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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
07 December 2017 (07.12.2017)

WIPO | PCT

(10) International Publication Number
WO 2017/206156 A1

(51) International Patent Classification:
G06K 9/66 (2006.01)

(21) International Application Number:
PCT/CN2016/084621

(22) International Filing Date:
03 June 2016 (03.06.2016)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant: **INTEL CORPORATION** [US/US]; 2200 Mission College Blvd., Santa Clara, California 95054 (US).

(72) Inventors; and

(71) Applicants *(for BZ only)*: **MA, Liwei** [CN/CN]; 8F Raycom A No. 2, Kexueyuan South Road, Haidian District, Beijing, 100080 (CN). **SONG, Jiqiang** [CN/CN]; 8F, Raycom Infotech Park A, No. 2, Kexueyuan South Road, Zhongguancun, Haidian District, Beijing 100190 (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))*

WO 2017/206156 A1

(54) **Title:** LOOK-UP CONVOLUTIONAL LAYER IN CONVOLUTIONAL NEURAL NETWORK

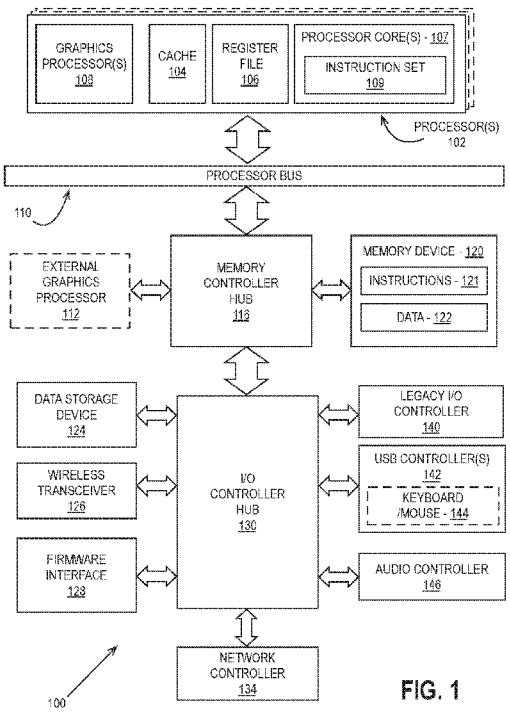


FIG. 1

(57) **Abstract:** Embodiments provide for a processor including logic to accelerate convolutional neural network processing, the processor including first logic to apply a convolutional layer to an image to generate a first convolution result and second logic to apply a look-up convolutional layer to the first convolution result to generate a second convolution result, the second convolution result associated with a location of the first convolution result within a global filter kernel.

LOOK-UP CONVOLUTIONAL LAYER IN CONVOLUTIONAL NEURAL NETWORK

TECHNICAL FIELD

Embodiments generally relate to image processing logic. More particularly, embodiments relate to image processing logic to perform autonomous localization using a convolutional neural network.

BACKGROUND

The ability to program machines perform visual recognition has significant implications in science, technology, and commerce. One technique used by computer vision algorithms is the convolutional neural network (CNN). CNNs are known in the art for many computer vision tasks, such as object classification and classification by inputting the raw RGB image. Researchers are applying CNN algorithms to many other applications, such as predicting camera pose (e.g., position and orientation) within an environment. Camera pose prediction can be used in many circumstances, such as robot localization, virtual reality games, and smart glasses. The traditional classification CNNs consist of several convolutional layers that extract local texture features and several full-connection layers that extract global hierarchical features. . In traditional convolutional neural networks, convolutional layers are used to reduce the input feature map, particularly in circumstances in which the size of input image is considered to be too large.

BRIEF DESCRIPTION OF THE DRAWINGS

The various advantages of the embodiments will become apparent to one skilled in the art by reading the following specification and appended claims, and by referencing the following drawings, in which:

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;

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;

Fig. 3 is a block diagram of one embodiment of a graphics processor which may be a discreet graphics processing unit, or may be graphics processor integrated with a plurality of processing cores;

Fig. 4 is a block diagram of an embodiment of a graphics processing engine for a graphics processor;

Fig. 5 is a block diagram of another embodiment of a graphics processor;

Fig. 6 is a block diagram of thread execution logic including an array of processing

elements;

Fig. 7 illustrates a graphics processor execution unit instruction format according to an embodiment;

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;

Fig. 9A is a block diagram illustrating a graphics processor command format according to an embodiment;

Fig. 9B is a block diagram illustrating a graphics processor command sequence according to an embodiment;

Fig. 10 illustrates exemplary graphics software architecture for a data processing system according to an embodiment;

Fig. 11 is a block diagram illustrating an IP core development system that may be used to manufacture an integrated circuit to perform operations according to an embodiment;

Fig. 12 is a block diagram illustrating an exemplary system on a chip integrated circuit that may be fabricated using one or more IP cores, according to an embodiment;

Fig. 13 is a block diagram illustrating an exemplary graphics processor of a system on a chip integrated circuit;

Fig. 14 is a block diagram illustrating an additional exemplary graphics processor of a system on a chip integrated circuit;

Fig. 15 is an illustration of primitive elements of a convolutional neural network;

Fig. 16A-B are illustrations of a conventional implementation of a convolutional neural network;

Fig. 17A-B illustrate a look-up convolutional neural network, according to embodiments described herein;

Fig. 18 illustrates a comparison between convolution and look-up convolution operations, according to an embodiment;

Fig. 19A-B illustrates look-up convolution kernels for camera pose estimation, according to embodiments;

Fig. 20 is a flow diagram of look-up convolution layer logic, according to an embodiment;

Fig. 21 is a flow diagram of camera pose estimation logic, according to an embodiment;

Fig. 22 is a block diagram of a graphics core including logic to perform operations associated with embodiments described herein; and

Fig. 23 is a block diagram of a computing device including logic to accelerate neural network processing for operations associated with embodiments described herein.

DESCRIPTION OF EMBODIMENTS

Embodiments described herein provide for a technique of using image and depth data to predict camera position and orientation within an environment, although some embodiments have wider application. In embodiments described herein, a bottleneck convolutional layer is used, followed by a look-up convolutional layer which expands the bottleneck layer to larger layer, which outputs a sparser set of features for input into the next layer. Embodiments described herein can be used as a general design technique in localization applications. Key aspects of the embodiments include not only the look-up convolutional layer described herein but also how these look-up convolutional layers are integrated to a convolutional neural network.

