March 29, 2010
RT0899
Communications; Computer Science 22 pages

Research Report

OPTICAL MEMORY RING AND ITS COMMAND ISSUE SCHEDULING METHOD

Atsuya Okazaki, Yasunao Katayama

IBM Research - Tokyo
IBM Japan, Ltd.
1623-14 Shimotsuruma, Yamato
Kanagawa 242-8502, Japan

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

APPLICATION FOR PATENT

ON

OPTICAL MEMORY RING AND ITS COMMAND ISSUE SCHEDULING METHOD

BY

ATSUYA OKAZAKI
2-2-23-303, TSUKIMINO,
YAMATO-SHI, KANAGAWA-KEN,
JAPAN
CITIZEN OF JAPAN

YASUNAO KATAYAMA
6-17-8, MINAMINO,
HACHIOUJI-SHI, TOKYO, JAPAN
CITIZEN OF JAPAN

OPTICAL MEMORY RING AND ITS COMMAND ISSUE SCHEDULING METHOD

TECHNICAL FIELD

[0001] The present disclosure generally relates to the field of computer technology, and more particularly to a memory apparatus connected in a ring configuration utilizing optical fibers.

BACKGROUND

[0002] There are increasing demands for higher memory bandwidths and capacities. Optically-attached memory systems may be utilized as an approach to satisfy such demands. An optically-attached memory system may comprise a plurality of memory nodes connected via optical fibers. Current optical technologies may allow processors (e.g., CPUs) and memory nodes to be placed meters away from each other without sacrificing bandwidth. Therefore, optically-attached memory systems may be utilized to increase memory bandwidths and/or capacities.

SUMMARY

[0003] The present disclosure is directed to a memory apparatus connected in a ring configuration utilizing optical fibers. The memory apparatus may comprise a plurality of memory nodes, each memory node comprising at least one storage device and an optical memory buffer for accessing the at least one storage device, the optical memory buffer of each memory node further comprising: an optical-to-electrical converter for converting an optical input to an electrical input; a transimpedance amplifier for

converting the electrical input to an input signal; a logic module for processing the input signal and generating an output signal based on the input signal; a laser diode driver coupled with a laser diode for producing an optical output based on the output signal; and a memory controller coupled with the plurality of memory nodes in a ring configuration utilizing optical fibers; wherein when a particular memory node of the plurality of memory nodes responds to a data request issued by the memory controller for the particular memory node, the output signal of the logic module of the particular memory node is generated by merging a response to the data request with the input signal to the logic module of the particular memory node.

[0004] It is to be understood that both the foregoing general description and the following detailed description are exemplary and explanatory only and are not necessarily restrictive of the present disclosure. The accompanying drawings, which are incorporated in and constitute a part of the specification, illustrate subject matter of the disclosure. Together, the descriptions and the drawings serve to explain the principles of the disclosure.

BRIEF DESCRIPTION OF THE DRAWINGS

[0005] The numerous advantages of the disclosure may be better understood by those skilled in the art by reference to the accompanying figures in which:

FIG. 1 is a block diagram of an optically-attached memory system utilizing a tree topology with optical modulation amplitude and average power of different fiber sections;

FIG. 2 is a block diagram of an optically-attached memory system utilizing a tree topology, wherein different memory nodes have different distances to a processor node;

FIG. 3 is a diagram illustrating loss of data in an optical-electrical or electrical-optical module;

FIG. 4 is a block diagram illustrating an optically-attached memory system utilizing a ring configuration;

FIG. 5 is a block diagram of a memory node;

FIG. 6 is a time table illustrating scheduling in a ring configuration;

FIG. 7 is a comparison of different optically-attached memory system topologies.

DETAILED DESCRIPTION

[0006] Reference will now be made in detail to the subject matter disclosed, which is illustrated in the accompanying drawings.

[0007] The present disclosure is directed to a memory apparatus connected in a ring configuration utilizing optical fibers. The memory apparatus may comprise a memory controller and a plurality of memory nodes, wherein the memory controller is coupled with the plurality of memory nodes in a ring configuration. The memory controller may receive read/write data commands from one or more processors. The memory controller may then issue corresponding data requests via the ring configuration. The data requests may circulate via the ring configuration, and particular memory nodes needing to communicate with the memory controller may respond via the ring configuration. The responses received by the memory controller may be returned to the corresponding processors.

[0008] Connecting the memory controller and the memory nodes in a ring configuration may improve certain disadvantages of optically-attached memory systems connected utilizing other connection topologies. For instance, FIG. 1 depicts an optically-attached memory system connected utilizing a tree topology (multi-drop link). In this configuration, the processor node may comprise a memory controller (e.g., an integrated controller) that connects to the memory nodes in a hierarchical manner. Optical couplers may be utilized in order to join the optical fibers to form the connection hierarchy. However, due to insertion losses of optical couplers, Optical Modulation Amplitude (OMA) and/or average power $P_{average}$ may drop at the endpoint receivers (memory nodes) in such optically-attached memory systems. As illustrated in FIG. 1, for example, fiber section A may carry an initial OAM and $P_{average}$ intensities. After sustaining insertion losses of optical couplers, the initial intensities may be noticeably reduced in fiber section B and further reduced in fiber section C.

