# Origin Servers

## Technical Overview of the Origin Family

Table of Contents


Introduction
Origin2000 Components
What Makes Origin2000 Different
Scalability and Modularity
Systems Interconnections
Crossbar
Distributed Shared Address Space(Memory and I/O)
System Bandwidth

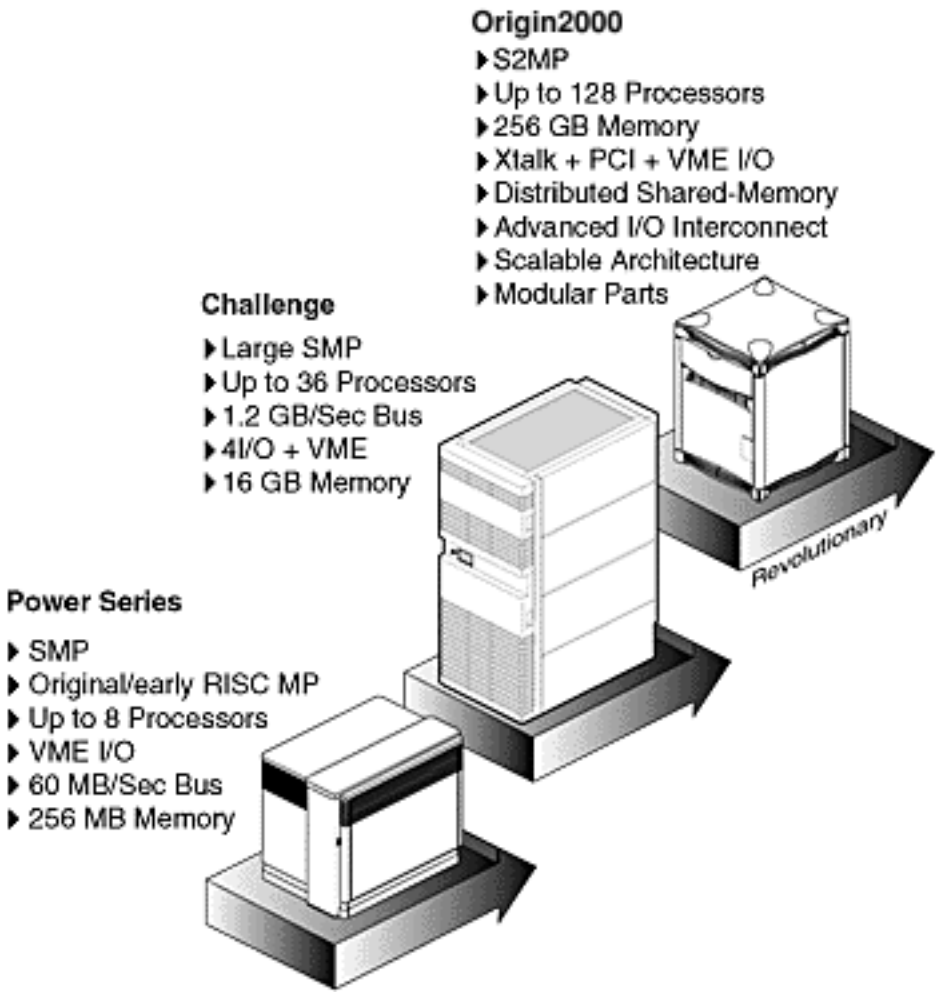

The guide describes the hardware architecture of the Origin family, and its specific implementations:

The entry-level Origin200 system, consisting of a maximum of two towers (up to four processors) that can be li together.

The deskside/rackmount Origin2000 system, consisting of from 1 to 128 processors, housed in deskside and ra cabinets.

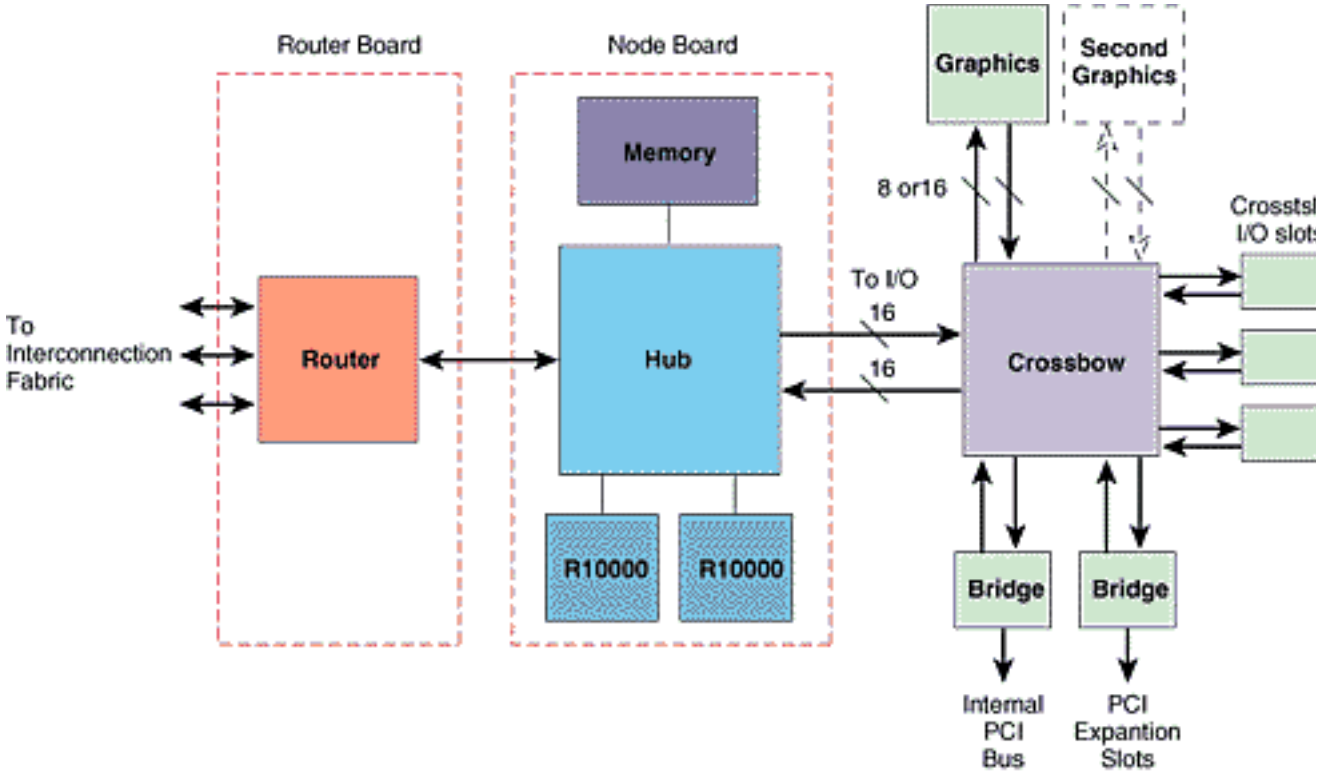
The Origin family is a revolutionary follow-on to the Challenge-class symmetric multiprocessing (SMP) system Silicon Graphics' distributed shared-memory multiprocessing architecture, called S2MP.

