

CSE 260 – Class #2

Larry Carter

carter@cs.ucsd.edu

www.cs.ucsd.edu/classes/fa01/cs260

Class time won't change

Office Hours: AP&M 4101

MW 10:00-11:00 or by appointment

Note slight change



First quizlet next Tuesday

- Vocabulary, concepts, and trends of parallel machines and languages
- 15 minutes – multiple choice and short answers

Reading Assignment

“High Performance Computing: Crays, Clusters, and Centers. What Next?”

Gordon Bell & Jim Gray

Microsoft Research Center Tech Report MSR-TR-2001-76

research.microsoft.com/scripts/pubs/view.asp?TR_ID=MSR-TR-2001-76

Recommended Talk

Google – Weds 9/26 (tomorrow) 11:00,

4301 APM

Some topics from of Class 1

- Why do parallel computation?
- Flynn's taxonomy (MIMD, SIMD, ...)
- Not all parallel computers are successful
- Vector machines
- Clusters and the Grid
 - Need to add term Beowulf cluster
 - Do-it-yourself cluster of (typically) Linux PC's or workstation (though Andrew Chien uses Windows NT).
 - Very popular recently

Scalability

An architecture is scalable if it continues to yield the same performance per processor (albeit on a larger problem size) as the number of processors increases

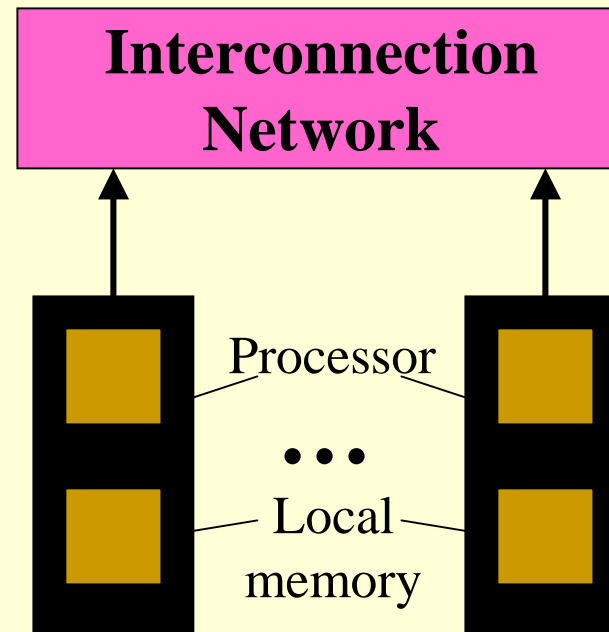
- Scalable MPPs designed so that larger versions of the same machine (i.e. versions with more nodes/CPU's) can be built or extended using the same design

A memory-centric taxonomy

- Multicomputers: Interconnected computers with separate address spaces. Also known as message-passing or distributed address-space computers.
- Multiprocessors*: Multiple processors having access to the same memory (shared address space or single address-space computers.)
- * Warning: Some use term "multiprocessor" to include multicomputers.

Multicomputer topology

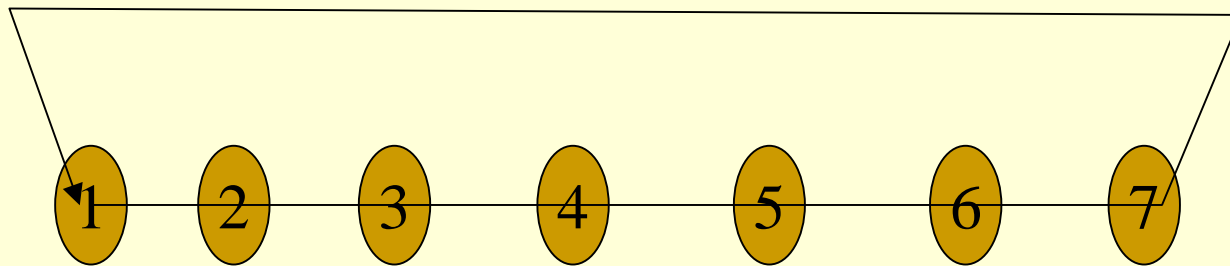
- Interconnection network should provide connectivity, low latency, high bandwidth
- Many interconnection networks developed over last 2 decades
 - Hypercube
 - Mesh, torus
 - Ring, etc.