For the purposes of explanation, numerous specific details are set forth to provide a thorough understanding of the various embodiments described below. However, it will be apparent to a skilled practitioner in the art that the embodiments 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, and to provide a more thorough understanding of embodiments. Although some of the following embodiments are described with reference to a graphics processor, the techniques and teachings described herein may be applied to various types of circuits or semiconductor devices, including general purpose processing devices or graphic processing devices. Reference herein to "one embodiment" or "an embodiment" indicate that a particular feature, structure, or characteristic described in connection or association with the embodiment can be included in at least one of such embodiments. However, the appearances of the phrase "in one embodiment" in various places in the specification do not necessarily all refer to the same embodiment.

In the following description and claims, the terms "coupled" and "connected," along with their derivatives, may be used. It should be understood that these terms are not intended as synonyms for each other. "Coupled" is used to indicate that two or more elements, which may or may not be in direct physical or electrical contact with each other, co-operate or interact with each other. "Connected" is used to indicate the establishment of communication between two or more elements that are coupled with each other.

In the description that follows, Figures 1-14 provide an overview of exemplary data processing system and graphics processor logic that incorporates or relates to the various embodiments. Figures 14-23 provide specific details of the various embodiments. Although some of the following embodiments are described with reference to a graphics processor, similar techniques and teachings can be applied to other types of circuits or semiconductor devices, as the teachings are applicable to any processor or machine that manipulates or processes image data.

System Overview

Fig. 1 is a block diagram of a processing system 100, according to an embodiment. In various embodiments the 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 system 100 is a processing platform incorporated within a system-on-a-chip (SoC) integrated circuit for use in mobile, handheld, or embedded devices.

An embodiment of 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 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.

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 a Digital Signal Processor (DSP).

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.

In some embodiments, processor 102 is coupled with a processor bus 110 to transmit communication signals such as address, data, or control signals between processor 102 and other components in system 100. In one embodiment the system 100 uses an exemplary 'hub' system architecture, including a memory controller hub 116 and an Input Output (I/O) controller hub 130. A memory controller hub 116 facilitates communication between a memory device and other components of system 100, while an I/O Controller Hub (ICH) 130 provides connections to I/O devices via a local I/O bus. In one embodiment, the logic of the memory controller hub 116 is integrated within the processor.

Memory device 120 can be a dynamic random access memory (DRAM) device, a static random access memory (SRAM) device, flash memory device, phase-change memory device, or some other memory device having suitable performance to serve as process memory. In one embodiment the memory device 120 can operate as system memory for the system 100, to store data 122 and instructions 121 for use when the one or more processors 102 executes an application or 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.

In some embodiments, ICH 130 enables peripherals to connect to memory device 120 and processor 102 via a high-speed I/O bus. The I/O peripherals include, but are not limited to, 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 140 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 with ICH 130. In some embodiments, a high-performance network controller (not shown) couples with processor bus 110. It will be appreciated that the system 100 shown is exemplary and not limiting, as other types of data processing systems that are differently configured may also be used. For example, the I/O controller hub 130 may be integrated within the one or more processor 102, or the memory controller hub 116 and I/O controller hub 130 may be integrated into a discreet external graphics processor, such as the external graphics processor 112.

Fig. 2 is a block diagram of an embodiment of a processor 200 having one or more processor cores 202A-202N, an integrated memory controller 214, and an integrated graphics processor 208. Those elements of **Fig. 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

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

The internal cache units 204A-204N 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 processor 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-204N.

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

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

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 core 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 a separate module coupled with the graphics processor via at least one interconnect, or may be integrated within the graphics processor 208 or system agent core 210.

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.

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 processor cores 202A-202N and graphics processor 208 use embedded memory modules 218 as a shared Last Level Cache.

In some embodiments, processor cores 202A-202N are homogenous cores executing the same instruction set architecture. In another embodiment, processor cores 202A-202N are heterogeneous in terms of instruction set architecture (ISA), where one or more of processor cores 202A-202N 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. In one embodiment processor cores 202A-202N are heterogeneous in terms of microarchitecture, where one or more cores having a relatively higher power consumption couple with one or more power cores having a lower power consumption. 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.

Fig. 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.

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.

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 graphics processing engine (GPE) 310. In some embodiments, GPE 310 is a compute engine for performing graphics operations, including three-dimensional

(3D) graphics operations and media operations.

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.

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.

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.

Graphics Processing Engine

Fig. 4 is a block diagram of a graphics processing engine 410 of a graphics processor in accordance with some embodiments. In one embodiment, the graphics processing engine (GPE) 410 is a version of the GPE 310 shown in **Fig. 3**. Elements of **Fig. 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. For example, the 3D pipeline 312 and media pipeline 316 of **Fig. 3** are illustrated. The media pipeline 316 is optional in some embodiments of the GPE 410 and may not be explicitly included within the GPE 410. For example and in at least one embodiment, a separate media and/or image processor is coupled to the GPE 410.

In some embodiments, GPE 410 couples with or includes a command streamer 403, which provides a command stream to the 3D pipeline 312 and/or media pipelines 316. In some embodiments, command streamer 403 is coupled with 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 312 and/or media pipeline 316. The commands are directives fetched from a ring buffer, which stores commands for the 3D pipeline 312 and media pipeline 316. In one embodiment, the ring buffer can additionally include batch command buffers storing batches of multiple commands. The commands for the 3D pipeline 312 can also include references to data stored in memory, such as but not limited to vertex and geometry data for the 3D pipeline 312 and/or image data and memory objects for the media pipeline 316. The 3D pipeline 312 and media pipeline 316 process the commands and data by performing operations via logic within the respective pipelines or by dispatching one or more execution threads to a graphics core array 414.

In various embodiments the 3D pipeline 312 can execute one or more shader programs, such as vertex shaders, geometry shaders, pixel shaders, fragment shaders, compute shaders, or other shader programs, by processing the instructions and dispatching execution threads to the graphics core array 414. The graphics core array 414 provides a unified block of execution resources. Multi-purpose execution logic (e.g., execution units) within the graphic core array 414 includes support for various 3D API shader languages and can execute multiple simultaneous execution threads associated with multiple shaders.

In some embodiments the graphics core array 414 also includes execution logic to perform media functions, such as video and/or image processing. In one embodiment, the execution units additionally include general-purpose logic that is programmable to perform parallel general purpose computational operations, in addition to graphics processing operations. The general purpose logic can perform processing operations in parallel or in conjunction with general purpose logic within the processor core(s) 107 of **Fig. 1** or core 202A-202N as in **Fig. 2.**

Output data generated by threads executing on the graphics core array 414 can output data to memory in a unified return buffer (URB) 418. The URB 418 can store data for multiple threads. In some embodiments the URB 418 may be used to send data between different threads executing on the graphics core array 414. In some embodiments the URB 418 may additionally be used for synchronization between threads on the graphics core array and fixed function logic within the shared function logic 420.

