism

- **Multiple Instruction Multiple Data (MIMD)**
  - **Tellers independently serve different customers at same time**

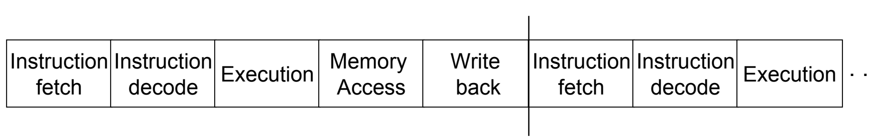


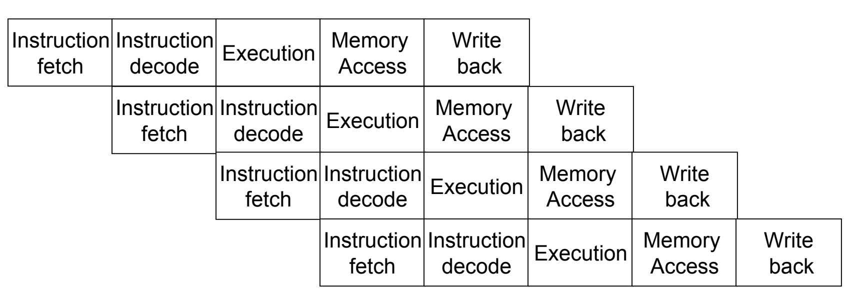
Pipeline Parallelism

- **Sources of pipeline parallelism**
  - **A number of tasks that may be partially overlapped**
  - **Tasks with multiple segments that may be overlapped**
- **Characteristics**
  - **Suitable for both fine-grain (instructions / vector) and coarse-grain (task) parallelism**
  - **Limited scalability**



Instruction-Level Pipeline Parallelism





SIMD Parallelism

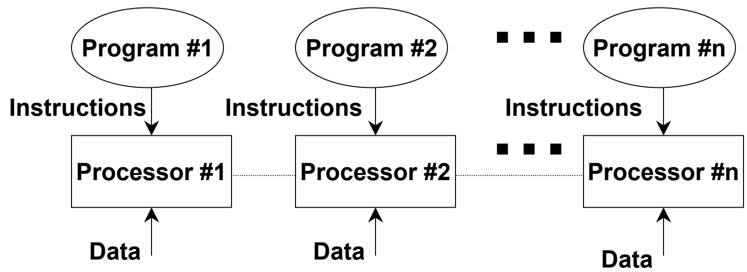
- **Sources of SIMD parallelism**
  - Identical operation executed on multiple data in parallel
- **Characteristics**
  - Single instruction stream for all processors
  - Limited flexibility
  - Suitable for vector, data-parallel computations

MIMD Parallelism

- **Sources of MIMD parallelism**
  - Any parallel computation
- **Characteristics**
  - Different instruction stream for each processor
  - Very flexible
  - Suitable for any (coarse-grain) parallel computation
    - Usually higher synchronization / communication costs
- **Types of MIMD parallelism**
  - MPMD – multiple program multiple data
  - SPMD – single program multiple data

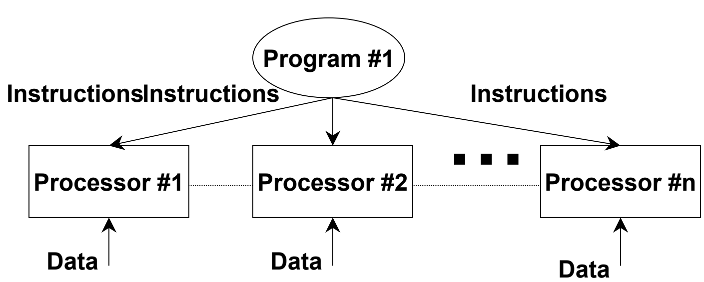
MPMD (Multiple Program Multiple Data)

- **Each processor has a different program to execute**
  - Greatest flexibility, complexity



SPMD (Single Program Multiple Data)

- **Each processor has same program to execute**
  - Can execute different instructions due to control flow
  - Moderate flexibility, lower complexity
  - Easier to program than MPMD in practice



Granularity of Parallelism

- **Granularity**
  - Refers to amount of work that can be executed independently
  - Coarse-grain parallelism → infrequent synchronization
  - Fine-grain parallelism → frequent synchronization
- **Categories**
  - Instruction level
    - A = 1, B = 2
  - Vector
    - A[1:100] = B[1:100] + C[1:100]
  - Loop level
    - Forall I = 1,100 { … }
  - Task level
    - Parallel case A: { … } B: { … }

