(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
28 July 2005 (28.07.2005)

PCT

(10) International Publication Number
WO 2005/069132 A1

(51) International Patent Classification[7]: G06F 9/445

(21) International Application Number:
PCT/EP2004/053205

(22) International Filing Date: 1 December 2004 (01.12.2004)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
10/752,632 7 January 2004 (07.01.2004) US

(71) Applicant *(for all designated States except US)*: **INTERNATIONAL BUSINESS MACHINES CORPORATION** [US/US]; New Orchard Road, Armonk, NY 10504 (US).

(71) Applicant *(for MG only)*: **IBM UNITED KINGDOM LIMITED** [GB/GB]; PO Box 41 North Harbour, Portsmouth, Hampshire PO6 3AU (GB).

(72) Inventors; and

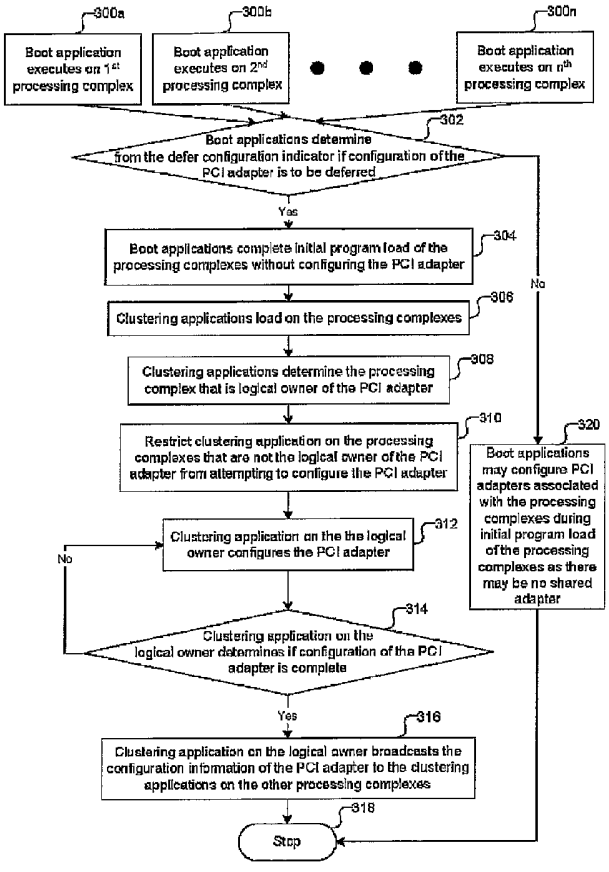
(75) Inventors/Applicants *(for US only)*: **HSU, Yu-Cheng** [US/US]; 2460 N. Aileen Avenue, Tucson, AZ 85715 (US). **MCCAULEY, John, Norbert** [US/US]; 8860 East Saddleback Drive, Tucson, AZ 85749 (US). **CHENG-CHUNG, Song** [US/US]; 3000 North Soldier Trail, Tucson, AZ 85749 (US). **SHERMAN, William, Griswold** [US/US]; 6916 East Via Dorado, Tucson, AZ 85715 (US).

(74) Agent: **BURT, Roger, James**; IBM United Kingdom Limited, Intellectual Property Law, Hursley Park, Winchester, Hampshire SO21 2JN (GB).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG,

[Continued on next page]

(54) Title: CONFIGURING A SHARED RESOURCE



(57) Abstract: Provided are a method, system, and article of manufacture, wherein in certain embodiments a determination is made as to whether a configuration indicator associated with a resource indicates a delayed configuration of the resource, wherein the resource is shared by a plurality of processing complexes via a bus, and wherein if the delayed configuration of the resource is indicated then the resource is prevented from being configured during initial program loads of the plurality of processing complexes. The resource is configured by only one of the plurality of processing complexes that shares the resource, in response to determining that the configuration indicator associated with the resource indicates the delayed configuration of the resource.

WO 2005/069132 A1

PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

CONFIGURING A SHARED RESOURCE

Technical Field

[001] The present disclosure relates to a method, system, and an article of manufacture for configuring a shared resource.

Background Art

[002] A multi-cluster system may couple a plurality of computing nodes together. The plurality of coupled computing nodes may collectively perform as a single computational system. The multi-cluster system may be used for parallel processing, load balancing, fault tolerance, etc., and implement a high-availability system or a redundant information technology system. For example, certain multi-cluster systems may store the same data in a plurality of computing nodes, where a computing node may be a computational unit, a storage unit, etc. When one computing node of the multi-cluster system is unavailable, an alternate computing node of the multi-cluster system may be used to substitute the unavailable computing node.

[003] A resource such as a Peripheral Component Interconnect (PCI) bus may be used to interconnect devices, with the local bus of a processor and main memory. In certain multi-cluster systems, a PCI adapter may be shared by a plurality of computing nodes via a PCI bus. Hosts may communicate with the computing nodes of the multi-cluster system via the shared PCI adapter.

[004] If a PCI adapter is shared among a plurality of computing nodes of a multi□

[005] cluster system, then the failure of one computing node may still allow the hosts to continue communications with other computing nodes of the multi-cluster system. Such hosts may be able to access the shared PCI adapter and access data associated with the computing nodes that have not failed.

[006] If more than one computing node is allowed to configure the shared resource then the shared resource may be in an erroneous state and may then not be shared amongst the computing nodes.