In some embodiments, graphics core array 414 is scalable, such that the array includes a variable number of graphics cores, each having a variable number of execution units based on

the target power and performance level of GPE 410. In one embodiment the execution resources are dynamically scalable, such that execution resources may be enabled or disabled as needed.

The graphics core array 414 couples with shared function logic 420 that includes multiple resources that are shared between the graphics cores in the graphics core array. The shared functions within the shared function logic 420 are hardware logic units that provide specialized supplemental functionality to the graphics core array 414. In various embodiments, shared function logic 420 includes but is not limited to sampler 421, math 422, and inter-thread communication (ITC) 423 logic. Additionally, some embodiments implement one or more cache(s) 425 within the shared function logic 420. A shared function is implemented where the demand for a given specialized function is insufficient for inclusion within the graphics core array 414. Instead a single instantiation of that specialized function is implemented as a stand-alone entity in the shared function logic 420 and shared among the execution resources within the graphics core array 414. The precise set of functions that are shared between the graphics core array 414 and included within the graphics core array 414 varies between embodiments.

Fig. 5 is a block diagram of another embodiment of a graphics processor 500. Elements of **Fig. 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.

In some embodiments, graphics processor 500 includes a ring interconnect 502, a pipeline front-end 504, a media engine 537, and graphics cores 580A-580N. 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.

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-580N. 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.

In some embodiments, graphics processor 500 includes scalable thread execution resources featuring modular cores 580A-580N (sometimes referred to as core slices), each having multiple sub-cores 550A-550N, 560A-560N (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-580N, each including a set of first sub-cores 550A-550N and a set of second sub-cores 560A-560N. Each sub-core in the set of first sub-cores 550A-550N includes at least a first set of execution units 552A-552N and media/texture samplers 554A-554N. Each sub-core in the set of second sub-cores 560A-560N includes at least a second set of execution units 562A-562N and samplers 564A-564N. In some embodiments, each sub-core 550A-550N, 560A-560N shares a set of shared resources 570A-570N. 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.

Execution Units

Fig. 6 illustrates thread execution logic 600 including an array of processing elements employed in some embodiments of a GPE. Elements of **Fig. 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.

In some embodiments, thread execution logic 600 includes a shader processor 602, a thread dispatcher 604, instruction cache 606, a scalable execution unit array including a plurality of execution units 608A-608N, 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 units 608A-608N. In some embodiments, each execution unit (e.g. 608A) is a stand-alone programmable general purpose computational unit that is capable of executing multiple simultaneous hardware threads while processing multiple data elements in parallel for each thread. In various embodiments, the array of execution units 608A-608N is scalable to include any number individual execution units.

In some embodiments, the execution units 608A-608N are primarily used to execute shader programs. A shader processor 602 can process the various shader programs and dispatch execution threads associated with the shader programs via a thread dispatcher 604. In one

embodiment the thread dispatcher includes logic to arbitrate thread initiation requests from the graphics and media pipelines and instantiate the requested threads on one or more execution unit in the execution units 608A-608N. For example, the geometry pipeline (e.g., 536 of **Fig. 5**) can dispatch vertex, tessellation, or geometry shaders to the thread execution logic 600 (**Fig. 6**) for processing. In some embodiments, thread dispatcher 604 can also process runtime thread spawning requests from the executing shader programs.

In some embodiments, the execution units 608A-608N support 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). Each of the execution units 608A-608N is capable of multi-issue single instruction multiple data (SIMD) execution and multi-threaded operation enables an efficient execution environment in the face of higher latency memory accesses. Each hardware thread within each execution unit has a dedicated high-bandwidth register file and associated independent thread-state. Execution is multi-issue per clock to pipelines capable of integer, single and double precision floating point operations, SIMD branch capability, logical operations, transcendental operations, and other miscellaneous operations. While waiting for data from memory or one of the shared functions, dependency logic within the execution units 608A-608N causes a waiting thread to sleep until the requested data has been returned. While the waiting thread is sleeping, hardware resources may be devoted to processing other threads. For example, during a delay associated with a vertex shader operation, an execution unit can perform operations for a pixel shader, fragment shader, or another type of shader program, including a different vertex shader.

Each execution unit in execution units 608A-608N 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-608N support integer and floating-point data types.

The execution unit instruction set includes 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.

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, a 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.

During execution, the graphics and media pipelines send thread initiation requests to thread execution logic 600 via thread spawning and dispatch logic. Once a group of geometric objects has been processed and rasterized into pixel data, pixel processor logic (e.g., pixel shader logic, fragment shader logic, etc.) within the shader processor 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, a pixel shader or fragment shader calculates the values of the various vertex attributes that are to be interpolated across the rasterized object. In some embodiments, pixel processor logic within the shader processor 602 then executes an application programming interface (API)-supplied pixel or fragment shader program. To execute the shader program, the shader processor 602 dispatches threads to an execution unit (e.g., 608A) via thread dispatcher 604. In some embodiments, pixel shader 602 uses texture sampling logic in the 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.

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.

Fig. 7 is a block diagram illustrating a graphics processor instruction formats 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-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.

In some embodiments, the graphics processor execution units natively support instructions in a 128-bit instruction 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. The native 128-bit instruction 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 a native instruction in the 128-bit instruction format 710.

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 or picture element. By default, the execution unit performs each instruction across all data channels of the operands. In some embodiments, instruction control field 714 enables control over certain execution options, such as channels selection (e.g., predication) and data channel order (e.g., swizzle). For instructions in the 128-bit instruction format 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.

Some execution unit instructions have up to three operands including two source operands, src0 720, 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.

In some embodiments, the 128-bit instruction format 710 includes an access/address mode field 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.

In some embodiments, the 128-bit instruction format 710 includes an access/address mode field 726, which specifies an address mode and/or an access mode for the instruction. In one embodiment the access mode is used to define a data access alignment for the instruction. Some

embodiments support access modes including a 16-byte aligned access mode and a 1-byte aligned access mode, where the byte alignment of the access mode determines the access alignment of the instruction operands. For example, when in a first mode, the instruction may use byte-aligned addressing for source and destination operands and when in a second mode, the instruction may use 16-byte-aligned addressing for all source and destination operands.

In one embodiment, the address mode portion of the access/address mode field 726 determines whether the instruction is to use direct or indirect addressing. When direct register addressing mode is used bits in the instruction directly provide the register address of one or more operands. When indirect register addressing mode is used, the register address of one or more operands may be computed based on an address register value and an address immediate field in the instruction.

