CRAY® T3D System Architecture Overview

HR-04033

Cray Research, Inc.

The work represented by this manual was partially supported by the Defense Advanced Research Projects Agency under agreement No. MDA972–92-H-0002, dated 21 January 1992.

Requests for copies of Cray Research, Inc. publications should be directed to:

CRAY RESEARCH, INC.
Distribution
2360 Pilot Knob Road
Mendota Heights, MN 55120
800-284-2729 extension 35907

Comments about this publication should be directed to:

CRAY RESEARCH, INC.
Hardware Publications and Training
890 Industrial Blvd.
Chippewa Falls, WI 54729

Record of Revision

Each time this manual is updated with a change packet, a change to part of a text page is indicated by a change bar in the margin directly opposite the change. A change bar in the footer of a text page indicates that most, if not all, of the text is new. A change bar in the footer of a page composed primarily of a table and/or figure may indicate that a change was made to that table/figure or, it could indicate that the entire table/figure is new. Change packets are assigned a numerical designator, which is indicated in the publication number on each page of the change packet.

Each time this manual is fully revised and reprinted, all change packets to the previous version are incorporated into the new version, and the new version is assigned an alphabetical revision level, which is indicated in the publication number on each page of the manual. A revised manual does not usually contain change bars.

REVISION	DESCRIPTION
	September 1993. Original Printing.

PREFACE

The *CRAY T3D System Architecture Overview* manual provides a basic description of the components of the CRAY T3D system. The manual is designed for people who are not familiar with the terminology associated with the CRAY T3D system. The manual is broken down into four sections:

"Overview" introduces and briefly describes the interconnect network, processing element nodes, and I/O gateways that comprise the CRAY T3D system.

"Interconnect Network" describes the characteristics and components of the interconnect network.

"Processing Element Nodes" describes the components and functions of a processing element node.

"I/O Gateways" describes the components, functions, and different types of I/O gateways. Included are descriptions of the low-speed (LOSP) and high-speed (HISP) channels.

If you have any comments or suggestions related to this manual, please fill out a reader comment form and send it back to Hardware Publications and Training. Reader comment forms are located in the front and back of this manual.

CONTENTS

1 OVERVIEW

2 INTERCONNECT NETWORK

3 PROCESSING ELEMENT NODES

3 PROCESSING ELEMENT NODES (continued)

4 I/O GATEWAYS

FIGURES

FIGURES (continued)

TABLES

GLOSSARY

1 OVERVIEW

Cray Research is implementing a three-phase massively parallel processor (MPP) program. Our goal is to reach a sustained performance on real customer code of one trillion floating-point operations per second. This manual describes the basic architecture of the first-phase MPP system, the CRAY T3D system.

The CRAY T3D system contains hundreds or thousands of microprocessors, each accompanied by a local memory. The system is designed to support different styles of MPP programming, such as data parallel, work-sharing, and message passing.

The CRAY T3D system connects to a host computer system. The host system provides support for applications running on the CRAY T3D system. All applications written for the CRAY T3D system are compiled on the host system but run on the CRAY T3D system.

The host system may be any Cray Research computer system that has an input/output subsystem model E (IOS-E). Host systems include the CRAY Y-MP E series computer systems, the CRAY Y-MP M90 series computer systems, and the CRAY C90 series computer systems.

The host system may reside in the same cabinet as the CRAY T3D system. This configuration is called a single-cabinet configuration. The host system may also reside in a separate cabinet that is cabled to the CRAY T3D system cabinet. This configuration is called a multiple-cabinet configuration.

The CRAY T3D system contains four types of components: processing element nodes, the interconnect network, I/O gateways, and a clock. Figure 1-1 shows a simplified model of the components of the CRAY T3D system.

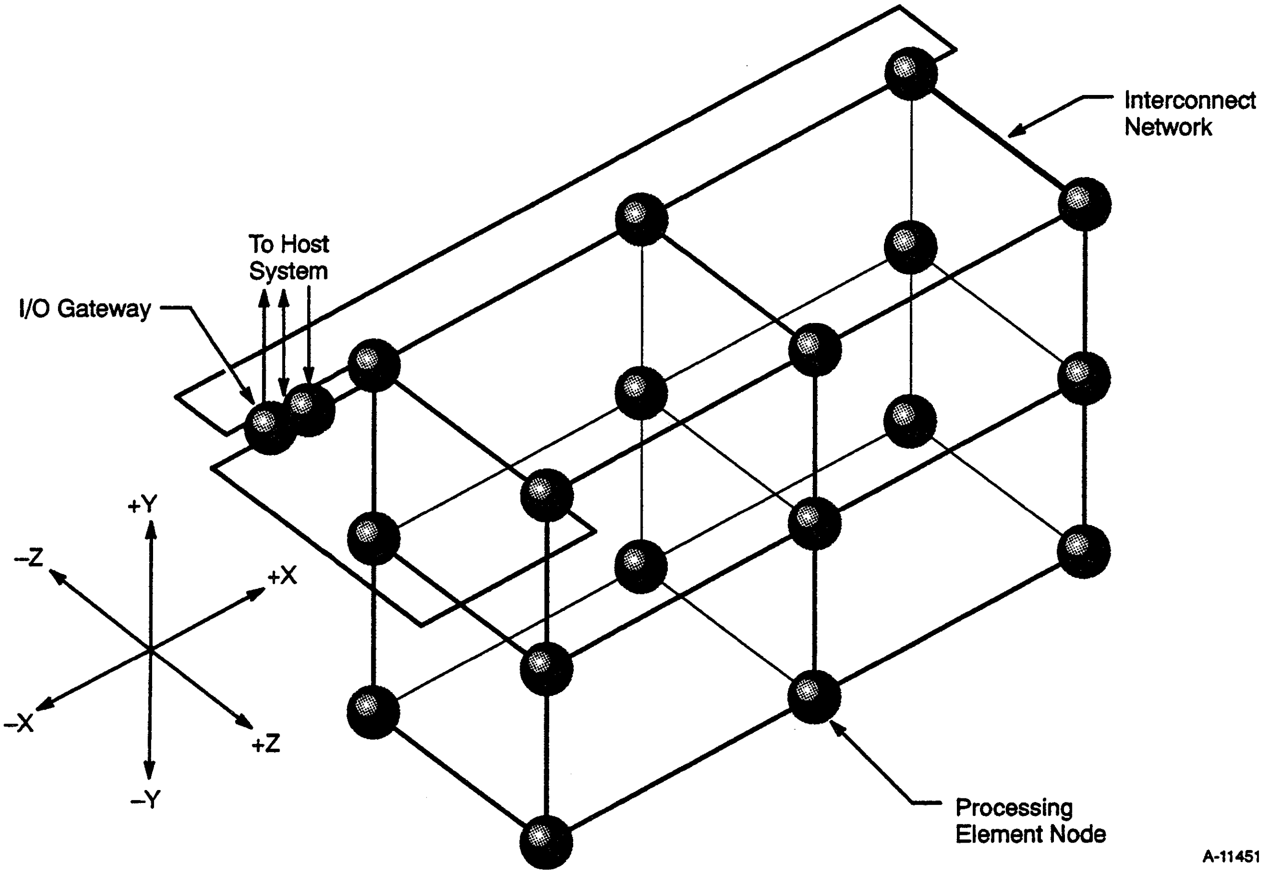


Figure 1-1. CRAY T3D System Components

The following subsections describe each component of the CRAY T3D system.

Processing Element Nodes

An MPP computer system contains hundreds or thousands of microprocessors, each accompanied by a local memory. Each microprocessor and local memory component is called a processing element (PE). In the CRAY T3D system, each PE contains a microprocessor, local memory, and support circuitry (refer to Figure 1-2). There are two PEs per processing element node.

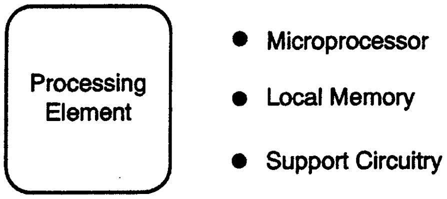


Figure 1-2. Processing Element Components

The microprocessor is a reduced instruction set computer (RISC) 64-bit microprocessor developed by Digital Equipment Corporation. The microprocessor performs arithmetic and logical operations on 64-bit integer and 64-bit floating-point registers [operations include the Institute of Electrical and Electronic Engineers (IEEE) floating point arithmetic]. The microprocessor also contains an internal instruction cache memory and data cache memory that each store 256 lines of data or instructions. Each line in the instruction and data cache memory is four 64-bit words wide.

Local memory consists of dynamic random access memory (DRAM) that stores system data. A low-latency, high-bandwidth data path connects the microprocessor to local memory in a PE. The size of local memory is 2 Mwords using 4-Mbit DRAM integrated circuits or 8 Mwords using 16-Mbit DRAM integrated circuits.

The local memory within each PE is part of a physically distributed, logically shared memory system. System memory is physically distributed because each PE contains local memory. System memory is logically shared because the microprocessor in one PE can access the memory of another PE without involving the microprocessor in that PE.

The support circuitry extends the control and addressing functions of the microprocessor. This includes performing data transfers to or from local memory.

A CRAY T3D system contains 32; 64; 128; 256; 512; 1,024; or 2,048 PEs, depending on the system configuration (excluding the PEs in the I/O gateways). The PEs reside in processing element nodes.

Each processing element node contains two PEs, a network interface, and a block transfer engine (refer to Figure 1-3). The following paragraphs briefly describe each of these components.

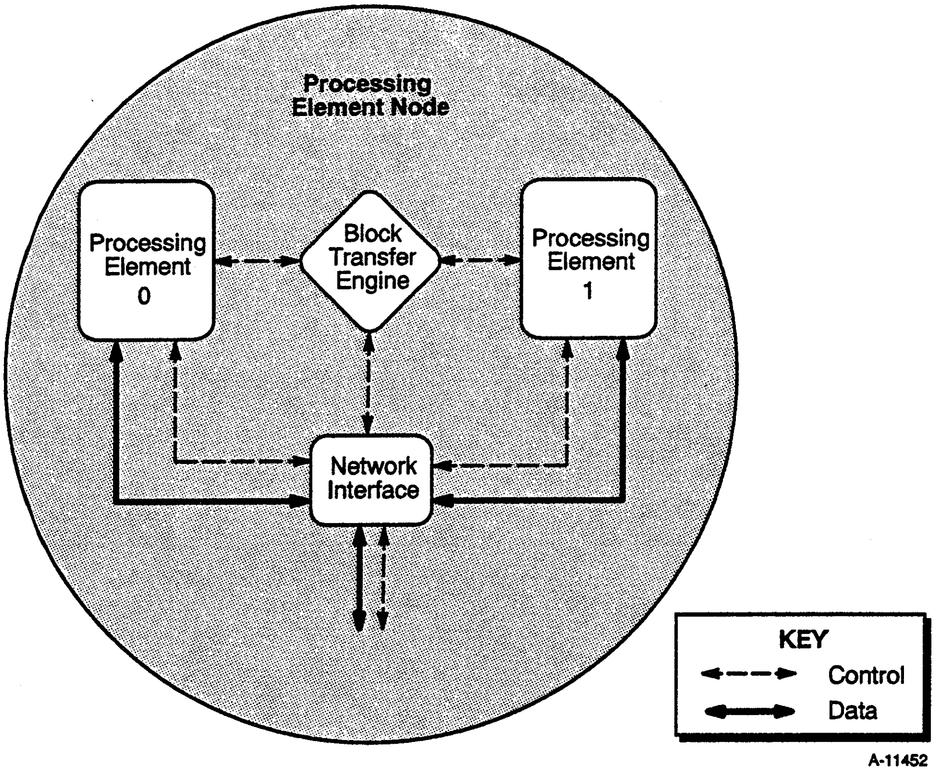


Figure 1-3. Processing Element Node

The two PEs in a processing element node are identical but function independently. Access to the block transfer engine and network interface is shared by the two PEs.

The network interface formats information before it is sent over the interconnect network to another processing element node or I/O gateway. The network interface also receives incoming information from another processing element node or I/O gateway and steers the information to PE 0 or PE 1 in the processing element node.

The block transfer engine (BLT) is an asynchronous direct memory access controller that redistributes system data. The BLT redistributes system data between the local memory in PE 0 or PE 1 and globally addressable system memory. The BLT can redistribute up to 65,536 64-bit words of data (or 65,536 4-word lines of data) without interruption from the PE.

Interconnect Network

The interconnect network provides communication paths among the processing element nodes and the I/O gateways in the CRAY T3D system. The interconnect network forms a three-dimensional matrix of paths that connect the nodes in the X, Y, and Z dimensions (refer again to Figure 1-1).

The interconnect network is composed of communication links and network routers. Figure 1-4 shows how the components of the interconnect network connect to a processing element node.

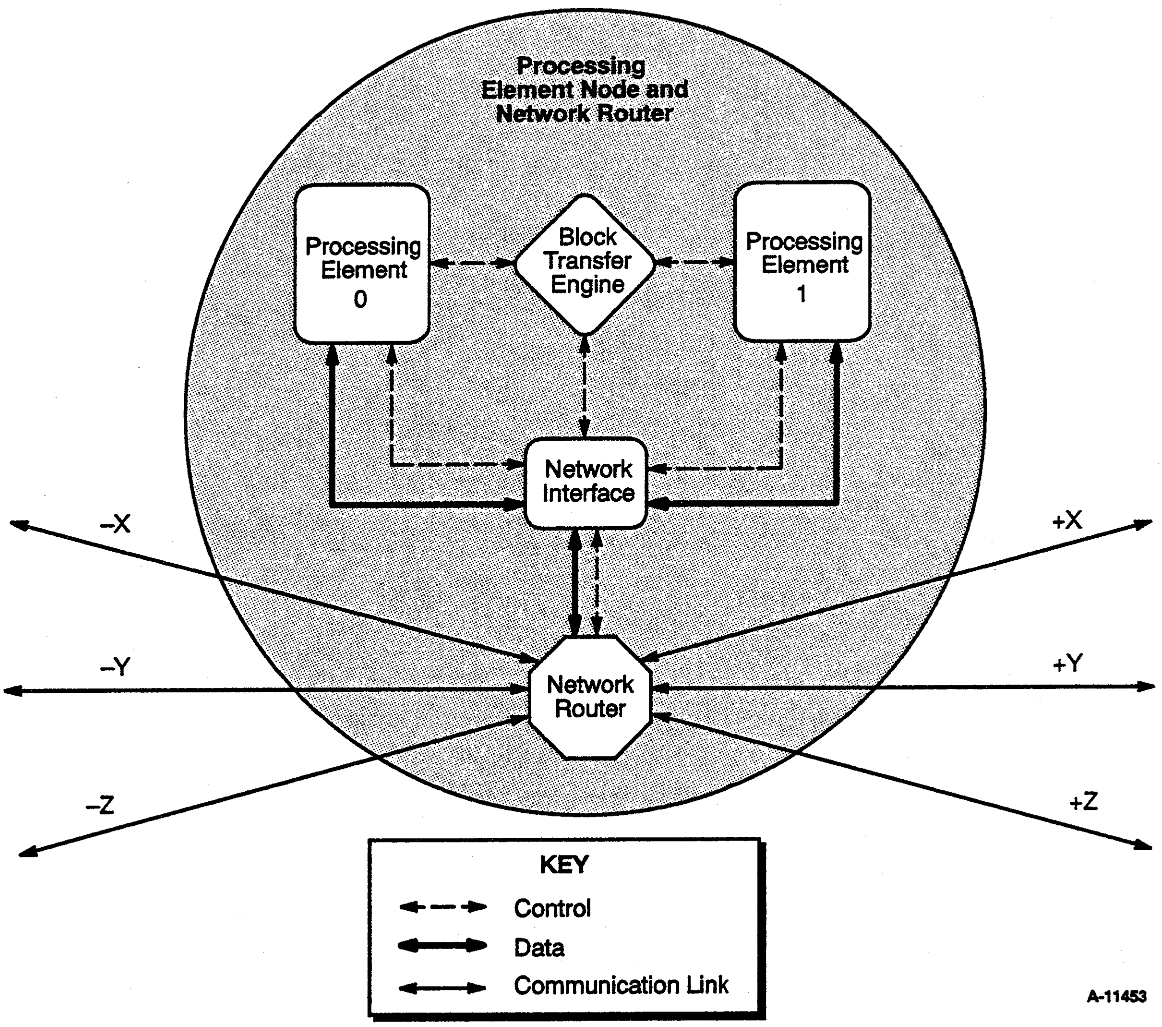


Figure 1-4. Interconnect Network Components

I/O Gateways

I/O gateways transfer system data and control information between the host system and the CRAY T3D system or between the CRAY T3D system and an input/output cluster (IOC). The I/O gateways connect to the interconnect network through network routers that have communication links in the X and Z dimensions only. [The I/O gateways do not have connections in the Y dimension because the Y dimension connectors on an I/O gateway circuit board were replaced with low-speed (LOSP) and high-speed (HISP) channel connectors.] An I/O gateway can transfer information to any PE in the interconnect network.

An I/O gateway contains an input node, an output node, and LOSP circuitry. Figure 1-5 shows the components of an I/O gateway.

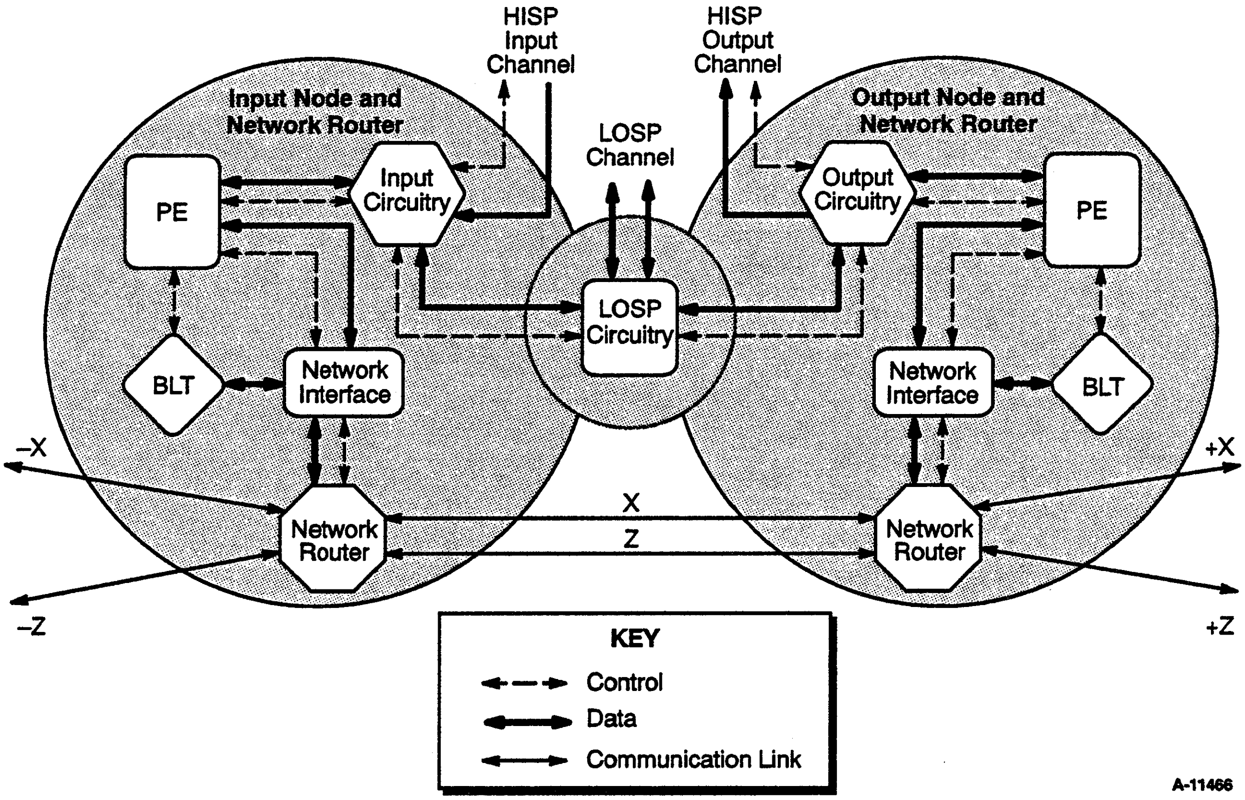


Figure 1-5. I/O Gateway

The input node contains one PE, a network interface, a BLT, and HISP input circuitry. The BLT and network interface in the input node are identical to the BLT and network interface used in the processing element node.