Basic Message Passing **Multicomputer**

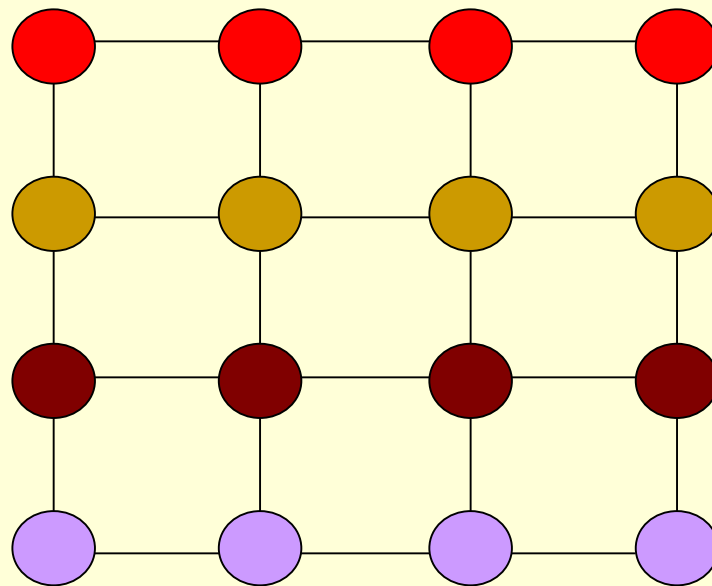
Lines and Rings

- Simplest interconnection network
- Routing becomes an issue
 - No direct connection between nodes



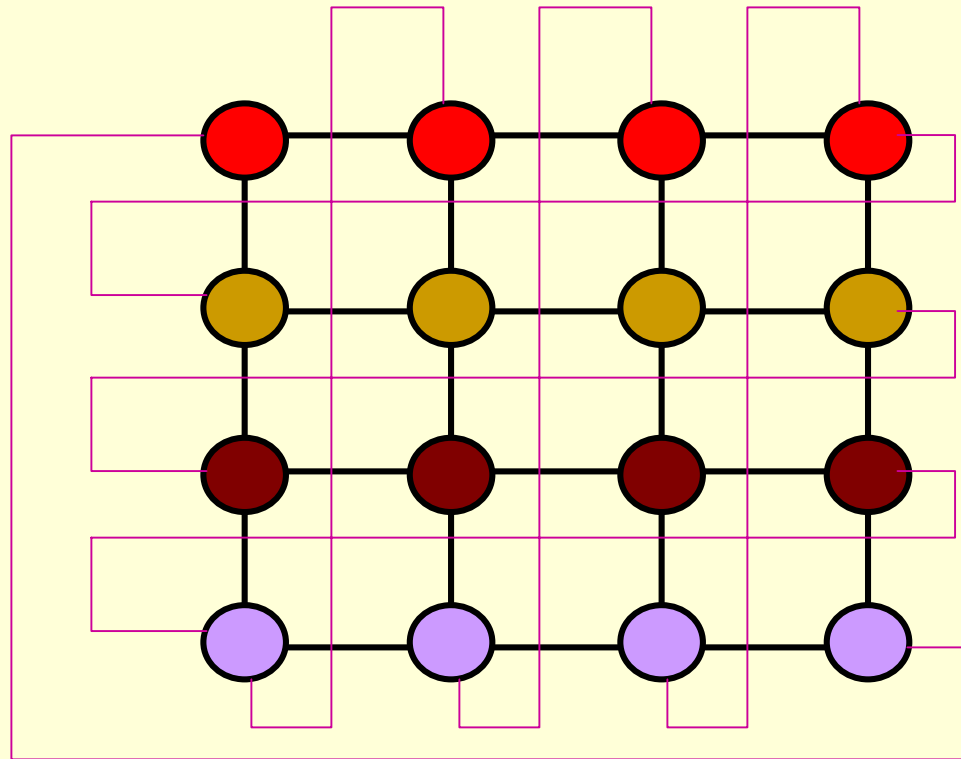
Mesh and Torus

- Generalization of line/ring to multiple dimensions
- 2D Mesh used on Intel Paragon; 3D Torus used on Cray T3D and T3E.