[0009] Another disadvantage associated with a memory system connected utilizing a tree topology may be clock signal synchronization. For instance, as illustrating in FIG. 2, the processor node may require capabilities to immediately synchronize clock signals transferred from every memory node. Since the distances from each memory node to the processor node are different for each memory node (as illustrated using dotted lines in FIG. 2), the clock signals transferred by the memory nodes may have differences in clock cycles and clock phase when received by the processor node. This disadvantage may appear not only in source synchronous clock schemes, but also in embedded clock schemes.

[0010] Still another disadvantage associated with existing optically-attached memory systems may be signal losses of a long constant input signal (such as

long streams of 0,0,0, ..., 0, or 1,1,1, ..., 1). As illustrated in FIG. 3, optical modules, such as optical-to-electrical (O/E) modules and/or electrical-to-optical (E/O) modules, may comprise an AC-coupled electrical interface for turning off or migrating into a squelch mode when the modules receive long constant input signals. When a long constant input signal is received as an input to the optical modules including AC-coupled high-speed electrical differential pair lines, the voltage gap between the differential pair lines may become small. As a result, the receivers may not detect the existence of the gap, consequently, may not determine if the data represents a '1' or a '0'. Additionally, the optical modules in off mode or squelch mode may need boot up time before fully functional, such delay may not satisfy timing requirements in memory access latency of the processor node.

[0011] Referring generally now to FIGS. 4 and 5, there is depicted a memory apparatus 400 of the present disclosure. The memory apparatus may comprise a memory controller 402 and a plurality of memory nodes connected via optical fibers. The memory controller 402 is coupled with the plurality of memory nodes in a ring configuration. For instance, the memory controller 402 may be directly connected to a first memory node 404, which may be connected to subsequent memory nodes. A last memory node 406 may be directly connected back to the memory controller 402.

[0012] The memory controller 402 may be coupled with a processor node 408 comprising one or more processors. In one embodiment, the memory controller 402 is configured as an on-chip integrated component of the processor node 408. Alternatively, the memory controller 402 may be a separate module in electrical connection with the processors of the processor node 408.

[0013] In one embodiment, each memory node may include one or more memory storage devices 502 (e.g., dynamic random access memories (DRAMs)) and an optical memory buffer 504 for accessing the storage devices. The optical memory buffer 504 may comprise an optical-to-electrical converter 506 for converting an optical input to an electrical input. The optical input may be input received via the optical fiber from the memory controller (e.g., when the memory node is the first memory node 404) or from the previous memory node. The converted electrical input may then pass through a current-to-voltage converter 508 (e.g., a transimpedance amplifier) to be converted to an input signal to a logic module 510.

[0014] The logic module 510 processes the input signal received, and generates an output signal based on the input signal. In one embodiment, if the memory node that the logic module 510 is associated with does not have any data response to be transmitted to the memory controller 402, the logic module 510 simply passes the input signal through as the output signal. Otherwise, if the memory node that the logic module 510 is associated with contains data responses to be transmitted to the memory controller 402, the logic module 510 may obtain the data responses and merge the responses with the input signal. The merged signal may then be transmitted as the output signal.

[0015] The output signal generated by the logic module 510 is received by a laser diode driver 512 (e.g., a current driver/control circuit) coupled with an electrical-to-optical converter 514 to produce an optical output based on the output signal. The electrical-to-optical converter 514 may be a laser diode. In one embodiment, the electrical-to-optical converter 514 is a

vertical-cavity surface-emitting laser (VCSEL). The converted optical output may be transmitted via optical fibers of the ring configuration to a subsequent memory node or to the memory controller 402 (e.g., when the current memory node is the last memory node 406).

[0016] In one embodiment, a data request (a packet) may be issued/generated by the memory controller corresponding to each read/write data command of the processor node. Each packet may comprise a command, an address, and a data portion (e.g., data to be stored). Cyclic redundancy check (CRC) and/or error correcting code (ECC) may also be optionally added to the packet. The packet may be serialized and transmitted via the optical fiber. The packet may be converted and processed by a logic module 510 of a particular memory node. A de-serializer 516 may be utilized to de-serialize the converted packet, and a decoder 518 may be utilized to decode information contained in the de-serialized converted packet. It is understood that the decoder 518 may access the storage devices of the particular memory node during the decoding process.

[0017] When the data request received is not related to this particular memory node, no further action may be necessary and the data request may be transmitted to the next node on the ring configuration. Otherwise, the data request is processed through a data bus interface 520 based on the command (e.g., read or write command), the address and the data portion specified in the data request. For example, if a read request was received previously, the memory node may be activated and data may be read from the storage devices at the address specified in the request. The data read utilizing the data bus interface 520 may be serialized by passing through a serializer 522 and merged with the input signal. The merged signal

comprising the data read in response to the data request and the input signal forms the output signal, which is then converted to optical signals and transmitted to the subsequent memory nodes or memory controller for further processing.