The PE in the input node is designed to interface with the HISP input circuitry. Because of this characteristic, the PE in the input node does not contain the circuitry to perform all of the operations that a PE in a processing element node performs. Instead, the circuitry is replaced with circuitry that interfaces with the HISP input circuitry. In addition, half of the local memory in the PE is replaced with HISP input circuitry that contains HISP channel buffers.

The HISP input circuitry receives incoming system data from the host system over the HISP channel. After receiving the data, the HISP input circuitry, PE, and BLT in the input node transfer the data to PEs in the CRAY T3D system.

Except for HISP output circuitry replacing the HISP input circuitry, the output node is identical to the input node. The HISP output circuitry transmits outgoing system data to the host system over the HISP channel. After the PE and BLT in the output node retrieve data from PEs in the CRAY T3D system, the HISP output circuitry transfers the data to the host system.

The LOSP circuitry transfers request and response information over the LOSP channel that connects the host system and the CRAY T3D system. LOSP request and response information is used to control the transfer of system data over the HISP channel.

There are two types of I/O gateways: a master I/O gateway and a slave I/O gateway. The two types of I/O gateway correspond to the two types of components connected by a HISP channel. The master I/O gateway is the master component of a HISP channel and sends the address information to the host system during a HISP transfer. The slave I/O gateway is the slave component of a HISP channel and receives the address information from the host system during a HISP transfer.

Clock

The CRAY T3D system contains a central clock that provides a 6.67-ns clock signal. The clock signal is fanned-out to all of the processing element nodes and I/O gateways in the system. The clock resides on one circuit board in the CRAY T3D system cabinet.

2 INTERCONNECT NETWORK

The interconnect network provides communication paths among the processing element (PE) nodes and the input/output gateways in the CRAY T3D system. The interconnect network forms a three-dimensional matrix of paths that connect the nodes in the X, Y, and Z dimensions (refer to Figure 2-1).

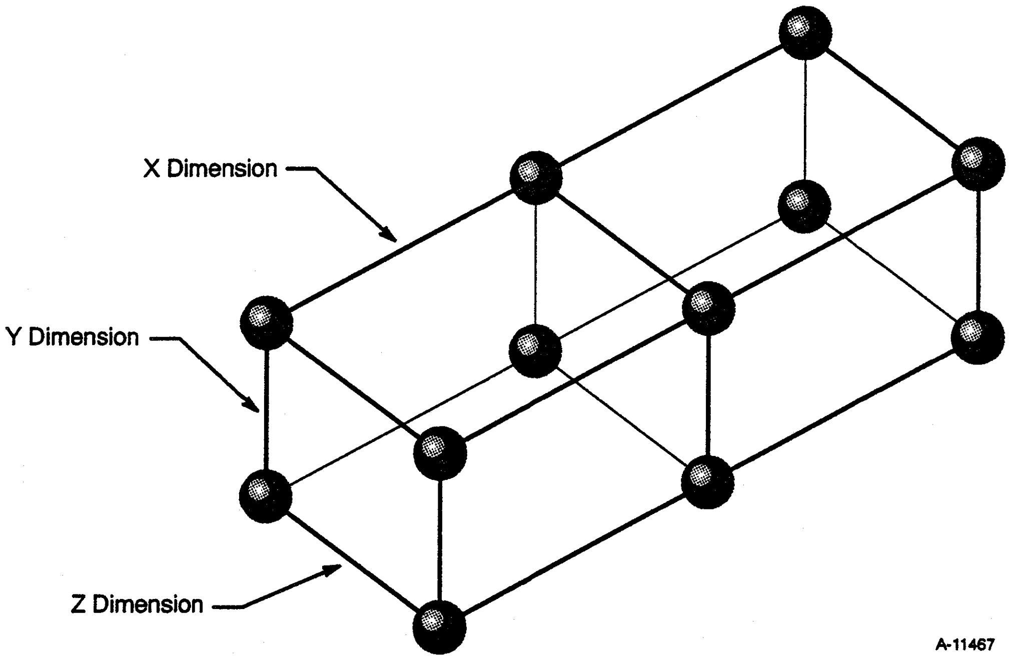


Figure 2-1. Three-dimensional Matrix of Nodes

This section describes the components and characteristics of the interconnect network. This information includes descriptions of the communication links, torus interconnect topology, interleaving, dimension order routing, virtual channels, packets, and network routers.

Communication Links

Communication links transfer data and control information between two network routers in the interconnect network. Each network router connects to a processing element node or an I/O gateway node. Each communication link connects two nodes in one dimension (refer to Figure 2-2).

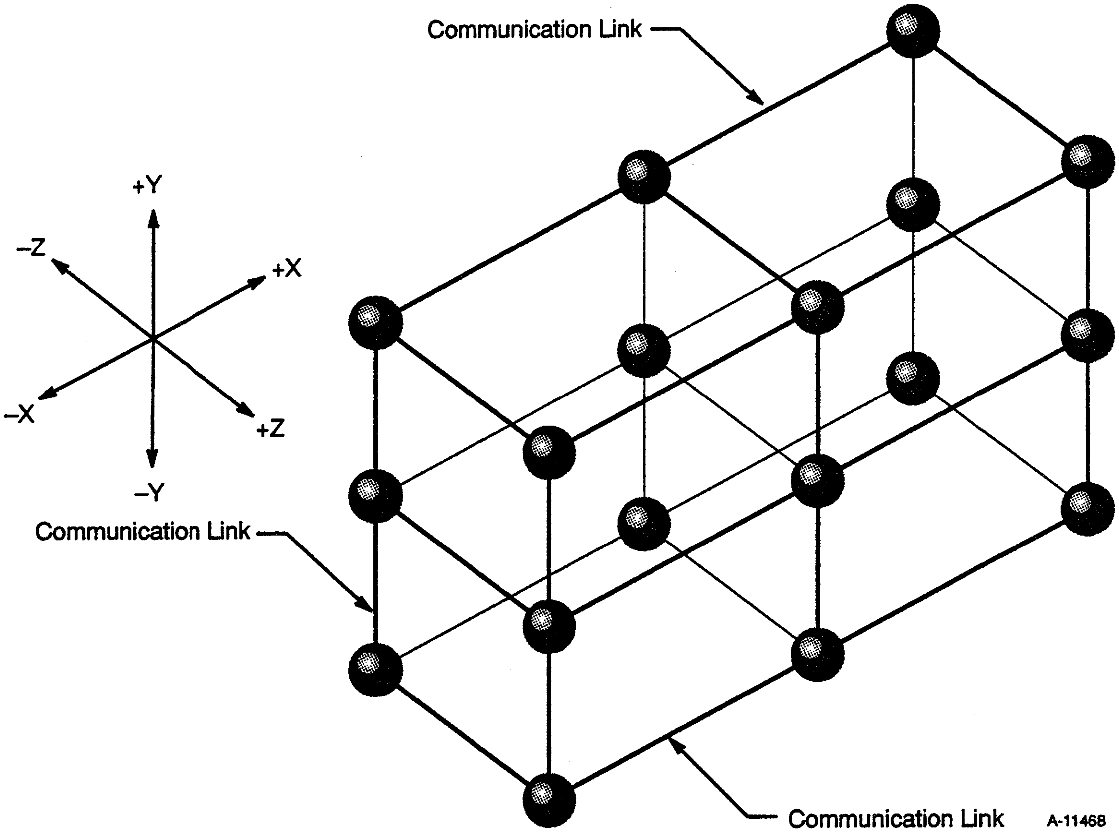


Figure 2-2. Communication Links

A communication link is actually two unidirectional channels. Each channel in the link contains data, control, and acknowledge signals. Figure 2-3 shows the signals for both channels in one communication link.

Data

Each channel contains 16 data signals. Data signals carry two types of information: requests and responses.

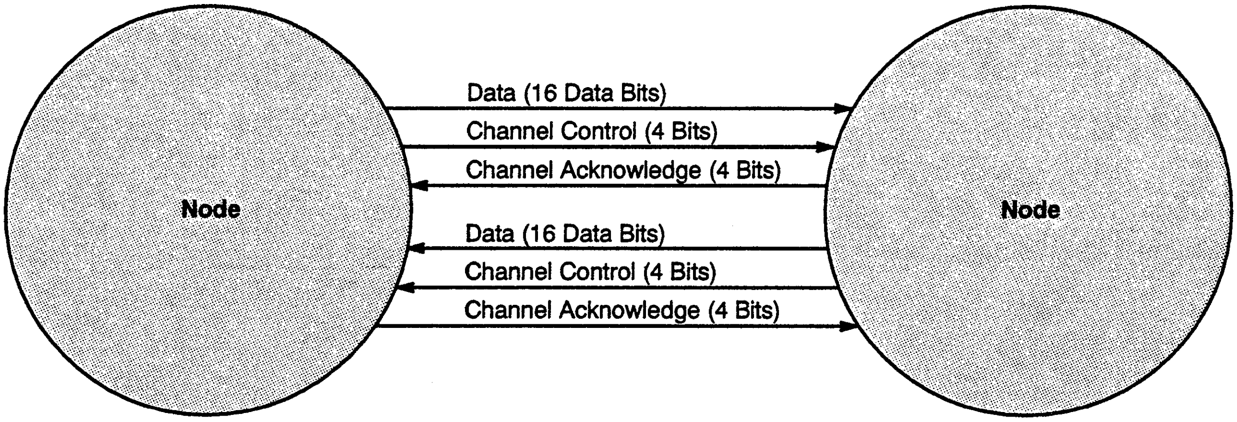


Figure 2-3. Communication Link Channel Signals

Requests contain information that requests a node to perform an activity. For example, a source node may send a request to a destination node to read data from memory in the destination node. This request is sent over one channel in the communication link.

Responses contain information that is the result of an activity. For example, after receiving a request for read data, a destination node sends the response back to the source node. The response contains the read data and is sent over the other channel in the communication link.

Requests and responses must be logically separated. This is done by providing separate buffers for requests and responses. The separate buffers are used for virtual channels. More information on virtual channels is provided later in this section.

Channel Control

The Channel Control signals are controlled by the node sending information over the link. The Channel Control signals are used to identify that the information is a response or a request and identify which virtual channel buffer in the receiving node will store the information.

Channel Acknowledge

The Channel Acknowledge signals are controlled by the node receiving information. The receiving node uses these signals to notify the sending node that the virtual channel buffers in the receiving node are empty.

Torus Interconnect Topology

The interconnect network is connected in a bidirectional torus. A torus contains communication links that connect the smallest numbered node in a dimension directly to the largest numbered node in the same dimension. This type of connection forms a ring where information can transfer from one node, through all of the nodes in the same dimension, and back to the original node.

Figure 2-4 shows a one-dimensional torus network in the X dimension. Information can transfer from node 00, through all of the nodes, and back to node 00 in a circular fashion. Each node has a communication link in both the positive and negative directions of the X dimension.

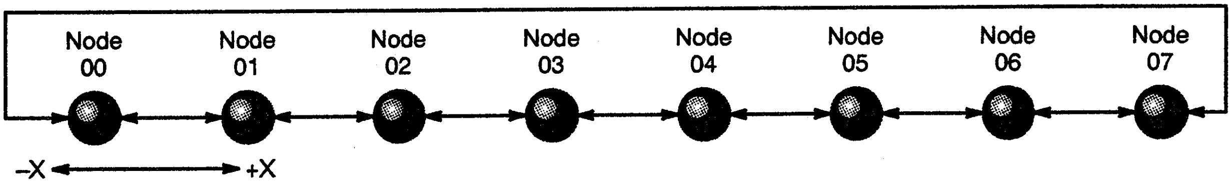


Figure 2-4. One-dimensional Torus Network

Torus networks offer several advantages for network communication. One advantage is speed of information transfers. For example, in Figure 2-4, node 07 can communicate directly with node 00 instead of sending information through all of the nodes in the X dimension.

Another advantage of the torus network is the ability to avoid bad communication links. For example, in Figure 2-4, if node 00 cannot transfer information directly to node 01 due to a bad communication link, node 00 can still communicate with node 01 by sending the information the long way around the network through the other nodes in the X dimension.

Figure 2-5 shows a two-dimensional torus network in the X and Y dimensions. Each node has communication links in both the positive and negative directions of the X and Y dimensions.

Figure 2-6 shows a three-dimensional torus network in the X, Y, and Z dimensions. Each node has communication links in both the positive and negative directions of the X, Y, and Z dimensions.

Several of the diagrams in this manual show three-dimensional network connections. For clarity, the communication link that completes the torus in each dimension is not shown. It is important to remember that although not shown in the diagrams, this communication link is always there.

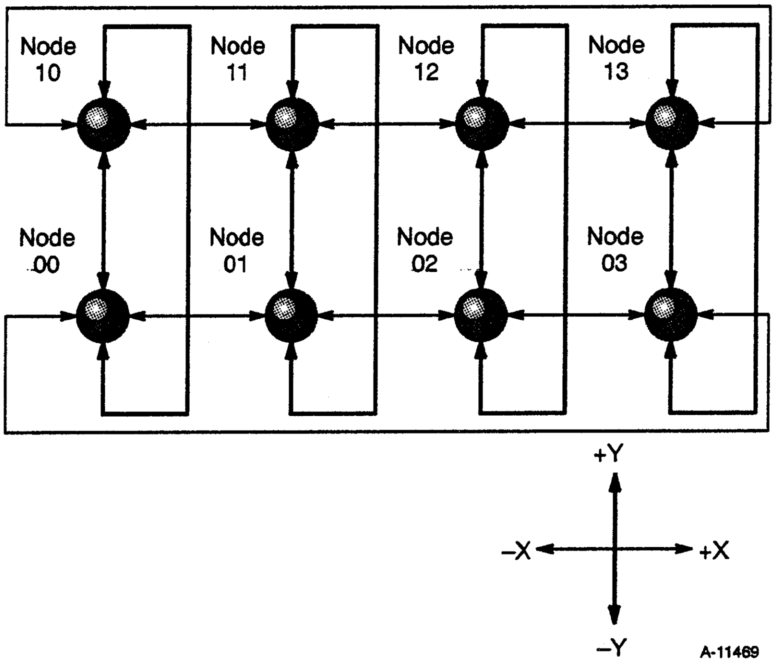


Figure 2-5. Two-dimensional Torus Network

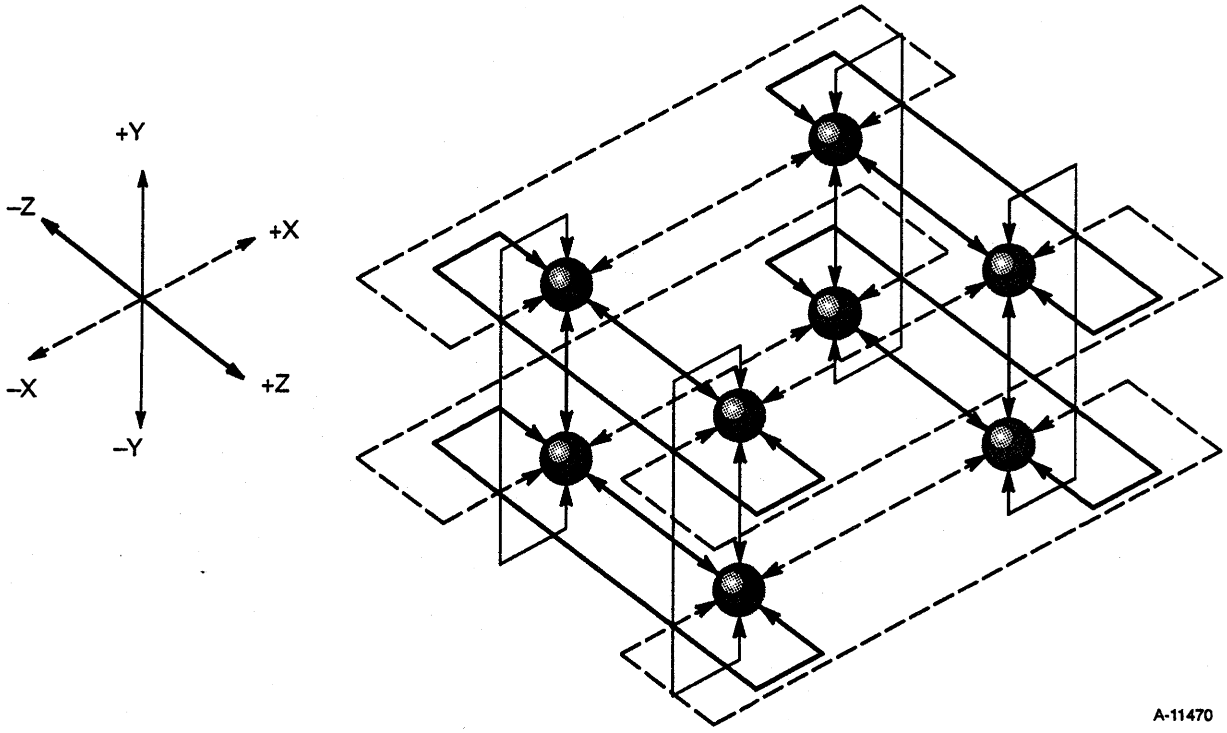


Figure 2-6. Three-dimensional Torus Network

Interleaving

The nodes in the interconnect network are interleaved. Interleaving is the physical placement of nodes so that the maximum wiring distance between nodes is minimized.

Figure 2-7 shows two one-dimensional torus networks. The eight nodes in the upper network are not interleaved. The eight nodes in the lower network are interleaved. In the interleaved network (also called a folded torus network), the physical length of the longest communication link is shorter than the physical length of the longest communication link in the non-interleaved network.

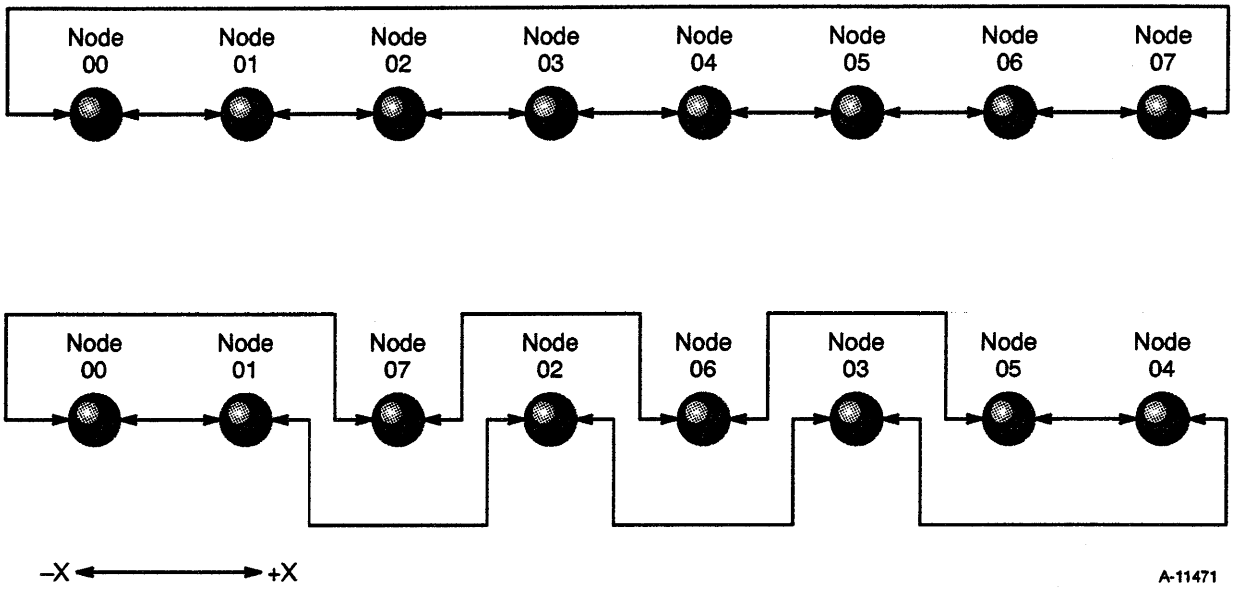


Figure 2-7. Interleaving

The X and Z dimensions of the network are interleaved. This minimizes the length of the physical communication links (wires) in the CRAY T3D system.

Several of the diagrams in this manual contain drawings of three-dimensional interconnect networks. For clarity, the communication links are shown logically and do not show the interleaving. It is important to remember that although not shown, the nodes in the network are physically interleaved.

Dimension Order Routing

When a node sends information to another node, the information may travel through several communication links in the network. Each transfer of information over a communication link is hereafter referred to as a hop.