Mesh and Torus

- Torus uses wraparound links to increase connectivity

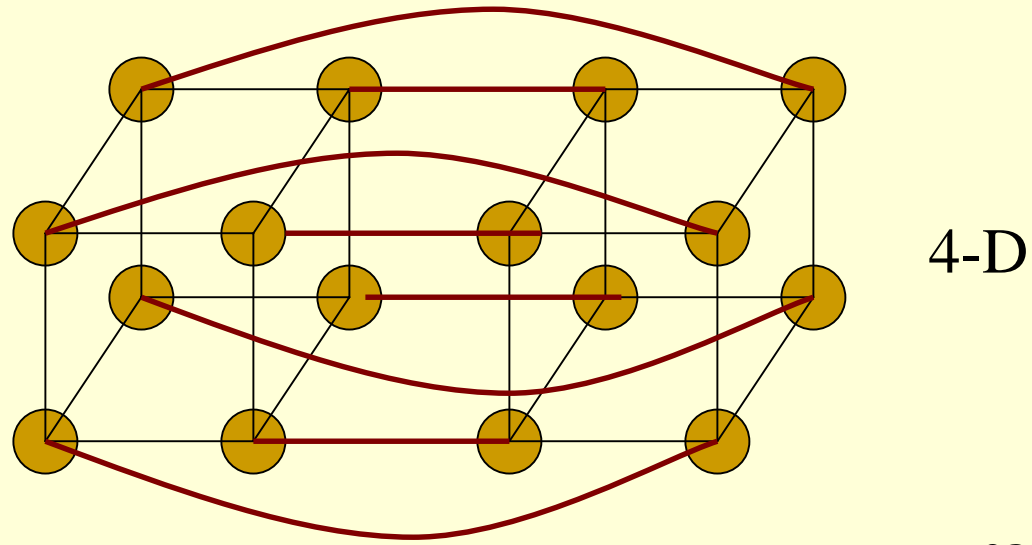


Hop Count

- Networks can be measured by diameter
 - This is the minimum number of hops that message must traverse for the two nodes that furthest apart
 - Line: Diameter = $N-1$
 - 2D (N x M) Mesh: Diameter = $(N-1) + (M-1)$
 - 2D (N x M) Torus: Diameter = $\lfloor N/2 \rfloor + \lfloor M/2 \rfloor$

Hypercube Networks

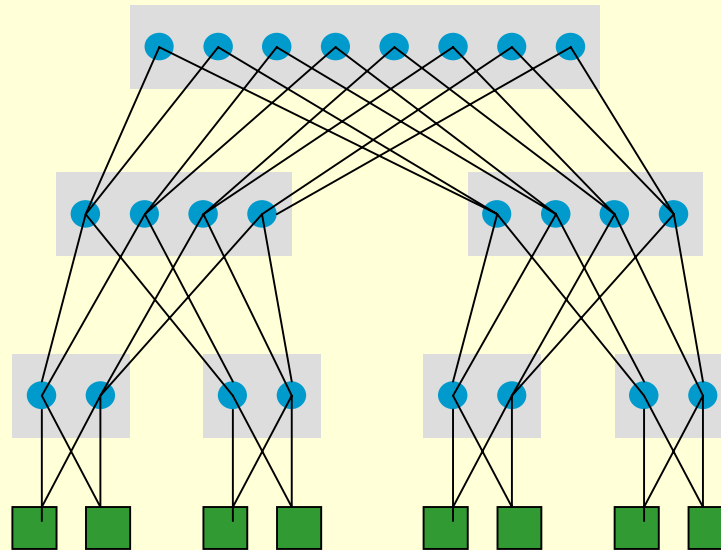
- Dimension N Hypercube is constructed by connecting the “corners” of two N-1 hypercubes
- Interconnect for Cosmic Cube (Caltech, 1985) and its offshoots (Intel iPSC, nCUBE), Thinking Machine’s CM2, and others.