The development path of Silicon Graphics' multiprocessor systems is shown in Figure 1-1.



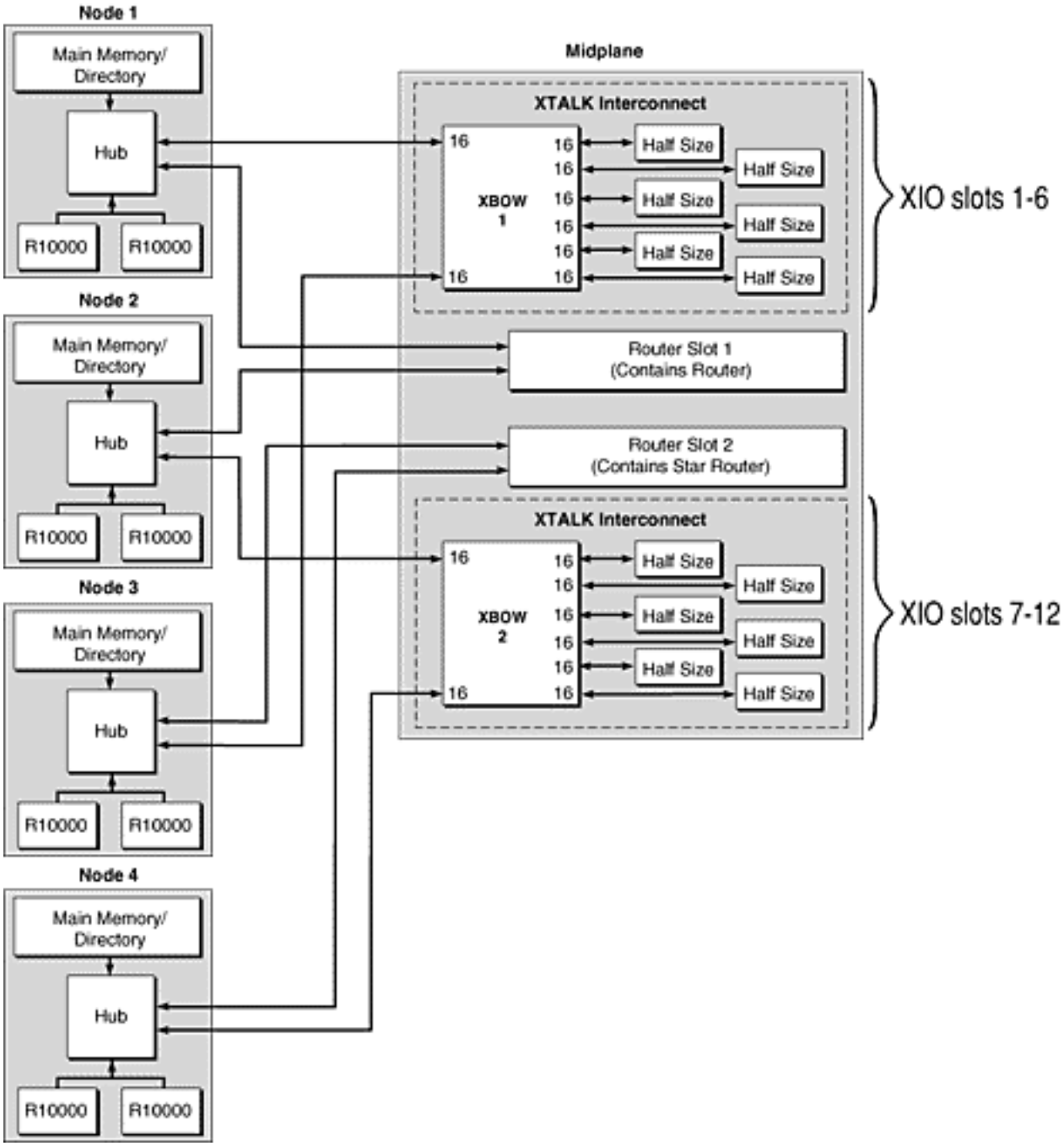

*Developmental Path of Silicon Graphics Multiprocessing Architectures*

Figure 1-2 is a block diagram of an Origin2000 system showing the central Node board, which can be viewed a system controller from which all other system components radiate.