After information leaves a node, it travels through the network in the X dimension first, then through the Y dimension, and finally through the Z dimension. When finished moving through the communication links in the Z dimension, the information arrives at the destination node. This method of information travel is called dimension order routing.

For example, if node A shown in Figure 2-8 sends request information to node B, the information first travels one hop in the +X direction. Because the information does not need to travel any farther in the X dimension, it switches direction to the Y dimension.

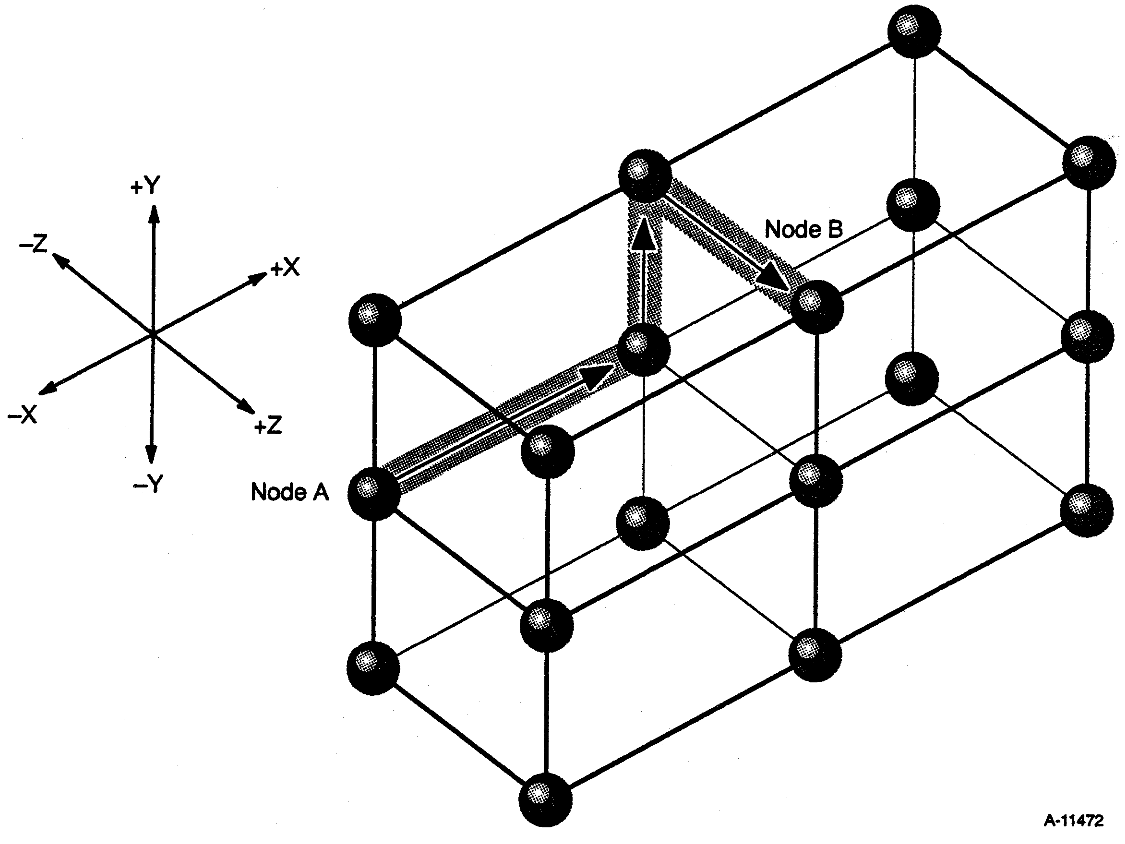


Figure 2-8. +X, +Y, and +Z Information Travel

After completing one hop in the +Y direction, the information switches direction to the Z dimension. After completing one hop in the +Z direction, the request information arrives at node B.

Information does not always travel in the positive direction of a dimension. For example, if node B in Figure 2-9 sends response information to node A, the information completes one hop in the –X direction and then changes direction into the Y dimension.

The information completes one hop in the –Y direction before changing direction into the Z dimension. After completing one hop in the –Z direction, the response information arrives at node A.

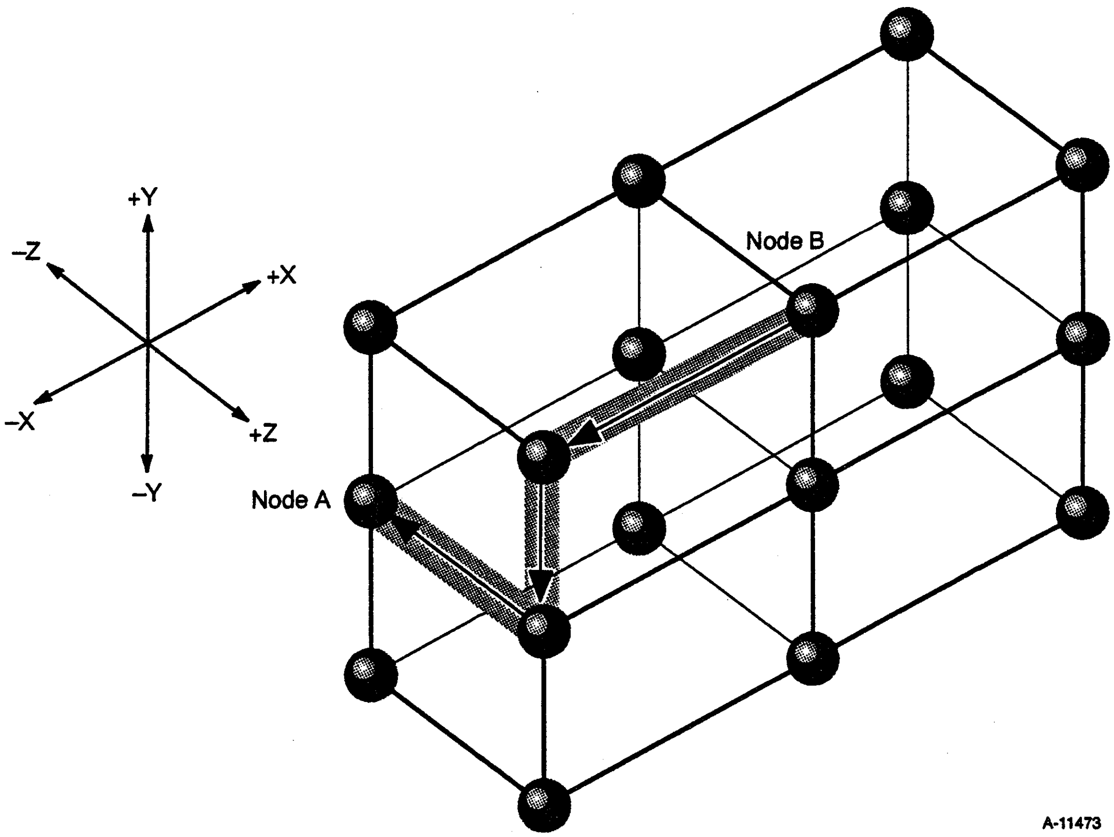


Figure 2-9. –X, –Y, –Z Information Travel

Because information can travel in either the positive or negative direction of a dimension, bad communication links can be avoided. For example, if node A in Figure 2-10 sends information to node B, the information completes one hop in the +X direction and then switches direction into the Y dimension. Suppose, due to a bad communication link, the information cannot complete a hop in the +Y direction. Instead, the information may be routed so it completes two hops in the –Y direction and travels the long way around the torus in the Y dimension. After switching directions into the Z dimension, the information completes one hop in the +Z direction and arrives at node B.

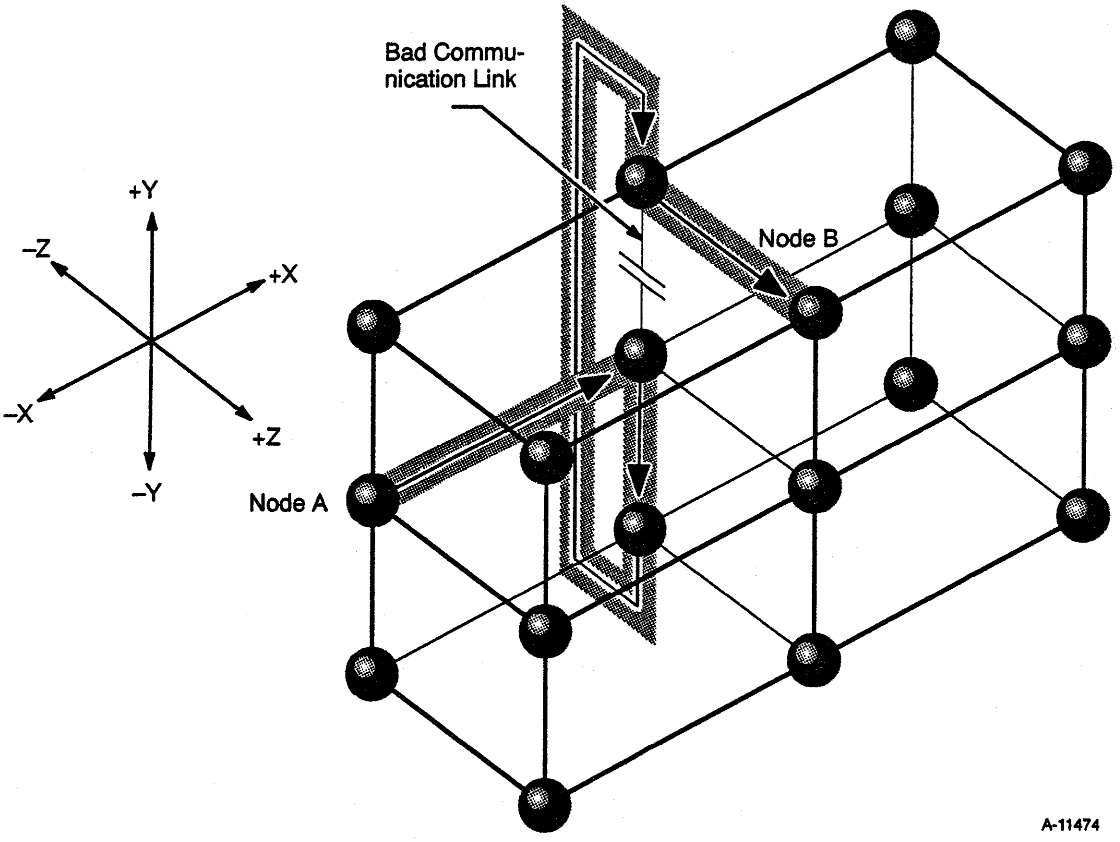


Figure 2-10. Avoiding a Bad Communication Link in the Y Dimension

Virtual Channels

The CRAY T3D system uses virtual channels to prevent communication deadlock conditions. A virtual channel is created when request and response information travels over the same physical communication link but is stored in different buffers. The CRAY T3D system contains four virtual channel buffers (refer to Table 2-1).

Table 2-1. Virtual Channel Buffers

Buffer Name	Definition
Virtual channel 0	Request buffer 0
Virtual channel 1	Request buffer 1
Virtual channel 2	Response buffer 0
Virtual channel 3	Response buffer 1

The virtual channel buffers prevent two types of communication deadlock conditions that may occur in the interconnect network. The following paragraphs describe these conditions.

The first condition would occur if all the nodes in the network contained only one type of buffer and two nodes attempted to simultaneously transfer request or response information to each other. Both nodes would wait until the buffer of the other node was empty before sending information. Because this empty state never occurs, the two nodes would be in a deadlock condition.

To prevent this condition from occurring, the CRAY T3D system contains request buffers and response buffers. These buffers provide separate destination buffers for request and response information.

Due to the characteristics of a torus interconnect network, a second type of deadlock condition could occur if the nodes contained only one set of request and response buffers. If all of the nodes in one dimension were to simultaneously send request or response information to the next node in the dimension, a deadlock condition could occur. For example, a deadlock condition may occur if all of the nodes in the X dimension send request information to the next node in the +X direction at the same time.

To prevent this condition from occurring, the CRAY T3D system contains two request buffers and two response buffers (refer again to Table 2-1). The buffers used when information travels through the network are determined by the dateline communication link. The dateline communication link is a single communication link in each dimension that software designates as the dateline communication link.

When information travels through a dimension, if the information will at some time use the dateline communication link in that dimension, the information always uses virtual channel 1 or virtual channel 3. If, when traveling through a dimension, the information never uses the dateline communication link in that dimension, the information always uses virtual channel 0 or virtual channel 2.

For example, Figure 2-11 shows four nodes in the X dimension. Each node is transferring request information to the node that is 2 hops away in the +X direction. The dateline communication link is the communication link that connects nodes 1 and 2.

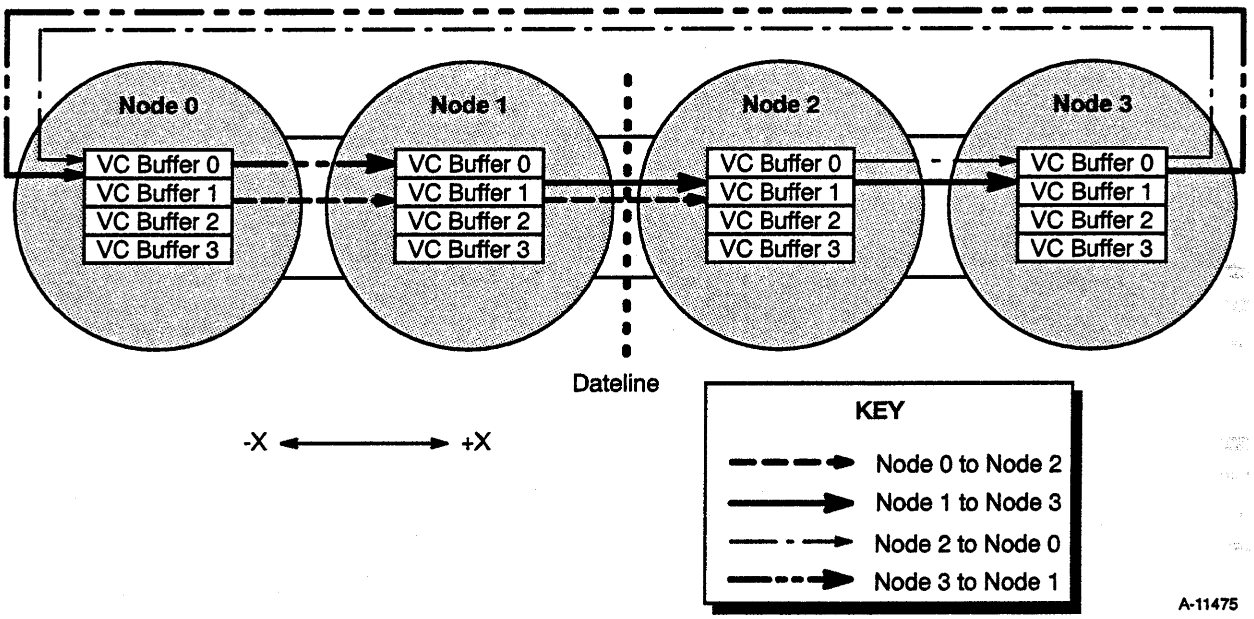


Figure 2-11. Dateline Communication Link

The request information that transfers from node 0 to node 2 and the request information that transfers from node 1 to node 3 will at some time use the dateline communication link. Because of this characteristic, this request information uses virtual channel buffer 1 (request buffer 1).

The request information that transfers from node 2 to node 0 and the request information that transfers from node 3 to node 1 will never use the dateline communication link. Because of this characteristic, this request information uses virtual channel buffer 0 (request buffer 0).

Each virtual channel buffer stores up to eight physical units (phits). Each phit contains 16 bits. A virtual channel is reserved until all of the phits of a packet have traveled through the virtual channel buffer. (More information on packets is provided on the next page.)

Packets

All information transfers over the data signals in a communication link in the form of a packet. A packet contains two parts: a header and a body (refer to Figure 2-12). The header and body have variable lengths and transfer over the communication link one 16-bit phit at a time.

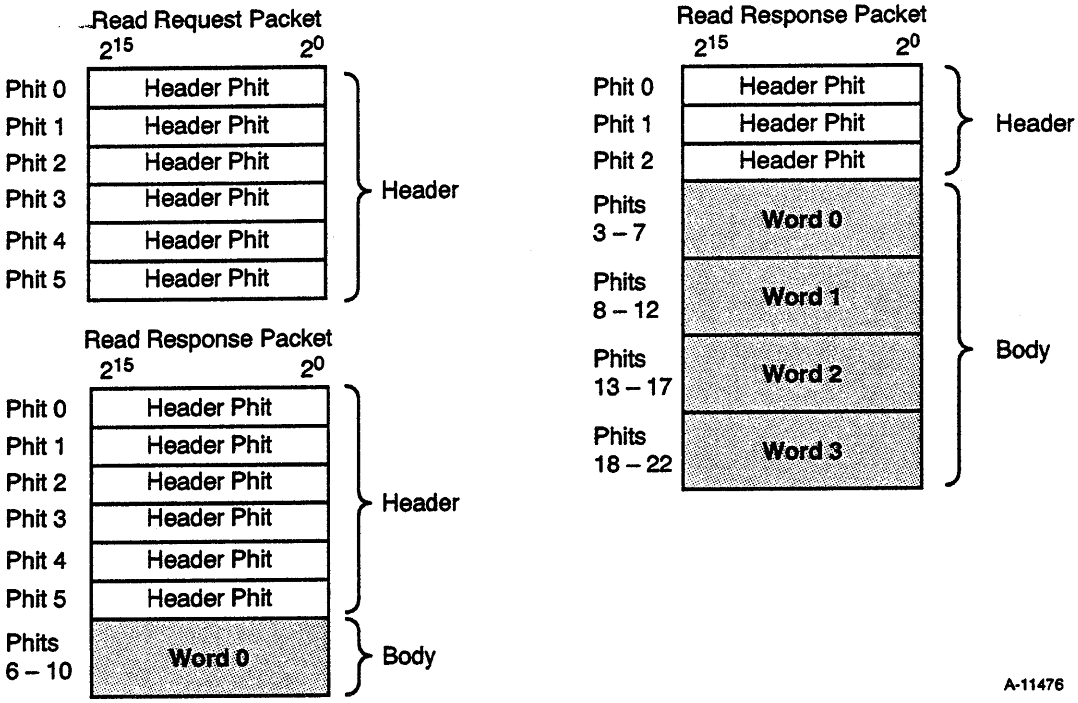


Figure 2-12. Generic Packet Format

Every packet contains a header. The header always contains routing information that steers the packet through the network, destination information that indicates which PE will receive the packet, and control information that instructs the PE that receives the packet to perform an operation. The header may also contain source information that indicates which PE created the packet and may contain memory address information.

When the network interface assembles a packet, it generates parity bits for the packet header information. Packet header information is checked for parity errors each time the network interface receives a packet.

A packet may or may not contain a body. The body of a packet contains one 64-bit word or four 64-bits words of system data. For example, the body of a read response packet contains 1 or 4 words of read data. Each microprocessor uses the check bits to perform error detection and correction on the data that it receives.

Network Routers

The network routers transfer packets through the communication links in the interconnect network. There are two types of network routers: processing element node network routers and I/O gateway network routers.

The processing element node network routers contain three components: an X-dimension switch, a Y-dimension switch, and a Z-dimension switch. Figure 2-13 shows the flow of packet information through a processing element node network router.

