



(51) International Patent Classification:

G06F 9/455 (2006.01) G06F 13/00 (2006.01)
G06F 9/50 (2006.01) H04L 29/08 (2006.01)
G06F 12/06 (2006.01)

(21) International Application Number:

PCT/US2017/029478

(22) International Filing Date:

25 April 2017 (25.04.2017)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

62/327,438 25 April 2016 (25.04.2016) US

(71) Applicant: NETLIST, INC. [US/US]; 175 Technology,
#150, Irvine, CA 92618 (US).

(72) Inventors: LEE, Hyun; c/o Netlist, Inc., 175 Technology,
#150, Irvine, CA 92618 (US). RYU, Junkil; c/o Netlist,
Inc., 175 Technology, #150, Irvine, CA 92618 (US).

(74) Agent: ZHENG, Jamie, J. et al.; MASCHOFF BREN-
NAN, 1389 Center Drive, Suite 300, Park City, UT 84098
(US).

(81) Designated States (unless otherwise indicated, for every
kind of national protection available): AE, AG, AL, AM,
AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ,
CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO,
DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN,
HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KH, KN, KP, KR,
KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG,
MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM,
PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC,

(54) Title: METHOD AND APPARATUS FOR UNIFORM MEMORY ACCESS IN A STORAGE CLUSTER

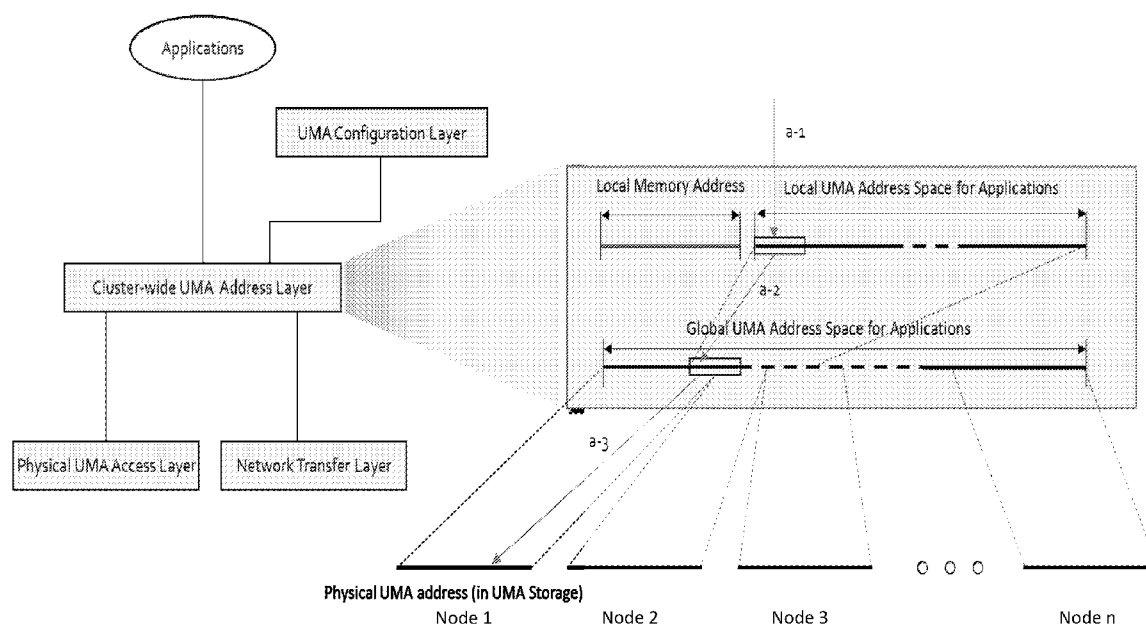


FIG. 8

(57) Abstract: The present application presents a Uniform Memory Access (UMA) network including a cluster of UMA nodes each having at least one UMA memory unit and a server local to the at least one UMA memory unit. A respective UMA memory unit in a respective UMA node comprises persistent memory; non-persistent memory, a node control device operatively coupled to the persistent memory and the non-persistent memory, a local interface for interfacing with the local server in the respective UMA node, and a network interface for interfacing with the UMA network. The node control device is configured to translate between a local unified memory access (UMA) address space accessible by applications running on the local server and a global UMA address space that is mapped to a physical UMA address space. The physical UMA address space includes physical address spaces associated with different UMA nodes in the cluster of UMA nodes.



SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR,
TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

- (84) Designated States** (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— with international search report (Art. 21(3))

Method and Apparatus for Uniform Memory Access in a Storage Cluster

FIELD

5 [0001] The present application is related to computer storage, and more particularly to methods and apparatus for uniform memory access (UMA) of a storage cluster in a computer network.

BACKGROUND

10 [0002] The Non Uniform Memory Access (NUMA) network architecture used in a server network produces unpredictable memory access time in the server network. In the NUMA architecture, each server (node) has its own set of memories controlled by that server. Other nodes (servers) can access the memory via the RDMA (Remote DMA) access with permission from the owner (server) of the memory. The access route needs
15 to go through a memory access controller (scheduler) under the control of the owner server, resulting in non-uniform memory access time and bandwidth. Furthermore, if the owner server fails, all memory accesses to its associated set of memories are blocked.

[0003] For an example, a NUMA network using Intel QPI (Quick Path Interconnect) presents significant degradation of memory access bandwidth when a
20 server accesses the memory owned by other servers. In some instances, the memory bandwidth can be degraded by as much as 60% when a server accesses the memory owned by its next neighboring server, and by 80% when the server needs to access the memory owned by a server that is next to the next neighboring server (so called “one-removed”).

25

SUMMARY

[0004] The present application presents apparatus and methods to support a novel shared memory network, i.e., the Uniform Memory Access (UMA) network, which removes non-uniformity in memory accesses across the network. According to certain
30 embodiments, the UMA network includes a cluster of UMA nodes. Each UMA node includes at least one UMA memory unit and may further include a server local to the at least one UMA memory unit. A respective UMA memory unit at a respective UMA node comprises persistent memory, non-persistent memory, a node control device operatively

coupled to the persistent memory and the non-persistent memory, a local interface for interfacing with the local server in the respective UMA node, and a network interface for interfacing with the UMA network.

[0005] In certain embodiments, the node control device comprises a set of control logic implemented using electronic circuits and software or firmware, including physical UMA access logic, cluster-wide UMA address management logic, and network transfer logic. The physical UMA access logic is configured to facilitate and control accesses by the local server of the persistent memory and/or the UMA network via the local interface. The cluster-wide UMA address management logic is configured to translate between a local unified memory access (UMA) address space and a global UMA address space.

[0006] In certain embodiments, the local UMA address space is accessible by applications running on the local server, while the global UMA address space is mapped to a physical UMA address space. The physical UMA address space includes physical address spaces associated with different UMA nodes in the cluster of UMA nodes. The network transfer logic is configured to communicate with the other nodes in the cluster of nodes command and data packets via the second interface. Thus, any server in the UMA network can access the physical address spaces associated with other UMA nodes across the UMA network without going through the servers at the other UMA nodes.

[0007] In certain embodiments, the global UMA address space is a shared address space managed by each individual UMA memory unit. The global UMA address space is not directly visible by the servers in the network, but each server can send a memory access request to the network to request access of any of the UMA memory units, as discussed above. Each server in the network can also broadcast a memory allocation request to the cluster. One or more UMA memory units, upon receiving the memory allocation request, would review its available memory space, and may respond with a memory allocation grant packet including an associated UMA address. The UMA memory units would also monitor the network for the memory allocation grant packet. If one UMA memory unit responds with such a packet, other UMA memory units would withdraw any pending response.

[0008] The requesting UMA node (requester) remembers the network address of the responder as a part of the UMA address, and it can manage the allocated UMA memory space without actually knowing the physical location or the memory density of the UMA memory unit. Thus, the UMA memory space can be constructed using a

combination of memories having different densities, different memory types (e.g., hard disk drives (HDD), solid state disc drives (SDD), dynamic random access memory (DRAM), etc.).

[0009] In certain embodiments, the UMA address space is a two-layer address
5 space, with a top layer comprised of the UMA network addresses of the UMA memory units in the cluster of UMA nodes, and a second layer comprised of the UMA memory space in each of cluster of UMA nodes. The top-layer network addresses can be visualized by the servers (requesters) connected to the UMA network. The second-layer address space is not seen by the requesters. However, each requester can access the
10 second layer address space using an offset address.

BRIEF DESCRIPTION OF THE DRAWINGS

[0010] FIG. 1 is a block diagram of a conventional server network in which each server in the network is associated with memory controlled by the each server.

15 [0011] FIGS. 2A-2D are block diagrams of exemplary UMA networks according to certain embodiments.

[0012] FIG. 3 is a block diagram of a plurality of network switches coupled to respective UMA nodes in a UMA network according to certain embodiments.

[0013] FIG. 4 is a flowchart illustrating a method performed by a network switch
20 coupled to a UMA node according to certain embodiments.

[0014] FIG. 5 is a block diagram of a UMA node according to certain embodiments.

[0015] FIGS. 6A and 6B are block diagrams of exemplary UMA memory units according to certain embodiments.

25 [0016] FIG. 7 is a block diagram of a UMA node control device according to certain embodiments.

[0017] FIG. 8 is a diagram illustrating address translation in a UMA network according to certain embodiments.

[0018] FIG. 9 is a flowchart illustrating a method performed by a UMA node
30 according to certain embodiments.

[0019] FIG. 10 is a flowchart illustrating a UMA process according to certain embodiments.

[0020] FIGS. 11A-11E are block diagrams of various UMA packets according to certain embodiments.

[0021] FIGS. 12A-12B are diagrams illustrating how UMA storage area for an application running on a server is distributed across different UMA nodes according to certain embodiments.

[0022] FIG. 13 is a diagram illustrating a process of inserting a new UMA storage device into a UMA network according to certain embodiments.

[0023] FIG. 14 is a diagram illustrating a process of allocating a new UMA address in a UMA network according to certain embodiments.

[0024] FIG. 15 is a diagram illustrating a scale-out storage system provided by the UMA storage devices in the UMA network according to certain embodiments.

DETAILED DESCRIPTION OF THE EMBODIMENTS

[0025] FIG. 1 is a block diagram of a conventional server network in which each server in the network is associated with memory controlled by the each server. FIG. 1 shows 4 servers, server 1, server 2, server 3 and server 4, each with local memory (MEM) that are connected via Intel QPI. When server 1 accesses the memory controlled by server 4, it needs to go through server 2 and server 4, as illustrated by the access path in dotted lines. Thus, the performance of server 1 is dependent on the memory access speed of server 4, which in turns depends on how busy server 4 is at the time. The performance of server 1 is also dependent of the data relay rate of server 2, which is dependent on how busy the central processing unit (CPU) of server 2 is at the time. In other words, the performance of server 1 is gated by what is going on with each server on the access path to the remote memory where the data is located. The UMA networks described below eliminate such limitations on the performance of a server in a network by removing the server CPU control from the network access path.

[0026] FIGS. 2A-2D are block diagrams of exemplary UMA networks according to certain embodiments. FIG. 2A shows a UMA network 200a having a cluster of 4 servers, server 1, server 2, server 3 and server 4, coupled to their respective UMA memory units, UMA1, UMA2, UMA3, and UMA4, respectively, and UMA access paths (illustrated using dotted lines) that run between the UMA memory units without going through any of the servers. In this case, any UMA memory unit can be directly accessed by any server in the cluster without going through another server. For example, server 1 can access UMA4, which is server 4's local memory, via the UMA network without going through server 2 or server 4. In a UMA network, a UMA memory unit, by itself or together with a local server, is sometimes referred to as a UMA node.

[0027] The UMA nodes in a UMA cluster can be connected with each other in any of a variety of configurations. For example, FIG. 2B illustrates a UMA cluster 200b where the UMA nodes are directly connected with each other via a daisy chain, and FIG. 2C illustrates a UMA cluster 200c where the UMA nodes are directly connected with each other via a star-network. The UMA nodes in a UMA cluster 200d may also be connected with each other via a local or wide-area network 201, as shown in FIG. 2D. Network 201 can be, for example, a packet-based network, such as the Internet. In certain embodiments, multiple UMA clusters can be included in a UMA network to support a large UMA memory space.