[0018] In one embodiment, the memory controller may provide dedicated slots in the packet stream to facilitate data merging. Dedicated slots may be empty slots or slots filled with an idle pattern (as place holders). The data to be transmitted to the memory controller may be merged into the input signal at provided dedicated slots and be transmitted to the memory controller. For example, the memory controller may utilize a certain number of optical fibers to carry commands and utilize the rest of the fibers to carry data packets.

[0019] FIG. 6 are time tables illustrating exemplary scheduling information of a read request (shown in (a)), a write request (shown in (b)), and a combination of two read and one write requests (shown in (c)). In these examples, the memory controller may utilize one fiber to carry commands and two fibers to carry data packets.

[0020] For instance, FIG. 6(a) is an example where a processor node reads data stored in a storage device of memory node 1. Upon receiving command from the process node, the memory controller issues an activation command at time (clock cycle) 1. The optical memory buffers of the memory nodes are always observing the packets on the fibers and pass them to the next memory node. The optical memory buffer of memory node 1 reads the activation command sent, and activates the connected DRAM device at time 2 if the address information in the activation command matches to the memory address range of the memory node. Subsequently,

at time 4, the memory controller issues a read command for memory node 1 and specifies the data address. The logic module of memory node 1 receives the command at time 5 and starts processing the read command. In this example, assuming that the storage device take three clock cycles (time slots) to respond to a data request. Thus, at time 8, a first 4 bits of the data requested is retrieved from the storage devices, and at time 9, the remaining 4 bits of the data request is retrieved from the storage devices. At this point, the data is ready to be transmitted to the memory controller by merging the data read to the dedicated slots on the optical fiber. It is understood that the data on the optical fiber may pass through subsequent memory nodes, which may merge additional data to the optical fiber to be transmitted to the memory controller.

[0021] A write request may be handled similarly as illustrated in FIG. 6(b), where the data bus interface may start writing data to memory node 1 following receipt of the data via optical fiber. Upon receiving command from the process node, the memory controller issues an activation command at time 1. The optical memory buffer of memory node 1 reads the command and issue activation command to its storage devices at time 2. From time 3, as an example, the processor may start transferring data "WD" to be written to the storage devices of memory node 1. A write command may be issued at time 4, and at time 5, memory node 1 may receive the write command and starts writing the received data to its storages device subsequently.

[0022] In another example as illustrated in FIG. 6(c), two read requests are issued to memory node 1 and a write request is issued to memory node 2. As shown in the schedule table, commands may arrive at each node sequentially via the ring configuration. Once the commands have been

received, data operations of memory node 1 may be carried out independently from data operations of memory node 2. Data in response to read command 1 and read command 3 may to be transmitted to the memory controller by merging the data to the dedicated slots on the optical fiber.

[0023] It is understood that clock cycles (time slots) may differ from the above examples based on the time instances the commands are issued as well as response times of the storage devices of each memory nodes (e.g., some storage devices may take longer to respond than others). It is also understood that data throughputs/widths illustrated in FIG. 6 are merely exemplary. Optical fibers and/or data bus interfaces of various data throughputs/widths may be utilized without departing from the scope and spirit of the present invention.

[0024] The memory apparatus of the present disclosure utilizing optical memory buffers connect in a ring configuration may maintain OMA and $P_{average}$ intensities as oppose to a tree topology. The optical-electrical and electrical-optical converting devices are able to maintain the OMA and $P_{average}$ of an incoming signal, and transfer the signal to the next memory node. Since the pass-through logic may transfer the incoming data substantially at close to wire speed to the next memory node, the latency overhead by optical-electrical and electrical-optical converting may be insignificant.

[0025] In addition, as illustrated in FIG. 7, the number of the fibers and the optical modules may be less utilizing a ring topology than a bi-directional topology such as a fully buffered dual in-line memory module (FB-DIMM). Therefore, the cost for the fibers and the optical modules may be reduced utilizing the ring configuration. Furthermore, when scheduling in a FB-DIMM,

the memory controller may need to wait for turnaround time to the last memory node. Therefore, the latency may increase as the number of the memory nodes on a channel increases. However, utilizing the ring configuration of the present disclosure, the memory controller may issue commands at anytime without additional latency overhead. There is no restriction on how many memory nodes may be connected on a ring.

[0026] The disadvantages associated with clock signal synchronization in a memory system connected utilizing a tree topology may also be addressed by the present disclosure. While the processor node needs to synchronize clock signals from every memory node connected in a tree topology, in a ring configuration of the present disclosure, the processor node may only need to synchronize with an output signal from the last memory node on the ring.