Disclosure of Invention

[007] The invention provides a method, system and computer program product, as claimed in the independent claims.

Brief Description of the Drawings

[008] Referring now to the drawings in which like reference numbers represent corresponding parts throughout:

[009] FIG 1 illustrates a block diagram of a first computing environment, in accordance with certain described embodiments of the invention;

[010] FIG 2 illustrates a block diagram of data structures associated with a shared PCI adapter, in accordance with certain described embodiments of the invention;

[011] FIG 3 illustrates logic for configuring a shared PCI adapter, in accordance with certain described embodiments of the invention;

[012] FIG 4 illustrates logic for transferring logical ownership of a shared PCI adapter, in accordance with certain described embodiments of the invention;

[013] FIG 5 illustrates a block diagram of a second computing environment, in accordance with certain described embodiments of the invention; and FIG 6 illustrates a block diagram of a computer architecture in which certain described aspects of the invention are implemented.

Best Mode for Carrying Out the Invention

[014] In the following description, reference is made to the accompanying drawings which form a part hereof and which illustrate several embodiments. It is understood that other embodiments may be utilized and structural and operational changes may be made without departing from the scope of the present embodiments.

[015] FIG 1 illustrates a block diagram of a first computing environment, in accordance with certain embodiments of the invention.

[016] In certain embodiments, a plurality of processing complexes 100a...100n are coupled to a PCI adapter 102 by a PCI bus 104. A processing complex, such as the processing complex 100a, may include one or more computing nodes, where the computing nodes may include uniprocessor or multiprocessor systems. In some embodiments, a processing complex 100a...100n may include a personal computer, a workstation, a server, a mainframe, a hand held computer, a palm top computer, a telephony device, a network appliance, a blade computer, a storage controller, etc.

[017] In certain embodiments, the PCI adapter 102 may be replaced by a PCI-X adapter, or any PCI or PCI-X bus compatible device. Furthermore, in alternative embodiments the PCI bus 104 may be replaced by a PCI-X bus or some other bus.

[018] The plurality of processing complexes 100a...100n include a plurality of boot applications 106a...106n and a plurality of clustering applications 108a....108n. For example, in certain embodiments, the processing complex 100a may include the boot application 106a and the clustering application 108a, the processing complex 100b may include the boot application 106b and the clustering application 108b, and the processing complex 100n may include the boot application 106n and the clustering ap-

plication 108n. The boot applications 106a...106n and the clustering applications 108a...108n may be implemented in software, firmware, or hardware or any combination thereof.

[019] A boot application 106a...106n when executed may perform an initial program load of the corresponding processing complex 100a...100n. For example, the boot application 106a may perform an initial program load of the processing complex 100a, the boot application 106b may perform an initial program load of the processing complex 100b, and the boot application 106n may perform an initial program load of the processing complex 100n. During an initial program load of a processing complex, the operating system (not shown), device drivers (not shown), etc., of the processing complex may be loaded, such that the processing complex becomes ready to execute other applications after the initial program load is completed.

[020] The clustering applications 108a...108n when executed may allow the processing nodes 100a...100n to share the PCI adapter 102. In certain embodiments, only one clustering application may configure the PCI adapter 102 and broadcast the configuration information of the PCI adapter 102 to the other clustering applications. For example, in certain embodiments the clustering application 108a may configure the PCI adapter 102 and broadcast the configuration information of the PCI adapter 102 to the 10 clustering application 108b....108n that may be executing in the processing complexes 100b...100n. Although a plurality of clustering applications 108a...108n are shown, in certain embodiments the plurality of clustering applications 108a...108n may be part of a distributed clustering application associated with the processing complexes 100a...100n.

[021] FIG 2 illustrates a block diagram of data structures associated with the PCI 20 adapter 102, in accordance with certain embodiments of the invention. In certain embodiments, a defer configuration indicator 200 may be associated with the PCI adapter 102. In some embodiments, the defer configuration indicator 200 may represent a single bit of data, where the single bit of data may indicate whether configuration of the PCI adapter 102 should be deferred until the completion of the initial program loads of the processing complexes 100a...100n.

[022] For example, in certain embodiments if the single bit of data in the defer configuration indicator 200 is set to one, then the boot applications 106a...106n may not configure the PCI adapter 102 during initial program loads of the processing complexes 100a...100n. In certain embodiments, only one of the clustering applications 108a...108n, such as, clustering application 108a, may configure the PCI

adapter 102. The clustering applications 108a...108n may start executing only after the completion of the initial program loads of the processing complexes 100a...100n.

[023] FIG. 3 illustrates logic for configuring a shared PCI adapter implemented in the processing complexes 100a...100n, in accordance with certain embodiments of the invention.

[024] Control starts at blocks 300a...300n, where the boot applications 106a...106n execute in the corresponding processing complexes 100a...100n. For example, the boot application 106a executes (at block 300a) in the processing complex 100a, the boot application 106b executes (at block 300b) in the processing complex 100b, and the boot application 106n executes (at block 300n) in the processing complex 100n. The execution of the boot applications 106a...106n in the processing complexes 100a...100n may be in parallel. As a result of the execution of the boot applications 106a...106n, the initial program loads start in the processing complexes 100a...100n.

