MIMD Overview

- MIMDs in the 1980s and 1990s
  - Distributed-memory multicomputers
    - Intel Paragon XP/S
    - Thinking Machines CM-5
    - IBM SP2
  - Distributed-memory multicomputers with hardware to look like shared-memory
    - nCUBE 3
    - Kendall Square Research KSR1
  - NUMA shared-memory multiprocessors
    - Cray T3D
    - Convex Exemplar SPP-1000
    - Silicon Graphics POWER & Origin
- General characteristics
  - 100s of powerful commercial RISC PEs
  - Wide variation in PE interconnect network
  - Broadcast / reduction / synch network

Intel Paragon XP/S Overview

- Distributed-memory MIMD multicomputer
- 2D array of nodes
  - Main memory physically distributed among nodes (16-64 MB / node)
  - Each node contains two Intel i860 XP processors: application processor to run user program, and message processor for inter-node communication

XP/S Nodes and Interconnection

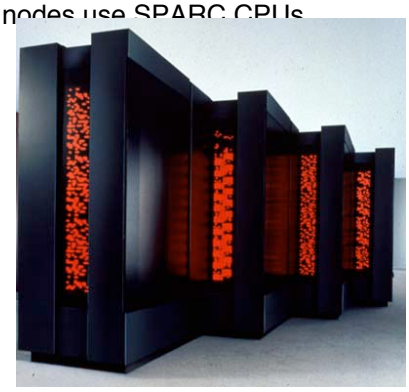
- Node composition
  - 16–64 MB of memory
  - Application processor
    - Intel i860 XP processor (42 MIPS, 50 MHz clock) to execute user programs
  - Message processor
    - Intel i860 XP processor
    - Handles details of sending / receiving a message between nodes, including protocols, packetization, etc.
    - Supports broadcast, synchronization, and reduction (sum, min, and, or, etc.)
- 2D mesh interconnection between nodes
  - Paragon Mesh Routing Chip (PMRC) / iMRC routes traffic in the mesh
    - 0.75 μm, triple-metal CMOS
    - Routes traffic in four directions and to and from attached node at > 200 MB/s

XP/S Usage

- System OS is based on UNIX, provides distributed system services and full UNIX to every node
  - System is divided into partitions, some for I/O, some for system services, rest for user applications
- Users have client/server access, can submit jobs over a network, or login directly to any node
- System has a MIMD architecture, but supports various programming models: SPMD, SIMD, MIMD, shared memory, vector shared memory
- Applications can run on arbitrary number of nodes without change
  - Run on more nodes for large data sets or to get higher performance

Thinking Machines CM-5 Overview

- Distributed-memory MIMD multicomputer
  - SIMD or MIMD operation
- Configurable with up to 16,384 processing nodes and 512 GB of memory
  - Divided into partitions, each managed by a control processor
  - Processing nodes use SPARC CPUs



CM-5 Partitions / Control Processors

- Processing nodes may be divided into (communicating) partitions, and are supervised by a control processor
  - Control processor broadcasts blocks of instructions to the processing nodes
    - SIMD operation: control processor broadcasts instructions and nodes are closely synchronized
    - MIMD operation: nodes fetch instructions independently and synchronize only as required by the algorithm
- Control processors in general
  - Schedule user tasks, allocate resources, service I/O requests, accounting, etc.
  - In a small system, one control processor may play a number of roles
  - In a large system, control processors are often dedicated to particular tasks (partition manager, I/O cont. proc., etc.)

CM-5 Nodes and Interconnection

- Processing nodes
  - SPARC CPU (running at 22 MIPS)
  - 8-32 MB of memory
  - (Optional) 4 vector processing units
- Each control processor and processing node connects to two networks
  - Control Network — for operations that involve all nodes at once
    - Broadcast, reduction (including parallel prefix), barrier synchronization
    - Optimized for fast response & low latency
  - Data Network — for bulk data transfers between specific source and destination
    - 4-ary hypertree
    - Provides point-to-point communication for tens of thousands of items simultaneously
    - Special cases for nearest neighbor
    - Optimized for high bandwidth

Tree Networks (Reference Material)