In some embodiments instructions are grouped based on opcode 712 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 and logic instructions are in the form of 0001xxxxb. 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.

Graphics Pipeline

Fig 8 is a block diagram of another embodiment of a graphics processor 800. Elements of **Fig. 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.

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.

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.

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.

In some embodiments, graphics pipeline 820 includes tessellation components to perform hardware-accelerated tessellation of 3D objects. In some embodiments, a 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 (e.g., hull shader 811, tessellator 813, and domain shader 817) can be bypassed.

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.

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 and depth test component 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 and depth test component 873 and access un-rasterized vertex data via a stream out unit 823.

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.

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. An associated render cache 878 and depth cache 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.

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 837 includes thread spawning functionality to spawn threads for dispatch to thread execution logic 850 via thread dispatcher 831.

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.

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), Open Computing Language (OpenCL), and/or Vulkan graphics and compute API, all from the Khronos Group. In some embodiments, support may also be provided for the Direct3D library from the Microsoft Corporation. In some embodiments, a combination of these libraries may be supported. 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.

Graphics Pipeline Programming

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 **Fig. 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 **Fig. 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.

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.

The flow diagram in **Fig. 9B** shows an exemplary graphics processor 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.

In some embodiments, the graphics processor 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.

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.

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.

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.

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.

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.

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.

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 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 commands to control pixel shading and pixel back end operations may also be included for those operations.

In some embodiments, the graphics processor 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 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 used to perform SIMD vector operations using computational shader programs that are not explicitly related to the rendering of graphics primitives.

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 a command queue before the media object commands 942. In some embodiments, media 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 of one or more pointers to "indirect" state elements that contain a batch of state settings.

In some embodiments, media object commands 942 supply pointers to media 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 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.

Graphics Software Architecture

Fig. 10 illustrates exemplary graphics software architecture for a data processing system 1000 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 operating system 1020 each execute in the system memory 1050 of the data processing system.

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 machine language suitable for execution by the general-purpose processor core 1034. The application also includes graphics objects 1016 defined by vertex data.

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. The operating system 1020 can support a graphics API 1022 such as the Direct3D API, the OpenGL API, or the Vulkan API. When the 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. In some embodiments, the shader instructions 1012 are provided in an intermediate form, such as a version of the Standard Portable Intermediate Representation (SPIR) used by the Vulkan API.

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 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.

IP Core Implementations

One or more aspects of at least one embodiment may be implemented by representative code stored on a machine-readable medium which represents and/or defines logic within an integrated circuit such as a processor. For example, the machine-readable medium may include instructions which represent various logic within the processor. When read by a machine, the instructions may cause the machine to fabricate the logic to perform the techniques described herein. Such representations, known as "IP cores," are reusable units of logic for an integrated circuit that may be stored on a tangible, machine-readable medium as a hardware model that describes the structure of the integrated circuit. The hardware model may be supplied to various customers or manufacturing facilities, which load the hardware model on fabrication machines that manufacture the integrated circuit. The integrated circuit may be fabricated such that the circuit performs operations described in association with any of the embodiments described herein.

Fig. 11 is a block diagram illustrating an IP core development system 1100 that may be used to manufacture an integrated circuit to perform operations according to an embodiment. The IP core development system 1100 may be used to generate modular, re-usable designs that can be incorporated into a larger design or used to construct an entire integrated circuit (e.g., an

SOC integrated circuit). A design facility 1130 can generate a software simulation 1110 of an IP core design in a high level programming language (e.g., C/C++). The software simulation 1110 can be used to design, test, and verify the behavior of the IP core using a simulation model 1112. The simulation model 1112 may include functional, behavioral, and/or timing simulations. A register transfer level (RTL) design can then be created or synthesized from the simulation model 1112. The RTL design 1115 is an abstraction of the behavior of the integrated circuit that models the flow of digital signals between hardware registers, including the associated logic performed using the modeled digital signals. In addition to an RTL design 1115, lower-level designs at the logic level or transistor level may also be created, designed, or synthesized. Thus, the particular details of the initial design and simulation may vary.

The RTL design 1115 or equivalent may be further synthesized by the design facility into a hardware model 1120, which may be in a hardware description language (HDL), or some other representation of physical design data. The HDL may be further simulated or tested to verify the IP core design. The IP core design can be stored for delivery to a 3rd party fabrication facility 1165 using non-volatile memory 1140 (e.g., hard disk, flash memory, or any non-volatile storage medium). Alternatively, the IP core design may be transmitted (e.g., via the Internet) over a wired connection 1150 or wireless connection 1160. The fabrication facility 1165 may then fabricate an integrated circuit that is based at least in part on the IP core design. The fabricated integrated circuit can be configured to perform operations in accordance with at least one embodiment described herein.

Exemplary System on a Chip Integrated Circuits

Figs 12-14 illustrated exemplary integrated circuits and associated graphics processors that may be fabricated using one or more IP cores, according to various embodiments described herein. In addition to what is illustrated, other logic and circuits may be included, including additional graphics processors/cores, peripheral interface controllers, or general purpose processor cores.

Fig. 12 is a block diagram illustrating an exemplary system on a chip integrated circuit 1200 that may be fabricated using one or more IP cores, according to an embodiment. Exemplary integrated circuit 1200 includes one or more application processor(s) 1205 (e.g., CPUs), at least one graphics processor 1210, and may additionally include an image processor 1215 and/or a video processor 1220, any of which may be a modular IP core from the same or multiple different design facilities. Integrated circuit 1200 includes peripheral or bus logic including a USB controller 1225, UART controller 1230, an SPI/SDIO controller 1235, and an I^2S/I^2C controller 1240. Additionally, the integrated circuit can include a display device 1245 coupled to one or more of a high-definition multimedia interface (HDMI) controller 1250 and a

mobile industry processor interface (MIPI) display interface 1255. Storage may be provided by a flash memory subsystem 1260 including flash memory and a flash memory controller. Memory interface may be provided via a memory controller 1265 for access to SDRAM or SRAM memory devices. Some integrated circuits additionally include an embedded security engine 1270.