Fat-tree Interconnect

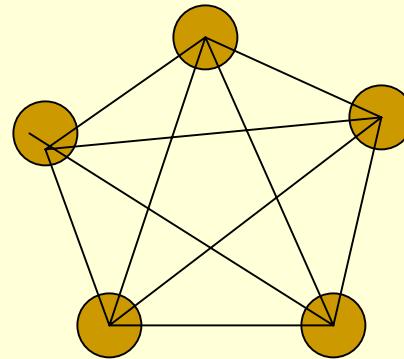
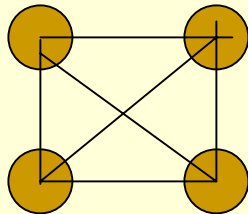
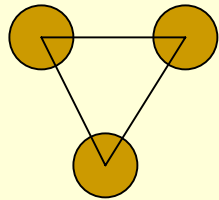
- Bandwidth is increased towards the root (but aggregate bandwidth decreases)
- Data network for TMC's CM-5 (a MI MD MPP)
 - 4 leaf nodes, internal nodes have 2 or 4 children
- To route from leaf A to leaf B, pick random switch C in the least common ancestor fat node of A and B, take unique tree route from A to C and from C to B

Binary fat-tree in which all internal nodes have two children



An Impractical Interconnection Topology

- Completely connected
 - Every node has a direct wire connection to every other node



$N \times (N-1)/2$ Wires

The MPP phenomenon

- In mid-90's, all major microprocessor were the engine for some MPP (Massively Parallel Processing systems, vaguely meaning >100 procs).
 - These replaced early-90's machines like CM-5 and KSR1 that had lots of proprietary hardware
- Examples:
 - IBM RS6000 & PowerPC -> SP1, SP2, SP
 - Dec Alpha -> Cray T3D and T3E
 - MIPS -> SGI Origin
 - Intel Pentium Pro -> Sandia ASCI Red
 - HP PA-RISC -> HP/Convex Exemplar
 - Sun SPARC -> CM-5

The MPP phenomenon

- Many of these have died or are dying out
 - IBM and SUN still doing well
- Being replaced by PC-based Beowulf clusters
- Next wave: clusters of playstations ???

Message Passing Strategies

- Store-and-Forward
 - Intermediate node receives entire message before sending it on to next link
- Cut through routing
 - Message divided into small "packets",
 - Intermediate nodes send on packets as they come in
 - Concern: what happens if destination isn't ready to receive a packet?
 - One possible answer: "Hot potato routing" - if destination isn't free, send it somewhere else! Used in Tera MTA.

Latency and Bandwidth

Bandwidth: number of bits per second that can be transmitted through the network

Latency: total time to send one ("zero-length") message through the network

- Fast Ethernet: BW = 10MB/sec (or 100 MB/sec for gigabit Ethernet), latency = 100usec.
- Myrinet: BW = 100's MB/sec, latency = 20 usec.
- SCI (Scalable Coherent Interface) - BW = ~400 MB/sec latency = 10 usec. (DSM interface)

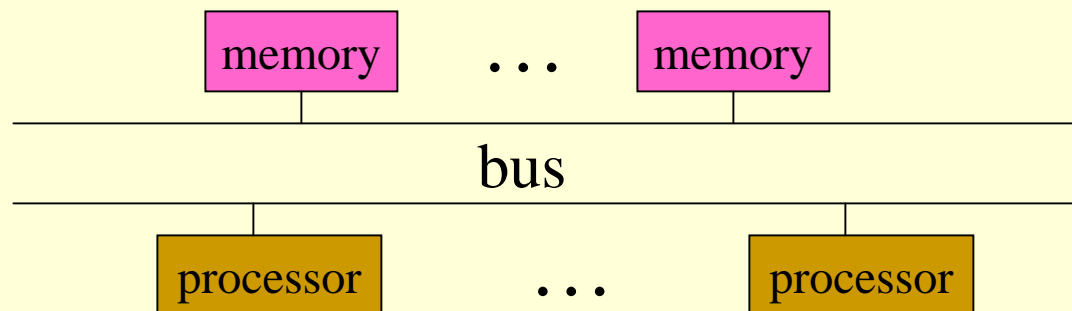
Latency is mostly time of software protocols

