US 20170078385A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2017/0078385 A1**

Dress (43) **Pub. Date: Mar. 16, 2017**

(54) **GROUP-COHERENT MEMORY**

(71) Applicant: **LightFleet Corporation**, Camas, WA (US)

(72) Inventor: **William Dress**, Camas, WA (US)

(21) Appl. No.: **15/262,391**

(22) Filed: **Sep. 12, 2016**

Related U.S. Application Data

(60) Provisional application No. 62/216,999, filed on Sep. 10, 2015, provisional application No. 62/217,001, filed on Sep. 10, 2015, provisional application No. 62/217,003, filed on Sep. 10, 2015, provisional application No. 62/217,004, filed on Sep. 10, 2015, provisional application No. 62/241,112, filed on Oct. 13, 2015.

Publication Classification

(51) **Int. Cl.**

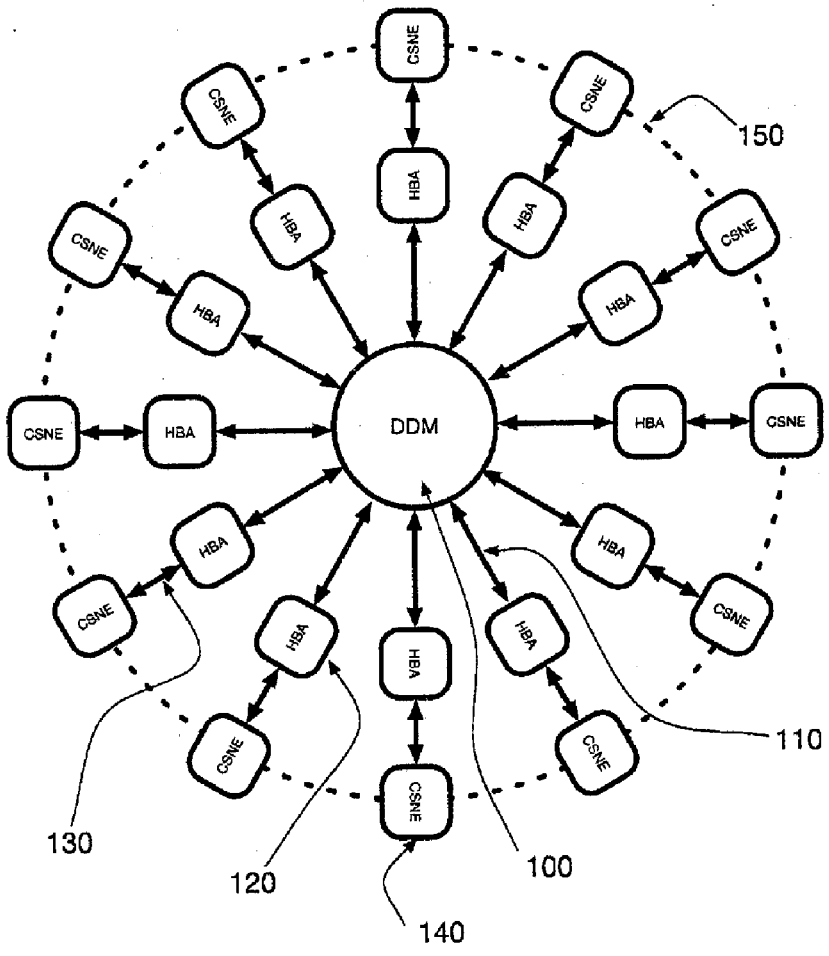
H04L 29/08	(2006.01)
G06F 3/06	(2006.01)
G06F 13/28	(2006.01)
H04L 12/18	(2006.01)
G06F 13/40	(2006.01)

(52) **U.S. Cl.**

CPC ***H04L 67/1095*** (2013.01); ***H04L 67/104*** (2013.01); ***H04L 12/1881*** (2013.01); ***G06F 13/4068*** (2013.01); ***G06F 13/28*** (2013.01); ***G06F 3/061*** (2013.01); ***G06F 3/065*** (2013.01); ***G06F 3/067*** (2013.01)

(57) **ABSTRACT**

Operating a data distribution including a data distribution module and a plurality of host-bus adapters coupled to the data distribution module can include defining a coherent group that includes a set of members that includes the plurality of host-bus adapters; providing a group-coherent memory area in each of the set of members; and initiating a one-to-all broadcast message from a one of the plurality of host-bus adapters to all of the set of members when the one of the plurality of host-bus adapters requests a write to its local group-coherent memory area.