[0028] The UMA networks according to certain embodiments can be operated with various arbitration types, such as “first come – first serve,” token ring, or other type of priority assignment, but there is no favor or a preset priority arrangement amongst the UMA nodes.

[0029] Thus, by removing server CPU control from network access paths, there is no preferred or designated access controller for any given UMA memory unit in a UMA network. Thus, each UMA unit schedules memory accesses by remote servers along with memory accesses by its local server, resulting in much faster data transfers across the network.

[0030] In certain embodiments, each UMA node has a network interface, which can be embedded in or attached to the UMA node and may include a network switch/repeater for adding more UMA nodes to the network to expand servers or UMA nodes. The media access control MAC in the network switch generates multiple sets of data bus pairs that connect to the UMA nodes. FIG. 3 is a block diagram of a plurality of network switches/repeaters coupled to (or embedded in) respective UMA nodes in a UMA network according to certain embodiments. FIG. 4 is a UMA network interface operational flowchart according to certain embodiments. As shown in FIG. 4, when the network interface receives (410) a UMA memory access packet, it determines (420) a target UMA node for the UMA memory access packet. If the target UMA node is the UMA node local to the network interface, it sends (430) the packet to the local UMA node. Otherwise, it repeats (450) the packet to the next network switch.

[0031] FIG. 5 is a block diagram of a UMA node 500 according to certain embodiments. As shown in FIG. 5, the UMA node includes server components, such as a processor, graphics card, input/output (I/O) interface, memory controller, memory bus, I/O controller, peripheral component interconnect (PCI) or PCI express (PCIe) interface,

etc. The UMA node also includes memory components, such as one or more memory modules coupled to the memory bus and storage coupled to the PCI or PCIe interface. Any one or more of the memory modules and the storage could be a UMA memory unit.

[0032] FIGS. 6A and 6B are block diagrams of exemplary UMA memory units

5 600, 600' according to certain embodiments. As shown in FIG. 6A, the UMA memory unit 600 includes a persistent storage memory unit 610 (e.g., Flash memory, hard disk drives (HDD), etc.), a non-persistent buffer memory unit 620 (e.g., DRAM); and a node control device 630 operatively coupled to the persistent memory unit 610 and the non-persistent memory unit 620. The UMA memory unit further includes a local interface
10 640 for interfacing with the server in the same UMA node (local server) via the server's memory bus or PCI/PCIe interface, and a UMA node interface 650 (or network interface) for interfacing with the UMA network through a backplane network connection 655. The UMA network interface 650 can be a standard Ethernet interface, optical interface, or any type of communication method including a serial bus interface.
15 The UMA memory unit further include control signal paths 671, 672, 673 (illustrated as solid lines), data signal paths 681, 682, 683 (illustrated as dashed lines), and network packet paths 691, 692 (illustrated as dash/dot lines).

[0033] The non-persistent buffer memory (DRAM) 620 in the UMA node is used as a buffer to queue read/write data from/to the persistent memory (storage) 610 so as to
20 match the data flow rate with the network interface. In certain embodiments, the UMA network interface 650 can be used by the local server to concurrently access multiple UMA nodes in the UMA network. So, the data transfer rate for the the UMA network interface can be much faster than the data retrieval rate at each individual UMA node.

[0034] In certain embodiments, as shown in FIG. 6B, the UMA memory unit 600' can be implemented as a dual inline memory module (DIMM) having a module board
25 601, with its components (i.e., the persistent memory unit 610, the non-persistent memory unit 620, and the node control device 630) mounted on the module board 601. The local interface 640 includes a main memory interface 641 having edge connections (not shown) on the module board for inserting into a module slot on a motherboard of
30 the local server so as to couple the memory module to the memory bus 660 of the local server. The UMA DIMM further includes a main memory component, which in turn includes dynamic random access memory (DRAM) and a DIMM control device (or register device). The local interface 640 further includes a data signal router 642 controlled by the UMA node control device 630 via control signal path 674. The data

signal router routes data signals between the memory bus 660 and the main memory component, or between the main memory and the UMA node control device 630, or between the memory bus 660 and the UMA node control device 630, depending on whether the local server is accessing the main memory or the persistent memory unit on this particular UMA DIMM or on any other UMA memory unit in the UMA network, as determined by the UMA node control device 630.

[0035] FIG. 7 is a block diagram of a UMA node control device 630 according to certain embodiments. As shown in FIG. 7, the UMA node control device 630 comprises a set of control logic including physical UMA access logic 710, network transfer logic 720, cluster-wide UMA address management logic 730; and UMA configuration logic 750. The physical UMA access logic is configured to control access by the local server of the persistent memory unit via the local interface. The network transfer logic 720 is configured to facilitate packet-based communication between this UMA node and the other UMA nodes in the cluster of UMA nodes.

[0036] Thus, the UMA network 200 or a UMA storage system (e.g., the UMA memory unit 600, 600') uses a UMA protocol that includes a physical UMA access layer, a Network transfer layer, a Cluster-wide UMA address layer, and a UMA configuration layer. The physical UMA access layer is a protocol to access the physical local UMA storage device. The Network transfer layer is the interface protocol to communicate among UMA storage nodes in the network, which can be IP, FC, PCI, etc. Media access protocol and traffic control protocol in the network transfer layer can follow existing standards, but the protocol for UMA command and data packets are defined proprietarily. The network transfer layer can be implemented in the UMA memory units (or UMA storage device) or separately from the UMA storage device. However, the UMA storage system includes software that recognizes both the UMA storage function and Network function and registers them into its logic for its usage.

[0037] In certain embodiments, applications running on the local server have access to a local memory address space corresponding to the memories associated with the local server and a local UMA address space corresponding the UMA storage space provided by the UMA cluster. The UMA node control monitors memory access requests from the local server via the control signal paths 671 and determines, for each memory access request, whether it is directed at the local memory space provided by, for example, the main memory, or the local UMA address space corresponding to memory units in the cluster of UMA nodes. FIG. 8 is a diagram illustrating address translation in

a UMA network according to certain embodiments. As shown in FIG. 8, upon determination that the local server has requested access (a-1) to the local UMA address space, the cluster-wide UMA address management logic 730 is configured to translate (a-2) between the local UMA address space and a global UMA address space. The global
5 UMA address space is then mapped (a-3) to a physical UMA address space. In certain embodiments, the physical UMA address space includes separate physical address spaces associated with respective UMA nodes (e.g., Node 1, Node 2, ..., Node n) in the cluster of UMA nodes. Thus, the UMA node control device provides access of the physical address spaces associated with the other UMA nodes for the applications
10 running on the local server.

[0038] In certain embodiments, the physical UMA access logic includes a UMA node arbiter that arbitrates the Flash/storage accesses either from the UMA network or from the local server (the system) via the main memory. The non-persistent buffer memory (DRAM) in the UMA node is not directly accessible by the system since its
15 address space is not defined in the local server main memory address space, but it is defined in the local UMA address space. The arbiter schedules the storage access in the UMA node based on fairness and quality of service (QoS) such that all access requests are given the same level of priority unless the requester requests for a higher priority.

[0039] The higher priority requests are processed such that a few initial packets
20 are sent out based on the elevated priority, but the priority level reduces to a lower level as more data packets are transferred. One other possible way of achieving fairness and QoS is use a credit based access priority. For example, all servers have the same number of priority credits that renews periodically, and a preset priority credits is assigned to each priority level. A server request for a UMA node access with higher priority needs to
25 spend more priority credit, and the server cannot request any more higher priority accesses when all priority credits are used up. It is also possible that the unused priority credits can be transferred to the next priority credit renewal period, or transferred up to other servers.

[0040] In certain embodiments, the local server can access the local UMA
30 memory unit without going through the network interface, but the access request still needs to be arbitrated by the UMA node arbiter. FIG. 9 is a flowchart illustrating a method 900 performed by the UMA node arbiter according to certain embodiments. As shown in FIG. 9, upon receiving (910) a access request from the local server, the UMA node arbitor cooperates with the cluster-wide UMA address logic to check (920) the

memory address/location associated with the access request, and determines (930) whether the memory address/location is in the local memory address space. In response to the the memory address/location being in the local memory address space, the UMA arbitor coordinates (950) local memory/storage access. In response to the the memory address/location not being in the local memory address space, the UMA arbitor starts the process of constructing (940) and processing UMA network packets.

[0041] FIG. 10 is a flowchart illustrating a UMA network communication process 1000 (referred sometimes in as “UMA process”) according to certain embodiments. This flow chart applies to various UMA access cases regardless of the network connection type or the network arbitration type. The UMA process 1000 starts with a server at a source UMA node (requester) requesting access to a memory address/location that is not in its local memory address space. In response, the UMA node control at the source UMA node constructs a UMA request packet and transmits (1010) the request packet to the UMA network via the UMA network interface at the source UMA node.

[0042] FIG. 11A is a block diagram of an exemplary UMA request packet 1110 according to certain embodiments. As shown in FIG. 11A, the request packet includes information such as a packet type (e.g., UMA request), a UMA node address, a source (requester) address, an access type, access priority, total byte size of data to be transferred, and error detection code (e.g., cyclic redundancy check (CRC) code), etc. The UMA node address in the request packet can be a specific UMA node address or a generic UMA node address. The access type can be Read start, Write_start, Status request, Standby, etc. The Standby access type is used to inform another UMA node to execute a ‘malloc’ operation for the total number of bytes. After the Standby access type, the requester may send another request packet for a completion status packet to ensure that the other UMA node has prepared the memory space. The requester may send multiple Write_Start type of request packets to the reserved UMA memory space.

[0043] After sending out the request packet, the source UMA node waits for a response from a target UMA node. When a generic UMA node address is sent out with a request packet, any available UMA node can respond. Thus, there could be more than one UMA node that responds to the request packet with a generic UMA node address. To minimize multiple responses from different UMA nodes, each UMA node listens to the network traffic, and if it detects any responses including requester address with zero total number of bytes and zero data packets transferred, it ignores the request packet since it is likely that the response is from a closest UMA node. In case the requester receives more

than one response, it can choose a UMA node as the target UMA node, and send out a request packet with the target UMA node.

[0044] In certain embodiments, the request packet is repeated until the target UMA node responds with an acknowledge packet indicating that it is available. Upon receiving (1020) such an acknowledgement packet from the target UMA node, the source UMA node sends out (1030) a command packet to tell the target UMA node to prepare for data packet(s). The target UMA node needs to ensure that the available input buffer is large enough to receive/send the total number of data bytes that are transferred with one or more data packets. FIG. 11B is a block diagram of an exemplary UMA command packet 1120 according to certain embodiments. As shown in FIG. 11B, the command packet 1120 includes information such as a packet type (e.g., UMA command), the target UMA node address, the source (requester) address, a start UMA memory address at the target UMA node, a CRC on/off indicator, dynamic real-time commands, and associated data size and CRC code, etc. The data size in the command packet can be the total number of data bytes to be transferred in response to the command packet.

[0045] The data size in the command packet is used to maximize transfer efficiency such that the requester can modulate the data packet size to maximize the data transfer success rate. For example, in a noisy system, the requester may reduce the packet size to optimize the bit error rate per transfer. The dynamic real time command is used for transfer interrupt, pause, stop, terminate, retransmit, etc. For example, if the requester receives a data packet with an error, it can send a command packet to the UMA node to retransmit the data from the start memory address and data packet size.

[0046] After transmitting the command packet to the target UMA node, the requester either sends or receives (1040) data packets. FIG. 11C is a block diagram of an exemplary UMA data packet 1130 according to certain embodiments. As shown in FIG. 11C, the data packet 1130 includes information such as a packet type (e.g., data), the target UMA node address, the source (requester) address, a transfer sequence number, data, and CRC code, etc. The transfer sequence number may be used to indicate an order of the data packet in a sequence of data packets associated with the command packet.

[0047] In certain embodiments, the target UMA node performs a UMA process that mirrors the UMA process at the source UMA node. As shown in FIG. 10, upon receiving (1010) the request packet, the target UMA node determines (1020) whether the UMA node address indicates that the request packet is directed at the target UMA node.

If the answer is no, it goes back to attend to the next request packet. If the answer is yes, it responds to the requester that it is ready and receives (1030) from or transmits to the requester the command packet. Afterwards, the target UMA node transmits or receives (1040) the data packets.

