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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
01 July 2021 (01.07.2021)

WIPO | PCT

(10) International Publication Number
WO 2021/133687 A1

(51) International Patent Classification:
G11C 5/06 (2006.01) *G06F 3/06* (2006.01)
G11C 7/10 (2006.01) *G11C 7/22* (2006.01)

(21) International Application Number:
PCT/US2020/066140

(22) International Filing Date:
18 December 2020 (18.12.2020)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/953,821 26 December 2019 (26.12.2019) US

(71) Applicant: **MICRON TECHNOLOGY, INC.** [US/US]; 8000 So. Federal Way, Boise, Idaho 83716-9632 (US).

(72) Inventor: **PAWLOWSKI, Joseph T.**; 12171 West Musket Drive, Boise, Idaho 83704 (US).

(74) Agent: **PERDOK, Monique M.** et al.; Schwegman Lundberg & Woessner, P.A., P.O. Box 2938, Minneapolis, Minnesota 55402 (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, IT, JO, 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, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, 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))*

(54) Title: TECHNIQUES FOR NON-DETERMINISTIC OPERATION OF A STACKED MEMORY SYSTEM

(57) Abstract: Techniques for non-deterministic operation of a stacked memory system are provided. In an example, a method of operating a memory package can include receiving a plurality of memory access requests for a channel at a logic die, returning first data to a host in response to a first memory access request of the plurality of memory access requests, returning an indication of data not ready to the host in response to a second memory access request of the plurality of memory access requests for second data, returning a first index to the host with the indication of data not ready, returning an indication data is ready with third data in response to a third memory access request of the plurality of memory access requests, and returning the first index with the indication of data ready.

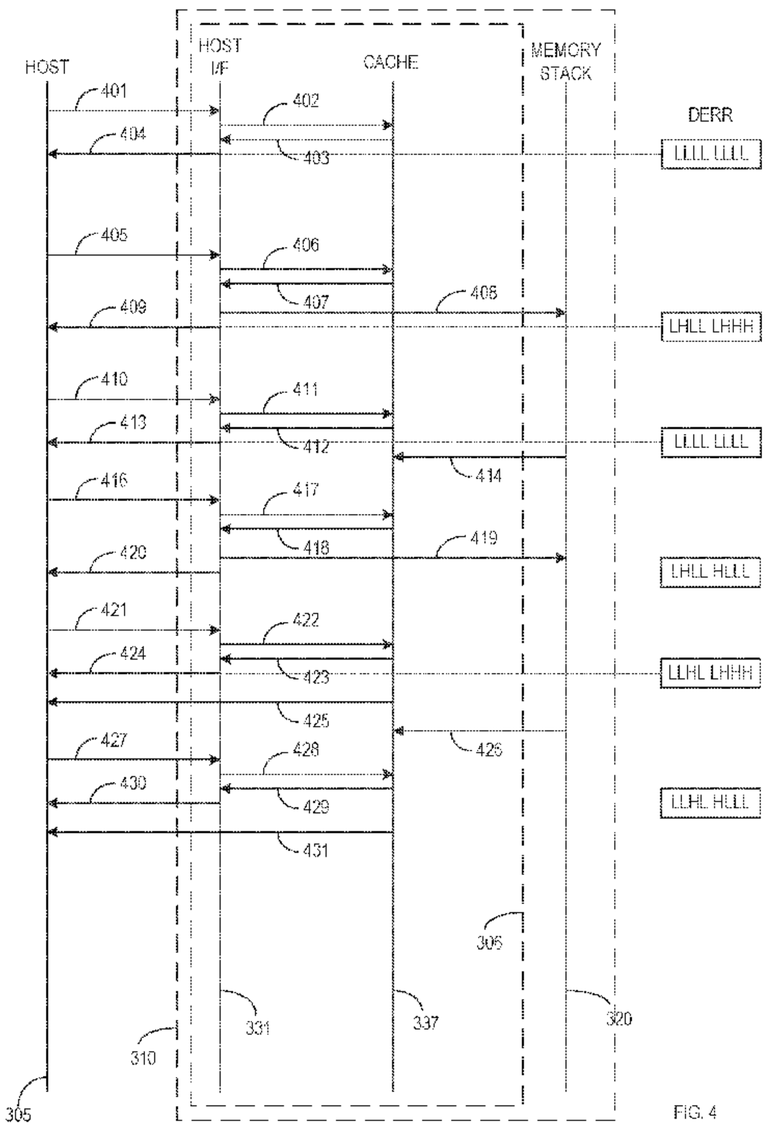

FIG. 4

TECHNIQUES FOR NON-DETERMINISTIC OPERATION OF A STACKED MEMORY SYSTEM

PRIORITY AND RELATED APPLICATIONS

This application claims the benefit of priority to Pawlowski, U.S. Provisional Patent Application No.: 62/953,821, titled, "TECHNIQUES FOR NON-DETERMINISTIC OPERATION OF A STACKED MEMORY SYSTEM, filed December 26, 2019, which is hereby incorporated by reference herein in its entirety.

TECHNICAL FIELD

The following relates generally to operating a memory array and more specifically to increasing bandwidth of a stacked memory device.

BACKGROUND

Memory devices are widely used to store information in various electronic devices such as computers, wireless communication devices, cameras, digital displays, and the like. Information is stored by programming different states of a memory device. For example, binary devices have two states, often denoted by a logic "1" or a logic "0." In other systems, more than two states may be stored. To access the stored information, a component of the electronic device may read, or sense, the stored state in the memory device. To store information, a component of the electronic device may write, or program, the state in the memory device.

Various types of memory devices exist, including magnetic hard disks, random-access memory (RAM), read only memory (ROM), DRAM, synchronous dynamic RAM (SDRAM), ferroelectric RAM (FeRAM), magnetic RAM (MRAM), resistive RAM (RRAM), flash memory, phase change memory (PCM), and others. Memory devices may be volatile or non-volatile.

Improving memory devices, generally, may include increasing memory cell density, increasing read/write speeds, increasing reliability, increasing data retention, reducing power consumption, or reducing manufacturing costs, among other metrics. Advancing memory technology has realized improvements for many of these metrics, however, as improvements in processing speed are developed, memory bandwidth can become a bottleneck to overall system performance improvements.

BRIEF DESCRIPTION OF THE DRAWINGS

In the drawings, which are not necessarily drawn to scale, like numerals may describe similar components in different views. Like numerals having different letter suffixes may represent different instances of similar components. The drawings illustrate generally, by way of example, but not by way of limitation, various embodiments discussed in the present document.

FIG. 1 illustrates an example of a memory die that supports features and operations in accordance with examples of the present disclosure.

FIG. 2 illustrates generally an example of a device that supports features and operations in accordance with examples of the present disclosure.

FIG. 3 illustrates generally an example storage system including a host device that can request and receive information from the storage system according to the present subject matter.

FIG. 4 illustrates generally and example time line of information flow between a host and a channel of an example memory package according to the present subject matter.

FIG. 5 illustrates generally and example method of operating a memory device according to the present subject matter.

FIG. 6 illustrates an example method of operating a host according to the present subject matter.

FIG. 7 illustrates generally a diagram of a system including a device that supports a storage system including stacked DRAM devices in accordance with aspects disclosed herein.

DETAILED DESCRIPTION

Techniques for non-deterministic operation of a stacked memory system are provided. In an example, a method of operating a memory package can include receiving a plurality of, or multiple, memory access requests for a channel at a logic die, returning first data to a host in response to a first memory access request of the plurality of memory access requests, returning an indication of data not ready to the host in response to a second memory access request of the plurality of memory access requests for second data, returning a first index to the host with the indication of data not ready, returning an indication data is ready with third data in response to a third memory access request of the plurality of memory access requests, and returning the first index with the indication of data ready.

FIG. 1 is a schematic diagram of an example memory package 110 including an interface (IF) chip 106, or circuit, and multiple of core chips 100. For example, the memory package 100 may be a 3D memory device, such as an high-bandwidth memory (HBM), a hyper memory cube (HMC), a Wide-IO DRAM etc. The memory package 110 is formed by stacking chips vertically, as shown in FIG. 1. The stacked chips may include two stacks 12 and 13 each assigned with a stack ID "0" and "1", respectively. Each stack 12 and 13 may include core chips 12 a to 12 d and 13 a to 13 d, respectively. In certain examples, each stack can have a number mutiple bit channels per chip. In certain examples, each multiple bit channel can include at least 128 bits for a width of 1024 bits or more across eight channels. The interface (IF) chip 11 of the memory package 110 may provide an interface with the multiple input/output channels. In certain examples, each of the channels can function independently of each other between the core chips 12 a to 12 d and 13 a to 13 d and a host device that may be a memory controller (not shown). The IF chip 106 may couple each channel to the host device via a number of data queues (DQs).

Each channel may include multiple memory cells and circuitries accessing the memory cells. For example, the memory cells may be DRAM memory cells.

FIG. 2 illustrates an apparatus or system 290 that supports channel routing for a memory package in accordance with various examples disclosed herein. The system 290 may include a host device 205 and multiple memory packages 210. In conventional systems, the plurality of memory devices are of the same type, such as DRAM memory devices. In certain examples, the memory devices can include a mix of capacitive based memory devices such as DRAM memory devices and cross-linked inverter memory devices such a SRAM memory devices. The present inventor has recognized that bandwidth improvements can be realized if the host has access to a second, faster type of memory, such as SRAM memory.

The host device 205 may be an example of a processor (e.g., a central processing unit (CPU), a graphics processing unit (GPU)), or a system on a chip (SoC). In some cases, the host device 205 may be a separate component from the memory device such that the host device 205 may be manufactured separately from the memory device. The host device 205 may be external to the memory device 210 (e.g., a laptop, server, personal computing device, smartphone, personal computer). In the system 290, the memory packages 210 may be configured to store data for the host device 205.

The host device 205 may exchange information with the memory packages 210 using signals communicated over signal paths. A signal path may be a path that a message or transmission may take from a transmitting component to a receiving component. In some cases, a signal path may be a conductor coupled with at least two components, where the conductor may selectively allow electrons to flow between the at least two components. The signal path may be formed in a wireless medium as in the case for wireless communications (e.g., radio frequency (RF) or optical). The signal paths may at least partially include a first substrate, such as an organic substrate of the memory device, and/or a second substrate, such as a package substrate (e.g., a second organic substrate) that may be coupled with at least one, if not both, of the memory device 210 and the host device 205. In some cases,

the memory package 210 may function as a slave-type device to the host device 205, which may function as a master-type device.