*Origin2000 Block Diagram*

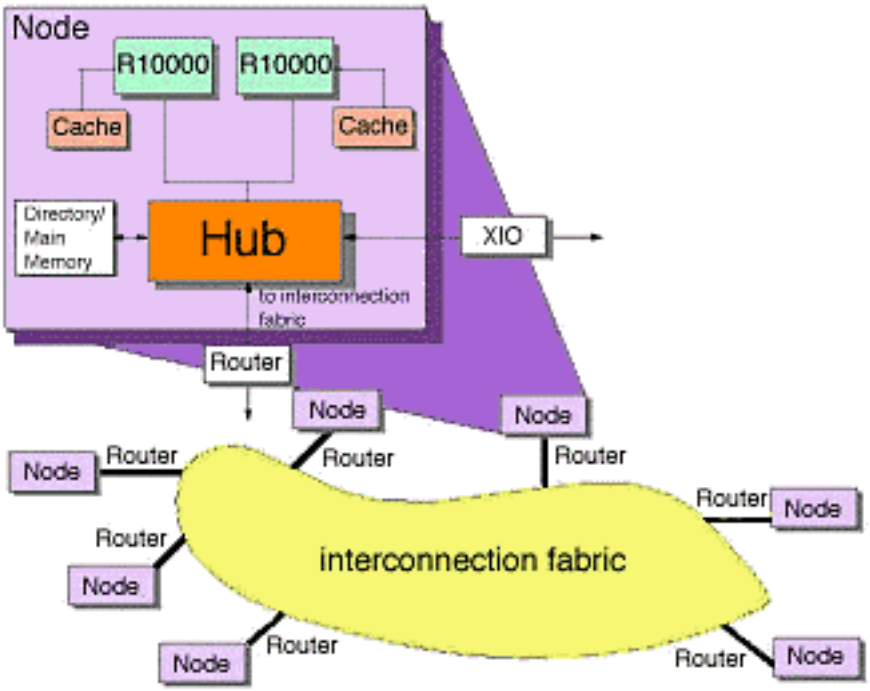
Figure 1-3 is a block diagram of a system with four Node boards; Nodes 1 and 3 connect to Crossbow 1, and N
and 4 connect to Crossbow 2. Crossbow 1 connects to XIO boards 1 through 6 and Crossbow 2 connects to XI
7 through 12.




*Block Diagram of a 4-Node System*

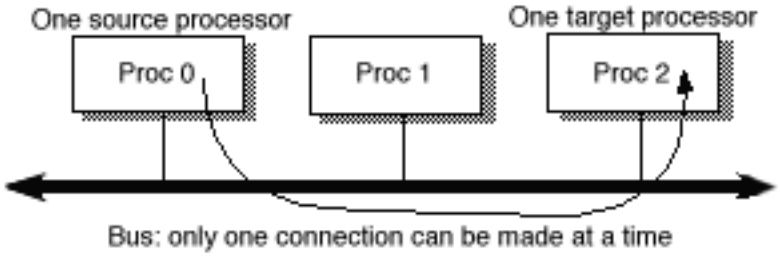
The components shown in Figure 1-2 are described further in this chapter, in the section titled ``Origin2000 Components.''

As illustrated in Figure 1-4, Origin2000 is a number of processing nodes linked together by an interconnection f Each processing node contains either one or two processors, a portion of shared memory, a directory for cache coherence, and two interfaces: one that connects to I/O devices and another that links system nodes through the interconnection fabric.



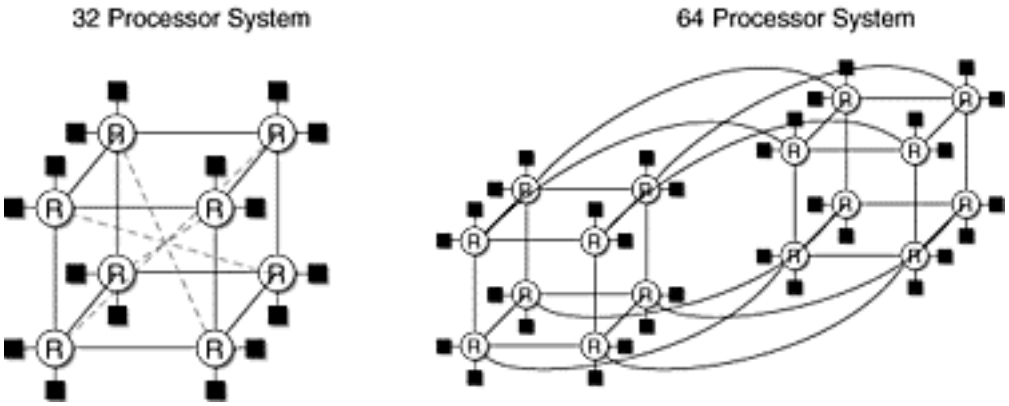

*Nodes in an Origin2000 System*

The interconnection fabric links nodes to each other, but it differs from a bus in several important ways. A bus i resource that can only be used by one processor at a time. The interconnection fabric is a mesh of multiple, simu dynamically-allocable -- that is, connections are made from processor to processor as they are needed -- transact This web of connections differs from a bus in the same way that multiple dimensions differ from a single dimen bus is a one-dimensional line, then the interconnection fabric is a multi-dimensional mesh.