Shared Address Space Multiprocessors

- 4 basic types of interconnection media:
 - Bus
 - Crossbar switch
 - Multistage network
 - Interconnection network with distributed shared memory (DSM)

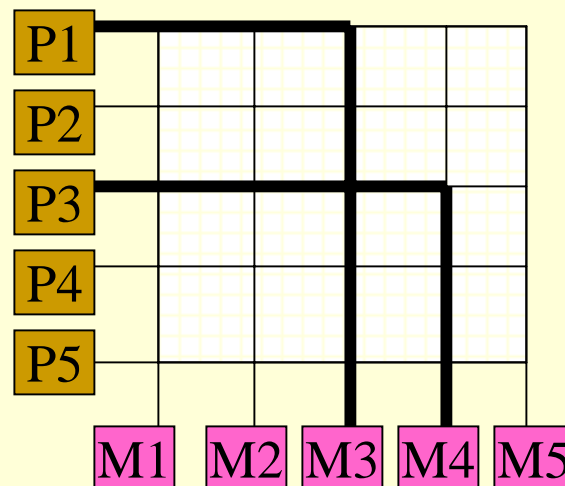
Bus architectures

- Bus acts as a “party line” between processors and shared memories
- Bus provides uniform access to shared memory (UMA = Uniform Memory Access) (SMP = symmetric multiprocessor)
- When bus saturates, performance of system degrades
- Bus-based systems do not scale to more than 32 processors [Sequent Symmetry, Balance]



Crossbar Switch

- Uses $O(mn)$ switches to connect m processors and n memories with distinct paths between each processor/memory pair
- UMA
- Scalable performance but not cost.
- Used in Sun Enterprise 10000 (like our "ultra")

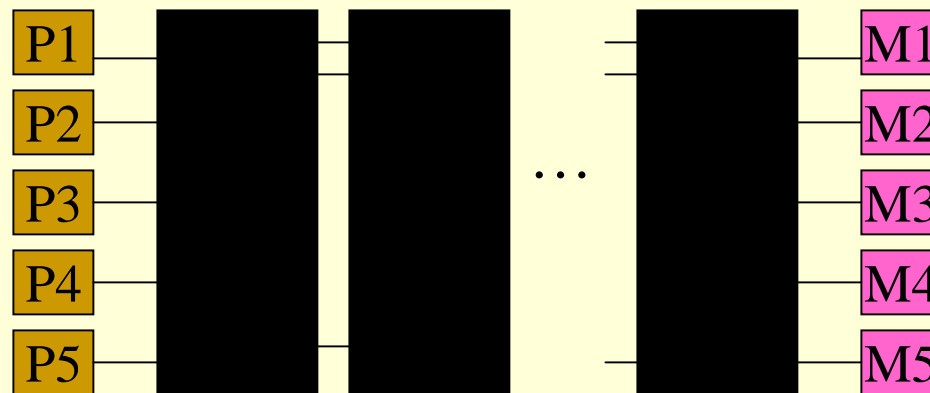


SUN Enterprise 10000

- We'll use 64-node E10000 ("ultra") at SDSC.
 - 400 MHz UltraSparc 2 CPU's, 2 floats/cycle.
 - UMA
 - 16 KB data cache (32 byte linesize), 4MB level 2 cache, 64 GB memory per processor
 - Front end processors ("gaos") are 336 MHz
 - Network: 10 GB/sec (aggregate), 600 ns latency