In some applications, the system 290 may benefit from a high-speed connection between the host device 205 and the memory devices 210. As such, some memory packages 210 support applications, processes, host devices, or processors that have multiple terabytes per second (TB/s) bandwidth needs. Satisfying such a bandwidth constraint within an acceptable energy budget may pose challenges in certain contexts.

The memory dies 200 of the memory packages 210 may be configured to work with multiple types of communication mediums 211 (e.g., substrates such as organic substrates and/or high-density interposers such as silicon interposers). The host device 205 may, in some cases, be configured with an interface or ball-out comprising a design (e.g., a matrix or pattern) of terminals.

In some cases, a buffer layer may be positioned between the memory dies 200 and the communication medium 211. The buffer layer may be configured to drive (e.g., redrive) signals to and from the memory dies 200. In some cases, the stack of memory dies 200 may be bufferless meaning that either no buffer layer is present or that a base layer does not include re-drivers, among other components. In certain examples of bufferless memory, a routing layer or logic die 206 may be positioned between the memory die 200, or stack of memory die 200 and the communication medium 211. In certain examples, the logic die 206 can form a lower layer of a memory die 200. In certain examples, a bufferless memory package 210 can include a lower most memory die 200 having a logic die layer 206.

FIG. 3 illustrates generally an example storage system 391 including a host device 305 and a memory package 310. The host 305 can request and receive information from a memory package 310 according to the present subject matter using a bus external to the memory package. The host device 305 may be, but is not limited to, a CPU, graphics processing unit (GPU), accelerated processing unit (GPU), digital signal processor (DSP), field-programmable gate array (FPGA), application specific integrated circuit (ASIC) and any other component of a larger

system that communicates with the storage system 310. In some embodiments, the device 305 may be multiple devices accessing the same storage system 310. The memory package 310 can include a logic die 306 integrated with a memory stack 320, such as a stack of dynamic random-access memory (DRAM) devices.

The logic die 306 can include a host interface 331 connected to a stacked DRAM control 332 and prefetch and cache logic 333. The stacked DRAM control 332 is connected to and interfaces with the memory stack 320. The prefetch and cache logic 333 can be connected with a prefetcher, prefetch buffers and a cache array 334. The prefetcher may be a hardware prefetcher. The prefetch buffers and cache array 334 may be, but is not limited to, an SRAM array, or any other memory array technology, or a register with faster access speeds than the tyoe of memory used in the memory stack 320.

The host interface 331 can include a command decoder 335 and interface registers 336. The host interface 331, and more specifically, the command decoder 335 can receive all incoming memory requests to the memory stack 320 from the host 305. The requests can be sent to the prefetch and cache logic 333, (for example, next-line, stride, and the like). The prefetch and cache logic 333 can monitor the incoming memory requests. Prefetched data can be placed into the prefetch buffers and cache array 334. The prefetch and cache logic 333 can also check any incoming memory requests against the data in the prefetch buffers and cache array 334. Any hits can be served directly from the prefetch buffers and cache array 334 without going to the stacked DRAM control 332. This can reduce service latencies for these requests, as well as reduce contention in the stacked DRAM control 332 of any remaining requests, (i.e., those that do not hit in the prefetch buffers and cache array 334).

The prefetcher may encompass any prefetching algorithm/method or combination of algorithms/methods. Due to the row-buffer-based organization of most memory technologies, (for example, DRAM), prefetch algorithms that exploit spatial locality, (for example, next-line, small strides and the like), have relatively low overheads because the prefetch requests will (likely) hit in the memory's row

buffer(s). Implementations may issue prefetch requests for large blocks of data, (i.e., more than one 64B cache line's worth of data), such as prefetching an entire row buffer, half of a row buffer, or other granularities.

The prefetch buffers and cache array 334 may be implemented as a direct-mapped, set-associative, to a fully-associative cache-like structure. In an embodiment, the prefetch buffers and cache array 334 may be used to service only read requests, (i.e., writes cause invalidations of prefetch buffer entries, or a write-through policy must be used). In another embodiment, the prefetch buffers and cache array 334 may employ replacement policies such as Least Recently Used (LRU), Least Frequency Used (LFU), or First-In-First-Out (FIFO). If the prefetch unit generates requests for data sizes larger than a cache line, the prefetch buffers and cache array 334 may also need to be organized with a correspondingly wider data block size. In some embodiments, sub-blocking may be used.

While described herein as being employed in a memory organization consisting of one logic chip and one or more memory chips, there are other physical manifestations. Although described as a vertical stack of a logic die with one or more memory chips, another embodiment may place some or all of the logic on a separate chip horizontally on an interposer or packaged together in a multi-chip module (MCM). More than one logic chip may be included in the overall stack or system.

As discussed above, the prefetch and cache logic 333 can also check any incoming memory requests against the data in the prefetch buffers and cache array 334. Any hits can be served directly from the prefetch buffers and cache array 334 without going to the stacked DRAM control 332. This can reduce service latencies for these requests, as well as reduce contention in the stacked DRAM control 332 of any remaining requests. However, prefetch or cache misses can still occur.

Conventional architectures provide a data error (DERR) indication at the host interface for each channel of the memory stack when a request cannot be serviced. When a read request, for example, of the host fails to hit data associated with the prefetch or the catch, the host interface indicates the cache miss by setting the data

error (DERR) indication for the corresponding channel to a respective logic level. Upon seeing the indication of the data error, the host typically moves to the next access request and then reissues the failed request at some later time.

The present inventor has recognized that a logic die having a memory-side SRAM cache as described above can operate to assist and improve accessing memory data after a missed access request. In this context a "missed access request" is one that the memory interface receives but is unable to process from the cache at that time. In certain examples, the system 391 can take advantage of the burst length of the memory system to encode additional information associated with missed access requests. Generally, burst length is the number of clock cycles a channel uses to exchange information across the data queues (DQs). In the example discussed above with respect to FIG. 1, a channel width of 32 can require eight clock cycles to provide a 128 bit channel length word, so the burst length is eight. In certain examples, each cycle of a burst length can use the physical data error (DERR) bit of each channel I/O to convey 8 bits worth of information about, for example, a true data error, whether a prior request is ready, and an index associated with a missed request. The eight bits can be transmitted serially during a burst using the physical data error bit of each channel I/O. As an example, the serial communicated bits of the physical data error I/O point (DERR) can be assigned as follows:

Bit	Function
0	data error
1	not Ready
2	alert
3-7	index

It is understood that other sequences of bits assignments are possible without departing from the scope of the present subject matter. The "data error" function a little different than conventional methods in that when a request cannot be serviced and will not be serviced, the data error bit can be activated. In such a scenario, upon

receiving a "data error" indication after making a memory access request, the host will need to re-issue the memory access request at least once.

The "not ready" function, or bit, can be activated when a request is a missed request. For example, the "not ready" function/bit, can be activated when the request is received, the request is not able to service by the cache, but the interface plans to attempt to read the data associated with the missed request to the cache in the near future. In addition to activating the "not ready" function/bit, the index bits (e.g., 3-7) can provide an index number associated with the missed request on that channel. The index number can be used by the host to track and later capture the corresponding data associated with the prior missed request.

The "alert" function, or bit, can be activated when data associated with a prior missed request is available in the cache. In addition, when the alert function/bit is activated, the index bits can provide an index number corresponding to the missed request. In some examples, the host interface can provide the data in an interval immediately following the activation of the corresponding "alert" function. In some example, the "alert" function can re-initiate the missed read access request associated with the index to get the corresponding data. In certain examples, the cache can include a list of index numbers associated with pending requests.

In the above example, the burst length allows for five index bits. As such, each channel can have up to 32 (e.g., 2^5) missed requests pending at any one time. In some examples, some of the index numbers can be assigned to convey other information and thus, the number of pending missed requests can be less. In addition, an example memory system can have a burst length or channel width different than the example above without departing from the scope of the present subject matter.

FIG. 4 illustrates generally and example time line 400 of information flow between a host 305 and a channel of an example memory package 310 according to the present subject matter. The memory package 310 can include logic die 306 and a stack of memory 320, such as a stack of DRAM. The logic die 306 can include a host interface 331 and a cache 337. In certain examples, the cache 337 can include

SRAM. In certain examples, the cache 337 can include buffers and tag maps associated with conventional stacked memory devices. At 401, the host 305 can request data from the memory system and the request can be received at the host interface 331. At 402, the host interface 331 can request the data from the cache 337. At 403, after determining the data is in the cache 337, the data can be received at the host interface 331 and at 404, the data can be passed to the host 305. At 405, the host 305 can request second data from the memory system and the request can be received at the host interface 331. At 406, the host interface 331 can request the data from the cache 337. At 407, after determining the data is not in the cache 337, the status of the cache request can be received at the host interface 331. At 408, the host interface 331 can request the second data from the memory stack 320, and at 409, the missed request can be reported to the host 305.

In certain examples, at 409, when reporting the missed request, the host interface 331 can use the physical data error (DERR) output point for the channel to indicate that the request is not ready, and also to provide an index number for the request. A compatible host 305 can use the index number to later retrieve the requested data as discussed below. In certain examples, since accessing the memory stack 320 is more time consuming than servicing data requests at the cache 337, the host interface 331 can make the channel available for further data requests while the second data is being retrieved. As such, at 410, the host 305 can request third data from the memory system and the request can be received at the host interface 331. At 411, the host interface 331 can request the data from the cache 337. At 412, after determining the data is in the cache 337, the data can be received at the host interface 331 and at 413, the data can be passed to the host 305. At 414, the second data can be passed from the memory stack 320 to the cache 337. It is understood that the host interface 331 may be able to receive more than one data request from the host 305 before receiving the data of a missed request without departing from the scope of the present subject matter.

At 416, the host 305 can request fourth data from the memory system and the request can be received at the host interface 331. At 417, the host interface 331

can request the data from the cache 337. At 418, after determining the data is not in the cache 337, the status of the cache request can be received at the host interface 331. At 419, the host interface 331 can request the fourth data from the memory stack 320, and at 420, the missed request can be reported to the host 305. As before, in certain examples, at 420, when reporting the missed request, the host interface 331 can use the physical data error (DERR) output point for the channel to indicate that the request is not ready, and also to provide an index number for the request. The index number can be different than the index number associated with reporting the missing request for the second data at 409.

At 421, the host 305 can request fifth data from the memory system and the request can be received at the host interface 331. At 422, the host interface 331 can request the data from the cache 337. At 423, after determining the data is in the cache 337, the data can be received at the host interface 331 and at 424, the data can be passed to the host 305. In addition, when the fifth data is passed to the host 305, the host interface 331 can activate the "data" ready bit, or "alert" bit discussed above, during the burst and can also provide the corresponding index number for the request for the second data as the second data is now available in the cache 337. In certain examples, the cache 337 at 425 can pass the second data to the host 305 via the host interface 331. In some examples, the host interface 331 may require that the host 305 resend a request for the second data before passing the second data. In some examples, the passing the second data via the host interface 331 may involve separate transactions between the cache 337 and the host interface 331 and between the host 305 interface and the host 305, and each transaction can be separated in time different than that illustrated without departing from the scope of the present subject matter. At 426, the fourth data can be passed from the memory stack 320 to the cache 337.