*Single Datapath Over a Bus*

As shown in Figure 1-5, a bus is a shared, common link that multiprocessors must contest for and that only a si processor can use at a time. The interconnection fabric allows many nodes to communicate simultaneously, as s Figure 1-6. Paths through the interconnection fabric are constructed as they are needed by Router ASICs, which switches.




*Multi-dimensional Datapaths Through an Interconnection Fabric*

Origin2000 is said to be scalable, because it can range in size from 1 to 128 processors. As you add nodes, you and scale the system bandwidth. Origin2000 is also modular, in that it can be increased in size by adding standa modules to the interconnection fabric. The interconnection fabric is implemented on cables outside of these mod

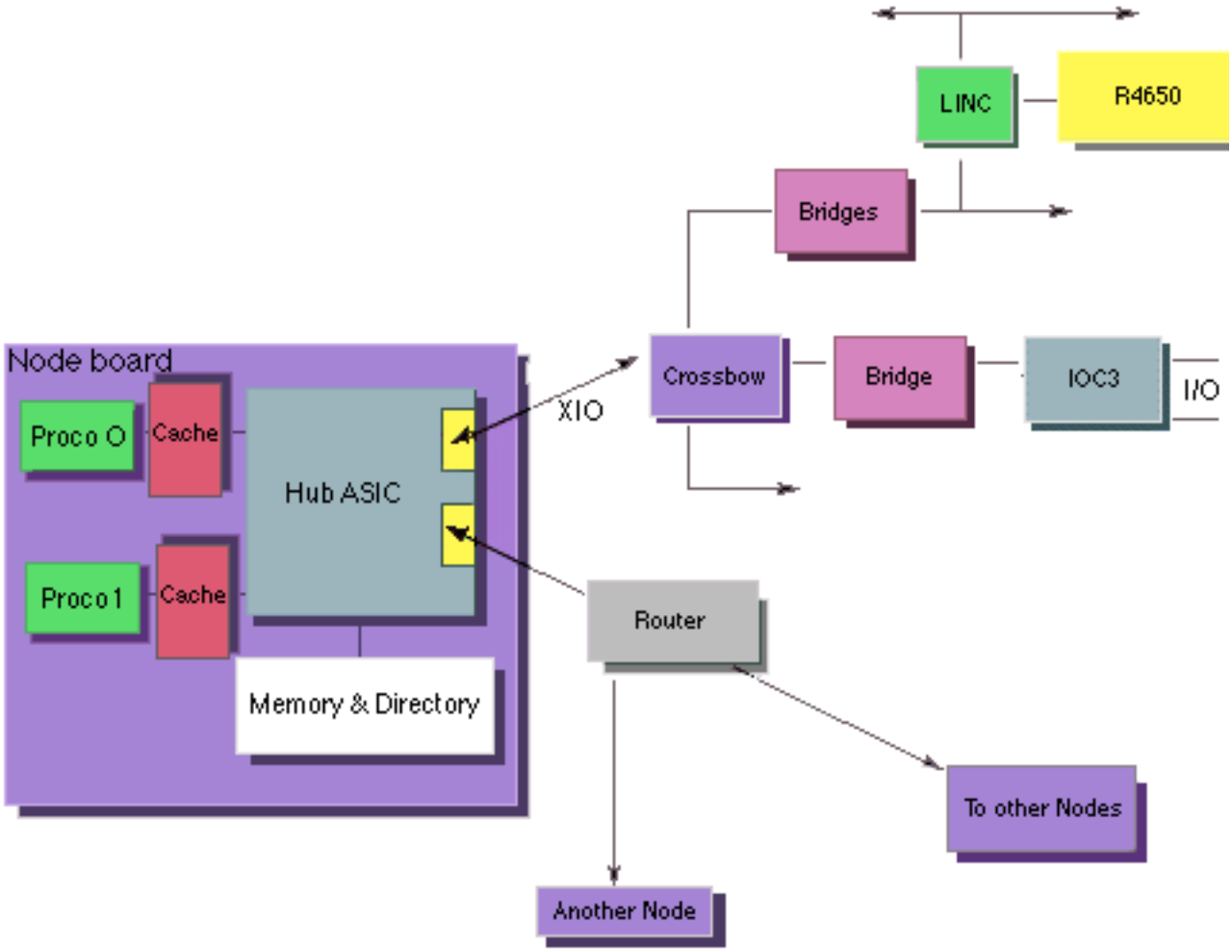
Origin2000 uses Silicon Graphics' new Scalable Shared-Memory Multiprocessor (S2MP) architecture to distribu shared memory amongst the nodes. This shared memory is accessible to all processors through the interconnecti and can be accessed with low latency.

## Origin2000 Components

An Origin2000 system has the following components:

- processor(s)
- memory
- I/O controllers
- distributed memory controller (Hub ASIC)
- directory memory for cache coherence
- CrayLink Interconnect
- XIO and Crossbow interfaces

These are linked as shown in Figure 1-7, and all of these are described in this section.




*Components in an Origin2000 System*

### Processor

Origin2000 uses the MIPS R10000, a high-performance 64-bit superscalar processor which supports dynamic scheduling. Some of the important attributes of the R10000 are its large memory address space, together with a for heavy overlapping of memory transactions -- up to twelve per processor in Origin2000.

### Memory

Each Node board added to Origin2000 is another independent bank of memory, and each bank is capable of sup up to 4 GB of memory. Up to 64 nodes can be configured in a system, which implies a maximum memory capa 256 GB.