5 [0048] The sender of the data packets transmits (1050) a completion status packet to the UMA network after transmitting the sequence of the data packets to the UMA network. For example, if the access type is for a memory write access, the completion status packet is generated by the requester. On the other hand, if the access type is for a memory read access, the completion status packet is generated by the target UMA node,
10 which has sent the data packets. FIG. 11D is a block diagram of an exemplary completion status packet 1140 according to certain embodiments. As shown in FIG. 11D, the completion status packet 1140 includes information such as a packet type (e.g., completion status), the target UMA node address, the source (requester) address, a total number of bytes transferred, a number of data packets transferred, and CRC code, etc.

15 [0049] The receiver of the completion status packet sends back 1060 the acknowledge packet. FIG. 11E is a block diagram of an exemplary acknowledge packet 1150 according to certain embodiments. As shown in FIG. 11E, the acknowledge packet 1150 includes information such as a packet type (e.g., acknowledge), the target UMA node address, the source (requester) address, a total number of bytes transferred, a
20 number of data packets transferred. The acknowledge packet may also include static information such as a total transfer time, average data payload in the data packets, error rate, sequence number(s) of any data packets that are erroneous (e.g., (write data packet with CRC error), and CRC code, etc. For example, in the case of memory write, if the requester receives an error data packet sequence number, it can retransmit with the
25 request packet after it receives the completion status packet, or just retransmit the command packet followed by the data packet that was reported as an error packet. In this case the command packet will indicate that it is a retransmit. In certain embodiments, if the requester needs to retransfer the error packets, it needs to restart with a request packet.

30 [0050] The requester UMA node completes the data transfer activity when it transmits (in the case of memory read) or receives (in the case of memory write) the acknowledge packet without any indication of error. The target UMA node completes the data transfer activity when it receives (in the case of memory read) or transmits (in the case of memory write) the acknowledge packet.

[0051] Thus, according to certain embodiments, applications running on a server or computing system in the UMA network can have its associated storage area distributed across UMA storages in different UMA nodes (computing nodes), as shown in FIGS. 12A-12B. As described above, the network functionality of the storage devices can be included internally in the UMA storages or externally in the Node system. One example of the physical interface between the UMA storage and the computing system can be memory interface (e.g., DIMM), PCIe interface and etc., as shown in FIG. 12A, where Application A uses UMA storage area A-#0 in computing node #0, UMA storage area A-#1 in computing node #1, and UMA storage area A-#2 in computing node #2. Hence, the UMA storage area for application A is $(A\text{-}\#0) + (A\text{-}\#1) + (A\text{-}\#2)$.

[0052] Another example of the physical interface between the UMA storages and the computing system can be PCI, Internet Protocol (IP), Fiber Channel (FC) network etc., wherein the UMA storages are separate from their associated computing systems, and are connected to the computing systems via a network, as shown in FIG. 12B. In this example, a cluster-wide memory address can be used.

[0053] Thus, the cluster-wide UMA address layer provides a single global address space among all UMA storage devices in the network and functions to translate between the Local UMA address for applications and the cluster-wide UMA address, or between the cluster-wide UMA address and the physical UMA address. For example, when an application accesses a part of the local UMA address space (a-1 in FIGS. 12A and 12B), the cluster-wide UMA address layer translates the local UMA address to a specific cluster-wide address and confirms which nodes have that specific cluster-wide address. Then, a request of the application is transferred to the designated target physical UMA addresses.

[0054] FIG. 13 is a diagram illustrating a process of inserting and registering a new UMA storage device into a UMA network according to certain embodiments. For example, when a new UMA storage device is inserted (1), the physical UMA access layer detects (2) the new UMA storage device and informs this detection to the UMA configuration layer. The UMA configuration layer recognizes (2) this detection and causes a user setup menu to be displayed. A user can determine how to allocate a cluster-wide UMA address space to the new device and causes a request for this address space to be generated. The cluster-wide UMA address layer broadcasts (4) the address reservation request, and each UMA storage responds (4). Then, the cluster-wide UMA address layer determines (5) the cluster-wide UMA address space for the new UMA storage device,

and the UMA configuration layer registers and broadcasts (6) this updated information to the UMA configuration layer in other nodes. In certain embodiments, the UMA configuration layer is used to define how new UMA storages join an existing UMA storage cluster, and how to allocate UMA address for a local application or computing node. For example, when a new UMA storage device joins, its address space can be allocated continuously or into separate groups in a global UMA address space. QoS and locality can be considered when a global UMA address space is allocated for local application of each node.

[0055] FIG. 14 is a diagram illustrating a process of allocating a UMA address for an application or a node in a UMA network according to certain embodiments. As shown in FIG. 14, after an application or node requests (1) a UMA storage area, the UMA configuration layer sends (2) this request to designated UMA storages, and the correspondent UMA storages would send responses (3). After receiving the responses, the UMA configuration layer sends a message "allocation completed" to the requestor. The UMA configuration layer then broadcasts this updated information to all UMA storages. In certain embodiments, the UMA configuration layer allows the user some choices (such as QoS, locality, etc.) on how the request is handled.

[0056] FIG. 15 is a diagram illustrating a scale-out storage system provided by the UMA storage devices in the UMA network according to certain embodiments. A Scale-out storage system can be made because the UMA storage system provides a cluster-wide UMA address space. To build a scale-out storage system based on the UMA storage devices, a (scale-out) storage device driver is devised. This storage device driver exports a block device interface to the file systems, data base system, and applications. In this example, each application does not need to know the entire UMA storage (cluster) address space. The UMA storage (cluster) address is managed by the UMA configuration layer, and each application only needs to manage its own allocated (targeted) storage space since there is no overlapping of the physical UMA address, and there is one-to-one mapping between the physical UMA address and the targeted address by each application node.

We Claim:

1. A memory module operable with a server in a network, the memory unit and the server forming a first node in a cluster of nodes in the network, the memory unit comprising:

5 persistent memory providing data storage associated with the first node;
non-persistent memory buffering data to or from the persistent memory;
a node control device operatively coupled to the persistent memory and the non-persistent memory;

10 a first interface for interfacing with the server; and
a second interface for interfacing with the network;

wherein the node control device comprises control logic including physical memory access logic, cluster-wide address management logic, and network transfer logic;

wherein the physical memory access logic is configured to control accesses by the server of the persistent memory via the first interface;

15 wherein the cluster-wide address management logic is configured to translate between a local unified memory access (UMA) address space and a global UMA address space, the local UMA address space being accessible by applications running on the server, the global UMA address space being mapped to a physical UMA address space, the physical UMA address space being associated with data storages in the cluster of
20 nodes;

wherein the network transfer logic is configured to communicate with the other nodes in the cluster of nodes command and data packets via the second interface.

2. The memory module of claim 1, wherein first interface includes a main memory interface coupled to a memory bus in the first server.

25 3. The memory module of claim 1, wherein the first interface is coupled to one of a PCIe interface, a Fiber Channel (FC) interface, and an Internet Protocol (IP) interface in the first server,

4. The memory module of claim 1, wherein the second interface is an Internet Protocol (IP) interface or a Fiber Channel (FC) interface.

30

5. The memory module of claim 1, wherein the cluster-wide address management logic is configured to determines a target node among the cluster of nodes in response to a first request from the server to access data storage at another node among the cluster of nodes.

6. The memory module of claim 5, wherein, in response to the first request, the network transfer logic is configured to form a request packet including a network address for the target node, and to transmit the request packet to the network via the second interface.

5 7. The memory module of claim 6, wherein the request packet includes an access type, and a size of data to be transferred in response to the first request.

8. The memory module of claim 6, wherein the network transfer logic is further configured to, in response to the target node acknowledging receipt of the request packet, form a command packet including the network address of the target node, a storage
10 address at the target node, and one or more dynamic real-time commands, and to transmit the command packet to the network via the second interface.

9. The memory module of claim 8, wherein the network transfer logic is further configured to communicate one or more data packets with the target node via the second interface subsequent to transmitting the command packet, each of the one or more data
15 packets identifying the target node, and the physical memory access logic is configured to communicate data in the one or more data packets with the first server.

10. The memory module of claim 9, wherein, subsequent to communicating the one or more data packets, the network transfer logic is further configured to transmit to or receive from the target node a completion status packet via the second interface.

20 11. The memory module of claim 10, wherein, subsequent to transmitting to or receiving from the target node the completion status packet, the network transfer logic is further configured to receive from or transmit to the target node an acknowledge packet.

12. The memory module of claim 1, wherein the control logic is configured to:
receive a request packet from the network via the second interface, the request
25 packet including a target node address and a requester address; and

in response to the target node address indicating that the first node is a target node for the request packet, receive a command packet from the network via the second interface, communicate one or more data packets with the network via the second interface in accordance with the command packet, and cause data in the one or more data
30 packets to be buffered in the non-persistent memory, the data being read from or written into the persistent memory.

13. The memory module of claim 1, wherein the control logic is configured to:
receive a request packet from the network via the second interface, the request
packet including a target node address and a requester address; and in response to the

target node address indicating that the first node is not a target node for the request data packet, repeat the request packet to a next node in the cluster of nodes.

14. The memory module of claim 1, wherein the control logic further includes configuration logic, and wherein the configuration logic is configured to modify the local
5 UMA address space and the global UMA address space in response to a new node being inserted into the cluster of node.

15. The memory module of claim 14, wherein the configuration logic is further configured to broadcast a reservation of a new space for the new node in the global UMA address space to other nodes in the cluster of nodes.

10 16. The memory module of claim 1, wherein the control logic further includes configuration logic, and wherein the configuration logic is configured to:

receive from the server a request for allocating a new UMA address to an application running on the server;

send the request to a designated node via the second interface;

15 receive a response from the designated node that the new UMA address has been allocated to the application;

inform the server or application that the new UMA address has been allocated to the application; and

20 broadcast to other nodes in the cluster of nodes that the new UMA address has been allocated to the application.

17. The memory module of claim 1, wherein the physical memory access logic is further configured to control accesses by the server of memory spaces in other nodes in the cluster of nodes.

18. The memory module of claim 1, further comprising a main memory unit, wherein
25 the first interface includes a main memory interface coupled to a memory bus in the first server and wherein the first interface further includes a data signal router coupled to the main memory interface and controlled by the node control device.

19. The memory module of claim 18, wherein the data router is configurable by the
30 node control device to route data signals between the memory bus and the main memory unit, or between the main memory unit and the node control device, or between the memory bus and the node control device, depending on whether the server is accessing the main memory unit, or the persistent memory unit, or memory space in another node in the cluster of nodes.

