



- (51) **International Patent Classification:**
G06F 12/02 (2006.01) *G06F 9/455* (2006.01)
- (21) **International Application Number:**
PCT/CN2014/094663
- (22) **International Filing Date:**
23 December 2014 (23.12.2014)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (71) **Applicant:** INTEL CORPORATION [US/US]; 2200 Mission College Boulevard, M/S: RNB-4-150, Santa Clara, California 95054 (US).
- (72) **Inventors; and**
- (71) **Applicants (for BZ only):** DONG, Yao Zu [CN/CN]; Apt. 10I (1009), Building #5, Yanping Road 123, Shanghai, 31, 200042 (CN). TIAN, Kun [CN/CN]; No 880 Zi Xing Road, Shanghai Zizhu Science Park, Shanghai, 31, 200241 (CN).
- (74) **Agent:** NTD PATENT AND TRADEMARK AGENCY LIMITED; 10th Floor, Block A, Investment Plaza, 27 Jinrongdajie, Xicheng District, Beijing 100033 (CN).
- (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, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, 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, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.
- (84) **Designated States** (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).
- Published:**
— with international search report (Art. 21(3))

(54) **Title:** APPARATUS AND METHOD FOR MANAGING VIRTUAL GRAPHICS PROCESSOR UNIT

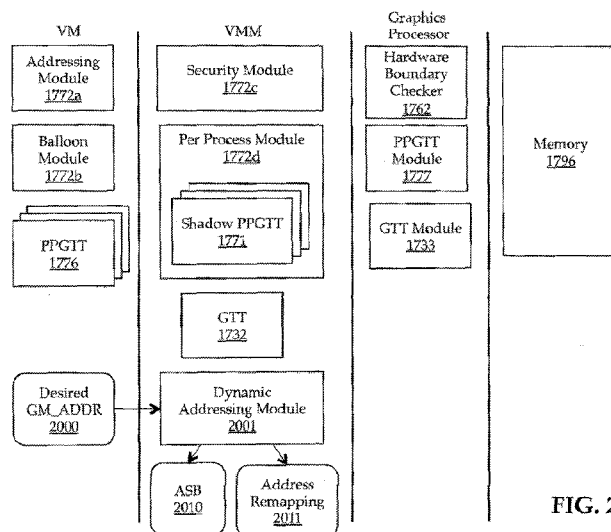


FIG. 20

(57) **Abstract:** An apparatus and method are described for managing a virtual graphics processor unit (GPU). For example, one embodiment of an apparatus comprises: a dynamic addressing module to map portions of an address space required by the virtual machine to matching free address spaces of a host if such matching free address spaces are available, and to select non-matching address spaces for those portions of the address space required by the virtual machine which cannot be matched with free address spaces of the host; and a balloon module to perform address space ballooning (ASB) techniques for those portions of the address space required by the virtual machine which have been mapped to matching address spaces of the host; and address remapping logic to perform address remapping techniques for those portions of the address space required by the virtual machine which have not been mapped to matching address spaces of the host.

APPARATUS AND METHOD FOR MANAGING
VIRTUAL GRAPHICS PROCESSOR UNIT

BACKGROUND

5 Field of the Invention

[0001] This invention relates generally to the field of computer processors. More particularly, the invention relates to a method and apparatus for managing a virtual graphics processor unit (GPU).

Description of the Related Art

10 **[0002]** Graphics processor unit (GPU) virtualization has become increasingly important as cloud and data centers share graphics (GFX) hardware among multiple virtual machines (VMs), such as with VDI (Virtual Desktop Infrastructure) use cases. Technology directed at GFX virtualization includes GRID[™] platforms developed by NVidia[™] and virtual Graphics Technology (vGT)[™] platforms developed by Intel[™]. Graphics virtualization has become
15 important in mobile devices as well, such as implemented in the virtualization-based Dual-OS project.

[0003] There are typically two methods to support efficient GFX virtualization, referred to as “pass-through” and “mediate pass-through.” Pass-through technologies are widely used with input/output memory management unit (IOMMU) technology such as Intel Virtualization
20 Technology for Directed I/O (VT-d) and/or Single Root I/O Virtualization (SRIOV), where the hardware can generate multiple instances of PCI express (PCIe) functions so that each guest can directly access the runtime hardware resources (e.g., memory mapped IO (MMIO) and direct memory access (DMA)). For example, IOMMU technology will remap the guest physical DMA address to the host. However, SRIOV-capable GFX devices require extensive hardware
25 modification and/or circuit budget and, as a result, examples of this technology have not appeared in any significant devices or platforms.

[0004] Mediate pass-through tweaks the solution a bit, by using the software to emulate the access of less performance-critical resources (e.g., MMIO registers) from the guest, but to pass-through performance-critical resource access, such as aperture and rendering commands.
30 For example, guest aperture access may go directly to hardware, and the guest rendering commands may be directly executed by the GPU.

BRIEF DESCRIPTION OF THE DRAWINGS

[0005] A better understanding of the present invention can be obtained from the following detailed description in conjunction with the following drawings, in which:

35 **[0006]** **FIG. 1** is a block diagram of an embodiment of a computer system with a processor

having one or more processor cores and graphics processors;

[0007] FIG. 2 is a block diagram of one embodiment of a processor having one or more processor cores, an integrated memory controller, and an integrated graphics processor;

[0008] FIG. 3 is a block diagram of one embodiment of a graphics processor which may be a discrete graphics processing unit, or may be graphics processor integrated with a plurality of processing cores;

[0009] FIG. 4 is a block diagram of an embodiment of a graphics-processing engine for a graphics processor;

[0010] FIG. 5 is a block diagram of another embodiment of a graphics processor;

[0011] FIG. 6 is a block diagram of thread execution logic including an array of processing elements;

[0012] FIG. 7 illustrates a graphics processor execution unit instruction format according to an embodiment;

[0013] FIG. 8 is a block diagram of another embodiment of a graphics processor which includes a graphics pipeline, a media pipeline, a display engine, thread execution logic, and a render output pipeline;

[0014] FIG. 9A is a block diagram illustrating a graphics processor command format according to an embodiment;

[0015] FIG. 9B is a block diagram illustrating a graphics processor command sequence according to an embodiment;

[0016] FIG. 10 illustrates exemplary graphics software architecture for a data processing system according to an embodiment;

[0017] FIGS. 11A-B are block diagrams of examples of graphics virtualization environments according to an embodiment;

[0018] FIG. 12 is a block diagram of an example of an address ballooning approach according to an embodiment;

[0019] FIG. 13 is a flowchart of an example of a method of managing virtual machines according to an embodiment;

[0020] FIG. 14 is a block diagram of an example of an architecture having a hardware boundary checker according to an embodiment;

[0021] FIGS. 15 and 16 are block diagrams of an example of a per process graphics virtualization approach according to an embodiment;

[0022] FIG. 17 is a block diagram of an example of a logic architecture according to an embodiment;

[0023] FIG. 18 is a block diagram of an example of a processor according to an embodiment;

[0024] FIG. 19 is a block diagram of an example of a system according to an embodiment;

[0025] FIG. 20 illustrates one embodiment of an architecture including a dynamic addressing module;

[0026] FIG. 21 illustrates a mapping in which an address space required by a virtual machine is mapped to matching free address spaces of a host;

[0027] FIG. 22 illustrates a mapping in which portions of an address space required by the virtual machine are mapped to non-matching free address spaces of a host;

[0028] FIG. 23 illustrates a mapping in which portions of an address space required by the virtual machine are mapped to matching free address spaces of a host; and

[0029] FIG. 24 illustrates a method in accordance with one embodiment of the invention.

DETAILED DESCRIPTION

[0030] In the following description, for the purposes of explanation, numerous specific details are set forth in order to provide a thorough understanding of the embodiments of the invention described below. It will be apparent, however, to one skilled in the art that the embodiments of the invention may be practiced without some of these specific details. In other instances, well-known structures and devices are shown in block diagram form to avoid obscuring the underlying principles of the embodiments of the invention.

EXEMPLARY GRAPHICS PROCESSOR ARCHITECTURES AND DATA TYPES

[0031] **System Overview**

[0032] **Figure 1** is a block diagram of a data processing system 100, according to an embodiment. Data processing system 100 includes one or more processors 102 and one or more graphics processors 108, and may be a single processor desktop system, a multiprocessor workstation system, or a server system having a large number of processors 102 or processor cores 107. In one embodiment, the data processing system 100 is a system-on-a-chip (SoC) integrated circuit for use in mobile, handheld, or embedded devices.

[0033] An embodiment of data processing system 100 can include, or be incorporated within a server-based gaming platform, a game console, including a game and media console, a mobile gaming console, a handheld game console, or an online game console. In some embodiments, data processing system 100 is a mobile phone, smart phone, tablet computing device or mobile Internet device. Data processing system 100 can also include, couple with, or be integrated within a wearable device, such as a smart watch wearable device, smart eyewear device, augmented reality device, or virtual reality device. In some embodiments, data processing system 100 is a television or set top box device having one or more processors 102 and a graphical interface generated by one or more graphics processors 108.

[0034] In some embodiments, the one or more processors 102 each include one or more processor cores 107 to process instructions which, when executed, perform operations for system and user software. In some embodiments, each of the one or more processor cores 107 is configured to process a specific instruction set 109. In some embodiments, instruction set 109 may facilitate Complex Instruction Set Computing (CISC), Reduced Instruction Set Computing (RISC), or computing via a Very Long Instruction Word (VLIW). Multiple processor cores 107 may each process a different instruction set 109, which may include instructions to facilitate the emulation of other instruction sets. Processor core 107 may also include other processing devices, such as a Digital Signal Processor (DSP).

[0035] In some embodiments, the processor 102 includes cache memory 104. Depending on the architecture, the processor 102 can have a single internal cache or multiple levels of internal cache. In some embodiments, the cache memory is shared among various components

of the processor 102. In some embodiments, the processor 102 also uses an external cache (e.g., a Level-3 (L3) cache or Last Level Cache (LLC)) (not shown), which may be shared among processor cores 107 using known cache coherency techniques. A register file 106 is additionally included in processor 102 which may include different types of registers for storing different types of data (e.g., integer registers, floating point registers, status registers, and an instruction pointer register). Some registers may be general-purpose registers, while other registers may be specific to the design of the processor 102.

[0036] In some embodiments, processor 102 is coupled to a processor bus 110 to transmit data signals between processor 102 and other components in system 100. System 100 uses an exemplary ‘hub’ system architecture, including a memory controller hub 116 and an input output (I/O) controller hub 130. Memory controller hub 116 facilitates communication between a memory device and other components of system 100, while I/O Controller Hub (ICH) 130 provides connections to I/O devices via a local I/O bus.

[0037] Memory device 120 can be a dynamic random access memory (DRAM) device, a static random access memory (SRAM) device, flash memory device, or some other memory device having suitable performance to serve as process memory. Memory 120 can store data 122 and instructions 121 for use when processor 102 executes a process. Memory controller hub 116 also couples with an optional external graphics processor 112, which may communicate with the one or more graphics processors 108 in processors 102 to perform graphics and media operations.

[0038] In some embodiments, ICH 130 enables peripherals to connect to memory 120 and processor 102 via a high-speed I/O bus. The I/O peripherals include an audio controller 146, a firmware interface 128, a wireless transceiver 126 (e.g., Wi-Fi, Bluetooth), a data storage device 124 (e.g., hard disk drive, flash memory, etc.), and a legacy I/O controller for coupling legacy (e.g., Personal System 2 (PS/2)) devices to the system. One or more Universal Serial Bus (USB) controllers 142 connect input devices, such as keyboard and mouse 144 combinations. A network controller 134 may also couple to ICH 130. In some embodiments, a high-performance network controller (not shown) couples to processor bus 110.

[0039] **Figure 2** is a block diagram of an embodiment of a processor 200 having one or more processor cores 202A-N, an integrated memory controller 214, and an integrated graphics processor 208. Those elements of **Figure 2** having the same reference numbers (or names) as the elements of any other figure herein can operate or function in any manner similar to that described elsewhere herein, but are not limited to such. Processor 200 can include additional cores up to and including additional core 202N represented by the dashed lined boxes. Each of

cores 202A-N includes one or more internal cache units 204A-N. In some embodiments each core also has access to one or more shared cache units 206.

[0040] The internal cache units 204A-N and shared cache units 206 represent a cache memory hierarchy within the processor 200. The cache memory hierarchy may include at least one level of instruction and data cache within each core and one or more levels of shared mid-level cache, such as a Level 2 (L2), Level 3 (L3), Level 4 (L4), or other levels of cache, where the highest level of cache before external memory is classified as the LLC. In some embodiments, cache coherency logic maintains coherency between the various cache units 206 and 204A-N.

[0041] In some embodiments, processor 200 may also include a set of one or more bus controller units 216 and a system agent 210. The one or more bus controller units manage a set of peripheral buses, such as one or more Peripheral Component Interconnect buses (e.g., PCI, PCI Express). System agent 210 provides management functionality for the various processor components. In some embodiments, system agent 210 includes one or more integrated memory controllers 214 to manage access to various external memory devices (not shown).

[0042] In some embodiments, one or more of the cores 202A-N include support for simultaneous multi-threading. In such embodiment, the system agent 210 includes components for coordinating and operating cores 202A-N during multi-threaded processing. System agent 210 may additionally include a power control unit (PCU), which includes logic and components to regulate the power state of cores 202A-N and graphics processor 208.

[0043] In some embodiments, processor 200 additionally includes graphics processor 208 to execute graphics processing operations. In some embodiments, the graphics processor 208 couples with the set of shared cache units 206, and the system agent unit 210, including the one or more integrated memory controllers 214. In some embodiments, a display controller 211 is coupled with the graphics processor 208 to drive graphics processor output to one or more coupled displays. In some embodiments, display controller 211 may be separate module coupled with the graphics processor via at least one interconnect, or may be integrated within the graphics processor 208 or system agent 210.

[0044] In some embodiments, a ring based interconnect unit 212 is used to couple the internal components of the processor 200. However, an alternative interconnect unit may be used, such as a point-to-point interconnect, a switched interconnect, or other techniques, including techniques well known in the art. In some embodiments, graphics processor 208 couples with the ring interconnect 212 via an I/O link 213.

[0045] The exemplary I/O link 213 represents at least one of multiple varieties of I/O interconnects, including an on package I/O interconnect which facilitates communication

between various processor components and a high-performance embedded memory module 218, such as an eDRAM module. In some embodiments, each of the cores 202-N and graphics processor 208 use embedded memory modules 218 as a shared Last Level Cache.

[0046] In some embodiments, cores 202A-N are homogenous cores executing the same instruction set architecture. In another embodiment, cores 202A-N are heterogeneous in terms of instruction set architecture (ISA), where one or more of cores 202A-N execute a first instruction set, while at least one of the other cores executes a subset of the first instruction set or a different instruction set.

[0047] In some embodiments, processor 200 is a part of, or implemented on, one or more substrates using any of a number of process technologies, for example, Complementary metal-oxide-semiconductor (CMOS), Bipolar Junction/Complementary metal-oxide-semiconductor (BiCMOS) or N-type metal-oxide-semiconductor logic (NMOS). Additionally, processor 200 can be implemented on one or more chips or as an SoC integrated circuit having the illustrated components, in addition to other components.

[0048] **Figure 3** is a block diagram of a graphics processor 300, which may be a discrete graphics processing unit, or may be a graphics processor integrated with a plurality of processing cores. In some embodiments, the graphics processor communicates via a memory mapped I/O interface to registers on the graphics processor and with commands placed into the processor memory. In some embodiments, graphics processor 300 includes a memory interface 314 to access memory. Memory interface 314 can be an interface to local memory, one or more internal caches, one or more shared external caches, and/or to system memory.

[0049] In some embodiments, graphics processor 300 also includes a display controller 302 to drive display output data to a display device 320. Display controller 302 includes hardware for one or more overlay planes for the display and composition of multiple layers of video or user interface elements. In some embodiments, graphics processor 300 includes a video codec engine 306 to encode, decode, or transcode media to, from, or between one or more media encoding formats, including, but not limited to Moving Picture Experts Group (MPEG) formats such as MPEG-2, Advanced Video Coding (AVC) formats such as H.264/MPEG-4 AVC, as well as the Society of Motion Picture & Television Engineers (SMPTE) 421M/VC-1, and Joint Photographic Experts Group (JPEG) formats such as JPEG, and Motion JPEG (MJPEG) formats.

[0050] In some embodiments, graphics processor 300 includes a block image transfer (BLIT) engine 304 to perform two-dimensional (2D) rasterizer operations including, for example, bit-boundary block transfers. However, in one embodiment, 2D graphics operations are performed using one or more components of the graphics-processing engine (GPE) 310. In some

embodiments, graphics-processing engine 310 is a compute engine for performing graphics operations, including three-dimensional (3D) graphics operations and media operations.

[0051] In some embodiments, GPE 310 includes a 3D pipeline 312 for performing 3D operations, such as rendering three-dimensional images and scenes using processing functions that act upon 3D primitive shapes (e.g., rectangle, triangle, etc.). The 3D pipeline 312 includes programmable and fixed function elements that perform various tasks within the element and/or spawn execution threads to a 3D/Media sub-system 315. While 3D pipeline 312 can be used to perform media operations, an embodiment of GPE 310 also includes a media pipeline 316 that is specifically used to perform media operations, such as video post-processing and image enhancement.

[0052] In some embodiments, media pipeline 316 includes fixed function or programmable logic units to perform one or more specialized media operations, such as video decode acceleration, video de-interlacing, and video encode acceleration in place of, or on behalf of video codec engine 306. In some embodiments, media pipeline 316 additionally includes a thread spawning unit to spawn threads for execution on 3D/Media sub-system 315. The spawned threads perform computations for the media operations on one or more graphics execution units included in 3D/Media sub-system 315.

[0053] In some embodiments, 3D/Media subsystem 315 includes logic for executing threads spawned by 3D pipeline 312 and media pipeline 316. In one embodiment, the pipelines send thread execution requests to 3D/Media subsystem 315, which includes thread dispatch logic for arbitrating and dispatching the various requests to available thread execution resources. The execution resources include an array of graphics execution units to process the 3D and media threads. In some embodiments, 3D/Media subsystem 315 includes one or more internal caches for thread instructions and data. In some embodiments, the subsystem also includes shared memory, including registers and addressable memory, to share data between threads and to store output data.

[0054] **3D/Media Processing**

[0055] **Figure 4** is a block diagram of a graphics processing engine 410 of a graphics processor in accordance with some embodiments. In one embodiment, the GPE 410 is a version of the GPE 310 shown in **Figure 3**. Elements of **Figure 4** having the same reference numbers (or names) as the elements of any other figure herein can operate or function in any manner similar to that described elsewhere herein, but are not limited to such.

[0056] In some embodiments, GPE 410 couples with a command streamer 403, which provides a command stream to the GPE 3D and media pipelines 412, 416. In some embodiments, command streamer 403 is coupled to memory, which can be system memory, or one or more of

internal cache memory and shared cache memory. In some embodiments, command streamer 403 receives commands from the memory and sends the commands to 3D pipeline 412 and/or media pipeline 416. The 3D and media pipelines process the commands by performing operations via logic within the respective pipelines or by dispatching one or more execution threads to an execution unit array 414. In some embodiments, execution unit array 414 is scalable, such that the array includes a variable number of execution units based on the target power and performance level of GPE 410.

[0057] In some embodiments, a sampling engine 430 couples with memory (e.g., cache memory or system memory) and execution unit array 414. In some embodiments, sampling engine 430 provides a memory access mechanism for execution unit array 414 that allows execution array 414 to read graphics and media data from memory. In some embodiments, sampling engine 430 includes logic to perform specialized image sampling operations for media.

[0058] In some embodiments, the specialized media sampling logic in sampling engine 430 includes a de-noise/de-interlace module 432, a motion estimation module 434, and an image scaling and filtering module 436. In some embodiments, de-noise/de-interlace module 432 includes logic to perform one or more of a de-noise or a de-interlace algorithm on decoded video data. The de-interlace logic combines alternating fields of interlaced video content into a single frame of video. The de-noise logic reduces or removes data noise from video and image data. In some embodiments, the de-noise logic and de-interlace logic are motion adaptive and use spatial or temporal filtering based on the amount of motion detected in the video data. In some embodiments, the de-noise/de-interlace module 432 includes dedicated motion detection logic (e.g., within the motion estimation engine 434).

[0059] In some embodiments, motion estimation engine 434 provides hardware acceleration for video operations by performing video acceleration functions such as motion vector estimation and prediction on video data. The motion estimation engine determines motion vectors that describe the transformation of image data between successive video frames. In some embodiments, a graphics processor media codec uses video motion estimation engine 434 to perform operations on video at the macro-block level that may otherwise be too computationally intensive to perform with a general-purpose processor. In some embodiments, motion estimation engine 434 is generally available to graphics processor components to assist with video decode and processing functions that are sensitive or adaptive to the direction or magnitude of the motion within video data.

[0060] In some embodiments, image scaling and filtering module 436 performs image-processing operations to enhance the visual quality of generated images and video. In some

embodiments, scaling and filtering module 436 processes image and video data during the sampling operation before providing the data to execution unit array 414.

[0061] In some embodiments, the GPE 410 includes a data port 444, which provides an additional mechanism for graphics subsystems to access memory. In some embodiments, data port 444 facilitates memory access for operations including render target writes, constant buffer reads, scratch memory space reads/writes, and media surface accesses. In some embodiments, data port 444 includes cache memory space to cache accesses to memory. The cache memory can be a single data cache or separated into multiple caches for the multiple subsystems that access memory via the data port (e.g., a render buffer cache, a constant buffer cache, etc.). In some embodiments, threads executing on an execution unit in execution unit array 414 communicate with the data port by exchanging messages via a data distribution interconnect that couples each of the sub-systems of GPE 410.