At 427, the host 305 can request sixth data from the memory system and the request can be received at the host interface 331. At 428, the host interface 331 can request the sixth data from the cache 337. At 429, after determining the sixth data is in the cache 337, the sixth data can be received at the host interface 331 and at 430,

the sixth data can be passed to the host 305. In addition, when the sixth data is passed to the host 305, the host interface 331 can activate a data readiness indication such as the "data ready" bit during the DERR burst and can also provide the corresponding index number for the request for the fourth data as the fourth data is now available in the cache 337. In certain examples, the cache 337 at 431 can pass the fourth data to the host 305 via the host interface 331. In some examples, the host interface 331 may require that the host 305 resend a request for the fourth data before passing the fourth data from the cache 337. In some examples, the passing the fourth data via the host interface 331 may involve separate transactions between the cache 337 and the host interface 331 and between the host interface 331 and the host 305, and each transaction can be separated in time different than that illustrated without departing from the scope of the present subject matter.

FIG. 5 illustrates generally and example method 500 of operating a memory device according to the present subject matter. At 501, a plurality of memory access requests can be received from a host at a memory system including a stack of DRAM memory devices. At 503, first data can be returned to the host in response to a first memory access request of the plurality of memory access requests. At 505, a first indication of data not ready can be returned to the host in response to a second memory access request of the plurality of memory access requests. At 507, an index can be returned with the first indication of data not ready. At 509, an indication of data ready can be returned with third data in response to a third memory access request of the plurality of memory access requests. At 511, the first index can be returned with the indication of data ready. At 513, second data can be returned to the host in response to the indication of data ready and the first index received with the third memory access request.

FIG. 6 illustrates an example method 600 of operating a host according to the present subject matter. At 601, multiple memory access requests can be sent to a memory system having a stack of memory devices. At 603, first data can be received in response to a first request of the multiple memory access requests. At 605, an indication of data not ready can be received in response to a second request

of the multiple memory access requests, wherein the second request is a request for second data. At 607, an index can be received with the indication of data not ready. At 609, third data can be received in response to a third request of the mutiple memory access requests. At 611, an indication of data ready and the index can be received with the third data. At 613, the second data can be received in response to the indication of data ready and the index.

FIG. 7 illustrates generally a diagram of a system 700 including a device 705 that supports a storage system including stacked DRAM devices in accordance with aspects disclosed herein. Device 705 may include components for bi-directional voice and data communications including components for transmitting and receiving communications, including memory controller 715, memory cells 720, basic input/output system (BIOS) component 725, processor 730, I/O controller 735, peripheral components 740, memory chip 755, system memory controller 760, encoder 765, decoder 770, and multiplexer 775. These components may be in electronic communication via one or more busses (e.g., bus 710). Bus 710, for example, may have a bus width of 16 data lines ("DQ" lines). Bus 710 may be in electronic communication with 32 banks of memory cells.

Memory controller 715 or 760 may operate one or more memory cells as described herein. Specifically, memory controller may be configured to support flexible multi-channel memory. In some cases, memory controller 715 or 760 may operate a row decoder, column decoder, or both, as described with reference to FIG. 1. Memory controller 715 or 760 may be in electronic communication with a host and may be configured to transfer data during each of a rising edge and a falling edge of a clock signal of the memory controller 715 or 760.

Memory cells 720 may store information (i.e., in the form of a logical state) as described herein. Memory cells 720 may represent, for example, memory cells 105 described with reference to FIG. 1. Memory cells 720 may be in electronic communication with memory controller 715 or 760, and memory cells 720 and memory controller 715 or 760 may be located on a chip 755, which may be one or

several planar memory devices as described herein. Chip 755 may, for example, be managed by system memory controller 715 or 760.

Memory cells 720 may represent a first array of memory cells with a plurality of regions coupled to a substrate. Each region of the plurality of regions may include a plurality of banks of memory cells and a plurality of channels traversing the first array of memory cells. At least one of the plurality of channels may be coupled to at least one region. Memory controller 715 or 760 may be configured to transfer data between the coupled region and the memory controller 715 or 760.

BIOS component 725 be a software component that includes BIOS operated as firmware, which may initialize and run various hardware components. BIOS component 725 may also manage data flow between a processor and various other components, e.g., peripheral components, input/output control component, etc. BIOS component 725 may include a program or software stored in read only memory (ROM), flash memory, or any other non-volatile memory.

Processor 730 may include an intelligent hardware device, (e.g., a general-purpose processor, a digital signal processor (DSP), a central processing unit (CPU), a microcontroller, an application-specific integrated circuit (ASIC), a field programmable gate array (FPGA), a programmable logic device, a discrete gate or transistor logic component, a discrete hardware component, or any combination thereof). In some cases, processor 730 may be configured to operate a memory array using a memory controller 715 or 760. In other cases, a memory controller 715 or 760 may be integrated into processor 730. Processor 730 may be configured to execute computer-readable instructions stored in a memory to perform various functions (e.g., functions or tasks supporting flexible multi-channel memory).

I/O controller 735 may manage input and output signals for device 705. I/O controller 735 may also manage peripherals not integrated into device 705. In some cases, I/O controller 735 may represent a physical connection or port to an external peripheral. I/O controller 735 may utilize an operating system such as iOS®, ANDROID®, MS-DOS®, MS-WINDOWS®, OS/2®, UNIX®, LINUX®, or

another known operating system. In other cases, I/O controller 735 may represent or interact with a modem, a keyboard, a mouse, a touchscreen, or a similar device. In some cases, I/O controller 735 may be implemented as part of a processor. A user may interact with device 705 via I/O controller 735 or via hardware components controlled by I/O controller 735.

Peripheral components 740 may include any input or output device, or an interface for such devices. Examples may include disk controllers, sound controller, graphics controller, Ethernet controller, modem, universal serial bus (USB) controller, a serial or parallel port, or peripheral card slots, such as peripheral component interconnect (PCI) or accelerated graphics port (AGP) slots.

Input 745 may represent a device or signal external to device 705 that provides input to device 705 or its components. This may include a user interface or an interface with or between other devices. In some cases, input 745 may be managed by I/O controller 735, and may interact with device 705 via a peripheral component 740.

Output 750 may also represent a device or signal external to device 705 configured to receive output from device 705 or any of its components. Examples of output 750 may include a graphics display, audio speakers, a printing device, another processor or printed circuit board, etc. In some cases, output 750 may be a peripheral element that interfaces with device 705 via peripheral component(s) 740. Output 750 may be managed by I/O controller 735.

System memory controller 715 or 760 may be in electronic communication with a first array of memory cells (e.g., memory cells 720). A host may be a component or device that controls or directs operations for a device of which memory controller 715 or 760 and corresponding memory array are a part. A host may be a component of a computer, mobile device, or the like. Or device 705 may be referred to as a host. In some examples, system memory controller 715 or 760 is a GPU.

Encoder 765 may represent a device or signal external to device 705 that provides performs error correction encoding on data to be stored to device 705 or its

components. Encoder 765 may write the encoded data to the at least one selected memory via the at least one channel and may also encode data via error correction coding.

Decoder 770 may represent a device or signal external to device 705 that sequences command signals and addressing signals to device 705 or its components. In some examples, memory controller 715 or 760 may be co-located within decoder 770.

Multiplexer 775 may represent a device or signal external to device 705 that multiplexes data to device 705 or its components. Multiplexer 775 may multiplex the data to be transmitted to the encoder 765 and de-multiplex data received from the encoder 765. A multiplexer 775 may be in electronic communication with the decoder 770. In some examples, multiplexer 775 may be in electronic communication with a controller, such as system memory controller 715 or 760.

The components of device 705 may include circuitry designed to carry out their functions. This may include various circuit elements, for example, conductive lines, transistors, capacitors, inductors, resistors, amplifiers, or other active or inactive elements, configured to carry out the functions described herein. Device 705 may be a computer, a server, a laptop computer, a notebook computer, a tablet computer, a mobile phone, a wearable electronic device, a personal electronic device, or the like. Or device 705 may be a portion or aspect of such a device. In some examples, device 705 is an aspect of a computer with high reliability, mission critical, or low latency constraints or parameters, such as a vehicle (e.g., an autonomous automobile, airplane, a spacecraft, or the like). Device 705 may be or include logic for artificial intelligence (AI), augmented reality (AR), or virtual reality (VR) applications.

In one example, a memory device may include an array of memory cells with a plurality of regions that may each may include a plurality of banks of memory cells, and a plurality of channels traversing the array of memory cells. Each of the channels may be coupled with a region of the array of memory cells and may

be configured to communicate signals between the plurality of banks of memory cells in the region with a host device.

In some examples, the memory device may further include I/O areas extending across the array of memory cells, the I/O areas occupying an area of the array of memory cells that may be devoid of memory cells. In some examples of the memory device, the I/O areas may include TSVs configured to couple the array of memory cells with a power node or a ground node.

In some examples, the memory device may further include a plurality of channel interfaces distributed in the array of memory cells. In some examples of the memory device, the plurality of channel interfaces may be bump-outs. In some examples of the memory device, a channel interface of the plurality of channel interfaces may be positioned in each quadrant of the array of memory cells.

In some examples, the memory device may further include a plurality of signal paths extending between memory cells of the region and a channel interface associated with the region. In some examples of the memory device, the channel interface may be positioned in the array of memory cells to minimize a length of the signal paths.

In some examples, the memory device may further include a second array of memory cells stacked on top of the array of memory cells. In some examples of the memory device, the second array of memory cells may have regions that may each include a plurality of banks of memory cells. In some examples, the memory device may further include a second plurality of channels traversing the second array of memory cells. In some examples of the memory device, each of the channels of the second plurality of channels may be coupled with a second region of the second array of memory cells and may be configured to communicate signals between the plurality of banks of memory cells in the second region with the host device.

In some examples, the memory device may further include TSVs extending through the array of memory cells to couple the second array of memory cells with the second plurality of channels. In some examples of the memory device, a channel may establish a point-to-point connection between the region and the host device. In

some examples of the memory device, each channel may include four or eight data pins. In some examples of the memory device, the region of the array of memory cells may include eight or more banks of memory cells.

In some examples, the memory device may further include an interface configured for bidirectional communication with the host device. In some examples of the memory device, the interface may be configured to communicate signals modulated using at least one of a NRZ modulation scheme or a PAM4 scheme, or both.