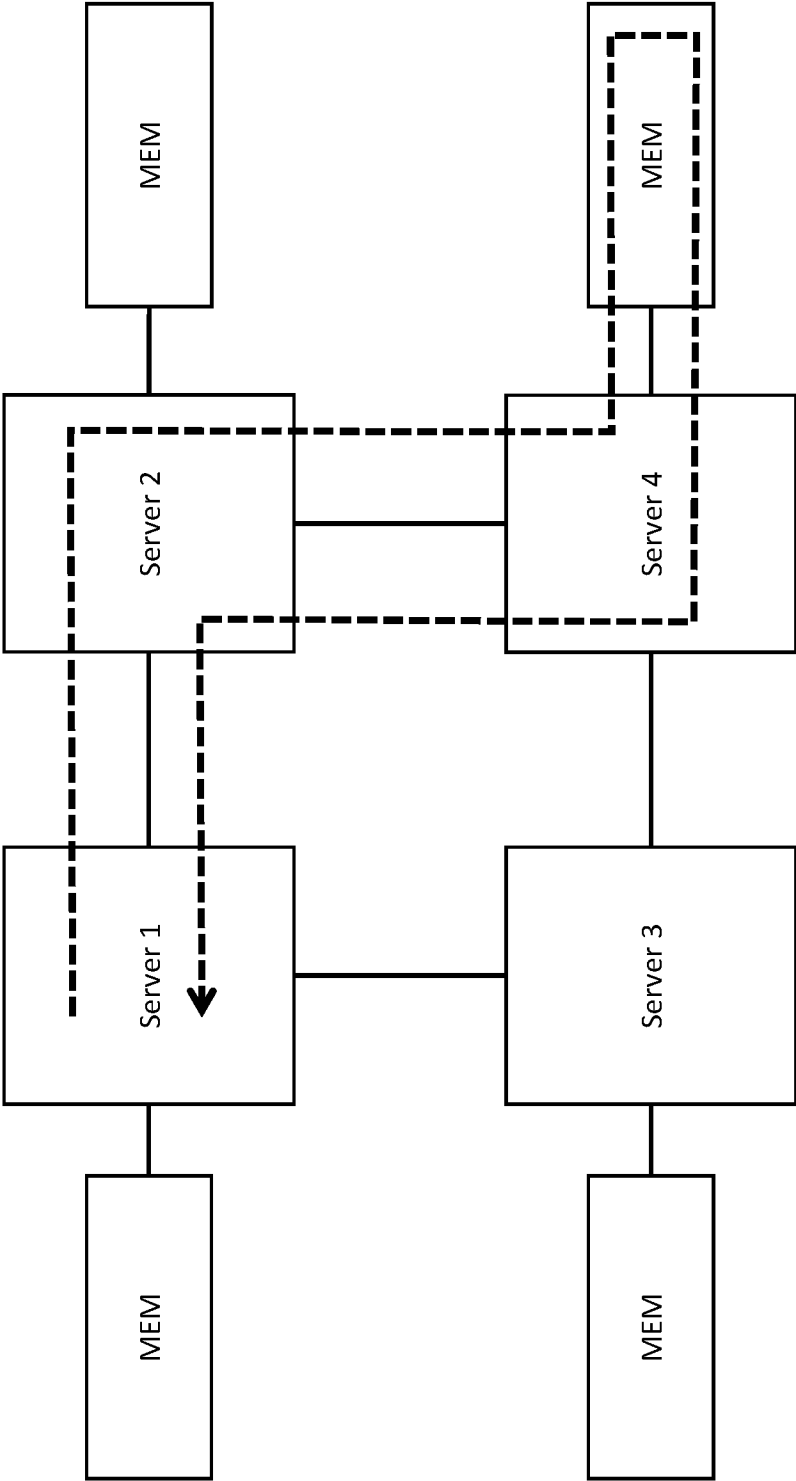


FIG. 1
(Prior Art)

