

20 MAR 2000

1

DYNAMIC I/O ALLOCATION IN A PARTITIONED COMPUTER SYSTEM**Background of the Invention**1. Field of the Invention

The present invention generally relates to multiprocessor computer systems, and more particularly to resource allocation among processors in a partitioned multiprocessor system. Still more particularly, this invention relates to methods for sharing a common set of I/O devices among the nodes of a multiprocessor computer system.

2. Description of the Prior Art

Multiprocessor computer systems are well known in the art, and provide for increased processing capability by allowing processing tasks to be divided among several different system processors. In conventional systems, each processor is able to access all of the system resources; i.e., all of the system resources, such as memory and I/O devices, are shared between all of the system processors. Typically, some parts of a system resource may be partitioned between processors, e.g., while each processor will be able to access a shared memory, this memory is divided such that each processor has its own workspace. In a Non-Uniform Memory Access (NUMA) system, each processor has its own memory, and can also access memory owned by other processors.

More recently, symmetric multiprocessor (SMP) systems have been partitioned to behave as multiple independent computer systems. For example, a single system having eight processors might be configured to treat each of the eight processors (or multiple groups of one or more processors) as a separate system for processing purposes. Each of these

"virtual" systems would have its own copy of the operating system, and may then be independently assigned tasks, or may operate together as a processing cluster, which provides for both high-speed processing and improved reliability. Typically, in a multiprocessor system, there is also a "service" processor, which manages the startup and operation of the overall system, including system configuration and data routing on shared buses and devices, to and from specific processors.

When several virtual systems, in a single multiprocessor system, are configured to operate as a cluster, software support must be provided to allow each cluster node to communicate with each other node in the multiprocessor system to perform quorum negotiation and validation, send "heartbeats," and perform other quorum functions using any cluster communication technique. When this is accomplished, if one of the processors fails, which would cause that node to become unavailable to the cluster, the jobs assigned to that node can be reassigned among the remaining processors (nodes), using standard cluster techniques.

Typically, when a multiprocessor system is divided into multiple virtual systems, each of the virtual systems has its own copy of the operating system, and the same operating system is used for each virtual system. Since each processor is running the same operating system, it has been relatively easy to provide for resource allocation among the processors.

One characteristic of large multiprocessor systems is that since they are typically used for large processing jobs, they have relatively little use for typical I/O devices such as keyboards, displays, removable media drives, etc. These devices cannot be removed, however, since there

may be occasions, however infrequent, when they are needed. Making these devices available within each of the nodes in a large multiprocessor system results in an expensive duplication of these seldom-used devices, and creates an unnecessary burden in managing and maintaining the equipment. Thus, there is a need for a means for several partitions or nodes of a large multiprocessor system to share a single set of I/O devices.

Summary of the Invention

It is therefore one object of the present invention to provide a system and method for the operation of a multiprocessor computer system.

It is another object of the present invention to provide a system and method for improved resource allocation within a multiprocessor computer system.

It is yet another object of the present invention to provide a system and method for sharing a common set of I/O devices among the nodes of a multiprocessor computer system.

Thus, there is provided a system and method for allowing multiple nodes of a multiprocessor system to share a set of I/O devices. A Cabinet Input/Output Controller (CI/OC) is provided which manages communications between the multiprocessor system nodes and the common I/O devices, allowing individual nodes to access one or more of its target devices exclusively. Each of the nodes communicates with the CI/OC via a service processor, and the CI/OC interconnects the various I/O devices and a node's USB controller. In an alternate embodiment, a USB to ISA bridge is also included to allow attachment of legacy I/O devices.

The above as well as additional objectives, features, and advantages of the present invention will become apparent in the following detailed written description.

Brief Description of the Drawings

The novel features believed characteristic of the invention are set forth in the appended claims. The invention itself, however, as well as a preferred mode of use, further objectives, and advantages thereof, will best be understood by reference to the following detailed description of an illustrative embodiment when read in conjunction with the accompanying drawings, wherein:

Figure 1 is a multiprocessor computer system in accordance with a preferred embodiment of the present invention;

Figure 2 depicts an SMP node connected to a CI/OC in accordance with a preferred embodiment of the present invention;

Figure 3 is a block diagram of a CI/OC in accordance with a preferred embodiment of the present invention;

Figure 4 depicts a block diagram of a CI/OC and service processor in accordance with a preferred embodiment of the present invention;

Figure 5 is a block diagram of several I/O devices connected to a hub, downstream of a CI/OC, in accordance with a preferred embodiment of the present invention; and

Figure 6 depicts a flowchart of the use of a common I/O system in accordance with a preferred embodiment of the present invention.

Description of the Preferred Embodiment

The preferred embodiment provides a Cabinet Input/Output Controller (CI/OC) which allows multiple nodes in a multiprocessor system to share common I/O devices. In the preferred embodiment, other run-time, performance related I/O such as an ethernet adapter and the associated network connection will continue to be present in each node.