[025] In certain embodiments, each of the boot applications 106a...106n may determine (at block 302) from the defer configuration indicator 200 whether the configuration of the PCI adapter 102 is to be deferred until the completion of the initial program loads of the processing complexes 100a...100n. If so, the boot applications 106a...106n complete (at block 304) the initial program loads of the processing complexes 100a...100n without configuring the PCI adapter 102.

[026] The clustering applications 108a...108n load (at block 306) in the corresponding processing complexes 100a...100n. The clustering applications 108a...108n may collectively determine (at block 308) a processing complex that is the logical owner of the PCI adapter 102, where the processing complex that is the logical owner is included in the plurality of processing complexes 100a...100n. For example, in certain embodiments the processing complex 100a may be determined as the logical owner of the PCI adapter 102. The processing complex that is the logical owner of the PCI adapter 102 assumes the responsibility of configuring the PCI adapter 102 and broadcasting the configuration information of the PCI adapter 102 to the other processing complexes.

[027] The clustering applications 108a...108n restrict (at block 310) those processing complexes that are not the logical owner of the PCI adapter 102 from attempting to configure the PCI adapter 102. The clustering application in the logical owner configures (at block 312) the PCI adapter 102. For example, in certain embodiments if processing complex 100a is the logical owner then the clustering application 108a may configure the PCI adapter 102.

[028] The clustering application in the logical owner determines (at block 314) whether the configuration of the PCI adapter 102 is complete. If so, the clustering application in the logical owner broadcasts (at block 316) the configuration information of the PCI adapter 102 to the clustering applications of the other processing complexes. For example, if processing complex 100a is determined to be the logical owner of the PCI adapter 102, then the clustering application 108a distributes the configuration information of the PCI adapter to the clustering applications 108b...108n. The process stops (at block 318) in response to the completion of the broadcast of the configuration information of the PCI adapter 102.

[029] If the clustering application in the logical owner determines (at block 314) that the configuration of the PCI adapter 102 is incomplete, then the clustering application in the logical owner continues configuring (at block 312) the PCI adapter 102.

[030] If the boot applications 106a...106n determine (at block 302) from the defer configuration indicator 200 that the configuration of the PCI adapter 102 is not to be deferred then the boot applications 106a...106n may configure (at block 320) one or more PCI adapters associated with the processing complexes 100a...100n during the initial program loads of the processing complexes 100a...100n as there may be no shared PCI adapter among the processing complexes 100a...100n. Subsequent to the configuration, the process stops (at block 318).

[031] Therefore, the logic of FIG. 3 illustrates certain embodiments in which configuration of the shared PCI adapter 102 is delayed until the completion of the initial program loads of the processing complexes 100a...100n. In response to the completion of the initial program loads of the processing complexes 100a...100n, only one of the plurality of processing complexes 100a...100n may configure the shared PCI adapter 102 and broadcast the configuration information of the shared PCI adapter 102 to the other processing complexes.

[032] FIG. 4 illustrates logic for transferring logical ownership of the shared PCI adapter 102, in accordance with certain described embodiments of the invention. In certain embodiments, the logic for transferring logical ownership of the shared PCI adapter may be implemented in the clustering applications 108a...108n that execute in the processing complexes 100a...100n.

[033] Control starts at block 400, where the processing complex that is the logical owner of the PCI adapter 102 fails. The failure may be as a result of a malfunctioning of the logical owner because of a software, hardware, or firmware error. Failures of the logical owner may also occur because of other reasons.

[034] The clustering applications in the processing complexes that are not the logical owner determine (at block 402) a new logical owner of the PCI adapter 102. For example, in certain embodiments if the logical owner that failed is the processing complex 100a, then the clustering applications 108b...108n determine a new logical owner of the PCI adapter 102 from the processing complexes 100b...100n.

[035] The clustering application in the new logical owner assumes (at block 404) responsibility for configuring or reconfiguring the PCI adapter 102 and broadcasting the configuration information of the PCI adapter 102 to the clustering applications in the other processing complexes. For example, if processing complex 100a had failed and the new logical owner is the processing complex 100b, then the new logical owner 100b may assume the responsibility for configuring or reconfiguring the PCI adapter 102 and broadcasting the configuration information of the PCI adapter 102 to the clustering applications 108c...108n in the processing complexes 100c...100n, where the processing complexes 100a, 100b, 100c,....100n share the PCI adapter 102.

[036] In certain embodiments, the clustering application in the new logical owner determines (at block 406) whether the failed processing complex has become functional and rejoined the processing complexes, where the processing complexes may form a multi-cluster environment. If so, in certain embodiments the clustering application in the new logical owner may transfer (at block 408) logical ownership of the shared PCI adapter 102 back to the old logical owner. For example, the new logical owner 100b may optionally transfer logical ownership of the shared PCI adapter 102 back to the original logical owner 100a, if the original logical owner 100a is no longer a failed processing complex. If the clustering application in the new logical owner determines (at block 406) that the failed processing complex has not become functional, then the new logical owner continues (at block 404) to assume responsibility for configuring the shared PCI adapter 102 and broadcasting the configuration information.

[037] Therefore, FIG. 4 illustrates certain embodiments in which in the event of a failure of a logical owner of the shared PCI adapter 102, the other processing complexes determine a new logical owner that assumes the responsibility of configuring the PCI adapter 102 and broadcasting the configuration information of the PCI adapter 102 to the other functioning processing complexes. In the event of a recovery from failure of the original logical owner, in certain embodiments the new logical owner may in transfer the logical ownership of the PCI adapter back to the original logical owner.