[0062] **Execution Units**

[0063] **Figure 5** is a block diagram of another embodiment of a graphics processor 500.

Elements of **Figure 5** having the same reference numbers (or names) as the elements of any other figure herein can operate or function in any manner similar to that described elsewhere herein, but are not limited to such.

[0064] In some embodiments, graphics processor 500 includes a ring interconnect 502, a pipeline front-end 504, a media engine 537, and graphics cores 580A-N. In some embodiments, ring interconnect 502 couples the graphics processor to other processing units, including other graphics processors or one or more general-purpose processor cores. In some embodiments, the graphics processor is one of many processors integrated within a multi-core processing system.

[0065] In some embodiments, graphics processor 500 receives batches of commands via ring interconnect 502. The incoming commands are interpreted by a command streamer 503 in the pipeline front-end 504. In some embodiments, graphics processor 500 includes scalable execution logic to perform 3D geometry processing and media processing via the graphics core(s) 580A-N. For 3D geometry processing commands, command streamer 503 supplies commands to geometry pipeline 536. For at least some media processing commands, command streamer 503 supplies the commands to a video front end 534, which couples with a media engine 537. In some embodiments, media engine 537 includes a Video Quality Engine (VQE) 530 for video and image post-processing and a multi-format encode/decode (MFX) 533 engine to provide hardware-accelerated media data encode and decode. In some embodiments, geometry pipeline 536 and media engine 537 each generate execution threads for the thread execution resources provided by at least one graphics core 580A.

[0066] In some embodiments, graphics processor 500 includes scalable thread execution resources featuring modular cores 580A-N (sometimes referred to as core slices), each having multiple sub-cores 550A-N, 560A-N (sometimes referred to as core sub-slices). In some embodiments, graphics processor 500 can have any number of graphics cores 580A through 580N. In some embodiments, graphics processor 500 includes a graphics core 580A having at least a first sub-core 550A and a second core sub-core 560A. In other embodiments, the graphics processor is a low power processor with a single sub-core (e.g., 550A). In some embodiments, graphics processor 500 includes multiple graphics cores 580A-N, each including a set of first sub-cores 550A-N and a set of second sub-cores 560A-N. Each sub-core in the set of first sub-cores 550A-N includes at least a first set of execution units 552A-N and media/textures samplers 554A-N. Each sub-core in the set of second sub-cores 560A-N includes at least a second set of execution units 562A-N and samplers 564A-N. In some embodiments, each sub-core 550A-N, 560A-N shares a set of shared resources 570A-N. In some embodiments, the shared resources include shared cache memory and pixel operation logic. Other shared resources may also be included in the various embodiments of the graphics processor.

[0067] **Figure 6** illustrates thread execution logic 600 including an array of processing elements employed in some embodiments of a GPE. Elements of **Figure 6** having the same reference numbers (or names) as the elements of any other figure herein can operate or function in any manner similar to that described elsewhere herein, but are not limited to such.

[0068] In some embodiments, thread execution logic 600 includes a pixel shader 602, a thread dispatcher 604, instruction cache 606, a scalable execution unit array including a plurality of execution units 608A-N, a sampler 610, a data cache 612, and a data port 614. In one embodiment the included components are interconnected via an interconnect fabric that links to each of the components. In some embodiments, thread execution logic 600 includes one or more connections to memory, such as system memory or cache memory, through one or more of instruction cache 606, data port 614, sampler 610, and execution unit array 608A-N. In some embodiments, each execution unit (e.g. 608A) is an individual vector processor capable of executing multiple simultaneous threads and processing multiple data elements in parallel for each thread. In some embodiments, execution unit array 608A-N includes any number individual execution units.

[0069] In some embodiments, execution unit array 608A-N is primarily used to execute “shader” programs. In some embodiments, the execution units in array 608A-N execute an instruction set that includes native support for many standard 3D graphics shader instructions, such that shader programs from graphics libraries (e.g., Direct 3D and OpenGL) are executed with a minimal translation. The execution units support vertex and geometry processing (e.g.,

vertex programs, geometry programs, vertex shaders), pixel processing (e.g., pixel shaders, fragment shaders) and general-purpose processing (e.g., compute and media shaders).

[0070] Each execution unit in execution unit array 608A-N operates on arrays of data elements. The number of data elements is the “execution size,” or the number of channels for the instruction. An execution channel is a logical unit of execution for data element access, masking, and flow control within instructions. The number of channels may be independent of the number of physical Arithmetic Logic Units (ALUs) or Floating Point Units (FPUs) for a particular graphics processor. In some embodiments, execution units 608A-N support integer and floating-point data types.

[0071] The execution unit instruction set includes single instruction multiple data (SIMD) instructions. The various data elements can be stored as a packed data type in a register and the execution unit will process the various elements based on the data size of the elements. For example, when operating on a 256-bit wide vector, the 256 bits of the vector are stored in a register and the execution unit operates on the vector as four separate 64-bit packed data elements (Quad-Word (QW) size data elements), eight separate 32-bit packed data elements (Double Word (DW) size data elements), sixteen separate 16-bit packed data elements (Word (W) size data elements), or thirty-two separate 8-bit data elements (byte (B) size data elements). However, different vector widths and register sizes are possible.

[0072] One or more internal instruction caches (e.g., 606) are included in the thread execution logic 600 to cache thread instructions for the execution units. In some embodiments, one or more data caches (e.g., 612) are included to cache thread data during thread execution. In some embodiments, sampler 610 is included to provide texture sampling for 3D operations and media sampling for media operations. In some embodiments, sampler 610 includes specialized texture or media sampling functionality to process texture or media data during the sampling process before providing the sampled data to an execution unit.

[0073] During execution, the graphics and media pipelines send thread initiation requests to thread execution logic 600 via thread spawning and dispatch logic. In some embodiments, thread execution logic 600 includes a local thread dispatcher 604 that arbitrates thread initiation requests from the graphics and media pipelines and instantiates the requested threads on one or more execution units 608A-N. For example, the geometry pipeline (e.g., 536 of **Figure 5**) dispatches vertex processing, tessellation, or geometry processing threads to thread execution logic 600 (**Figure 6**). In some embodiments, thread dispatcher 604 can also process runtime thread spawning requests from the executing shader programs.

[0074] Once a group of geometric objects has been processed and rasterized into pixel data, pixel shader 602 is invoked to further compute output information and cause results to be

written to output surfaces (e.g., color buffers, depth buffers, stencil buffers, etc.). In some embodiments, pixel shader 602 calculates the values of the various vertex attributes that are to be interpolated across the rasterized object. In some embodiments, pixel shader 602 then executes an API-supplied pixel shader program. To execute the pixel shader program, pixel shader 602
5 dispatches threads to an execution unit (e.g., 608A) via thread dispatcher 604. In some embodiments, pixel shader 602 uses texture sampling logic in sampler 610 to access texture data in texture maps stored in memory. Arithmetic operations on the texture data and the input geometry data compute pixel color data for each geometric fragment, or discards one or more pixels from further processing.

10 **[0075]** In some embodiments, the data port 614 provides a memory access mechanism for the thread execution logic 600 output processed data to memory for processing on a graphics processor output pipeline. In some embodiments, the data port 614 includes or couples to one or more cache memories (e.g., data cache 612) to cache data for memory access via the data port.

[0076] **Figure 7** is a block diagram illustrating a graphics processor instruction formats

15 700 according to some embodiments. In one or more embodiment, the graphics processor execution units support an instruction set having instructions in multiple formats. The solid lined boxes illustrate the components that are generally included in an execution unit instruction, while the dashed lines include components that are optional or that are only included in a sub-set of the instructions. In some embodiments, instruction format 700 described and illustrated are macro-
20 instructions, in that they are instructions supplied to the execution unit, as opposed to micro-operations resulting from instruction decode once the instruction is processed.

[0077] In some embodiments, the graphics processor execution units natively support instructions in a 128-bit format 710. A 64-bit compacted instruction format 730 is available for some instructions based on the selected instruction, instruction options, and number of operands.

25 The native 128-bit format 710 provides access to all instruction options, while some options and operations are restricted in the 64-bit format 730. The native instructions available in the 64-bit format 730 vary by embodiment. In some embodiments, the instruction is compacted in part using a set of index values in an index field 713. The execution unit hardware references a set of compaction tables based on the index values and uses the compaction table outputs to reconstruct
30 a native instruction in the 128-bit format 710.

[0078] For each format, instruction opcode 712 defines the operation that the execution unit is to perform. The execution units execute each instruction in parallel across the multiple data elements of each operand. For example, in response to an add instruction the execution unit performs a simultaneous add operation across each color channel representing a texture element
35 or picture element. By default, the execution unit performs each instruction across all data

channels of the operands. In some embodiments, instruction control field 712 enables control over certain execution options, such as channels selection (e.g., predication) and data channel order (e.g., swizzle). For 128-bit instructions 710 an exec-size field 716 limits the number of data channels that will be executed in parallel. In some embodiments, exec-size field 716 is not available for use in the 64-bit compact instruction format 730.

[0079] Some execution unit instructions have up to three operands including two source operands, src0 722, src1 722, and one destination 718. In some embodiments, the execution units support dual destination instructions, where one of the destinations is implied. Data manipulation instructions can have a third source operand (e.g., SRC2 724), where the instruction opcode 712 determines the number of source operands. An instruction's last source operand can be an immediate (e.g., hard-coded) value passed with the instruction.

[0080] In some embodiments, the 128-bit instruction format 710 includes access/address mode information 726 specifying, for example, whether direct register addressing mode or indirect register addressing mode is used. When direct register addressing mode is used, the register address of one or more operands is directly provided by bits in the instruction 710. When indirect register addressing mode is used, the register address of one or more operands may be computed based on field 726, which specifies an address register value mode and/or an address immediate field in the instruction. In one embodiment the access/address mode information 726 may determine the operand mode to define a data access alignment for the instruction. For example, when in a first mode, the instruction 710 may use byte-aligned addressing for source and destination operands and, when in a second mode, the instruction 710 may use 16-byte-aligned access mode and 1-byte aligned addressing for all source and destination operands. access mode, where the byte alignment of the access mode determines the access alignment of the instruction operands.

[0081] In some embodiments instructions are grouped based on opcode bit-fields to simplify Opcode decode 740. For an 8-bit opcode, bits 4, 5, and 6 allow the execution unit to determine the type of opcode. The precise opcode grouping shown is merely an example. In some embodiments, a move and logic opcode group 742 includes data movement and logic instructions (e.g., move (mov), compare (cmp)). In some embodiments, move and logic group 742 shares the five most significant bits (MSB), where move (mov) instructions are in the form of 0000xxxxb (e.g., 0x0x) and logic instructions are in the form of 0001xxxxb (e.g., 0x01). A flow control instruction group 744 (e.g., call, jump (jmp)) includes instructions in the form of 0010xxxxb (e.g., 0x20). A miscellaneous instruction group 746 includes a mix of instructions, including synchronization instructions (e.g., wait, send) in the form of 0011xxxxb (e.g., 0x30). A parallel math instruction group 748 includes component-wise arithmetic instructions (e.g., add,

multiply (mul)) in the form of 0100xxxxb (e.g., 0x40). The parallel math group 748 performs the arithmetic operations in parallel across data channels. The vector math group 750 includes arithmetic instructions (e.g., dp4) in the form of 0101xxxxb (e.g., 0x50). The vector math group performs arithmetic such as dot product calculations on vector operands.

5

[0082] Graphics Pipeline

[0083] Fig 8 is a block diagram of another embodiment of a graphics processor 800.

Elements of **Figure 8** having the same reference numbers (or names) as the elements of any other figure herein can operate or function in any manner similar to that described elsewhere herein, but are not limited to such.

10

[0084] In some embodiments, graphics processor 800 includes a graphics pipeline 820, a media pipeline 830, a display engine 840, thread execution logic 850, and a render output pipeline 870. In some embodiments, graphics processor 800 is a graphics processor within a multi-core processing system that includes one or more general purpose processing cores. The graphics processor is controlled by register writes to one or more control registers (not shown) or via commands issued to graphics processor 800 via a ring interconnect 802. In some embodiments, ring interconnect 802 couples graphics processor 800 to other processing components, such as other graphics processors or general-purpose processors. Commands from ring interconnect 802 are interpreted by a command streamer 803, which supplies instructions to individual components of graphics pipeline 820 or media pipeline 830.

15

20

[0085] In some embodiments, command streamer 803 directs the operation of a vertex fetcher 805 that reads vertex data from memory and executes vertex-processing commands provided by command streamer 803. In some embodiments, vertex fetcher 805 provides vertex data to a vertex shader 807, which performs coordinate space transformation and lighting operations to each vertex. In some embodiments, vertex fetcher 805 and vertex shader 807 execute vertex-processing instructions by dispatching execution threads to execution units 852A, 852B via a thread dispatcher 831.

25

[0086] In some embodiments, execution units 852A, 852B are an array of vector processors having an instruction set for performing graphics and media operations. In some embodiments, execution units 852A, 852B have an attached L1 cache 851 that is specific for each array or shared between the arrays. The cache can be configured as a data cache, an instruction cache, or a single cache that is partitioned to contain data and instructions in different partitions.

30

[0087] In some embodiments, graphics pipeline 820 includes tessellation components to perform hardware-accelerated tessellation of 3D objects. In some embodiments, a

35

programmable hull shader 811 configures the tessellation operations. A programmable domain shader 817 provides back-end evaluation of tessellation output. A tessellator 813 operates at the direction of hull shader 811 and contains special purpose logic to generate a set of detailed geometric objects based on a coarse geometric model that is provided as input to graphics pipeline 820. In some embodiments, if tessellation is not used, tessellation components 811, 813, 817 can be bypassed.

[0088] In some embodiments, complete geometric objects can be processed by a geometry shader 819 via one or more threads dispatched to execution units 852A, 852B, or can proceed directly to the clipper 829. In some embodiments, the geometry shader operates on entire geometric objects, rather than vertices or patches of vertices as in previous stages of the graphics pipeline. If the tessellation is disabled the geometry shader 819 receives input from the vertex shader 807. In some embodiments, geometry shader 819 is programmable by a geometry shader program to perform geometry tessellation if the tessellation units are disabled.

[0089] Before rasterization, a clipper 829 processes vertex data. The clipper 829 may be a fixed function clipper or a programmable clipper having clipping and geometry shader functions. In some embodiments, a rasterizer 873 in the render output pipeline 870 dispatches pixel shaders to convert the geometric objects into their per pixel representations. In some embodiments, pixel shader logic is included in thread execution logic 850. In some embodiments, an application can bypass the rasterizer 873 and access un-rasterized vertex data via a stream out unit 823.

[0090] The graphics processor 800 has an interconnect bus, interconnect fabric, or some other interconnect mechanism that allows data and message passing amongst the major components of the processor. In some embodiments, execution units 852A, 852B and associated cache(s) 851, texture and media sampler 854, and texture/sampler cache 858 interconnect via a data port 856 to perform memory access and communicate with render output pipeline components of the processor. In some embodiments, sampler 854, caches 851, 858 and execution units 852A, 852B each have separate memory access paths.

[0091] In some embodiments, render output pipeline 870 contains a rasterizer and depth test component 873 that converts vertex-based objects into an associated pixel-based representation. In some embodiments, the rasterizer logic includes a windower/masker unit to perform fixed function triangle and line rasterization. Associated render and depth buffer caches 878, 879 are also available in some embodiments. A pixel operations component 877 performs pixel-based operations on the data, though in some instances, pixel operations associated with 2D operations (e.g. bit block image transfers with blending) are performed by the 2D engine 841, or substituted at display time by the display controller 843 using overlay display planes. In some

embodiments, a shared L3 cache 875 is available to all graphics components, allowing the sharing of data without the use of main system memory.

[0092] In some embodiments, graphics processor media pipeline 830 includes a media engine 837 and a video front end 834. In some embodiments, video front end 834 receives pipeline commands from the command streamer 803. In some embodiments, media pipeline 830 includes a separate command streamer. In some embodiments, video front-end 834 processes media commands before sending the command to the media engine 837. In some embodiments, media engine 337 includes thread spawning functionality to spawn threads for dispatch to thread execution logic 850 via thread dispatcher 831.

[0093] In some embodiments, graphics processor 800 includes a display engine 840. In some embodiments, display engine 840 is external to processor 800 and couples with the graphics processor via the ring interconnect 802, or some other interconnect bus or fabric. In some embodiments, display engine 840 includes a 2D engine 841 and a display controller 843. In some embodiments, display engine 840 contains special purpose logic capable of operating independently of the 3D pipeline. In some embodiments, display controller 843 couples with a display device (not shown), which may be a system integrated display device, as in a laptop computer, or an external display device attached via a display device connector.

[0094] In some embodiments, graphics pipeline 820 and media pipeline 830 are configurable to perform operations based on multiple graphics and media programming interfaces and are not specific to any one application programming interface (API). In some embodiments, driver software for the graphics processor translates API calls that are specific to a particular graphics or media library into commands that can be processed by the graphics processor. In some embodiments, support is provided for the Open Graphics Library (OpenGL) and Open Computing Language (OpenCL) from the Khronos Group, the Direct3D library from the Microsoft Corporation, or support may be provided to both OpenGL and D3D. Support may also be provided for the Open Source Computer Vision Library (OpenCV). A future API with a compatible 3D pipeline would also be supported if a mapping can be made from the pipeline of the future API to the pipeline of the graphics processor.

[0095] **Graphics Pipeline Programming**

[0096] **Fig 9A** is a block diagram illustrating a graphics processor command format 900 according to some embodiments. **Fig 9B** is a block diagram illustrating a graphics processor command sequence 910 according to an embodiment. The solid lined boxes in **Figure 9A** illustrate the components that are generally included in a graphics command while the dashed lines include components that are optional or that are only included in a sub-set of the graphics commands. The exemplary graphics processor command format 900 of **Figure 9A** includes data

fields to identify a target client 902 of the command, a command operation code (opcode) 904, and the relevant data 906 for the command. A sub-opcode 905 and a command size 908 are also included in some commands.

[0097] In some embodiments, client 902 specifies the client unit of the graphics device that processes the command data. In some embodiments, a graphics processor command parser examines the client field of each command to condition the further processing of the command and route the command data to the appropriate client unit. In some embodiments, the graphics processor client units include a memory interface unit, a render unit, a 2D unit, a 3D unit, and a media unit. Each client unit has a corresponding processing pipeline that processes the commands. Once the command is received by the client unit, the client unit reads the opcode 904 and, if present, sub-opcode 905 to determine the operation to perform. The client unit performs the command using information in data field 906. For some commands an explicit command size 908 is expected to specify the size of the command. In some embodiments, the command parser automatically determines the size of at least some of the commands based on the command opcode. In some embodiments commands are aligned via multiples of a double word.

[0098] The flow diagram in **Figure 9B** shows an exemplary command sequence 910. In some embodiments, software or firmware of a data processing system that features an embodiment of a graphics processor uses a version of the command sequence shown to set up, execute, and terminate a set of graphics operations. A sample command sequence is shown and described for purposes of example only as embodiments are not limited to these specific commands or to this command sequence. Moreover, the commands may be issued as batch of commands in a command sequence, such that the graphics processor will process the sequence of commands in at least partially concurrence.

[0099] In some embodiments, sample command sequence 910 may begin with a pipeline flush command 912 to cause any active graphics pipeline to complete the currently pending commands for the pipeline. In some embodiments, the 3D pipeline 922 and the media pipeline 924 do not operate concurrently. The pipeline flush is performed to cause the active graphics pipeline to complete any pending commands. In response to a pipeline flush, the command parser for the graphics processor will pause command processing until the active drawing engines complete pending operations and the relevant read caches are invalidated. Optionally, any data in the render cache that is marked 'dirty' can be flushed to memory. In some embodiments, pipeline flush command 912 can be used for pipeline synchronization or before placing the graphics processor into a low power state.

[00100] In some embodiments, a pipeline select command 913 is used when a command sequence requires the graphics processor to explicitly switch between pipelines. In some embodiments, a pipeline select command 913 is required only once within an execution context before issuing pipeline commands unless the context is to issue commands for both pipelines. In some embodiments, a pipeline flush command 912 is required immediately before a pipeline switch via the pipeline select command 913.

[00101] In some embodiments, a pipeline control command 914 configures a graphics pipeline for operation and is used to program the 3D pipeline 922 and the media pipeline 924. In some embodiments, pipeline control command 914 configures the pipeline state for the active pipeline. In one embodiment, the pipeline control command 914 is used for pipeline synchronization and to clear data from one or more cache memories within the active pipeline before processing a batch of commands.

[00102] In some embodiments, return buffer state commands 916 are used to configure a set of return buffers for the respective pipelines to write data. Some pipeline operations require the allocation, selection, or configuration of one or more return buffers into which the operations write intermediate data during processing. In some embodiments, the graphics processor also uses one or more return buffers to store output data and to perform cross thread communication. In some embodiments, the return buffer state 916 includes selecting the size and number of return buffers to use for a set of pipeline operations.

[00103] The remaining commands in the command sequence differ based on the active pipeline for operations. Based on a pipeline determination 920, the command sequence is tailored to the 3D pipeline 922 beginning with the 3D pipeline state 930, or the media pipeline 924 beginning at the media pipeline state 940.

[00104] The commands for the 3D pipeline state 930 include 3D state setting commands for vertex buffer state, vertex element state, constant color state, depth buffer state, and other state variables that are to be configured before 3D primitive commands are processed. The values of these commands are determined at least in part based the particular 3D API in use. In some embodiments, 3D pipeline state 930 commands are also able to selectively disable or bypass certain pipeline elements if those elements will not be used.

[00105] In some embodiments, 3D primitive 932 command is used to submit 3D primitives to be processed by the 3D pipeline. Commands and associated parameters that are passed to the graphics processor via the 3D primitive 932 command are forwarded to the vertex fetch function in the graphics pipeline. The vertex fetch function uses the 3D primitive 932 command data to generate vertex data structures. The vertex data structures are stored in one or more return buffers. In some embodiments, 3D primitive 932 command is used to perform

vertex operations on 3D primitives via vertex shaders. To process vertex shaders, 3D pipeline 922 dispatches shader execution threads to graphics processor execution units.

