SPP 2000 Architecture

Beth Richardson
NCSA Consultant and Trainer
bethr@ncsa.uiuc.edu

Outline

- Overview
- Processor
- Cache
- Node
- Memory
- Interconnect
- I/O
- Data Mover

Overview of the SPP 2000

Second generation of SPP from HP/Convex
Maintains SPP 1200 binary compatibility
Uses HP PA-8000 processors
Bigger, faster, more scalable than SPP 1200
Higher bandwidth and lower latency interconnect

Building Blocks of the SPP 2000

- Processor
  - the smallest building block
- Node
  - an SMP collection of 16 processors
  - just one node have an S class machine
  - multiple nodes have an X class machine
- Wall
  - a collection of 32 nodes (512 processors)

Processor

- Named the PA-8000
- Clock speed is 180 Mhz
- Can execute 4 flops peak per clock
- This gives rise to 720 Mflops peak
- Does prefetching and out of order execution
- 3.8 million transistors
- Size of chip is 338 mm^2

NCSA
University of Illinois at Urbana-Champaign

Processor (Cont.)

Functional Units

- 2 floating point
- 2 divide/square root
- 2 integer
- 2 shift/merge
- 2 load store

NCSA
University of Illinois at Urbana-Champaign

Super Scalar Execution

- 4 way superscalar
- In any cpu cycle can have:
  - Up to 2 floating point instructions
  - Up to 2 integer instructions
  - Up to 2 load/stores
  - Max of 2 integer and load/stores combined

Cache

- Size of Cache
  - 1 Mbyte data cache
  - 1 Mbyte instruction cache
- Cache is
  - external
  - direct mapped
  - dual ported
- Cache line is
  - 32 bytes wide
- Cache coherency is directory based

Cache (Cont.)

- Cache latency is
  - 3 clocks (1 when pipelined)
- Bandwidth of data from cache to a register
  - 2.88 Gbytes/sec
- Bandwidth of data from memory to cache
  - 960 Mbytes/sec

Processor Follow-On

- Will be named the PA-8500
- 1.5 Mbyte of on-chip cache
  - .5 Mbyte for instructions
  - 1.0 Mbyte for data
  - cache is 4 way set associative
- CPU core shrinks to $80mm^2$
- Clock speed˝estimated at 400 Mhz
- Production date estimated as 2H98

Processor - Next Generation

- Named the Merced chip
- Collaboration between HP and Intel

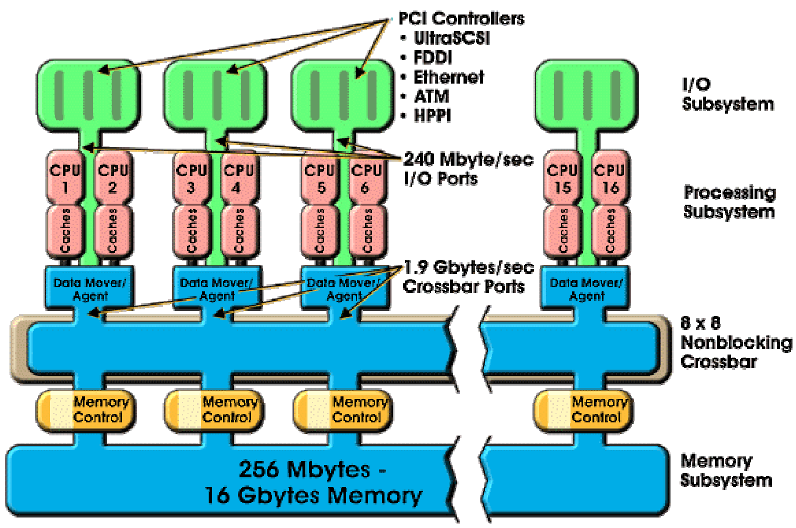
Node

- 16 processor boards
  - Every 2 processors share an agent chip
  - Where the agent chip is the traffic manager
- 8 memory boards
- Crossbar switch connects processors and memory
- 24 PCI controllers
- Power supply
- Cooling
- Nodes are configurable into subcomplexes

Node Physicals

Weighs approximately 500 pounds
Measures 29 x 36 x 35 inches
Air cooled
Nodes can be stacked 2 high (maximum)