[038] FIG. 5 illustrates a block diagram of a second computing environment 500, in

accordance with certain described embodiments of the invention. In the second computing environment 500, the plurality of processing complexes 100a...100n may comprise a multi-cluster system 502, where a processing complex in the plurality of processing complexes 100a...100n is a node of the multi-cluster system 502.

[039] The processing complexes 100a...100n in the multi-cluster system 502 may be coupled to a shared resource 504 via a bus 506. In certain embodiments, the shared resource 504 may include the PCI adapter 102 and the bus 506 may be a PCI bus. In certain embodiments, other resources besides the PCI adapter 102 may comprise the shared resource 504. The shared resource 504 may include a defer configuration indicator 508 that indicates whether the configuration of the shared resource 504 should be delayed until the processing complexes 100a...100n of the multi-cluster system 502 have completed initial program loads. Subsequent to the completion of the initial program loads of the processing complexes 100a...100n of the multi-cluster system 502, only one of the processing complexes 100a...100n may configure the shared resource 504 and broadcast the configuration information of the shared resource to the other processing complexes.

[040] A host 510 that is coupled to the multi-cluster system 502 may access data in the processing complexes 100a...100n via the shared resource 504. Even if one or more processing complexes 100a...100n fail, the other processing complexes may still be accessed by the host 510 via the shared resource 504.

[041] An attribute that indicates that configuration of the shared PCI adapter should be deferred until the completion of the initial program loads of the processing complexes is associated with the shared PCI adapter. A clustering application implemented in the plurality of processing complexes may coordinate the plurality of processing complexes, such that, only a single processing complex of the plurality of processing complexes configures the shared PCI adapter and broadcasts the configuration information of the shared PCI adapter to the other processing complexes. If more than a single processing complex of the plurality of processing complexes were to configure the shared PCI adapter then the shared PCI adapter may be in an erroneous state and may not be shared among the plurality of processing complexes. The embodiments allow the PCI adapter to be shared among the plurality of processing complexes by configuring the PCI adapter with only one of the plurality of processing complexes.

[042] Additional Implementation Details

[043] The described techniques may be implemented as a method, apparatus or article of manufacture using standard programming and/or engineering techniques to produce

software, firmware, hardware, or any combination thereof. The term "article of manufacture" as used herein refers to code or logic implemented in hardware logic (e.g., an integrated circuit chip, Programmable Gate Array (PGA), Application Specific Integrated Circuit (ASIC), etc.) or a computer readable medium (e.g., magnetic storage medium, such as hard disk drives, floppy disks, tape), optical storage (e.g., CD-ROMs, optical disks, etc.), volatile and non-volatile memory devices (e.g., EEPROMs, ROMs, PROMs, RAMs, DRAMs, SRAMs, firmware, programmable logic, etc.). Code in the computer readable medium is accessed and executed by a processor. The code in which

[044] embodiments are made may further be accessible through a transmission media or from a file server over a network. In such cases, the article of manufacture in which the code is implemented may comprise a transmission media, such as a network transmission line, wireless transmission media, signals propagating through space, radio waves, infrared signals, etc. Of course, those skilled in the art will recognize that many modifications may be made to this configuration without departing from the scope of the embodiments, and that the article of manufacture may comprise any information bearing medium known in the art.

[045] FIG. 6 illustrates a block diagram of a computer architecture in which certain aspects of the invention are implemented. Any of the processing complexes 100a...100n or the host 510 may implement a computer architecture 600 having a processor 602, a memory 604 (e.g., a volatile memory device), and storage 606 (e.g., a non-volatile storage, magnetic disk drives, optical disk drives, tape drives, etc.). The storage 606 may comprise an internal storage device, an attached storage device or a network accessible storage device. Programs in the storage 606 may be loaded into the memory 604 and executed by the processor 602 in a manner known in the art. The architecture may further include a network card 608 to enable communication with a network. The architecture may also include at least one input 610, such as a keyboard, a touchscreen, a pen, voice-activated input, etc., and at least one output 612, such as a display device, a speaker, a printer, etc.

[046] The logic of FIGs. 3 and 4 describes specific operations occurring in a particular order. Further, the operations may be performed in parallel as well as sequentially. In alternative embodiments, certain of the logic operations may be performed in a different order, modified or removed and still implement embodiments of the present invention. Moreover, steps may be added to the above described logic and still conform to the embodiments. Yet further steps may be performed by a single process

or distributed processes.

[047] Many of the software and hardware components have been described in separate modules for purposes of illustration. Such components may be integrated into a fewer number of components or divided into a larger number of components. Additionally, certain operations described as performed by a specific component may be performed by other components.

[048] Therefore, the foregoing description of the embodiments has been presented for the purposes of illustration and description. It is not intended to be exhaustive or to limit the invention to the precise form disclosed.

Claims

[001] A method, comprising: determining whether a configuration indicator (200) associated with a resource (102) indicates a delayed configuration of the resource, wherein the resource is shared by a plurality of processing complexes (100a, 100b, 100n) via a bus (104), and wherein if the delayed configuration of the resource is indicated then the resource is prevented from being configured during initial program loads of the plurality of processing complexes; and configuring the resource by only one of the of plurality of processing complexes that shares the resource, in response to determining that the configuration indicator associated with the resource indicates the delayed configuration of the resource.