Fig. 13 is a block diagram illustrating an exemplary graphics processor 1310 of a system on a chip integrated circuit that may be fabricated using one or more IP cores, according to an embodiment. Graphics processor 1310 can be a variant of the graphics processor 1210 of **Fig. 12**. Graphics processor 1310 includes a vertex processor 1305 and one or more fragment processor(s) 1315A-1315N. Graphics processor 1310 can execute different shader programs via separate logic, such that the vertex processor 1305 is optimized to execute operations for vertex shader programs, while the one or more fragment processor(s) 1315A-1315N execute fragment (e.g., pixel) shading operations for fragment or pixel shader programs. The vertex processor 1305 performs the vertex processing stage of the 3D graphics pipeline and generates primitives and vertex data. The fragment processor(s) 1315A-1315N use the primitive and vertex data generated by the vertex processor 1305 to produce a framebuffer that is displayed on a display device. In one embodiment, the fragment processor(s) 1315A-1315N are optimized to execute fragment shader programs as provided for in the OpenGL API, which may be used to perform similar operations as a pixel shader program as provided for in the Direct 3D API.

Graphics processor 1310 additionally includes one or more memory management units (MMUs) 1320A-1320B, cache(s) 1325A-1325B, and circuit interconnect(s) 1330A-1330B. The one or more MMU(s) 1320A-1320B provide for virtual to physical address mapping for integrated circuit 1300, including for the vertex processor 1305 and/or fragment processor(s) 1315A-1315N, which may reference vertex or image/texture data stored in memory, in addition to vertex or image/texture data stored in the one or more cache(s) 1325A-1325B. In one embodiment the one or more MMU(s) 1325A-1325B may be synchronized with other MMUs within the system, including one or more MMUs associated with the one or more application processor(s) 1205, image processor 1215, and/or video processor 1220 of **Fig. 12**, such that each processor 1205-1220 can participate in a shared or unified virtual memory system. The one or more circuit interconnect(s) 1330A-1330B enable graphics processor 1310 to interface with other IP cores within the SoC, either via an internal bus of the SoC or via a direct connection, according to embodiments.

Fig. 14 is a block diagram illustrating an additional exemplary graphics processor 1410 of a system on a chip integrated circuit that may be fabricated using one or more IP cores, according to an embodiment. Graphics processor 1410 can be a variant of the graphics processor

1210 of **Fig. 12**. Graphics processor 1410 includes the one or more MMU(s) 1320A-1320B, caches 1325A-1325B, and circuit interconnects 1330A-1330B of the integrated circuit 1300 of Fig. 13.

Graphics processor 1410 includes one or more shader core(s) 1415A-1415N, which provides for a unified shader core architecture in which a single core or type or core can execute all types of programmable shader code, including vertex shaders, fragment shaders, and compute shaders. The exact number of shader cores present can vary among embodiments and implementations. Additionally, graphics processor 1410 includes an inter-core task manager 1405, which acts as a thread dispatcher to dispatch execution threads to one or more shader cores 1415A-1415N and a tiling unit 1418 to accelerate tiling operations for tile-based rendering, in which rendering operations for a scene are subdivided in image space, for example to exploit local spatial coherence within a scene or to optimize use of internal caches.

Look-Up Convolutional Layer In A Convolutional Neural Network

Embodiments described herein provide for a technique of using image and depth data to predict camera position and orientation within an environment, although some embodiments have wider application. One embodiment may be utilized as a general design technique for use in autonomous localization applications, including the look-up convolutional layer described herein and the integration of the look-up convolutional layer into a convolutional neural network for image processing and autonomous localization.

Convolutional Neural Networks

Fig. 15 is an illustration of primitive elements of a convolutional neural network. A convolutional neural network (CNN) includes a number of convolution and subsampling layers optionally followed by one or more fully connected layers. The convolutional layers are the core building blocks of a CNN, and is where the majority of the computational operations of the CNN are performed. Within a convolutional layer, an original image 1502 having some data to be analyzed is processed by a set of convolution kernels that apply each apply a different filter 1504A, 1504B to the original image 1502. The filters 1504A, 1504B are learnable and typically much smaller than the original image to which the filters will be applied. The convolution kernels output a set of feature maps 1506A, 1506B that contain the features searched for by the convolution kernels. Results of the filter operations may be summed together to provide an output from the convolutional layer **902** to a subsequent layer, such as a pooling layer or a fully connected neural network.

Fig. 16A-B are illustrations of a conventional implementation of a convolutional neural network 1600. The convolutional neural network 1600 illustrated in Fig. 16A analyzes red, green, and blue (RGB) components 1602 of an image. In general, a CNN can be used to analyze

an m x m x r image, where m is the height and width of the image and r is the number of channels. For example, an RGB image has r=3 channels. The convolutional layers 1604A, 1604B can have k filters (or kernels) of size n x n x q, where n is smaller than the dimension of the image and q can either be the same as the number of channels r or smaller, and may vary for each kernel. The size of the filters gives rise to the locally connected structure which are each convolved with the image to produce k feature maps of size m–n+1. Some implementations use convolution kernels in pixel sizes of 11*11, 9*9, 7*7, 5*5, 3*3 and/or 1*1, which are each much smaller than the input images or feature maps sizes.

Fig. 16B illustrates a conventional CNN implementation, in which the size of input image or feature maps can be defined as (m*m), the convolution kernel size is defined as (n*n), producing a convolution result size of (m - n+1)*(m – n + 1) without padding zeros rounding the input, where m is significantly greater than n (e.g., (m >> n)). In some implementations the convolution result can be further reduced using down-sampling or pooling layers, which can shrink the original input image to multiple feature maps with sizes of less than 7*7 pixels, as in some instances a highly compact feature map can be more information intensive.

While the conventional CNN of Fig. 16A-B may be suitable for computer vision operations such as image classification, there are differences between the image classification use case and other computer vision use cases, such as camera pose prediction.

In image classification, the objective of the classification is to find a subset of the input image and to identity the name of the object in the image subset. For pose prediction, the objective is to find the input image as a subset of a larger environment to identify the input image pose within the environment. The classification CNN uses small convolution kernels, skip step and max-pooling techniques to shrink the original image to a series of small feature maps. The classification CNN may then use a full-connection layer to perform the actual classification.

In the instance of camera pose prediction, the series of small feature map should be matched in a larger environment filter to perform localization operations, such as for use in autonomous robot or autonomous vehicle localization in which the robot or vehicle can construct or use a navigational map or floor plan to localize itself within an environment. For such implementation, it may be advantageous to use one or more look-up convolutional layers to represent environment filters, which are larger convolution kernels that may be significantly larger than the input feature maps.

Using look-up convolutional layers it is possible to match input feature maps in a series of larger environment filters and to produce larger feature maps, which are more descriptive to location and pose parameters used in localization and pose prediction.