- Binary Tree
  - 2^k–1 nodes arranged into complete binary tree of depth k–1
  - Diameter is 2(k–1)
  - Bisection width is 1
- Hypertree
  - Low diameter of a binary tree plus improved bisection width
  - Hypertree of degree k and depth d
    - From “front”, looks like k-ary tree of height d
    - From “side”, looks like upside-down binary tree of height d
    - Join both views to get complete network
  - 4-ary hypertree of depth d
    - 4^d leaves and $2^d(2^{d+1}–1)$ nodes
    - Diameter is 2d
    - Bisection width is 2^{d+1}

IBM SP2 Overview

- Distributed-memory MIMD multicomputer
- Scalable POWERparallel 1 (SP1)
- Scalable POWERparallel 2 (SP2)
  - RS/6000 workstation plus 4–128 POWER2 processors
  - POWER2 processors used IBM's in RS 6000 workstations, compatible with existing software

SP2 System Architecture

- RS/6000 as system console
- SP2 runs various combinations of serial, parallel, interactive, and batch jobs
  - Partition between types can be changed
  - High nodes — interactive nodes for code development and job submission
  - Thin nodes — compute nodes
  - Wide nodes — configured as servers, with extra memory, storage devices, etc.
- A system "frame" contains 16 thin processor or 8 wide processor nodes
  - Includes redundant power supplies, nodes are hot swappable within frame
  - Includes a high-performance switch for low-latency, high-bandwidth communication

SP2 Processors and Interconnection

- POWER2 processor
  - RISC processor, load-store architecture, various versions from 20 to 62.5 MHz
  - Comprised of 8 semi-custom chips: Instruction Cache, 4 Data Cache, Fixed-Point Unit, Floating-Point Unit, and Storage Control Unit
- Interconnection network
  - Routing
    - Packet switched = each packet may take a different route
    - Cut-through = if output is free, starts sending without buffering first
    - Wormhole routing = buffer on subpacket basis if buffering is necessary
  - Multistage High Performance Switch (HPS) network, scalable via extra stages to keep bw to each processor constant
  - Guaranteed fairness of message delivery

nCUBE 3 Overview

- Distributed-memory MIMD multicomputer (with hardware to make it look like shared-memory multiprocessor)
  - If access is attempted to a virtual memory page marked as "non-resident", the system will generate messages to transfer that page to the local node
- nCUBE 3 could have 8–65,536 processors and up to 65 TB memory
  - Can be partitioned into "subcubes"
- Multiple programming paradigms: SPMD, inter-subcube processing, client /server

nCUBE 3 Processor and Interconnect

- Processor
  - 64-bit custom processor
    - 0.6 μm, 3-layer CMOS, 2.7 million transistors, 50 MHz, 16 KB data cache, 16 KB instruction cache, 100 MFLOPS
    - ALU, FPU, virtual memory management unit, caches, SDRAM controller, 18-port message router, and 16 DMA channels
      - ALU for integer operations, FPU for floating point operations
    - Argument against off-the-shelf processor: shared memory, vector floating-point units, aggressive caches are necessary in workstation market but superfluous here
- Interconnect
  - Hypercube interconnect
    - Wormhole routing + adaptive routing around blocked or faulty nodes

nCUBE 3 I/O

- ParaChannel I/O array
  - Separate network of nCUBE processors
  - 8 computational nodes connect directly to one ParaChannel node
  - ParaChannel nodes can connect to RAID mass storage, SCSI disks, etc.
    - One I/O array can be connected to more than 400 disks

MediaCUBE Overview

- For delivery of interactive video to client devices over a network (from LAN -based training to video-on-demand to homes)
  - MediaCUBE 30 = 270 1.5 Mbps data streams, 750 hours of content
  - MediaCUBE 3000 = 20,000 & 55,000

Kendall Square Research KSR1 Overview and Processor

- COMA distributed-memory MIMD multicomputer (with hardware to make it look like shared-memory multiprocessor)
- Multiple variations
  - 8 cells ($500K): 320 MFLOPS, 256 MB memory, 210 GB disk, 210 MB/s I/O
  - 1088 cells ($30M): 43 GFLOPS, 34 GB memory, 15 TB disk, 15 GB/s I/O
- Each APRD (ALLCACHE Processor, Router, and Directory) Cell contains:
  - Custom 64-bit integer and floating-point processors (1.2 μm, 20 MHz, 450,000 transistors, on a 8x13 printed circuit board)
  - 32 MB of local cache
  - Support chips for cache, I/O, etc.