Pipeline

SIMD

MIMD

Parallel Architecture Overview

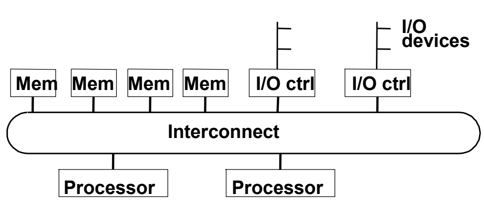
- **Motivation**
- **Forms of parallelism**
  - Pipeline, SIMD, MIMD
- **Types of parallel architectures**
  - Shared memory multiprocessor (SMP)
  - Distributed memory multiprocessors (MPP)
  - Cluster
  - Constellation
- **Summary**

Shared-Memory Multiprocessor (SMP)

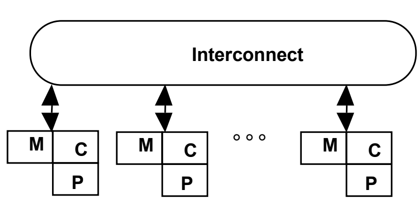
- **Multiple processors sharing single memory**
  - HP SuperDome, Sun SunFire, IBM Regatta
- **Types**
  1. UMA (Uniform Memory Access)
     - Physically shared memory, shared bus
     - Also known as SMP (Symmetric Multi Processor)
  2. NUMA (Non-Uniform Memory Access)
     - Physically distributed memory, shared interconnect
     - Coherent cache (SGI Origin 2000)
       - Snoops the system bus to maintain cache coherence
     - Non-coherent cache (Cray T3E)
  3. COMA (Cache Only Memory Architecture)
     - NUMA that treats all memory as cache

Models of Shared-Memory Multiprocessors

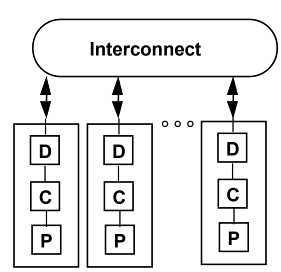
Uniform Memory Access (UMA) Model



Interconnect: Bus, Crossbar, Multistage network
P: Processor
M: Memory
C : Cache
D: Cache directory

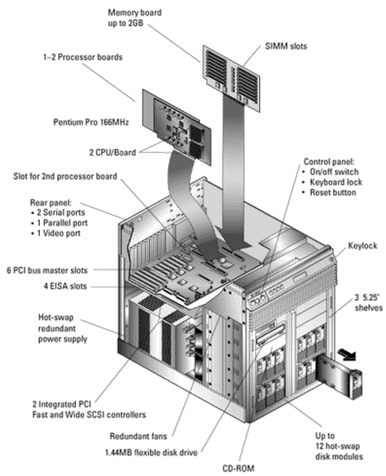


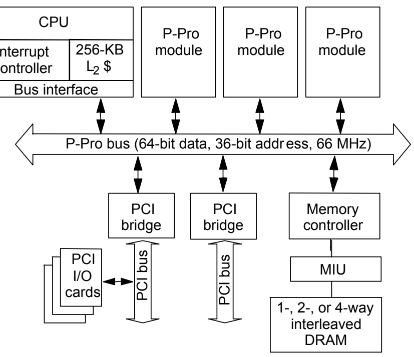
Distributed memory (NUMA) Model



Cache-Only Memory Architecture (COMA)

SMP – Intel Pentium Pro Quad





- 4 processors
- All coherence and multiprocessing glue in processor module

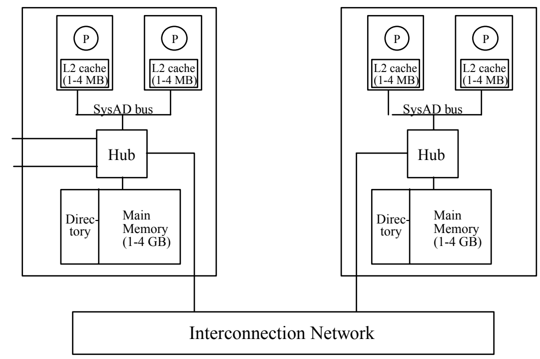
Cache Coherence

Time	Event	Cache for A	Cache for B	Memory for x
0				1
1	CPU A reads X	1		1
2	CPU B reads X	1	1	1
3	CPU A writes 0 to X	0	1	0

invalid value

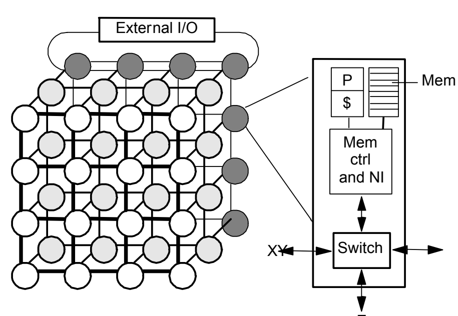
- Caches are coherent if all processors have same value
- Bus snooping
  - All processors listen on shared bus for updates to local cache
- Directory-based
  - Processors maintain directory of all processors with data
  - Send messages to invalidate or update remote cache entries