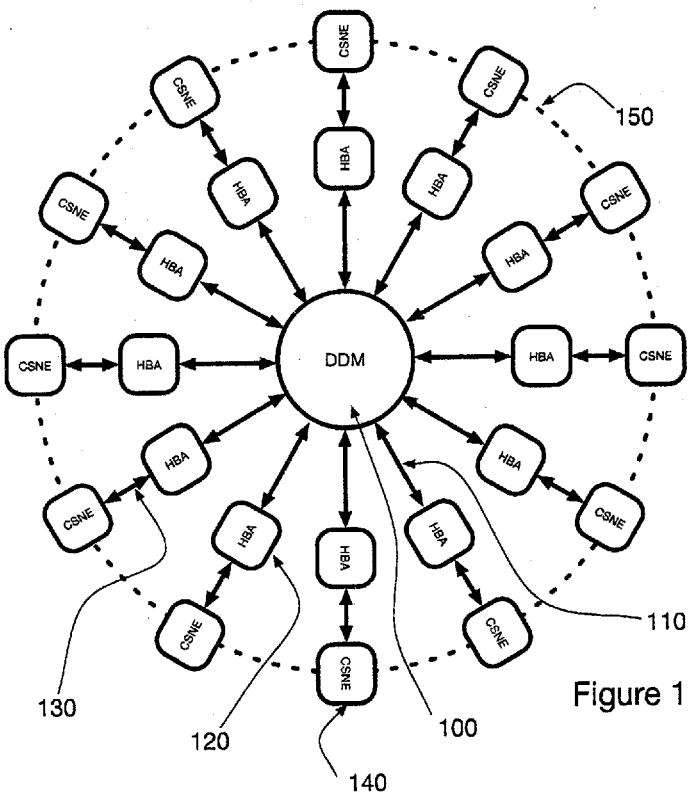


Figure 1

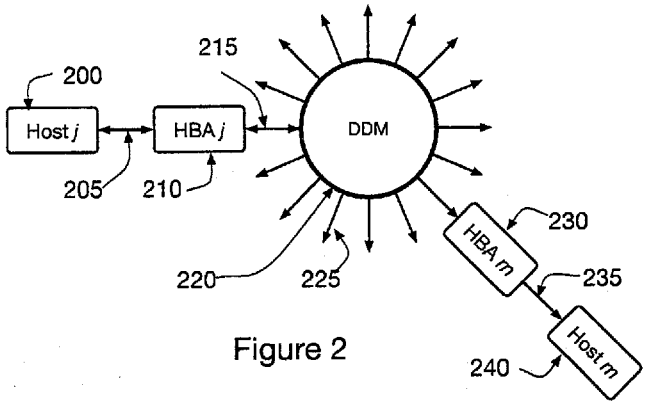


Figure 2

GROUP-COHERENT MEMORY

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] Referring to the application data sheet filed herewith, this application claims a benefit of priority under 35 U.S.C. 119(e) from co-pending, commonly-assigned provisional patent applications U.S. Ser. No. 62/216,999, filed Sep. 10, 2015, U.S. Ser. No. 62/217,001, filed Sep. 10, 2015, U.S. Ser. No. 62/217,003, filed Sep. 10, 2015, U.S. Ser. No. 62/217,004, filed Sep. 10, 2015 and U.S. Ser. No. 62/241,112, filed Oct. 13, 2015, the entire contents of all of which are hereby expressly incorporated herein by reference for all purposes. This application is related to co-pending utility patent application U.S. Ser. No. 15/175,685, filed Jun. 7, 2016, the entire contents of which are hereby expressly incorporated herein by reference for all purposes.