In one example, a memory device may include an array of memory cells with regions that each include a plurality of banks of memory cells, I/O areas extending across the array of memory cells, the I/O areas may include a plurality of terminals configured to route signals to and from the array of memory cells, and a plurality of channels positioned in the I/O areas of the array of memory cells, each of the channels may be coupled with a region of the array of memory cells and may be configured to communicate signals between the plurality of banks of memory cells in the region with a host device.

In some examples, the memory device may further include a plurality of channel interfaces positioned in the I/O areas of the array of memory cells, signal paths couple the regions with the plurality of channel interfaces. In some examples of the memory device, the I/O areas may include TSVs configured to couple a second array of memory cells stacked on top of the array of memory cells with a channel interface.

In some examples of the memory device, a channel interface of the region may be positioned within an I/O area that bisects the region serviced by the channel interface. In some examples of the memory device, the I/O areas may include TSVs configured to couple the array of memory cells with a power node or a ground node.

In some examples of the memory device, the I/O areas may occupy an area of the array of memory cells that may be devoid of memory cells. In some examples of the memory device, the array of memory cells may be bisected by two I/O areas. In

some examples of the memory device, the array of memory cells may be bisected by four I/O areas.

In one example, a system may include a host device, a memory device including a memory die with a plurality of regions that may each include a plurality of banks of memory cells, and a plurality of channels configured to communicatively couple the host device and the memory device, each of the channels may be coupled with a region of the memory die and may be configured to communicate signals between the plurality of banks of memory cells in the region with the host device.

In some examples, the system may include an interface configured for bidirectional communication with the host device. In some examples of the system, the interface may be configured to communicate signals modulated using at least one of a NRZ modulation scheme or a PAM4 scheme, or both. In some examples of the system, the host device may be an example of a GPU. In some examples of the system, the memory device may be positioned in a same package as the host device.

In one example, a memory device may include an array of memory cells with a plurality of regions that each include a plurality of banks of memory cells, and a plurality of channels traversing the array of memory cells, each of the channels may be coupled to at least one region of the array of memory cells and each channel may include two or more data pins and one or more command/address pin.

In some examples of the memory device, each channel may include two data pins. In some examples of the memory device, each channel may include one command/address pin. In some examples of the memory device, each region of the array may include four banks of memory cells. In some examples of the memory device, each channel may include four data pins. In some examples of the memory device, each channel may include two command/address pins. In some examples of the memory device, each region of the array may include eight banks of memory cells. In some examples of the memory device, each bank of memory cells may be contiguous with a channel.

In some examples of the memory device, a first set of banks of each plurality may be contiguous with a channel and a second set of banks of each plurality may be contiguous with another bank and non-contiguous with a channel. In some examples, the memory device may include 128 data pins and configured with a ratio of two, four, or eight data pins per channel.

In some examples, the memory device may include one, two, three, four, or six command/address pins per channel. In some examples, the memory device may include 256 data pins and configured with a ratio of two, four, or eight data pins per channel. In some examples, the memory device may include one, two, three, four, or six command/address pins per channel. In some examples of the memory device, the array may include a plurality of memory dice that each may include a plurality of channels.

In some examples of the memory device, each memory die of the plurality may be coupled with a different channel of the plurality of channels. In some examples, the memory device may include a buffer layer coupled with array. In some examples, the memory device may include an organic substrate underlying the array.

In some examples of the memory device, the array may be configured for a pin rate of 10, 16, 20, or 24 Gbps. In some examples, the memory device may include an interface configured for bidirectional communication with a host device. In some examples of the memory device, the interface may be configured for at least one of a binary modulation signaling or pulse-amplitude modulation, or both.

In one example, a system may include at least one memory die that may include a plurality of regions that each may include a plurality of banks of memory cells, one or more channels associated with each memory die, each of the channels may be coupled to at least one region of the die of memory cells and each channel may include two or more data pins, and an organic substrate that underlies the memory die.

In some examples, the system may include a host device, and an interface configured for bidirectional communication with the host device, the interface

supports at least one of a NRZ signaling or a PAM4, or both. In some examples of the system, the host device may include a GPU.

In some examples, the system may include a plurality of memory arrays that each may include 128 or 256 data pins and configured with a ratio of two, four, or eight data pins per channel. In some examples, the system may include a buffer layer positioned between the at least one memory die and the organic substrate.

Information and signals described herein may be represented using any of a variety of different technologies and techniques. For example, data, instructions, commands, information, signals, bits, symbols, and chips that may be referenced throughout the above description may be represented by voltages, currents, electromagnetic waves, magnetic fields or particles, optical fields or particles, or any combination thereof. Some drawings may illustrate signals as a single signal; however, it will be understood by a person of ordinary skill in the art that the signal may represent a bus of signals, where the bus may have a variety of bit widths.

As may be used herein, the term "virtual ground" refers to a node of an electrical circuit that is held at a voltage of approximately zero volts (0V) but that is not directly connected with ground. Accordingly, the voltage of a virtual ground may temporarily fluctuate and return to approximately 0V at steady state. A virtual ground may be implemented using various electronic circuit elements, such as a voltage divider consisting of operational amplifiers and resistors. Other implementations are also possible. "Virtual grounding" or "virtually grounded" means connected to approximately 0V.

The may be used herein, the term "electronic communication" and "coupled" refer to a relationship between components that support electron flow between the components. This may include a direct connection between components or may include intermediate components. Components in electronic communication or coupled to one another may be actively exchanging electrons or signals (e.g., in an energized circuit) or may not be actively exchanging electrons or signals (e.g., in a de-energized circuit) but may be configured and operable to exchange electrons or signals upon a circuit being energized. By way of example, two components

physically connected via a switch (e.g., a transistor) are in electronic communication or may be coupled regardless of the state of the switch (i.e., open or closed).

The term "layer" used herein refers to a stratum or sheet of a geometrical structure. Each layer may have three dimensions (e.g., height, width, and depth) and may cover some or all of a surface. For example, a layer may be a three-dimensional structure where two dimensions are greater than a third, e.g., a thin-film. Layers may include different elements, components, and/or materials. In some cases, one layer may be composed of two or more sublayers. In some of the appended figures, two dimensions of a three-dimensional layer are depicted for purposes of illustration. Those skilled in the art will, however, recognize that the layers are three-dimensional in nature.

As used herein, the term "electrode" may refer to an electrical conductor, and in some cases, may be employed as an electrical contact to a memory cell or other component of a memory array. An electrode may include a trace, wire, conductive line, conductive layer, or the like that provides a conductive path between elements or components of a memory array.

The term "isolated" refers to a relationship between components in which electrons are not presently capable of flowing between them; components are isolated from each other if there is an open circuit between them. For example, two components physically connected by a switch may be isolated from each other when the switch is open.

The devices discussed herein, including a memory array, may be formed on a semiconductor substrate, such as silicon, germanium, silicon-germanium alloy, gallium arsenide, gallium nitride, etc. In some cases, the substrate is a semiconductor wafer. In other cases, the substrate may be a silicon-on-insulator (SOI) substrate, such as silicon-on-glass (SOG) or silicon-on-sapphire (SOP), or epitaxial layers of semiconductor materials on another substrate. In some examples, the substrate may be an organic build up substrate formed from materials such as ABF or BT. The conductivity of the substrate, or sub-regions of the substrate, may be controlled through doping using various chemical species including, but not

limited to, phosphorous, boron, or arsenic. Doping may be performed during the initial formation or growth of the substrate, by ion-implantation, or by any other doping means.

A transistor or transistors discussed herein may represent a field-effect transistor (FET) and comprise a three terminal device including a source, drain, and gate. The terminals may be connected to other electronic elements through conductive materials, e.g., metals. The source and drain may be conductive and may comprise a heavily-doped, e.g., degenerate, semiconductor region. The source and drain may be separated by a lightly-doped semiconductor region or channel. If the channel is n-type (i.e., majority carriers are electrons), then the FET may be referred to as a n-type FET. If the channel is p-type (i.e., majority carriers are holes), then the FET may be referred to as a p-type FET. The channel may be capped by an insulating gate oxide. The channel conductivity may be controlled by applying a voltage to the gate. For example, applying a positive voltage or negative voltage to an n-type FET or a p-type FET, respectively, may result in the channel becoming conductive. A transistor may be "on" or "activated" when a voltage greater than or equal to the transistor's threshold voltage is applied to the transistor gate. The transistor may be "off" or "deactivated" when a voltage less than the transistor's threshold voltage is applied to the transistor gate.

The various illustrative blocks and modules described in connection with the disclosure herein may be implemented or performed with a general-purpose processor, a DSP, an ASIC, an FPGA or other programmable logic device, discrete gate or transistor logic, discrete hardware components, or any combination thereof designed to perform the functions described herein. A general-purpose processor may be a microprocessor, but in the alternative, the processor may be any processor, controller, microcontroller, or state machine.

A processor may also be implemented as a combination of computing devices (e.g., a combination of a DSP and a microprocessor, multiple microprocessors, one or more microprocessors in conjunction with a DSP core, or any other such configuration).

In a first example, Example 1, a storage system can include a stack of first memory devices configured to store data, the stack including multiple memory die of a first storage type, and a logic die. The logic die can include second memory of a second storage type, and an interface circuit. The interface circuit can be configured to receive a multiple of memory requests from an external host using an external bus, to relay data between the external host and multiple channels of the stack of first memory devices via the second memory, to provide an indication of a data error on a first output bit of a corresponding channel during a single clock cycle of multiple clock cycles used to exchange the data with the corresponding channel of the stack of first memory devices in response to a first respective memory request of the multiple on memory requests, to provide an indication of data readiness associated with a second respective memory request of the multiple of memory requests, and to provide a first index configured to identify data associated with the second respective memory request.

In Example 2, the second memory of Example 1 optionally comprises a list configured to store multiple indexes including the first index.

In Example 3, the indication of data readiness of any one or more of Examples 1-2 optionally is encoded on the first output bit during the multiple clock cycles.

In Example 4, the indication of data readiness of any one or more of Examples 1-3 optionally includes an indication data is not ready.

In Example 5, the indication of data readiness of any one or more of Examples 1-4 optionally includes an indication data is ready.

In Example 6, the first index of any one or more of Examples 1-5 optionally is encoded on the first output bit during the multiple clock cycles.

In Example 7, the first storage type of any one or more of Examples 1-6 optionally is dynamic random-access memory (DRAM).

In Example 8, the second storage type of any one or more of Examples 1-7 optionally is static random-access memory (SRAM).