Cache Coherent NUMA – SGI Origin 2000



- Two processors on each node, up to 512 processors
- Hub implements the directory-based cache coherence protocol
- Scalable interconnection network

Non-Cache Coherent NUMA – Cray T3E



- Scale up to 1024 processors, 480MB/s links
- Memory controller generates communication requests for non-local references
- No hardware mechanism for coherence

Shared Memory Multiprocessor (SMP)

- **Advantages**
  - Processor can directly reference any memory location
    - Communication occurs implicitly as loads and stores
  - Convenient:
    - Simple programming model
- **Disadvantages**
  - May introduce (data race) errors dependent on execution order
  - SMP is not scalable – contention for shared interconnect

Distributed-Memory Multiprocessor (MPP)

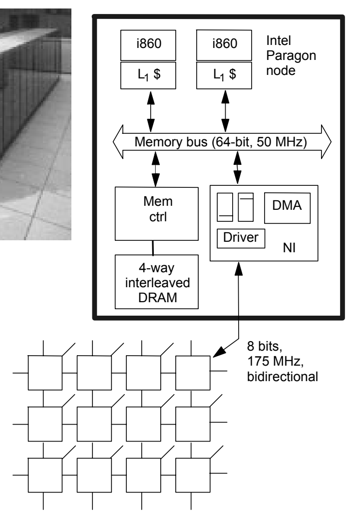
- **Large collection of tightly connected processor nodes**
  - Intel Paragon, IBM SP-2
- **Also known as “message-passing” computers**
- **Characteristics**
  - Distributed memory and separate address spaces
  - Non-local memory accesses expensive
  - Large memory & high scalability

Distributed-Memory Multiprocessor (MPP)

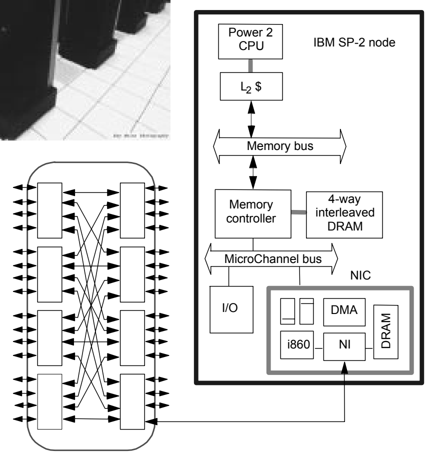
- **Architecture**
  - **Comprised of multiple autonomous computers (nodes)**
  - **Each node consists of a processor, local memory, attached storage and I/O peripherals (approx single PC or workstation)**
  - **Local memory is only accessible by local processors**
  - **Inter-node communication is carried out by sending messages through the interconnection network**
- **Programming**
  - **Explicit messages for interprocessor communication**

MPP – Intel Paragon

Sandia' s Intel Paragon XP/S-based Supercomputer



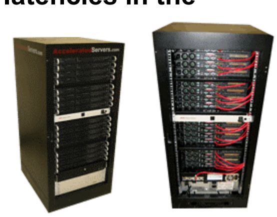
MPP – IBM SP-2



- Made out of essentially complete IBM RS6000 workstations
- Network interface integrated in I/O bus (bandwidth limited by I/O bus)

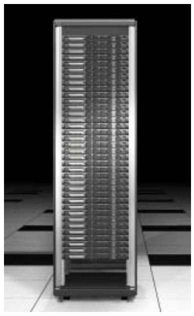
Cluster Architecture

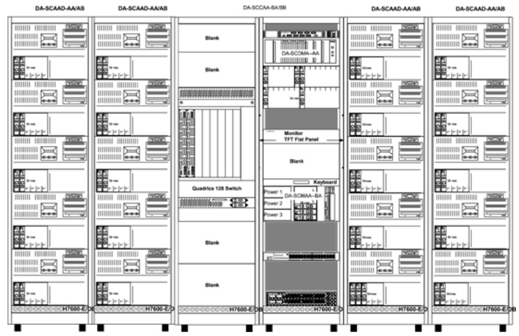
- Large collections of loosely connected PC / SMP nodes
  - Beowulf cluster (Linux PC), HP AlphaServer SC (Alpha)
  - High-performance interconnects
    - Couple hundred of megabytes/s of bandwidth (300-400) and latencies in the order of a few microseconds (2-5)
    - Quadrics QsNet, Myrinet
  - Low-performance interconnects
    - Tens of megabytes/s of bandwidth and latencies in the order of tens of microseconds (20-50)
    - Gigabit Ethernet
- Characteristics
  - Large memory & very high scalability
  - Hybrid memory model if SMP nodes
  - Communicate using message passing