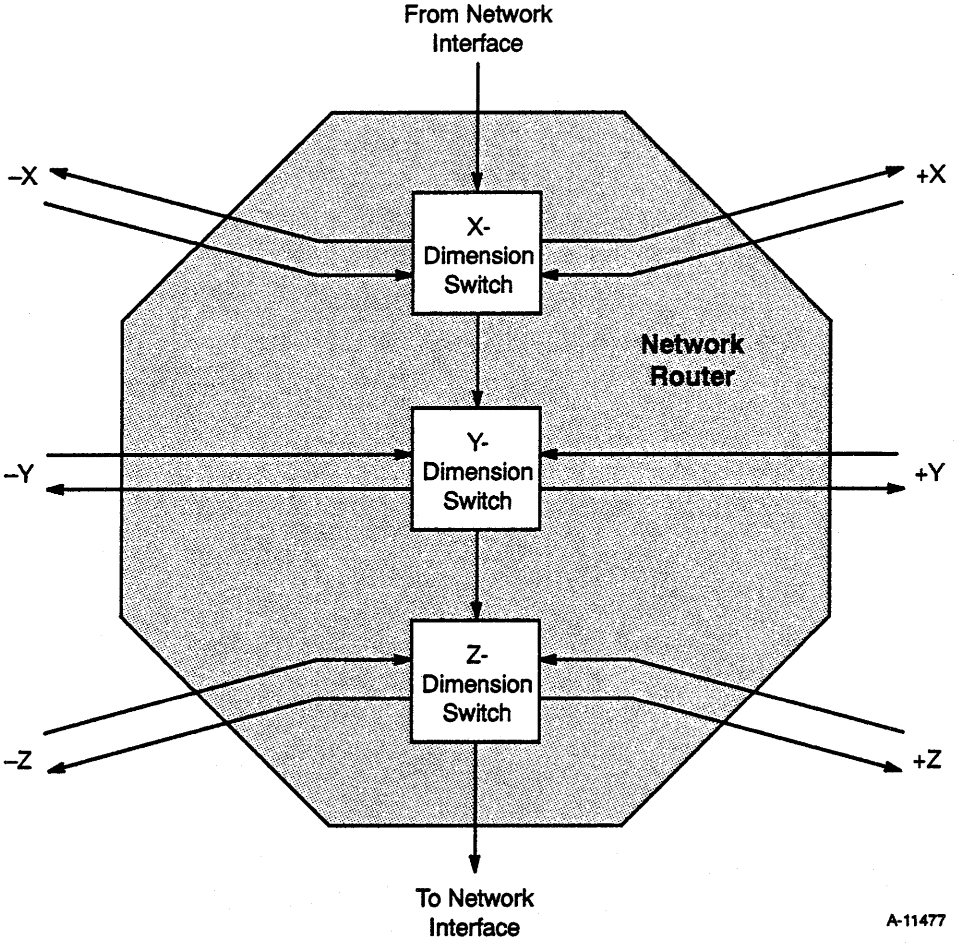


Figure 2-13. Processing Element Node Network Router

The X-dimension switch controls the flow of packets through the X-dimension communication links. Using the routing information in the packet and information received from the channel control signals, the

X-dimension switch steers packets from one X-dimension communication link to the other, or from one X-dimension communication link to the Y-dimension switch.

The Y- and Z-dimension switches function identically to the X-dimension switch. The Y- and Z-dimension switches transfer packets over the Y- and Z-dimension communication links, respectively.

The I/O gateway network routers operate similarly to the processing element node network routers; however, the I/O gateway network routers do not contain a Y-dimension switch. Figure 2-14 shows the components of the input node network router.

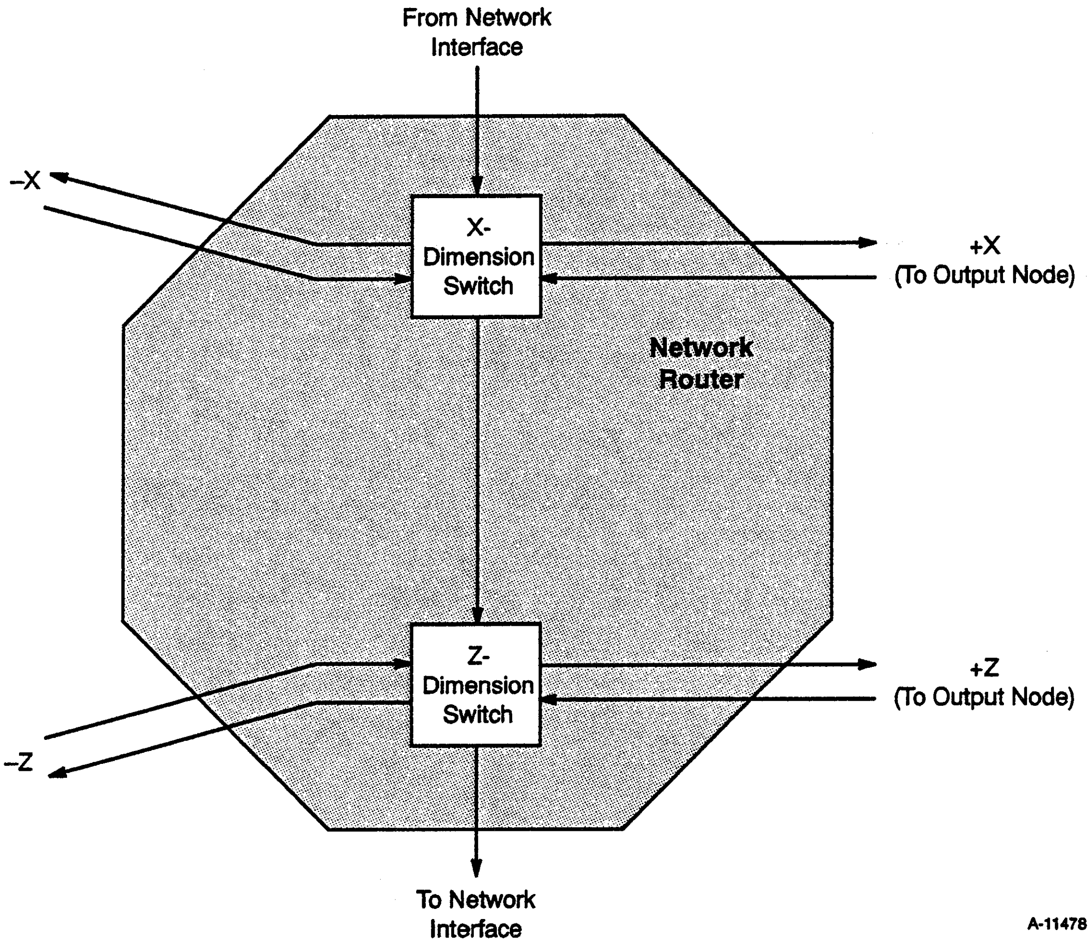


Figure 2-14. I/O Gateway Network Router

The two network routers for an I/O gateway are connected to each other. The +X and +Z communication links from the input node network router connect to the output node network router. The –X and –Z communication links from the output node network router connect to the input node network router.

3 PROCESSING ELEMENT NODES

Processing element nodes perform all program instructions and store system data. After describing processing element (PE) numbering, this section describes physical nodes, and the components of a processing element node.

PE Numbering

Depending on the context, a PE is identified by one of three types of numbers: a physical number, a logical number, or a virtual number. All three types of numbers consist of a PE bit, which identifies whether the PE is PE 0 or PE 1 in a node, and a field containing the node number or node coordinates.

Physical PE Number

Every PE in the CRAY T3D system is assigned a unique number that indicates where the PE is physically located in a system. This number is the physical PE number.

The support circuitry in each PE contains a register called the physical PE register. When a circuit board is placed in the system cabinet, hardware automatically sets the bits of the physical PE register to indicate where the PE is located in the cabinet.

Logical PE Number

Not all of the physical PEs in a CRAY T3D system are part of the logical configuration of a CRAY T3D system. For example, a 512-PE CRAY T3D system contains 520 physical PEs (not including PEs in the I/O gateways). Of these 520 PEs, 512 PEs are used in the logical system and 8 PEs (in 4 spare PE nodes) are used as spare PEs.

Each physical PE used in a logical system is assigned a unique logical PE number. The logical PE number identifies where in the logical system a PE is located.

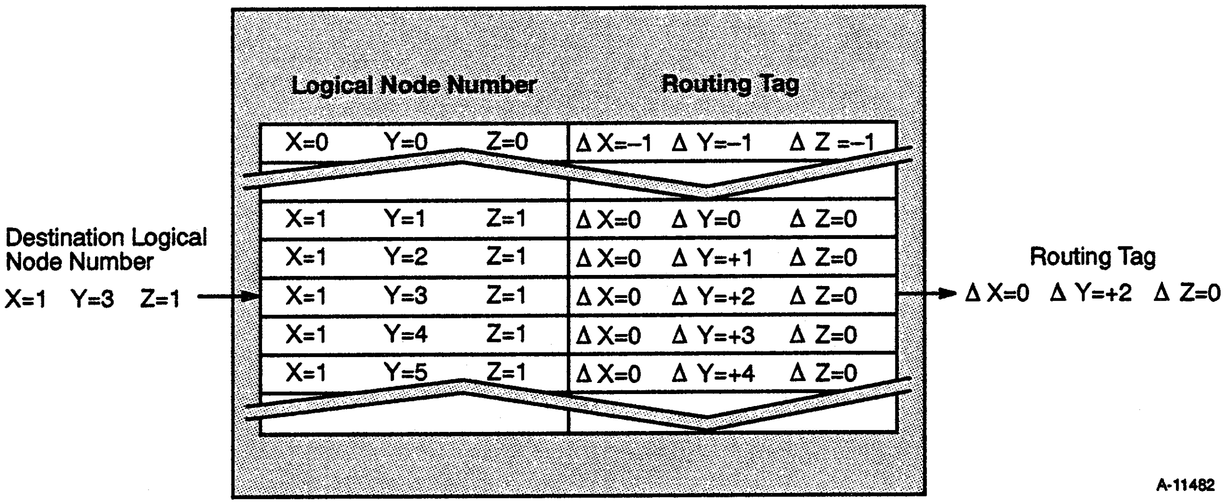
The logical nodes form a three-dimensional matrix of nodes. For example, Figure 3-2 shows the logical PE nodes for a 128-PE CRAY T3D system. Although the system actually contains 68 physical PE nodes, only 64 of the nodes are used in the logical system. The remaining 4 spare physical nodes are physically connected to the interconnect network but are not given logical node numbers.

This type of configuration enables a spare node to logically replace a failing node. When this occurs, the spare node obtains a logical number and the failing nodes does not receive a new logical node number.

For example, if logical node Z=0 Y=2 X=3 fails to operate properly, the physical node assigned to this number may be removed from the logical system. A spare node is then assigned the logical node number Z=0 Y=2 X=3, and the failing node does not receive a logical node number. Information is then rewritten into the routing tag look-up table of each node.

The routing tag look-up table contains information each node uses to create the routing tag in the header of a packet. Because the logical node number may correspond to any of the physical nodes, hardware in the nodes cannot use the logical node number to route data from one node to another.

Each PE node in the CRAY T3D system uses a look-up table to obtain the routing tag. Circuitry in the node enters the logical node number into the routing tag look-up table. The routing tag look-up table then provides the routing tag for a packet (refer to Figure 3-1). The routing tag steers the packet from the physical source node to the physical destination node.



Logical Node Number	Routing Tag
X=0 Y=0 Z=0	Δ X=–1 Δ Y=–1 Δ Z =–1
X=1 Y=1 Z=1	Δ X=0 Δ Y=0 Δ Z=0
X=1 Y=2 Z=1	Δ X=0 Δ Y=+1 Δ Z=0
X=1 Y=3 Z=1	Δ X=0 Δ Y=+2 Δ Z=0
X=1 Y=4 Z=1	Δ X=0 Δ Y=+3 Δ Z=0
X=1 Y=5 Z=1	Δ X=0 Δ Y=+4 Δ Z=0

Destination Logical Node Number
X=1 Y=3 Z=1
Routing Tag
Δ X=0 Δ Y=+2 Δ Z=0
A-11482

Figure 3-1. Routing Tag Look-up Table for Logical Node X=1, Y=1, Z=1

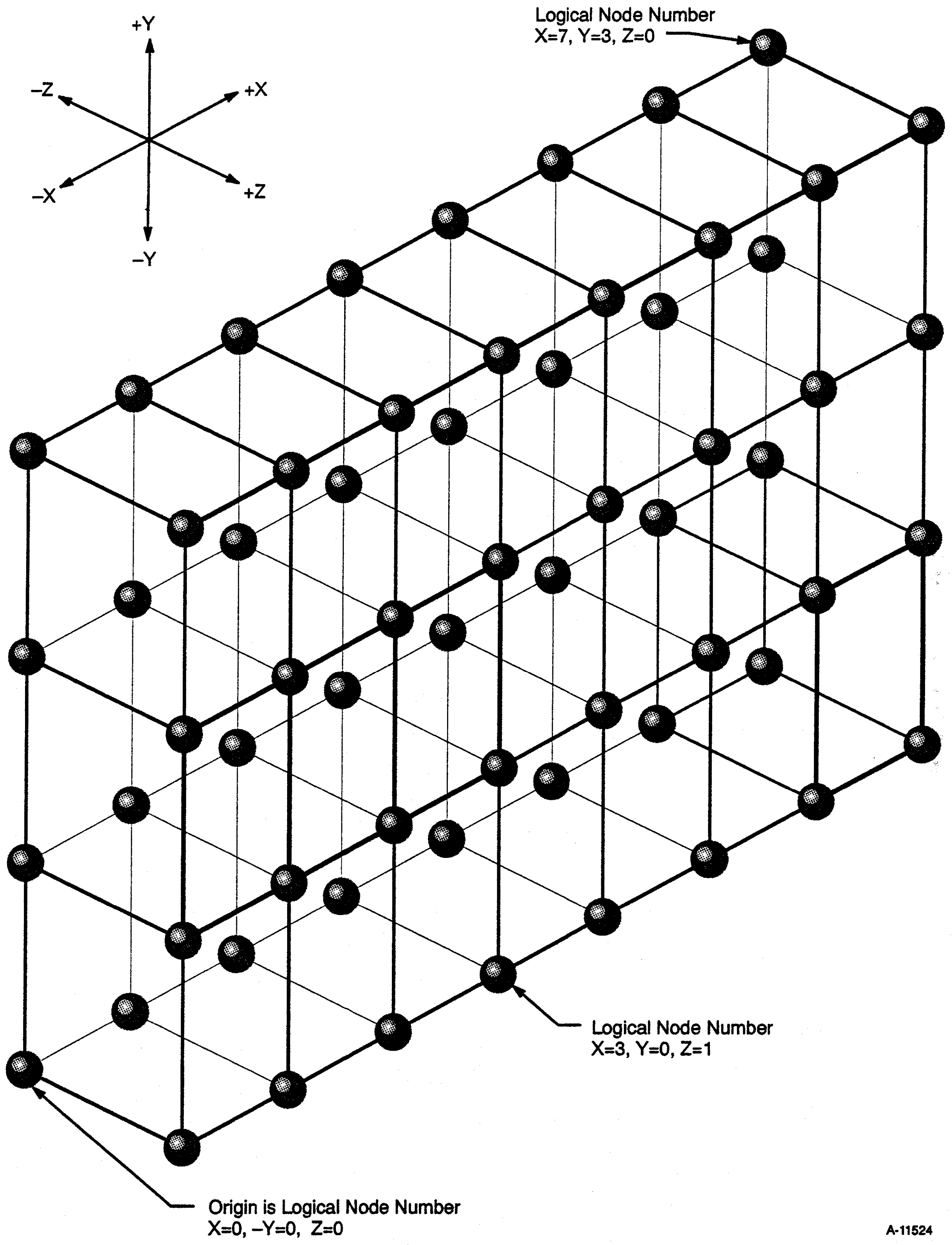


Figure 3-2. Logical Node Numbers

Virtual PE Number

When an MPP application initiates, the support software running on the host system determines the resources needed for the application and creates a partition for the application to run in. A partition is a group of PEs and a portion of the barrier synchronization resources that are assigned to one application. (More information on barrier synchronization is provided later in this section.) The application uses virtual PE numbers to reference the PEs in a partition.

There are two types of partitions: an operating system partition and a hardware partition. In an operating system partition, when the application transfers data between PEs, the operating system must be involved with the transfer. The operating system converts the virtual PE numbers used by the application into logical PE numbers.

In a hardware partition, when the application transfers data between PEs, the operating system is not involved with the transfer. Hardware in each PE node converts the virtual PE numbers used by the application into logical PE numbers.

The virtual PE number contains two parts: the virtual node number and the PE bit. The virtual node number ranges from 1 to 10 bits and indicates which processing element node in a hardware partition the PE resides in. The PE bit indicates whether the PE is PE 0 or PE 1 in the node.

The virtual node number has 0 to 3 bits assigned to the X dimension, 0 to 4 bits assigned to the Y dimension, and 0 to 3 bits assigned to the Z dimension. By assigning bits of the virtual node number to the appropriate dimensions, software arranges the virtual nodes into one of several shapes. For example, a three-bit virtual node number indicates there are eight nodes in the hardware partition. These nodes may be arranged in one of 10 shapes.

Table 3-1 lists the possible node shapes for a three-bit virtual node number. For each shape, the number of nodes in each dimension is limited to powers of two (1, 2, 4, 8, 16, etc.).

Table 3-1. Eight-node Partition Shapes

Eight-node Array Shape in X, Y, Z	Description
8, 1, 1	One-dimensional array in the X dimension
1, 8, 1	One-dimensional array in the Y dimension
1, 1, 8	One-dimensional array in the Z dimension
1, 2, 4	Two-dimensional array in the Y-Z plane
1, 4, 2	Two-dimensional array in the Y-Z plane
2, 1, 4	Two-dimensional array in the X-Z plane
4, 1, 2	Two-dimensional array in the X-Z plane
2, 4, 1	Two-dimensional array in the X-Y plane
4, 2, 1	Two-dimensional array in the X-Y plane
2, 2, 2	Three-dimensional array

Figure 3-3 shows three of the eight-node partition shapes in a 128-PE CRAY T3D system.

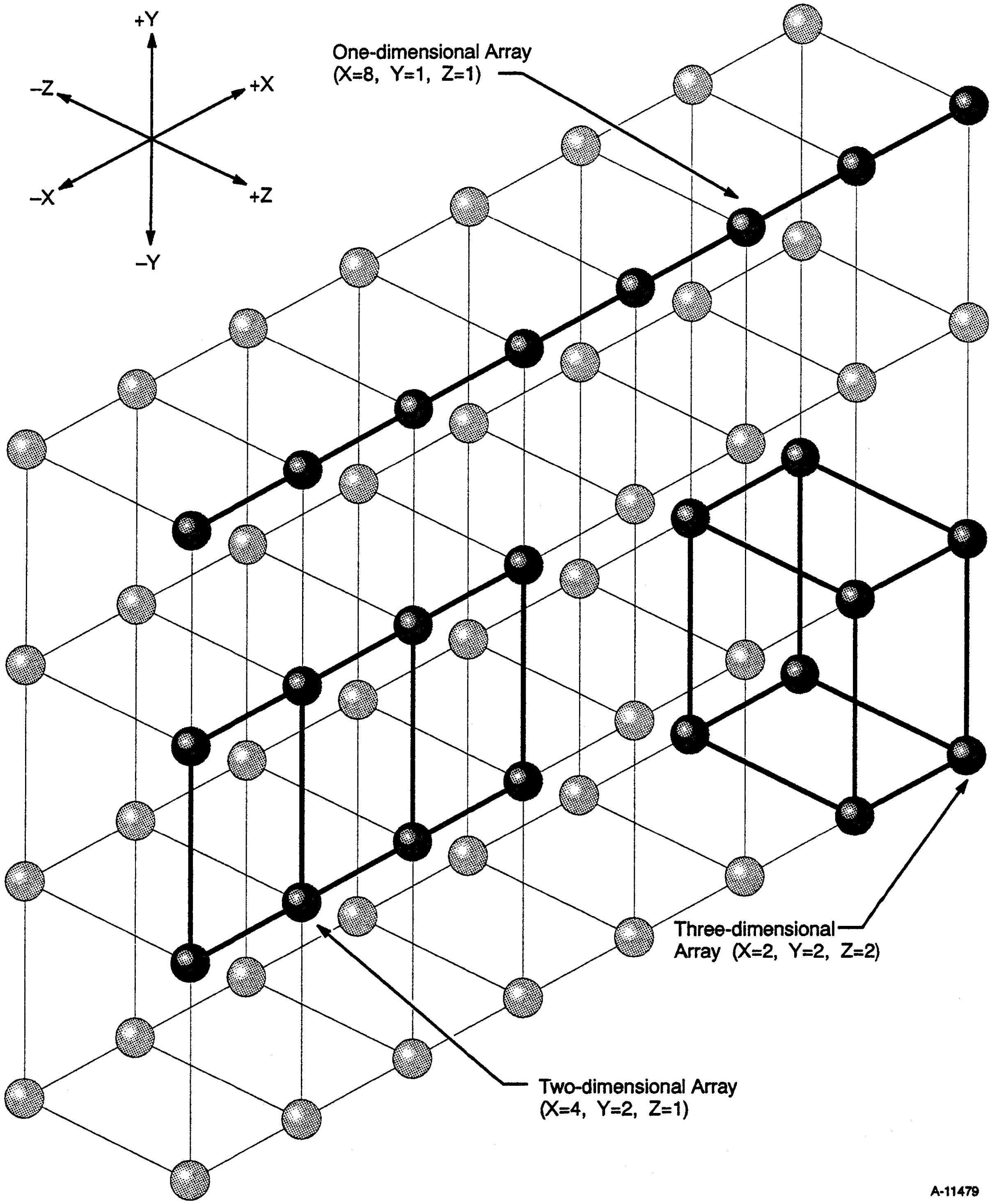


Figure 3-3. Three 8-node Partition Shapes in a 128-PE CRAY T3D System

As an example of virtual PE numbers, Figure 3-4 shows a two-dimensional, eight-node partition that contains 8 nodes. Each node in the partition is referred to by the 3-bit virtual node numbers shown in Figure 3-4.