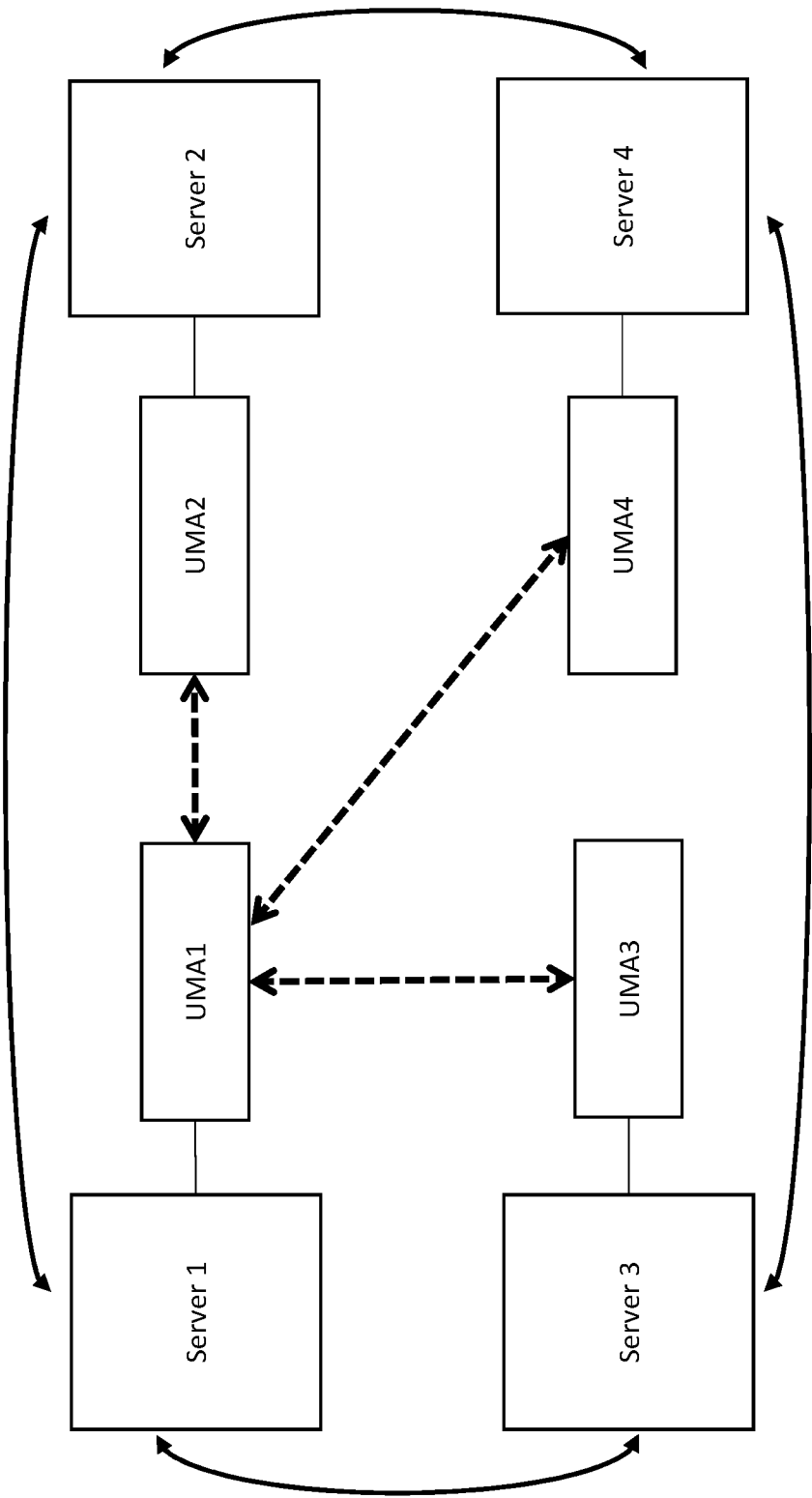


FIG. 2A

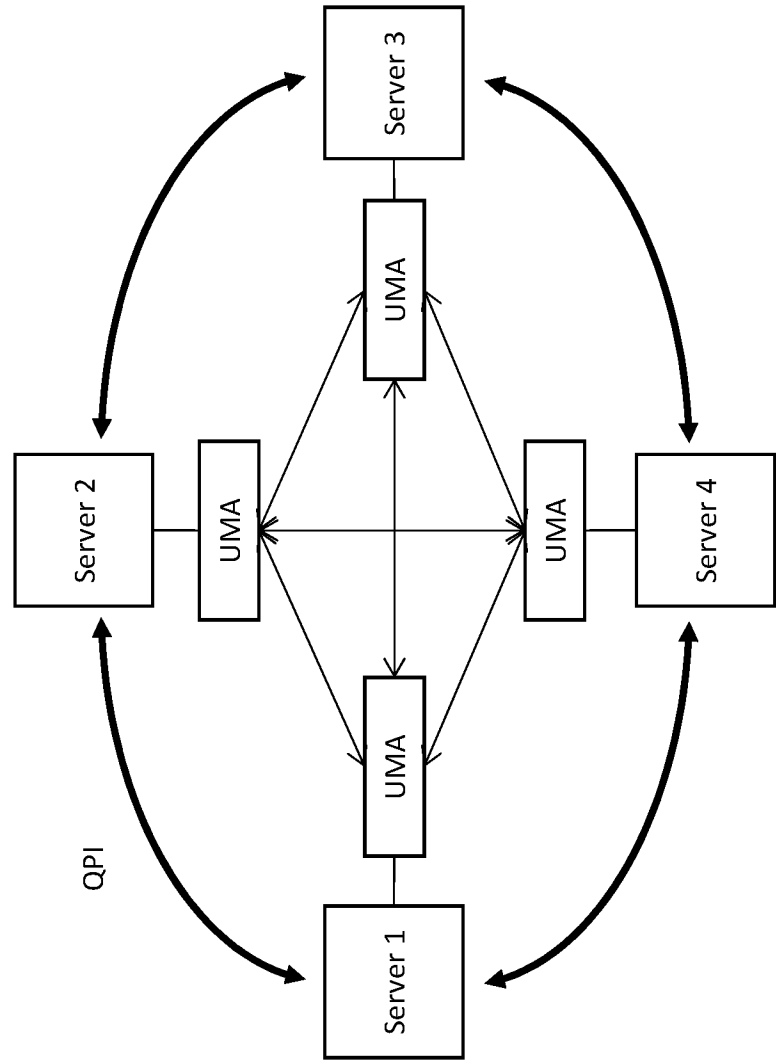


FIG. 2C

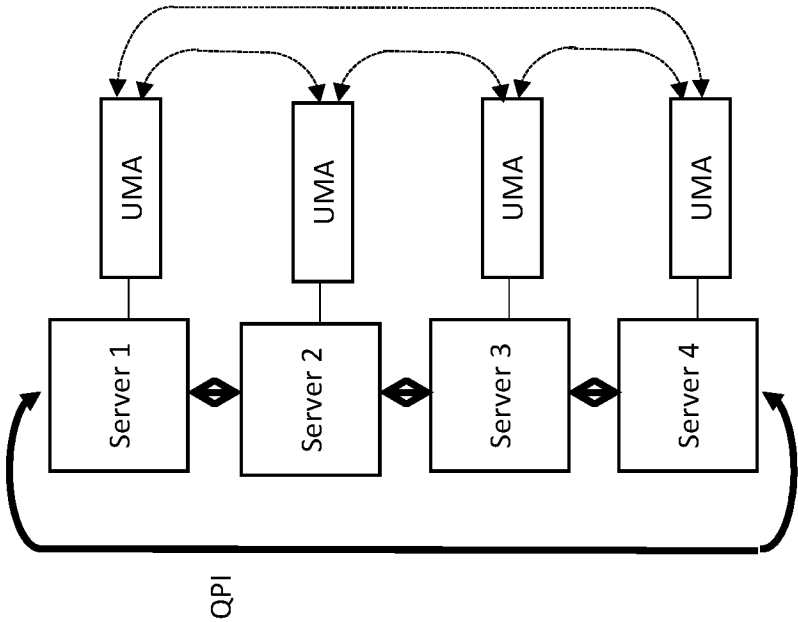


FIG. 2B

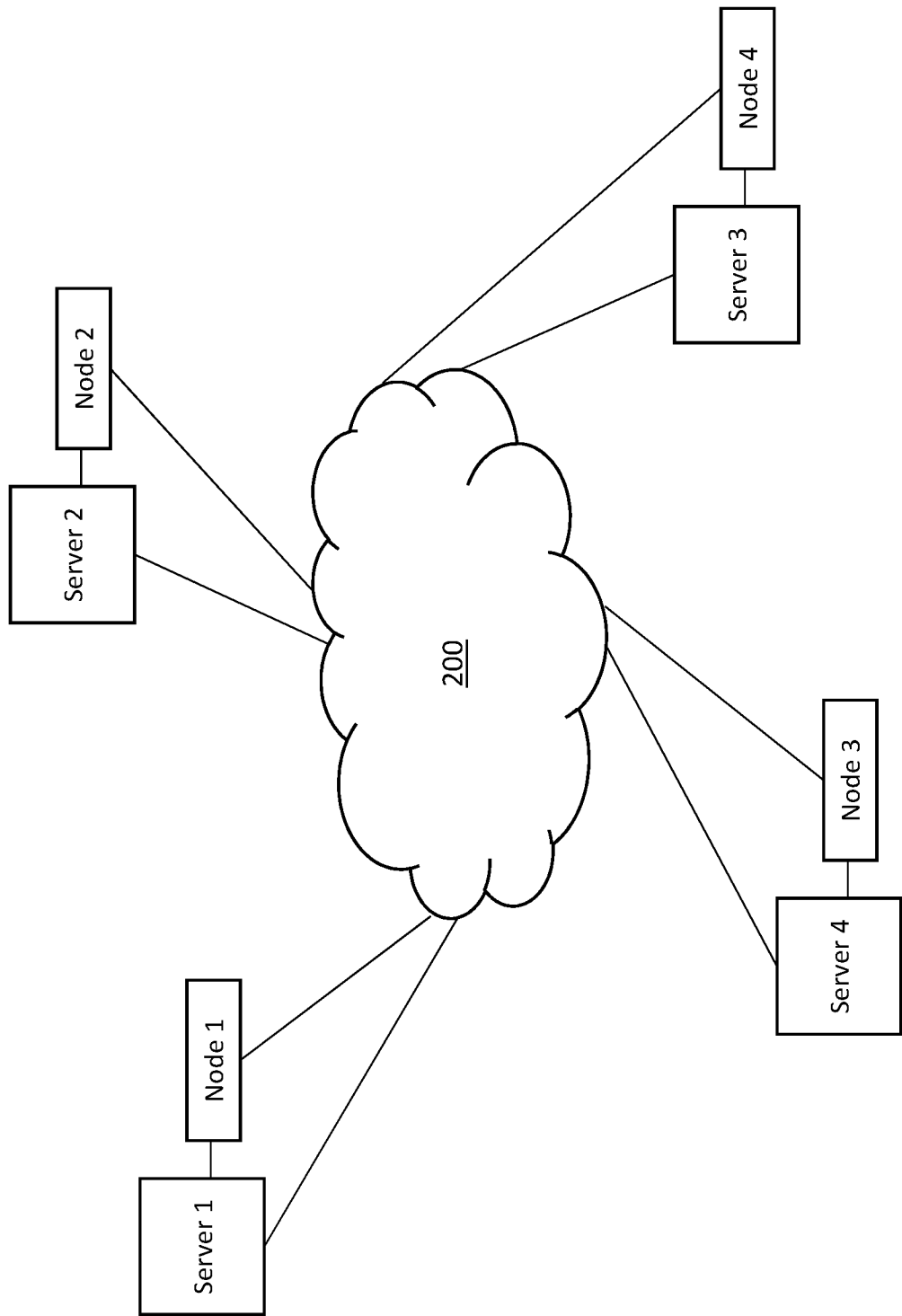


FIG. 2D

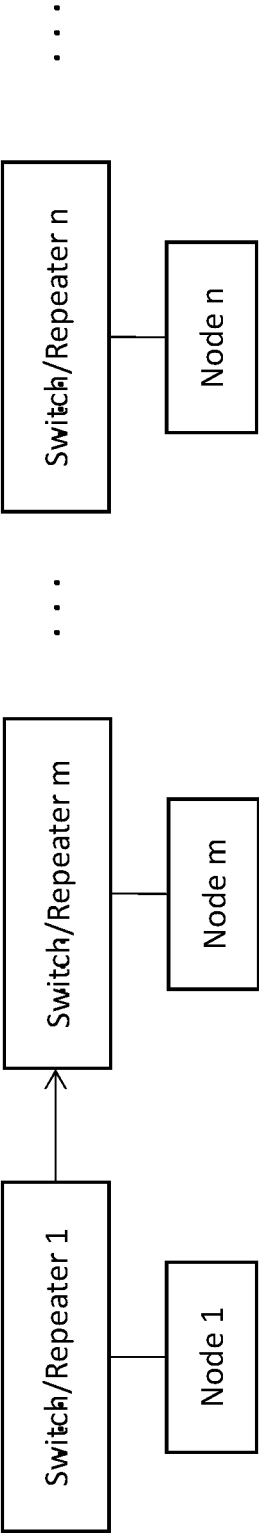


FIG. 3

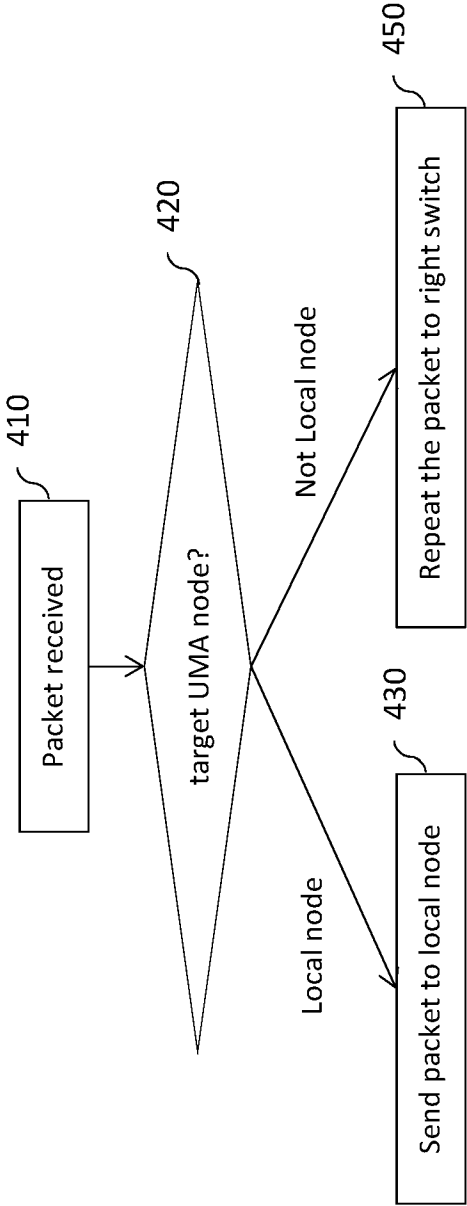