BACKGROUND

[0002] A particular memory model is advantageous in computing architectures for multiple tasks, whose threads are distributed across multiple and separate hosts, requiring access to the same data. Such tasks may be viewed as a working group, dealing with different aspects of the same problem while reading from and writing to the same relative memory locations within each individual host. In a cluster-wide shared architecture, this expanded requirement is often handled by specialized hardware and software in addition to the usual networking hardware for interconnecting the cluster. The goal is to move data between different hosts across the entire cluster such that a portion of local memory in each host is maintained as "mirror image" of the same relative memory in any other host.

[0003] Moving memory contents around to meet coherency needs can materially slow a parallel application. What is desired is a method of automatically updating mirrored copies across a computing cluster with without the addition of specialized hardware.

SUMMARY

[0004] There is a need for the following embodiments of the present disclosure. Of course, the present disclosure is not limited to these embodiments.

[0005] According to an embodiment of the present disclosure, a method comprises operating a data distribution system including a data distribution module and a plurality of host-bus adapters coupled to the data distribution module including defining a coherent group that includes a set of members that includes the plurality of host-bus adapters; providing a group-coherent memory area in each of the set of members; and initiating a one-to-all broadcast message from a one of the plurality of host-bus adapters to all of the set of members when the one of the plurality of host-bus adapters requests a write to its local group-coherent memory area. According to another embodiment of the present disclosure, an apparatus comprises a data distribution system including a data distribution module and a plurality of host-bus adapters coupled to the data distribution module, wherein operating the data distribution system includes defining a coherent group that includes a set of members that includes the plurality of host-bus adapters; providing a group-coherent memory area in each of the set of members; and initiating a one-to-all broadcast message from a one of the plurality of host-bus adapters to each of the set of members ensuring that when the one of the plurality of host-bus adapters request a write, to update its local group-coherent memory area, the one-to-all broadcast message maintains temporal memory coherency across all of the set of members of the coherent group.

[0006] These, and other, embodiments of the present disclosure will be better appreciated and understood when considered in conjunction with the following description and the accompanying drawings. It should be understood, however, that the following description, while indicating various embodiments of the present disclosure and numerous specific details thereof, is given for the purpose of illustration and does not imply limitation. Many substitutions, modifications, additions and/or rearrangements may be made within the scope of embodiments of the present disclosure, and embodiments of the present disclosure include all such substitutions, modifications, additions and/or rearrangements.

BRIEF DESCRIPTION OF THE DRAWINGS

[0007] The drawings accompanying and forming part of this specification are included to depict certain embodiments of the present disclosure. A clearer concept of the embodiments described in this application will be readily apparent by referring to the exemplary, and therefore nonlimiting, embodiments illustrated in the drawings (wherein identical reference numerals (if they occur in more than one view) designate the same elements). The described embodiments may be better understood by reference to one or more of these drawings in combination with the following description presented herein. It should be noted that the features illustrated in the drawings are not necessarily drawn to scale.

[0008] FIG. **1** illustrates the data-distribution system (DDS) tight-cluster interconnect.

[0009] FIG. **2** depicts a coherent memory update process.

DETAILED DESCRIPTION

[0010] The invention relates generally to the field of methods and devices for maintaining coherence between mirrored copies of a task's or program's memory within a cluster of interconnected computers or host devices.

[0011] The invention presented in this disclosure provides a coherent group memory implemented specifically for a tight cluster of servers or other host devices such as found in database installations, high-performance computing applications, and anywhere parallel or cooperative programming may be needed. The coherency mechanism described in this disclosure is supported by the same hardware used to interconnect the cluster and is implemented by defining a coherent memory area in each host and ensuring that a write to that area by any thread in any host initiates a one-to-all broadcast message from the host requesting the write to all hosts in the cluster, including itself. That is, the method is based on a single group whose task it is to maintain temporal memory coherency across the cluster.