[0016] It is contemplated that the memory controller of the present disclosure may issue storage device (DRAM) refresh commands for not only a specified memory node but also all memory nodes at a time (i.e., broadcasting the refresh command). In response to the refresh command, a memory node may refresh the DRAM devices of itself, and at the same time, transfer (pass through) the refresh command to the next memory node. Utilizing this broadcast scheme, issuing a single refresh command may refresh the DRAM devices in all memory nodes, as oppose to issuing multiple refresh commands for each memory node. Consequently, this broadcast scheme may also improve bandwidth efficiency of the connection.

[0017] The broadcast scheme may be appreciate in reducing signal losses of a long constant input signal. Storage device refresh command may be utilized to prevent the AC-coupled optical modules from turning off or

migrating into squelch mode by the long constant input. One requirement for the periodic refresh time t_{REFI} of the DRAM devices is that t_{REFI} needs to be smaller than the transitional time of the AC coupling capacitor (in the order of tens of microseconds). This requirement may be satisfied utilizing the memory apparatus of the present disclosure through the broadcast scheme, which signals all storage devices in all memory nodes to refresh substantially simultaneously (time spent for issuing commands is effectively minimized), as long as the time needed for refreshing of DRAM devices themselves satisfies the requirement (e.g., t_{REFI} = 7.8 μsec in Micron DDR2 SDRAM, which meets the requirements).

[0018] In the present disclosure, the methods disclosed may be implemented as sets of instructions or software readable by a device. Further, it is understood that the specific order or hierarchy of steps in the methods disclosed are examples of exemplary approaches. Based upon design preferences, it is understood that the specific order or hierarchy of steps in the method can be rearranged while remaining within the disclosed subject matter. The accompanying method claims present elements of the various steps in a sample order, and are not necessarily meant to be limited to the specific order or hierarchy presented.

[0019] It is believed that the present disclosure and many of its attendant advantages will be understood by the foregoing description, and it will be apparent that various changes may be made in the form, construction and arrangement of the components without departing from the disclosed subject matter or without sacrificing all of its material advantages. The form described is merely explanatory, and it is the intention of the following claims to encompass and include such changes.

CLAIMS

What is claimed is:

1. An apparatus, comprising:

a plurality of memory nodes, each memory node comprising at least one storage device and an optical memory buffer for accessing the at least one storage device, the optical memory buffer of each memory node further comprising:

an optical-to-electrical converter for converting an optical input to an electrical input;

a transimpedance amplifier for converting the electrical input to an input signal;

a logic module for processing the input signal and generating an output signal based on the input signal;

a laser diode driver coupled with a laser diode for producing an optical output based on the output signal; and

a memory controller coupled with the plurality of memory nodes in a ring configuration utilizing optical fibers;

wherein when a particular memory node of the plurality of memory nodes responds to a data request issued by the memory controller for the particular memory node, the output signal of the logic module of the particular memory node is generated by merging a response to the data request with the input signal to the logic module of the particular memory node.

OPTICAL MEMORY RING AND ITS COMMAND ISSUE SCHEDULING METHOD

ABSTRACT

The present disclosure is directed to a memory apparatus connected in a ring configuration utilizing optical fibers. The memory apparatus may comprise a plurality of memory nodes, each memory node comprising at least one storage device and an optical memory buffer for accessing the at least one storage device; a memory controller coupled with the plurality of memory nodes in a ring configuration utilizing optical fibers, wherein when a particular memory node of the plurality of memory nodes responds to a data request issued by the memory controller for the particular memory node, the optical memory buffer of the particular memory node merges a response to the data request with an input signal to the optical memory buffer.

