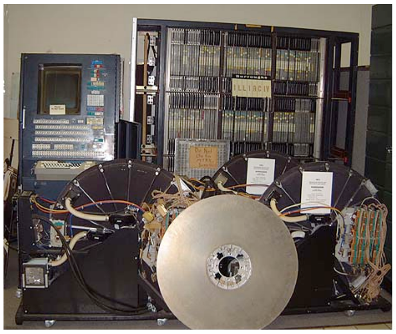
SIMD+ Overview

- Early machines
  - Illiac IV (first SIMD)
  - Cray-1 (vector processor, not a SIMD)
- SIMDs in the 1980s and 1990s
  - Thinking Machines CM-2 (1980s)
- General characteristics
  - Host computer to interact with user and execute scalar instructions, control unit to send parallel instructions to PE array
  - 100s or 1000s of simple custom PEs, each with its own private memory
  - PEs connected by 2D torus, maybe also by row/column bus(es) or hypercube
  - Broadcast / reduction network

Illiac IV History

- First massively parallel (SIMD) computer
- Sponsored by DARPA, built by various companies, assembled by Burroughs, under the direction of Daniel Slotnick at the University of Illinois
  - Plan was for 256 PEs, in 4 quadrants of 64 PEs, but only one quadrant was built
  - Used at NASA Ames Research Center in mid-1970s



Illiac IV Architectural Overview

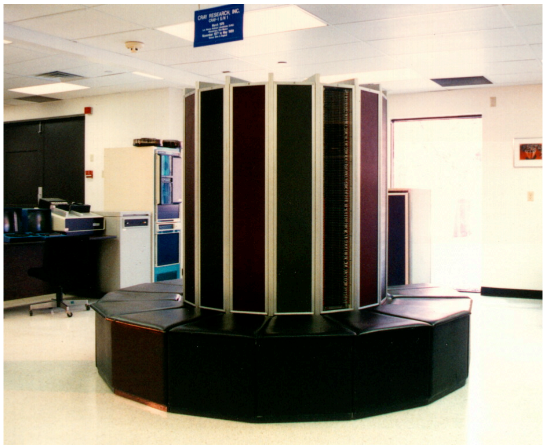
- CU (control unit) +
  64 PUs (processing units)
  - PU = 64-bit PE (processing element) + PEM (PE memory)
- CU operates on scalars,
  PEs operate on vector-aligned arrays
  (A[1] on PE 1, A[2] on PE2, etc.)
  - All PEs execute the instruction broadcast by the CU, if they are in active mode
  - Each PE can perform various arithmetic and logical instructions on data in 64-bit, 32-bit, and 8-bit formats
  - Each PEM contains 2048 64-bit words
- Data routed between PEs various ways
- I/O is handled by a separate Burroughs B6500 computer (stack architecture)

Illiac IV Routing and I/O

- Data routing
  - CU bus —instructions or data can be fetched from a PEM and sent to the CU
  - CDB (Common Data Bus) — broadcasts information from CU to all PEs
  - PE Routing network — 2D torus
- Laser memory
  - 1 Tb write-once read-only laser memory
  - Thin film of metal on a polyester sheet, on a rotating drum
- DFS (Disk File System)
  - 1 Gb, 128 heads (one per track)
- ARPA network link (50 Kbps)
  - Illiac IV was a network resource available to other members of the ARPA network

Cray-1 History

- First famous vector (not SIMD) processor
- In January 1978 there were only 12 non-Cray-1 vector processors worldwide:
  - Illiac IV, TI ASC (7 installations), CDC STAR 100 (4 installations)



Cray-1 Vector Operations

- Vector arithmetic
  - 8 vector registers, each holding a 64-element vector (64 64-bit words)
  - Arithmetic and logical instructions operate on 3 vector registers
    - Vector C = vector A + vector B
    - Decode the instruction once, then pipeline the load, add, store operations
- Vector chaining
  - Multiple functional units
    - 12 pipelined functional units in 4 groups: address, scalar, vector, and floating point
    - Scalar add = 3 cycles, vector add = 3 cycles, floating-point add = 6 cycles, floating-point multiply = 7 cycles, reciprocal approximation = 14 cycles
  - Use pipelining with data forwarding to bypass vector registers and send result of one functional unit to input of another

Cray-1 Physical Architecture

- Custom implementation
  - Register chips, memory chips, low-speed and high-speed gates
- Physical architecture
  - “Cylindrical tower (6.5’ tall, 4.5’ diameter) with 8.5’ diameter seat
    - Composed of 12 wedge-like columns in 270° arc, so a “reasonably trim individual” can get inside to work
  - World’s most expensive love-seat”
    - “Love seat” hides power supplies and plumbing for Freon cooling system
- Freon cooling system
  - Vertical cooling bars line each wall, modules have a copper heat transfer plate that attaches to the cooling bars
  - Freon is pumped through a stainless steel tube inside an aluminum casing

Thinking Machines Corporation’s Connection Machine CM-2

- Distributed-memory SIMD (bit-serial)
- Thinking Machines Corp. founded 1983
  - CM-1, 1986 (1000 MIPS, 4K processors)
  - CM-2, 1987 (2500 MFLOPS, 64K…)
- Programs run on one of 4 Front-End Processors, which issue instructions to the Parallel Processing Unit (PE array)
  - Control flow and scalar operations run on Front-End Processors, while parallel operations run on the PPU
  - A 4x4 crossbar switch (Nexus) connects the 4 Front-Ends to 4 sections of the PPU
  - Each PPU section is controlled by a Sequencer (control unit), which receives assembly language instructions and broadcasts micro-instructions to each processor in that PPU section

CM-2 Nodes / Processors

- CM-2 constructed of "nodes", each with:
  - 32 processors (implemented by 2 custom processor chips), 2 floating-point accelerator chips, and memory chips
- 2 processor chips (each 16 processors)
  - Contains ALU, flag registers, etc.
  - Contains NEWS interface, router interface, and I/O interface
    - 16 processors are connected in a 4x4 mesh to their N, E, W, and S neighbors
- 2 floating-point accelerator chips
  - First chip is interface, second is FP execution unit
- RAM memory
  - 64Kbits, bit addressable

CM-2 Interconnect

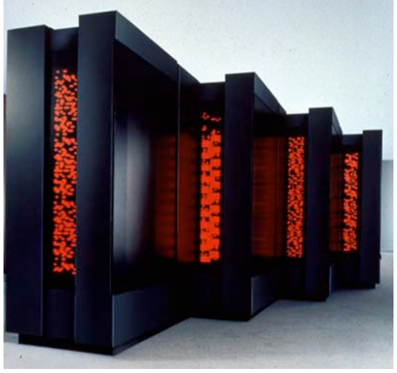
- Broadcast and reduction network
  - Broadcast, Spread (scatter)
  - Reduction (e.g., bitwise OR, maximum, sum), Scan (e.g., collect cumulative results over sequence of processors such as parallel prefix)
  - Sort elements
- NEWS grid can be used for nearest-neighbor communication
  - Communication in multiple dimensions: 256x256, 1024x64, 8x8192, 64x32x32, 16x16x16x16, 8x8x4x8x8x4
- The 16-processor chips are also linked by a 12-dimensional hypercube
  - Good for long-distance point-to-point communication

MIMD Overview

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

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

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

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

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