In Example 9, is a method of operating a memory package having a logic die and stack of memory devices, the memory package can be configured to communicate with a host using multiple independent channels. The method can include receiving multiple memory access requests for a channel at the logic die, returning first data to the host in response to a first memory access request of the multiple memory access requests, returning an indication of data not ready to the host in response to a second memory access request of the multiple memory access requests for second data, returning a first index to the host with the indication of data not ready, returning an indication data is ready with third data in response to a third memory access request of the multiple memory access requests, returning the first index with the indication of data ready, and returning the second data to the host in response to the indication of data ready and the first index returned with the third memory access request.

In Example 10, the returning an indication of data not ready of any one or more of Examples 1-9 optionally includes encoding the indication of data not ready on a single output bit of the channel over a first single cycle of a burst of cycles used to exchange data of a respective memory access request of the multiple memory access requests with the host.

In Example 11, the returning an indication of data ready of any one or more of Examples 1-10 optionally includes encoding the indication of data ready on the single output bit of the channel over a second single cycle of the burst of cycles.

In Example 12, the returning the first index of any one or more of Examples 1-11 optionally includes encoding the first index on the single output bit of the channel over multiple cycles of the burst of cycles.

In Example 13, the multiple cycles of any one or more of Examples 1-12 optionally does not include the first single cycle or the second single cycle.

In Example 14, the method of any one or more of Examples 1-13 optionally includes returning a data error (DERR) indication when the memory package is unable to service a memory access request of the multiple memory access requests, wherein the DERR indication is encoded on the single output bit.

In Example 15, a method of operating a host configured to exchange information using multiple independent channels of a memory package including a stack of memory devices can include sending a plurality of memory access requests to the memory package using a single channel, receiving first data via the single channel in response to a first memory access request of the plurality of memory access requests, receiving an indication of data not ready via the single channel in response to a second memory access request of the plurality of memory access requests for second data, receiving a first index with the indication of data not ready, receiving an indication data is ready with third data via the single channel in response to a third memory access request of the plurality of memory access requests, receiving the first index with the indication of data ready, and receiving the second data via the single channel in response to the indication of data ready and the first index received in response to the third memory access request.

In Example 16, the receiving an indication of data not ready of any one or more of Examples 1-15 optionally includes decoding the indication of data not ready on a single output bit of the single channel over a first single cycle of a burst of cycles, the burst of cycles used to exchange data of a respective memory access request of the plurality of memory access requests between the host and the memory package.

In Example 17, the receiving an indication of data ready of any one or more of Examples 1-16 optionally includes decoding the indication of data ready on the single output bit of the single channel over a second single cycle of the burst of cycles.

In Example 18, the receiving the first index of any one or more of Examples 1-17 optionally includes decoding the first index on the single output bit of the single channel over a plurality of cycles of the burst of cycles.

In Example 19, the plurality of cycles of any one or more of Examples 1-18 optionally does not include the first single cycle or the second single cycle.

In Example 20, the method of any one or more of Examples 1-11 optionally includes receiving a data error (DERR) indication when the memory package is

unable to service a memory access request of the plurality of memory access requests, wherein the DERR indication is encoded on the single output bit.

Example 21 can include or use, or can optionally be combined with any portion or combination of any portions of any one or more of Examples 1 through 20 to include or use, subject matter that can include means for performing any one or more of the functions of Examples 1 through 20, or a machine-readable medium including instructions that, when performed by a machine, cause the machine to perform any one or more of the functions of Examples 1 through 20.

Claims

What is claimed is:

1. A storage system comprising:

a stack of first memory devices configured to store data, the stack including multiple memory die of a first storage type; and

a logic die including:

second memory of a second storage type; and

an interface circuit configured to,

receive a multiple of memory requests from an external host using an external bus,

relay data between the external host and multiple channels of the stack of first memory devices via the second memory,

provide an indication of a data error on a first output bit of a corresponding channel during a single clock cycle of multiple clock cycles used to exchange the data with the corresponding channel of the stack of first memory devices in response to a first respective memory request of the multiple on memory requests,

provide an indication of data readiness associated with a second respective memory request of the multiple of memory requests, and

provide a first index configured to identify data associated with the second respective memory request.

2. The storage system of claim 1, wherein second memory comprises a list configured to store multiple indexes including the first index.

3. The storage system of claim 1, wherein the indication of data readiness is encoded on the first output bit during the multiple clock cycles.

4. The storage system of claim 3, wherein the indication of data readiness includes an indication data is not ready.

5. The storage system of claim 3, wherein the indication of data readiness includes an indication data is ready.

6. The storage system of claim 1, wherein the first index is encoded on the first output bit during the multiple clock cycles.

7. The storage system of claim 1, wherein the first storage type is dynamic random-access memory (DRAM).

8. The storage system of claim 1, wherein the second storage type is static random-access memory (SRAM).

9. A method of operating a memory package having a logic die and stack of memory devices, the memory package configured to communicate with a host using multiple independent channels, the method comprising:

receiving multiple memory access requests for a channel at the logic die;

returning first data to the host in response to a first memory access request of the multiple memory access requests;

returning an indication of data not ready to the host in response to a second memory access request of the multiple memory access requests for second data;

returning a first index to the host with the indication of data not ready;

returning an indication data is ready with third data in response to a third memory access request of the multiple memory access requests;

returning the first index with the indication of data ready; and

returning the second data to the host in response to the indication of data ready and the first index returned with the third memory access request.

10. The method of claim 9, wherein returning an indication of data not ready includes encoding the indication of data not ready on a single output bit of the channel over a first single

cycle of a burst of cycles used to exchange data of a respective memory access request of the multiple memory access requests with the host.

11. The method of claim 10, wherein returning an indication of data ready includes encoding the indication of data ready on the single output bit of the channel over a second single cycle of the burst of cycles.

12. The method of claim 11, wherein returning the first index includes encoding the first index on the single output bit of the channel over multiple cycles of the burst of cycles.

13. The method of claim 12, wherein the multiple cycles does not include the first single cycle or the second single cycle.

14. The method of claim 10, including returning a data error (DERR) indication when the memory package is unable to service a memory access request of the multiple memory access requests, wherein the DERR indication is encoded on the single output bit.

15. A method of operating a host configured to exchange information using multiple independent channels of a memory package including a stack of memory devices, the method comprising:

sending a plurality of memory access requests to the memory package using a single channel;

receiving first data via the single channel in response to a first memory access request of the plurality of memory access requests;

receiving an indication of data not ready via the single channel in response to a second memory access request of the plurality of memory access requests for second data;

receiving a first index with the indication of data not ready;

receiving an indication data is ready with third data via the single channel in response to a third memory access request of the plurality of memory access requests;

receiving the first index with the indication of data ready; and

receiving the second data via the single channel in response to the indication of data ready and the first index received in response to the third memory access request.

16. The method of claim 15, wherein receiving an indication of data not ready includes decoding the indication of data not ready on a single output bit of the single channel over a first single cycle of a burst of cycles, the burst of cycles used to exchange data of a respective memory access request of the plurality of memory access requests between the host and te memory package.

17. The method of claim 16, wherein receiving an indication of data ready includes decoding the indication of data ready on the single output bit of the single channel over a second single cycle of the burst of cycles.

18. The method of claim 17, wherein receiving the first index includes decoding the first index on the single output bit of the single channel over a plurality of cycles of the burst of cycles.

19. The method of claim 18, wherein the plurality of cycles does not include the first single cycle or the second single cycle.

20. The method of claim 16, including receiving a data error (DERR) indication when the memory package is unable to service a memory access request of the plurality of memory access requests, wherein the DERR indication is encoded on the single output bit.

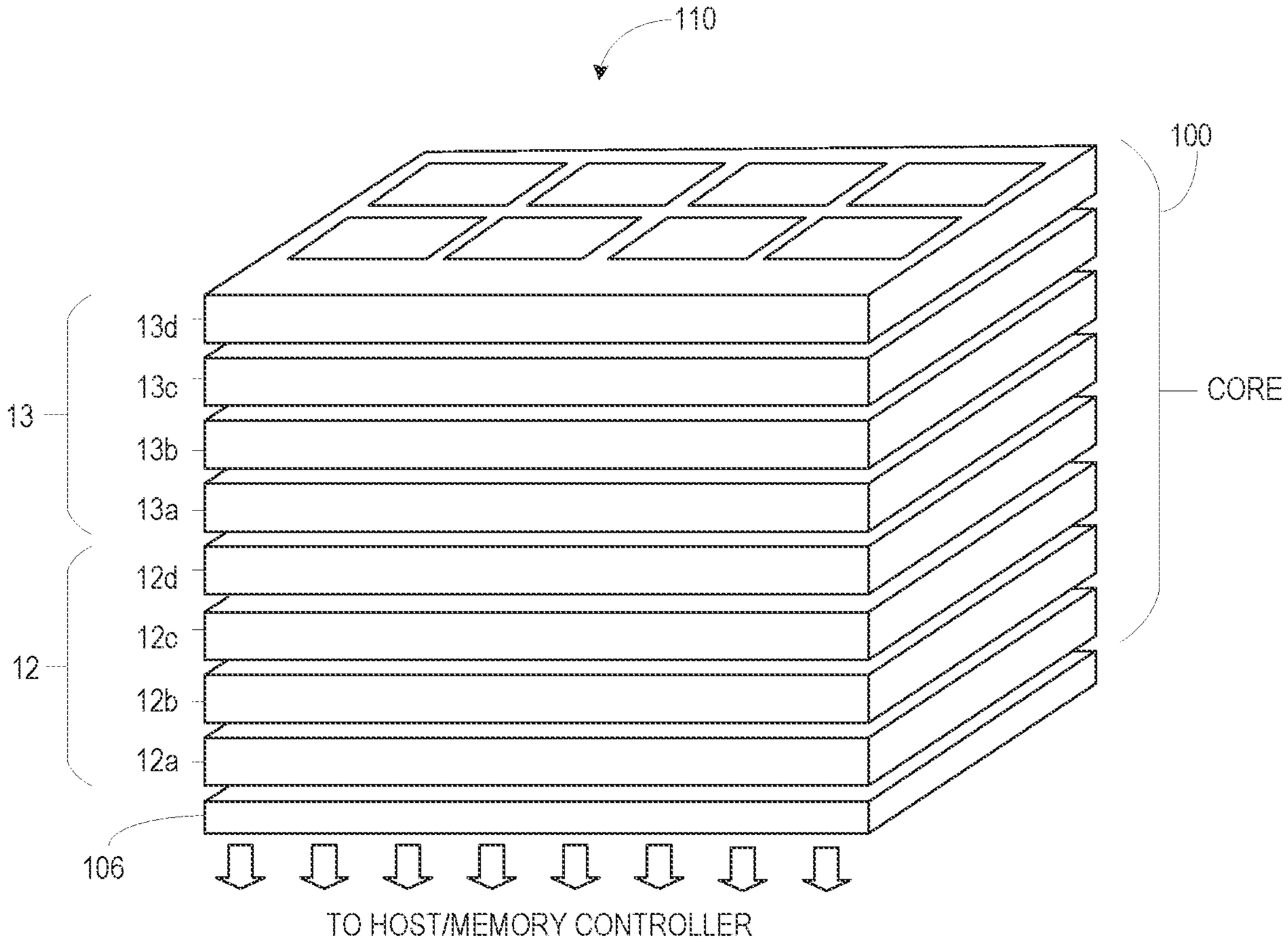
110
100
13d
13c
13b
13a
13
12d
12c
12b
12a
12
CORE
106
TO HOST/MEMORY CONTROLLER