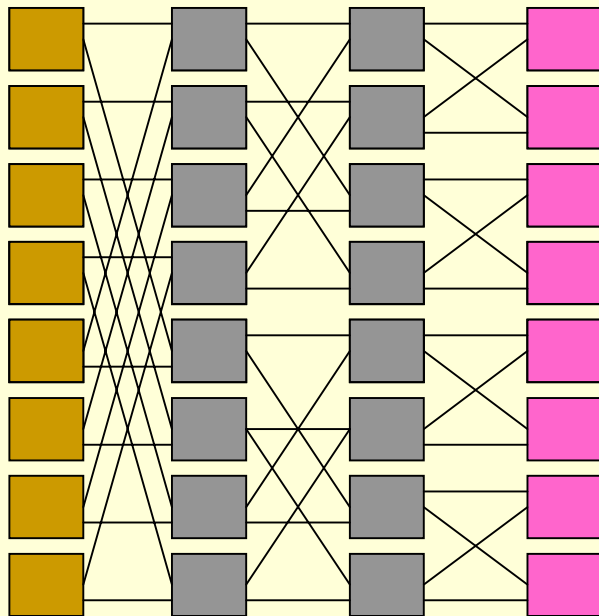
Multistage Networks

- Multistage networks provide more scalable performance than bus but at less cost than crossbar
- Typically $\max\{\log n, \log m\}$ stages connect n processors and m shared memories
- Memory still considered “centralized” (as opposed to “distributed”). Also called “dancehall” architecture.

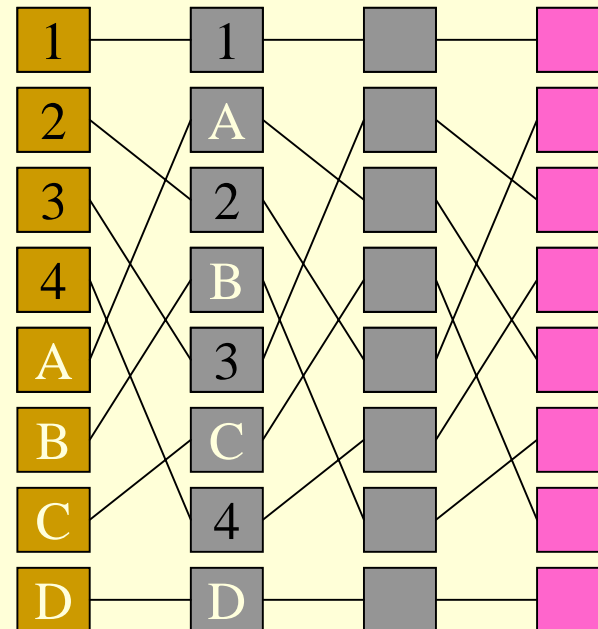


Some Multistage Networks

- Butterfly multistage



- Shuffle multistage



Distributed Shared Memory (DSM)

- Rather than having all processors on one side of network and all memory on the other, DSM has some memory at each processor (or group of processors).
- NUMA (Non-uniform memory access)
- Example: HP/Convex Exemplar (late 90's)
 - 3 cycles to access data in cache
 - 92 cycles for local memory (shared by 16 procs)
 - 450 cycles for non-local memory

Cache Coherency

If processors in a multiprocessor have caches, they must be kept coherent (according to some chosen consistency model, e.g. sequential consistency.)

The problem: If P1 and P2 both have a copy of a variable X in cache, and P1 modifies X, future accesses by P2 should get the new value.

Typically done on bus-based systems with hardware “snooping” – all processors watch bus activity to decide whether their data is still valid.

Multistage networks and DSM use directory-based methods.

MESI coherency protocol

Used by IBM Power PC, Pentiums, ...

Four states for data in P1's cache:

Modified: P1 has changed data; not reflected in memory
(Data is "dirty". Other processors must invalidate copies.)

Exclusive: P1 is only cache with data, same as in memory

Shared: P1 and other proc's have (identical) copies of data

Invalid: Data is unusable since other proc has changed

P1 initiates bus traffic when:

P1 changes data (from S to M state), so P2 knows to make I

P2 accesses data that has been modified by P1 (P1 must write block back before P2 can load it).

Multiprocessor memory characteristics

UMA (uniform memory access) computer

Also known as SMP (symmetric multiprocessor)

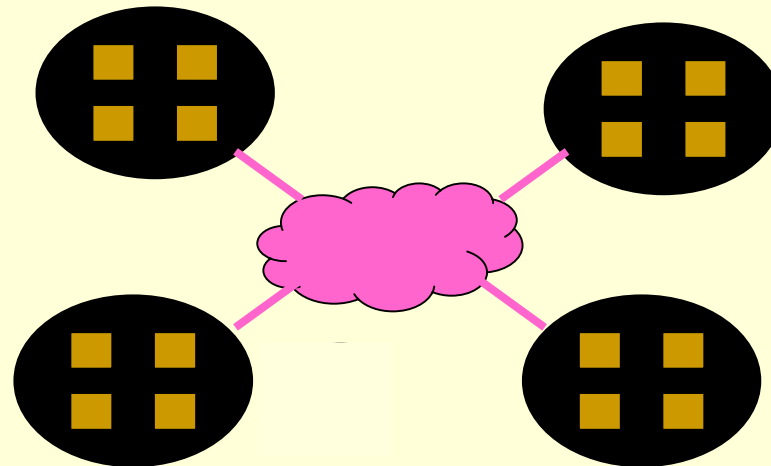
- Sequent, Sun Enterprise 10000 (E10000)