[00106] In some embodiments, 3D pipeline 922 is triggered via an execute 934 command or event. In some embodiments, a register write triggers command execution. In some
5 embodiments execution is triggered via a ‘go’ or ‘kick’ command in the command sequence. In one embodiment command execution is triggered using a pipeline synchronization command to flush the command sequence through the graphics pipeline. The 3D pipeline will perform geometry processing for the 3D primitives. Once operations are complete, the resulting geometric objects are rasterized and the pixel engine colors the resulting pixels. Additional
10 commands to control pixel shading and pixel back end operations may also be included for those operations.

[00107] In some embodiments, sample command sequence 910 follows the media pipeline 924 path when performing media operations. In general, the specific use and manner of programming for the media pipeline 924 depends on the media or compute operations to be
15 performed. Specific media decode operations may be offloaded to the media pipeline during media decode. In some embodiments, the media pipeline can also be bypassed and media decode can be performed in whole or in part using resources provided by one or more general purpose processing cores. In one embodiment, the media pipeline also includes elements for general-purpose graphics processor unit (GPGPU) operations, where the graphics processor is
20 used to perform SIMD vector operations using computational shader programs that are not explicitly related to the rendering of graphics primitives.

[00108] In some embodiments, media pipeline 924 is configured in a similar manner as the 3D pipeline 922. A set of media pipeline state commands 940 are dispatched or placed into in a command queue before the media object commands 942. In some embodiments, media
25 pipeline state commands 940 include data to configure the media pipeline elements that will be used to process the media objects. This includes data to configure the video decode and video encode logic within the media pipeline, such as encode or decode format. In some embodiments, media pipeline state commands 940 also support the use one or more pointers to “indirect” state elements that contain a batch of state settings.

[00109] In some embodiments, media object commands 942 supply pointers to media
30 objects for processing by the media pipeline. The media objects include memory buffers containing video data to be processed. In some embodiments, all media pipeline states must be valid before issuing a media object command 942. Once the pipeline state is configured and media object commands 942 are queued, the media pipeline 924 is triggered via an execute
35 command 944 or an equivalent execute event (e.g., register write). Output from media pipeline

924 may then be post processed by operations provided by the 3D pipeline 922 or the media pipeline 924. In some embodiments, GPGPU operations are configured and executed in a similar manner as media operations.

[00110] Graphics Software Architecture

5 **[00111]** **Figure 10** illustrates exemplary graphics software architecture 1000 for a data processing system according to some embodiments. In some embodiments, software architecture includes a 3D graphics application 1010, an operating system 1020, and at least one processor 1030. In some embodiments, processor 1030 includes a graphics processor 1032 and one or more general-purpose processor core(s) 1034. The graphics application 1010 and
10 operating system 1020 each execute in the system memory 1050 of the data processing system.

[00112] In some embodiments, 3D graphics application 1010 contains one or more shader programs including shader instructions 1012. The shader language instructions may be in a high-level shader language, such as the High Level Shader Language (HLSL) or the OpenGL Shader Language (GLSL). The application also includes executable instructions 1014 in a
15 machine language suitable for execution by the general-purpose processor core 1034. The application also includes graphics objects 1016 defined by vertex data.

[00113] In some embodiments, operating system 1020 is a Microsoft® Windows® operating system from the Microsoft Corporation, a proprietary UNIX-like operating system, or an open source UNIX-like operating system using a variant of the Linux kernel. When the
20 Direct3D API is in use, the operating system 1020 uses a front-end shader compiler 1024 to compile any shader instructions 1012 in HLSL into a lower-level shader language. The compilation may be a just-in-time (JIT) compilation or the application can perform shader pre-compilation. In some embodiments, high-level shaders are compiled into low-level shaders during the compilation of the 3D graphics application 1010.

25 **[00114]** In some embodiments, user mode graphics driver 1026 contains a back-end shader compiler 1027 to convert the shader instructions 1012 into a hardware specific representation. When the OpenGL API is in use, shader instructions 1012 in the GLSL high-level language are passed to a user mode graphics driver 1026 for compilation. In some
30 embodiments, user mode graphics driver 1026 uses operating system kernel mode functions 1028 to communicate with a kernel mode graphics driver 1029. In some embodiments, kernel mode graphics driver 1029 communicates with graphics processor 1032 to dispatch commands and instructions.

VIRTUALIZATION WITH ADDRESS

BALLOONING AND ADDRESS SPACE REMAPPING

[00115] Turning now to **Figure 11A**, a graphics virtualization environment is shown in which a plurality of virtual machines (VMs) 1110 (1110a-1110d) access graphics hardware 1112 via a virtual machine monitor (VMM) 1114. The graphics hardware 1112 may include, for example, registers and/or memory devices that support the performance of graphics related tasks.

5 Each VM 1110 may function as a self-contained platform, running its own operating system (OS) and/or application software. In the illustrated example, the VM 1110a includes a graphics driver 1116 (e.g., guest software) that conducts various operations 1118 (1118a, 1118b) with respect to the graphics hardware 1112. Other VMs may also include guest software that is specific to the OS configuration of the respective VM.

10 **[00116]** The operations 1118 may generally relate to the memory mapped input output (MMIO) address operation of graphics related applications such as, for example, gaming applications, video applications, and so forth. In general, graphics memory address space in the graphics hardware 1112 may be partitioned between the VMs 1110 so that the addresses used by each VM 1110 may be directly used by the graphics hardware 1112. In one example, the VM
15 1110a includes efficiency logic 1122 to identify an assigned address space of the VM 1110a for MMIO operations and balloon the remaining address space that is not assigned to the VM 1110a (e.g., address space assigned to the other VMs). Thus, “ballooning” may involve the VM 1110a cooperatively refraining from using the remaining address space. As will be discussed in greater detail, the VMM 1114 may ensure that the VM 1110a complies with the policy of refraining
20 from using the remaining address space. Such an approach may enable the view of the assigned address space by the VM 1110a (and indeed, the entire graphics memory address space) to be identical to the view of that assigned address space by the VMM 1114 (e.g., VMM address space). The other VMs may also include efficiency logic to balloon their non-assigned address space in a similar fashion. As a result, conventional address remapping and/or address fixing
25 solutions may be eliminated, which may in turn significantly enhance efficiency.

[00117] More particularly, a first set 1118a of the operations 1118 (e.g., performance critical operations) may be provided direct access to an aperture region of the graphics hardware 1112, whereas a second set 1118b of the operations 1118 (e.g., non- or less-performance critical operations) may be processed by a trap-and-emulate module 1120 of the VMM 1114 prior to
30 being given access to the MMIO address space of the graphics hardware 1112. The “aperture” region may be part of a graphics processing unit (GPU) graphics memory address space that is mapped to a system memory address space for central processing unit (CPU) accesses. Thus, the entire graphics memory address space (aperture and non-aperture regions) may be accessed by the GPU in the illustrated approach. The operations 1118 may involve, for example, frame
35 buffer access and other guest virtual OS activities. As will be discussed in greater detail, the

aperture regions may be partitioned among the VMs 10 so that address remapping and/or address fixing is not needed.

[00118] **Figure 11B** shows a graphics virtualization environment in which the plurality of VMs 1110 generate commands 1119 that are passed through directly to a graphics processor such as, for example, a GPU 1121. The commands 1119 may include rendering commands, media commands, blitter commands, and so forth, that may be placed in a batch buffer and/or ring buffer 1123 by a host processor such as, for example, a CPU (not shown) running the VMs 1110 and/or the VMM 1114. The GPU 1121 may fetch the commands 1119 from the batch buffer and/or ring buffer 1123 for execution. The batch buffer and/or ring buffer 1123 may therefore enable the CPU to safely send commands to the GPU 1121. In general, the commands 1119 may contain graphic memory addresses (GMAs) that may be considered non-aperture regions of the graphics hardware.

[00119] As in the case of the aperture regions associated with the MMIO operations 1118 (Figure 11A), the non-aperture regions associated with the commands 1119 may be partitioned among the VMs 1110 so that address remapping and/or address fixing is not needed. Moreover, the illustrated efficiency logic 1122 may identify an assigned address space of the VM 1110a for commands and balloon the remaining address space that is not assigned to the VM 1110a (e.g., address space assigned to the other VMs). As already noted, the other VMs may also include efficiency logic to balloon their non-assigned address space in a similar fashion.

[00120] **Figure 12** shows a partitioning approach in which a VM view 1236 of a graphics memory address space assigned to the VM is identical to a VMM view 1226 of the graphics memory address space. In the illustrated example, a particular VM (“VM2”) may issue commands for direct access to an aperture region of the graphics memory address space and direct access to a non-aperture region of the graphics memory address space. In the illustrated example, a graphics translation table (GTT) 1232 maps addresses in the graphics memory address space to addresses in a physical address space 1234 (e.g., system address space) on a system bus associated with the graphics hardware/memory. The illustrated partitioning approach may also apply to other VMs in the system (e.g., VM0, VM1, VM3). For example, the address space not assigned to VM2 may be partitioned among and assigned to VM0, VM1, VM3, etc., so that the other VMs may also issue commands for direct access to an aperture region of the graphics hardware and direct access to a non-aperture region of the graphics hardware.

[00121] More particularly, an address ballooning approach is shown in which the VM view 1236 of the graphics memory address space assigned to the VM includes ballooned address spaces. For example, remaining address space other than an assigned address space 1238 in an aperture region may be ballooned so that the VM corresponding to the assigned address space

1238 ("VM2") will not use the remaining address space in the aperture region. Similarly, remaining address space other than an assigned address space 1240 in the non-aperture region may be ballooned so that the VM in question will not use the remaining address space in the non-aperture region. The VMM view 1226 demonstrates that the assigned address spaces 1238, 1240 may be the same from the perspective of both the VM and the VMM. Although the illustrated example shows a single address space in the aperture and non-aperture regions, each VM may also have multiple segments of address space in each region.

[00122] Turning now to **Figure 13**, a method 1344 of managing virtual machines is shown. The method 1344 may be implemented in a computing device as a set of logic instructions and/or firmware stored in a machine- or computer-readable medium such as random access memory (RAM), read only memory (ROM), programmable ROM (PROM), flash memory, etc., in configurable logic such as, for example, programmable logic arrays (PLAs), field programmable gate arrays (FPGAs), complex programmable logic devices (CPLDs), in fixed-functionality logic hardware using circuit technology such as, for example, application specific integrated circuit (ASIC), complementary metal oxide semiconductor (CMOS) or transistor-transistor logic (TTL) technology, or any combination thereof. For example, computer program code to carry out operations shown in the method 1344 may be written in any combination of one or more programming languages, including an object oriented programming language such as C++ or the like and conventional procedural programming languages, such as the "C" programming language or similar programming languages. Moreover, the method 1344 may be implemented using any of the aforementioned circuit technologies.

[00123] Illustrated VM block 1346 identifies an assigned address space of a VM, wherein the assigned address space is associated with a graphics memory address space. The VM and a VMM associated with the VM may use a paravirtualized mechanism to convey information indicating the assigned address space and/or the guest software in the VM may query the VMM (e.g., through hypercall) for the assigned address space. VM block 1348 balloons the remaining address space that is not assigned to that VM. As already noted, such an approach may enable the view of the assigned and ballooned address space by the VM to be identical to the view of the host (or VMM) address space. VM block 1350 may issue a command such as, for example, a frame buffer access, rendering, media and/or blitter command, to the graphics hardware through a batch buffer and/or ring buffer such as, for example the ring buffer 1323 (FIG. 11B), as already discussed.

[00124] Illustrated VMM block 1352 provides for identifying the command from the VM, wherein the command may include a rendering command, media command, blitter command, and so forth. In the illustrated example, a scan of the command may be conducted at VMM

block 1354, wherein the scan may be conducted via the VMM and/or a hardware boundary checker. A purely VMM-based solution may be considered relatively “lightweight” because no address conversion would be involved.

[00125] If it is determined at VMM block 1356 that the command attempts to access the remaining address space in the graphics hardware that is not assigned to the VM in question, illustrated VMM block 1358 may generate an alert (e.g., host processor, central processing unit/CPU interrupt, etc.). If the VM and/or guest software is trusted, the command scan at VMM block 1354 and the determination at VMM block 1356 may be bypassed, in order to improve performance.

[00126] With regard to CPU memory accesses, an extended page table (EPT) may be used to protect against accesses outside the assigned address space. For GPU memory accesses, however, a GTT and/or per process GTT (PPGTT) may be used to protect against accesses outside the assigned address space.

[00127] **Figure 14** shows a graphics processor architecture that uses a hardware boundary checker 1462 to scan VM commands. In the illustrated example, a command parser 1464 parses commands and distributes them to various graphics modules (e.g., rendering engines) 1466, which may in turn issue those commands to a memory access module 1468 containing the hardware boundary checker 1462. In one implementation, the boundary checker 1462 uses lower/upper pair address registers for guest aperture (lower) and non-aperture (upper) regions of the graphics memory address space of the current VM. These address registers may be programmed/restored by the VMM at rendering engine/owner switch time. Thus, the below pseudo code may provide for the alert decision making process.

[00128] If $((GMA \geq \text{lower_aperture}) \& \& (GMA < \text{upper_aperture}))$, access OK;

[00129] Else if $((GMA \geq \text{lower_gmadr}) \& \& (GMA < \text{upper_gmadr}))$, access OK;

[00130] Else access is denied

[00131] Where “GMA” may refer to the address of a graphics memory access attempt, “lower_aperture” and “upper_aperture” may refer to the lower and upper address boundaries, respectively, in the aperture region of the graphics memory assigned to the VM in question, and “lower_gmadr” and “upper_gmadr” may refer to the lower and upper boundaries, respectively, in the non-aperture region of the graphics memory.

[00132] In one example, the hardware boundary checker 1462 only checks accesses from the graphics modules 1466. In such a case, the check of host processor accesses (e.g., through aperture) may be conducted using an EPT of the VMM, as already noted.

[00133] In another implementation, the hardware scan may verify the command against multiple aperture segments and multiple non-aperture segments in the assigned address space.

Thus, the boundary checker 1462 might use multiple sets of lower/upper pair address registers to support multiple segments of aperture/gmadr regions as follows.

```

[00134]      {
[00135]      For (i=0; i<num1; i++)
5  [00136]      If ((GMA>=lower_aperture[i])&&(GMA<upper_aperture[i]), access OK;
[00137]          If (i==num1)
[00138]          For (i=0; i<num2; i++)
[00139]          If ((GMA>=lower_gmadr[i])&&(GMA<upper_gmadr[i])), access OK;
[00140]          If (i==num2)
10 [00141]          access is denied.
[00142]      }

```

[00143] **Figure 15** shows a per process graphics virtualization environment 1580. In this regard, GTT address space partitioning as described herein may co-exist with per process GTT (PPGTT) based full address space switches. In general, a PPGTT address space may be
15 designed for massive rendering workload usage under current process contexts and may be unable or not suitable to be used as a global address space for performance and security reasons. Although other embodiments are considered, the environment 1580 may include two primary data paths, such as a render data path and a display data path.

[00144] For example, the environment 1580 may represent a graphics data path in an
20 advanced scheduler model, wherein, an advanced scheduler may be part of an OS or a graphics VM, such as the VMs 1510 (FIG. 1). The illustrated environment 1580 includes a display path 1582 and a render path 1584, wherein a graphics module 1590 may use graphics memory addresses to access data. The graphics module 1590 may include, for example, rendering engines, media (e.g., audio and video) engines, GPGPU (general purpose GPU) engines, and so forth.
25 The graphics memory addresses may be translated to system memory addresses by a global GTT 1588. The graphics module 1590 may also use a PPGTT including a page directory 1592 and a page table 1594 to translate graphics memory addresses to physical addresses, depending on the addressing mode the command uses. Thus, the GTT translated address space, including aperture regions, may be accessed by the GPU as a frame buffer and/or used for other rendering engine
30 purposes. Alternatively, the PPGTT translated address space may also be used as a frame buffer.

[00145] Accordingly, **Figure 16** demonstrates a virtualization approach 1670 in which the mapping of the PPGTT may not be required to be present all the time (e.g., when the context is “scheduled out”), or may be dynamically switched between guest PPGTTs. In one example,
35 each PPGTT entry includes a page directory index into a page directory (e.g., page table base,

part of the GTT), a page table index into a page table (e.g., page address, part of system memory), and an offset into a page in memory (e.g., final address). In such a case, the page directory index and the page table index may constitute the system address, a guest page number (GPN) from the guest's perspective and a host page number (HPN) from the host's perspective. Thus, the illustrated approach 1670 demonstrates that a PPGTT 1676 may be identified, a shadow PPGTT 1671 may be maintained for page table entries (PTEs), and a partitioned GTT 1673 may be maintained for aperture regions and page directory entries (PDEs) of the PPGTT 1676.

[00146] Turning now to **Figure 17**, a logic architecture 1772 (1772a-1772d) is shown. The logic architecture 1772 may generally implement one or more aspects of the method 1744 (FIG. 3), and may function as the efficiency logic 1722 (FIG. 1), already discussed. Accordingly, an addressing module 1772a (e.g., various applications seeking access to graphics memory) may identify an assigned address space of a VM, wherein the assigned address space is to be associated with a graphics memory. The logic architecture 1772 may also include a balloon module 1772b to balloon a remaining address space that is not assigned to the VM. As already discussed, the view of the graphics memory address space by the VM may be identical to the view of the VMM address space.

[00147] In one example, the logic architecture 1772 also includes a security module 1772c that identifies commands from the VM, uses the VMM to conduct scans of the commands and generates an alert (e.g., host processor/CPU interrupt), and or stops executing the command, if the commands attempt to access the remaining address space that is not assigned to the VM (e.g., the ballooned out address space). Once the scan is complete, the VMM may issue the command to the hardware device for execution.

[00148] The VMM portion of the illustrated logic architecture 1772 also includes a per process module 1772d that identifies a PPGTT 1776, maintains a shadow PPGTT 1771 for page table entries (PTEs), and maintains a partitioned GTT 1732 for aperture regions and page directory entries (PDEs) of the PPGTT 76.

[00149] A graphics processor may alternatively use a hardware boundary checker 1762 to conduct the scans of the commands, wherein the hardware boundary checker 1762 may generate an alert and/or stop executing the command if one or more of the commands attempt to access the remaining address space. Moreover, the scan may verify the commands against multiple aperture segments and multiple non-aperture segments in the assigned address space.

[00150] The graphics processor may also include a PPGTT module 1777 and a GTT module 1733 to support the address translation, wherein the scan by the hardware boundary checker 1762 may be conducted subsequent to the address translation.

[00151] **Figure 18** illustrates a processor core 1800 according to one embodiment. The processor core 1800 may be the core for any type of processor, such as a micro-processor, an embedded processor, a digital signal processor (DSP), a GPU, a network processor, or other device to execute code. Although only one processor core 1800 is illustrated in Figure 8, a processing element may alternatively include more than one of the processor core 1800 illustrated in Figure 8. The processor core 1800 may be a single-threaded core or, for at least one embodiment, the processor core 1800 may be multithreaded in that it may include more than one hardware thread context (or “logical processor”) per core.

[00152] **Figure 18** also illustrates a memory 1870 coupled to the processor 1800. The memory 1270 may be any of a wide variety of memories (including various layers of memory hierarchy) as are known or otherwise available to those of skill in the art. The memory 1870 may include one or more code 1813 instruction(s) to be executed by the processor 1800 core, wherein the code 1813 may implement one or more aspects of the method 1344 (FIG. 3), already discussed. The processor core 1800 follows a program sequence of instructions indicated by the code 1813. Each instruction may enter a front end portion 1810 and be processed by one or more decoders 1820. The decoder 1820 may generate as its output a micro operation such as a fixed width micro operation in a predefined format, or may generate other instructions, microinstructions, or control signals which reflect the original code instruction. The illustrated front end 1810 also includes register renaming logic 1825 and scheduling logic 1830, which generally allocate resources and queue the operation corresponding to the convert instruction for execution.

[00153] The processor 1800 is shown including execution logic 1850 having a set of execution units 1855-1 through 1855-N. Some embodiments may include a number of execution units dedicated to specific functions or sets of functions. Other embodiments may include only one execution unit or one execution unit that can perform a particular function. The illustrated execution logic 1850 performs the operations specified by code instructions.

[00154] After completion of execution of the operations specified by the code instructions, back end logic 1860 retires the instructions of the code 1813. In one embodiment, the processor 1800 allows out of order execution but requires in order retirement of instructions. Retirement logic 1865 may take a variety of forms as known to those of skill in the art (e.g., re-order buffers or the like). In this manner, the processor core 1800 is transformed during execution of the code 1813, at least in terms of the output generated by the decoder, the hardware registers and tables utilized by the register renaming logic 1825, and any registers (not shown) modified by the execution logic 1850.

[00155] Although not illustrated in FIG. 18, a processing element may include other elements on chip with the processor core 1800. For example, a processing element may include memory control logic along with the processor core 1800. The processing element may include I/O control logic and/or may include I/O control logic integrated with memory control logic.

5 The processing element may also include one or more caches.

[00156] Referring now to Figure 19, shown is a block diagram of a system 1900 embodiment in accordance with an embodiment. Shown in Figure 19 is a multiprocessor system 1900 that includes a first processing element 1970 and a second processing element 1980. While two processing elements 1970 and 1980 are shown, it is to be understood that an embodiment of the system 1900 may also include only one such processing element.

10 [00157] The system 1900 is illustrated as a point-to-point interconnect system, wherein the first processing element 1970 and the second processing element 1980 are coupled via a point-to-point interconnect 1950. It should be understood that any or all of the interconnects illustrated in Figure 19 may be implemented as a multi-drop bus rather than point-to-point interconnect.