[002] The method of claim 1, wherein determining whether the configuration indicator associated with the resource indicates the delayed configuration of the resource is performed during the initial program loads of the processing complexes, and wherein the method further comprises: completing the initial program loads of the plurality of processing complexes that share the resource, in response to determining that the configuration indicator associated with the resource indicates the delayed configuration of the resource, wherein configuring the resource by only one of the plurality of processing complexes is subsequent to completing the initial program loads of the plurality of processing complexes that share the resource.

[003] The method of claim 1, wherein the only one of the plurality of processing complexes is a logical owner of the resource, and wherein the method further comprises: broadcasting, by the logical owner, a configuration information of the configured resource to other processing complexes of the plurality of processing complexes.

[004] The method of claim 1, further comprising: determining a failure of the only one of the plurality of processing complexes that configured the resource, wherein the only one of the plurality of processing complexes is a logical owner of the resource; and determining a new logical owner of the resource from the plurality of processing complexes, in response to determining the failure, wherein the new logical owner is responsible for a subsequent configuration of the resource.

[005] The method of claim 1, wherein the only one of the plurality of processing complexes is an original logical owner of the resource, and wherein the method

further comprises: determining a new logical owner of the resource from the plurality of processing complexes, in response to a failure of the original logical owner; determining that the original logical owner has recovered from the failure; and transferring logical ownership of the resource from the new logical owner to the original logical owner, in response to determining that the original logical owner has recovered from the failure.

[006] The method of claim 1, wherein determining whether the configuration indicator associated with the resource indicates a delayed resource is performed during the initial program loads of the plurality of processing complexes, and wherein the method further comprises: deferring a configuration of the resource until the initial program loads are completed for the plurality of processing complexes, in response to determining that the configuration indicator associated with the resource indicates the delayed configuration of the resource.

[007] The method of claim 1, wherein the plurality of processing complexes comprise a multi-cluster system, and wherein the plurality of processing complexes are accessed by a host via the configured resource that is shared by the plurality of processing complexes.

[008] The method of claim 1, wherein determining whether a configuration indicator associated with a resource indicates a delayed configuration of the resource is performed by a boot application implemented in a first processing complex of the plurality of processing complexes, wherein configuring the resource by only one of the plurality of processing complexes that shares the resource is performed by a clustering application implemented in the only one of the plurality of processing complexes.

[009] The method of claim 1, wherein configuring the resource is coordinated by a clustering application that spans the plurality of processing complexes.

[010] The method of claim 1, wherein the shared resource is a PCI adapter, wherein the bus is a PCI bus, and wherein the configuration indicator is implemented in the PCI adapter.

[011] A system, comprising: a plurality of processing complexes; a bus coupled to the plurality of processing complexes; a resource shared by the plurality of processing complexes via the bus; a configuration indicator associated with the resource; means for determining whether the configuration indicator associated with the resource indicates a delayed configuration of the resource, and wherein if the delayed configuration of the resource is indicated then the resource is

prevented from being configured during initial program loads of the plurality of processing complexes; and means for configuring the resource by only one of the of plurality of processing complexes, in response to determining that the configuration indicator indicates the delayed configuration of the resource.

[012] The system of claim 11, wherein the means for determining whether the configuration indicator associated with the resource indicates the delayed configuration of the resource performs during the initial program loads of the processing complexes, and wherein the system further comprises: means for completing the initial program loads of the plurality of processing complexes that share the resource, in response to determining that the configuration indicator associated with the resource indicates the delayed configuration of the resource, wherein configuring the resource by only one of the plurality of processing complexes is subsequent to completing the initial program loads of the plurality of processing complexes that share the resource.

[013] The system of claim 11: a logical owner of the resource, wherein the only one of the plurality of processing complexes is the logical owner of the resource; and means for broadcasting, by the logical owner, a configuration information of the configured resource to other processing complexes of the plurality of processing complexes.

[014] The system of claim 11, further comprising: means for determining a failure of the only one of the plurality of processing complexes that configured the resource, wherein the only one of the plurality of processing complexes is a logical owner of the resource; and means for determining a new logical owner of the resource from the plurality of processing complexes, in response to determining the failure, wherein the new logical owner is responsible for a subsequent configuration of the resource.

[015] The system of claim 11, wherein the only one of the plurality of processing complexes is an original logical owner of the resource, and wherein the system further comprises: means for determining a new logical owner of the resource from the plurality of processing complexes, in response to a failure of the original logical owner; means for determining that the original logical owner has recovered from the failure; and means for transferring logical ownership of the resource from the new logical owner to the original logical owner, in response to determining that the original logical owner has recovered from the failure.

[016] The system of claim 11, wherein determining whether the configuration indicator

associated with the resource indicates a delayed resource is performed during the initial program loads of the plurality of processing complexes, and wherein the system further comprises: means for deferring a configuration of the resource until the initial program loads are completed for the plurality of processing complexes, in response to determining that the configuration indicator associated with the resource indicates the delayed configuration of the resource.

[017] The system of claim 11, wherein the plurality of processing complexes comprise a multi-cluster system, and wherein the plurality of processing complexes are accessed by a host via the configured resource that is shared by the plurality of processing complexes.