NUMA = non-uniform memory access

- Cray T3E (uses remote loads & stores rather than cache coherency)
- COMA = cache-only memory access
 - Kendall Square Research's KSR1
- CC-NUMA = cache coherent NUMA
 - Stanford DASH, SGI Origin

Multi-tiered computers

- Cluster of SMP's.
 - or Multiprocessor Multicomputer
 - Each "node" has multiple processors with cache-coherent shared memory
 - Nodes connected by high-speed network.
- Used in the biggest MPP's
 - IBM SP (e.g. Blue Horizon), Intel ASCI Red, ...



Message Passing vs. Shared Memory

- Message Passing:
 - Requires software involvement to move data
 - More cumbersome to program
 - More scalable
- Shared Memory:
 - Subtle programming and performance bugs
- Multi-tiered
 - Best(?) Worst(?) of both worlds

Other terms for classes of computers

- Special Purpose
 - Signal processors, Deep Blue, Sony gameboys & playstations
- Bit Serial
 - CM-2, DAP
- COTS (Commercial Off-The Shelf)
- Heterogeneous – different model procs
 - Grid, many clusters, partially upgraded MPP's,...

Some notable architects

- Seymore Cray
 - CDC 6600, Cray Research vector machines, then moved to Cray, killed in auto accident.
- John Cocke
 - Many IBM supercomputers, prime inventor of RISC (though Patterson coined term), resisted MPP's for years.
- Burton Smith
 - HEP, Tera MTA, recently acquired Cray Research, changed name to Cray Inc.

Cray Computers

- Cray is almost synonymous with supercomputer
- Superscalar-like machines (before term invented):
 - CDC 6600, 7600
- Multiprocessor vector machines without caches:
 - Cray 1, Cray X-MP, Y-MP, C90, T90, (J90)
- MPP's (Massively Parallel Processors)
 - T3D, T3E ("T3" = 3-D Torus)
- Recent offerings:
 - SV1 (vector multiprocessor + cache), Tera MTA, assorted other servers via mergers and spinoffs.

Today's fastest supercomputers include:

(<http://www.netlib.org/benchmark/top500/top500.list.html>)
(doesn't include secret machines, nor commercial ones like google)

Sandia's ASCI Red [9216 Intel Pentiums Pro's] – first to achieve 1 TFLOP/S speed (1997) (bigger now).

Livermore's ASCI White ('2000) [8192 IBM SP Power3's] – today's fastest computer.

Los Alamos' ASCI Blue-Mountain ('98) – [48 128-proc cc-NUMA SGI Origins, connected via HIPPI.]