15 [00158] As shown in **Figure 19**, each of processing elements 1970 and 1980 may be multicore processors, including first and second processor cores (i.e., processor cores 1974a and 1974b and processor cores 1984a and 1984b). Such cores 1974, 1974b, 1984a, 1984b may be configured to execute instruction code in a manner similar to that discussed above in connection with Figure 18.

20 [00159] Each processing element 1970, 1980 may include at least one shared cache 1896a, 1896b. The shared cache 1896a, 1896b may store data (e.g., instructions) that are utilized by one or more components of the processor, such as the cores 1974a, 1974b and 1984a, 1984b, respectively. For example, the shared cache 1996a, 1996b may locally cache data stored in a memory 1932, 1934 for faster access by components of the processor. In one or more embodiments, the shared cache 1896a, 1896b may include one or more mid-level caches, such as level 2 (L2), level 3 (L3), level 4 (L4), or other levels of cache, a last level cache (LLC), and/or combinations thereof.

25 [00160] While shown with only two processing elements 1970, 1980, it is to be understood that the scope of the embodiments are not so limited. In other embodiments, one or more additional processing elements may be present in a given processor. Alternatively, one or more of processing elements 1970, 1980 may be an element other than a processor, such as an accelerator or a field programmable gate array. For example, additional processing element(s) may include additional processors(s) that are the same as a first processor 1970, additional processor(s) that are heterogeneous or asymmetric to processor a first processor 1970,

35

accelerators (such as, e.g., graphics accelerators or digital signal processing (DSP) units), field programmable gate arrays, or any other processing element. There can be a variety of differences between the processing elements 1970, 1980 in terms of a spectrum of metrics of merit including architectural, micro architectural, thermal, power consumption characteristics, and the like. These differences may effectively manifest themselves as asymmetry and heterogeneity amongst the processing elements 1970, 1980. For at least one embodiment, the various processing elements 1970, 1980 may reside in the same die package.

[00161] The first processing element 1970 may further include memory controller logic (MC) 1972 and point-to-point (P-P) interfaces 1976 and 1978. Similarly, the second processing element 1980 may include a MC 1982 and P-P interfaces 1986 and 1988. As shown in FIG. 19, MC's 1972 and 1982 couple the processors to respective memories, namely a memory 1932 and a memory 1934, which may be portions of main memory locally attached to the respective processors. While the MC 1972 and 1982 is illustrated as integrated into the processing elements 1970, 1980, for alternative embodiments the MC logic may be discrete logic outside the processing elements 1970, 1980 rather than integrated therein.

[00162] The first processing element 1970 and the second processing element 1980 may be coupled to an I/O subsystem 1990 via P-P interconnects 1076 1086, respectively. As shown in FIG. 9, the I/O subsystem 1990 includes P-P interfaces 1994 and 1998. Furthermore, I/O subsystem 1990 includes an interface 1992 to couple I/O subsystem 1990 with a high performance graphics engine 1938. In one embodiment, the graphics engine 1938 includes a GPU and a bus 1949 may be used to couple the graphics engine 1938 to the I/O subsystem 1990. Alternately, a point-to-point interconnect may couple these components.

[00163] In turn, I/O subsystem 1990 may be coupled to a first bus 1916 via an interface 1996. In one embodiment, the first bus 1916 may be a Peripheral Component Interconnect (PCI) bus, or a bus such as a PCI Express bus or another third generation I/O interconnect bus, although the scope of the embodiments are not so limited.

[00164] As shown in Figure 19, various I/O devices 1914 (e.g., cameras) may be coupled to the first bus 1916, along with a bus bridge 1918 which may couple the first bus 1916 to a second bus 1920. In one embodiment, the second bus 1920 may be a low pin count (LPC) bus. Various devices may be coupled to the second bus 1920 including, for example, a keyboard/mouse 1912, network controllers/communication device(s) 1926 (which may in turn be in communication with a computer network), and a data storage unit 1919 such as a disk drive or other mass storage device which may include code 1930, in one embodiment. The code 1930 may include instructions for performing embodiments of one or more of the methods described above. Thus, the illustrated code 1930 may implement the method 44 (FIG. 3), and may be

similar to the code 213 (FIG. 8), already discussed. Further, an audio I/O 1924 may be coupled to second bus 1920.

[00165] Note that other embodiments are contemplated. For example, instead of the point-to-point architecture of FIG. 9, a system may implement a multi-drop bus or another such communication topology. Also, the elements of FIG. 9 may alternatively be partitioned using more or fewer integrated chips than shown in FIG. 9.

[00166] Additional Notes and Examples:

[00167] Example 1 may include an apparatus to manage virtual machines, comprising an addressing module to identify an assigned address space of a virtual machine (VM), wherein the assigned address space is to be associated with a graphics memory address space. The apparatus may also include a balloon module to balloon a remaining address space that is not assigned to the VM, wherein a view of the graphics memory address space by the VM is identical to a virtual machine monitor (VMM) address space.

[00168] Example 2 may include the apparatus of Example 1, further including a security module to identify a command from the VM, use the VMM to conduct a scan of the command and generate an alert if the command attempts to access the remaining address space that is not assigned to the VM.

[00169] Example 3 may include the apparatus of Example 1, further including a security module to identify a command from one or more of the VM or a graphics module, the security module having a hardware boundary checker to conduct a scan of the command and generate an alert if the command attempts to access the remaining address space.

[00170] Example 4 may include the apparatus of Example 3, wherein the scan is to verify the command against multiple aperture segments and multiple non-aperture segments in the assigned address space.

[00171] Example 5 may include the apparatus of Example 1, further including a per process module to identify a per process graphics translation table (PPGTT), maintain a shadow PPGTT for page table entries (PTEs), and maintain a partitioned graphics translation table (GTT) for aperture regions and page directory entries (PDEs) of the PPGTT.

[00172] Example 6 may include the apparatus of Example 1, further including a graphics translation table (GTT) to map a command address in the assigned address space to a physical address in the graphics memory.

[00173] Example 7 may include the apparatus of any one of Examples 1 to 6, wherein the assigned address space is to include an aperture region and a non-aperture region.

[00174] Example 8 may include a method of managing virtual machines, comprising identifying an assigned address space of a VM, wherein the assigned address space is associated

with a graphics memory address space. The method may also provide for ballooning a remaining address space in the graphics memory that is not assigned to the VM, wherein a view of the graphics memory address space by the VM is identical to a virtual machine monitor (VMM) address space.

5 **[00175]** Example 9 may include the method of Example 8, further including identifying a command from the VM, and using the VMM to conduct a scan of the command and generate an alert if the command attempts to access the remaining address space that is not assigned to the VM.

10 **[00176]** Example 10 may include the method of Example 8, further including identifying a command from one or more of the VM or a graphics module, and using a hardware boundary checker to conduct a scan of the command and generate an alert if the command attempts to access the remaining address space.

15 **[00177]** Example 11 may include the method of Example 10, wherein the scan is to verify the command against multiple aperture segments and multiple non-aperture segments in the assigned address space.

[00178] Example 12 may include the method of Example 8, further including identifying a per process graphics translation table (PPGTT), maintaining a shadow PPGTT for page table entries (PTEs), and maintaining a partitioned graphics translation table (GTT) for aperture regions and page directory entries (PDEs) of the PPGTT.

20 **[00179]** Example 13 may include the method of Example 8, wherein a graphics translation table (GTT) maps the assigned address space to a physical address in the graphics memory.

[00180] Example 14 may include the method of any one of Examples 8 to 13, wherein the assigned address space includes an aperture region and a non-aperture region.

25 **[00181]** Example 15 may include at least one computer readable storage medium comprising a set of instructions which, if executed by a computing device, cause the computing device to identify an assigned address space of a VM, wherein the assigned address space is to be associated with a graphics memory address space. The instructions, if executed, may also cause a computing device to balloon a remaining address space that is not assigned to the VM, wherein a view of the graphics memory address space by the VM is identical to a virtual
30 machine monitor (VMM) address space.

[00182] Example 16 may include the at least one computer readable storage medium of Example 15, wherein the instructions, if executed, cause a computing device to identify a command from the VM, and use the VMM to conduct a scan of the command and generate an alert if the command attempts to access the remaining address space that is not assigned to the
35 VM.

[00183] Example 17 may include the at least one computer readable storage medium of Example 15, wherein the instructions, if executed, cause a computing device to identify a command from one or more of the VM or a graphics module, and use a hardware boundary checker to conduct a scan of the command and generate an alert if the command attempts to access the remaining address space.

[00184] Example 18 may include the at least one computer readable storage medium of Example 17, wherein the scan is to verify the command against multiple aperture segments and multiple non-aperture segments in the assigned address space.

[00185] Example 19 may include the at least one computer readable storage medium of Example 15, wherein the instructions, if executed, cause a computing device to identify a per process graphics translation table (PPGTT), maintain a shadow PPGTT for page table entries (PTEs), and maintain a partitioned graphics translation table (GTT) for aperture regions and page directory entries (PDEs) of the PPGTT.

[00186] Example 20 may include the at least one computer readable storage medium of Example 15, wherein a graphics translation table (GTT) is to map the assigned address space to a physical address space in the graphics memory.

[00187] Example 21 may include the at least one computer readable storage medium of any one of Examples 15 to 20, wherein the assigned address space is to include an aperture region and a non-aperture region.

[00188] Example 22 may include an apparatus to manage virtual machines, comprising means for performing the method of any one of Examples 8 to 14.

APPARATUS AND METHOD FOR MANAGING A VIRTUAL GRAPHICS PROCESSOR UNIT (GPU)

[00189] As described above, there are currently two techniques which support efficient GFX virtualization, referred to as “pass-through” and “mediate pass-through.” Pass-through technologies are widely used with input/output memory management unit (IOMMU) technology such as Intel Virtualization Technology for Directed I/O (VT-d) and/or Single Root I/O Virtualization (SRIOV), where the hardware can generate multiple instances of PCI express (PCIe) functions so that each guest can directly access the runtime hardware resources (e.g., memory mapped IO (MMIO) and direct memory access (DMA)). For example, IOMMU technology will remap the guest physical DMA address to the host. However, SRIOV-capable GFX devices require extensive hardware modification and/or circuit budget and, as a result, examples of this technology have not appeared in any significant devices or platforms.

[00190] Mediate pass-through uses the software to emulate the access of less performance-critical resources (e.g., MMIO registers) from the guest, but to pass-through

performance-critical resource access, such as aperture and rendering commands. For example, guest aperture access may go directly to hardware as described above, and the guest rendering commands may be directly executed by the GPU. In addition, there are two implementations of mediate pass-through provided by different guest software stacks: a command-level forwarding based implementation such as the VMWare virtual SVGA (vSVGA), and MMIO-level trap-and-forwarding based approach, such as XenGT developed by IntelTM. The architectures are based on Intel GEN hardware, but the same underlying techniques are applicable to other GPU architectures.

[00191] To reuse guest commands for direct hardware execution, there is an additional challenge. The graphic memory addresses (GM_ADDR) used in guest commands need to be scanned and converted to host GM_ADDRs. Address space ballooning (ASB) is discussed above for the GGTT (global graphics translation table)-mapped address space, where each guest may allocate a portion of the address space, at very beginning, and use that portion of address space only to construct commands for direct execution. Because a portion of the address space is allocated ahead of time, ASB eliminates the need to perform guest GM_ADDR conversion which involves scanning/reading of the guest graphics memory, and updating of the graphics memory, which typically requires a cache flush operation resulting in a performance hit.

[00192] However, ASB poses additional challenges to VM migration. The ballooned address space may be different in the target VM compared with that in source VM, and therefore the migration may fail. A simple solution is to remove the ASB, and reuse the address remapping with shadow page tables (e.g., GGTT and PPGTT). However, this requires scanning and translation of the address from the guest GM_ADDR to the host GM_ADDR all the time, which results in a significant performance penalty.

[00193] One embodiment of the invention employs a hybrid solution to dynamic switching between ASB and address remapping. If the target VM can allocate the same address space, the scan and conversion of the guest CMDs can be skipped for performance (i.e., ASB is utilized). Otherwise, this embodiment of the invention falls back to address remapping with scan and conversion of the guest CMDs (incurring a performance penalty).

[00194] In addition, one embodiment of the invention employs resource pool management, attempting to find a target node with the desired resource (i.e., to create an identical mapping here). This works well for IT virtualization platforms where the physical node and the VM are deployed universally by the IT department, in a private cloud (e.g., such as Huawei IT VM).

[00195] Another embodiment of the invention employs a best effort ASB implementation for a mixed hybrid solution which distinguishes between the identically-mapped guest GM_ADDR space and the non-identical host GM_ADDR space. For the identically mapped

GM_ADDRs, translation does not need to be performed because ASB may be used.

Consequently, cache flushes can be avoided and less access to UC memory. For those non-identically mapped GM_ADDRs, address remapping may be used with scan and conversion of the guest CMDs (with a performance penalty).

5 **[00196]** As illustrated in **Figure 20**, one embodiment of the invention includes a dynamic addressing module 2001 which implements these techniques. In particular, the dynamic addressing module 2001 attempts to partition the address space such that the desired guest GM_ADDR space 2000 maps to the same GM_ADDR space in the host. In such a case, ASB 2010 techniques may be used and scan/conversion of the guest commands may be skipped, 10 resulting in higher performance. If, however, some or all of the desired guest GM_ADDR space 2000 is not available, then the dynamic addressing module 2001 will perform a non-identical mapping between the guest GM_ADDR space and host GM_ADDR space. In such a case, the dynamic addressing module 2001 will choose to perform address remapping 2011 in which the VMM must scan and convert the graphic address (GM_ADDR) in guest commands. Note, 15 however, that while the ASB techniques may not be employed, the resource ballooning mechanism described herein may be maintained.

[00197] In one embodiment, the dynamic addressing module 2001 may take a mixed hybrid approach in which it maps certain portions of the guest GM_ADDR space to use ASB 2010 while other portions of the guest GM_ADDR space use address remapping techniques 20 2011. For example, the dynamic addressing module 2001 may map the desired guest GM_ADDR space to the same GM_ADDR space where it is possible to do so (e.g., where the desired portions of the host GM_ADDR space are free) and revert to address remapping techniques for the remainder of the GM_ADDR space which is not identically mapped.

[00198] The above techniques are illustrated in greater detail in **Figures 21-23**. In 25 particular, **Figure 21** shows an example in which free space 2101-2102 in the initial host GM_ADDR view 2120 can be mapped to the required GM_ADDR space 2110-2111. As such, the dynamic addressing module 2001 maps the required GM_ADDR space 2110-2111 to the free space 2101-2102, resulting in a post-partitioning host GM_ADDR view 2121 in which the matching address space 2103-2104 is mapped to the virtual machine (VM2 in the example). As 30 such, ASB techniques may be used, resulting in improved performance.

[00199] **Figure 22** shows an alternate example in which the free space 2201-2202 in the initial host GM_ADDR view cannot be mapped to the required GM_ADDR space 2210-2211. As such, the dynamic addressing module 2001 cannot match the required GM_ADDR space 2210-2211 to the free space 2201-2202, resulting in a post-partitioning host GM_ADDR view 35 2221 in which there is a mismatch between the address space 2203-2204 mapped to the virtual

machine and the required GM_ADDR space 2210-2211. As such, in this embodiment, address remapping techniques must be employed including scan/conversion of guest commands.

[00200] Figure 23 shows another example in which only a portion of the required GM_ADDR space 2310-2311 can be mapped to the free space 2301-2302 in the initial host

GM_ADDR view 2320. In particular, in this embodiment, the dynamic addressing module 2001 performs mapping to construct a post-partitioning host GM_ADDR view 2321 in which a first portion 2303, 2305 of the GM_ADDR view 2321 is matched to a portion of the required GM_ADDR space 2310-2311 and a second portion 2304, 2306 of the GM_ADDR view 2321 is not matched to the required GM_ADDR space. Consequently, in this embodiment, ASB techniques may be employed for guest addresses corresponding to the first portion 2303, 2305 of the GM_ADDR view 2321 and address remapping techniques may be employed for the remainder of the guest addresses (e.g., for those portions which do not match, the VMM may perform scan/conversion of guest commands).

[00201] Figure 24 illustrates a method in accordance with one embodiment of the invention. The method may be implemented within the context of the system architectures described above, but is not limited to any particular system architecture.

[00202] At 2401, the required guest GM_ADDR space is determined as is the free space in the initial (pre-allocated) host GM_ADDR view. At 2402, a determination is made as to whether all or a portion of the required guest GM_ADDR space can be matched to the free space in the host view. If none of the GM_ADDR space can be matched, then at 2403 a partitioning such as illustrated in Figure 22 is performed in which there is a mismatch between the guest/host GM_ADDR spaces. Consequently, at 2406, address remapping techniques are performed (e.g., requiring scan/conversion of guest commands) in lieu of ASB.

[00203] If all of the required guest GM_ADDR space can be matched, then at 2405 an allocation such as illustrated in Figure 21 is performed in which there is complete matching of the guest/host GM_ADDR space. As such, at 2408 ASB techniques may be performed in lieu of address remapping, resulting in improved performance.

[00204] If only a portion of the guest GM_ADDR space can be matched, then at 2404 an allocation such as illustrated in Figure 23 is performed in which portions of the required guest GM_ADDR space are matched to the host GM_ADDR space while other portions are mismatched. Consequently, at 2408, the matched portions may employ ASB techniques while the mismatched portions may employ address remapping techniques.

[00205] The end result of the foregoing techniques is that the best available performance can be achieved. That is, ASB is used wherever it is possible to do so (e.g., where the guest/host addresses can be matched) while the system will fall back to address remapping whenever ASB

cannot be used.

[00206] In the foregoing specification, the invention has been described with reference to specific exemplary embodiments thereof. It will, however, be evident that various modifications and changes may be made thereto without departing from the broader spirit and scope of the invention as set forth in the appended claims. The specification and drawings are, accordingly, to be regarded in an illustrative rather than a restrictive sense.

[00207] Embodiments of the invention may include various steps, which have been described above. The steps may be embodied in machine-executable instructions which may be used to cause a general-purpose or special-purpose processor to perform the steps. Alternatively, these steps may be performed by specific hardware components that contain hardwired logic for performing the steps, or by any combination of programmed computer components and custom hardware components.

[00208] As described herein, instructions may refer to specific configurations of hardware such as application specific integrated circuits (ASICs) configured to perform certain operations or having a predetermined functionality or software instructions stored in memory embodied in a non-transitory computer readable medium. Thus, the techniques shown in the Figures can be implemented using code and data stored and executed on one or more electronic devices (e.g., an end station, a network element, etc.). Such electronic devices store and communicate (internally and/or with other electronic devices over a network) code and data using computer machine-readable media, such as non-transitory computer machine-readable storage media (e.g., magnetic disks; optical disks; random access memory; read only memory; flash memory devices; phase-change memory) and transitory computer machine-readable communication media (e.g., electrical, optical, acoustical or other form of propagated signals – such as carrier waves, infrared signals, digital signals, etc.). In addition, such electronic devices typically include a set of one or more processors coupled to one or more other components, such as one or more storage devices (non-transitory machine-readable storage media), user input/output devices (e.g., a keyboard, a touchscreen, and/or a display), and network connections. The coupling of the set of processors and other components is typically through one or more busses and bridges (also termed as bus controllers). The storage device and signals carrying the network traffic respectively represent one or more machine-readable storage media and machine-readable communication media. Thus, the storage device of a given electronic device typically stores code and/or data for execution on the set of one or more processors of that electronic device. Of course, one or more parts of an embodiment of the invention may be implemented using different combinations of software, firmware, and/or hardware. Throughout this detailed description, for the purposes of explanation, numerous specific details were set forth in order to provide a thorough

understanding of the present invention. It will be apparent, however, to one skilled in the art that the invention may be practiced without some of these specific details. In certain instances, well known structures and functions were not described in elaborate detail in order to avoid obscuring the subject matter of the present invention. Accordingly, the scope and spirit of the invention

5 should be judged in terms of the claims which follow

CLAIMS

What is claimed is:

1. An apparatus comprising:

5 a dynamic addressing module to map portions of an address space required by the virtual machine to matching free address spaces of a host if such matching free address spaces are available, and to select non-matching address spaces for those portions of the address space required by the virtual machine which cannot be matched with free address spaces of the host; and

10 a balloon module to perform address space ballooning (ASB) techniques for those portions of the address space required by the virtual machine which have been mapped to matching address spaces of the host; and

address remapping logic to perform address remapping techniques for those portions of the address space required by the virtual machine which have not been mapped to matching
15 address spaces of the host.

2. The apparatus as in claim 1 wherein ASB techniques comprise ballooning a remaining address space that is not assigned to the virtual machine.

20 3. The apparatus of claim 2 further including a security module to identify a command from the virtual machine, conducting a scan of the command and generating an alert if the command attempts to access the remaining address space that is not assigned to the virtual machine.

25 4. The apparatus as in claim 1 wherein if all portions of the address space required by the virtual machine can be mapped to matching free address spaces of the host, then performing ASB techniques for all of the address space required by the virtual machine.

5. The apparatus as in claim 1 wherein if no portion of the address space required by
30 the virtual machine can be mapped to matching free address spaces of the host, then performing address remapping techniques for all of the address space required by the virtual machine.

6. The apparatus as in claim 1 wherein the host comprises a virtual machine monitor (VMM) to provide a virtualized execution environment for the virtual machine.

35

7. The apparatus as in claim 6 further comprising:

a graphics translation table (GTT) usable by the VMM to map addresses in a graphics memory address space, including those portions of the address space required by the virtual machine, to addresses in a physical address space.

5

8. The apparatus of claim 1, further including a security module to identify a command from the virtual machine, the security module having a hardware boundary checker to conduct a scan of the command and generate an alert if the command attempts to access the remaining address space, wherein the scan is to verify the command against multiple aperture segments and multiple non-aperture segments in the assigned address space.