Fig. 17A-B illustrate a look-up convolutional neural network 1700, according to embodiments described herein. The convolutional neural network 1700 illustrated in Fig. 17A can analyze RGB components 1702 of an input image using both conventional convolutional layers 1704A, 1704B and at least one look-up convolutional layer 1705. Unlike conventional convolutional layers 1704A, 1704B, which shrink the input feature maps to more compact feature maps, the look-up convolutional layer 1705 enlarges the input feature maps to wider feature maps. The output of the look-up convolutional layer 1705 can be input into a set of fully connected layers 1706 of a fully connected neural network to produce scores used to perform the camera pose prediction.

Fig. 17B illustrates the size of input image or feature maps, where an input image or feature map is defined as ($m * m$), where m is the height and width of the image the convolution kernel size ($n * n$) and the result size ($n - m + 1$) * ($n - m + 1$) without padding zeros rounding the input, where m is significantly smaller than n (e.g., ($m << n$)). The larger convolution kernel represents the environment filter maps and the convolution operation will match the current feature map to the environment filters to produce an output feature map. The optimized ratio for n to m can be determined by the ratio of the environment panorama frame to the camera frame.

Fig. 18 illustrates a comparison 1800 between convolution and look-up convolution operations, according to an embodiment. A convolutional layer 1810 in a CNN of one embodiment is applies a convolution kernel 1804 to a feature map 1802 to compress the feature map 1802 into an information dense convolution result 1806. The look-up convolution layer 1820 of such embodiment is used to determine a location of a dense feature map 1812 within a global filter kernel (e.g., look-up convolution kernel 1814), to produce a convolution result 1816 that is an expansion of the feature map 1812 that is input into the look-up convolution kernel 1814.

Fig. 19A-B illustrates look-up convolution kernels for camera pose estimation 1900, according to embodiments. In one embodiment the look-up convolution kernel and associated environment maps are mapped to a shape defined by a view angle of an imaging module used to capture the image used for localization and/or camera pose estimation.

Fig. 19A illustrates a cylindrical convolution kernel 1914 for a convolution look-up layer of one embodiment. A cylindrical convolution kernel 1914 used in one embodiment to represent a possible view for a robot in a room, where the camera sweep can form a closed loop around the robot. The robot camera maintains a constant horizontal angle with only minor variations in the vertical angle. The robot can move and/or turn around in a complete 360 degree sweep. In this configuration, the robot's view form a cylindrical shape, thus the use of the cylindrical

convolution kernel 1914. Applying the cylindrical convolution kernel 1914 to an $m * m$ feature map 1912 produces a cylinder shaped convolution result 1916 of size $(n1 - m + 1) * n2$.

Fig. 19B illustrates a spherical convolution kernel 1924 for a convolution look-up layer of one embodiment. In such embodiment, the possible view for a robot in a room can forms a sphere, as the robot's optical sensor or camera can vary across large vertical ranges or can rotate while moving vertically. The robot itself can also move or turn around in large angles. To localize the robot within an environment, a spherical convolution kernel 1924 can be applied to a feature map 1922 to produce a spherical convolution result 1926. Multiple spherical convolution kernels 1924 can be used to create a spherical convolution look-up layer to represent the environment layers used to localization. In one embodiment, where the feature map 1922 is convoluted by the spherical convolution kernel 1924, some interpolation may be needed need to fit the feature map 1922 to the spherical surface represented by the spherical convolution kernel 1924.

While embodiments utilizing square ($m * m$) feature maps are illustrated thus far, a feature map can also be rectangular ($m1 * m2$) or one dimensional ($1 * m2$, $m1=1$). In the case of a one dimensional feature map, a one dimensional convolution kernel can also be used. Accordingly, a convolution kernel can be defined where $n1 = n2$; $n1 != n2$; or $n1=1$.

Additionally, embodiments generally utilize multiple convolutional kernels in one layer, both for conventional convolutional layers and look-up convolutional layers. For each type of look-up convolutional kernels, the actual number of kernel in one layer is great than one, usually tens or hundreds of kernels. With these kernels, the look-up layer can represent not only a sphere-like space, but also linear, circular or any shape of environments.

Experimental data gathered using one embodiment shows that the addition of a look-up convolutional layer results in increased accuracy localization. Using the convolution kernel 1714 of Fig. 17, experiments were performed using a public dataset including seven scenes of data and associated camera poses for each scene. All data was used as training data, and the reported validation data is shown in **Table 1** below.

Table 1: experimental results with a public dataset

	Dataset	**Position(m)**		**Rotation(°)**	
scene		No look-up layer (error)	With look-up layer (error)	No look-up layer (error)	With look-up layer
chess	6K	0.25	0.20	2.69	2.48

fire	4K	0.19	0.17	2.42	2.20
heads	2K	0.14	0.15	2.00	1.86
office	10K	0.36	0.25	3.45	2.80
Pump.	6K	0.26	0.21	2.57	2.31
Redk	12K	0.37	0.28	3.31	2.83
stairs	3K	0.20	0.15	1.84	1.63
ALL	43K	0.60	0.42	5.09	4.16

Fig. 17, experiments were performed using a public dataset including seven scenes of data and associated camera poses for each scene. All data was used as training data, and the reported validation data is shown in Table 1 below.

The values of Table 1 are the average regression error, where a lower value represents a better result. The results show that with the addition of a look-up convolutional layer to the CNN, validation error is reduced. Table 1 shows the position and rotation estimation error in the analysis results of the dataset without the use of a convolutional look-up layer and with the use of a convolutional look-up layer, for each of seven scenes from the Ms7scenes public database. Additionally, analysis was performed using a combination of all scenes (e.g., ALL), which is a mix of the scene data from all seven scenes. This mix is not possible in reality, but is a useful test of the algorithm.

The results also indicate that network models with a look-up convolutional layer can be used to search larger environment areas without the corresponding reduction in accuracy associated with large environments. In existing implementations, as environment area increases, the accuracy of camera pose estimation drops. The use of the look-up convolutional layer enables both increased accuracy for the same environmental area, as well as the ability to process a larger environmental area without a corresponding reduction in accuracy.

Fig. 20 is a flow diagram of look-up convolution layer logic 2000, according to an embodiment. The look-up convolution layer logic 2000 can be performed by processing logic described herein, for example, as compute shader or pixel shader logic executed on a graphics processor or image processor according to an embodiment described herein.

In one embodiment the look-up convolution layer logic 2000 can select a set of look-up convolution kernels based on camera configuration of an imaging unit associated with an autonomous localization module, as shown at 2002. For an autonomous localization module

having a camera configuration in which the imaging unit captures a cylindrical view of a local environment, cylindrical or otherwise cylinder based convolution kernels may be used, such as the cylinder convolution kernel 1914 of Fig. 19A. For an autonomous localization module having a camera configuration in which the imaging unit captures a spherical view of a local environment, spherical or semi-spherical convolution kernels may be used, such as the spherical convolution kernel 1924 of Fig. 19B.