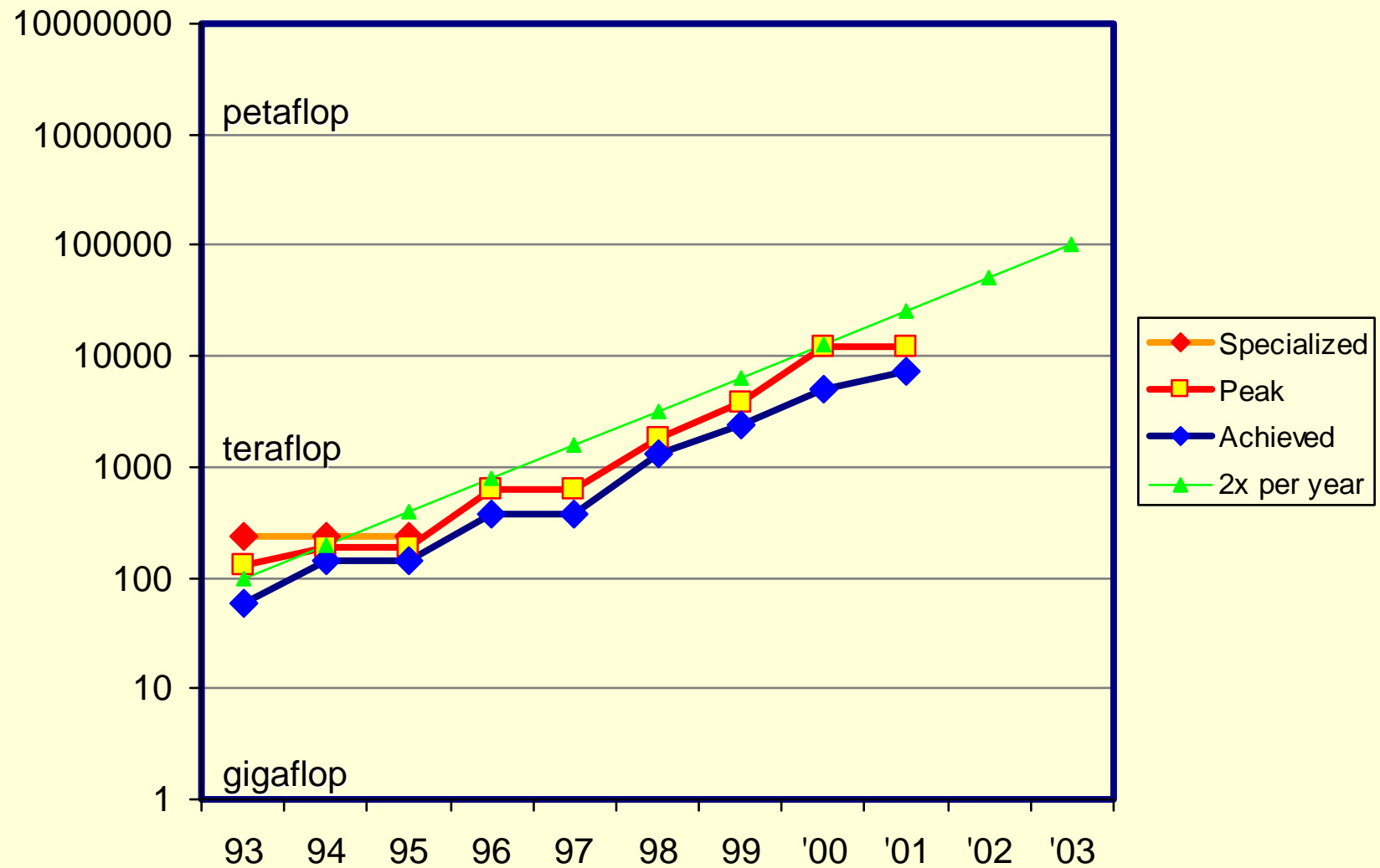
ASCI (Advanced Strategic Computer Initiative) is a big DOE (Department of Energy) project for replacing nuclear tests by simulation

... and, after 5 more IBM, 2 Hitachi, 1 NEC, 1 T3E, ...

SDSC's Blue Horizon: 1152-proc IBM SP
World's 13th fastest computer (June 2001 listing)
Fastest computer available to US academics.



Biggest supercomputers



Selected Computers

- SISD

- Scalar: CDC 6600 (and 7600)
- Vector: Cray X-MP (and Y-MP), T90

- SIMD

- Illiack V

“MP”=multiprocessor

• Special purpose machines:
IBM Deep Blue
Sony Playstations

- MIMD

- Distributed Address Space
 - Vendor-assembled (IBM SP, Cray T3E, TMC CM-5)
 - Clusters (e.g. Beowulf)
- Shared Address Space
 - UMA (Sun E10000, Cray/Tera MTA)
 - NUMA (SGI Origin)
- Clusters of SMP's (e.g. ASCI red/white/blue)

e.g. Blue Horizon at SDSC

Possible Mini-studies

- Extend previous chart to earlier years, using some objective measure of performance.
- Make list of 10 most notable computers by some criterion (fastest, best cost/performance, most profitable, ...)