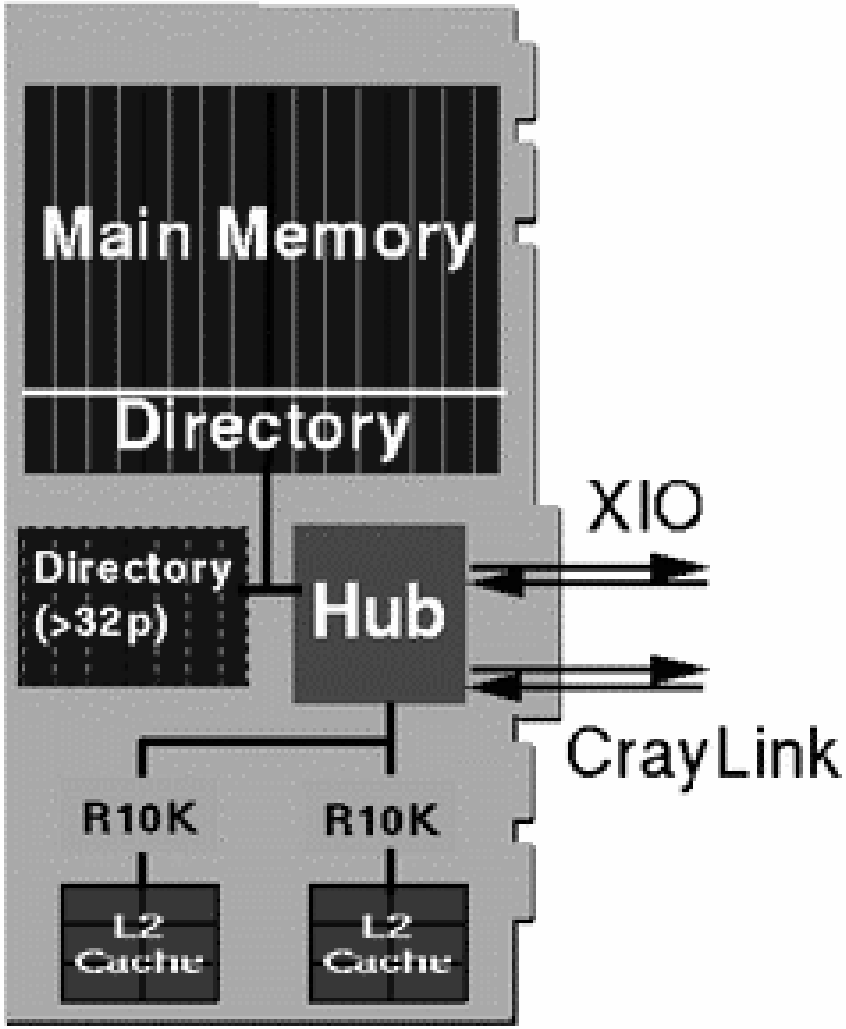
Origin2000 Architecture

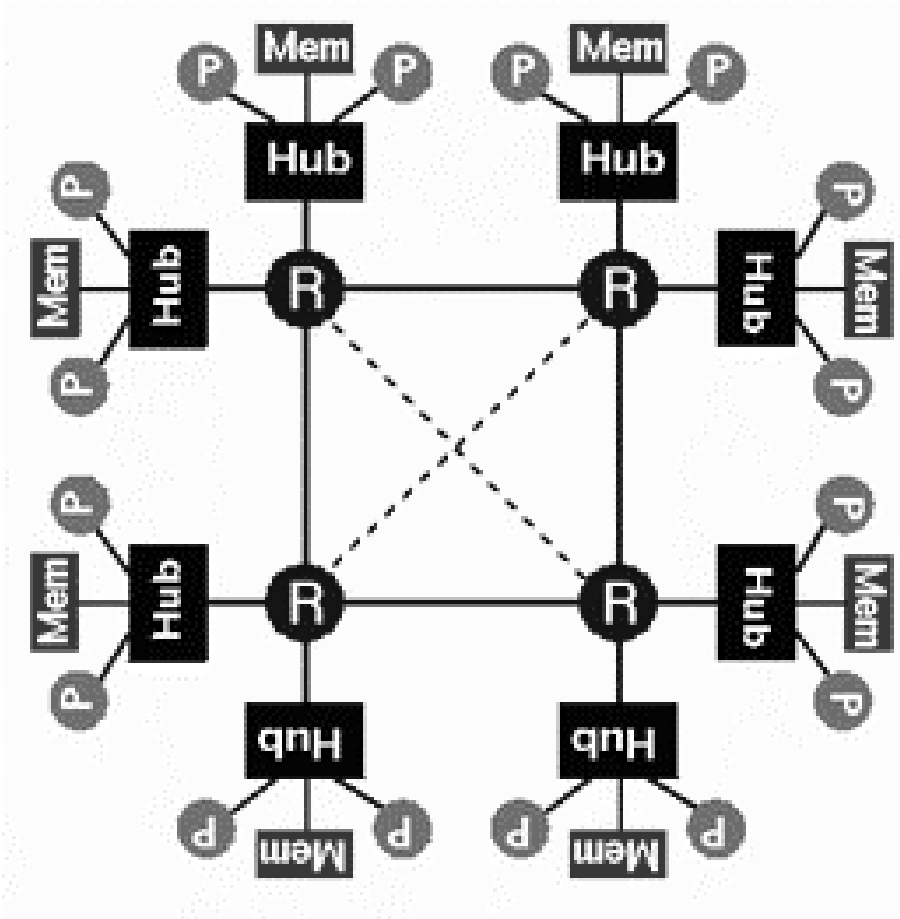
Distributed Shared Memory System

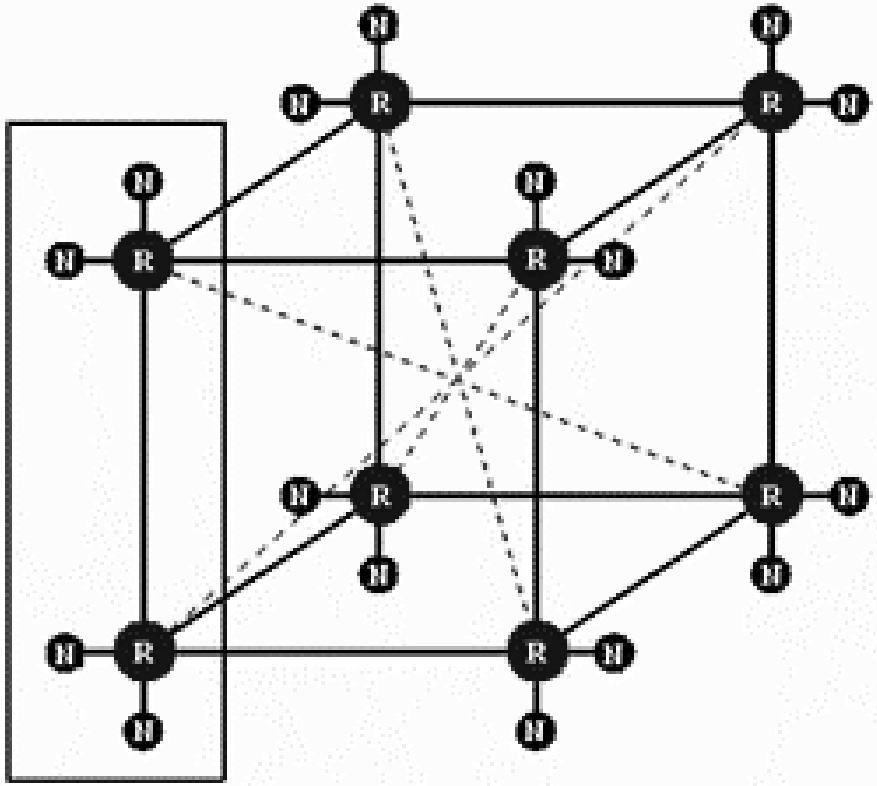
- **SMP nodes (2 processors nodes boards)**
- **High speed interconnect**
  - **proprietary internal network called CrayLink**
  - **fat hypercube topology**