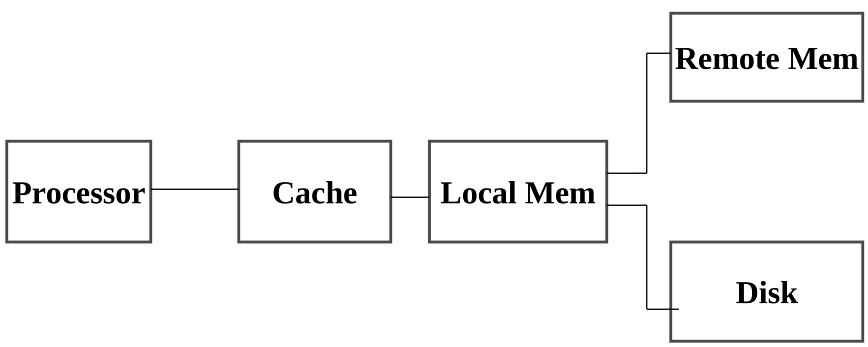
Node Graphic



Global Shared Memory

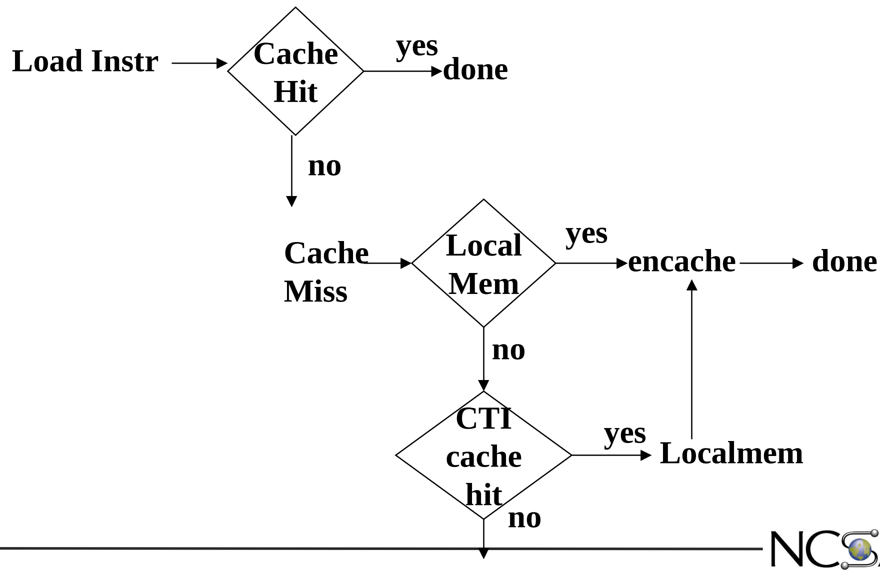
- True hardware based DSM
  - memory is physically distributed and virtually shared
- Memory is ccNUMA
  - cache coherent non-uniform memory access
- Load instruction can access data
  - in cache (means local to your processor)
  - in local memory (means local to your node)
  - in remote memory (means on another node)
  - on disk (in virtual memory)
- Named GSM for Global Shared Memory

Global Shared Memory Graphic



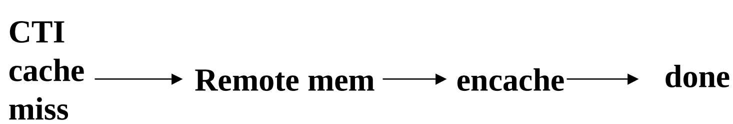
NCSA
University of Illinois at Urbana-Champaign

Global Shared Memory



University of Illinois at Urbana-Champaign

Global Shared Memory (Cont.)



University of Illinois at Urbana-Champaign

Memory

Uses SDRAMS

4 Gbyte/node

4 way interleave per memory board

Memory latencies

- to retrieve a cache line is 105 nanoseconds
- to local memory for a cache line is .51 microsecs
- to remote memory for a cache line is 1.56 microsecs

Interconnect

The 16 processors in a node are interconnected with a crossbar switch

Multiple nodes are interconnected in a 2D torus

A router supports alternate routing paths between nodes

Crossbar

- 8x8 non-blocking crossbar
- Each processor has 2 ports to the crossbar
  - one for input, one for output
- Hence the crossbar is actually 2 crossbars
  - input and output on the crossbar are concurrent
- The data path
  - is 64 bits wide
- The crossbar bandwidth
  - is 960MB/sec per crossbar port

CTI Interconnect

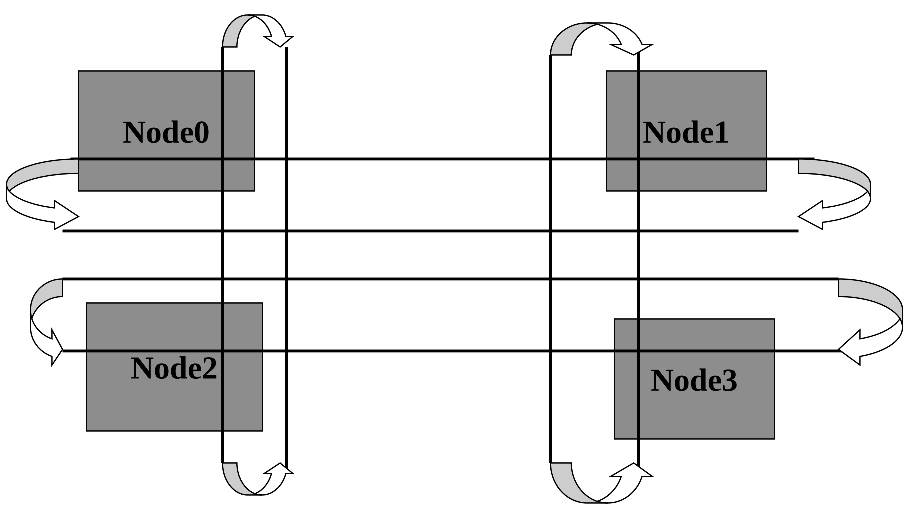
- CTI is Coherent Toroidal Interconnect
- It’s a proprietary SCI network
- In 2D it’s a torus
- Rings
  - 8 y direction links
  - 8 x direction links
  - Each link connects 2 processors/node
  - 32 bit wide data path
  - the multiple rings spread out the load on the interconnect

NCSA
University of Illinois at Urbana-Champaign

CTI Interconnect (Cont.)

- CTI controller has routing tables to permit packet redirection
- Link bandwidth is 480 Mbyte/sec
- Node to node bandwidth is 3.84 Gbyte/sec

CTI Rings Graphic



I/O

- Each of the agent chips has an I/O channel
- Each I/O channel supports a 120 MB/sec PCI bus
- I/O path is 32 bits wide
- Each PCI bus can have up to 3 controllers
- Each PCI controller supports 10 Ultra-SCSI disks
- Each disk has 9 Gbytes capacity
- Estimated bandwidth is 30-35 MB/sec per controller

Data Mover

- A hardware bcopy engine
- Does copying, zero-ing and messaging
- Implemented in the agent chip, so 8 data movers per node
- Works asynchronously with the processors
- Works both inter and intra node
- Rated at 500 MB/sec
- Used by the OS
- User invoked by calling ail_bcopy instead of bcopy in message passing programs