### I/O Controllers

Origin2000 supports a number of high-speed I/O interfaces, including Fast, Wide SCSI, Fibrechannel, 100BAS ATM, and HIPPI-Serial. Internally, these controllers are added through XIO cards, which have an embedded P PCI-64 bus. Thus, in Origin2000 I/O performance is added one bus at a time.

### Hub

This ASIC is the distributed shared-memory controller. It is responsible for providing all of the processors and I devices a transparent access to all of distributed memory in a cache-coherent manner

### Directory Memory

This supplementary memory is controlled by the Hub. The directory keeps information about the cache status of memory within its node. This status information is used to provide scalable cache coherence, and to migrate data node that accesses it more frequently than the present node.

### CrayLink Interconnect

This is a collection of very high speed links and routers that is responsible for tying together the set of hubs that the system. The important attributes of CrayLink Interconnect are its low latency, scalable bandwidth, modularit fault tolerance.

### XIO and Crossbow

These are the internal I/O interfaces originating in each Hub and terminating on the targeted I/O controller. XIO u same physical link technology as CrayLink Interconnect, but uses a protocol optimized for I/O traffic. The Cross ASIC is a crossbar routing chip responsible for connecting two nodes to up to six I/O controllers.

## What Makes Origin2000 Different

The following characteristics make Origin2000 different from previous system architectures:

Origin2000 is scalable.

Origin2000 is modular.

Origin2000 uses an interconnection fabric to link system nodes and internal crossbars within the system (Hub, Router, Crossbow).

Origin2000 has distributed shared-memory and distributed shared-I/O.

Origin2000 shared memory is kept cache coherent using directories and a directory-based cache coherer protocol.

Origin2000 uses page migration and replication to improve memory latency.

**Scalability.** Origin2000 is easily scaled by linking nodes together over an interconnection fabric, and system bandwidth scales linearly with an increase in the number of processors and the associated switching fabric. This Origin2000 can have a low entry cost, since you can build a system upward from an inexpensive Origin200 configuration.

In contrast, traditional bus-based systems are only scalable in the amount of its processing and I/O power. The I interconnect is the E-bus, which has a fixed bandwidth and is the same size from entry-level to high-end.

**Modularity.** A system is comprised of standard processing nodes. Each node contains processor(s), memory, directory for cache coherence, an I/O interface, and a system interconnection. Node boards are placed in one of t following three types of system modules: entry-level, deskside, or rack.

Traditional bus-based systems are not as modular; there is a fixed number of slots in each deskside or rack syste this number cannot be changed.

**System interconnections.** Origin2000 uses an interconnection fabric and crossbars. The interconnection fab web of dynamically-allocated switch-connected links that attach nodes to one another. Crossbars are part of the interconnection fabric, and are located inside several of the ASICs--the Crossbow, the Router, and the Hub. Cr dynamically link ASIC input ports with their output ports.

In traditional bus-based systems, processors access memory and I/O interfaces over a shared system bus (E-bus a fixed size and a fixed bandwidth.

**Distributed shared-memory (DSM) and I/O.** Origin2000 memory is physically dispersed throughout the for faster processor access. Page migration hardware moves data into memory closer to a processor that frequen it. This page migration scheme reduces memory latency -- the time it takes to retrieve data from memory. Althou memory is distributed, it is universally accessible and shared between all the processors in the system. Similarly devices are distributed among the nodes, and each device is accessible to every processor in the system.

Traditional bus-based systems have shared memory, but their memory is concentrated, not distributed, and they distribute I/O. All I/O accesses, and those memory accesses not satisfied by the cache, incur extra latencies whe traversing the E-bus.

**Directory-based cache coherence.** Origin2000 uses caches to reduce memory latency. Cache coherence is supported by a hardware directory that is distributed among the nodes along with main memory. Cache coheren applied across the entire system and all memory. In a snoopy protocol, every cache-line invalidation must be br all CPUs in the system, whether the CPU has a copy of the cache line or not. In contrast, a directory protocol re point-to-point messages that are only sent those CPUs actually using the cache line. This removes the scalability problems inherent in the snoopy coherence scheme used by bus-based systems such as Everest. A directory-bas protocol is preferable to snooping since it reduces the amount of coherence traffic that must be sent throughout t system.

Traditional bus-based systems use a snoopy coherence protocol.

**Page migration and replication.** To provide better performance by reducing the amount of remote memory Origin2000 uses a process called page migration. Page migration moves data that is often used by a processor in memory close to that processor.

Traditional bus-based systems do not support page migration.

## Scalability and Modularity

Origin2000 scalability and modularity allow one to start with a small system and incrementally add modules to n system as large as needed. An entry-level Origin200 module can hold from one to four MIPS R10000 processo Origin2000 deskside module can hold from one to eight R10000 processors. A series of these deskside module mounted in racks, scaling the system up to the following maximum configuration:

- 128 processors out the system for faster processor access. Page migration hardware moves data into me closer to a processor that frequently uses it. This page migration scheme reduces memory latency -- the takes to retrieve data from memory. Although main memory is distributed, it is universally accessible a between all the processors in the system. Similarly, I/O devices are distributed among the nodes, and ea is accessible to every processor in the system.