32 processor Xeon Beowulf cluster

Cluster Architecture – HP AlphaServer SC





- 4-processor 1.2 GHz EV68 Alpha nodes
- 2-32 GB memory / node
- 280 MB/sec AlphaServer SC PCI adaptor
- 32 GB/sec cross-section bandwidth
  - The amount of information that all the nodes can pass to one another and back again, all at one time, and all in one second

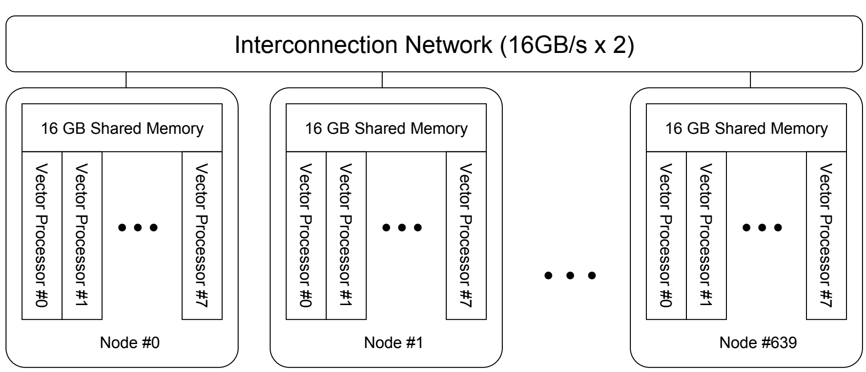
Constellation Architecture

- Small / medium collections of fast vector nodes
  - NEC Earth Simulator, Cray X1
    - 4-8 vector processor nodes
  - IBM BlueGene
    - 64-processor SMP nodes
- Characteristics
  - Large memory & moderate scalability
  - High performance nodes
  - Hybrid memory model

Constellation – Vector Nodes

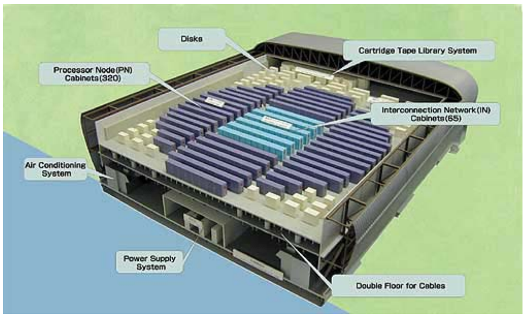
- **Custom vector processors**
  - **NEC SX-6, Cray SV-1**
- **Characteristics**
  - **Vector operations on vector registers**
  - **High-bandwidth pipelined memory access**
    - **Memory bandwidth matches processing rate**
    - **No cache coherence needed for vector operations**
    - **Achieves high percentage of peak performance**
  - **Global shared memory (PVP)**
    - **Very limited scalability in processor count**
  - **Easy, but different, programming model**
  - **Expensive**

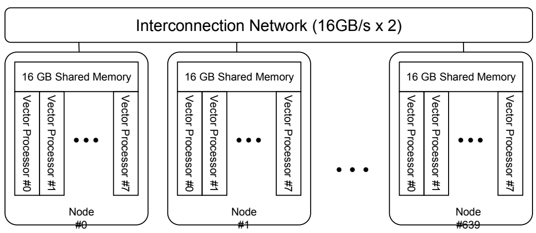
Constellation – NEC Earth Simulator



Number of Processors/Node	**8**	**Peak Performance/Processor**	**8 Gflops**
Total Number of Processors	**5120**	**Peak Performance/Node**	**64 Gflops**
Total Number of Nodes	**640**	**Total Peak Performance**	**40Tflops**
Shared Memory/Node	**16 GB**	**Total Main Memory**	**10TB**

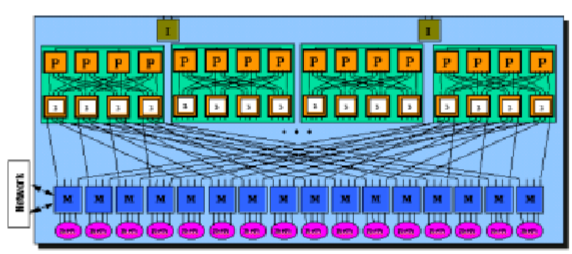
Constellation – NEC Earth Simulator





- Three levels of parallelism
  - Vector processors
  - Shared memory
  - Message passing
- Currently world's most powerful computer
  - As fast as next 5 fastest computers combined