With reference now to the figures, and in particular with reference to **Figure 1**, a block diagram of a data processing system in which a preferred embodiment of the present invention may be implemented is depicted. Data processing system 100 may be, for example, one of the server models of computers available from International Business Machines Corporation of Armonk, New York. Data processing system 100 includes processors 101 and 102, which in the exemplary embodiment are each connected to level two (L2) caches 103 and 104, respectively, which are connected in turn to a system bus 106.

Also connected to system bus 106 is system memory 108 and Primary Host Bridge (PHB) 122. PHB 122 couples I/O bus 112 to system bus 106, relaying and/or transforming data transactions from one bus to the other. In the exemplary embodiment, data processing system 100 includes graphics adapter 118 connected to I/O bus 112, receiving user interface information for display 120. Peripheral devices such as nonvolatile storage 114, which may be a hard disk drive, and keyboard/pointing device 116, which may include a conventional mouse, a trackball, or the like, are connected via an Industry Standard Architecture (ISA) bridge 121 to

I/O bus 112. PHB 122 is also connected to PCI slots 124 and Universal Serial Bus controller 126 via I/O bus 112.

The exemplary embodiment shown in **Figure 1** is provided solely for the purposes of explaining the invention and those skilled in the art will recognize that numerous variations are possible, both in form and function. For instance, data processing system 100 might also include a compact disk read-only memory (CD-ROM) or digital video disk (DVD) drive, a sound card and audio speakers, and numerous other optional components. All such variations are believed to be within the spirit and scope of the present invention. Data processing system 100 and the CI/OC architecture examples below are provided solely as examples for the purposes of explanation and are not intended to imply architectural limitations.

Referring now to **Figure 2**, a node 200 is shown which can be used as a building block for a large multiprocessor system. Node 200 comprises SMP processors 202 and associated memory 204 (which can be shared with other processors). SMP processors 202 are connected to Primary Host Bridge (PHB) 206, which is then connected to Universal Serial Bus (USB) controller 208 and PCI slots 210. Connected to (or, typically, plugged into) PCI slots 210 are I/O devices 212, which, in this embodiment, are not shared with other nodes. USB controller 208 is the preferred node input connection for CI/OC 216, which is controlled by Service Processor (SP) 214. CI/OC 216 is connected to other nodes via node inputs 218, and allows them to share I/O devices 220. It should be noted that although this diagram only shows details of one exemplary node, node 200 and other

nodes 218 are each connected to CI/OC 216 by identical node inputs of the CI/OC 216.

The large multiprocessor system can be arranged as a number of smaller, independent partitions or arranged as NUMA or cluster. If the large multiprocessor system is arranged as NUMA or cluster, the optional Node Interconnect hardware 222 may be used to obtain the desired interconnect configuration. The CI/OC 216 of the preferred embodiment is used to control and interconnect the nodes with a single, common collection of devices. For example, a large multiprocessor system housed in a single rack would contain multiple computer nodes but the entire rack would only require a single operator terminal, diskette drive, etc. Preferably, a Service Processor (SP) 214 manages the CI/OC configuration. Each node contains a USB controller 208.

With reference now to **Figure 3**, a high-level view of a CI/OC according to the preferred embodiment is depicted. Here, CI/OC 300 is shown connected to its service processor 302. Exemplary node inputs 306 and 308 are shown, which allow connection of nodes such as shown in **Figure 2**. The CI/OC 300 is also connected to common I/O devices 304. Preferably there is only one CI/OC 300 and one SP 302 per large multiprocessor system, and all the nodes share the common I/O devices 304 connected downstream of the CI/OC 300. The number of node inputs on the CI/OC 300 is implementation dependent.

Referring now to **Figure 4**, a more detailed view of the CI/OC 400 is shown as comprising a series of switches 408 for connecting node inputs 406 to the downstream I/O devices 404. The SP 402 switches a maximum of one node onto the common I/O devices connection 404. Because the preferred

embodiment provides that all connections are USB-compliant, the devices are hot-pluggable, i.e., they can be attached and removed from a node while that node is operating. Activating a switch 408 is the equivalent of an Attach of the downstream connection. Deactivating a switch is the equivalent of a Removal of the downstream connection. The SP 402 will connect the downstream devices 404 to the node inputs 406 as they are needed.

With reference now to Figure 5, the common I/O system which contains the base I/O used to enable the common I/O connectivity, as well as a variety of exemplary I/O devices, is shown. In this figure, CI/OC 500 is connected to hub 502, which in the preferred embodiment is a USB hub. The hub 502 allows communications from the CI/OC to be passed to any attached devices. These devices can include a keyboard 504 and a mouse 506. If native USB devices are not available to meet the requirements, the I/O devices can include commercially available USB to ISA conversion logic 508. By doing so, legacy devices such as an ISA floppy drive 510, a serial port 512 and a parallel port 514 can be attached.