- 256 GB of memory
- 64 I/O interfaces with 192 I/O controllers (or 184 XIO and 24 PCI-64)
- 128 3.5-inch Ultra-SCSI devices and 16 6.25-inch devices

As one adds nodes to the interconnection fabric, bandwidth and performance scale linearly without significantly impacting system latencies. This is a result of the following design decisions:

- replacing the fixed-size, fixed-bandwidth bus of traditonal bus-based systems with the scalable intercon fabric whose bisection bandwidth (the bandwidth through the center of CrayLink Interconnect) scales li with the number of nodes in the system
- reducing system latencies by replacing the centrally-located main memory of traditional bus based syste the tightly-integrated but distributed shared-memory S2MP architecture of Origin2000.

## System Interconnections

Origin2000 replaces traditional bus-based system's shared, fixed-bandwidth bus with the following:

- a scalable interconnection fabric, in which processing nodes are linked by a set of routers
- crossbar switches, which implement the interconnection fabric. Crossbars are located in the following p
    - within the Crossbow ASIC connecting the I/O interfaces to the nodes
    - within the Router ASIC forming the interconnection fabric itself,
    - within the Hub ASIC which interconnects the processors, memory, I/O, and interconnection fa interfaces within each node.

These internal crossbars maximize the throughput of the major system components and concurrent operations.

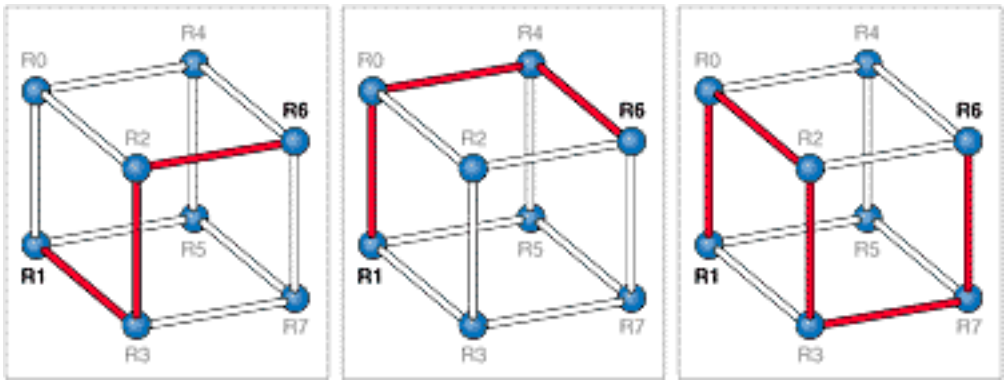
### Interconnection Fabric

Origin2000 nodes are connected by an interconnection fabric. The interconnection fabric is a set of switches, cal routers, that are linked by cables in various configurations, or topologies. The interconnection fabric differs fror standard bus in the following important ways:

- The interconnection fabric is a mesh of multiple point-to-point links connected by the routing switches. links and switches allow multiple transactions to occur simultaneously.
- The links permit extremely fast switching. Each bidirectional link sustains as much bandwidth as the en Everest bus.
- The interconnection fabric does not require arbitration nor is it as limited by contention, while a bus mu contested for through arbitration.
- More routers and links are added as nodes are added, increasing the interconnection fabric's bandwidth shared bus has a fixed bandwidth that is not scalable.
- The topology of the CrayLink Interconnect is such that the bisection bandwidth grows linearly with the of nodes in the system.

The interconnection fabric provides a minimum of two separate paths to every pair of Origin2000 nodes. This redundancy allows the system to bypass failing routers or broken interconnection fabric links. Each fabric link i additionally protected by a CRC code and a link-level protocol, which retry any corrupted transmissions and pro tolerance for transient errors.

Earlier in this chapter, Figure 1-5 and Figure 1-6 showed how an interconnection fabric differs from an ordinary bus. Figure 1-8 amplifies this difference by illustrating an 8-node hypercube with its multiple datapaths. Simulta R1 can communicate with R0, R2 to R3, R4 to R6, and R5 to R7, all without having to interface with any other