The look-up convolution layer logic 2000 can receive one or more feature maps from a convolutional layer of a convolutional neural network, as shown at 2004. The one or more feature maps have greater information density relative to an original image. The greater information density of the feature map relative to the original image is due to information compression performed by the filter kernels of the convolutional layer.

The look-up convolution layer logic 2000 can apply one or more look-up convolution kernels to the one or more feature maps to generate one or more look-up convolution results, as shown at 2006. The look-up convolution results have reduced information density (e.g., are more sparse) relative to the one or more feature maps received from the convolution layer, as the look-up convolutional layer is configured to determine the location of the information dense feature map in a global look-up kernel. The global look-up kernel is associated with environment maps that are used to perform camera pose estimation. The output from the look-up convolutional layer can be provided to a set of fully connected layers of a neural network, which can output scores for use in estimating the camera position and orientation of the original image, as shown at 2008. The fully-connected neural network can be a portion of the convolutional neural network or a separate neural network.

Fig. 21 is a flow diagram of camera pose estimation logic 2100, according to an embodiment. The camera pose estimation logic 2100 as described herein can apply a convolutional neural network including a look-up convolutional layer to generate a first feature map, as shown at 2102. In one embodiment, multiple filter kernels may be applied to generate multiple feature maps.

The camera pose estimation logic 2100 can then apply a look-up convolution layer to the first feature map to generate a second feature map, as shown at 2104. The second feature map can be associated with a location of the first feature map within a look-up convolution kernel, where the look-up convolution kernel represents an environment map in which the first feature map is to be located. The location of the first feature map in an environment map can be used to estimate a position and orientation of a camera associated with the image, as shown at 2106.

One having skill in the art will understand that, for the camera pose estimation logic 2100 to operate properly, the convolutional provided that the neural network is trained with proper

environmental maps associated with the environment in which imaging logic associated with the camera pose estimation logic 2100 is to operate.

Fig. 22 is a block diagram of a graphics core 2200 including logic to perform operations associated with embodiments described herein. The graphics core 2200 includes execution resources sufficient to perform the processing operations associated with a convolutional neural network and autonomous localization. In one embodiment the graphics core 2200 (e.g., slice) includes a cluster of sub-cores 2206A-2206C, which may be variants of the sub-cores 550A-550N. In one embodiment the graphics core includes shared resources 2201, such as the shared function logic 420 of Fig. 4. However, in the illustrated embodiment each of the sub-cores 2206A-2206C includes sampler resources 2203A-2203C and a sampler cache 2213A-2213C. In one embodiment the shared resources 2201 include of a set of fixed function units 2202, for example, to support media, two-dimensional graphics functionality, and pixel back end operations for graphics and image processing. For programmable graphics and computational processing, a thread dispatcher 2204 can dispatch execution threads to the various sub-cores 2206A-2206C, where a local dispatch unit 2208A-2208C dispatches execution threads to the execution unit groups 2210A-2210C in each of the sub-cores. The number of execution units in each of the execution unit groups 2210A-2210C can vary among embodiments. Execution units within each group 2210A-C can also be dynamically enabled or disabled based on workload, power, or thermal conditions.

In one embodiment, a level three (L3) data cache 2220 is shared between each of the sub-cores 2206A-C. The L3 data cache 2220 can include an atomics & barriers unit 2222 and shared local memory 2224. The atomics & barriers unit 2222 includes dedicated logic to support implementation of barriers across groups of threads and is available as a hardware alternative to pure compiler or software based barrier implementations. Additionally, the atomics & barriers unit 2222 enables a suite of atomic read-modify-write memory operations to the L3 data cache 2220 or to the shared local memory 2224. Atomic operations to global memory can be supported via the L3 data cache 2220.

In one embodiment, the shared local memory 2224 supports programmer managed data for sharing amongst hardware threads, with access latency similar to the access latency to the L3 data cache 2220. In one embodiment, the shared local memory 2224 sharing is limited to between threads within the same sub-core 2206A-C, however, not all embodiments share such limitation.

Fig. 23 is a block diagram of a computing device 2300 including a graphics processor 2304, according to an embodiment. The computing device 2300 can be a computing device such as the data processing system 100 as in of Fig. 1. The computing device 2300 may also be or be

included within a communication device such as a set-top box (e.g., Internet-based cable television set-top boxes, etc.), global positioning system (GPS)-based devices, etc. The computing device 2300 may also be or be included within mobile computing devices such as cellular phones, smartphones, personal digital assistants (PDAs), tablet computers, laptop computers, e-readers, smart televisions, television platforms, wearable devices (e.g., glasses, watches, bracelets, smartcards, jewelry, clothing items, etc.), media players, etc. For example, in one embodiment, the computing device 2300 includes a mobile computing device employing an integrated circuit ("IC"), such as system on a chip ("SoC" or "SOC"), integrating various hardware and/or software components of computing device 2300 on a single chip.

The computing device 2300 includes a graphics processor 2304, which may be any graphics processor described herein. The graphics processor 2304 includes neural network logic 2324 to accelerate convolution and/or look-up convolution operations associated with a convolutional neural network. The graphics processor also includes one or more graphics engine(s) 2354, which may include one or more instances of the graphics core 2200 of Fig. 22, or any graphics execution logic described herein, such as the execution logic 600 of Fig. 6. The graphics engine(s) 2354 include execution resources which may work in conjunction with the neural network logic 2324. The graphics processor 2304 also includes localization logic 2344 to leverage the neural network logic 2324 to perform autonomous localization, for example, for autonomous robots or autonomous vehicles.

The graphics processor 2304 also includes a display engine 2334 to couple the graphics processor to a display device. Data that is processed by the graphics processor 2304 is stored in a buffer within a hardware graphics pipeline and state information is stored in memory 2308. The resulting image is then transferred to a display controller of the display engine 2334 for output via a display device, such as the display device 320 of Fig. 3. The display device may be of various types, such as Cathode Ray Tube (CRT), Thin Film Transistor (TFT), Liquid Crystal Display (LCD), Organic Light Emitting Diode (OLED) array, etc., and may be configured to display information to a user.