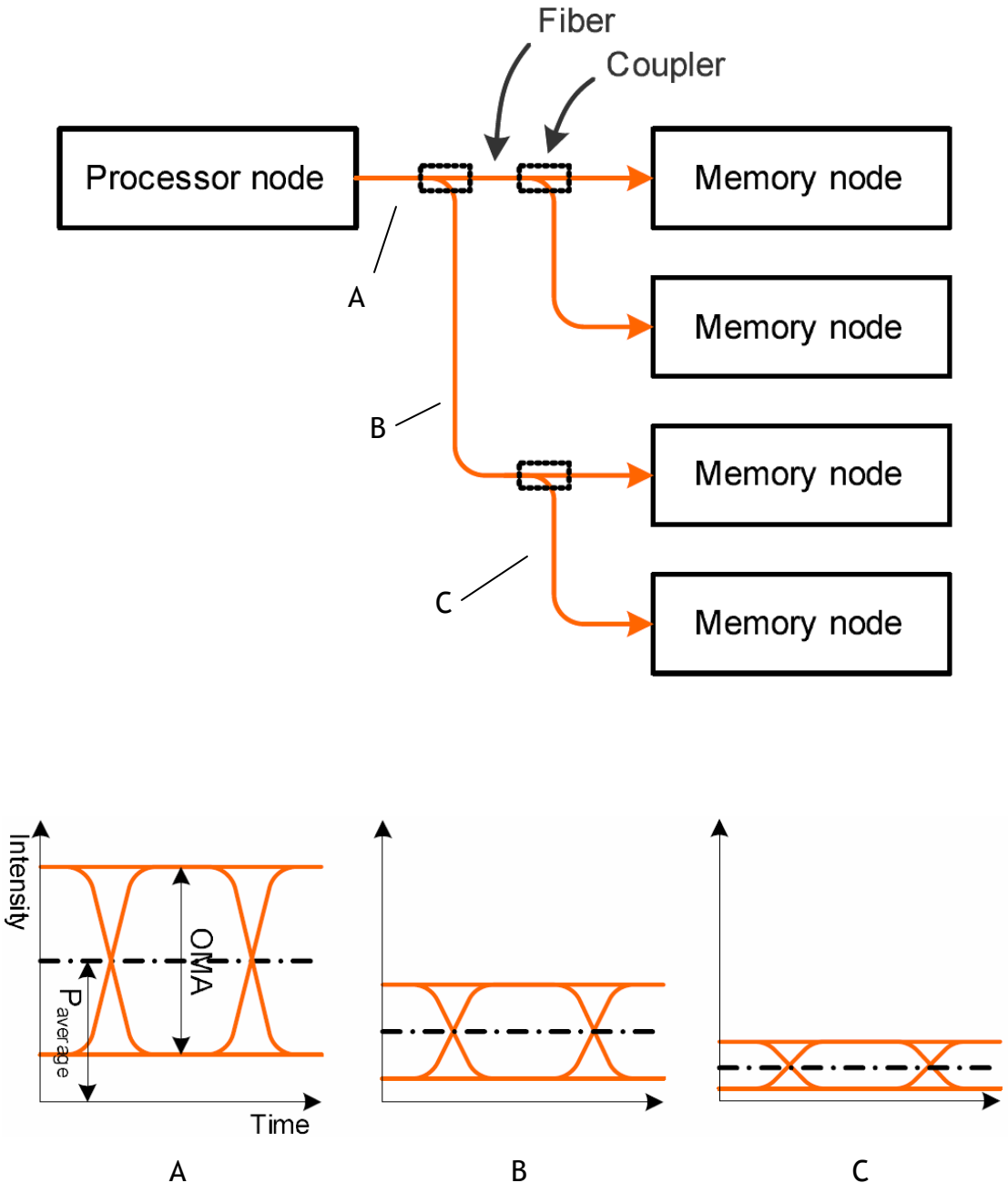

Fiber
Coupler
Processor node
Memory node
Memory node
Memory node
Memory node
A
B
C
Intensity
OMA
P average
Time
A
B
C