Node Board


Main Memory
Directory
Directory
(>32p)
Hub
XIO
CrayLink
R10K
R10K
L2
Cache
L2
Cache

16 and 32 Processor Systems





CrayLink Interconnect

- **HUB is a crossbar switch**
- **Point-to-Point fabric**
  - **each HUB can only send or receive data with one neighbor at a time**
  - **each router is a crossbar switch**
  - **multiple between any two HUBs**

R10000 Processors

- 195 Mhz Clock
- MIPS IV instruction set
- 4-way superscalar
- 390 Mflop/s peak
  - 1 mult-add/cycle (chained) OR
  - 2 FP instructions
- 32KB floating point and 32KB integer/instruction data caches
- 4MB secondary cache
  - 2-way set associatve

R10000 Processors

- 64 bit processor
  - 64 bit registers
  - 64 bit integer and floating point operations
  - 64 bit virtual addresses
  - 40 bit physical addresses (1 TB)
  - Binary compatible with earlier MIPS processor

R10000 Processors

- **Superscalar -- up to 4 instructions per cycle**
  - **up to 2 floating point inst per cycle**
  - **up to 2 integer inst per cycle**
  - **up to 1 load/stores per cycle**
- **5 independent pipelined functional units**
- **Out-of-order instruction execution**
- **64 entry TLB**

R10000 Processors

- **Peak Performance (at 195 Mhz)**
  - **390 Mflop/s: 2 floating point instructions per cycle**
  - **780 Mips: 2 integer + 2 floating point instructions per cycle**

Cache R(10k)

- **External (L2) Cache**
  - **secondary floating point cache**
  - **secondary instruction and integer data cache**
  - **4MB per processor**
  - **2-way set associative**
  - **cache line 128 bytes**
  - **write-back policy**
  - **cache clock rate is 2/3 that of CPU**

Cache (R10k)

- **External Cache**
  - **2 way interleaved -- even/odd doubleword (64 bits per doubleword)**
  - **1.06 GB/sec/bank cache-to-processor**
    - **1 doubleword per bank per cycle**
  - **2.13 GB/sec overall possible**
    - **2 doublewords per cycle**
  - **miss penalties**
    - **roughly 60 cycles to local memory**
    - **roughly 100 cycles + 20 cycles/router hop to remote memory**

Cache (R10k)

- **External Cache**
  - **coherence maintained in hardware**
  - **directory-based scheme**

Bandwidths

- **CrayLink Interconnect - 624MB/s bi-directional sustainable for user data**
- **I/O Buses -- 624MB/sec sustainable to XIO interface board**
  - **supports, Ultra Fast Wide SCSI, Fibrechannel, 100BASE-T, ATM, and HIPPI-Serial**
  - **each controller has ~600MB/sec bi-directional bandwidth**

IRIX 6

- 64-bit UNIX (IRIX 6.4+)
  - IRIX 6.1 -- full 64 bit file system
- standards compliant
  - System V Release 4 ABI
    - System V Interface Definition 3 (SVID3)
  - also MIPS ABI Group
- upwardly compatible revision of IRIX

IRIX 6

- **Process Scheduling**
  - **non-preemptive (only kernel preempts)**
  - **processor cache affinity heuristic**
  - **gang scheduling**
- **Lightweight Processes**
  - **sproc -- shared virtual address space**
  - **POSIX compliant threads**
    - **Fully supported in IRIX 6.4**

IRIX 6 -- High Performance I/O

- **Asynchronous I/O (Posix)**
  - **multiple concurrent I/O ops per process**
- **Memory Mapped I/O (SVID 3)**
  - **disk files visible in program address space**
  - **no explicit I/O ops required**
- **Direct I/O (SGI defined)**
  - **unbuffered transfers to/from device**
  - **performance approaches raw device interface**