Intel Paragon XP/S Overview

- Scalable, heterogeneous, distributed-memory multicomputer, MIMD
- 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

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
- Network Interface Controller (NIC)
  - Connects node to its PMRC
  - Parity-checked, full-duplexed router with error checking

Paragon XP/S Nodes

- Application processor
  - Intel i860 XP processor (42 MIPS, 50 MHz clock) to execute user programs
- Message processor
  - Intel i860 XP processor dedicated to inter-node communication
  - Handles communication traffic for node
  - Handles all details of sending / receiving a message, including protocols, packetization, etc.
  - Division of node into application and message processors reduces cache turbulence and avoids context switches; messaging software may even remain in cache
  - Supports global operations including broadcast, synchronization, sum, min, and, or,etc.

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)

Connection Machine CM-5 Overview

- Hundreds or thousands of processing nodes, each with its own memory
  - 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

Connection Machine 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
- Each control processor and parallel processing node connects to 2 networks
  - Control Network: for operations that involve all nodes at once, such as synchronization or broadcast
  - Data Network: for bulk data transfers between specific source and destination
  - Diagnostic Network : checks the network
- Same mechanism throughout for transferring data between partitions and I/O devices

CM-5 Nodes

- Processing nodes
  - 8-32 MB of memory, 128 MIPS/MFLOPS
  - Optional vector units for high-speed arithmetic
- Control processors
  - 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, individual control processors are often dedicated to particular tasks (partition manager, I/O control processor, etc.)

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
  - Loosely coupled communication services
  - 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)

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