FIG. 1

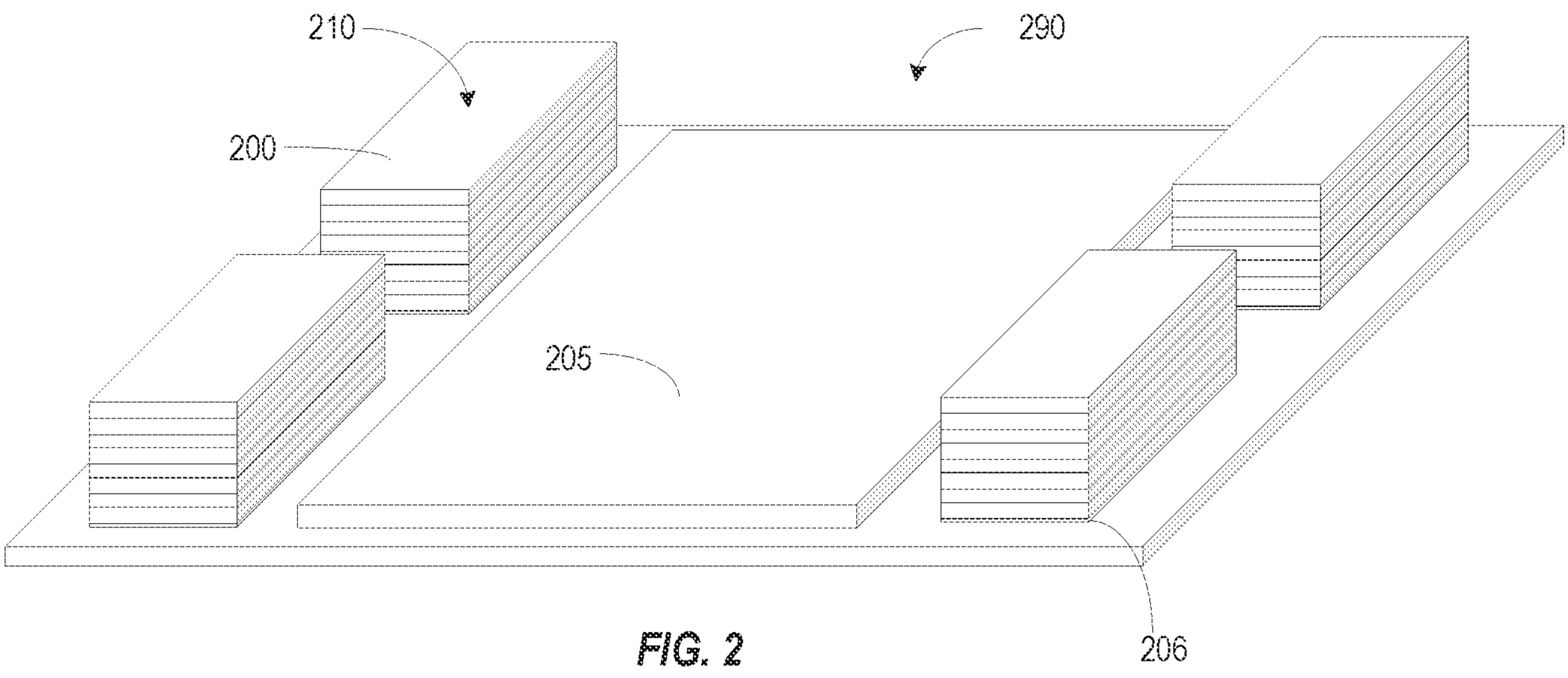

FIG. 2

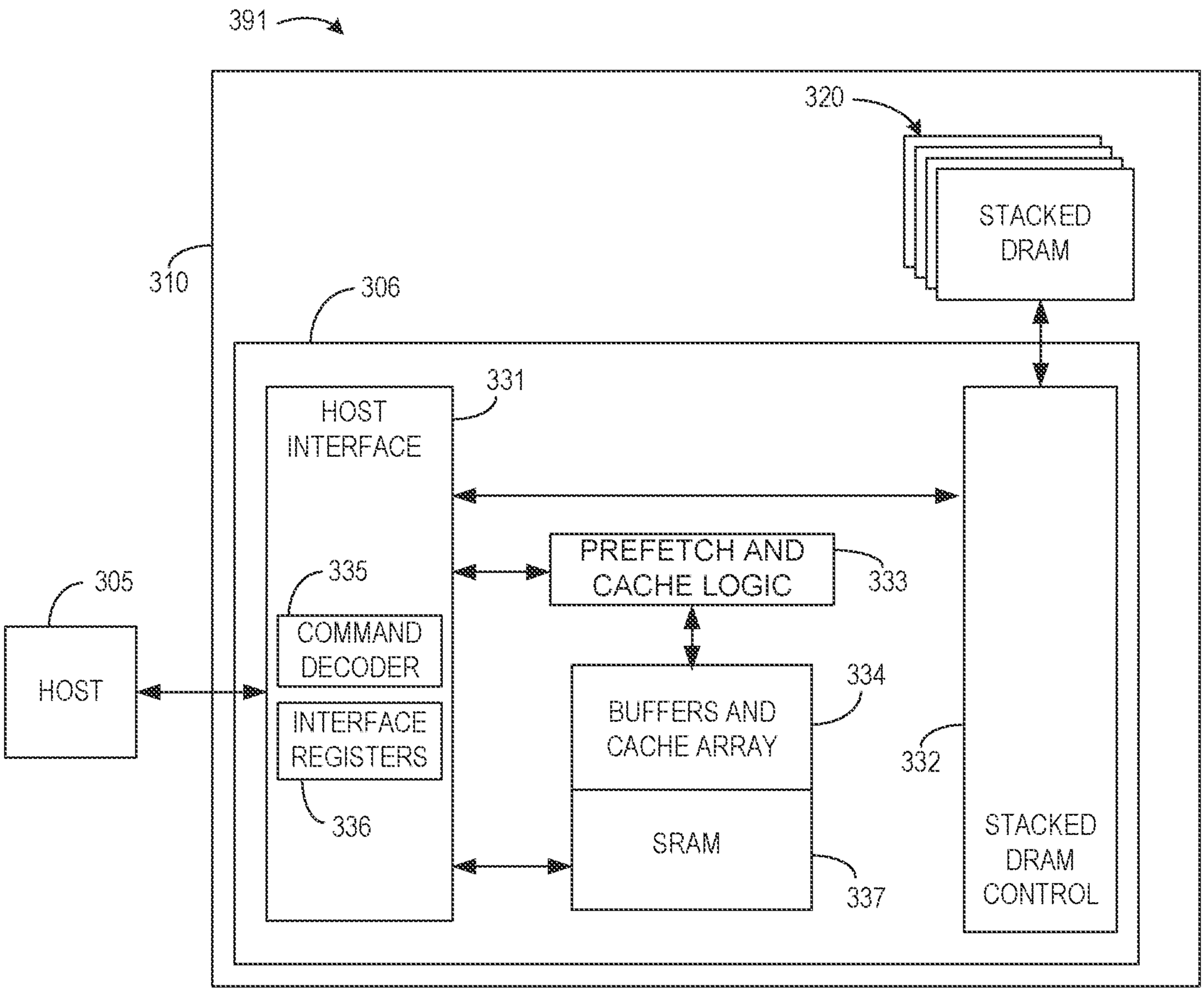
391
320
STACKED DRAM
310
306
331
HOST INTERFACE
305
HOST
335
COMMAND DECODER
INTERFACE REGISTERS
336
PREFETCH AND CACHE LOGIC
333
334
BUFFERS AND CACHE ARRAY
SRAM
337
332
STACKED DRAM CONTROL

FIG. 3

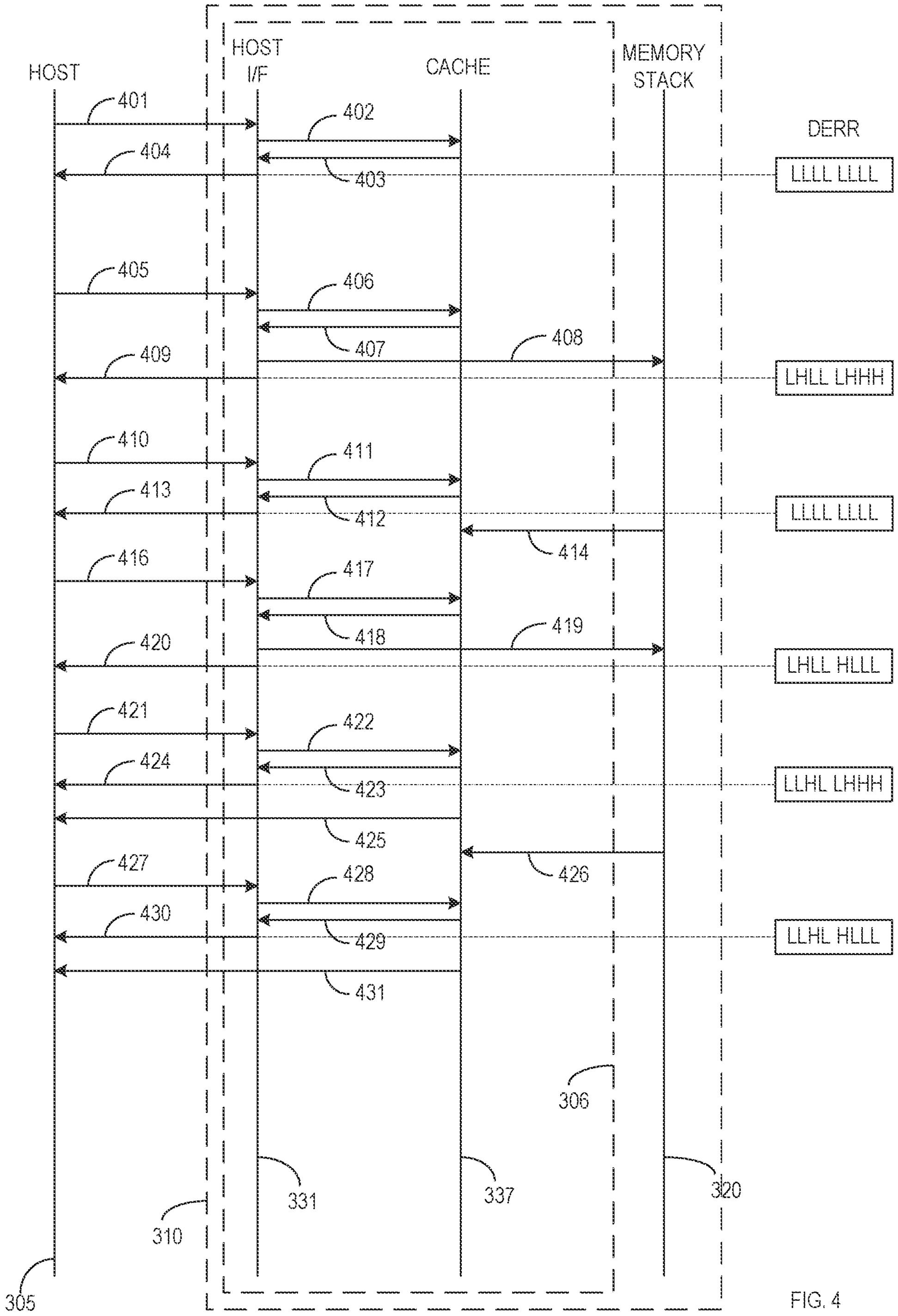
HOST
HOST I/F
CACHE
MEMORY STACK
DERR
401
402
404
403
LLLL LLLL
405
406
407
408
409
LHLL LHHH
410
411
413
412
LLLL LLLL
414
416
417
418
419
420
LHLL HLLL
421
422
424
423
LLHL LHHH
425
426
427
428
430
429
LLHL HLLL
431
306
305
310
331
337
320