[0012] The interconnect mechanism is based on U.S. Ser. No. 62/216,999, filed Sep. 10, 2015 and U.S. Ser. No. 15/175,685, filed Jun. 7, 2016 that describes a message-distribution system or DDS consisting of host bus adapters (HBAs) and a data-distribution module (DDM) where the HBAs provide an interface mechanism between a host and the DDM. The concept, which may also be called "virtual

shared memory", is based on the unique multicast mechanism inherent in the DDS, which can include a data-distribution module (DDM) with host-bus adapters (HBAs) mediating between hosts and the DDM.

[0013] The following sections describe how to achieve a coherent memory mirrored across multiple stations when defined by a single group. The physical memory in the DDS is mirrored (each host has a copy of the coherent memory) where each host in the system belongs to the single coherence group that, while physically mirrored across all hosts or servers, is accessed as if it were a single memory asset shared by the entire cluster. The difference is that no semaphores or access enables are required to achieve a temporal coherence of the cluster's "shared" memory.

[0014] At the architectural level, the DDS as described in U.S. Ser. No. 62/216,999, filed Sep. 10, 2015 and U.S. Ser. No. 15/175,685, filed Jun. 7, 2016 is based on multicast; that is, multicast is the native operational mode of the DDS in that every message injected into the system is treated as multicast having one or more destinations. This native multicast mechanism is built in at the hardware level or physical layer of the DDS. Messages are guided through the DDM (from HBA to HBA) by means of a destination field in the start-of-message (SOM) header, which directs the message along internal paths from the input port to the specified output ports.

[0015] In addition to the multicast, multiple priority levels are built in at the architectural design stage. The lowest and highest priorities are reserved for system functions such as flow control and system maintenance. The message priorities are maintained as separate information channels from a transmitting HBA through the DDM to the receiving HBA.

[0016] A third design feature of the MDA is its support for "fast-priority messages" or FPMs which are single-frame messages whose purpose is to maintain efficient control over system functions. In the output stage of the DDM and HBA, these messages are assigned priority PO which means that, if queued in a PO queue, they have transmission priority (either to the HBA or to the DDM).

[0017] Any message sent by a host that is meant to update the coherent memory is assigned to channel P**1** while any other messages involving memory-to-memory are assigned priority channels such as a P**2** or P**3** channels with lower priority than channel P**1**. All flow control and other critical semaphores take place on the PO channel. Such control frames bypass the internal distribution mechanism of the DDM and carry out their specific functions without interfering with normal message traffic, other than introducing a one-frame delay. These features are discussed in detail in the above-referenced patent application.

[0018] Based on the above description, a set of mirrored memory locations distributed over the cluster is maintained in a coherent fashion by ensuring that any host updating its local copy of the coherent memory also sends out a one-to-all broadcast to all other hosts in the cluster.

[0019] The memory mechanism proposed here is meant to ensure a read-coherence across mirrored copies of a single, privileged group and the method properly supports coherence only in a tight computing cluster. There are no read requests to the local copy other that the usual reads supported by the kernel. Additionally, there are no special write locks beyond normal memory DMA writes.

[0020] The tight-cluster constraint may be relaxed by a simple handshake mechanism between the HBA and the host receiving a memory update and the single group may be expanded to multiple groups while retaining the single coherence group for maintaining coherent memory for a single cooperative-computing task. However, these enhancements require architectural changes to the interconnect and must be supported by software enhancements to the application programming interface.

[0021] In summary, the memory model disclosed here maintains a group-based, system-wide coherent memory for those hosts having membership in the special coherent group. Coherence in this special group is achieved at the architectural level of the DDM by operationally restricting the P**1** channel to accept only group-memory updates in the form one-to-all messages. These updates are effectively broadcast from the host issuing the update to all other hosts in the system. Due to the priority mechanism, any memory update is sent from the updating HBA to the DDM where it is distributed to the remaining hosts without interference or delays from any other system messages. A feature inherent in the way the DDM processes multicast ensures that the receiving HBAs will simultaneously receive the same update information as long as they are restricted to the P**1** channel. Each HBA will pass on its copy of the update at the correct priority to the host by means of a DMA action, ensuring that the memory images across the cluster will be updated at the same time.