[018] The system of claim 11, further comprising: a clustering application implemented in the only one of the plurality of processing complexes; and a boot application implemented in a first processing complex of the plurality of processing complexes, wherein determining whether a configuration indicator associated with a resource indicates a delayed configuration of the resource is performed by the boot application, wherein configuring the resource by only one of the plurality of processing complexes that shares the resource is performed by the clustering application.

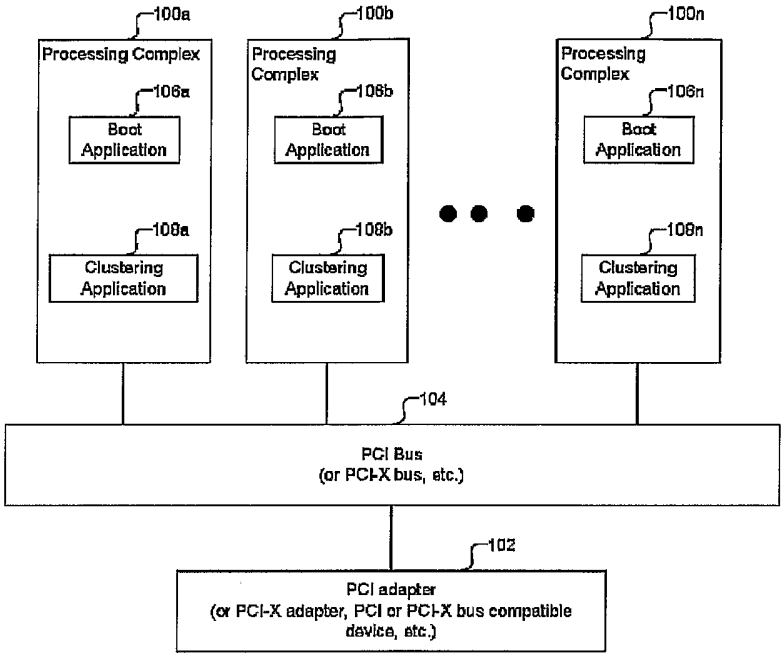
[019] The system of claim 11, further comprising: a clustering application that spans the plurality of processing complexes, wherein configuring the resource is coordinated by the clustering application.

[020] The system of claim 11, wherein the shared resource is a PCI adapter, wherein the bus is a PCI bus, and wherein the configuration indicator is implemented in the PCI adapter.

[021] A computer program product stored on a computer readable storage medium for, when run on a computer system, instructing the computer system to carry out the method of any preceding method claim.

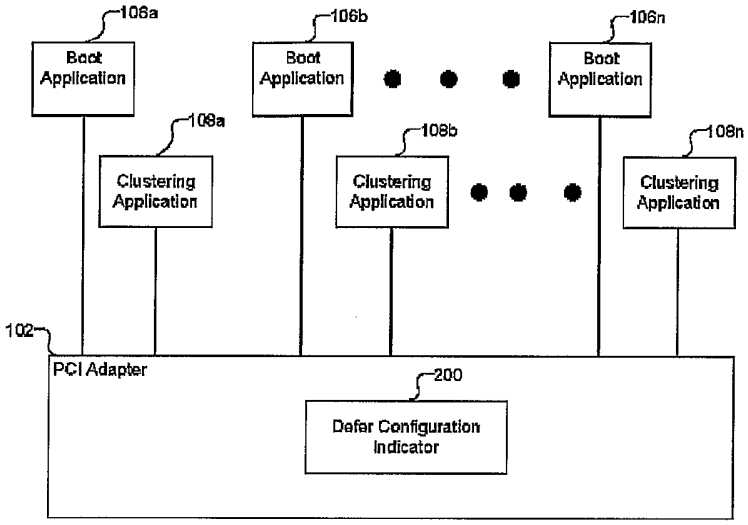
[Fig.]

FIG. 1


100a
Processing Complex
106a
Boot Application
108a
Clustering Application
100b
Processing Complex
106b
Boot Application
108b
Clustering Application
100n
Processing Complex
106n
Boot Application
108n
Clustering Application
104
PCI Bus
(or PCI-X bus, etc.)
102
PCI adapter
(or PCI-X adapter, PCI or PCI-X bus compatible device, etc.)

[Fig.]

FIG. 2


106a
Boot Application
106b
Boot Application
106n
Boot Application
108a
Clustering Application
108b
Clustering Application
108n
Clustering Application
102
PCI Adapter
200
Defer Configuration Indicator

[Fig.]