10

9. The apparatus of claim 1, further including a per process module to identify a per process graphics translation table (PPGTT), maintain a shadow PPGTT for page table entries (PTEs), and maintain a partitioned graphics translation table (GTT) for aperture regions and page directory entries (PDEs) of the PPGTT.

15

10. The apparatus of claim 1, further including a graphics translation table (GTT) to map a command address in the address space required by the virtual machine to a physical address in a graphics memory.

20

11. A method comprising:

mapping portions of an address space required by a virtual machine to matching free address spaces of a host if such matching free address spaces are available;

selecting non-matching address spaces for those portions of the address space required by the virtual machine which cannot be matched with free address spaces of the host;

25

performing address space ballooning (ASB) techniques for those portions of the address space required by the virtual machine which have been mapped to matching address spaces of the host; and

performing address remapping techniques for those portions of the address space required by the virtual machine which have not been mapped to matching address spaces of the host.

30

12. The method as in claim 11 wherein ASB techniques comprise ballooning a remaining address space that is not assigned to the virtual machine.

35

13. The method of claim 12 further comprising:

identifying a command from the virtual machine;
conducting a scan of the command; and
generating an alert if the command attempts to access the remaining address space that is not assigned to the virtual machine.

5

14. The method as in claim 11 wherein if all portions of the address space required by the virtual machine can be mapped to matching free address spaces of the host, then performing ASB techniques for all of the address space required by the virtual machine.

10 15. The method as in claim 11 wherein if no portion of the address space required by the virtual machine can be mapped to matching free address spaces of the host, then performing address remapping techniques for all of the address space required by the virtual machine.

15 16. The method as in claim 11 wherein the host comprises a virtual machine monitor (VMM) to provide a virtualized execution environment for the virtual machine.

17. The method as in claim 16 further comprising:
using a graphics translation table (GTT) by the VMM to map addresses in a graphics memory address space, including those portions of the address space required by the virtual machine, to addresses in a physical address space.

20

18. The method of claim 11, further comprising:
identifying a command from the virtual machine with a hardware boundary checker to conduct a scan of the command and generate an alert if the command attempts to access the remaining address space, wherein the scan is to verify the command against multiple aperture segments and multiple non-aperture segments in the assigned address space.

25

19. The method of claim 11 further comprising:
mapping a command address in the address space required by the virtual machine to a physical address in a graphics memory using a graphics translation table (GTT).

30

21. A machine-readable medium having program code stored thereon which, when executed by a machine, causes the machine to perform the operations of:
mapping portions of an address space required by a virtual machine to matching free address spaces of a host if such matching free address spaces are available;

35

selecting non-matching address spaces for those portions of the address space required by the virtual machine which cannot be matched with free address spaces of the host;

performing address space ballooning (ASB) techniques for those portions of the address space required by the virtual machine which have been mapped to matching address spaces of

5 the host; and

performing address remapping techniques for those portions of the address space required by the virtual machine which have not been mapped to matching address spaces of the host.

22. The machine-readable medium as in claim 21 wherein ASB techniques comprise
10 ballooning a remaining address space that is not assigned to the virtual machine.

23. The machine-readable medium as in claim 22 comprising additional program code to cause the machine to perform the operations of:

identifying a command from the virtual machine;

15 conducting a scan of the command; and

generating an alert if the command attempts to access the remaining address space that is not assigned to the virtual machine.

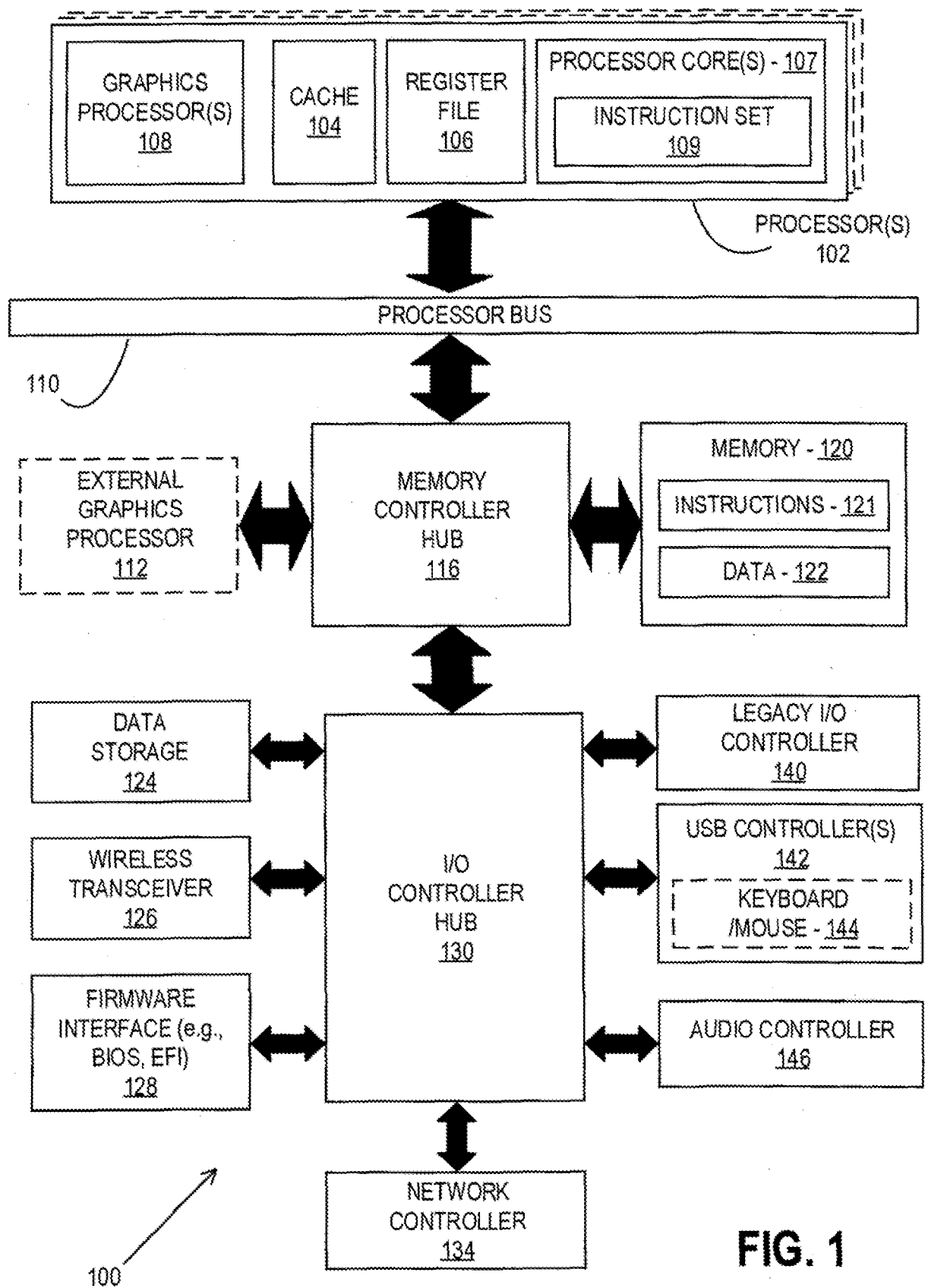


FIG. 1

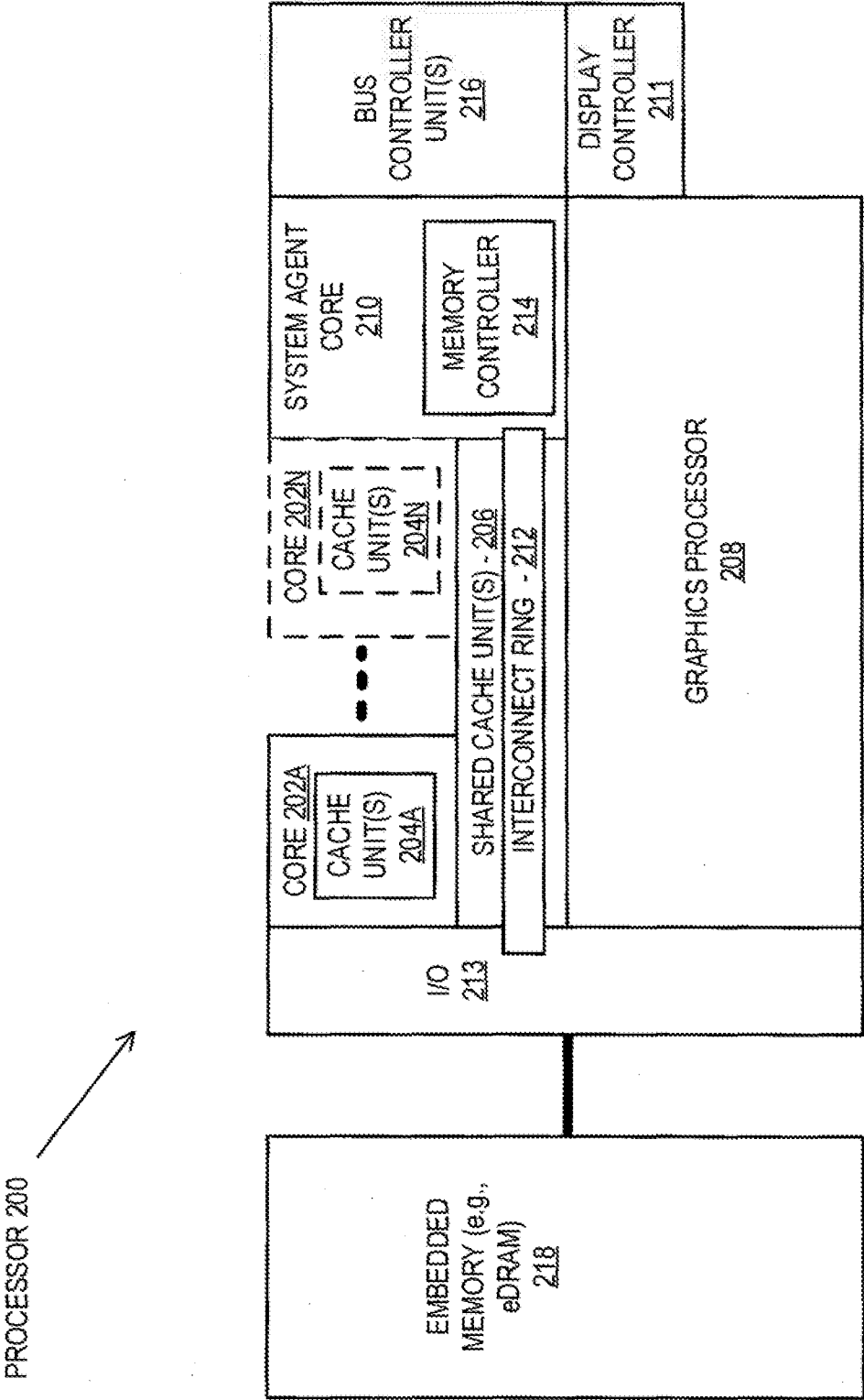
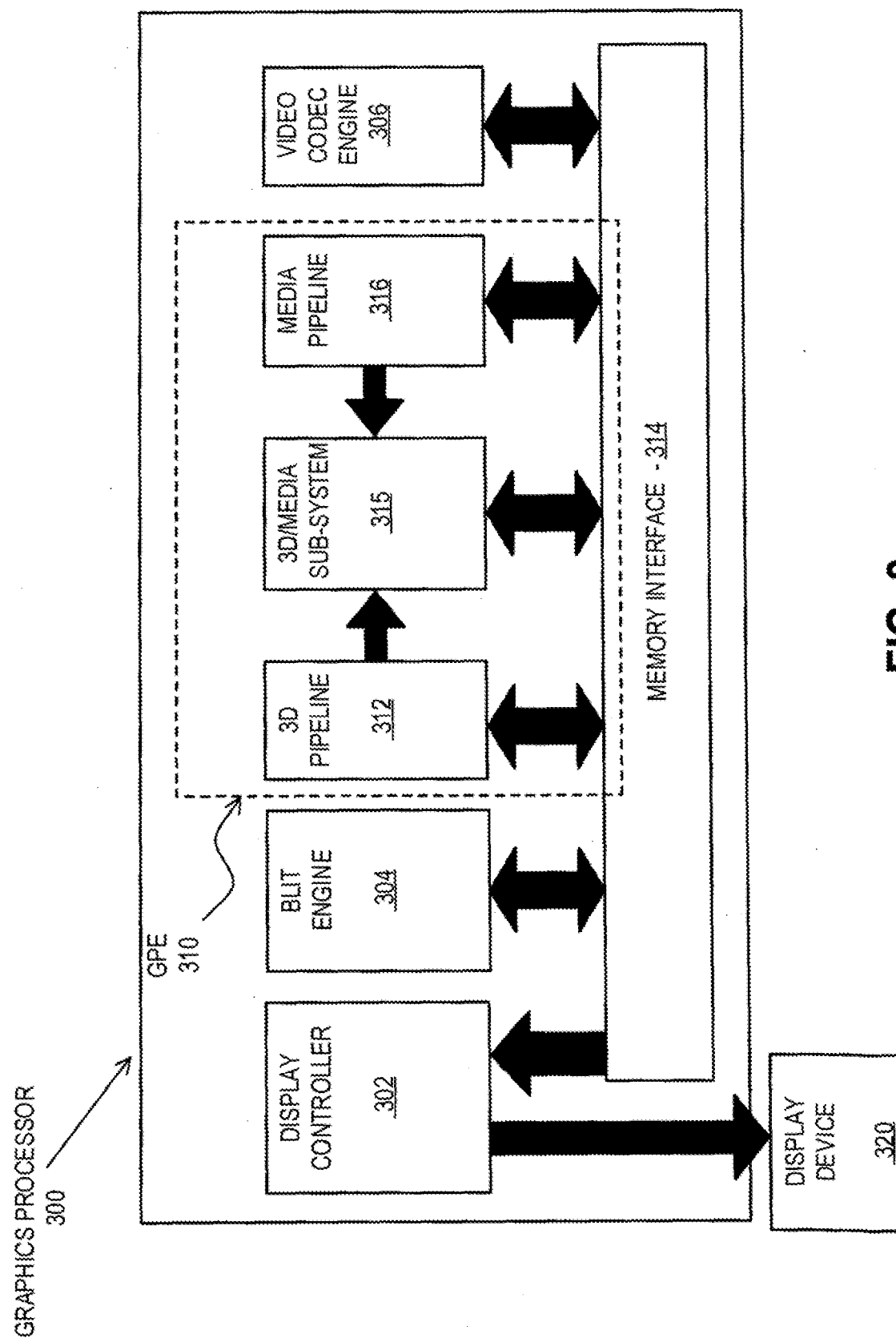