[0022] There is no need for locks and semaphores in this basic model. Thus, coherency is maintained across all system hosts without effort beyond reserving the highest message priority for memory updates. The other priority channels are free to handle whatever message traffic is required and do not interfere with the coherency of the memory update. Thus, other memory locations may receive messages on the lower-priority channels, but such are not guaranteed to be coherent across their respective groups.

[0023] Otherwise stated, the basic concept supports multiple groups allowing other modes of communication between hosts (such as maintenance and reporting functions). The coherence group is a preferred group whose only access is by means of the P**1** priority channel and this channel may be reserved only for coherent memory updates. The memory images of the other groups are consistent in that a group write updates all images within a group just as done for the coherent group. However, these updates are not guaranteed to maintain a strict read coherency as does the privileged group by virtue of its temporal coherency.

[0024] Referring to FIG. **1**, the DDS tight-cluster consists of a central data-distribution module (DDM) **100** connected to the host-bus adapters (HBAs) **120** via fiber connections **110**. Each HBA **120** connects to a host or CSNE **140** (computing, storage, or network element) such that these elements are fully connected to each other by means of DDM **100**. Each host or CSNE **140** contains an identical region of memory, perhaps with different absolute addresses within each host. It is this collection of memory images that is required to remain coherent with precisely the same contents at any instant in time. There are several classes of "coherent" memory. The most common one may be termed "read coherent" in that any time any host is allowed to read the same relative memory location, the same value is obtained; enforcing read coherence usually requires the use of locks and semaphores. The most stringent type of coherent memory is temporally coherent in that reads are not delayed or controlled by semaphores, yet a random read by

any host to its local copy of the coherent memory is guaranteed to return the same value.

[0025] Refer to FIG. **2** and assume for now that there is no other traffic on priority channel P**1** through DDM **220**. Then a memory-update message broadcast from Host j **200** will enter HBA j via connection **205** and be prepared according to the prescription disclosed in the above-referenced patent application. Each host receiving the update is a member of a "coherent group" that is defined as a working or cooperative group of processes (or threads or tasks), one residing in each host belonging to the group, such that group communication is restricted to and reserved for priority channel P**1** as described in the above-referenced patent application.

[0026] Since the group-coherent update is a broadcast message wherein the transmitting host is numbered among the recipients, the message is prepared with a group index referring to the subscription table entry where all bits are set, indicating that all exits in the DDM are to transmit copies of the update. This message then enters the DDM via connection **215** with the start-of-message header (SOM) prepared with the aforementioned group index as its destination and a priority indicating the P**1** priority channel and the offset into the mirrored copies indicating the location in each mirrored image that is to receive the update. The end-of-message (EOM) is prepared as normal with a priority designation of P**1**, a source index of j, and a CRC-32 computed over the data portion of the update as described in detail in the above-referenced patent disclosure.

[0027] The DDM behaves as described in the referenced disclosure and the coherent update is distributed to each P**1** output FIFO simultaneously as there are no prior messages causing delays in any of the paths. Copies of the update then leave all exit ports along connections **225** to each connected HBA **230**, arriving in all HBAs **230**, including HBA **210**, simultaneously since there is no other traffic in the cluster to differentially delay any of the updates. In a similar manner, since HBAs **230** (and **210** as well) are free to process any messages from the DDM without delay and such messages are written by direct-memory access (DMA) directly into the targeted memory areas at the uniformly specified offset positions from the base address in each target Host **240** (and **200** as well), the updates arrive simultaneously in all copies of the group memory (within a small differential jitter of a few nanoseconds due to inhomogeneities in the various physical paths).

[0028] The mechanism for the coherent update occurs when a process running a host completes a calculation or receives a message (not directed to its group-coherent memory) that is to be written out to the coherent group. The write in this case may be trapped by the operating-system kernel in one possible implementation of the process. The kernel then initiates a broadcast to all members of the coherent group, including a send to self. This update then undergoes the process described above so that all copies of the group-coherent memory are updated synchronously or nearly so within the time required for a single update to traverse the cluster.