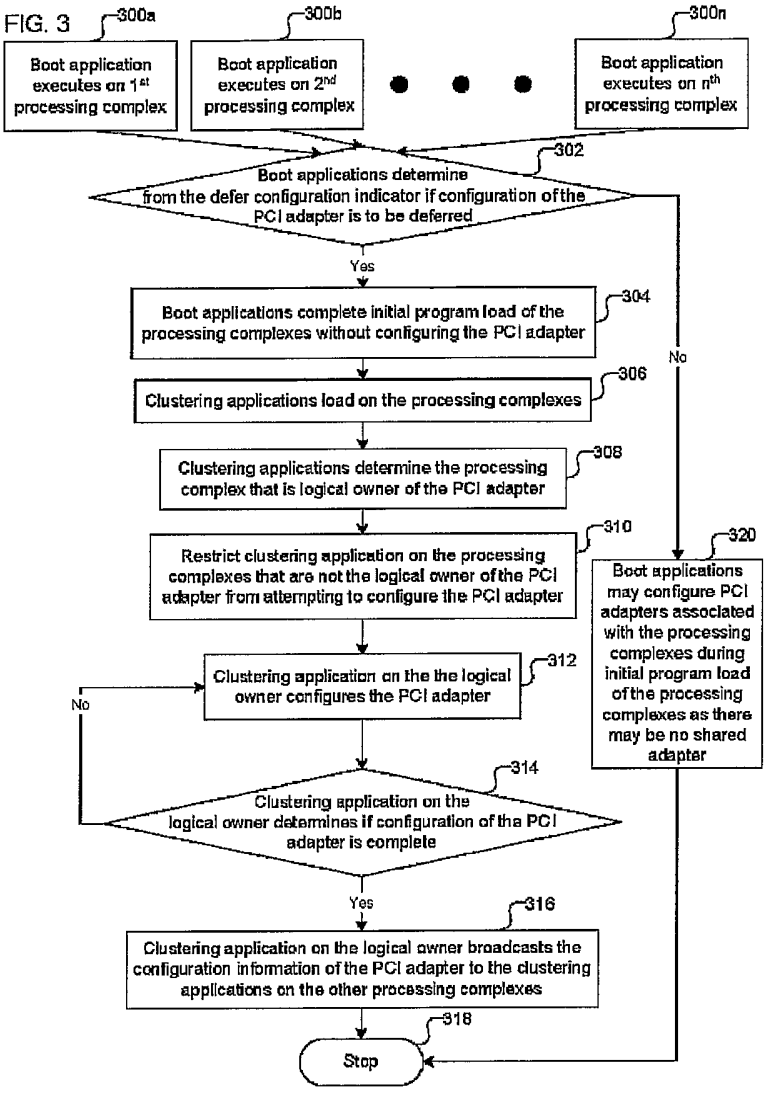
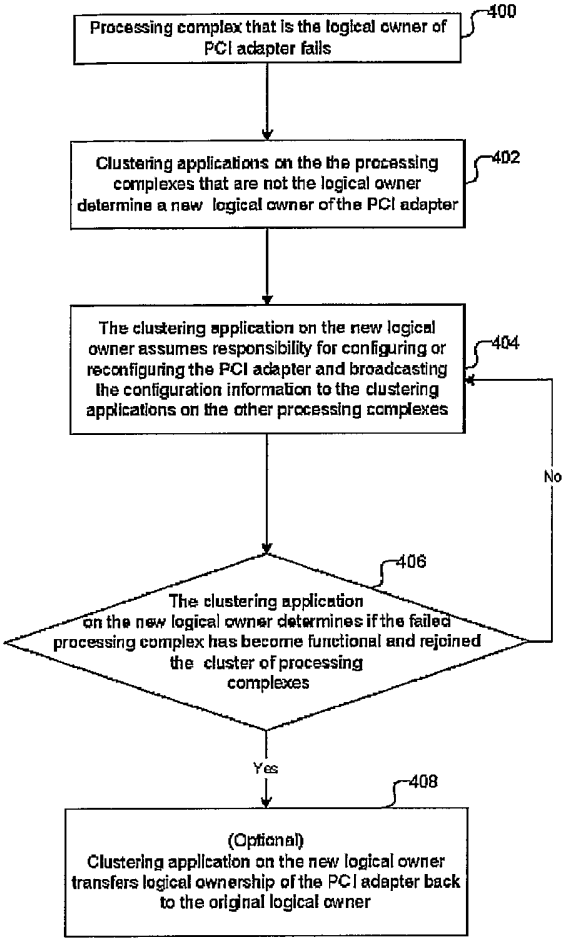


FIG. 3
300a
300b
300n
Boot application executes on 1st processing complex
Boot application executes on 2nd processing complex
Boot application executes on nth processing complex
302
Boot applications determine from the defer configuration indicator if configuration of the PCI adapter is to be deferred
Yes
No
304
Boot applications complete initial program load of the processing complexes without configuring the PCI adapter
306
Clustering applications load on the processing complexes
308
Clustering applications determine the processing complex that is logical owner of the PCI adapter
310
Restrict clustering application on the processing complexes that are not the logical owner of the PCI adapter from attempting to configure the PCI adapter
320
Boot applications may configure PCI adapters associated with the processing complexes during initial program load of the processing complexes as there may be no shared adapter
312
Clustering application on the the logical owner configures the PCI adapter
No
314
Clustering application on the logical owner determines if configuration of the PCI adapter is complete
Yes
316
Clustering application on the logical owner broadcasts the configuration information of the PCI adapter to the clustering applications on the other processing complexes
318
Stop