FIG. 1

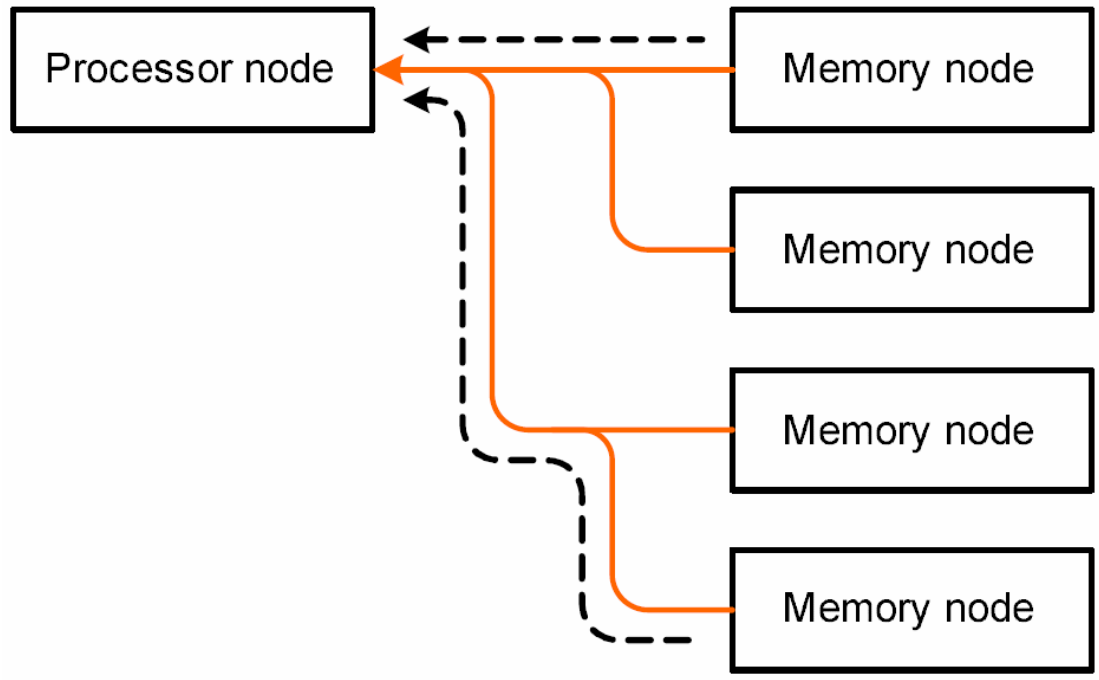

Processor node
Memory node
Memory node
Memory node
Memory node

FIG. 2

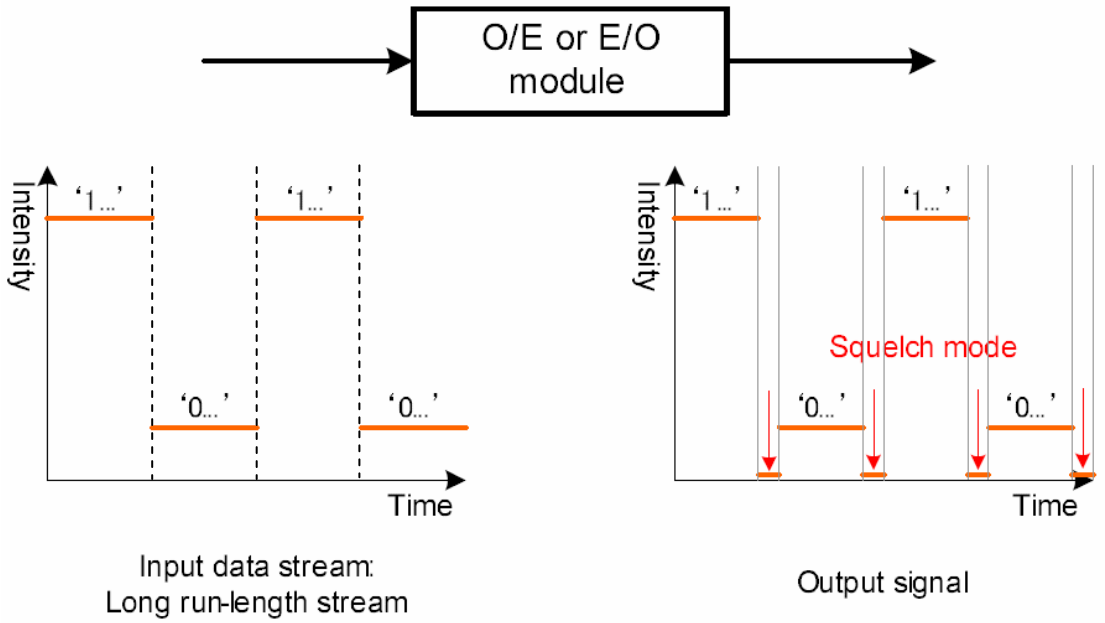

O/E or E/O
module
Intensity
'1...'
'1...'
'0...'
'0...'
Time
Input data stream:
Long run-length stream
Intensity
'1...'
'1...'
Squelch mode
'0...'
'0...'
Time
Output signal