[0029] The process disclosed herein does not prevent the use of semaphores based on the fast-priority message for maintaining flow control or other system functions described in the above-referenced patent disclosure. The process allows multiple simultaneous coherent updates from different hosts. Such updates will, perhaps, reach the DDM simultaneously where an arbitration mechanism, disclosed in the above-referenced patent application, will maintain message order such that each message arrives intact in a serial fashion. Precedence of simultaneous coherence updates to same mirrored location must be resolved at the application level by, perhaps, including a priority of some sort within the body of the message. Such issues do not alter the method or effectiveness of the disclosed process.

[0030] Embodiments of this disclosure can include a method of maintaining a coherent memory for an interconnect system having the capability to (1) define a coherent group and allocate a corresponding mirror memory area in each member of the group; (2) broadcast messages such a group such that they arrive at their destinations simultaneously or nearly so; and (3) such that the group member initiating the message also sends the same message to itself through the same mechanism. Embodiments of this disclosure can include a coherent memory update that takes place over the same DDS that is used to interconnect the tight cluster as disclosed in U.S. Ser. No. 62/216,999, filed Sep. 10, 2015 and U.S. Ser. No. 15/175,685, filed Jun. 7, 2016. Embodiments of this disclosure can include a method of coherent updates that are carried out by means of a special coherent group defined as in U.S. Ser. No. 62/216,999, filed Sep. 10, 2015 and U.S. Ser. No. 15/175,685, filed Jun. 7, 2016. Embodiments of this disclosure can include a method of coherent memory update across a tight cluster that is lock- and semaphore-free, wherein updates take place over a high-priority channel exclusively reserved for such updates. Embodiments of this disclosure can include a method of coherent updates that are initiated by a write request to a local copy of the coherent memory. Embodiments of this disclosure can include a coherent memory update that can be used in conjunction with existing tight-cluster interconnects as an additional, add-on system requiring another set of HBAs, connections, and the DDM adjacent to and parallel with an existing switched interconnect.

Definitions

[0031] The phrase end-to-end partitioning of message pathways is intended to mean partitioning of the message pathways from a CSME (computing, storage, or network element) to another CSME, for instance a priority channel from a computing element through a host-bus adapter through a data distribution module through another data distribution module then through another host-bus adapter and then to a storage element. The phrase multiple priority levels is intended to mean three or more priority levels, for instance five priority levels including a highest priority channel reserved specifically for fast priority messages and a channel reserved specifically for maintenance functions. The phrase above-referenced patent application is intended to mean U.S. Ser. No. 62/216,999, filed Sep. 10, 2015 and U.S. Ser. No. 15/175,685, filed Jun. 7, 2016. The terms program and software and/or the phrases program elements, computer program and computer software are intended to mean a sequence of instructions designed for execution on a computer system (e.g., a program and/or computer program, may include a subroutine, a function, a procedure, an object method, an object implementation, an executable application, an applet, a servlet, a source code, an object code, a shared library/dynamic load library and/or other sequence of instructions designed for execution on a computer or computer system).

[0032] The term uniformly is intended to mean unvarying or deviate very little from a given and/or expected value (e.g., within 10% of). The term substantially is intended to mean largely but not necessarily wholly that which is specified. The term approximately is intended to mean at least close to a given value (e.g., within 10% of). The term generally is intended to mean at least approaching a given state. The term coupled is intended to mean connected, although not necessarily directly, and not necessarily mechanically.

[0033] The terms first or one, and the phrases at least a first or at least one, are intended to mean the singular or the plural unless it is clear from the intrinsic text of this document that it is meant otherwise. The terms second or another, and the phrases at least a second or at least another, are intended to mean the singular or the plural unless it is clear from the intrinsic text of this document that it is meant otherwise. Unless expressly stated to the contrary in the intrinsic text of this document, the term or is intended to mean an inclusive or and not an exclusive or. Specifically, a condition A or B is satisfied by any one of the following: A is true (or present) and B is false (or not present), A is false (or not present) and B is true (or present), and both A and B are true (or present). The terms a and/or an are employed for grammatical style and merely for convenience.