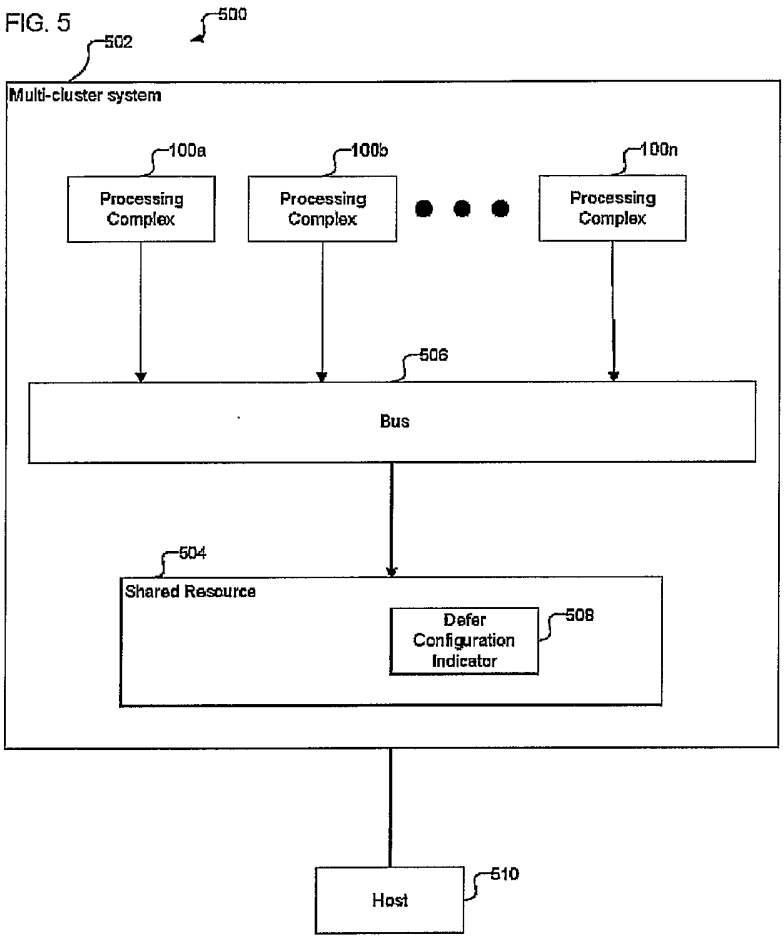
[Fig.]

FIG. 4


Processing complex that is the logical owner of PCI adapter fails
400
Clustering applications on the the processing complexes that are not the logical owner determine a new logical owner of the PCI adapter
402
The clustering application on the new logical owner assumes responsibility for configuring or reconfiguring the PCI adapter and broadcasting the configuration information to the clustering applications on the other processing complexes
404
No
406
The clustering application on the new logical owner determines if the failed processing complex has become functional and rejoined the cluster of processing complexes
Yes
408
(Optional)
Clustering application on the new logical owner transfers logical ownership of the PCI adapter back to the original logical owner

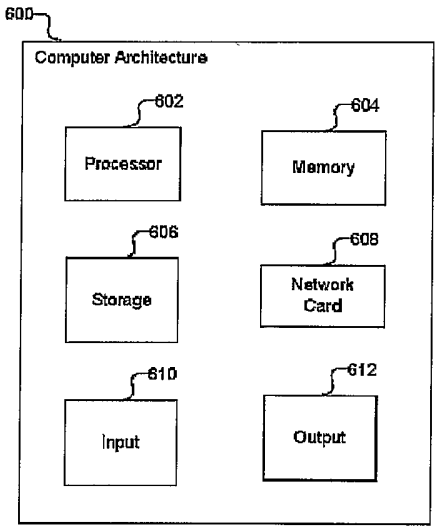
[Fig.]

FIG. 5


500
502
Multi-cluster system
100a
Processing Complex
100b
Processing Complex
100n
Processing Complex
506
Bus
504
Shared Resource
Defer Configuration Indicator
508
Host
510

[Fig.]

FIG. 6


600
Computer Architecture
602
Processor
604
Memory
606
Storage
608
Network Card
610
Input
612
Output

INTERNATIONAL SEARCH REPORT

International Application No
PCT/EP2004/053205

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06F9/445

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, WPI Data, PAJ, INSPEC

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 6 401 120 B1 (GAMACHE ROD ET AL) 4 June 2002 (2002-06-04) column 4, lines 15-57 column 5, lines 10-22 column 6, lines 55-63 column 7, lines 15-26 column 8, lines 17-50 column 9, lines 22-55 column 11, lines 20-25 column 12, lines 28-65 column 13, lines 1-33 column 18, lines 20-29 column 19, lines 30-45 figures 2-11 ----- -/--	1-21

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
8 April 2005	20/04/2005

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Noll, J

4

Form PCT/ISA/210 (second sheet) (January 2004)

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2003/188108 A1 (DAMRON TIMOTHY M ET AL) 2 October 2003 (2003-10-02) paragraph '0002! paragraph '0004! paragraphs '0011!, '0012! paragraph '0063! figures 1-4 claim 8	1-21
A	WILKINS R S ET AL: "Disaster tolerant wolfpack geo-clusters" CLUSTER COMPUTING, 2002. PROCEEDINGS. 2002 IEEE INTERNATIONAL CONFERENCE ON 23-26 SEPT. 2002, PISCATAWAY, NJ, USA,IEEE, 23 September 2002 (2002-09-23), pages 222-227, XP010621879 ISBN: 0-7695-1745-5 page 223, left-hand column - page 226, left-hand column	1-21

INTERNATIONAL SEARCH REPORT

Information on patent family members

Internal Application No
PCT/EP2004/053205

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 6401120 B1	04-06-2002	AU 3729500 A	16-10-2000
		EP 1222540 A1	17-07-2002
		WO 0058824 A1	05-10-2000
		US 2002161889 A1	31-10-2002
US 2003188108 A1	02-10-2003	NONE	