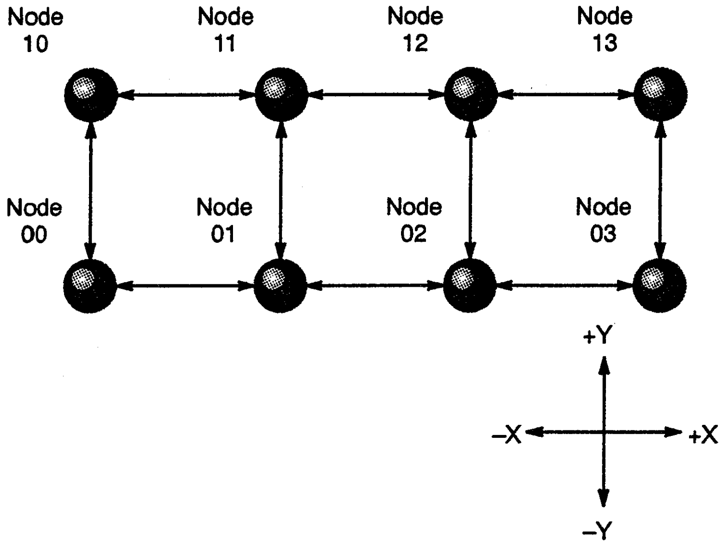


Figure 3-4. Virtual Node Numbers of a Two-dimensional Array

This two-dimensional array of eight nodes may actually correspond to one of many two-dimensional, eight-node arrays in the logical system. For example, Figure 3-5 shows two examples of how this two-dimensional array may be placed in the logical system of nodes in a 128-PE CRAY T3D system.

A virtual node number does not always correspond to the same logical node number. For example, Figure 3-5 shows how virtual node Y=1 X=2 from Figure 3-4 may correspond to either logical node number Z=1 Y=2 X=2 or logical node number Z=1 Y=3 X=6.

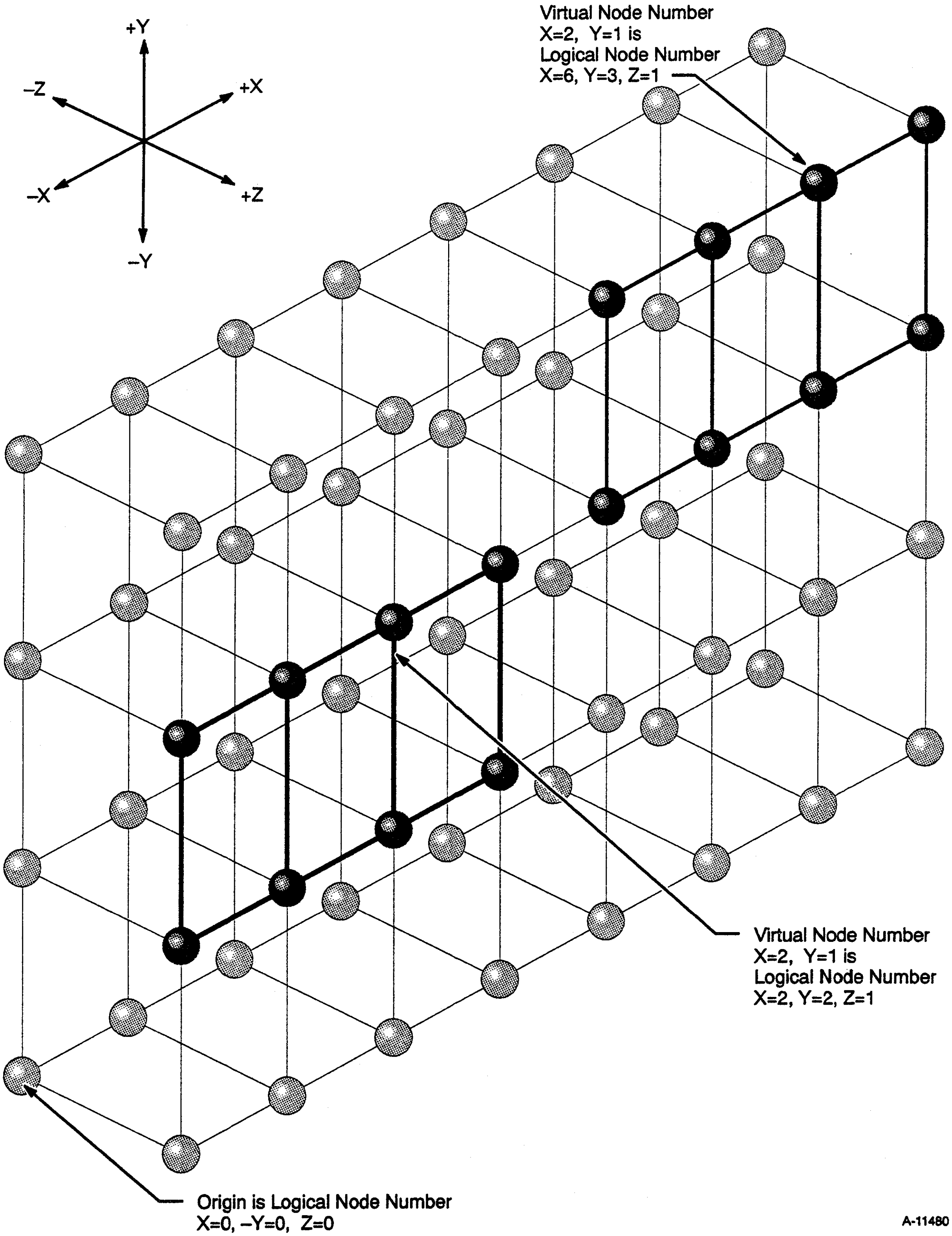


Figure 3-5. Virtual and Logical Node Numbers

Physical Nodes

Physically, each processing element node resides on half of a circuit board in the CRAY T3D system cabinet (refer to Figure 3-6). The integrated circuits above the dashed line are used for the components of one processing element node. The integrated circuits below the dashed line are used for the components of another processing element node.

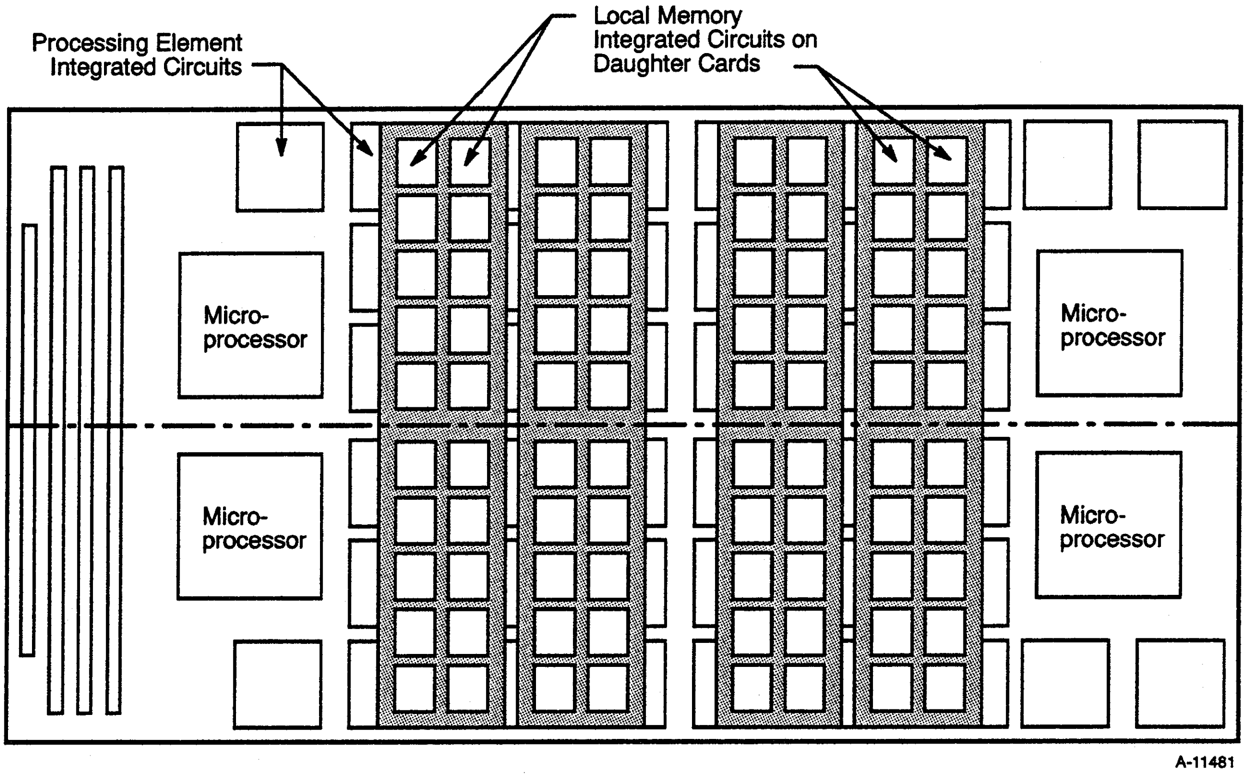

Processing Element Integrated Circuits
Local Memory Integrated Circuits on Daughter Cards
Micro-processor
Micro-processor
Micro-processor
Micro-processor
A-11481

Figure 3-6. Processing Element Node Circuit Board

Components

A processing element node is composed of four components: two PEs, a BLT, and a network interface (refer to Figure 3-7).

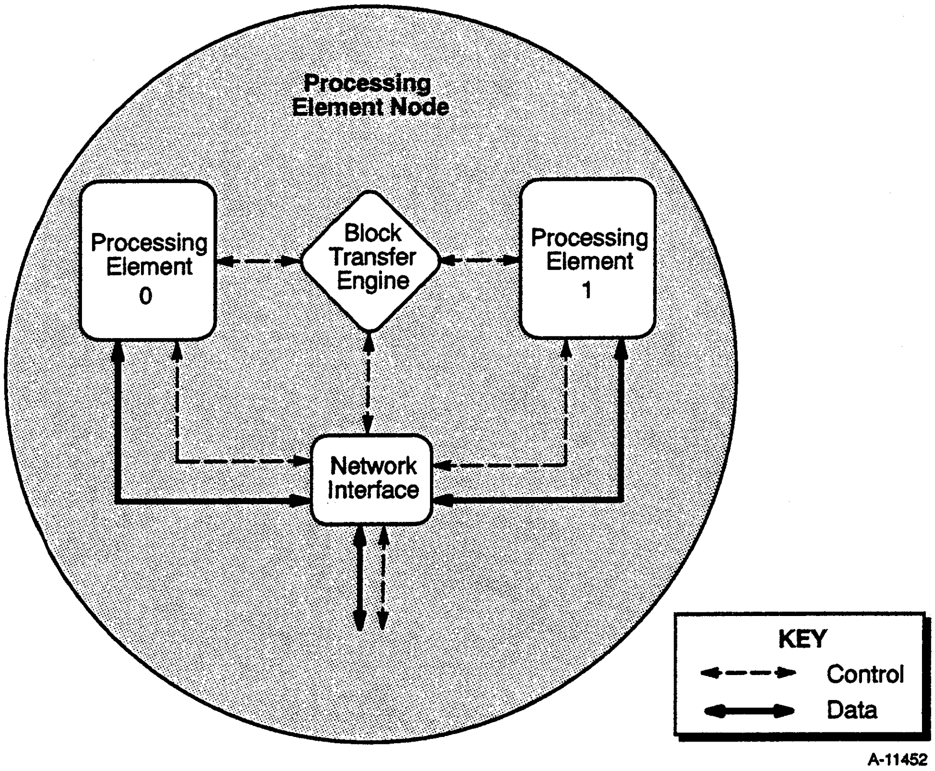


Figure 3-7. Processing Element Node

Figure 3-8 is a functional block diagram of the components in a processing element node. The following subsections describe these components.

Processing Elements

Each processing element node contains two PEs: PE 0 and PE 1. Each PE contains a microprocessor, local memory, and support circuitry.

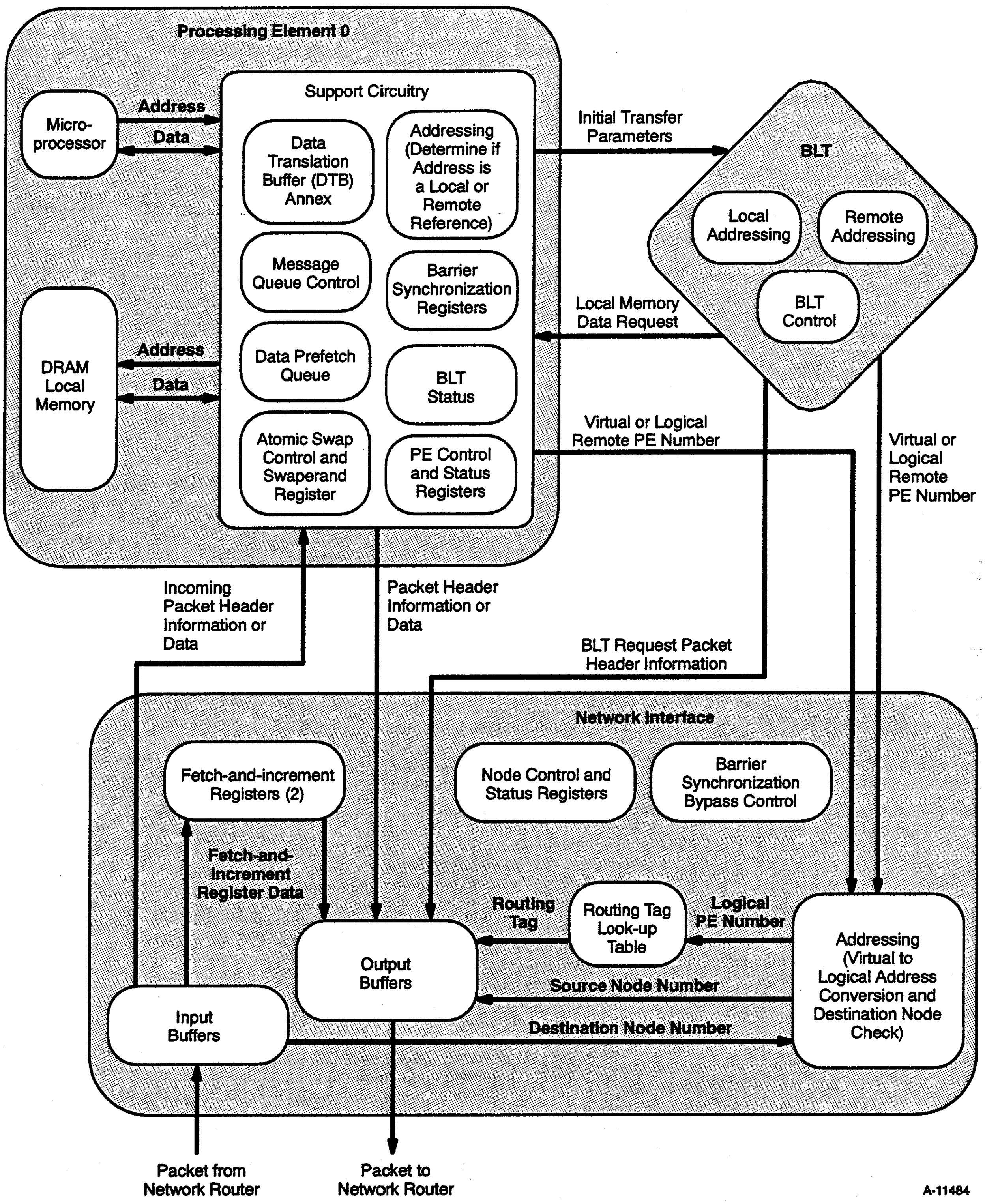

Processing Element 0
Micro-processor
Address
Data
Support Circuitry
Data Translation Buffer (DTB) Annex
Addressing (Determine if Address is a Local or Remote Reference)
Message Queue Control
Barrier Synchronization Registers
Data Prefetch Queue
BLT Status
Atomic Swap Control and Swaperand Register
PE Control and Status Registers
DRAM Local Memory
Address
Data
Initial Transfer Parameters
BLT
Local Addressing
Remote Addressing
BLT Control
Local Memory Data Request
Virtual or Logical Remote PE Number
Virtual or Logical Remote PE Number
Incoming Packet Header Information or Data
Packet Header Information or Data
BLT Request Packet Header Information
Network Interface
Fetch-and-increment Registers (2)
Node Control and Status Registers
Barrier Synchronization Bypass Control
Fetch-and-increment Register Data
Routing Tag
Routing Tag Look-up Table
Logical PE Number
Output Buffers
Addressing (Virtual to Logical Address Conversion and Destination Node Check)
Source Node Number
Input Buffers
Destination Node Number
Packet from Network Router
Packet to Network Router
A-11484

Figure 3-8. Processing Element Node Functional Block Diagram

Microprocessor

The microprocessor is a reduced instruction set computer (RISC) 64-bit microprocessor developed by Digital Equipment Corporation. The microprocessor comprises a central control unit, an integer execution unit, a floating-point execution unit, an address generation and bus interface unit, a data cache memory, and an instruction cache memory (refer to Figure 3-9). The following paragraphs describe each of these components.

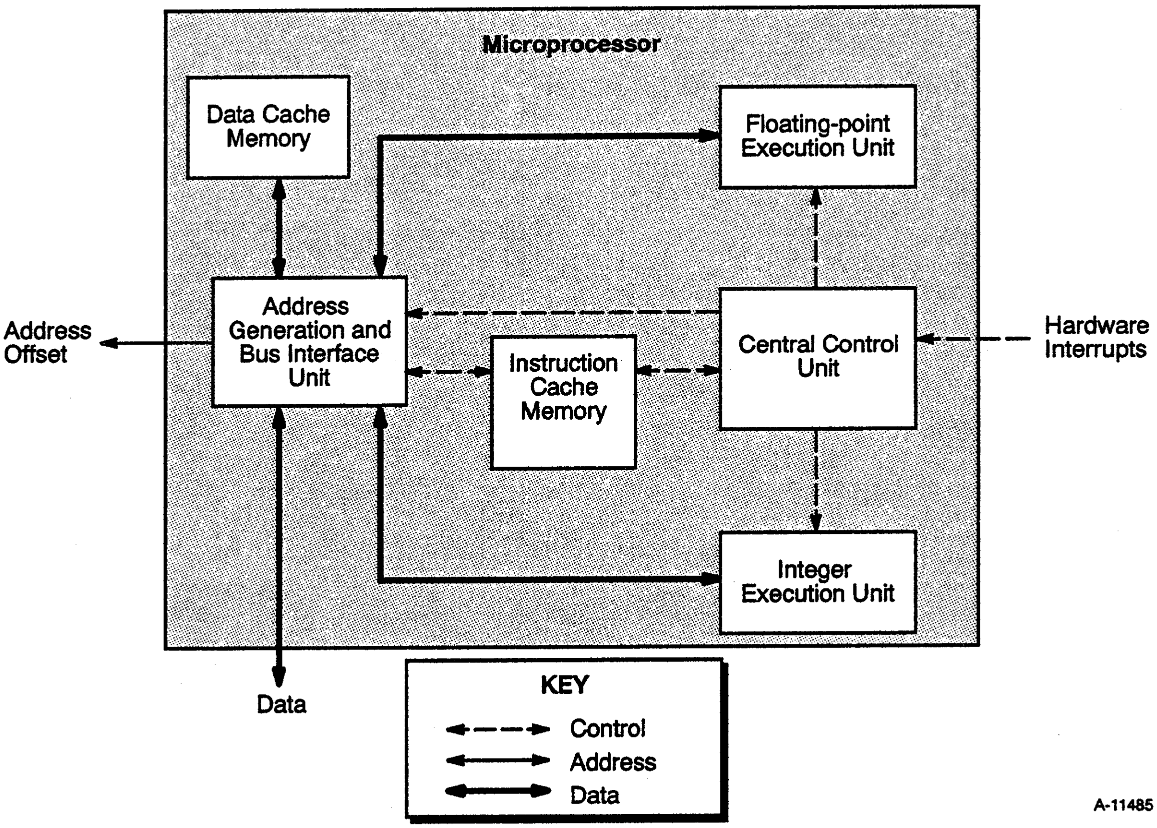


Figure 3-9. Microprocessor

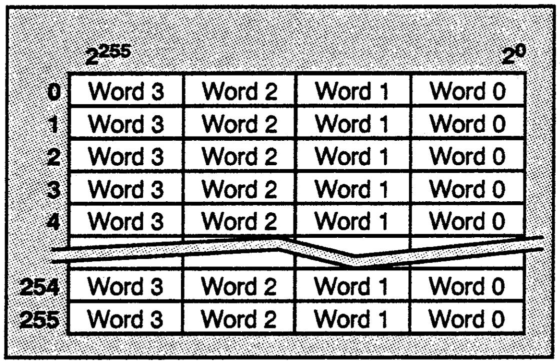
The central control unit issues instructions to the integer execution unit, the floating-point execution unit, and the address generation and bus interface unit. The central control unit also receives hardware interrupt signals from external PE circuitry. These interrupts include barrier synchronization interrupts, messaging interrupts, BLT interrupts, and error interrupts.