FIG. 2



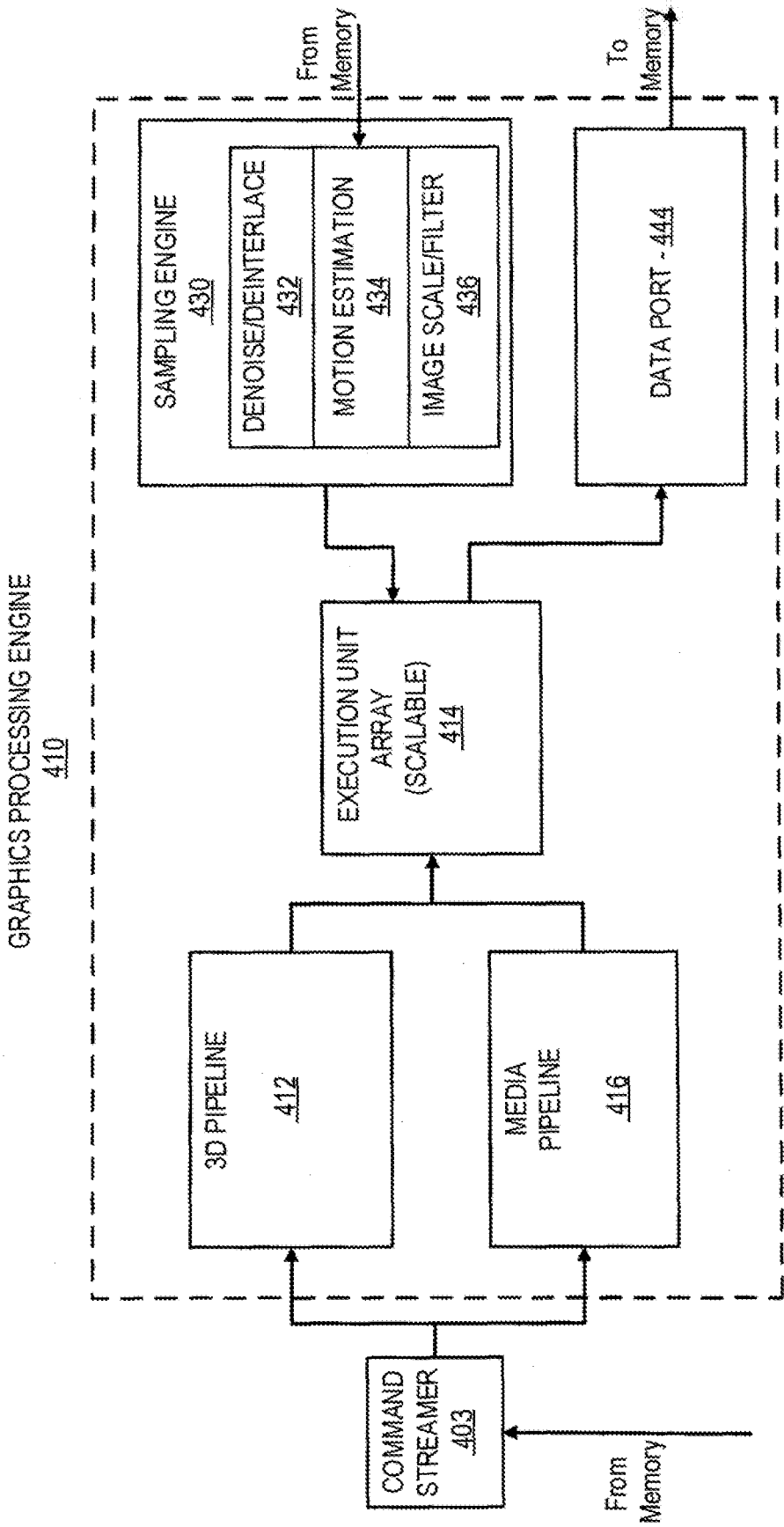


FIG. 4

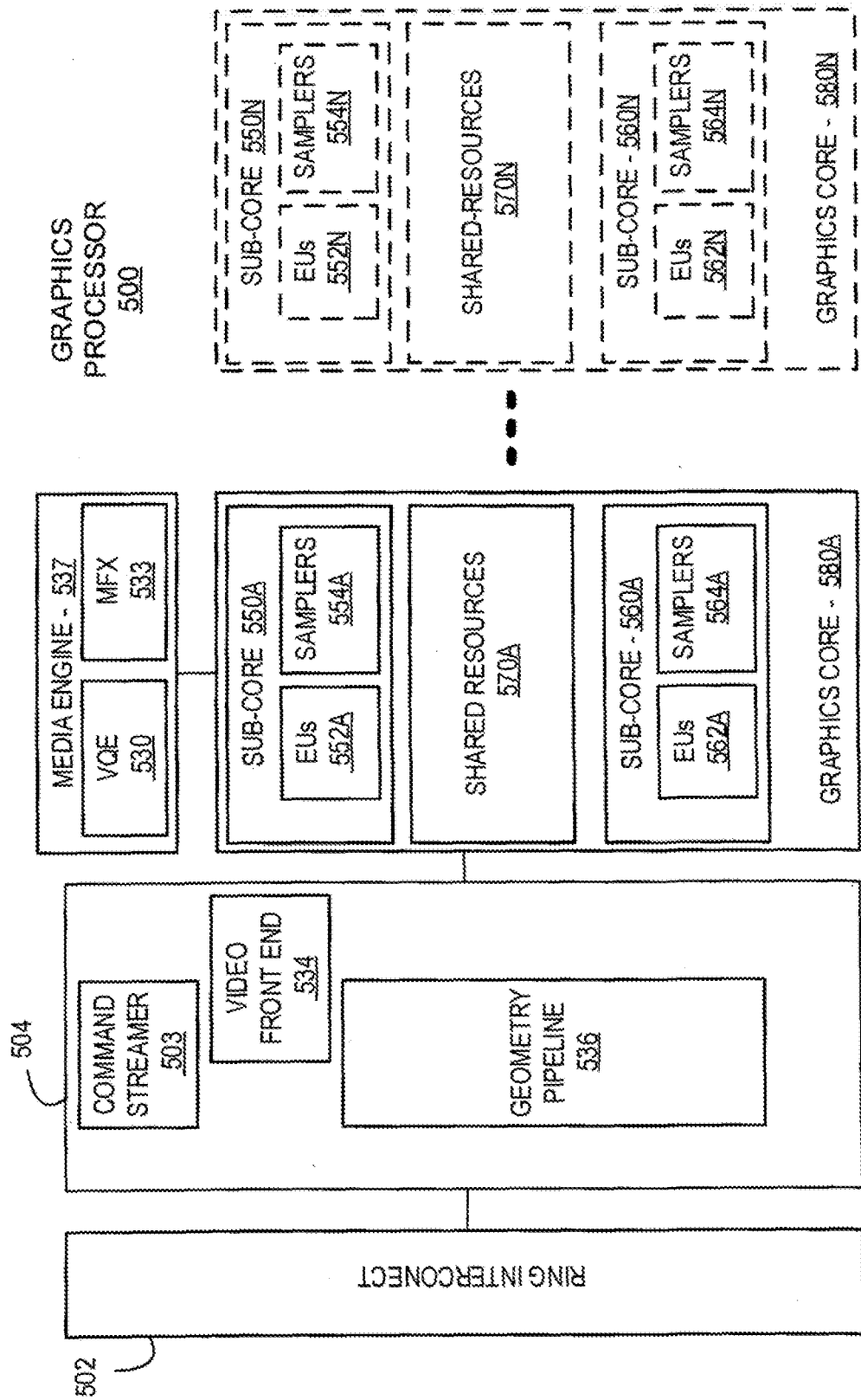


FIG. 5

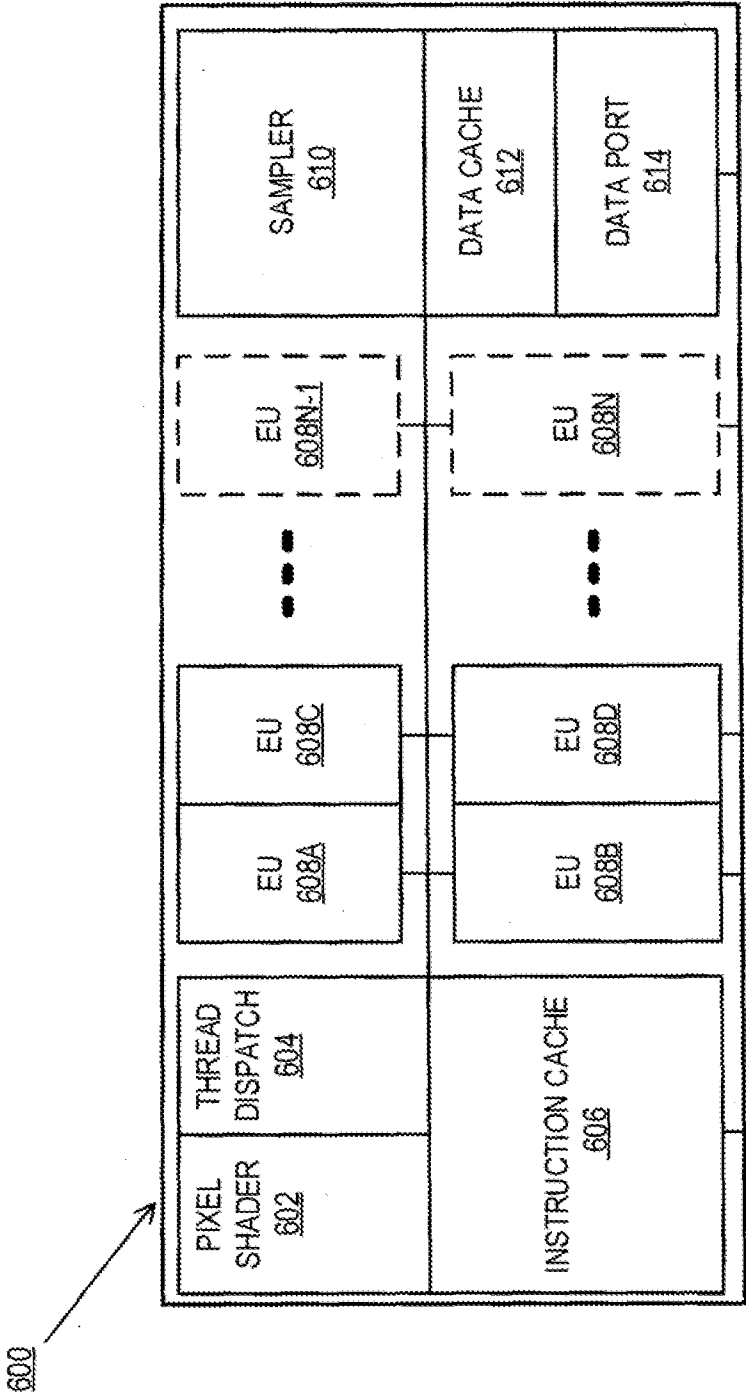


FIG. 6

GRAPHICS CORE INSTRUCTION
FORMATS

700

128-BIT INSTRUCTION

710



64-BIT COMPACT
INSTRUCTION

730



OPCODE DECODE

740

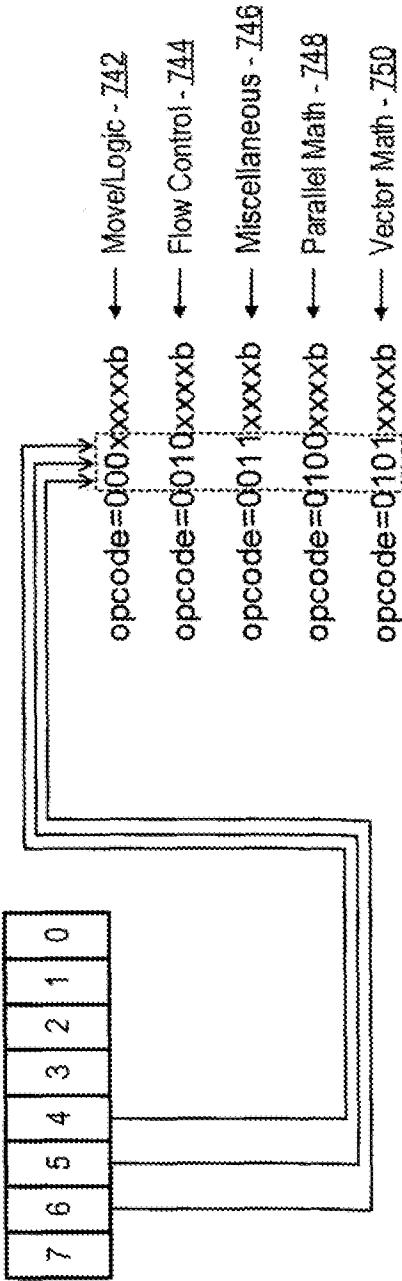


FIG. 7

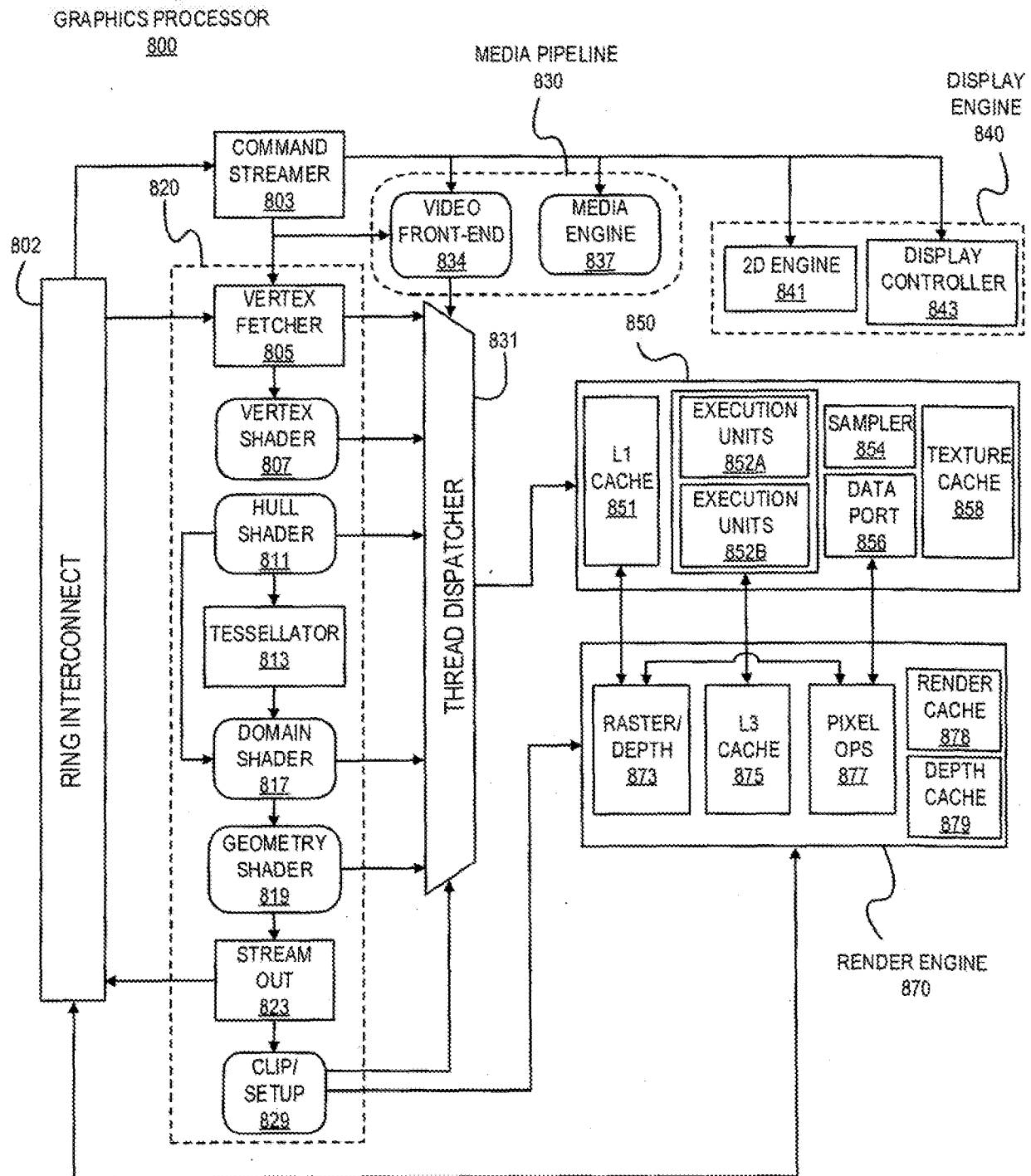


FIG. 8

FIG. 9A

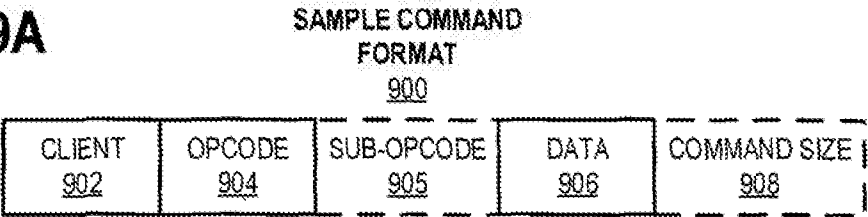
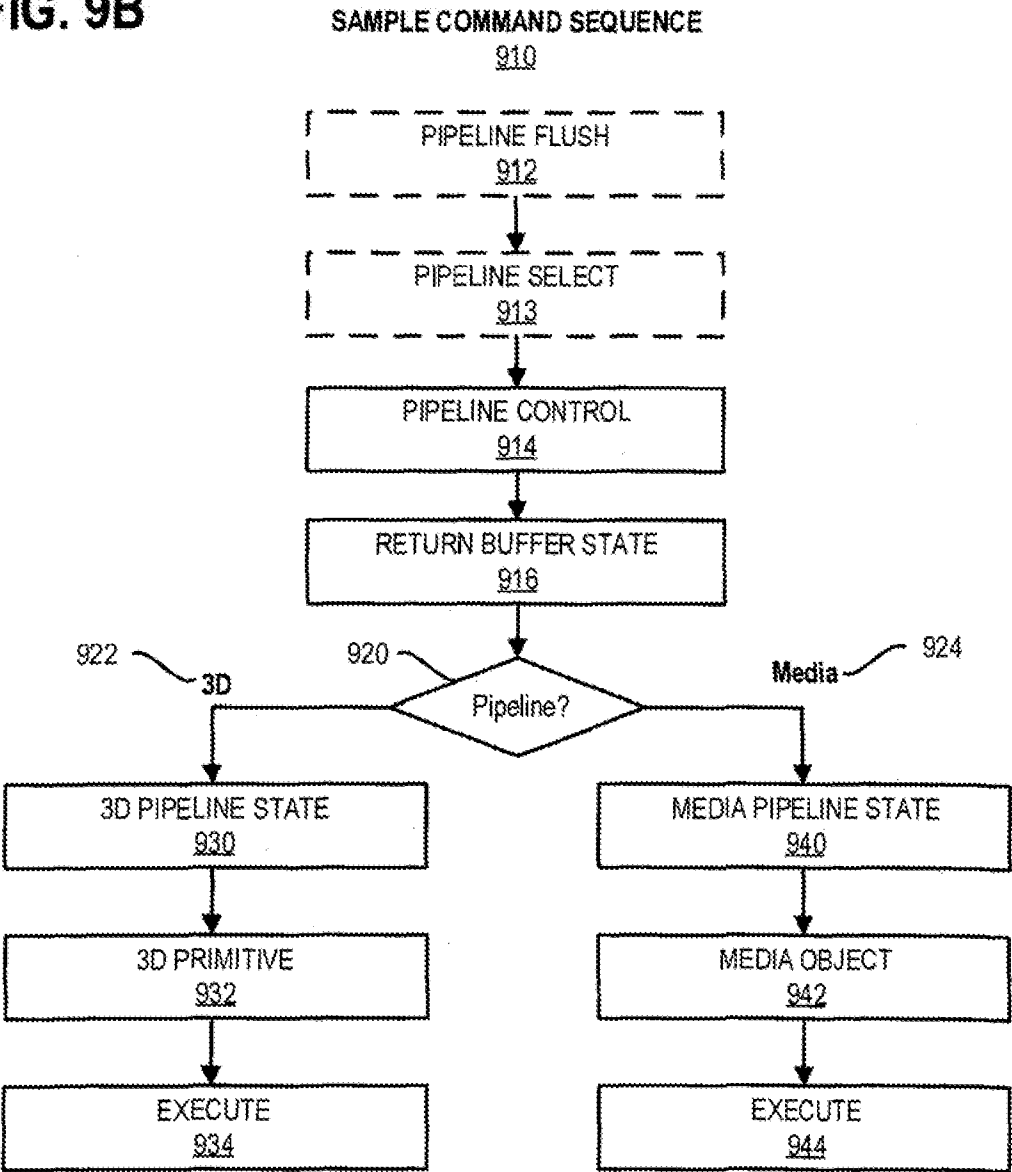
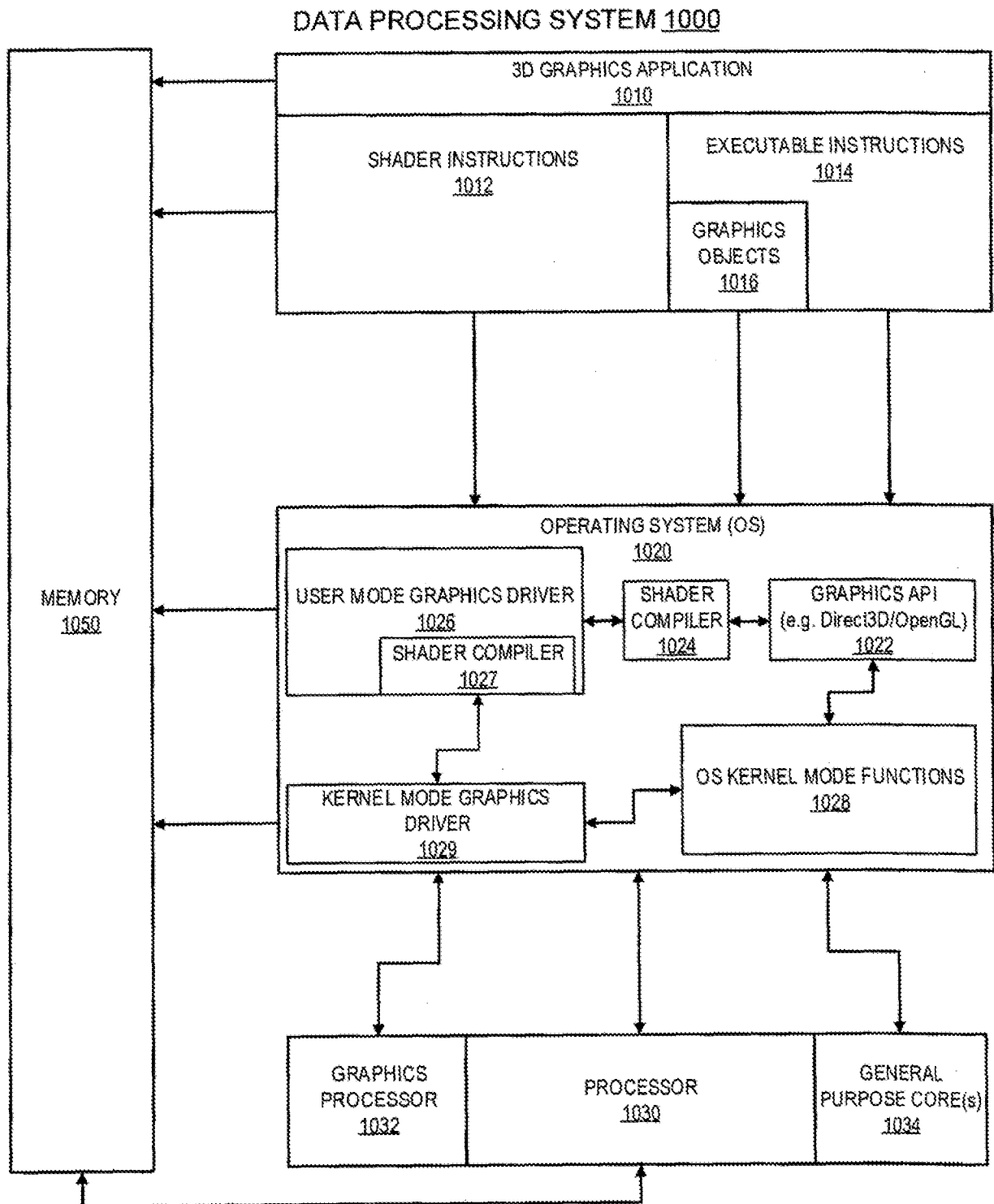