FIG. 4

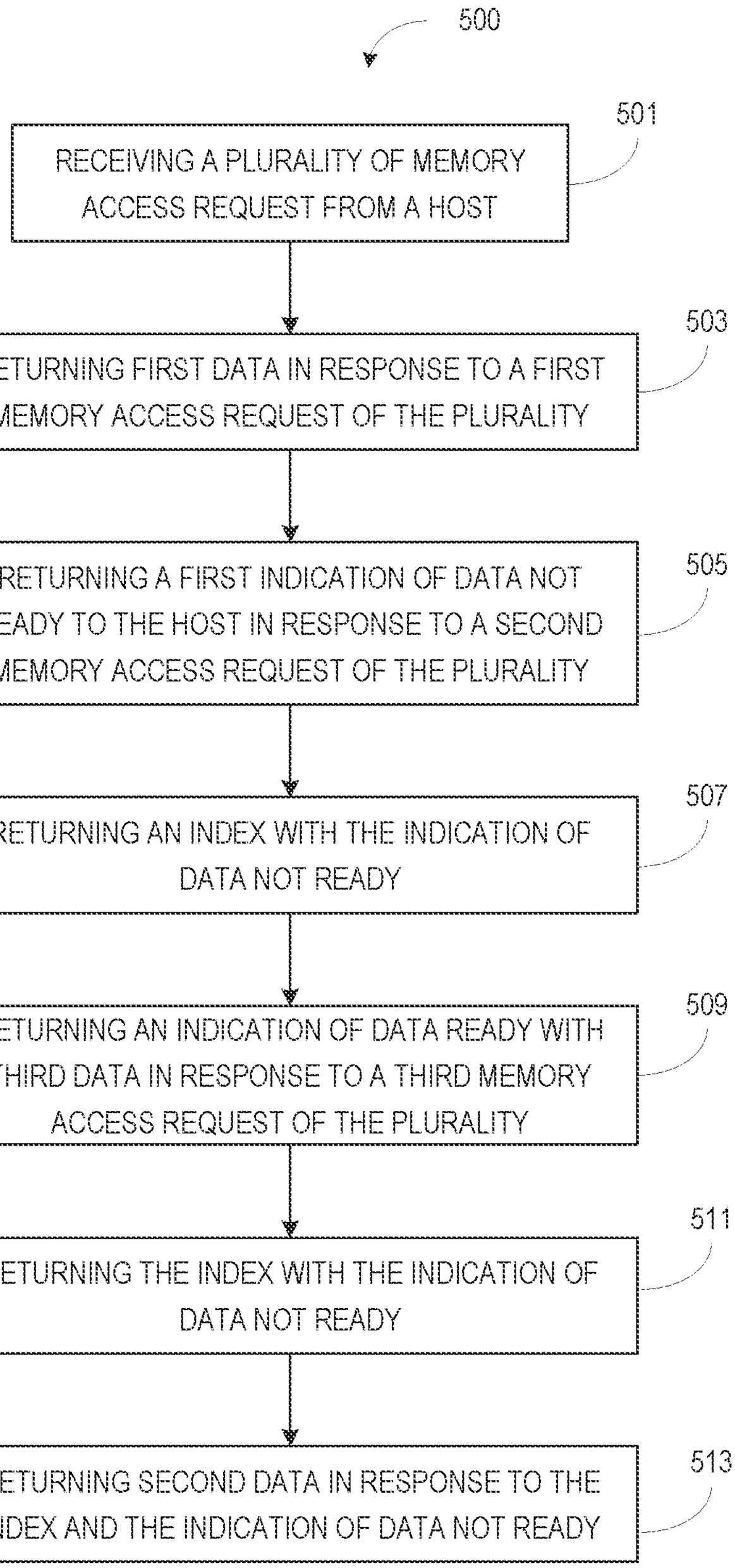
500
501
RECEIVING A PLURALITY OF MEMORY ACCESS REQUEST FROM A HOST
503
RETURNING FIRST DATA IN RESPONSE TO A FIRST MEMORY ACCESS REQUEST OF THE PLURALITY
505
RETURNING A FIRST INDICATION OF DATA NOT READY TO THE HOST IN RESPONSE TO A SECOND MEMORY ACCESS REQUEST OF THE PLURALITY
507
RETURNING AN INDEX WITH THE INDICATION OF DATA NOT READY
509
RETURNING AN INDICATION OF DATA READY WITH THIRD DATA IN RESPONSE TO A THIRD MEMORY ACCESS REQUEST OF THE PLURALITY
511
RETURNING THE INDEX WITH THE INDICATION OF DATA NOT READY
513
RETURNING SECOND DATA IN RESPONSE TO THE INDEX AND THE INDICATION OF DATA NOT READY

FIG. 5

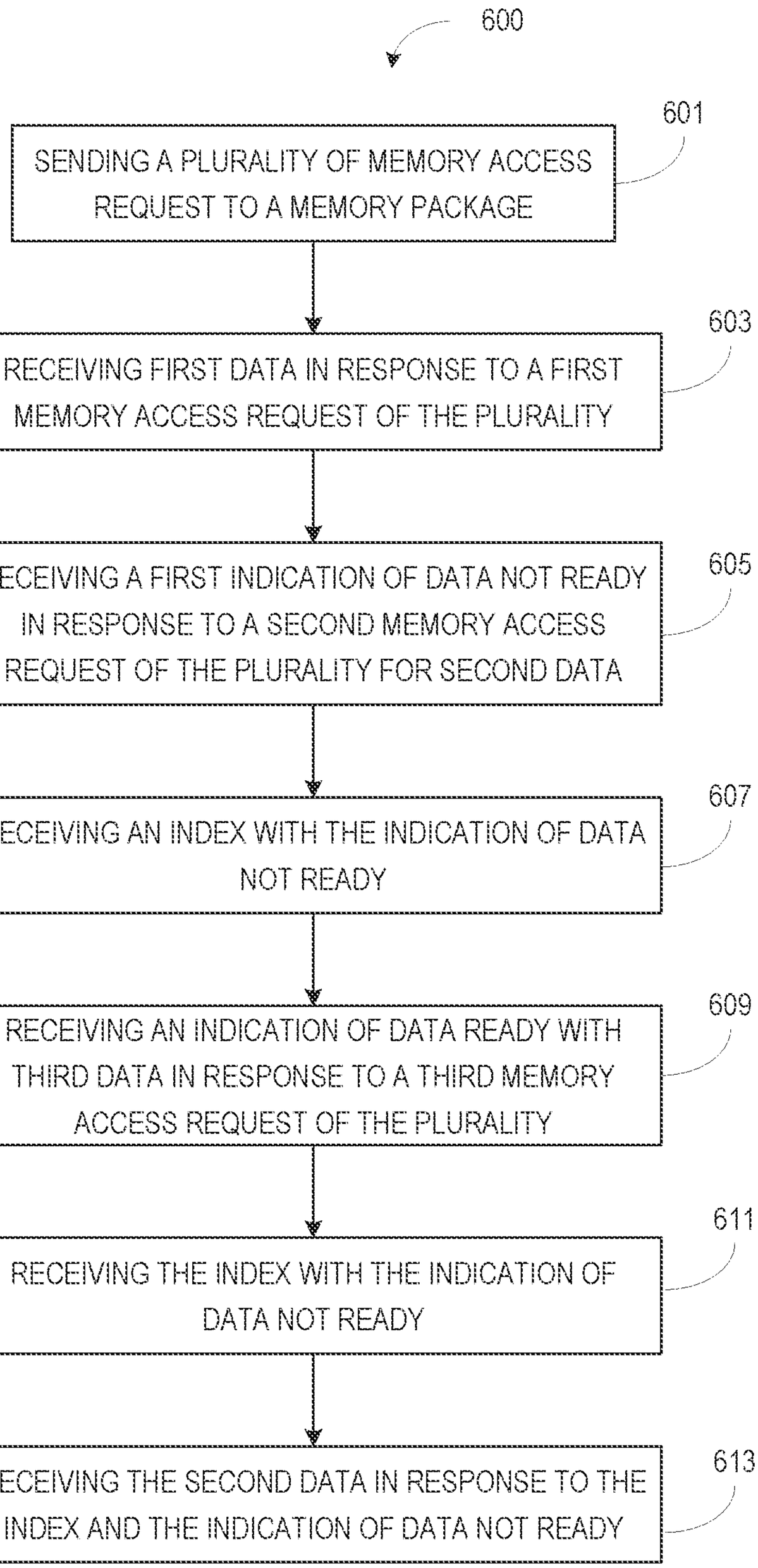
600
601
SENDING A PLURALITY OF MEMORY ACCESS REQUEST TO A MEMORY PACKAGE
603
RECEIVING FIRST DATA IN RESPONSE TO A FIRST MEMORY ACCESS REQUEST OF THE PLURALITY
605
RECEIVING A FIRST INDICATION OF DATA NOT READY IN RESPONSE TO A SECOND MEMORY ACCESS REQUEST OF THE PLURALITY FOR SECOND DATA
607
RECEIVING AN INDEX WITH THE INDICATION OF DATA NOT READY
609
RECEIVING AN INDICATION OF DATA READY WITH THIRD DATA IN RESPONSE TO A THIRD MEMORY ACCESS REQUEST OF THE PLURALITY
611
RECEIVING THE INDEX WITH THE INDICATION OF DATA NOT READY
613
RECEIVING THE SECOND DATA IN RESPONSE TO THE INDEX AND THE INDICATION OF DATA NOT READY