The integer execution unit performs integer operations on 64-bit integer registers. Integer operations include arithmetic, compare, logical, and shift operations. There are a total of 32 integer registers.

The floating-point execution unit performs floating-point operations on 64-bit floating-point registers. The floating-point operations include IEEE arithmetic instructions, plus instructions for performing conversions between floating-point and integer quantities. There are 32 floating-point registers.

The address generation and bus interface unit generates memory addresses and steers data and control information. The address generation circuitry converts the virtual address it receives from the compiler into address information used by the support circuitry in the PE. The bus interface circuitry responds to read or write instructions for the data cache, instruction cache, and data bus.

Data cache memory is a small, high-speed random access memory that temporarily stores frequently or recently accessed data. The data cache memory is internal to the microprocessor and stores 256 32-byte lines (four 64-bit words in a line) of data (refer to Figure 3-10).



	2^{255}			2^0
0	Word 3	Word 2	Word 1	Word 0
1	Word 3	Word 2	Word 1	Word 0
2	Word 3	Word 2	Word 1	Word 0
3	Word 3	Word 2	Word 1	Word 0
4	Word 3	Word 2	Word 1	Word 0
254	Word 3	Word 2	Word 1	Word 0
255	Word 3	Word 2	Word 1	Word 0

A-11486

Figure 3-10. Data Cache and Instruction Cache Memory Organization

The instruction cache memory operates similarly to the data cache memory, but stores instructions. The instruction cache memory is also internal to the microprocessor and stores 256 32-byte lines of data.

The microprocessor also performs single-error correction/double-error detection (SECDED[†]). After receiving 128 data bits and 28 check bits from the PE circuitry, the microprocessor generates a new set of check bits. If the new set of check bits is not identical to the original set of check bits, 1 or more of the system data or original check bits changed value during the data transfer.

If only 1 bit changed value, hardware in the microprocessor corrects the value of the incorrect bit. If more than 1 bit changed value, the microprocessor is interrupted.

[†] Hamming, R. W. "Error Detection and Correcting Codes." *Bell System Technical Journal*. 29.2 (1950): 147–160.

Each microprocessor is a 431-pin pin grid array (PGA) integrated circuit (refer again to Figure 3-6). Table 3-2 lists the specifications for the microprocessor.

Table 3-2. Microprocessor Specifications

Characteristic	Specification
Number of microprocessors per processing element node	Two (One in each PE)
Microprocessor type	Reduced instruction set computer (RISC) 64-bit microprocessor
Bi-directional data bus	128 data bits plus 28 check bits
Data error protection	Single error correction/double error detection
Address bus	34 bits of address information
Clock speed	6.67 ns
Issue rate	2 instructions per clock period maximum
Internal data cache memory size	8 Kbytes (256 32-byte lines)
Internal instruction cache memory size	8 Kbytes (256 32-byte lines)
Latency of data cache to CPU transfers	3 clock periods
Bandwidth of data cache to CPU transfers	64 bits per clock period
Floating-point instruction unit	Supports the Institute of Electrical and Electronic Engineers (IEEE) floating-point arithmetic instructions plus instructions for converting between floating-point and integer quantities
Total number of floating-point registers	32
Floating-point register size	64 bits
Integer instruction unit	Integer arithmetic, compare, logical, and shift instructions
Total number of integer registers	32
Integer register size	64 bits
Integrated circuit type	Complimentary metal-oxide semiconductor (CMOS)
Integrated circuit size	14.1 mm X 16.8 mm
Total number of pins	431
Number of signal pins	291
Typical power dissipation	23 W

Local Memory

Each PE contains local memory. Local memory consists of dynamic random access memory (DRAM) that stores system data. A low-latency, high-bandwidth data path connects the microprocessor to local memory in a PE.

System data is stored in a physically distributed, logically shared memory. Memory is physically distributed because each PE contains a local memory. Memory is logically shared because any microprocessor can access data in the local memory of any PE without involving the microprocessor in that PE.

Figure 3-11 illustrates the physical distribution of memory in a 256-PE CRAY T3D system. The local memory in each PE stores a set number of 64-bit words (represented by the variable *m* in Figure 3-11). The size of local memory depends on the type of DRAM integrated circuits used in the system.

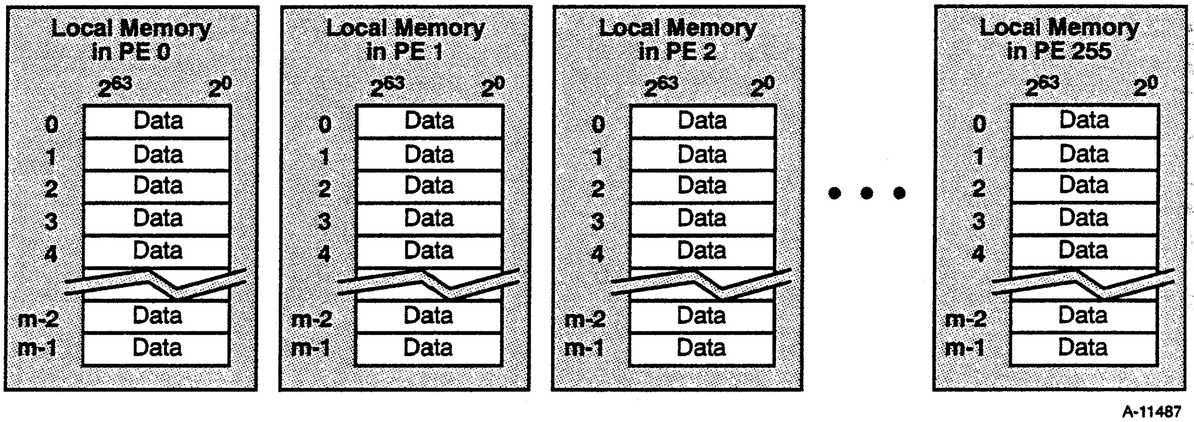


Figure 3-11. Physical Distribution of Memory

The total size of shared memory is the size of local memory in one PE multiplied by the total number of PEs in the system. For example, a CRAY T3D system with 512 PEs, each with 2 Mwords of local memory, has a total system memory of 1 Gword (8 Gbytes).

Each microprocessor uses memory addressing that references any word in shared memory. This address, the virtual address, is initially generated by the program compiler. As previously described, the virtual address is converted into a logical node number, PE number, and address offset by the microprocessor and other components in the processing element node.

Local memory comprises DRAM integrated circuits that are mounted on daughter-card printed circuit boards. The DRAM daughter cards plug into the PE circuit board and reside on top of the integrated circuits used in the PEs (refer to Figure 3-12 and again to Figure 3-6)

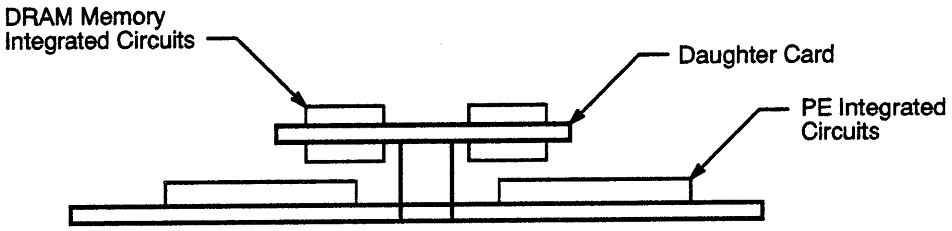


Figure 3-12. Side View of PE Circuit Board

Table 3-3 lists the specifications for local memory.

Table 3-3. Local Memory Specifications

Characteristic	Specification
Local memory size per PE	2 Mwords (4 Mbit DRAMs) or 8 Mwords (16 Mbit DRAMs)
Total memory per processing element node	4 Mwords (4 Mbit DRAMs) or 16 Mwords (16 Mbit DRAMs)
Total logically shared memory in a CRAY T3D system	Local memory size per PE multiplied by the total number of PEs in the system

Support Circuitry

The support circuitry extends the control and addressing functions of the microprocessor. These functions include:

- Address interpretation
- Reads and writes
- Data prefetch
- Messaging
- Barrier synchronization
- Fetch and increment
- Status

Address Interpretation

In the CRAY T3D system, the address pins of the microprocessor do not directly address a physical memory. Instead, the support circuitry in the PE interprets the address and routes data between the microprocessor and either local memory, memory-mapped registers, or memory in a remote PE.

The support circuitry uses part of the address generated by the microprocessor as an index into a 32-entry table called the DTB annex. Each entry in the DTB annex contains a virtual or logical PE number and a function code. The PE number is the number of the destination PE. The function code indicates what type of memory function the support circuitry will perform.

The support circuitry compares the PE number received from the DTB annex with the number of the PE that contains the support circuitry. If they match, the microprocessor is addressing local memory. If they do not match, the microprocessor is addressing memory in another PE.

When the microprocessor requests a data transfer with local memory, the support circuitry uses the address from the microprocessor as an address offset for data in local memory. The support circuitry then transfers data between the microprocessor and local memory.

When the microprocessor requests a data transfer with remote memory, the support circuitry sends the remote PE number along with the address offset and control information to the network interface for use in the header of a request packet.

When the microprocessor addresses a register in the support circuitry, the support circuitry routes the information to the appropriate register. In some cases, the support circuitry also performs a function related to the register.

Reads

Read operations transfer data from system memory to a register in the microprocessor. Depending on the type of read operation, the microprocessor may also update information in the data cache.

After receiving address and cycle request information from the microprocessor, the support circuitry retrieves a function code and PE number from an entry in the DTB annex. The value of the function code determines which read operation the support circuitry performs. There are two main types of read operations: noncacheable reads and cached reads.

During noncacheable reads, the microprocessor does not place a copy of the read data in the data cache. During cached reads, the microprocessor does place a copy of the read data in the data cache. Cached reads may be used to reduce the latency of subsequent read operations to specified memory addresses. If the data in the data cache is valid, the microprocessor reads data from the data cache instead of from system memory.

During noncacheable memory read operations, as the support circuitry transfers the read data to the microprocessor, the support circuitry signals the microprocessor not to update a line in the data cache. There are two types of noncacheable read operations: normal noncacheable read and noncacheable atomic swap read.

Normal noncacheable read operations transfer data from memory to a register in the microprocessor without updating the data cache. Noncacheable atomic swap read operations transfer a 64-bit word from memory to the microprocessor and then transfer another 64-bit word into the same memory location in an indivisible operation.

Before initiating an atomic swap operation, the microprocessor loads a 64-bit word into a register called the swaperand register. During the atomic swap operation, the support circuitry transfers a word from memory to the microprocessor, then transfers the word from the swaperand register to the same memory location.

During cached read operations, as the support circuitry transfers the read data to the microprocessor, the support circuitry signals the microprocessor to update a line in the data cache. There are three types of cached read operations: a normal cached read, a cached atomic swap read, and a cached read ahead.

Normal cached reads and cached atomic swap reads operate the same as a noncacheable normal read and noncacheable atomic swap read except, when transferring data to the microprocessor, the support circuitry signals the microprocessor to update a line in the data cache.

Cached read aheads are used to hide the latency of local memory reads. The following paragraphs describe a cached read ahead.

The support circuitry contains a local memory read stage. After the support circuitry performs a cached read ahead operation, the local memory read stage buffers a four-word block of data (or instruction fetches) read from local memory. When the microprocessor issues any type of cached or noncacheable read operation (except atomic swaps) with an address that matches the buffered block of data, data transfers from the local memory read stage to the microprocessor. This action prevents the support circuitry from having to access DRAM memory to retrieve the block of data and decreases the latency for the read operation.

During a cached read ahead operation, the support circuitry retrieves a block of data from local memory (or the local memory read stage). The support circuitry then sends the block of data to the microprocessor and signals the microprocessor to update the data cache.

Immediately after sending the data to the microprocessor, the support circuitry retrieves the next sequential block of data from local memory and buffers the data in the local memory read stage of the support circuitry. The data buffered in the support circuitry remains in the support circuitry until a memory-barrier instruction is issued or a data or instruction read operation from a different local memory address occurs.

Writes

Write operations transfer data from the microprocessor to system memory. To initiate a write operation, the microprocessor provides the support circuitry with an address and with cycle request information. The microprocessor may then continue issuing program instructions while the support circuitry completes the write operation.

After receiving the address and cycle request information from the microprocessor, the support circuitry retrieves a function code and PE number from an entry in the DTB annex. The support circuitry then checks the value of the PE number read from the DTB annex. If the PE number is set to the local PE, the support circuitry writes up to four 64-bit words into local memory. If the PE number is set to a remote PE, the support circuitry creates a write request packet that contains up to 4 words of data and sends the request packet to the remote PE.

After creating a write request packet, the support circuitry increments a counter (called the outstanding write request counter) that counts the number of write request packets created and sent to remote PEs. After receiving a write response packet, the support circuitry in the PE that requested the write operation decrements the outstanding write request counter. This action completes a write operation.

Data Prefetch

When requested by the microprocessor, the support circuitry performs a data prefetch operation. A data prefetch operation transfers one 64-bit word of data from memory in a remote PE to the data prefetch queue, which is located in the local PE support circuitry.

The microprocessor initiates a data prefetch operation when it encounters a prefetch instruction in a program. A programmer may place the prefetch instruction several instructions before the instruction that actually uses the prefetch data.

When issuing the prefetch instruction, the microprocessor signals the support circuitry that the next data transfer is a prefetch operation. After the microprocessor issues the prefetch instruction, the microprocessor continues with other program instructions.

The support circuitry assembles information for a prefetch read request packet and sends the information to the remote PE over the interconnect network. After receiving the request packet, the support circuitry in the destination PE creates a prefetch read response packet that contains the word of data and indicates that the data is a prefetch response. The support circuitry in the remote PE then sends the response packet to the local PE over the interconnect network.

The support circuitry in the local PE receives the prefetch response packet and stores the word of data in the data prefetch queue. When the microprocessor issues the instruction that uses the data, the microprocessor reads the data from the data prefetch queue instead of creating a read request packet and waiting for a response.

The data prefetch queue stores a maximum of 16 words. The microprocessor can issue up to 16 data prefetch instructions before reading the data out of the prefetch queue 1 word at a time.

Messaging

The support circuitry also controls the messaging facility. The messaging facility transfers a special packet, called a message, from one PE to another PE. After receiving a message, the support circuitry in a PE interrupts the microprocessor and places the message in a message queue. The microprocessor may then read the message from the message queue.

The message queue is located in a reserved portion of local memory. The message queue stores up to 4,080 message packets and includes 16 reserved locations for a small amount of overflow (total of 256K bytes of information). The support circuitry places message packets in the message queue in the order that they are received.

To create a message, the microprocessor fills one of its internal write buffer lines with 4 words of data. The microprocessor then transfers the data from the write buffer to the support circuitry. During the transfer, the microprocessor also provides the support circuitry with an address and with cycle request information.

After receiving the address and cycle request information from the microprocessor, the support circuitry retrieves a function code and PE number from an entry in the DTB annex. The function code indicates that the support circuitry should perform a message write. The support circuitry then creates a message packet and sends the packet to the destination PE.

After receiving the message packet, the support circuitry in the destination PE attempts to store the message in the message queue. If the message queue can accept the message, the support circuitry stores the message in the queue and sets the message hardware interrupt for the microprocessor. The support circuitry in the destination PE then creates a message acknowledge packet and sends the packet to the PE that created the message.

If the message queue in the destination PE cannot accept the message (full message queue), the support circuitry returns the message to the requesting PE by creating a no-acknowledge (NACK) packet. After receiving the NACK, the requesting PE can resend the message. Because of this feature, message delivery is guaranteed regardless of the amount of system message traffic.

In addition to message packets, the support circuitry may receive error "messages" and store the error messages in the message queue. The network interface generates error messages if it receives a misrouted packet or if it receives a packet that contains parity errors. If a network error occurs, the network interface turns the packet it received into an error message and sends the error message to the appropriate PE in the node.

Barrier Synchronization

The support circuitry also controls barrier synchronization operations. There are two types of barrier synchronization operations: barriers and eurekas.

A barrier is a point in program instructions where a microprocessor must wait until all other microprocessors associated with the barrier have finished their part of the program instructions. A programmer may use a barrier to ensure that all of the microprocessors associated with a distributed, parallel loop in a program finish the instructions for the loop before continuing with other program instructions.

The support circuitry in each PE contains two 8-bit registers called barrier register 0 and barrier register 1. Each bit in the barrier registers is connected to a separate barrier synchronization circuit. For example, Figure 3-13 shows the barrier synchronization circuit for bit 2 of barrier register 0 in a simplified CRAY T3D system.

All of the barrier synchronization circuits function identically and independently. The following paragraphs describe the operation of the barrier synchronization circuit connected to bit 2 of barrier register 0. Before the barrier synchronization process begins, bit 2 of barrier register 0 in each PE is reset to a logical 0.

Each barrier synchronization circuit in the CRAY T3D system is actually an AND-tree and fan-out tree circuit (refer again to Figure 3-13). The AND-tree circuit receives an input from all of the PEs. The fan-out-tree circuit sends a copy of the final AND gate output to all of the PEs.

The first layer of the AND tree contains four AND gates. Each AND gate receives signals from two PEs. For example, one AND gate receives signals from bit 2 of barrier register 0 in PE 0 and bit 2 of barrier register 0 in PE 1. When all of the microprocessors set bit 2 of barrier register 0 to 1, the output of each of the four AND gates is 1.

The second layer of the AND tree contains two AND gates. Each AND gate receives signals from two of the AND gates in the first layer of the AND tree. When the output of all the AND gates in the first layer of the AND tree is 1, the output of both the AND gates in the second layer of the AND tree is 1.

The third layer of the AND tree contains the final AND gate. This AND gate receives signals from both AND gates in the second layer of the AND tree. When the output of both AND gates in the second layer of the AND tree is 1, the output of the final AND gate is 1. The output of the final AND gate connects to the fan-out tree circuit.