FIG. 3

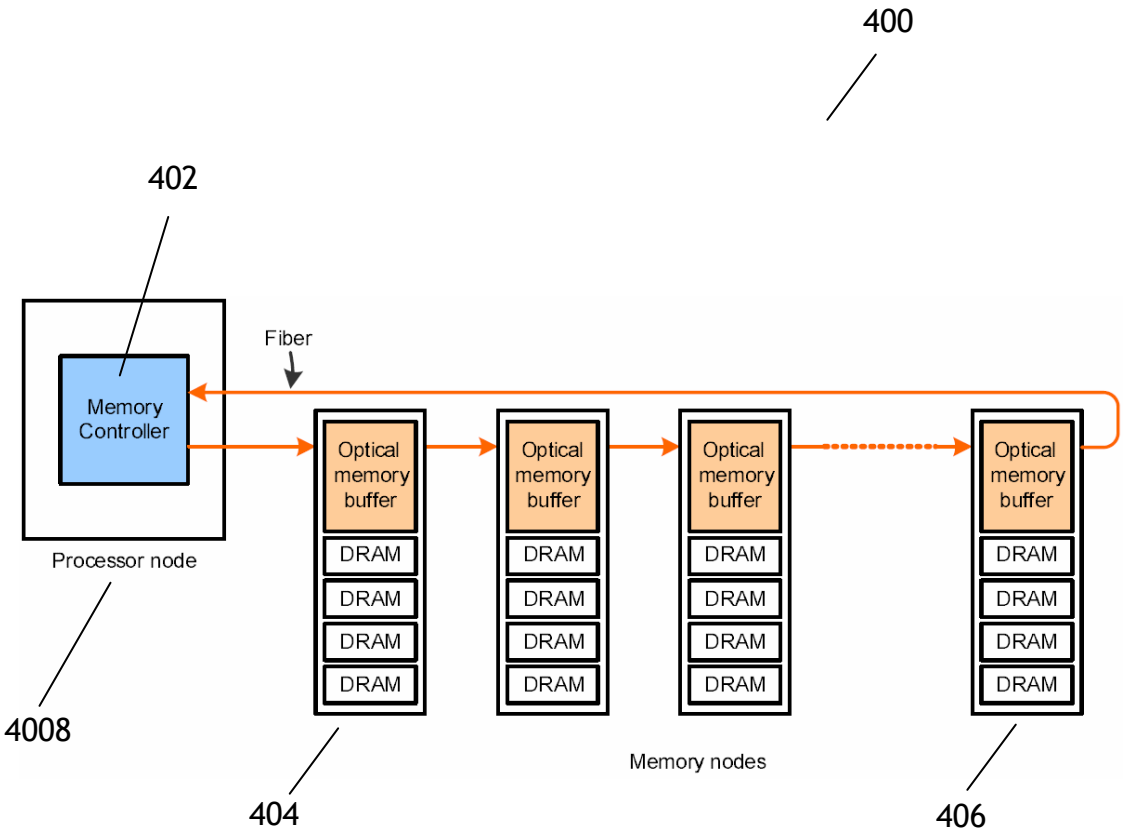

400
402
Fiber
Memory Controller
Processor node
4008
Optical memory buffer
DRAM
DRAM
DRAM
DRAM
Optical memory buffer
DRAM
DRAM
DRAM
DRAM
Optical memory buffer
DRAM
DRAM
DRAM
DRAM
Optical memory buffer
DRAM
DRAM
DRAM
DRAM
Memory nodes
404
406

FIG. 4

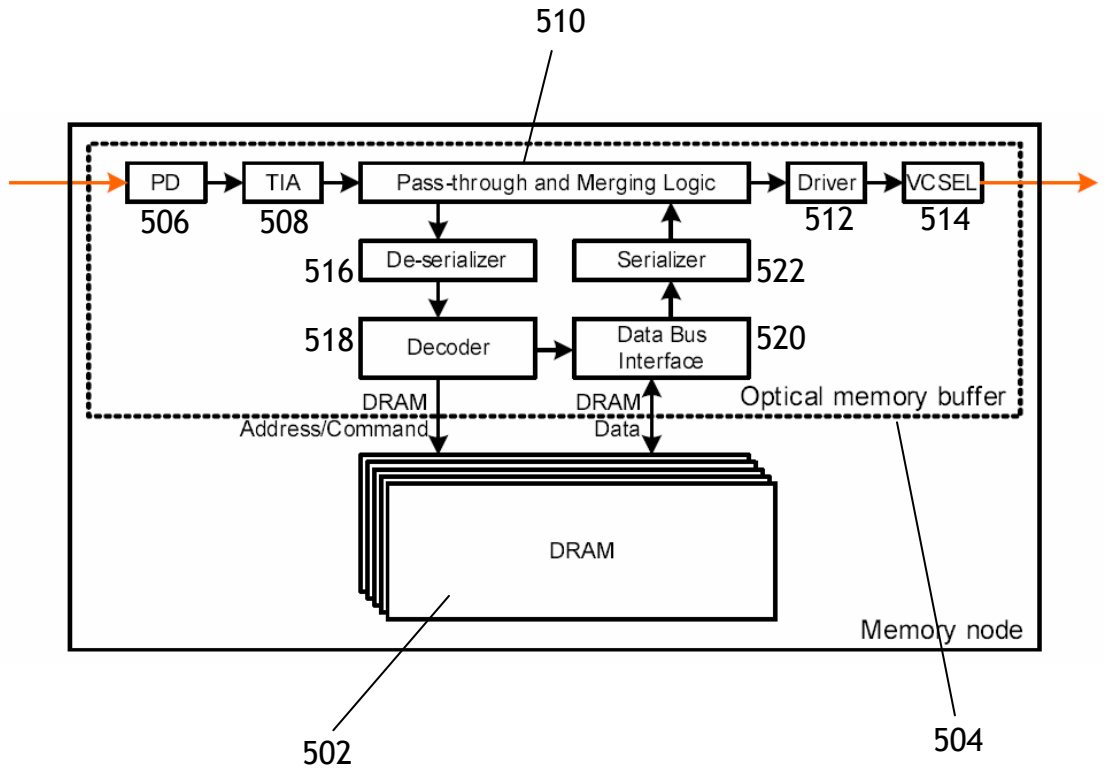
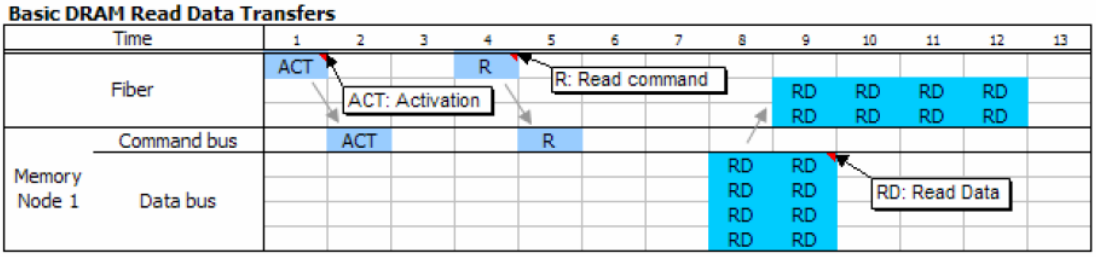
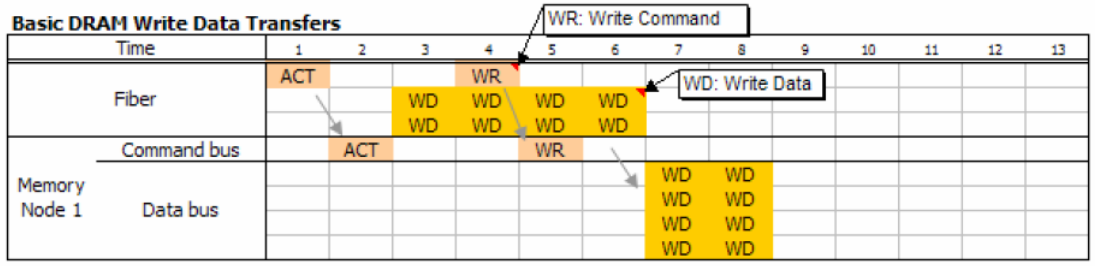