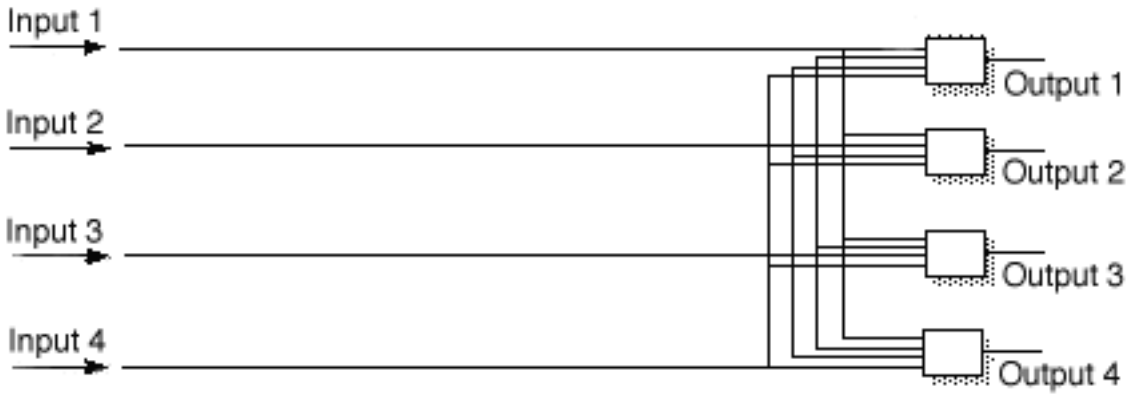



*Datapaths in an Interconnection Fabric*

## Crossbar

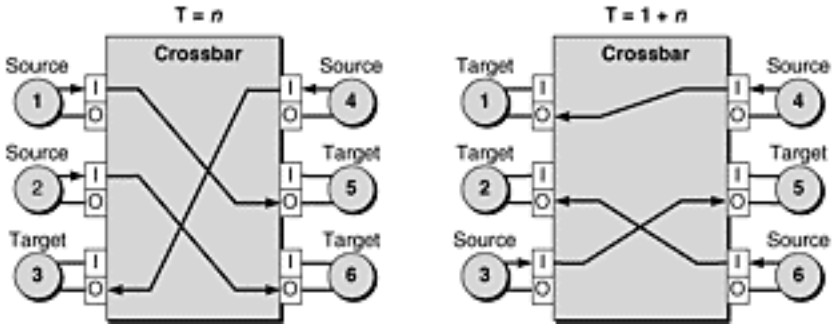
Several of the ASICs (Hub, Router, and Crossbow) use a crossbar for linking on-chip inputs with on-chip outp interfaces. For instance, an 8-way crossbar is used on the Crossbow ASIC; this crossbar creates direct point-to- links between one or more nodes and multiple I/O devices. The crossbar switch also allows peer-to-peer commu in which one I/O device can speak directly to another I/O device.

The Router ASIC uses a similar 6-way crossbar to link its six ports with the interconnection fabric, and the Hub links its four interfaces with a crossbar. A logical diagram of a 4-way (also referred to as four-by-four, or 4 x 4) is given in Figure 1-9; note that each output is determined by multiplexing the four inputs.




*Logical illustration of a Four-by-Four (4 x 4) Crossbar*

Figure 1-10 shows a 6-way crossbar at work. In this example, the crossbar connects six ports, and each port ha input (I) and an output (O) buffer for flow control. Since there must be an output for every input, the six ports c connected as six independent, parallel paths. The crossbar connections are shown at two clock intervals: Time= Time=n+1.




*Crossbar Operation*

At clock T=n, the ports independently make the following parallel connections:

- from port 1 to port 5
- from port 2 to port 6
- from port 3 to port 4
- from port 4 to port 2
- from port 5 to port 3
- from port 6 to port 1

Figure 1-10 shows the source (Input) and target (Output) for each connection, and arrows indicate the direction When a connection is active, its source and target are not available for any other connection over the crossbar.

At the next clock, T=n+1, the ports independently reconfigure themselves into six new data links: 1-to-5, 2-to-4 4-to-1, 5-to-3 and 6-to-2. At clock intervals, the ports continue making new connections as needed. Connection decisions are based on algorithms that take into account flow control, routing, and arbitration.

## Distributed Shared Address Space (Memory and I/O)

Origin2000 memory is located in a single shared address space. Memory within this space is distributed amongs processors, and is accessible over the interconnection fabric. This differs from a traditional bus-based system, i memory is centrally located on and only accessible over a single shared bus. By distributing Origin2000's mem among processors memory latency is reduced: accessing memory near to a processor takes less time than access remote memory. Although physically distributed, main memory is available to all processors.

I/O devices are also distributed within a shared address space; every I/O device is universally accessible through system.

### Origin2000 Memory Hierarchy

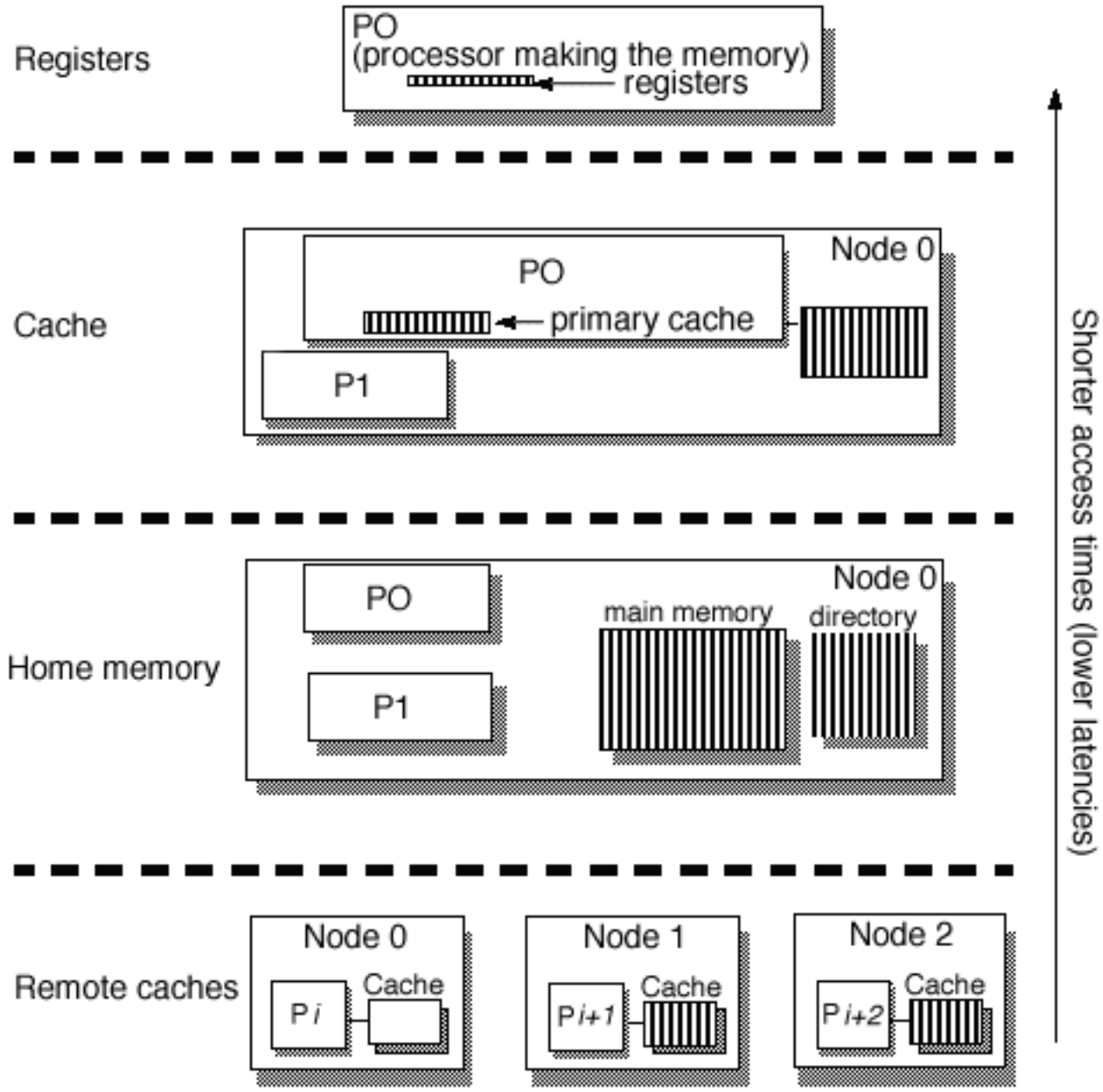
Memory in Origin2000 is organized into the following hierarchy:

- At the top, and closest to the processor making the memory request, are the processor registers. Since t physically on the chip they have the lowest latency -- that is, they have the fastest access times. In Figu these are on the processor labelled P0.

- The next level of memory hierarchy is labelled **cache**. In Figure 1-11, these are the primary and second caches located on P0. Aside from the registers, caches have the lowest latency in Origin2000, since they on the R10000 chip (primary cache) or tightly-coupled to its processor on a daughterboard (secondary c

- The next level of memory hierarchy is called **home memory**, which can be either local or remote. The **local** if the address of the memory reference is to address space on the same node as the processor. The is **remote** if the address of the memory reference is to address space on another node. In Figure 1-11, memory is the block of main memory on Node 0, which means it is local to Processor 0.

- The next level of memory hierarchy consists of the **remote caches** that may be holding copies of a giv memory block. If the requesting processor is writing, these copies must be invalidated. If the processor reading, this level exists if another processor has the most up-to-date copy of the requested location. In 1-11, remote cache is represented by the blocks labelled ``cache" on Nodes 1 and 2.



*Memory Hierarchy, Based on Relative Latencies and Data Capacities*

Caches are used to reduce the amount of time it takes to access memory -- also known as a memory's **latency** - moving faster memory physically close to, or even onto, the processor. This faster memory is generally some v static RAM, or SRAM.

The DSM structure of Origin2000 also creates the notion of local memory. This memory is close to the processc reduced latency compared to bus-based systems, where all memory must be accessed through a shared bus.

While data only exists in either local or remote memory, copies of the data can exist in various processor caches these copies consistent is the responsibility of the logic of the various hubs. This logic is collectively referred to *cache-coherence protocol.*

## System Bandwidth

There are three types of bandwidths:

- Peak bandwidth, which is a theoretical number derived by multiplying the clock rate at the interface by t width of the interface.
- Sustained bandwidth, which is derived by subtracting the packet header and any other immediate overh the peak bandwidth. This best-case figure, sometimes called Peak Payload bandwidth, does not take int contention and other variable effects.
- Bisection bandwidth, which derived by dividing the interconnection fabric in half, and measuring the da across this divide. This figure is useful for measuring data rates when the data is not optimally placed.

Table 1-1 gives a comparison between peak and sustained data bandwidths of Origin2000.

**Table 1-1** ***Peak and Sustained Bandwidths***

| Interface | Sustained Bandwidth [Peak BW in brackets] |
|---|---|
| Memory | 780 MB per second [780] |
| Node Card | 1.25 GB per second [1.56 GB] |
| Crossbow | 2.5 GB per second [3.12 GB] |
| Module (deskside) | 5.0 GB per second [6.24 GB] |
| Rack | 80 GB per second [100 GB] |

Table 1-2 lists the bisection bandwidths of various Origin2000 configurations both with and without Express Li

**Table 1-2** ***System Bisection Bandwidths***

| System Size (number of CPUs) | Sustained Bisection Bandwidth without Express Links [Peak BW] | Sustained Bisection Bandwidth with Express Links [Peak BW] |
|---|---|---|
| 8 | 1.25 GB per second [1.56 GB] | 2.5 GB per second [3.12 GB]* |
| 16 | 2.5 GB per second [3.12 GB] | 5.0 GB per second [6.24 GB] |

| | | |
|---|---|---|
| 32 | 5.0 GB per second [6.24 GB] | 10 GB per second [12.5 GB] |
| 64 | 10 GB per second [12.5 GB] | N/A |
| 128 | 20 GB per second [25.0] | N/A |

* With Star Router

Table 1-3 lists the bandwidths of Ultra SCSI and FibreChannel devices.

**Table 1-3** ***Peripheral Bandwidths***

| Peripheral | Bandwidth |
|---|---|
| Ultra SCSI | 40 MB per second |
| FibreChannel | 100MB per second |

Copyright © 1996 Silicon Graphics, Inc. All Rights Reserved. Trademark Information webmaster@www.sgi.com.