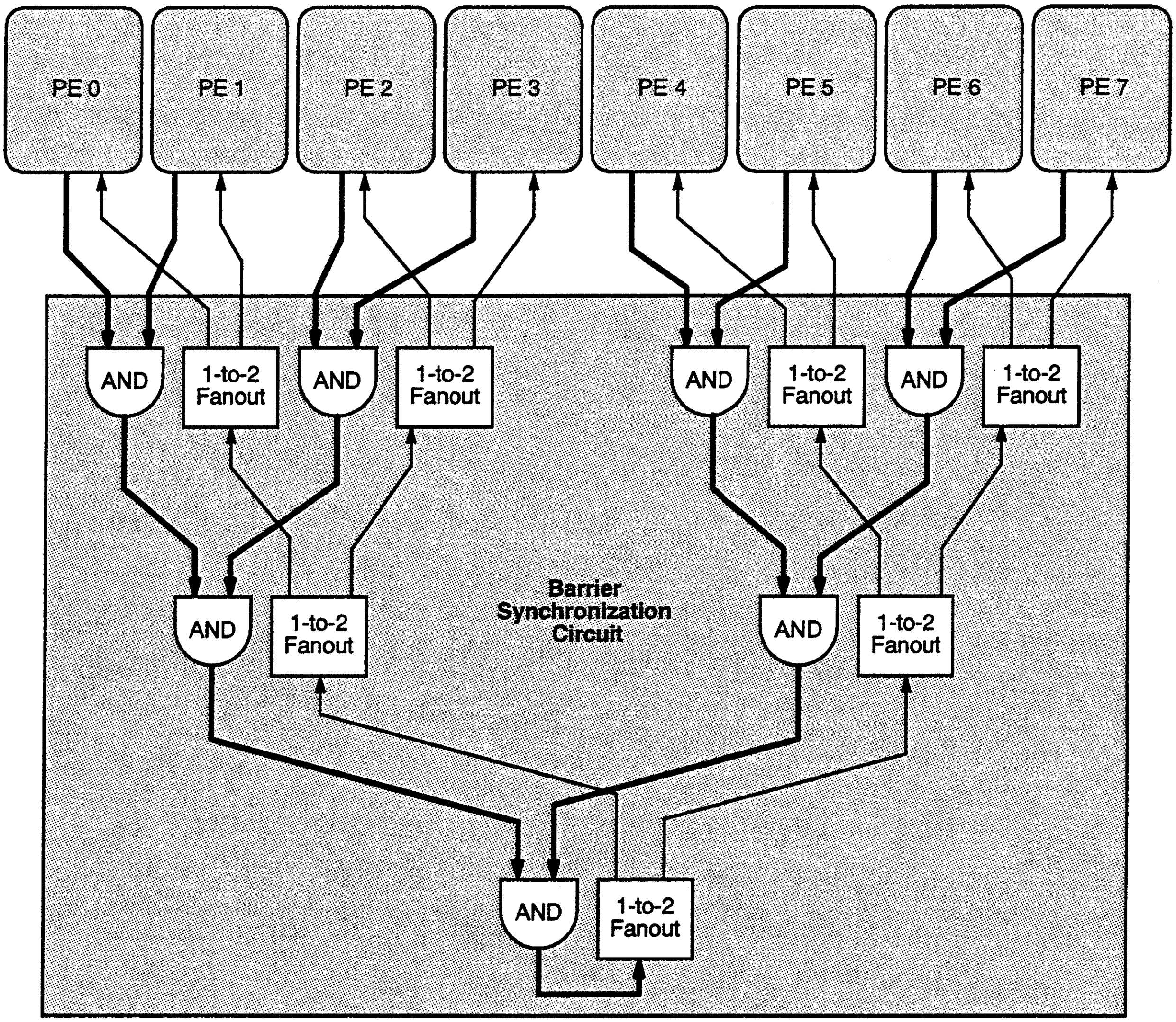


Figure 3-13. Simplified Barrier Synchronization Circuit

The first fan-out block in the fan-out tree receives a 1 from the final AND gate. After creating two copies of the 1, the first fan-out block sends the logical 1's to two fan-out blocks in the second layer of the fan-out tree.

The two fan-out blocks in the second layer of the fan-out tree each create two copies of the 1. The two fan-out blocks in the second layer of the fan-out tree then send the 1's to four fan-out blocks in the third layer of the fan-out tree.

The four fan-out blocks in the third layer of the fan-out tree each create two copies of the 1. The fan-out blocks in the third layer of the fan-out tree then send the 1's to the support circuitry in each of the eight PEs.

The microprocessor monitors the barrier synchronization circuit using one of two methods. In the first method, after the microprocessor sets bit 2 of barrier register 0 to 1, the microprocessor enters a loop that continuously checks the value of bit 2 of barrier register 0. After receiving a 1 from the fan-out circuitry, the support circuitry resets bit 2 of barrier register 0 to 0. Because the microprocessor constantly checks the value of bit 2 of barrier register 0, the microprocessor continues with program instructions as soon as bit 2 of barrier register 0 is reset to 0.

In the second method, after the microprocessor sets bit 2 of barrier register 0 to 1, the microprocessor enables a hardware interrupt. The microprocessor may then issue program instructions that are not associated with the barrier. After receiving a 1 from the fan-out circuitry, the support circuitry resets bit 2 of barrier register 0 to 0 and sets the hardware interrupt. This interrupt indicates to the microprocessor that all of the microprocessors have reached the barrier.

Each of the barrier synchronization circuits may also be used for eureka synchronization. Eureka synchronization uses a point in program instructions where a microprocessor is informed when the first microprocessor associated with the eureka has finished its part of the program instructions. Eureka synchronization functions like a global logical OR operation.

Eureka synchronization has several uses, including database searches. Using eureka synchronization, a programmer can stop a database search as soon as any microprocessor finds the data rather than waiting for all of the microprocessors to exhaust the search.

When used for eureka synchronization, a barrier synchronization circuit operates differently than when used for barrier synchronization. Before the eureka synchronization begins, each microprocessor associated with the eureka sets the appropriate bit of one of the barrier registers to a 1. For example, each microprocessor may set bit 2 of barrier register 0 to 1.

As soon as bit 2 of barrier register 0 in each PE is set to 1, the eureka synchronization begins (this event is usually controlled with a separate barrier synchronization operation). When a microprocessor finishes the program instructions associated with the eureka, the microprocessor resets bit 2 of barrier register 0 to 0. Because all of the inputs to the barrier synchronization circuit are not 1, the output of the final AND gate resets to 0.

When the output of the final AND gate is 0, the fan-out circuitry sends a 0 to the support circuitry in each PE. The support circuitry then resets bit 2 of barrier register 0 and, if enabled, sets the microprocessor hardware interrupt. This signals the microprocessor that the eureka synchronization is complete.

Each barrier synchronization circuit receives an input from all of the PEs in the CRAY T3D system; however, not all of the PEs in the system must participate in the same barrier or eureka operation. The support software running in the host system may divide a barrier synchronization circuit into smaller circuits that each receive inputs from a limited number of PEs. This is done so the smaller barrier synchronization circuits more closely match user partitions.

Each AND gate in the AND tree is paired with a fan-out block in the fan-out tree. An AND gate and fan-out block pair is called a bypass point (refer again to Figure 3-13). The support software can redirect the output of an AND gate in a bypass point so that the output of the AND gate connects to the fan-out block in the bypass point. For example, Figure 3-14 shows a bypass point when the output of the AND gate is not redirected to the fan-out block and when the output of the AND gate is redirected to the fan-out block.

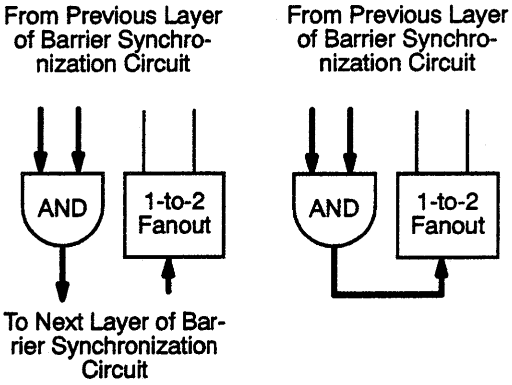

From Previous Layer of Barrier Synchronization Circuit
From Previous Layer of Barrier Synchronization Circuit
AND
1-to-2 Fanout
AND
1-to-2 Fanout
To Next Layer of Barrier Synchronization Circuit

Figure 3-14. Bypass Points

By redirecting the output of AND gates in different level bypass points, the support software may divide a physical barrier synchronization circuit into a combination of barrier partitions. Because the number and shape of the PEs in a barrier partition may not exactly match the number and shape of the PEs in a user partition, the barrier synchronization circuit may be further partitioned by software using a barrier mask register.

Using the barrier mask register and the bypass points, the support software may match a barrier partition to a user PE node partition so that all the PEs in a user partition use the same barrier synchronization circuitry. For more information on user partitions, refer again to "Virtual PE Number" at the beginning of this section.

Fetch and Increment

The support circuitry also performs read or write operations to the fetch-and-increment registers. A fetch-and-increment register is a special register where after reading information from a fetch-and-increment register, hardware in the PE node automatically increments the contents of the register by one (refer to Figure 3-15).

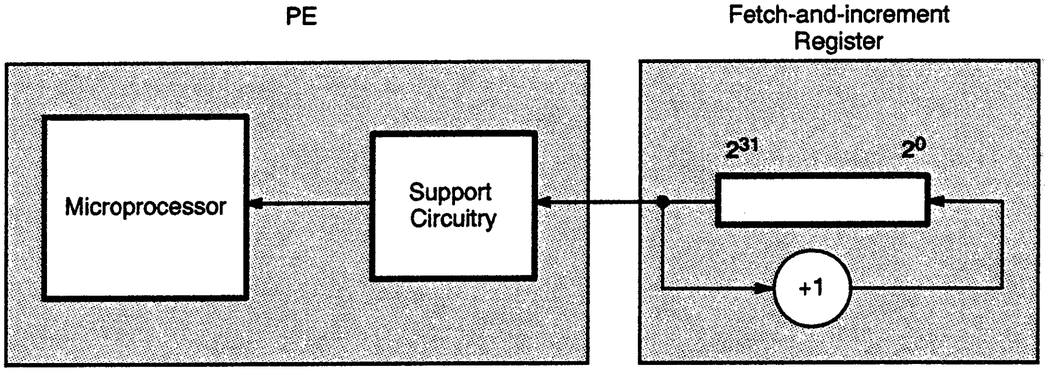


Figure 3-15. Fetch-and-increment

Each processing element node contains two fetch-and-increment registers, which are each 32 bits in size. This size is large enough to contain a loop index value or an address offset value.

Although each processing element node contains two fetch-and-increment registers, the registers function independently of the PEs. Any PE may use any of the fetch-and-increment registers in a partition.

The fetch-and-increment register may be used to dynamically distribute independent iterations of a program loop to more than one microprocessor. For example, four independent iterations of a program loop may be distributed among four different microprocessors.

Status

Each PE contains registers that indicate the status of PE operations. The status information includes error information and outstanding request information. The status information is used by the operating system.

BLT

The block transfer engine (BLT) is an asynchronous direct memory access device that redistributes system data. The BLT redistributes system data between globally addressable system memory and local memory in either of the PEs in a processing element node. The BLT can create up to 65,536 packets that contain one 64-bit word of data or up to 65,536 packets that contain four 64-bit words of data without interruption from the PE.

The BLT performs four types of data transfer operations: constant stride read, constant stride write, gather, and scatter. A constant stride read operation transfers data from fixed increment address locations in system memory to fixed increment address locations in local memory. A constant stride write operation transfers data from fixed increment address locations in local memory to fixed increment address locations in system memory.

A gather operation transfers data from nonsequential memory locations in system memory to fixed increment address locations in local memory. A scatter operation transfers data from fixed increment address locations in local memory to nonsequential address locations in system memory.

The BLT receives initial transfer parameters from one of the PEs in a processing element node and then functions independently from the PEs. The BLT contains three main components: system addressing, local addressing, and control.

The system addressing portion of the BLT generates addresses that are used to reference a location in system memory (which includes local or remote memory). For example, the system addressing circuitry may generate addresses that point to a word of read data in a remote PE during a BLT constant stride read or gather operation.

The local addressing portion of the BLT generates addresses that are used to reference a location in the local PE. For example, the local addressing circuitry may generate addresses that point to the location in local memory where data from a BLT read response packet will be stored.

The control portion of the BLT controls the BLT transfer. In addition, the control circuitry provides an interrupt to the PE when an error occurs, when the BLT is free to start a transfer, and when the BLT transfer is complete.

Network Interface

The network interface assembles outgoing request and response packets and steers incoming request and response packets to the correct PE in the node. The network interface also contains the fetch-and-increment registers and barrier synchronization bypass point control.

When assembling an outgoing request or response packet, the network interface receives packet header information from PE 0, PE 1, or the BLT. In addition, the network interface receives a virtual or logical PE number from a PE or the BLT and may receive data from a PE.

If the PE number is a virtual PE number, the network interface converts the virtual PE number into a logical PE number. The network interface then converts the logical PE number into a routing tag that is used in the outgoing packet.

When receiving an incoming packet, the network interface checks the destination node number in the packet header. If the destination node number is the same as the number of the node that the network interface is in, the packet arrived at the correct node. The network interface then sends the packet to the destination PE. If the destination node number is not correct, the network interface converts the packet into an error message and sends the error message to one of the PEs in the node.

4 I/O GATEWAYS

All input and output communication between the CRAY T3D system and the host system is performed through the I/O gateways. As was described in Section 1, "Overview," each I/O gateway contains an input node, an output node, and LOSP circuitry (refer to Figure 4-1).

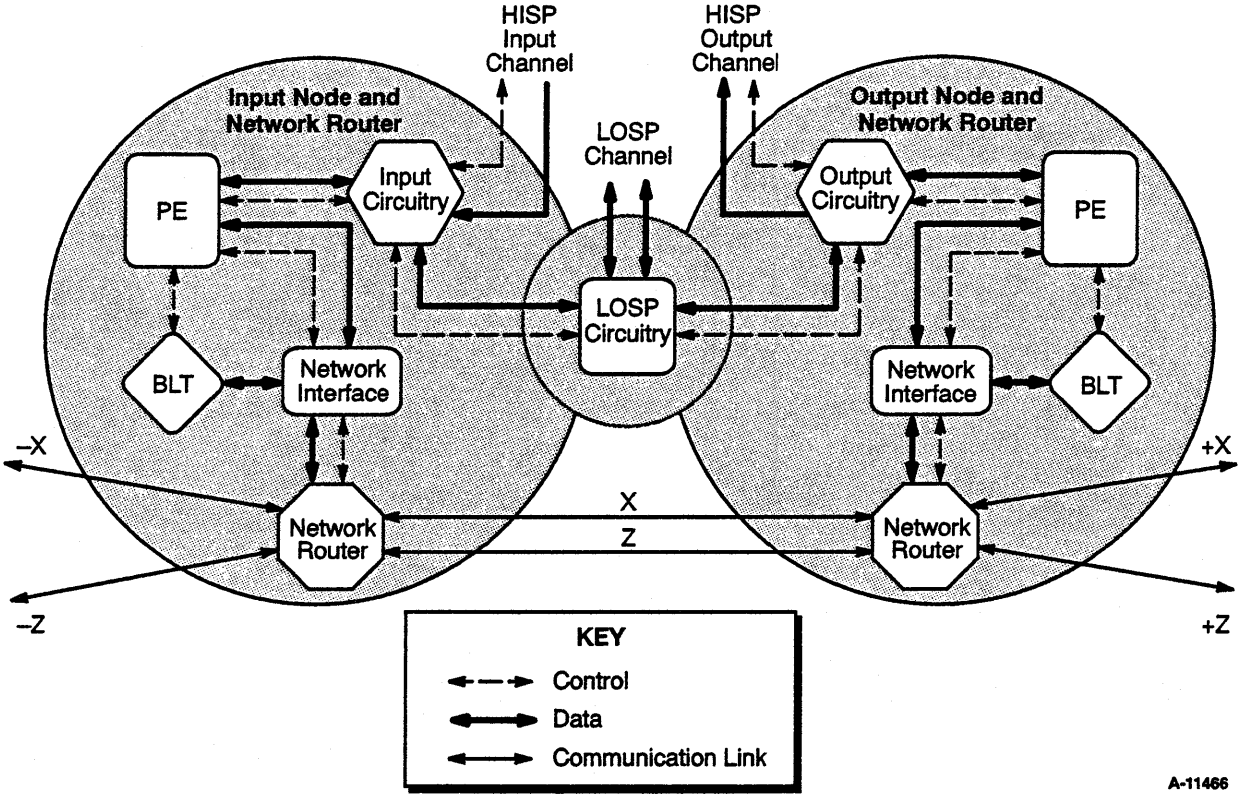


Figure 4-1. I/O Gateway

This section describes LOSP and HISP channels, physical I/O gateways, and the master and slave I/O gateways.

LOSP Channel

LOSP channels transfer request and response information between the CRAY T3D system and the host system or an input/output cluster (IOC). Either the CRAY T3D system or the host system (or IOC) can initiate a transfer of information over the LOSP channel.

Each LOSP channel is actually a pair of unidirectional channels. Figure 4-2 shows the signals used in a LOSP channel.

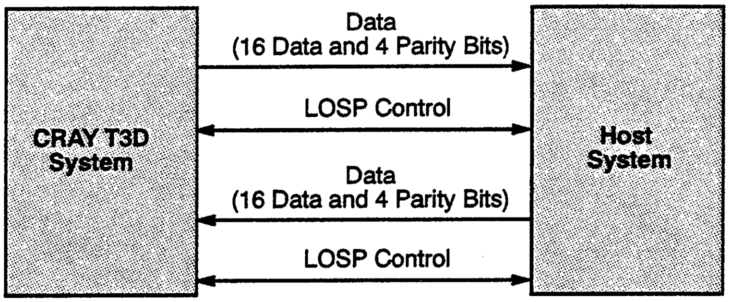

CRAY T3D System
Data (16 Data and 4 Parity Bits)
LOSP Control
Data (16 Data and 4 Parity Bits)
LOSP Control
Host System

Figure 4-2. LOSP Channel Signals

LOSP data is transferred over the LOSP channel in 16-bit parcels. The data contains information used to control the HISP channel. The 16-bits of data are protected by 4 parity bits that are used to check the data for errors.

The data transfer rate of a LOSP channel is 6 Mbytes/s in each direction. The LOSP channel uses control signals that indicate when data is on the channel, when data is received, and when a transfer of data is finished.

When information transfers over the LOSP channel from the host system (or IOC) to the CRAY T3D system, the most significant bit of the first parcel transferred directs the information to the appropriate node. If this bit is 0, the information is for the output node. If this bit is 1, the information is for the input node.

When information transfers over the LOSP channel from the CRAY T3D system to the host system (or IOC), the input node and output node share the LOSP channel. The first node to request a transfer over the LOSP channel controls the channel until a disconnect is sent.

HISP Channel

HISP channels transfer system data between the CRAY T3D system and the host system (or IOC). The HISP channel connects two components: a master and a slave. The master controls the HISP channel by providing address information to the slave.

The data transfer rate of a HISP channel is 200 Mbytes/s in each direction; however, by modifying parameters in the I/O gateway memory-mapped registers, software may change the HISP channel transfer rate to 100 Mbytes/s. This enables the CRAY T3D system to operate with a host system that uses either 100 Mbytes/s or 200 Mbytes/s HISP channel protocol. Figure 4-3 shows the signals used in a HISP channel.

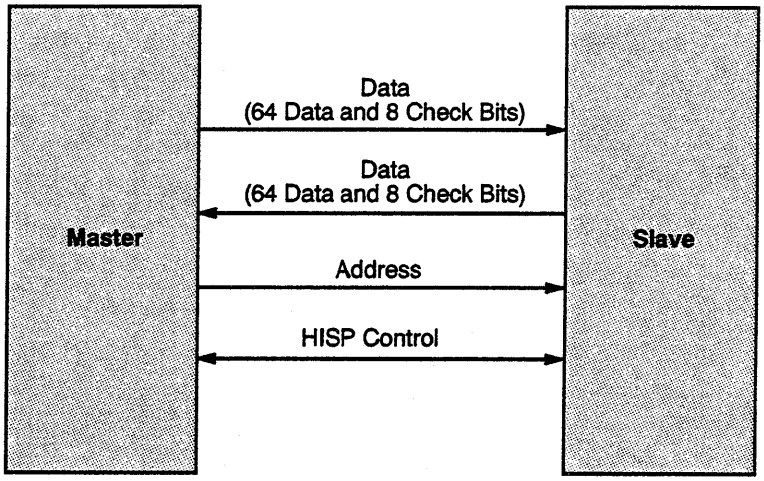


Figure 4-3. HISP Channel Signals

System data transfers between the master and slave in 64-bit words. The 64-bits of data are protected by 8 check bits used to check the data for errors and correct any single-bit errors.

Address and block length information is sent from the master to the slave. The address contains information on where data will be stored or read in the slave's memory. The block length indicates the total number of words that will be transferred.

HISP protocol uses control signals that clear the HISP channel, control when a data transfer starts, and indicate when the last 64-bit word of data is transferring over the HISP channel. Error signals are also sent to indicate whether a data error occurred during the transfer.

Physical I/O Gateways

Physically, each I/O gateway resides on one circuit board in the CRAY T3D system cabinet (refer to Figure 4-4). The integrated circuits to the left of the dashed line are used for components of the input node. The integrated circuits to the right of the dashed line are used for components of the output node.

