Cray T3D Overview

- NUMA shared-memory MIMD multiprocessor
- DEC Alpha 21064 processors arranged into a virtual 3D torus (hence the name)
  - 32–2048 processors, 512MB–128GB of memory, 4.2–300 GFLOPS
  - Parallel vector processor (Cray Y-MP or C90) acts as a host computer, and runs the scalar / vector parts of the program
  - Each processor has a local memory, but the memory is globally addressable
  - 3D torus is virtual — the physical layout includes redundant nodes that can be mapped into the torus if a node fails
- System contains:
  - Processing element nodes, Interconnect network, I/O gateways

Cray T3D

T3D Processing Element Nodes

- Node contains 2 PEs; each PE contains:
  - DEC Alpha 21064 microprocessor
    - 150 MHz, 64 bits, double issue, 8/8 KB L1 caches
    - Support for L2 cache, eliminated 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.
- Node also contains:
  - Network interface
  - Block transfer engine (BLT) — asynchronously distributes data between PE memories and rest of system (overlap computation and communication)

T3D Interconnection Network and I/O Gateways

- Interconnection network
  - Between PE nodes and I/O gateways
  - 3D torus between routers, each router connecting to a PE node or I/O gateway
  - Dimension-order routing: when a message leaves a node, it first travels in the X dimension, then Y, then Z
- I/O gateways
  - Between host and T3D, or between T3D and an I/O cluster (workstations, tape drives, disks, and / or networks)
  - Hide latency:
    - Alpha has a FETCH instruction that can initiate a memory prefetch
    - Remote stores are buffered — 4 words accumulate before store is performed
    - BLT redistributes data asynchronously

Cray T3D Usage

- Processors can be divided into partitions
  - System administrator can define a set of processors as a pool, specifying batch use, interactive use, or both
  - User can request a specific number of processors from a pool for an application, MAX selects that number of processors and organizes them into a partition
- A Cray Y-MP C90 multiprocessor is used as a UNIX server together with the T3D
  - OS is MAX, Massively Parallel UNIX
  - All I/O is attached to the Y-MP, and is available to T3D
  - Shared file system between Y-MP & T3D
  - Some applications run on the Y-MP, others on the T3D, some on both

Cray T3D Programming

- Programming models
  - Message passing based on PVM
  - High-Performance FORTRAN — implicit remote accessing
  - MPP FORTRAN — implicit and explicit communication
- OS is UNICOS MAX, a superset of UNIX
  - Distributed OS with functionality divided between host and PE nodes (microkernels only on PE nodes)

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) has 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, up to 25 GFLOPS
  - 256 MB – 32 GB memory
  - Logical view: processors – crossbar – shared memory – crossbar – peripherals
- System made up of "hypernodes", each of which contains 8 processors and 4 cache memories (each 64–512MB) connected by a crossbar switch
  - Hypernodes connected in a ring via Coherent Toroidal Interconnect, an implementation of IEEE 1596-1992, Scalable Coherency Interface
    - Hardware support for remote memory access
    - Keeps the caches at each procesor consistent with each other

Cray Exemplar SPP-1000

Convex Exemplar SPP-1000 Processors

- HP PA7100 RISC processor
  - 555,000 transistors, 0.8µ
- 64 bit wide, external 1MB data cache & 1MB instruction cache
  - Reads take 1 cycle, writes take 2
- Can execute one integer and one floating-point instruction per cycle
- Floating Point Unit can multiply, divide / square root, etc. as well as multiply-add and multiply-subtract
  - Most fp operations take two cycles, divide and square root take 8 for single precision and 15 for double precision
- Supports multiple threads, and hardware semaphore & synchronization operations

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 109 GFLOPS, 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 POWER CHALLENGEarray Programming

- Fine- to medium-grained parallelism
  - Shared-memory techniques within a POWERnode, using parallelizing FORTRAN and C compilers
- Medium- to coarse-grained parallelism
  - Shared-memory within a POWERnode or message-passing between POWERnode
  - Applications based on message-passing will run within a POWERnode, and libraries such as MPI or PVM will use the shared-memory instead
- Large applications
  - Hierarchical programming, using a combination of the two techniques

Silicon Graphics Origin 2000 Overview

- ccNUMA shared-memory MIMD
- Various models, 2–128 MIPS R10000 processors, 16 GB – 1 TB memory, 6–200 GB/s peak I/O bandwidth
  - Largest model also supports 96 PCI cards and 160 Ultra SCSI devices
  - Processing node board contains two R10000 processors, part of the shared memory, directory for cache coherence, node interface, and I/O interface
- ccNUMA (SGI says they supply 95% of ccNUMA systems worldwide)
  - Crossbar switches that scale upwards
- Packaged solutions for business (file serving, data mining), Internet (media serving), & high-performance computing

SGI Origin 2000 @ Boston University