[0034] The term plurality is intended to mean two or more than two. The term any is intended to mean all applicable members of a set or at least a subset of all applicable members of the set. The phrase any integer derivable therein is intended to mean an integer between the corresponding numbers recited in the specification. The phrase any range derivable therein is intended to mean any range within such corresponding numbers. The term means, when followed by the term "for" is intended to mean hardware, firmware and/or software for achieving a result. The term step, when followed by the term "for" is intended to mean a (sub) method, (sub)process and/or (sub)routine for achieving the recited result. Unless otherwise defined, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this present disclosure belongs. In case of conflict, the present specification, including definitions, will control.

[0035] The described embodiments and examples are illustrative only and not intended to be limiting. Although embodiments of the present disclosure can be implemented separately, embodiments of the present disclosure may be integrated into the system(s) with which they are associated. All the embodiments of the present disclosure disclosed herein can be made and used without undue experimentation in light of the disclosure. Embodiments of the present disclosure are not limited by theoretical statements (if any) recited herein. The individual steps of embodiments of the present disclosure need not be performed in the disclosed manner, or combined in the disclosed sequences, but may be performed in any and all manner and/or combined in any and all sequences. The individual components of embodiments of the present disclosure need not be combined in the disclosed configurations, but could be combined in any and all configurations.

[0036] Various substitutions, modifications, additions and/or rearrangements of the features of embodiments of the present disclosure may be made without deviating from the scope of the underlying inventive concept. All the disclosed elements and features of each disclosed embodiment can be combined with, or substituted for, the disclosed elements and features of every other disclosed embodiment except where such elements or features are mutually exclusive. The scope of the underlying inventive concept as defined by the appended claims and their equivalents cover all such substitutions, modifications, additions and/or rearrangements.

[0037] The appended claims are not to be interpreted as including means-plus-function limitations, unless such a limitation is explicitly recited in a given claim using the phrase(s) "means for" or "mechanism for" or "step for". Sub-generic embodiments of this disclosure are delineated by the appended independent claims and their equivalents. Specific embodiments of this disclosure are differentiated by the appended dependent claims and their equivalents.

What is claimed is:

1. A method, comprising operating a data distribution system including a data distribution module and a plurality of host-bus adapters coupled to the data distribution module including defining a coherent group that includes a set of members that includes the plurality of host-bus adapters;

providing a group-coherent memory area in each of the set of members; and

initiating a one-to-all broadcast message from a one of the plurality of host-bus adapters to all of the set of members when the one of the plurality of host-bus adapters requests a write to its local group-coherent memory area.

2. The method of claim **1**, wherein the one-to-broadcast message is transmitted on a priority channel.

3. The method of claim **2**, wherein the priority channel is lock-free and semaphore-free.

4. The method of claim **1**, further comprising ensuring that when the one of the plurality of host-bus adapters requests the write the one-to-all broadcast message maintains temporal memory coherency across all of the set of members of the coherent group including the one of the plurality of host-bus adapters that request the write.

5. A non-transitory computer readable media comprising executable programming instructions for performing the method of claim **1**.

6. An apparatus, comprising: a data distribution system including a data distribution module and a plurality of host-bus adapters coupled to the data distribution module, wherein operating the data distribution system includes

defining a coherent group that includes a set of members that includes the plurality of host-bus adapters;

providing a group-coherent memory area in each of the set of members; and

initiating a one-to-all broadcast message from a one of the plurality of host-bus adapters to each of the set of members ensuring that when the one of the plurality of host-bus adapters request a write, to update its local group-coherent memory area, the one-to-all broadcast message maintains temporal memory coherency across all of the set of members of the coherent group.

7. The apparatus of claim **6**, further comprising a computing, storage or networking element coupled to each of the at least two host-bus adapters.

8. The apparatus of claim **6**, further comprising another data distribution module coupled to the data distribution module.

9. A network, comprising the apparatus of claim **6**.

10. An interconnect fabric, comprising the apparatus of claim **6**

* * * * *