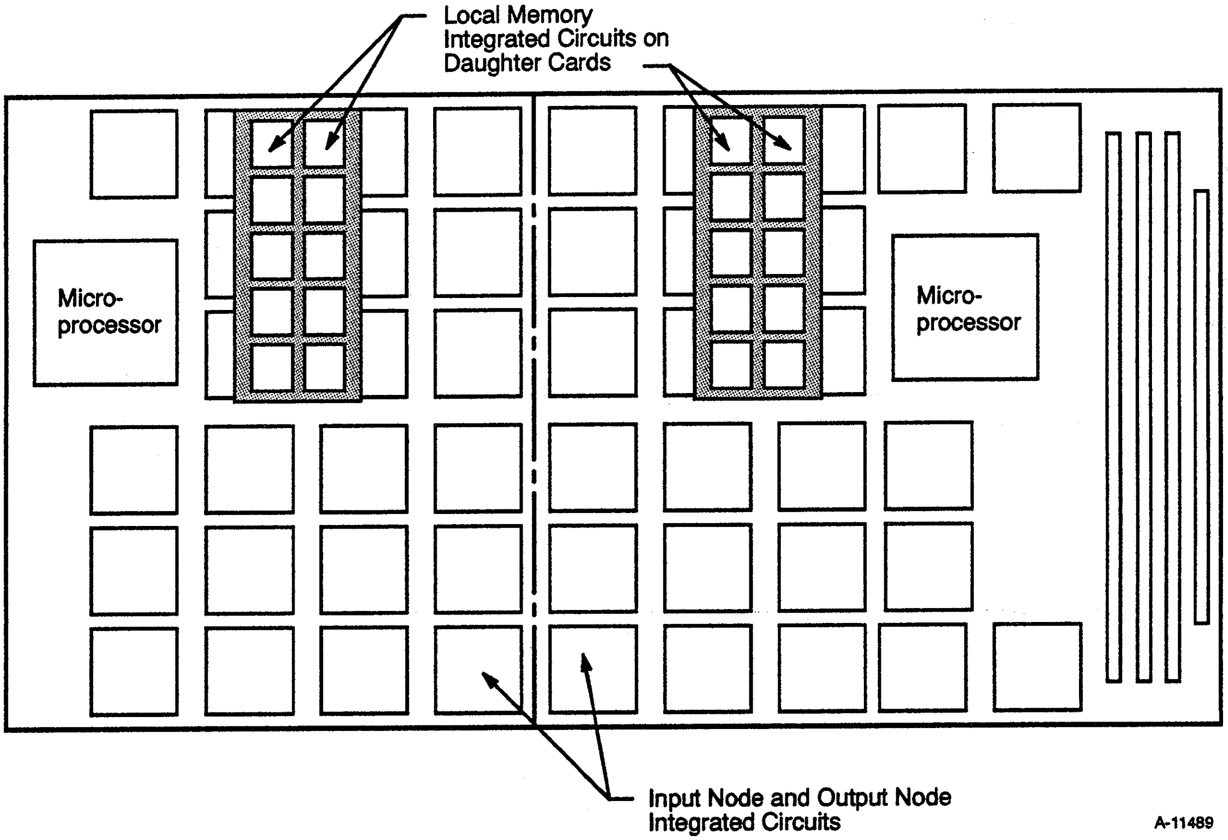

Local Memory Integrated Circuits on Daughter Cards
Micro-processor
Micro-processor
Input Node and Output Node Integrated Circuits
A-11489

Figure 4-4. I/O Gateway Circuit Board

I/O Gateway Configurations

The I/O gateways connect the CRAY T3D system to the host system in three types of cabling configurations: phase 1, phase 2, and phase 3. Tape, network, or other non-disk device I/O is managed by the host system. Disk I/O is managed by the host system for phases 1 and 2, and is managed by the CRAY T3D system in phase 3.

Phase 1 Configuration

Phase 1, which is provided with initial CRAY T3D systems, connects a master I/O gateway to the host system over the HISP channel. All configurations of the CRAY T3D system must have at least one phase 1 channel configuration (at least one master I/O gateway).

The HISP and LOSP channels from the master I/O gateway connect to the circuitry on a CPU module or shared I/O module in the host system. Figure 4-5 shows the phase 1 channel configuration of a master I/O gateway.

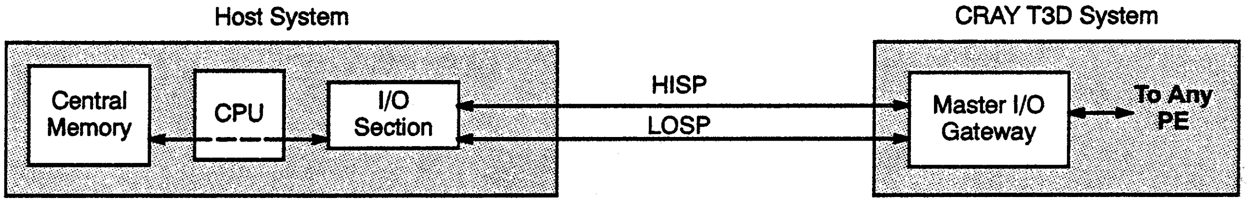


Figure 4-5. Phase 1 I/O Gateway Channel Configuration

When the CRAY T3D system connects to the host system through a master I/O gateway, the CPU in the host system controls all input and output. For example, the CPU controls tape, network, and disk device input and output. Data passes through the CPU and to the master I/O gateway.

Phase 2 Configuration

Phase 2, which will be available in the first half of 1994, connects a slave I/O gateway to an IOC that is also connected to a CPU. Figure 4-6 shows the phase 2 channel configuration of a slave I/O gateway.

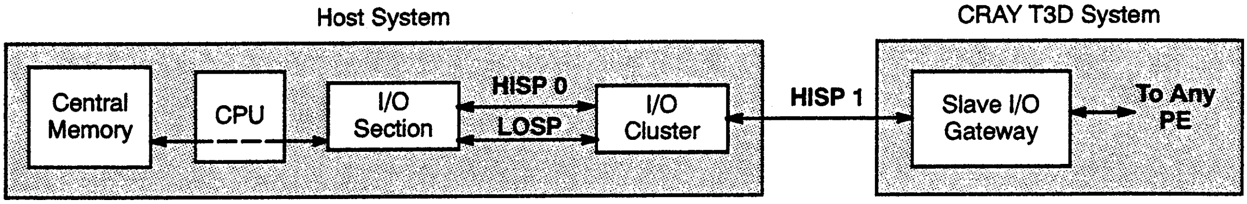


Figure 4-6. Phase 2 Slave I/O Gateway Channel Configuration

When the CRAY T3D system connects to the host system using the phase 2 channel configuration, the CPU in the host system controls all input and output. For example, the CPU controls tape, network, and disk device I/O.

Although the CPU controls the input and output, a HISP data path connects the CRAY T3D system to the IOC. This provides a path for data to travel between the CRAY T3D system and disk devices without traveling through the circuitry on a CPU module.

The phase 2 configuration uses the back-door HISP software support that is also used for the SSD solid state storage device model E (SSD-E). If the slave I/O gateway is connected to an IOC in this configuration, the SSD-E in the host system cannot be configured with back-door capability to the same IOC.

Phase 3 Configuration

Phase 3, which will be available in 1995, connects a slave I/O gateway to an IOC. Figure 4-7 shows the phase 3 configuration of a slave I/O gateway.

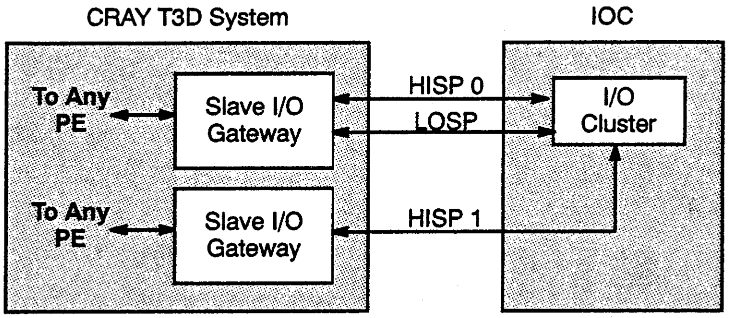


Figure 4-7. Phase 3 I/O Gateway Channel Configuration

When the CRAY T3D system connects to an IOC using a phase 3 channel configuration, the CRAY T3D system controls disk device input and output. In addition, a HISP data path connects the CRAY T3D system to the IOC.

GLOSSARY

A

ACK

An ACK is a special packet sent by a destination PE to the source PE that indicates the destination PE accepted a message sent by the source PE.

Atomic Swap

An atomic swap is an operation that transfers a 64-bit word of data from memory to the microprocessor, then stores a new word into the same memory location in an indivisible operation.

B

Barrier Synchronization

Barrier synchronization is an event initiated by software that prevents the PEs in a partition from continuing to issue new program instructions until all of the PEs in a partition have reached the same point in the program. The CRAY T3D system contains hardware that provides a low-latency method of performing barrier synchronization.

Block Transfer Engine (BLT)

The block transfer engine is an asynchronous direct memory access controller that redistributes system data. The BLT redistributes system data between the local memory in PE 0 or PE 1 and globally addressable system memory..

Body

The body is the part of a packet that contains data.

Bypass Point

A bypass point is a component of the barrier synchronization circuitry that contains one AND gate and one fan-out block.

C

Communication Link

A communication link is a component of the interconnect network that transfers data and control information between two network routers in the interconnect network. A communication link is actually two unidirectional channels (in one dimension).

Cycle Request

Cycle request information is control signals that the microprocessor provides to the support circuitry to indicate what type of instruction is being issued.

D

Data Cache Memory The data cache memory is a small, high-speed, random access memory in each microprocessor that temporarily stores frequently or recently accessed data.

Data Prefetch A data prefetch is an operation that transfers a 64-bit word of data from memory to the data prefetch queue.

Data Prefetch Queue The data prefetch queue is a set of hardware registers that are located in the support circuitry. The data prefetch queue stores up to 16 64-bit words of data.

Dateline The dateline is a single communication link in each dimension that is designated by software as the dateline communication link. The dateline communication link is used to indicate which network packets will use a second set of virtual channels to avoid communication deadlock conditions.

Daughter Card A daughter board is a small printed circuit board that connects to a larger printed circuit board via pin connectors. In the CRAY T3D system, local memory is located on daughter cards.

Deadlock Condition A deadlock condition occurs when two components that use the same set of channels cannot communicate because of unresolvable channel conflicts.

Destination PE The destination PE is the PE that receives a request. For example, the destination PE contains the read data for a read request.

Dimension Order Routing Dimension Order Routing is a process of steering information through the interconnect network where the information travels through the X dimension first, then through the Y dimension, and finally through the Z dimension.

Data Translation Buffer (DTB) The DTB is a component in the microprocessor that is used to translate addressing information.

DTB Annex The DTB annex is a component of the support circuitry that extends the functions of the data translation buffer (DTB) in the microprocessor. The DTB annex is a 32-entry table. Each entry in the table contains a virtual or logical PE number and a function code.

E

Eureka A eureka is an event initiated by software that indicates when at least one of the PEs in a partition has reached a specified point in a program.

F

Fetch-and-increment A fetch-and-increment is an operation that transfers a 32-bit halfword from a fetch-and-increment register to the microprocessor and then increments the value of the fetch-and-increment register by one.

Fetch-and-increment Register A fetch-and-increment register is a hardware register that contains a 32-bit value. Each time the fetch-and-increment register is read, hardware automatically increments the value stored in the fetch-and-increment register by one.

Folded Torus Network A folded torus network consists of nodes that are physically placed so that the maximum wiring distance between the nodes is minimized. This type of network is also referred to as an interleaved network.

Function Code The function code is information stored in the DTB annex that indicates what type of memory function (read, write, etc.) the support circuitry will perform.

G

Globally Addressable Memory Globally addressable memory is a type of memory where any location in memory can be referenced using the same address format by any microprocessor in the system.

Global Memory Global memory is the total amount of system memory in the CRAY T3D system and is equal to the size of local memory multiplied by the total number of PEs in the system (not including the PEs in the I/O gateways).

H

Header The header is the part of a packet that contains control and address information.

High-speed (HISP) Channel A HISP channel is a channel that transfers system data between the CRAY T3D system and the host system. The HISP channel operates an 100 Mbytes/s or 200 Mbytes/s.

Hop A hop is a transfer of information over a single communication link.

Host System The host system is a Cray Research, Inc. computer system that provides support for applications running on the CRAY T3D system.

I

Input Node An input node is a component of an I/O gateway that controls the transfer of information from the host system to the CRAY T3D system over the HISP input channel.

Input/Output Cluster (IOC) An input/output cluster is a component of an input/output subsystem model E (IOS-E) that transfers system data between computer systems or peripheral devices.

Instruction Cache Memory The instruction cache memory is a small, high-speed random access memory in each microprocessor that temporarily stores frequently or recently accessed instructions.

Interconnect Network The interconnect network is a three-dimensional matrix of channels that provide communication paths among the processing element nodes and I/O gateways in the CRAY T3D system.

Interleaving Interleaving is the physical placement of nodes so that the maximum wiring distance between nodes is minimized.

I/O Gateway An I/O gateway is a component of the CRAY T3D system that transfers system data and control information between the host system and any PE in the CRAY T3D system or between any PE in the CRAY T3D system and an input/output cluster (IOC).

L

Local Memory Local memory is memory that is physically located near a microprocessor in a processing element. Each PE contains a low-latency, high-bandwidth data path that connects the microprocessor to local memory.

Logical PE Number Logical PE numbers are numbers assigned to all of the PEs in a CRAY T3D system that run applications. The logical numbering of PEs enable spare PEs to replace failing PEs.

Logically Shared Memory Logically shared memory is a memory system that enables any microprocessor in the system to access the memory in another PE without involving the microprocessor in that PE.

Low-speed (LOSP) Channel A LOSP channel is a channel that transfers request and response information between the CRAY T3D system and the host system. Losp channels operate at 6 MBytes/s.

M

Massively Parallel Processor (MPP)

An MPP computer system contains hundreds or thousands of microprocessor that are each accompanied by a local memory and communicate using an interconnect network.

Master I/O Gateway

The master I/O gateway controls a HISP channel and sends address information to the host system during a HISP transfer.

Memory-mapped Registers

Memory-mapped registers are registers that are physically located in the components of the processing element node; however, the microprocessor addresses the registers as if they were located in a local memory. Memory-mapped registers are used to obtain system status information and set function parameters.

Message

A message is a special type of packet that the support circuitry in the receiving PE stores in the message queue. After storing a message in the message queue, the support circuitry interrupts the microprocessor and the microprocessor reads the message from the message queue.

Message Queue

The message queue is a portion of local memory that is reserved for message packets.

Multiple-cabinet Configuration

A configuration of the CRAY T3D system that contains CRAY T3D system modules and the host system modules in separate cabinets.

N

NACK

A NACK is a special packet sent by a destination PE to the source PE that indicates the destination PE did not accept a message sent by the source PE.

Network Interface

The network interface is a component of a processing element node that formats data and control information into a packet before it is sent through the interconnect network. The network interface also buffers incoming packet information from the interconnect network.

Network Router

The network router is a component of the interconnect network that receives and sends packets through the communication links in the interconnect network.

Node

A node is a point in the interconnect network where information can transfer from one communication link to another communication link.

Noncacheable

Noncacheable is a type of operation that transfers data out of or into the microprocessor, but does not read from or write to the data cache in the microprocessor.

O

Output Node An output node is a component of an I/O gateway that controls the transfer of information from the CRAY T3D system to the host system over the HISP output channel.

P

Packet A packet contains information that is transferred through the interconnect network. Each packet has a header, which contains control and addressing information, and may have a body, which contains data. All information is transferred through the network in the form of a packet.

Partition A partition is a group of PEs and a portion of the barrier synchronization resources that are assigned to one application.

Physical PE Number The physical PE number is a number that is assigned by hardware to each physical PE in the CRAY T3D system and indicates where the PE is physically located in the system.

Physically Distributed Memory Physically distributed memory is a memory system that is divided into physical segments where each segment is located in a different PE.

Processing Element (PE) A processing element is a component of the CRAY T3D system that contains a microprocessor, local memory, and support circuitry.

Processing Element Node A processing element node is a component of the CRAY T3D system that contains two processing elements, a block transfer engine, and a network interface.

R

Read Ahead A read ahead is an operation that retrieves a 4-word block of data from local memory, sends the block of data to the microprocessor, and then buffers the next sequential 4-word block of data in the local memory read stage of the support circuitry.

Request A request is information that requests a PE to perform an activity. For example, a request may be to read data from the memory in a PE.

Response A response is information that is the result of an activity. For example, a response to a read request is the read data.

Routing Tag The routing tag is part of a packet header that indicates the path a packet will follow when traveling through the interconnect network.

S

Shared Memory Shared memory is the total amount of system memory in the CRAY T3D system and is equal to the size of local memory multiplied by the total number of PEs in the system (not including the PEs in the I/O gateways).

Single-cabinet Configuration A configuration of the CRAY T3D system that contains the host system modules and the CRAY T3D system modules in the same cabinet.

Slave I/O Gateway The slave I/O gateway receives address information from the host system, which is controlling the HISP channel.

Source PE The source PE is a PE that generates a request. For example, the source PE may requests read data from a destination PE.

Spare PE A spare PE is a PE that is physically connected in the interconnect network but is not part of the software-configured logical CRAY T3D system. Spare PEs are used to replace failing PEs and prevent prolonged system down time.

Support Circuitry The support circuitry is a component in a PE that extends the control and addressing functions of the microprocessor in the PE.

Swaperand Register The swaperand register is a hardware register that stores a 64-bit word of data that will be written into memory during an atomic swap operation.

Switch A switch is a component of the network router that controls the flow of packets through the communication links in one dimension only.

T

Torus A torus is a type of interconnect network connection where the largest numbered node in a dimension is connected to the smallest numbered node in the dimension. This type of connection forms a ring where information can transfer from one node, through all of the nodes in the same dimension, and back to the original node.

V

Virtual Channel A virtual channel is a channel that is created when request and response information travels over the same physical channel but is stored in different buffers.

Virtual PE Number The virtual PE number is a number that an application uses to reference each PE in a partition.

W

Word A word contains 64-bits of information.

X

X-dimension Switch The X-dimension switch is a component of the network router that controls the flow of packets through the communication links in the X dimension only.

Y

Y-dimension Switch The Y-dimension switch is a component of the network router that controls the flow of packets through the communication links in the Y dimension only.

Z

Z-dimension Switch The Z-dimension switch is a component of the network router that controls the flow of packets through the communication links in the Z dimension only.

Reader Comment Form

Title: CRAY T3D System Architecture Overview **Number: HR-04033**

Your feedback on this publication will help us provide better documentation in the future. Please take a moment to answer the few questions below.

For what purpose did you primarily use this manual?

_______Troubleshooting
_______Tutorial or introduction
_______Reference information
_______Classroom use
_______Other - please explain ______________________________

Using a scale from 1 (poor) to 10 (excellent), please rate this manual on the following criteria and explain your ratings:

_______Accuracy ______________________________
_______Organization ______________________________
_______Readability ______________________________
_______Physical qualities (binding, printing, page layout) ______________________________
_______Amount of diagrams and photos ______________________________
_______Quality of diagrams and photos ______________________________

Completeness (Check one)

_______Too much information ______________________________
_______Too little information ______________________________
_______Just the right amount of information

Your comments help Hardware Publications and Training improve the quality and usefulness of your publications. Please use the space provided below to share your comments with us. When possible, please give specific page and paragraph references. We will respond to your comments in writing within 48 hours.

NAME ______________________________
JOB TITLE______________________________
FIRM______________________________
ADDRESS______________________________
CITY__________________STATE________ZIP__________
DATE______________________________

[or attach your business card]

CRAY RESEARCH, INC.

CUT ALONG THIS LINE

Fold

NO POSTAGE
NECESSARY
IF MAILED
IN THE
UNITED STATES

BUSINESS REPLY CARD

FIRST CLASS PERMIT NO 6184 ST. PAUL, MN

POSTAGE WILL BE PAID BY ADDRESSEE

CRAY
RESEARCH, INC.

Attn: Hardware Publications and Training
890 Industrial Boulevard
Chippewa Falls, WI 54729

Fold

STAPLE

Cray Research, Inc.
Hardware Publications and Training
890 Industrial Boulevard
Chippewa Falls, WI 54729