FIG. 4

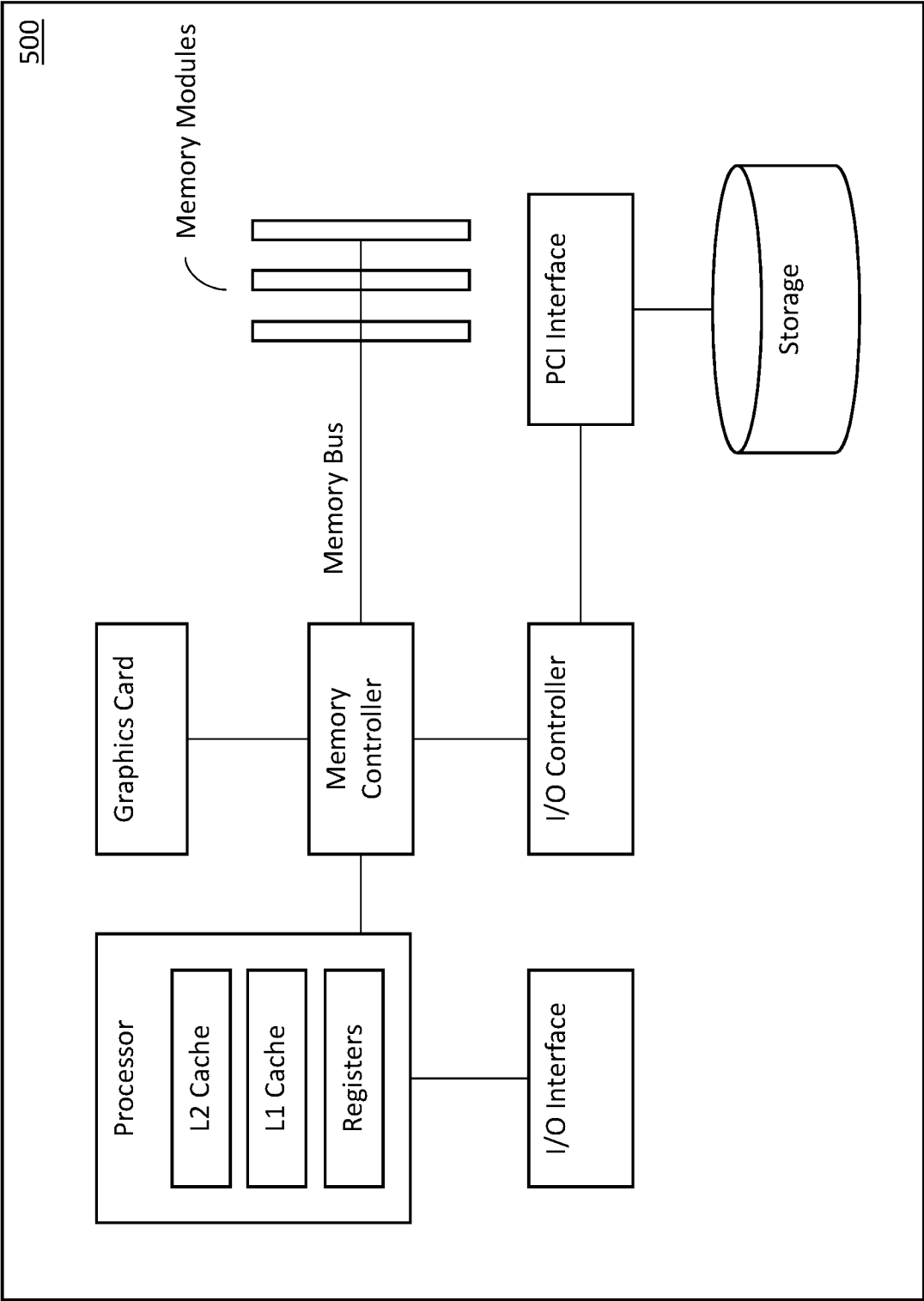


FIG. 5

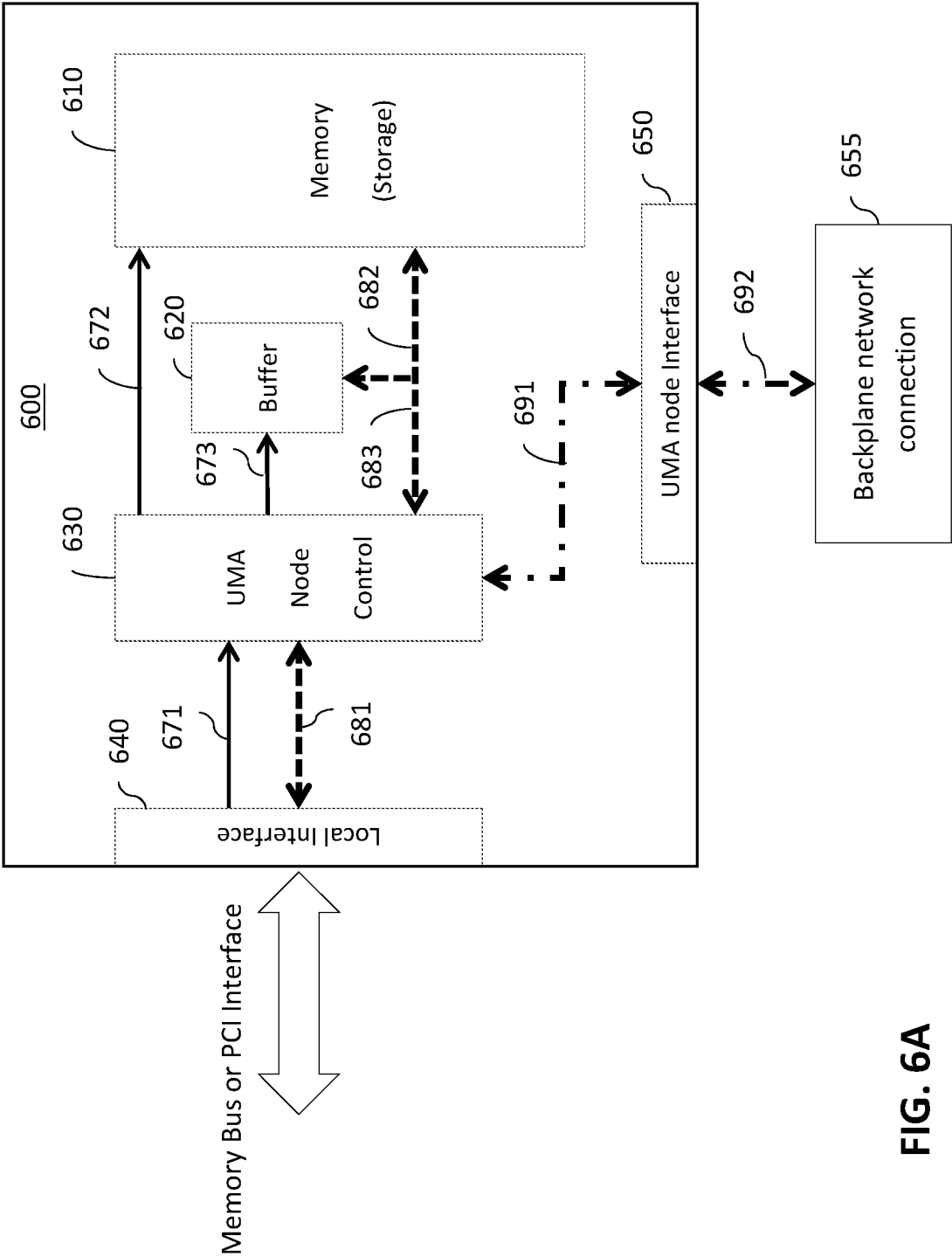


FIG. 6A

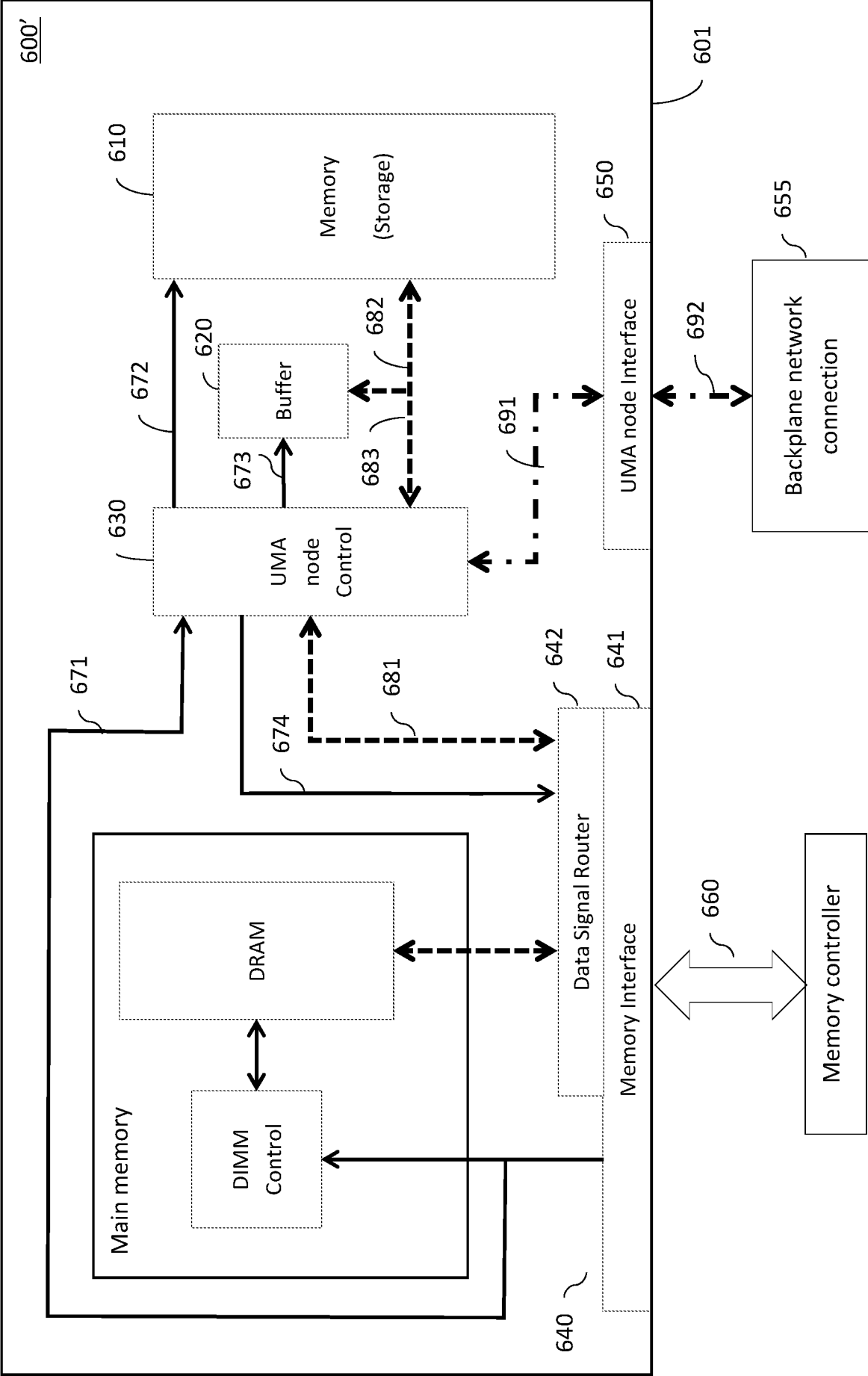


FIG. 6B

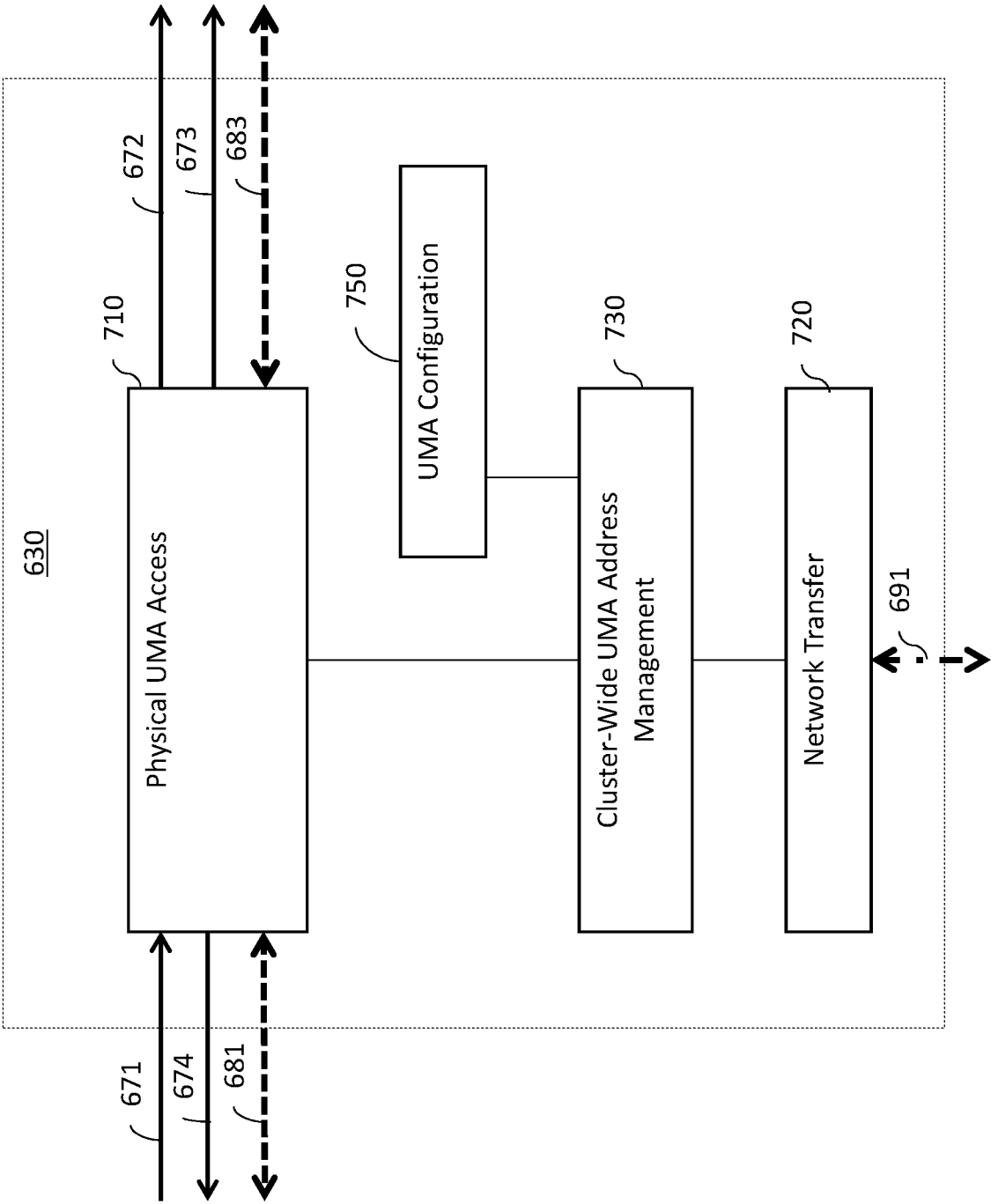


FIG. 7

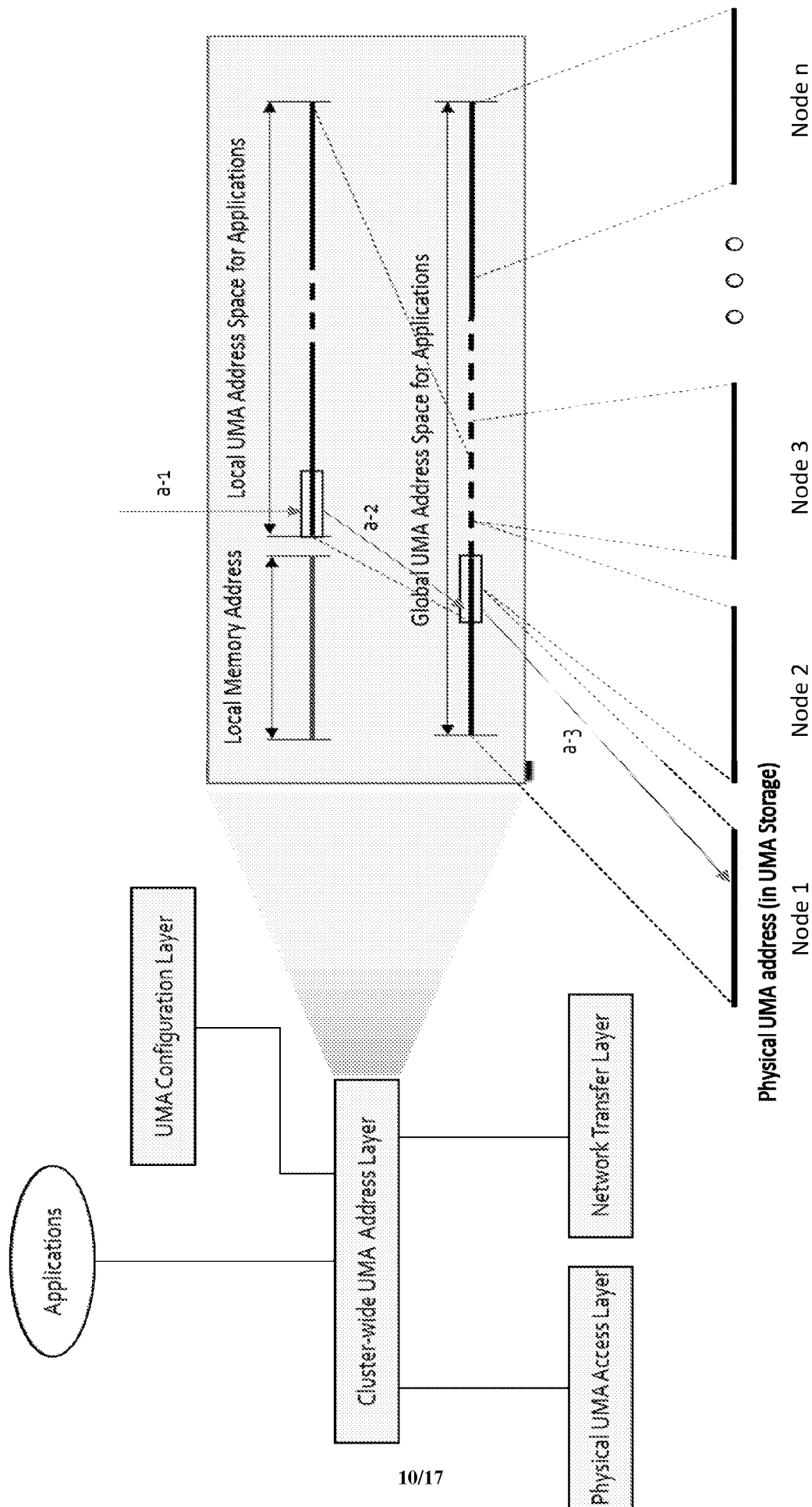
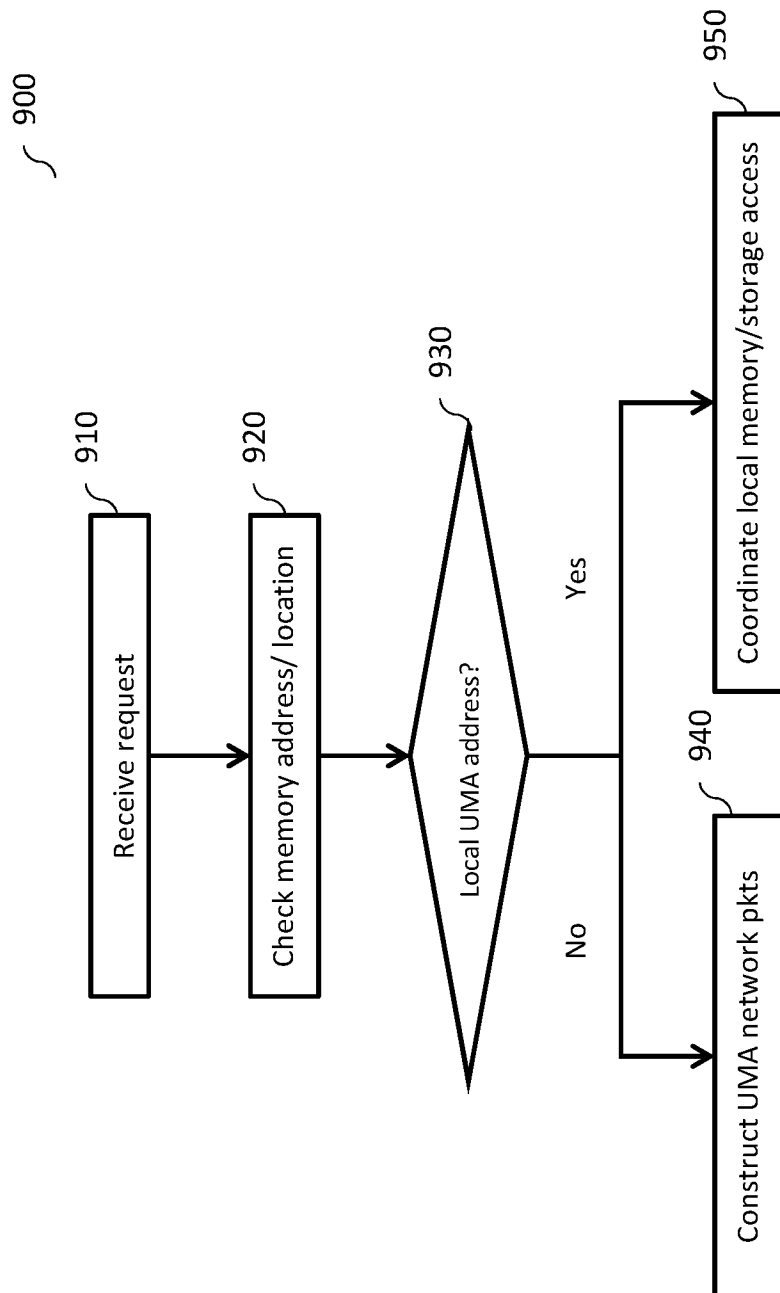