510
PD
506
TIA
508
Pass-through and Merging Logic
Driver
512
VCSEL
514
516
De-serializer
Serializer
522
518
Decoder
Data Bus Interface
520
DRAM Address/Command
DRAM Data
Optical memory buffer
DRAM
Memory node
502
504

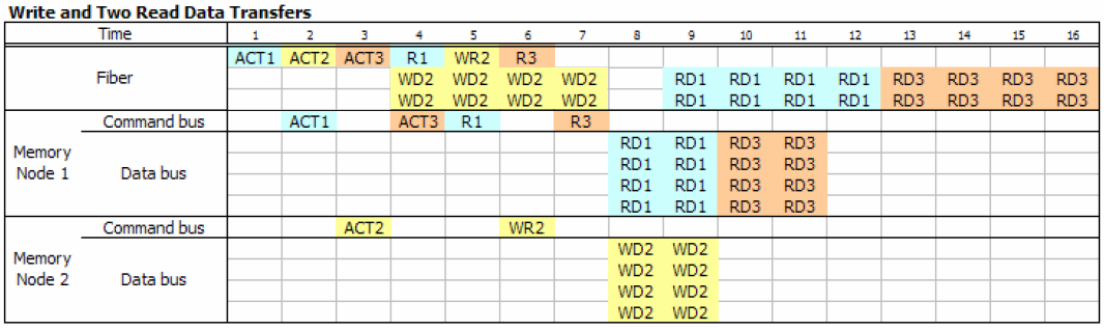
FIG. 5


Basic DRAM Read Data Transfers
Time
1 2 3 4 5 6 7 8 9 10 11 12 13
Fiber
ACT
R
R: Read command
ACT: Activation
RD RD RD RD
RD RD RD RD
Command bus
ACT
R
Memory Node 1
Data bus
RD RD
RD RD
RD RD
RD RD
RD: Read Data

(a)

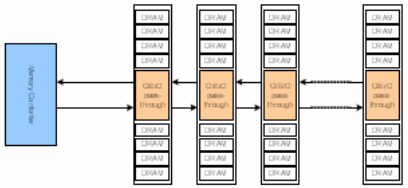
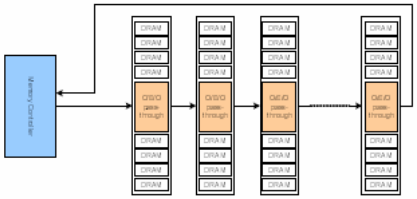
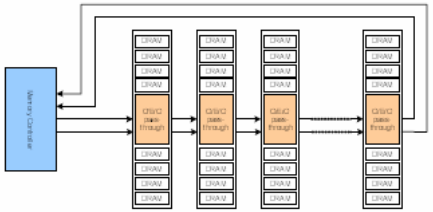

Basic DRAM Write Data Transfers
WR: Write Command
Time
1 2 3 4 5 6 7 8 9 10 11 12 13
Fiber
ACT
WR
WD: Write Data
WD WD WD WD
WD WD WD WD
Command bus
ACT
WR
Memory Node 1
Data bus
WD WD
WD WD
WD WD
WD WD

(b)


Write and Two Read Data Transfers
Time
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Fiber
ACT1 ACT2 ACT3 R1 WR2 R3
WD2 WD2 WD2 WD2 RD1 RD1 RD1 RD1 RD3 RD3 RD3 RD3
WD2 WD2 WD2 WD2 RD1 RD1 RD1 RD1 RD3 RD3 RD3 RD3
Command bus
ACT1 ACT3 R1 R3
Memory Node 1
Data bus
RD1 RD1 RD3 RD3
RD1 RD1 RD3 RD3
RD1 RD1 RD3 RD3
RD1 RD1 RD3 RD3
Command bus
ACT2 WR2
Memory Node 2
Data bus
WD2 WD2
WD2 WD2
WD2 WD2
WD2 WD2

(c)

FIG. 6

Topology	 Bi directional (FB-DIMM)	 Ring	 Ring
# of fiber	2n	n + 1	2(n + 1)
# of OE/EO	2n	n	2n

n: # of memory node

FIG. 7