KSR1 System Architecture

- The ALLCACHE system moves an address set requested by a processor to the Local Cache on that processor
  - Provides the illusion of a single sequentially-consistent shared memory
- Memory space consists of all the 32 KB local caches
  - No permanent location for an “address”
  - Addresses are distributed and based on processor need and usage patterns
  - Each processor is attached to a Search Engine, which finds addresses and their contents and moves them to the local cache, while maintaining cache coherence throughout the system
    - 2 levels of search groups for scalability

Cray T3D Overview

- NUMA shared-memory MIMD multiprocessor
  - Each processor has a local memory, but the memory is globally addressable
- DEC Alpha 21064 processors arranged into a virtual 3D torus (hence the name)
  - 32–2048 processors, 512MB–128GB of memory
  - Parallel vector processor (Cray Y-MP / C90) used as host computer, runs the scalar / vector parts of the program
  - 3D torus is virtual, includes redundant nodes

T3D Nodes and Interconnection

- Node contains 2 PEs; each PE contains:
  - DEC Alpha 21064 microprocessor
    - 150 MHz, 64 bits, 8 KB L1 I&D caches
    - Support for L2 cache, not used in favor of improving latency to main memory
  - 16–64 MB of local DRAM
    - Access local memory: latency 87–253ns
    - Access remote memory: 1–2μs (~8x)
  - Alpha has 43 bits of virtual address space, only 32 bits for physical address space — external registers in node provide 5 more bits for 37 bit phys. addr.
- 3D torus connections PE nodes and I/O gateways
  - Dimension-order routing: when a message leaves a node, it first travels in the X dimension, then Y, then Z

Cray T3E Overview

- T3D = 1993,
  T3E = 1995 successor (300 MHz, $1M),
  T3E-900 = 1996 model (450 MHz, $.5M)
- T3E system = 6–2048 processors, 3.6 –1228 GFLOPS,1–4096 GB memory
  - PE = DEC Alpha 21164 processor (300 MHz, 600 MFLOPS, quad issue), local memory, control chip, router chip
    - L2 cache is on-chip so can't be eliminated, but off-chip L3 can and is
    - 512 external registers per process
  - GigaRing Channel attached to each node and to I/O devices and other networks
  - T3E-900 = same w/ faster processors, up to 1843 GFLOPS
- Ohio Supercomputer Center (OSC) had a T3E with 128 PEs (300 MHz), 76.8 GFLOPS, 128 MB memory / PE

Convex Exemplar SPP-1000 Overview

- ccNUMA shared-memory MIMD
  - 4–128 HP PA 7100 RISC processors, 256 MB – 32 GB memory
  - Hardware support for remote memory access
- System is comprised of up to 16 "hypernodes", each of which contains 8 processors and 4 cache memories (each 64–512MB) connected by a crossbar switch
  - Hypernodes are connected in a ring
  - Hardware keeps caches consistent with each other

Silicon Graphics
POWER CHALLENGEarray Overview

- ccNUMA shared-memory MIMD
- "Small" supercomputers
  - POWER CHALLENGE — up to 144 MIPS R8000 processors or 288 MISP R1000 processors, with up to 128 GB memory and 28 TB of disk
  - POWERnode system — shared-memory multiprocessor of up to 18 MIPS R8000 processors or 36 MIPS R1000 processors, with up to 16 GB of memory
- POWER CHALLENGEarray consists of up to 8 POWER CHALLENGE or POWERnode systems
  - Programs that fit within a POWERnode can use the shared-memory model
  - Larger program can span POWERnodes

Silicon Graphics
Origin 2000 Overview

- ccNUMA shared-memory MIMD
  - SGI says they supply 95% of ccNUMA systems worldwide
- Various models, 2–128 MIPS R10000 processors, 16 GB – 1 TB memor
  - Processing node board contains two R10000 processors, part of the shared memory, directory for cache coherence, plus node and I/O interface
- File serving, data mining, media serving, high-performance computing