The Universal Serial Bus specification, which is available at <http://www.usb.org> (as of the filing date of this application) and hereby incorporated by reference, specifies that the USB transfers signal and power over four wires comprised of VBUS, GND, D+, and D-. The signaling occurs over two wires, D+ and D-. In the preferred embodiment, hub 502 is a powered hub which supplies power to attached devices, so that it does not require VBUS and GND from the node USB controller. In this embodiment, the interconnect from a node USB controller 208 to the CI/OC 216, as shown in Figure 2, and then to the hub 502, as shown

in Figure 5, consists of USB signals D+ and D-, with VBUS and GND omitted.

Referring now to Figure 6, a flowchart of the preferred operation of the common I/O system is shown. When a node requires a common I/O device (step 600), it sends a request to the service processor (step 610), requesting an Attach operation. If no other node is currently using the I/O channel (step 620), the SP switches the CI/OC to allow that node to communicate with the downstream devices (step 630). The node uses the I/O device (step 640), and when it is finished, instructs the SP to switch the CI/OC connection off, detaching the I/O devices (step 650). The node will then continue to operate normally (step 660).

If the I/O channel is in use by another node (step 620), the connection is denied (step 670). The node resumes normal operation, and may retry to establish the connection as many times as necessary. Because of the nature of the multiprocessor system, these sort of device conflicts will be relatively unusual, and more sophisticated arbitration techniques are not necessary, although they may be implemented by one of skill in the art.

Modifications and Variations

While the invention has been particularly shown and described with reference to a preferred embodiment, it will be understood by those skilled in the art that various changes in form and detail may be made therein without departing from the spirit and scope of the invention. For example, the CI/OC blocks can be cascaded in the event a large node count must be supported, so that each node is connected to one CI/OC, and can communicate via a chain of CI/OC devices to the peripheral devices. This, and other

modifications, are considered within the scope of the claims below.

**CLAIMS:**

What is claimed is:

1. A computer system, comprising
a plurality of system nodes, each having at least one system processor and an associated memory, said system nodes each connected to communicate over a respective device port;
a device controller, connected to at least one of said device ports; and
at least one peripheral device connected to said device controller,
wherein said system is configured to perform the steps of:
sending a request, from a system node, to said device controller, over a direct connection;
establishing, by said controller, an exclusive connection between said system node and said peripheral device; and
operating said peripheral device by said system node.
2. The system of claim 1, wherein said system nodes operate as symmetric multiprocessor systems.
3. The system of claim 1, wherein each of said system nodes share said peripheral device.
4. The system of claim 1, wherein said connections are according to the Universal Serial Bus specification.



5. The system of claim 1, wherein said system is also configured to perform the step of preventing a connection between a second system node and said external device when said system node and said external device are connected.

6. The system of claim 1, wherein said controller is connected to a plurality of computer systems, and said peripheral device is shared between said systems.

7. The system of claim 1, wherein said computer system is a rack-mounted system.

8. The system of claim 1, wherein when multiple peripheral devices are connected, all of said peripherals are connected to said system node during said establishing step.

9. A multiprocessor computer system, comprising:
a plurality of virtual computer systems, each system having at least one system processor and a memory connected to be written to and read from by said processor;
a common input/output controller, connected to said computer systems; and
at least one peripheral device, connected to said controller,
wherein said controller establishes an exclusive connection between one of said computer systems and said peripheral device, when requested by said computer system.

10. The system of claim 9, wherein said virtual computer systems are partitions in a symmetric multiprocessor system.



11. The system of claim 9, wherein said controller is connected to provide a connection to said peripheral devices for said plurality of virtual computer systems.

12. The system of claim 9, wherein each of said virtual computer systems is connected to said controller via a universal serial bus, and each peripheral is connected to said controller via a universal serial bus hub.

13. The system of claim 9, wherein each of said virtual computer systems comprises a universal serial bus controller.

14. The system of claim 9, wherein when multiple peripheral devices are connected, all of said peripherals are connected to said virtual system when said peripheral device is connected.

15. A device controller, comprising:
a communications controller;
a plurality of node inputs, each operatively connected to said communications controller; and
a plurality of device connections, each operatively connected to said communications controller,
wherein said communications controller arbitrates node ownership of said device connections, and provides an exclusive connection between a given node input and all of said device connections when that node input is granted ownership of the device connections.

16. The device controller of claim 15, further comprising an integrated universal serial bus hub connected to said device connections.



17. The device controller of claim 15, wherein said device connections are universal serial bus connections.

18. The device controller of claim 15, further comprising an industry standard architecture device connection operatively connected to said communications controller.