Intel Paragon XP/S Overview

- Distributed-memory MIMD multicomputer
- 2D array of nodes, performing both OS functionality as well as user computation
  - Main memory physically distributed among nodes (16-64 MB / node)
  - Each node contains two Intel i860 XP processors: application processor for user's program, message processor for inter-node communication
- Balanced design: speed and memory capacity matched to interconnection network, storage facilities, etc.
  - Interconnect bandwidth scales with number of nodes
  - Efficient even with thousands of processors

Intel MP Paragon XP/S 150 @ Oak Ridge National Labs

Paragon XP/S Nodes

- Network Interface Controller (NIC)
  - Connects node to its PMRC
  - Parity-checked, full-duplexed router with error checking
- Message processor
  - Intel i860 XP processor
  - Handles all details of sending / receiving a message between nodes, including protocols, packetization, etc.
  - Supports global operations including broadcast, synchronization, sum, min, and, or, etc.
- Application processor
  - Intel i860 XP processor (42 MIPS, 50 MHz clock) to execute user programs
- 16–64 MB of memory

Paragon XP/S Node Interconnection

- 2D mesh chosen after extensive analytical studies and simulation
- 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
    - 40 ns to make routing decisions and close appropriate switches
    - Transfers are parity checked, router is pipelined, routing is deadlock-free
  - Backplane is active backplane of router chips rather than mass of cables

Paragon XP/S Usage

- 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
- Applications can run on arbitrary number of nodes without change
  - Run on larger number of nodes to process larger data sets or to achieve required performance
- Users have client/server access, can submit jobs over a network, or login directly to any node
  - Comprehensive resource control utilities for allocating, tracking, and controlling system resources

Paragon XP/S Programming

- MIMD architecture, but supports various programming models: SPMD, SIMD, MIMD, shared memory, vector shared memory
- CASE tools including:
  - Optimizing compilers for FORTRAN, C, C++, Ada, and Data Parallel FORTRAN
  - Interactive Debugger
  - Parallelization tools: FORGE, CAST
  - Intel's ProSolver library of equation solvers
  - Intel's Performance Visualization System (PVS)
  - Performance Analysis Tools (PAT)

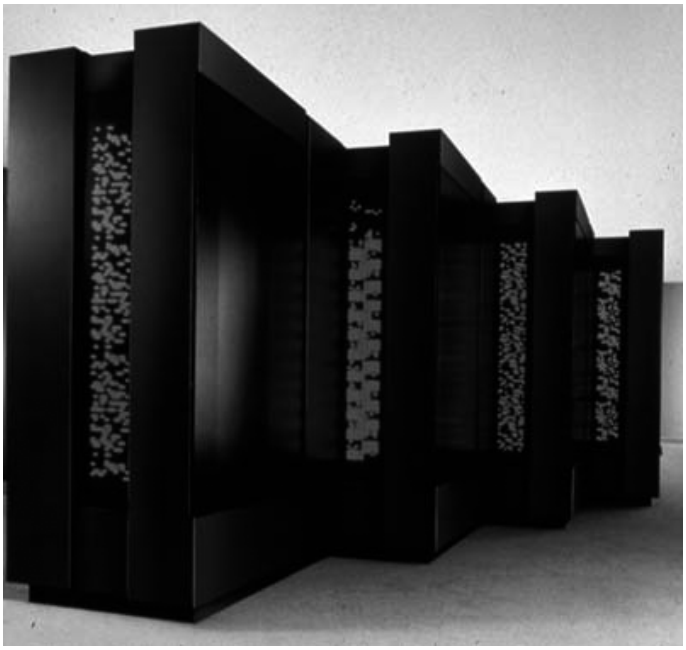
Thinking Machines CM-5 Overview

- Distributed-memory MIMD multicomputer
  - SIMD or MIMD operation
- Processing nodes are supervised by a control processor, which runs UNIX
  - Control processor broadcasts blocks of instructions to the processing nodes, and initiates execution
  - SIMD operation: nodes are closely synchronized, blocks broadcast as needed
  - MIMD operation: nodes fetch instructions independently and synchronize only as required by the algorithm
- Nodes may be divided into partitions
  - One control processor, called the partition manager, per partition
  - Partitions may exchange data

Thinking Machines CM-5 Overview (cont.)

- Other control processors, called I/O Control Processors, manage the system's I/O devices
  - Scale to achieve necessary I/O capacity
  - DataVaults to provide storage
- Control processors in general
  - Scheduling user tasks, allocating resources, servicing I/O requests, accounting, security, etc.
  - May execute some code
  - No arithmetic accelerators, but additional I/O connections
  - In small system, one control processor may play a number of roles
  - In large system, control processors are often dedicated to particular tasks (partition manager, I/O cont. proc., etc.)

Thinking Machines CM-5@ NCSA



CM-5 Nodes

- Processing nodes
  - SPARC CPU
    - 22 MIPS
  - 8-32 MB of memory
  - Network interface
  - (Optional) 4 pipelined vector processing units
    - Each can perform up to 32 million double-precision floating-point operations per second
    - Including divide and square root
- Fully configured CM-5 would have
  - 16,384 processing nodes
  - 512 GB of memory
  - Theoretical peak performance of 2 teraflops

CM-5 Networks

- Control Network
  - Tightly coupled communication services
  - Optimized for fast response, low latency
  - Functions: synchronizing processing nodes, broadcasts, reductions, parallel prefix operations
- Data Network
  - 4-ary hypertree, optimized for high bandwidth
  - Functions: point-to-point commn. for tens of thousands of items simultaneously
  - Responsible for eventual delivery of messages accepted
  - Network Interface connects nodes or control processors to the Control or Data Network (memory-mapped control unit)

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}

CM-5 Usage

- Runs Cmost, enhanced vers. of SunOS
- User task sees a control processor acting as a Partition Manager (PM), a set of processing nodes, and inter-processor communication facilities
  - User task is a standard UNIX process running on the PM, and one on each of the processing nodes
  - The CPU scheduler schedules the user task on all processors simultaneously
- User tasks can read and write directly to the Control Network and Data Network
  - Control Network has hardware for broadcast, reduction, parallel prefix operations, barrier synchronization
  - Data Network provides reliable, deadlock-free point-to-point communication