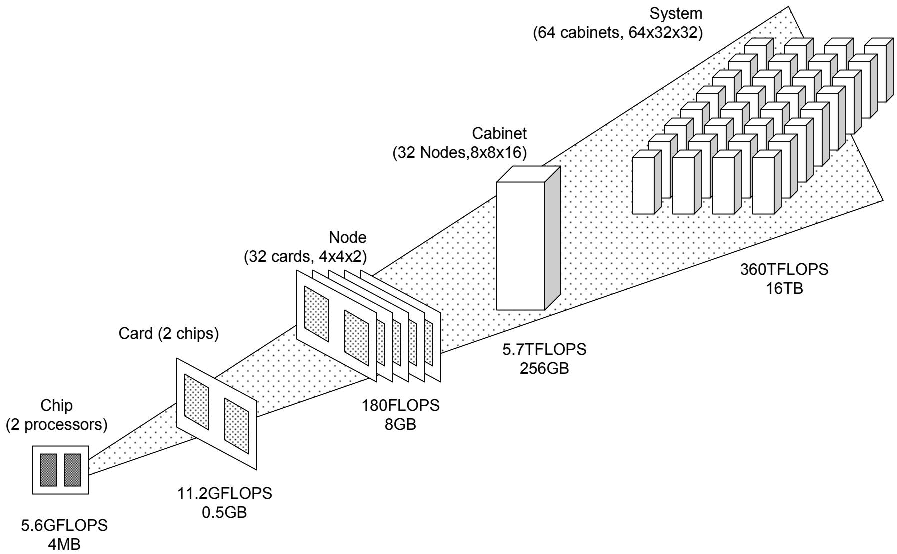
Constellation – Cray X-1



Cray X1 node

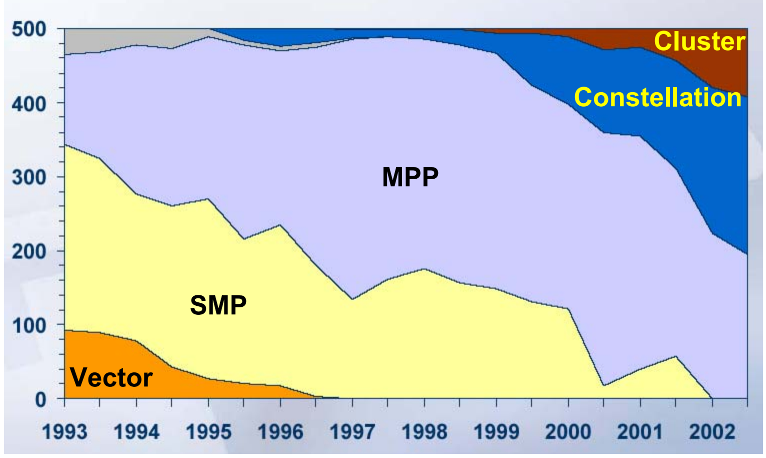
- Custom vector processors, 1.86 Gflop / sec peak
- 4 processors / node, 16 nodes / chassis, 4096 processors max
- 32 Gb – 64 Tb, cache coherent, physically distributed, globally addressable memory (7x – 40x bandwidth of PC clusters)
- Modified 2D torus interconnect, 400 Gb/sec peak bandwidth

Constellation – IBM BlueGene



Parallel Architectures – Trends

- **Top 500 List**
  - **Based on performance on Linpack benchmark**
  - **Graph shows number of architecture of each type in Top 500**



Parallel Architecture Trends

- **Faster interconnects**
  - Fabrics improving, but adapters are bottleneck
  - Not keeping up with node performance in clusters
- **Larger memories**
  - SMPs with 64 - 1024 GB memories
  - Deeper cache hierarchies
- **Processor / memory integration**
  - Processor-in-memory (PIM), and memory controllers
  - Relative cost of off-chip communication increases

Parallel Architecture Summary

- **Parallel architectures provide large performance boost**
- **Different forms of parallelism may be exploited**
  - Pipeline, SIMD, MIMD
- **Many different parallel architectures**
  - Shared memory, distributed memory, cluster, constellation
  - Typically built from commodity parts
- **Challenges**
  - Extracting parallel performance
    - Sustained performance usually small fraction of peak
  - Programming model
    - Parallel programming much more labor intensive