FIG. 9B



**FIG. 10**

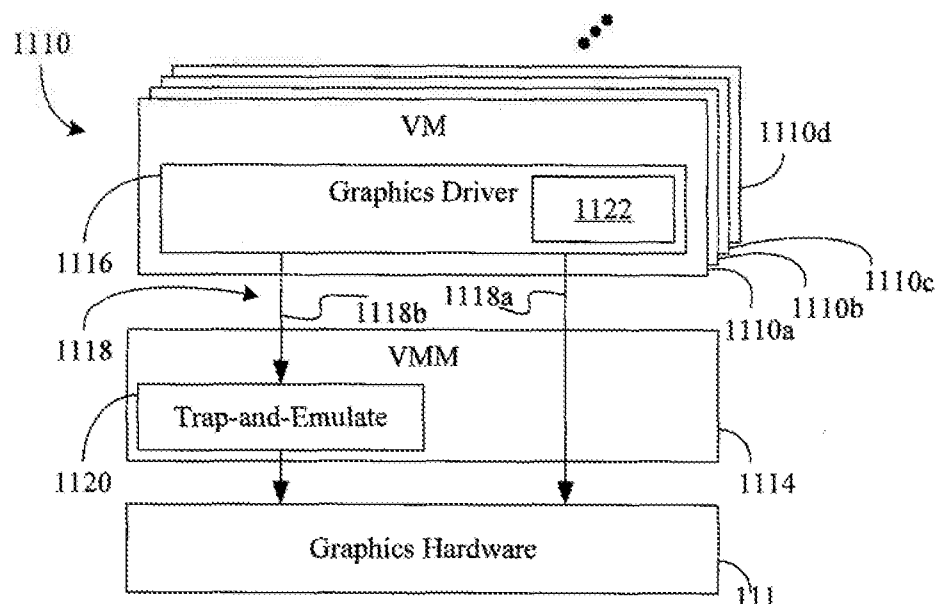


FIG. 11A

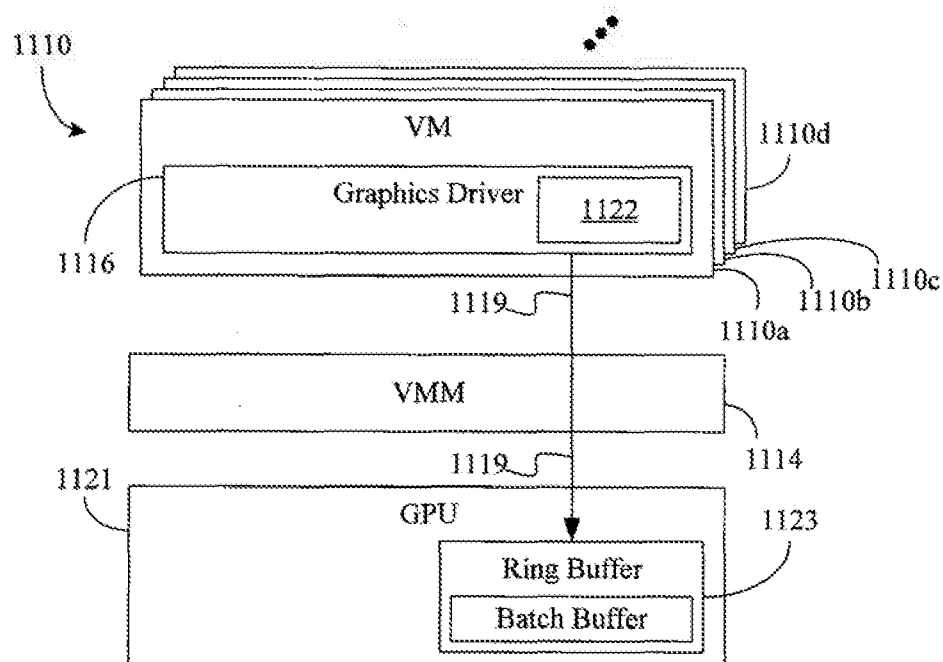


FIG. 11B

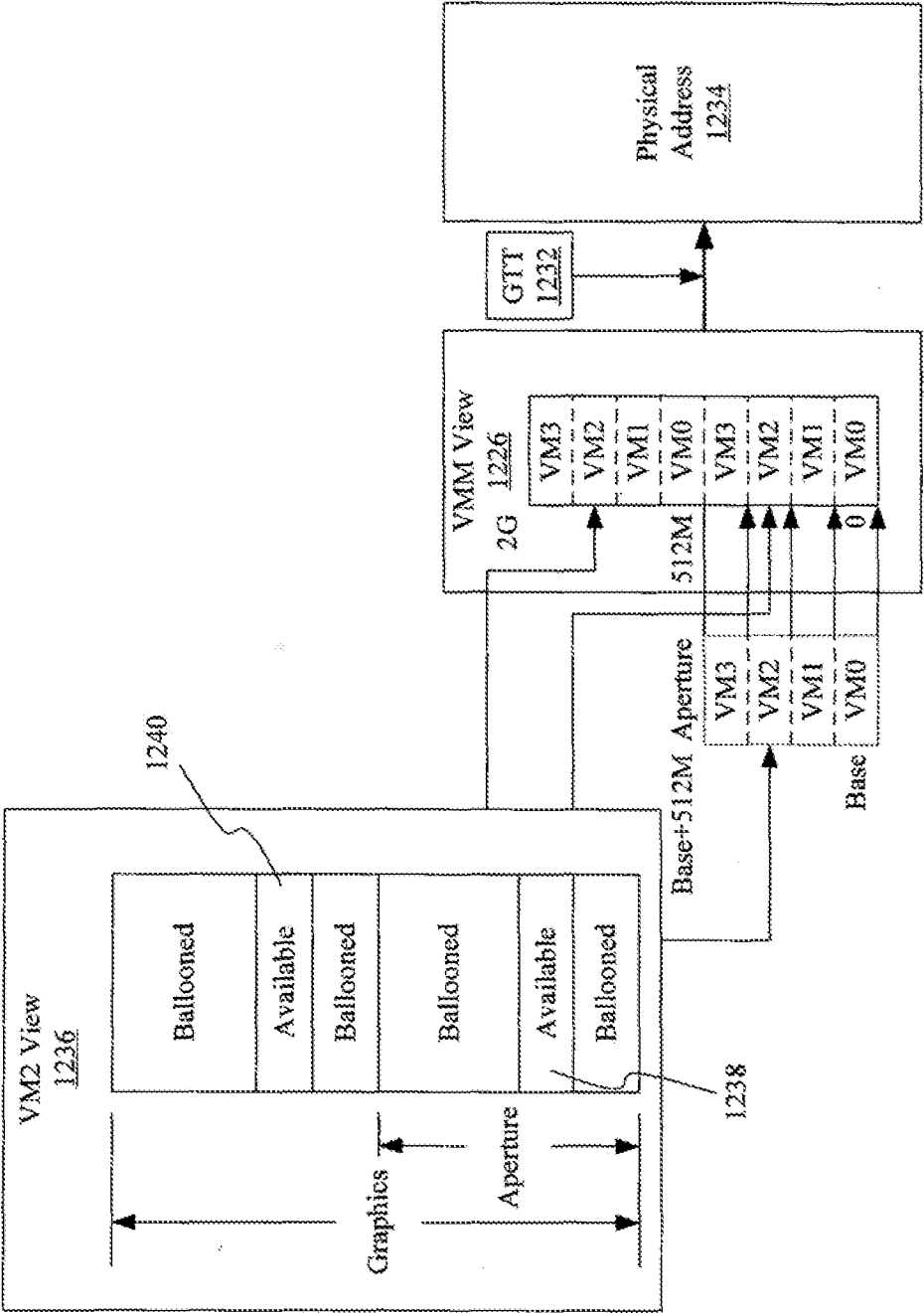


FIG. 12

FIG. 13

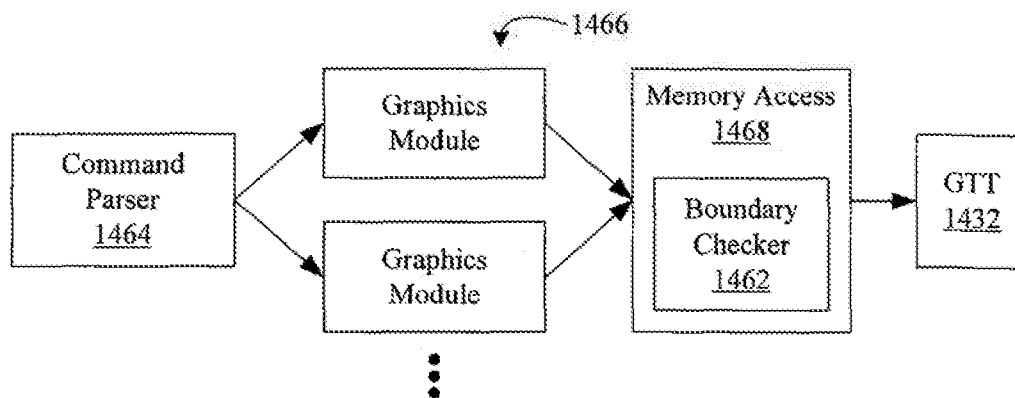
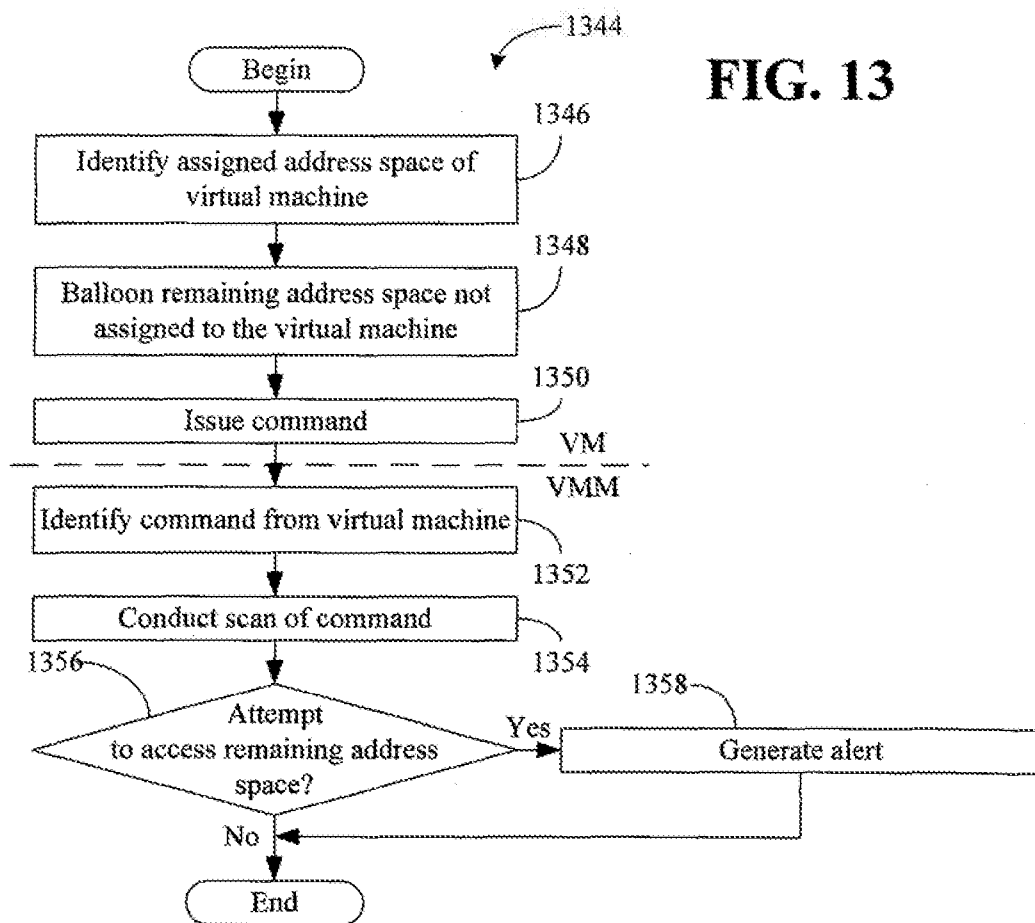


FIG. 14

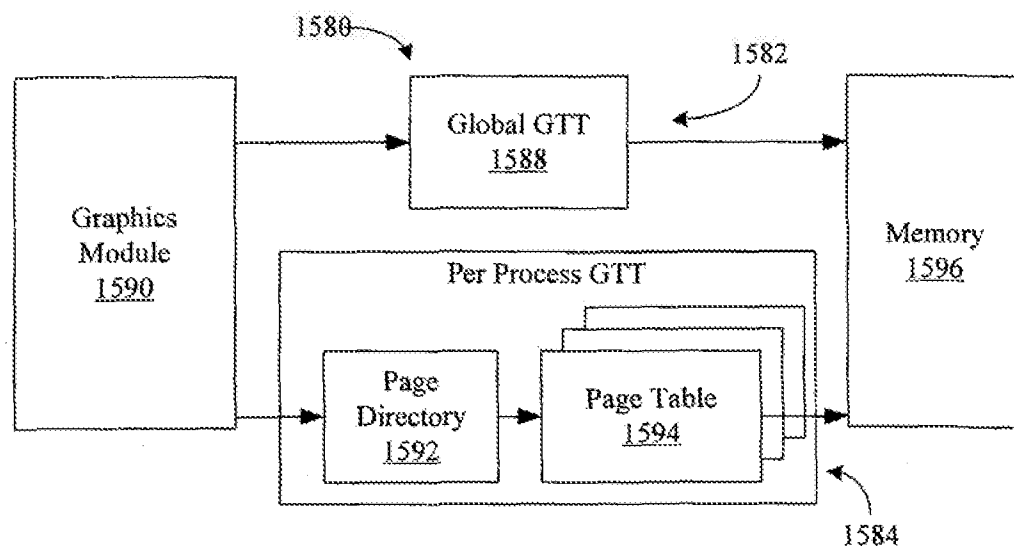


FIG. 15

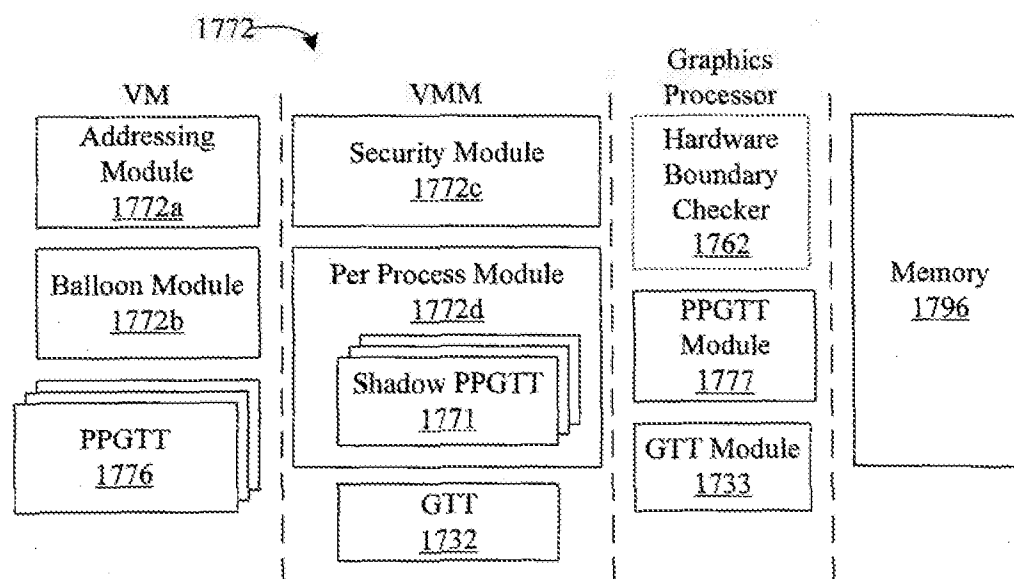


FIG. 17

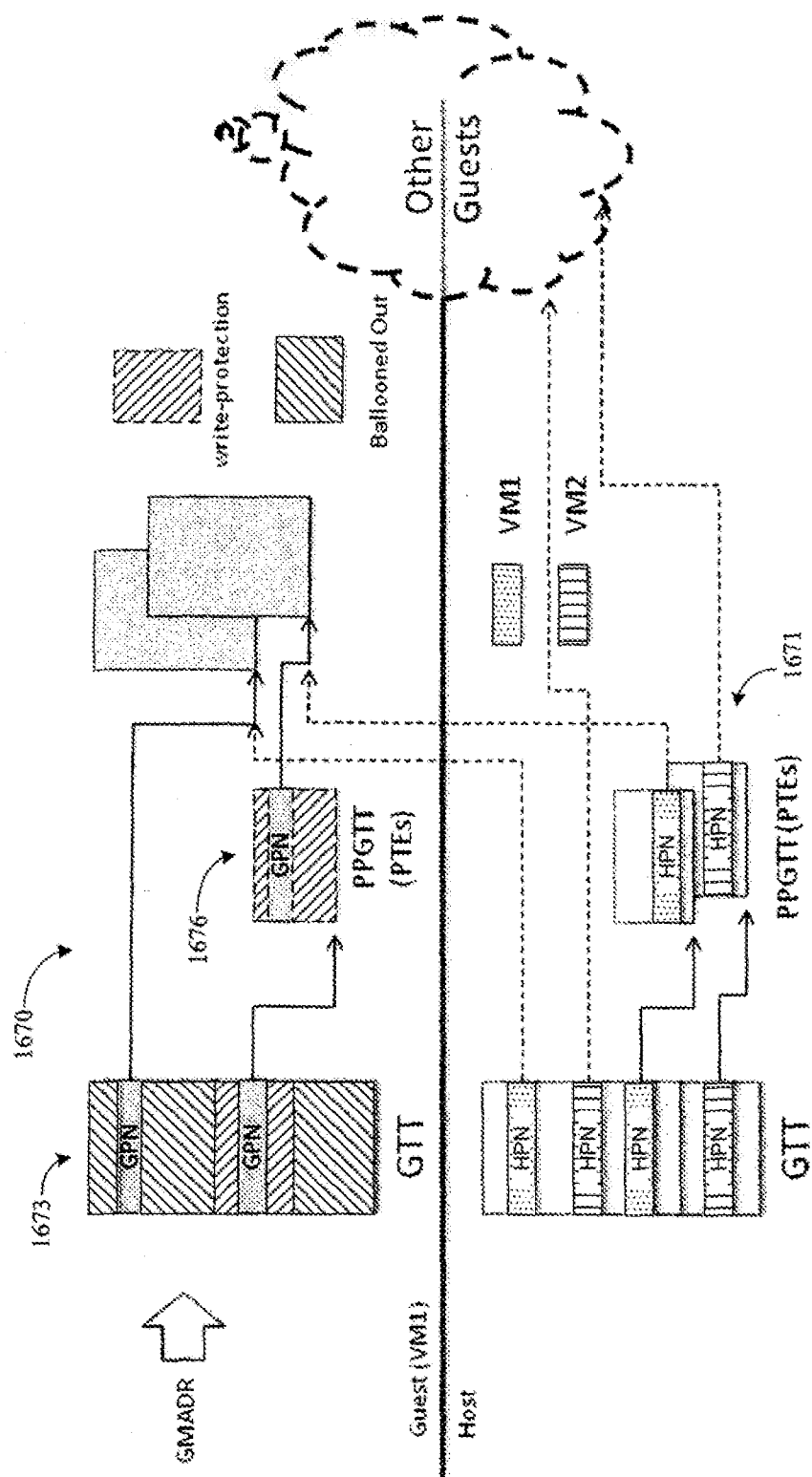


FIG. 16

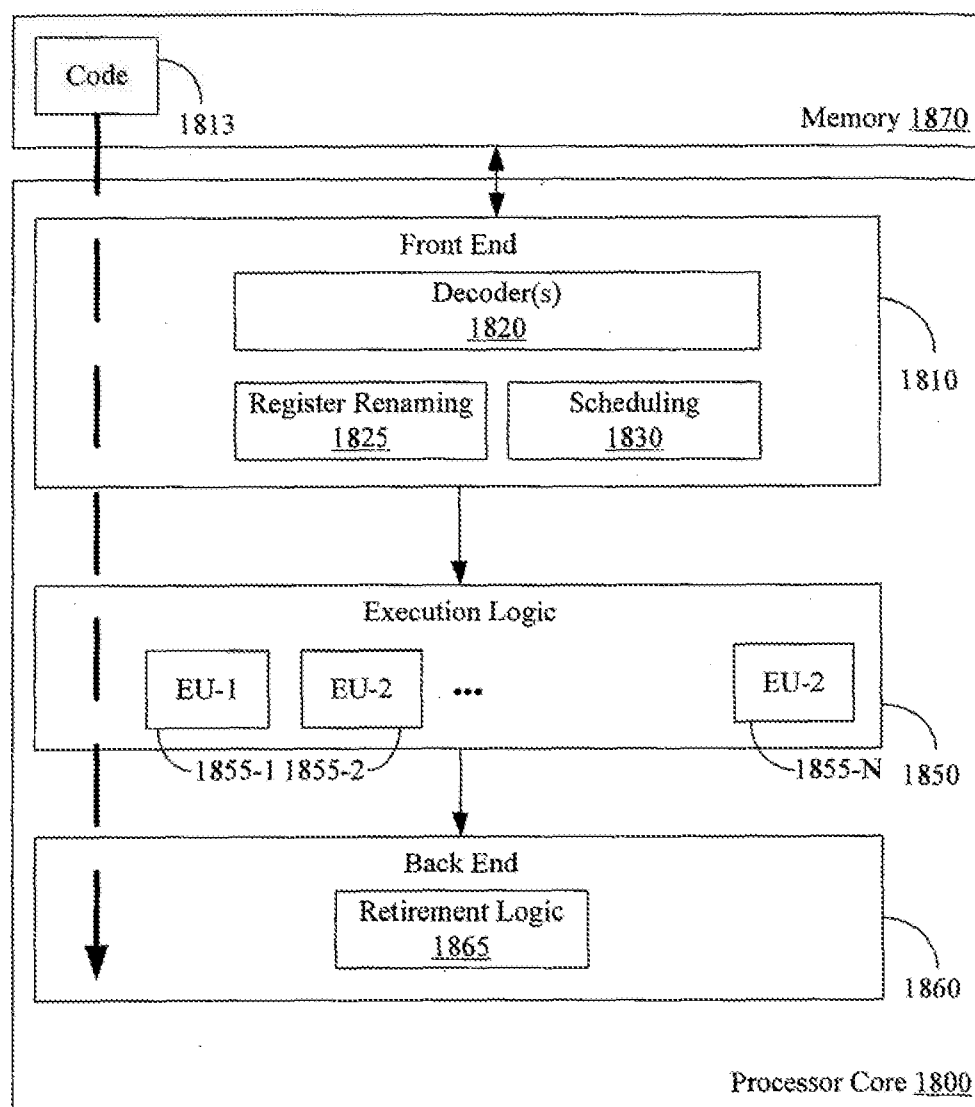
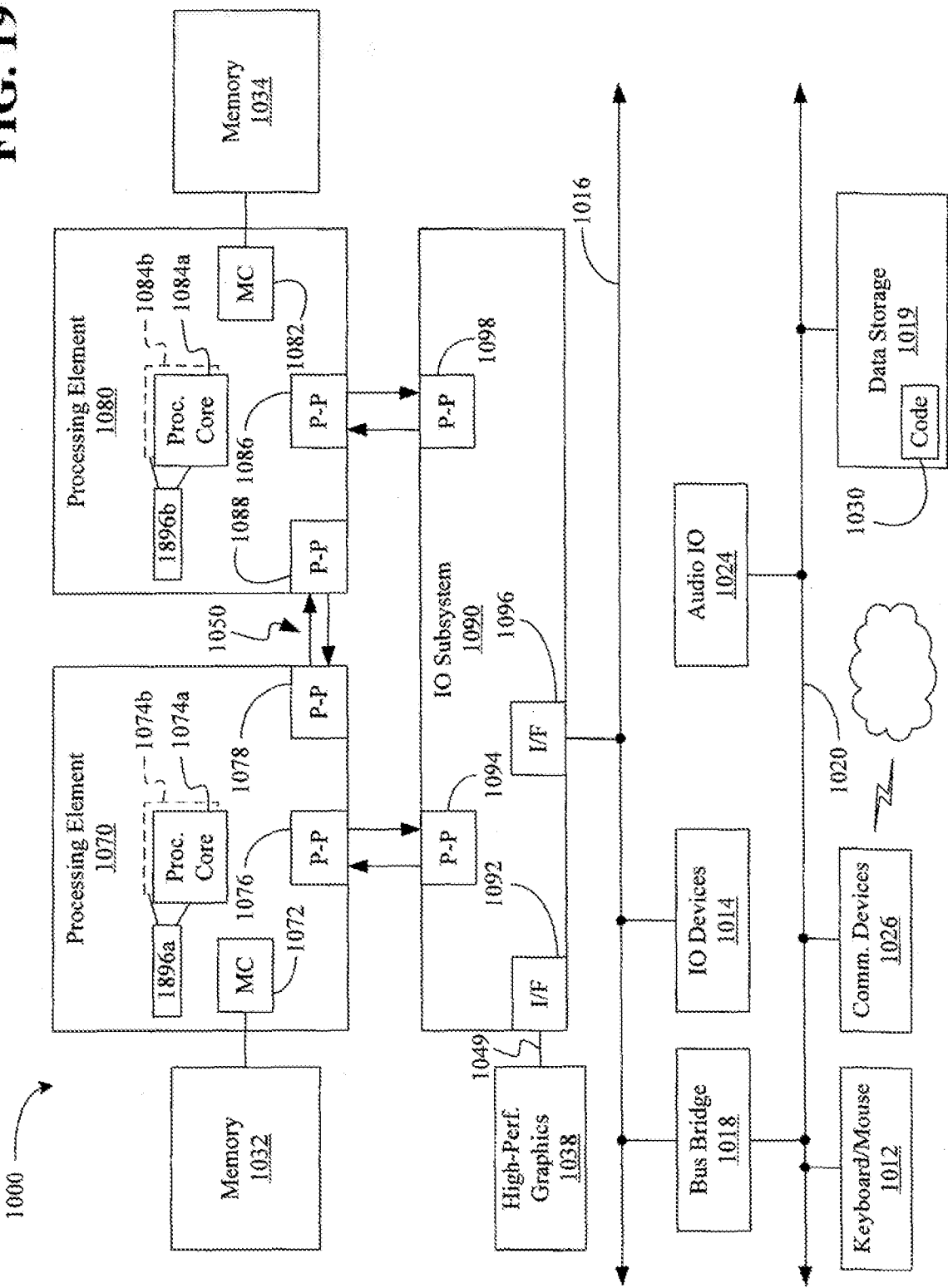
**FIG. 18**

