IBM SP2 Overview

- Distributed-memory MIMD multicomputer
- Scalable POWERparallel 1 (SP1)
  - Development started February 1992, delivered to users in April 1993
- Scalable POWERparallel 2 (SP2)
  - 120-node systems delivered 1994
    - 4–128 nodes: RS/6000 workstation with POWER2 processor, 66.7 MHz, 267 MFLOPS
    - POWER2 used in RS 6000 workstations, gives compatibility with existing software
  - 1997 version (NUMA):
    - High Node (SMP node, 16 nodes max): 2–8 PowerPC 604, 112 MHz, 224 MFLOPS, 64MB–2GB memory
    - Wide Node (128 nodes max): 1 P2SC (POWER2 Super Chip, 8 chips on one chip), 135 MHz, 640 MFLOPS, 64MB–2GB memory

IBM SP2 @ Oak Ridge National Labs

IBM SP2 Overview (cont.)

- 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

IBM SP2 Processors

- POWER2 processor
  - Various versions from 20 to 62.5 MHz
  - RISC processor, load-store architecture
    - Floating point multiple & add instruction with latency of 2 cycles, pipelined for initiation of new one each cycle
    - Conditional branch to decrement and test a “count register” (without fixed-point unit involvement), good for loop closings
- POWER 2 processor chip set
  - 8 semi-custom chips: Instruction Cache Unit, four Data Cache Units, Fixed-Point Unit (FXU), Floating-Point Unit (FPU), and Storage Control Unit
    - 2 execution units per FXU and FPU
    - Can execute 6 instructions per cycle: 2 FXU, 2 FPU, branch, condition register
    - Options: 4-word memory bus with 128 KB data cache, or 8-word with 256 KB

IBM SP2 Interconnection Network

- General
  - Multistage High Performance Switch (HPS) network, with extra stages added to keep bw to each processor constant
  - Message delivery
    - PIO for short messages with low latency and minimal message overhead
    - DMA for long messages
  - Multi-user support — hardware protection between partitions and users, guaranteed fairness of message delivery
- 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

IBM SP2 AIX Parallel Environment

- Parallel Operating Environment — based on AIX, includes Desktop interface
  - Partition Manager to allocate nodes, copy tasks to nodes, invoke tasks, etc.
  - Program Marker Array — (online) squares graphically represent program tasks
  - System Status Array — (offline) squares show percent of CPU utilization
- Parallel Message Passing Library
- Visualization Tool — view online and offline performance
  - Group of displays for communications characteristics or performance (connectivity graph, inter-processor communication, message status, etc.)
- Parallel Debugger

nCUBE Overview

- Distributed-memory MIMD multicomputer (with hardware to make it look like shared-memory multiprocessor)
- History
  - nCUBE 1 — 1985
  - nCUBE 2 — 1989
    - 34 GFLOPS, scalable
    - ?–8192 processors
  - nCUBE 3 — 1995
    - 1–6.5 TFLOPS, 65 TB memory,
24 TB/s hypercube interconnect, 1024
3 GB/s I/O channels, scalable
    - 8–65,536 processors
- Operation
  - Can be partitioned into “subcubes”
  - Programming paradigms: SPMD, inter-subcube processing, client/server

nCUBE 3 Processor

- 0.6 µm, 3-layer CMOS, 2.7 million transistors, 50 MHz, 16 KB data cache, 16 KB instruction cache, 100 MFLOPS
  - Argument against off-the-shelf processor: shared memory, vector floating-point units, aggressive caches are necessary in workstation market but superfluous here
- 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, both 64 bit
    - Most integer operations execute in one 20ns clock cycle
    - FPU can complete two single- or double-precision operations in one clock cycle
  - Virtual memory pages can be marked as “non-resident”, the system will generate messages to transfer page to local node

nCUBE 3 Interconnect

- Hypercube interconnect
  - Added hypercube dimension allows for double the processors, but processors can be added in increments of 8
  - Wormhole routing + adaptive routing around blocked or faulty nodes
- ParaChannel I/O array
  - Separate network of nCUBE processors for load distribution and I/O sharing
  - 8 computational nodes (nCUBE processors plus local memory) connect directly to one ParaChannel node, and can also communicate with those nodes via the regular hypercube network
  - ParaChannel nodes can connect to RAID mass storage, SCSI disks, etc.
    - One I/O array can be connected to more than 400 disks

nCUBE 3 Software

- Parallel Software Environment
  - nCX microkernel OS — runs on all compute nodes and I/O nodes
  - UNIX functionality
  - Programming languages including FORTRAN 90, C, C++, as well as HPF, Parallel Prolog, and Data Parallel C

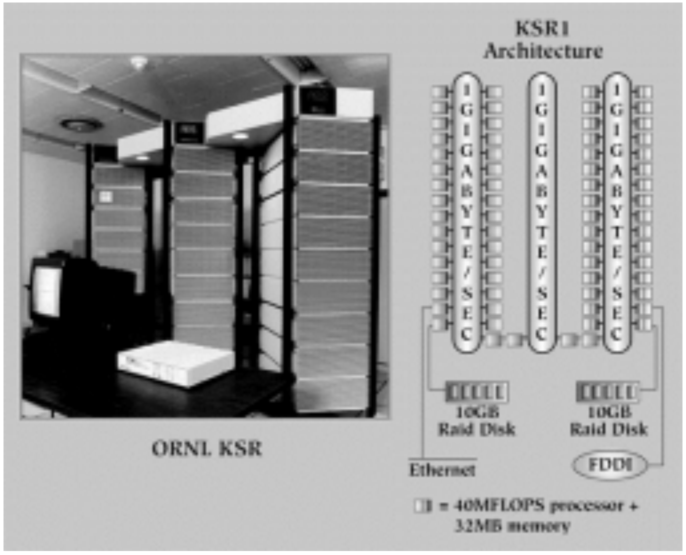
MediaCUBE Overview

- Emphasized on their web page;
  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

- COMA distributed-memory MIMD multicomputer (with hardware to make it look like shared-memory multiprocessor)
- 6 years in development, 36 variations in 1992 (8 cells for $500k, 1088 for $30m)
  - 8 cells: 320 MFLOPS, 256 MB memory, 210 GB disk, 210 MB/s I/O
  - 1088 cells: 43 GFLOPS, 34 GB memory, 15 TB disk, 15 GB/s I/O
- Each system includes:
  - Processing Modules, each containing up to 32 APRD Cells including 1GB of ALLCACHE memory
  - Disk Modules, each containing 10 GB
  - I/O adapters
  - Power Modules, with battery backup

KSR1 @ Oak Ridge National Labs



Kendall Square Research KSR1 Processor Cells

- Each APRD (ALLCACHE Processor, Router, and Directory) Cell contains:
  - 64-bit Floating Point Unit, 64-bit Integer Processing Unit
  - Cell Execution Unit for address gen.
  - 4 Cell Interconnection Units, External I/O Unit
  - 4 Cache Control Units
  - 32 MB of Local Cache, 512 KB of subcache
- Custom 64-bit processor: 1.2 µm, each up to 450,000 transistors, packaged in 8x13x1 printed circuit board
  - 20 MHz clock
  - Can execute 2 instructions per cycle

Kendall Square Research KSR1 ALLCACHE System

- 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

Kendall Square Research KSR1 Programming Environment

- KSR OS = enhanced OSF/1 UNIX
  - Scalable, supports multiple computing modes including batch, interactive, OLTP, and database management and inquiry
- Programming languages
  - FORTRAN with automatic parallelization
  - C
  - PRESTO parallel runtime system that dynamically adjusts to number of available processors and size of the current problem