FIG. 8

**FIG. 9**

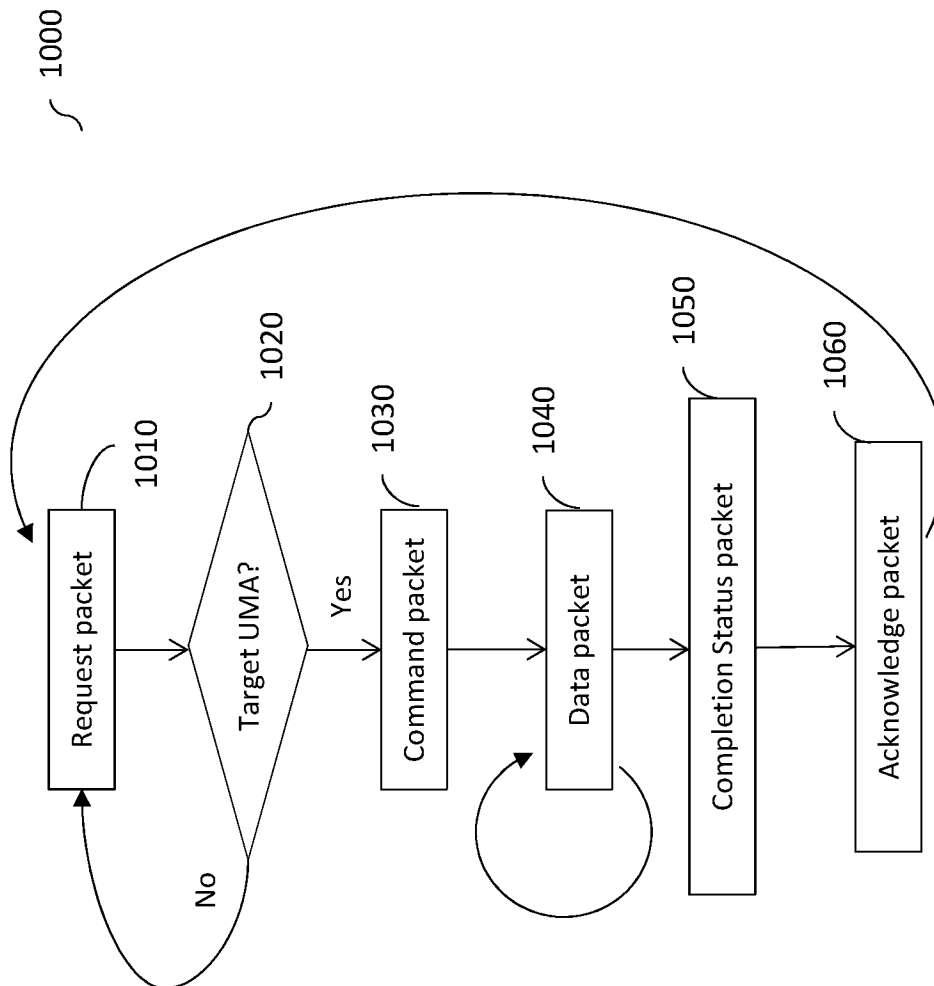


FIG. 10

Pkt type	UMA node address	Requester address	Access Type	Access priority	Total Byte size to transfer	CRC
----------	------------------	-------------------	-------------	-----------------	-----------------------------	-----

1110

FIG. 11A

Request packet

Pkt type	UMA node address	Requester address	UMA memory address	CRC on/off	Dynamic real time commands	Data size	CRC
----------	------------------	-------------------	--------------------	------------	----------------------------	-----------	-----

1120

FIG. 11B

Command packet

Pkt type	UMA node address	Requester address	Seq No.	Data	Data CRC	CRC
----------	------------------	-------------------	---------	------	----------	-----

1130

FIG. 11C

Data packet

Pkt type	UMA node address	Requester address	Total Number of bytes transferred	No. Data pkts transferred	Reserved	CRC
----------	------------------	-------------------	-----------------------------------	---------------------------	----------	-----

1140

FIG. 11D

Completion Status packet

Pkt type	UMA node address	Requester address	Total Number of bytes transferred	No. Data pkts transferred	Error Data pkts Seq No.	CRC
----------	------------------	-------------------	-----------------------------------	---------------------------	-------------------------	-----