FIG. 19



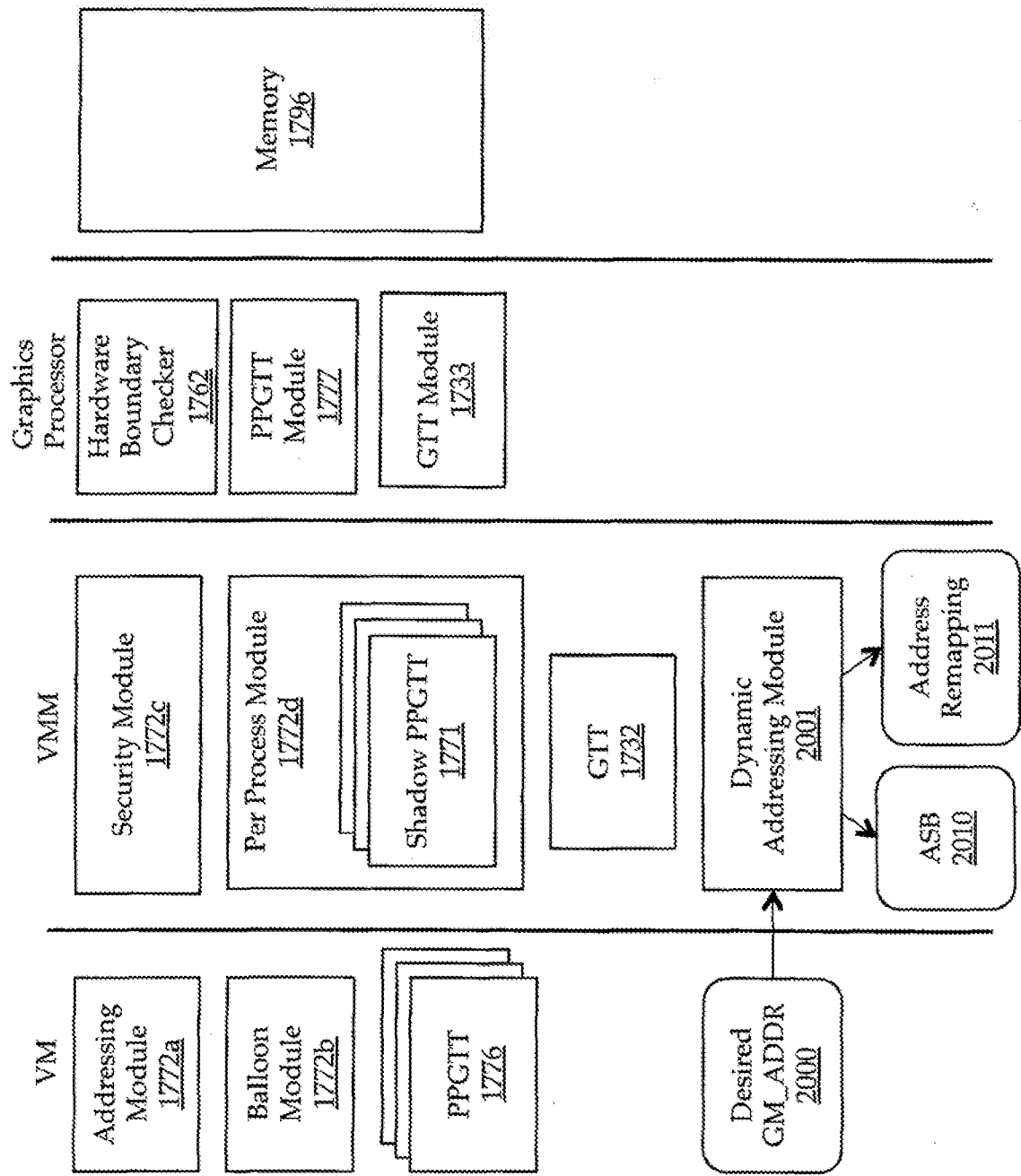


FIG. 20

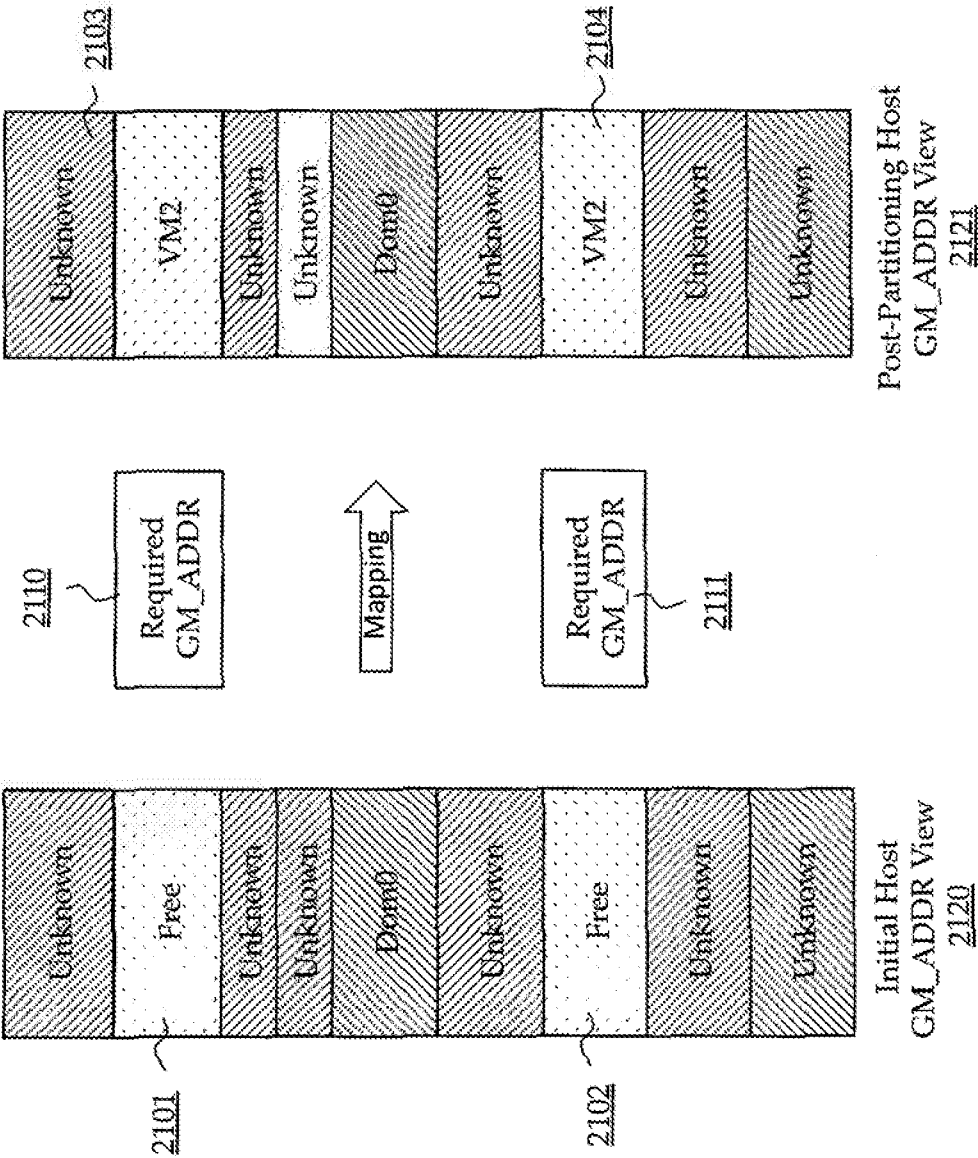


FIG. 21

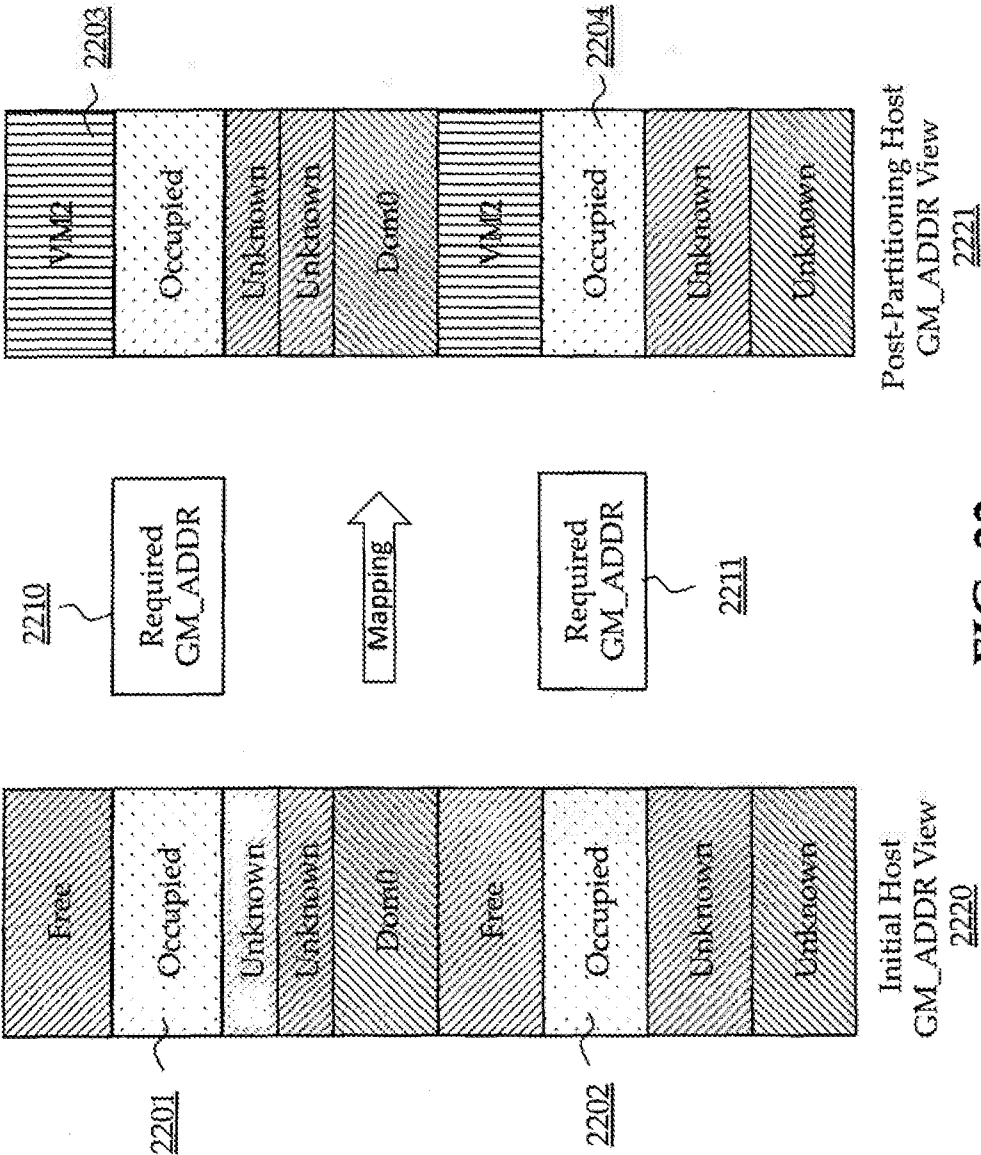


FIG. 22

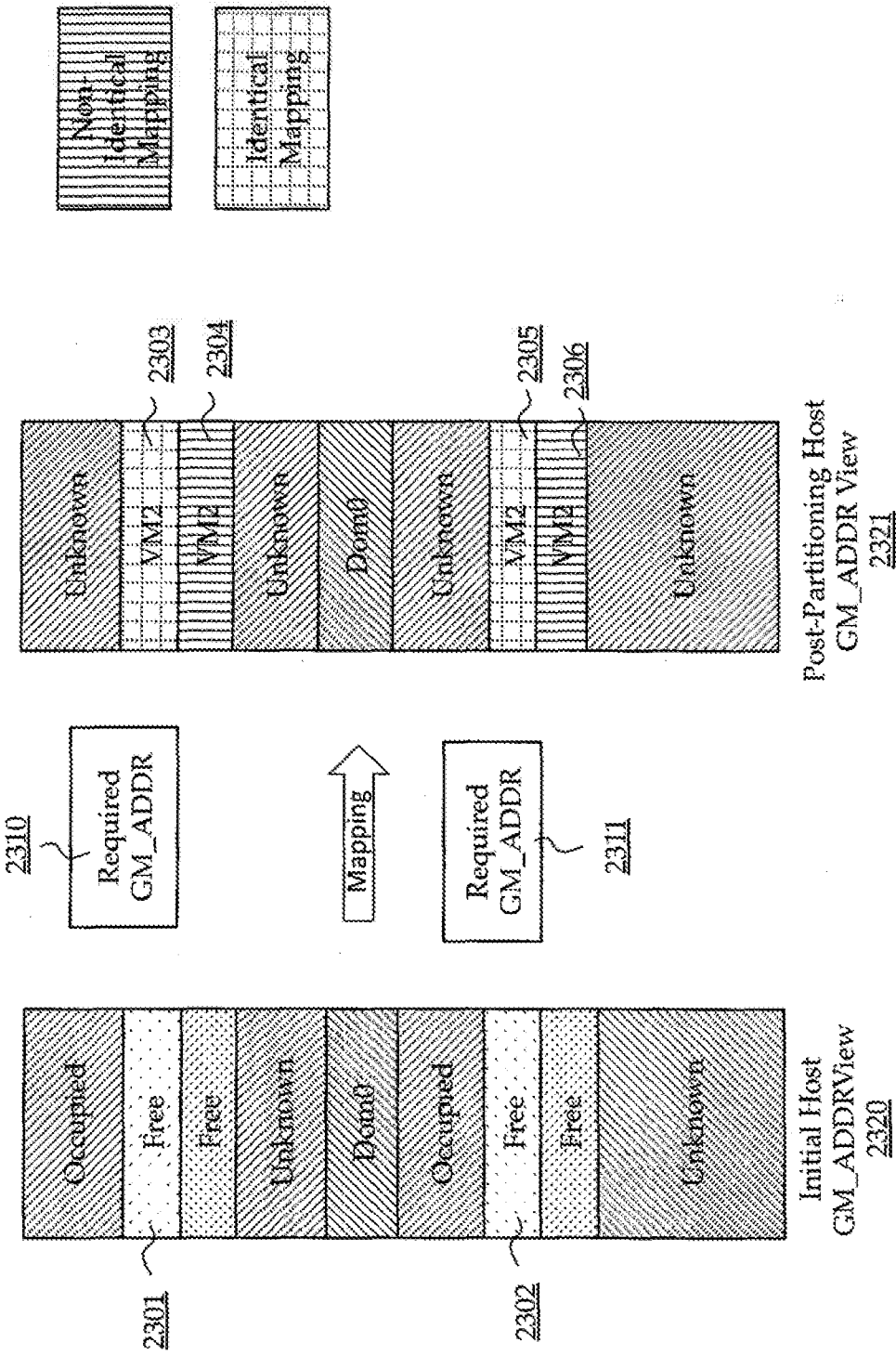


FIG. 23

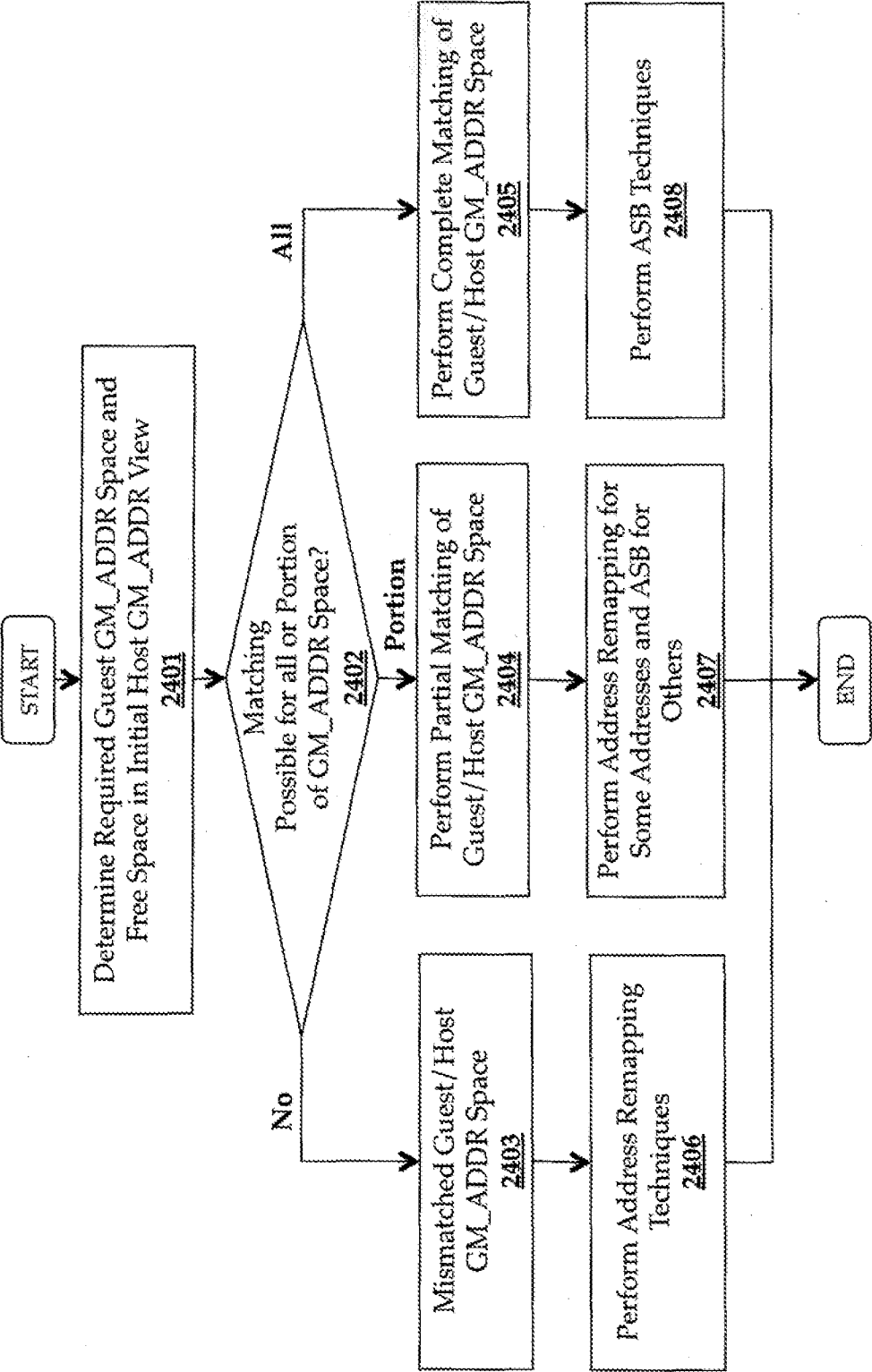


FIG. 24

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2014/094663

A. CLASSIFICATION OF SUBJECT MATTER

G06F 12/02(2006.01)i; G06F 9/455(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)

G06F

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

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

WPI, EPODOC, CNKI, CNPAT, IEEE, GOOGLE: virtual, machine, memory, allocate, dynamic, address, map, match, balloon

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	TIAN, Kun et al., "A Full GPU Virtualization Solution with Mediated Pass-Through," <i>Proceedings of USENIX ATC' 14: 2014 USENIX Annual Technical Conference</i> , Vol. 1, No. 1, 20 June 2014 (2014-06-20), ISSN: /, page 122, left column, paragraph 2 to page 123, left column, paragraph 1, page 128, left column, paragraph 5 and figures 5-7	1-19, 21-23
A	US 7702843 B1 (VMWARE, INC.) 20 April 2010 (2010-04-20) the whole document	1-19, 21-23
A	CN 103389945 A (HANGZHOU H3C TECHNOLOGIES CO., LTD.) 13 November 2013 (2013-11-13) the whole document	1-19, 21-23
A	CN 104111896 A (UNIV. YUNNAN) 22 October 2014 (2014-10-22) the whole document	1-19, 21-23
A	US 2013339568 A1 (VMWARE, INC.) 19 December 2013 (2013-12-19) the whole document	1-19, 21-23



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

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

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

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

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

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

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

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

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

"&" document member of the same patent family

Date of the actual completion of the international search

01 September 2015

Date of mailing of the international search report

23 September 2015

Name and mailing address of the ISA/CN

STATE INTELLECTUAL PROPERTY OFFICE OF THE
P.R.CHINA
6, Xitucheng Rd., Jimen Bridge, Haidian District, Beijing
100088, China

Facsimile No. (86-10)62019451

Authorized officer

JIANG, Xiaoqing

Telephone No. (86-10)61648093

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2014/094663

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
US	7702843	B1	20 April 2010	US 8095931 B1	10 January 2012
				US 2013132957 A1	23 May 2013
				US 2013145377 A1	06 June 2013
				US 2012110577 A1	03 May 2012
				US 7716446 B1	11 May 2010
				US 2014189195 A1	03 July 2014
				US 2010241785 A1	23 September 2010
				US 2011320682 A1	29 December 2011
CN	103389945	A	13 November 2013	None	
CN	104111896	A	22 October 2014	None	
US	2013339568	A1	19 December 2013	None	