FIG. 6

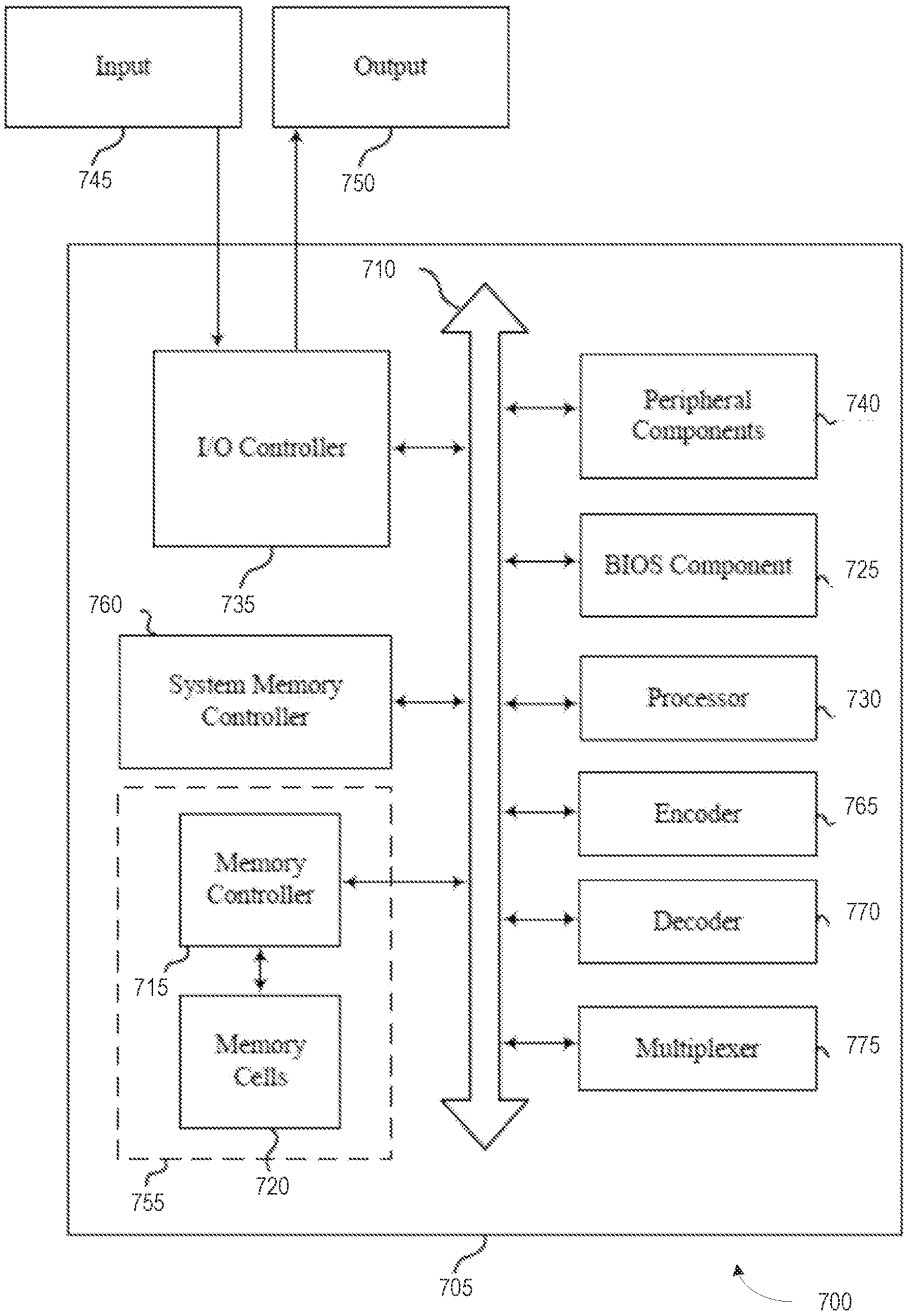
Input
Output
745
750
710
I/O Controller
735
Peripheral Components
740
BIOS Component
725
760
System Memory Controller
Processor
730
Encoder
765
Memory Controller
715
Decoder
770
Memory Cells
Multiplexer
775
755
720
705
700

FIG. 7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2020/066140

A. CLASSIFICATION OF SUBJECT MATTER

G11C 5/06(2006.01)i; **G11C 7/10**(2006.01)i; **G06F 3/06**(2006.01)i; **G11C 7/22**(2006.01)i

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)

G11C 5/06(2006.01); G06F 11/10(2006.01); G06F 12/06(2006.01); G06F 13/16(2006.01); G06F 13/40(2006.01); G11C 5/02(2006.01); G11C 5/04(2006.01); G11C 7/10(2006.01); H01L 23/48(2006.01); H04L 1/18(2006.01)

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

Korean utility models and applications for utility models
Japanese utility models and applications for utility models

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

eKOMPASS(KIPO internal) & keywords: stack, memory die, dymamic random-access memory (DRAM), static random-access memory (SRAM), clock cycle, data error (DERR) indicator, index

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2019-0102330 A1 (MICRON TECHNOLOGY, INC.) 04 April 2019 (2019-04-04) paragraphs [0082]-[0083], [0086], [0247]-[0248], [0273]-[0278]; and figure 6	9,15
A		1-8,10-14,16-20
Y	US 2015-0347226 A1 (MICRON TECHNOLOGY, INC.) 03 December 2015 (2015-12-03) paragraphs [0045], [0051]; and claims 1, 4-5	9,15
A	US 2014-0195715 A1 (MOSAID TECHNOLOGIES INCORPORATED) 10 July 2014 (2014-07-10) paragraphs [0093]-[0094]; and figure 10	1-20
A	US 2018-0189133 A1 (TEXAS INSTRUMENTS INCORPORATED) 05 July 2018 (2018-07-05) claim 9	1-20
A	US 2012-0063190 A1 (YONG-NAM KOH) 15 March 2012 (2012-03-15) paragraphs [0035]-[0039]; and figures 9-10	1-20

☐ 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
"D" document cited by the applicant in the international application
"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	Date of mailing of the international search report
14 April 2021	**15 April 2021**
Name and mailing address of the ISA/KR **Korean Intellectual Property Office** **189 Cheongsa-ro, Seo-gu, Daejeon** **35208, Republic of Korea** Facsimile No. **+82-42-481-8578**	Authorized officer **YANG, JEONG ROK** Telephone No. **+82-42-481-5709**

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/US2020/066140

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
US 2019-0102330 A1	04 April 2019	CN 109599141 A	09 April 2019
		EP 3673483 A1	01 July 2020
		KR 10-2020-0040313 A	17 April 2020
		WO 2019-070459 A1	11 April 2019
US 2015-0347226 A1	03 December 2015	CN 106471485 B	08 January 2019
		CN 109032516 A	18 December 2018
		CN 110262751 A	20 September 2019
		EP 3149602 A1	05 April 2017
		EP 3149602 B1	22 May 2019
		KR 10-1796413 B1	01 December 2017
		KR 10-2017-0005498 A	13 January 2017
		TW 201610687 A	16 March 2016
		TW 201610688 A	16 March 2016
		TW 201614476 A	16 April 2016
		US 10146457 B2	04 December 2018
		US 10540104 B2	21 January 2020
		US 10572164 B2	25 February 2020
		US 10921995 B2	16 February 2021
		US 2015-0347015 A1	03 December 2015
		US 2015-0347019 A1	03 December 2015
		US 2019-0102095 A1	04 April 2019
		US 2020-0097190 A1	26 March 2020
		US 2020-0150884 A1	14 May 2020
		US 9600191 B2	21 March 2017
		US 9747048 B2	29 August 2017
		US 9823864 B2	21 November 2017
		WO 2015-187572 A1	10 December 2015
		WO 2015-187574 A1	10 December 2015
		WO 2015-187578 A1	10 December 2015
US 2014-0195715 A1	10 July 2014	CN 101506895 A	12 August 2009
		CN 101675478 A	17 March 2010
		CN 102760476 A	31 October 2012
		EP 2118903 A1	18 November 2009
		EP 2121603 A2	25 November 2009
		JP 2013-077375 A	25 April 2013
		JP 2014-063523 A	10 April 2014
		JP 2015-165448 A	17 September 2015
		JP 2016-076291 A	12 May 2016
		JP 5334869 B2	06 November 2013
		JP 5960322 B2	02 August 2016
		JP 6145023 B2	07 June 2017
		KR 10-1492383 B1	12 February 2015
		KR 10-2009-0130093 A	17 December 2009
		KR 10-2012-0110157 A	09 October 2012
		TW 200929245 A	01 July 2009
		TW 201112249 A	01 April 2011
		TW 201316351 A	16 April 2013
		US 2014-0325178 A1	30 October 2014
		US 2015-0009761 A1	08 January 2015
		US 2015-0255167 A1	10 September 2015

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/US2020/066140

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
		US 2017-0322730 A1	09 November 2017
		US 2018-0314424 A1	01 November 2018
		US 2019-0163365 A1	30 May 2019
		US 2019-0189225 A1	20 June 2019
		US 2020-0110535 A1	09 April 2020
		US 2020-0357477 A1	12 November 2020
		US 8619473 B2	31 December 2013
		US 8626958 B2	07 January 2014
		US 9245640 B2	26 January 2016
		US 9384847 B2	05 July 2016
		US 9928918 B2	27 March 2018
		US 9971518 B2	15 May 2018
US 2018-0189133 A1	05 July 2018	CN 110352407 A	18 October 2019
		EP 3566139 A1	13 November 2019
		JP 2020-514869 A	21 May 2020
		US 10372531 B2	06 August 2019
		US 10838808 B2	17 November 2020
		US 2019-0317855 A1	17 October 2019
		WO 2018-129246 A1	12 July 2018
US 2012-0063190 A1	15 March 2012	KR 10-2012-0028484 A	23 March 2012
		US 8611123 B2	17 December 2013