1150

FIG. 11E

Acknowledge packet

A. UMA Storage in each computing node

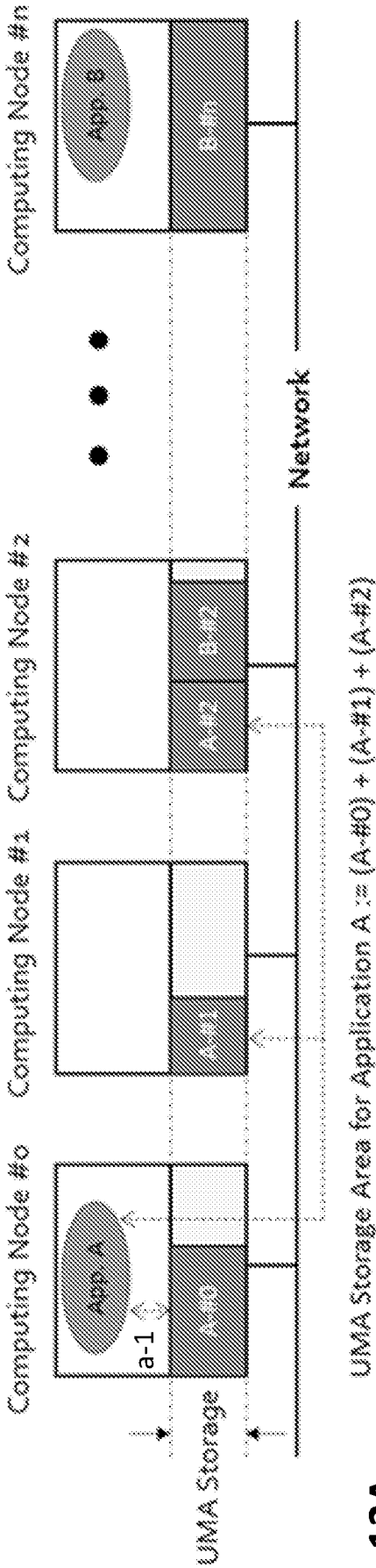


FIG. 12A

B. UMA Storage separate from computing node

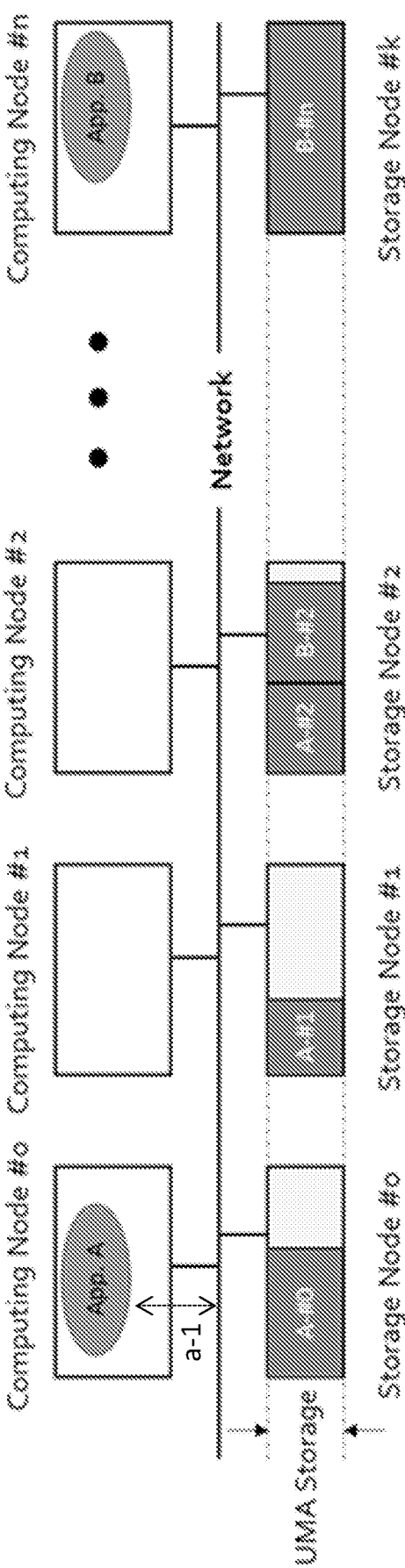


FIG. 12B

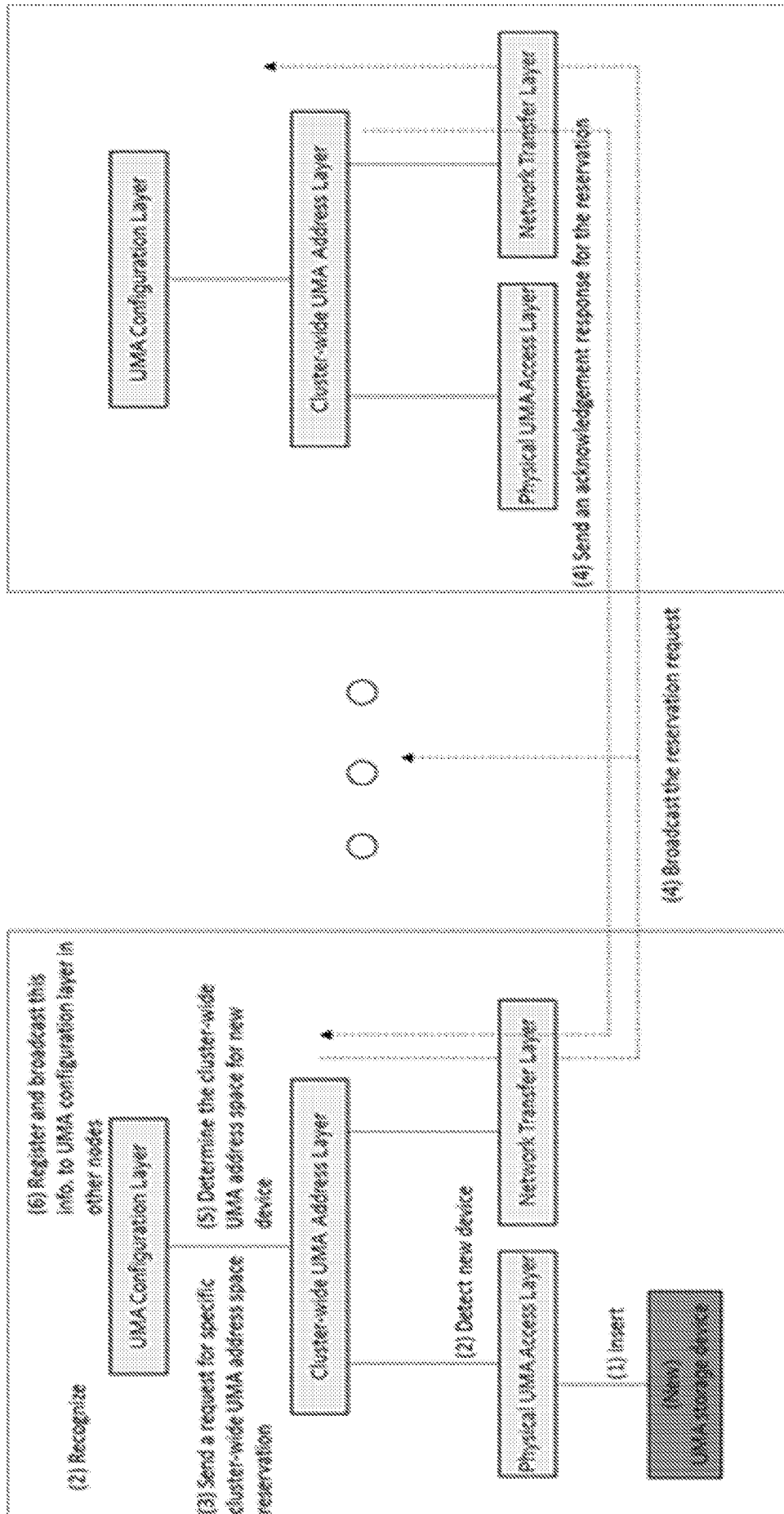


FIG. 13

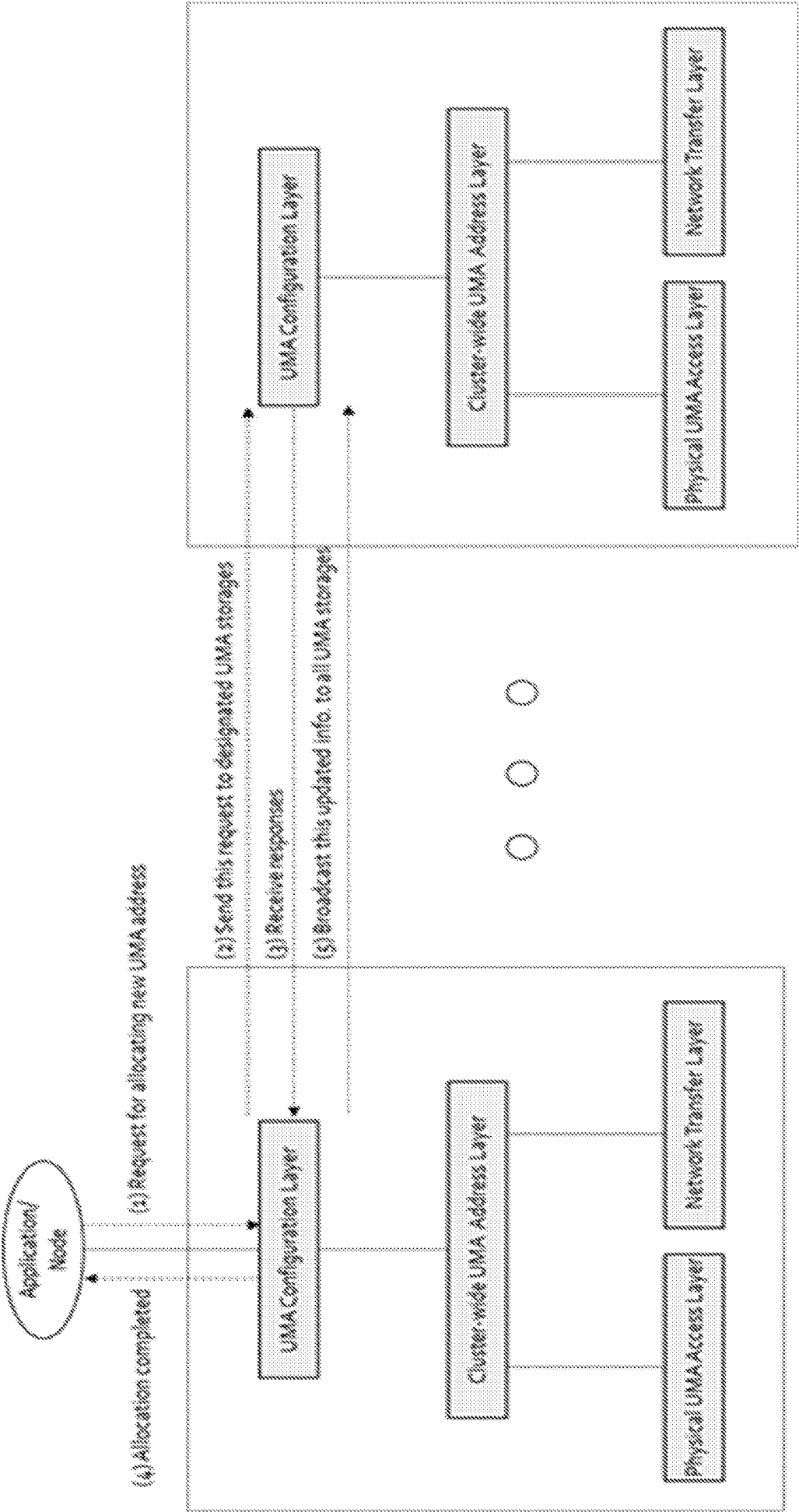


FIG. 14

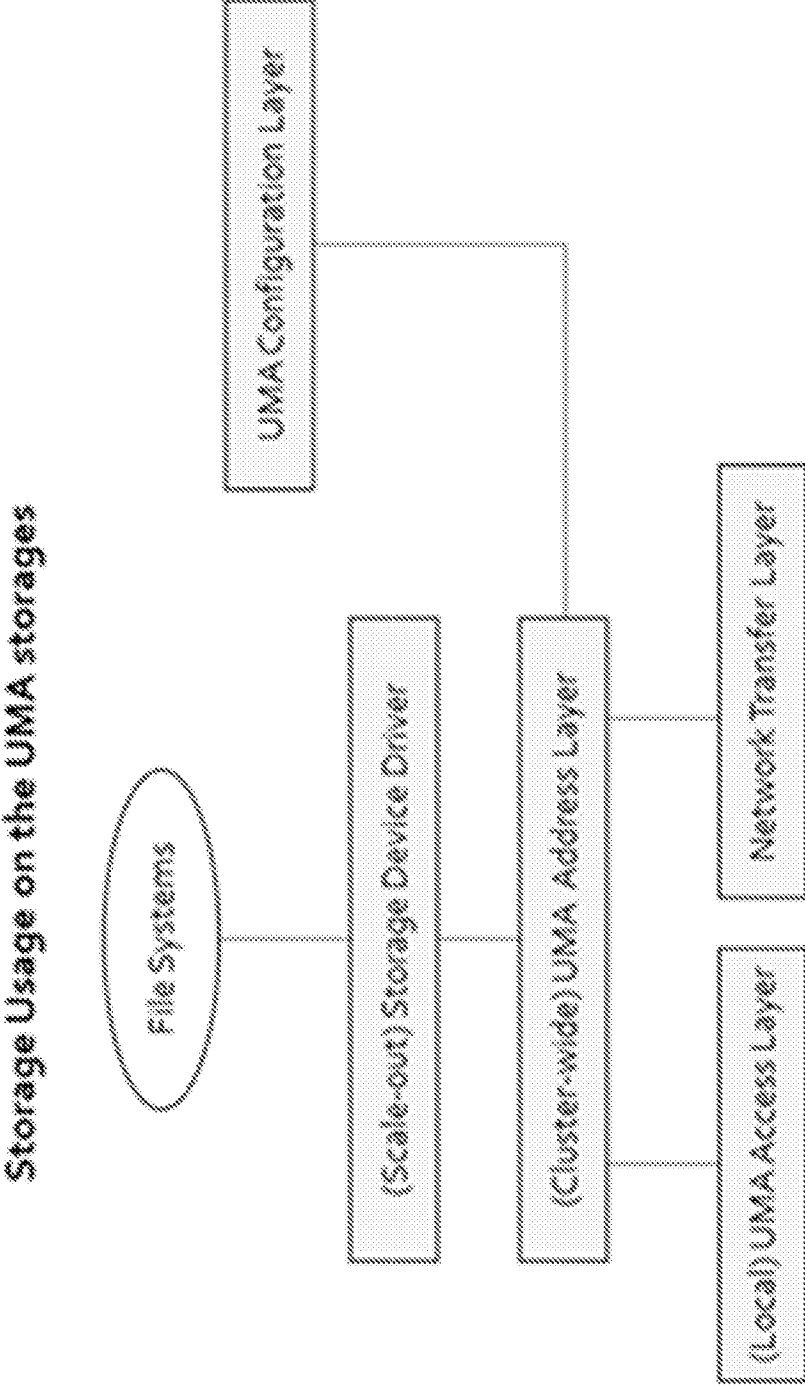


FIG. 15

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US2017/029478

A. CLASSIFICATION OF SUBJECT MATTER

IPC(8) - G06F 9/455; G06F 9/50; G06F 12/06; G06F 13/00; H04L 29/08 (2017.01)

CPC - G06F 9/45533; G06F 9/50; G06F 9/5016; G06F 9/5027; G06F 9/5072; G06F 9/5077; G06F 9/5083; G06F 9/5088; G06F 12/0284; G06F 12/08; G06F 12/0811; G06F 13/00; G06F 15/173; G06F 2209/504; G06F 2209/542; G06F 2212/251; G06F 2212/2515; G06F 2212/253; G06F 2212/302; G06F 2213/0038; H04L 67/1008; H04L 67/1097 (2017.02)

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)

See Search History document

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

USPC - 345/537; 345/572; 709/205; 709/214; 711/1 (keyword delimited)

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

See Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2014/0207871 A1 (MILOUSHEV et al) 24 July 2014 (24.07.2014), entire document	1-19
Y	EP 1 988 473 B1 (HOFMEISTER) 08 February 2012 (08.02.2012), entire document	1-19
Y	US 2012/0075319 A1 (DALLY) 29 March 2012 (29.03.2012), entire document	1-19
Y	US 2014/0247832 A1 (MELLANOX TECHNOLOGIES LTD.) 04 September 2014 (04.09.2014), entire document	5-13
A	US 2004/0260905 A1 (CYPHER et al) 23 December 2004 (23.12.2004), entire document	1-19
A	US 6,275,900 B1 (LIBERTY) 14 August 2001 (14.08.2001), entire document	1-19
A	US 2011/0078406 A1 (NICKOLLS et al) 31 March 2011 (31.03.2011), entire document	1-19
A	US 2005/0044301 A1 (VASILEVSKY et al) 24 February 2005 (25.02.2005), entire document	1-19

☐ Further documents are listed in the continuation of Box C.☐ See patent family 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 application or patent 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

11 June 2017

Date of mailing of the international search report

30 JUN 2017

Name and mailing address of the ISA/US

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
P.O. Box 1450, Alexandria, VA 22313-1450
Facsimile No. 571-273-8300

Authorized officer

Blaine R. Copenheaver

PCT Helpdesk: 571-272-4300
PCT OSP: 571-272-7774