As illustrated, in one embodiment, in addition to a graphics processor 2304, the computing device 2300 may further include any number and type of hardware components and/or software components, such as (but not limited to) an application processor 2306, memory 2308, and input/output (I/O) sources 2310. The application processor 2306 can interact with a hardware graphics pipeline, as illustrated with reference to Fig. 3, to share graphics pipeline functionality. The application processor 2306 can include one or processors, such as processor(s) 102 of Fig. 1, and may be the central processing unit (CPU) that is used at least in part to execute an operating system (OS) 2302 for the computing device 2300. The OS 2302 can serve as an interface

between hardware and/or physical resources of the computer device 2300 and a user. The OS 2302 can include driver logic 2322 for various hardware devices in the computing device 2300. The driver logic 2322 can include graphics driver logic 2323 such as the user mode graphics driver 1026 and/or kernel mode graphics driver 1029 of Fig. 10. In one embodiment the graphics driver logic 2323 is configured to receive data from an imaging application programming interface (API) 2320 and/or a compute AP 2321 to perform convolutional neural network and autonomous localization operations via the graphics processor 2304. In one embodiment, convolution kernels and look-up convolution kernels can be provided to the neural network logic via the compute API 2321. In one embodiment images for processing can be captured by or provided by the imaging API 2320. In one embodiment, a localization API associated with the localization logic can utilize both the imagine API 2320 and the compute API 2321 to provide a unified processing interface for image data to enable autonomous localization via camera input.

It is contemplated that in some embodiments, the graphics processor 2304 may exist as part of the application processor 2306 (such as part of a physical CPU package) in which case, at least a portion of the memory 2308 may be shared by the application processor 2306 and graphics processor 2304, although at least a portion of the memory 2308 may be exclusive to the graphics processor 2304, or the graphics processor 2304 may have a separate store of memory. The memory 2308 may comprise a pre-allocated region of a buffer (e.g., framebuffer); however, it should be understood by one of ordinary skill in the art that the embodiments are not so limited, and that any memory accessible to the lower graphics pipeline may be used. The memory 2308 may include various forms of random access memory (RAM) (e.g., SDRAM, SRAM, etc.) comprising an application that makes use of the graphics processor 2304 to render a desktop or 3D graphics scene. A memory controller hub, such as memory controller hub 116 of Fig. 1, may access data in the memory 2308 and forward it to graphics processor 2304 for graphics pipeline processing. The memory 2308 may be made available to other components within the computing device 2300. For example, any data (e.g., input graphics data) received from various I/O sources 2310 of the computing device 2300 can be temporarily queued into memory 2308 prior to their being operated upon by one or more processor(s) (e.g., application processor 2306) in the implementation of a software program or application. Similarly, data that a software program determines should be sent from the computing device 2300 to an outside entity through one of the computing system interfaces, or stored into an internal storage element, is often temporarily queued in memory 2308 prior to its being transmitted or stored.

The I/O sources can include devices such as touchscreens, touch panels, touch pads, virtual or regular keyboards, virtual or regular mice, ports, connectors, network devices, or the like, and

can attach via an input/output (I/O) control hub (ICH) 130 as referenced in Fig. 1. Additionally, the I/O sources 2310 may include one or more I/O devices that are implemented for transferring data to and/or from the computing device 2300 (e.g., a networking adapter); or, for a large-scale non-volatile storage within the computing device 2300 (e.g., hard disk drive). User input devices, including alphanumeric and other keys, may be used to communicate information and command selections to graphics processor 2304. Another type of user input device is cursor control, such as a mouse, a trackball, a touchscreen, a touchpad, or cursor direction keys to communicate direction information and command selections to GPU and to control cursor movement on the display device. Camera and microphone arrays of the computer device 2300 may be employed to observe gestures, record audio and video and to receive and transmit visual and audio commands.

I/O sources 2310 configured as one or more network interface(s) can provide access to a network, such as a LAN, a wide area network (WAN), a metropolitan area network (MAN), a personal area network (PAN), Bluetooth, a cloud network, a cellular or mobile network (e.g., 3rd Generation (3G), 4th Generation (4G), etc.), an intranet, the Internet, etc. Network interface(s) may include, for example, a wireless network interface having one or more antenna(e). Network interface(s) may also include, for example, a wired network interface to communicate with remote devices via network cable, which may be, for example, an Ethernet cable, a coaxial cable, a fiber optic cable, a serial cable, or a parallel cable.

Network interface(s) may provide access to a LAN, for example, by conforming to IEEE 802.11 standards, and/or the wireless network interface may provide access to a personal area network, for example, by conforming to Bluetooth standards. Other wireless network interfaces and/or protocols, including previous and subsequent versions of the standards, may also be supported. In addition to, or instead of, communication via the wireless LAN standards, network interface(s) may provide wireless communication using, for example, Time Division, Multiple Access (TDMA) protocols, Global Systems for Mobile Communications (GSM) protocols, Code Division, Multiple Access (CDMA) protocols, and/or any other type of wireless communications protocols.

It is to be appreciated that a lesser or more equipped system than the example described above may be preferred for certain implementations. Therefore, the configuration of the computing device 2300 may vary from implementation to implementation depending upon numerous factors, such as price constraints, performance requirements, technological improvements, or other circumstances. Examples include (without limitation) a mobile device, a personal digital assistant, a mobile computing device, a smartphone, a cellular telephone, a handset, a one-way pager, a two-way pager, a messaging device, a computer, a personal

computer (PC), a desktop computer, a laptop computer, a notebook computer, a handheld computer, a tablet computer, a server, a server array or server farm, a web server, a network server, an Internet server, a work station, a mini-computer, a main frame computer, a supercomputer, a network appliance, a web appliance, a distributed computing system, multiprocessor systems, processor-based systems, consumer electronics, programmable consumer electronics, television, digital television, set top box, wireless access point, base station, subscriber station, mobile subscriber center, radio network controller, router, hub, gateway, bridge, switch, machine, or combinations thereof.

Embodiments may be implemented as any one or a combination of: one or more microchips or integrated circuits interconnected using a parent-board, hardwired logic, software stored by a memory device and executed by a microprocessor, firmware, an application specific integrated circuit (ASIC), and/or a field programmable gate array (FPGA). The term "logic" may include, by way of example, software or hardware and/or combinations of software and hardware.

Embodiments may be provided, for example, as a computer program product which may include one or more machine-readable media having stored thereon machine-executable instructions that, when executed by one or more machines such as a computer, network of computers, or other electronic devices, may result in the one or more machines carrying out operations in accordance with embodiments described herein. A machine-readable medium may include, but is not limited to, floppy diskettes, optical disks, CD-ROMs (Compact Disc-Read Only Memories), and magneto-optical disks, ROMs, RAMs, EPROMs (Erasable Programmable Read Only Memories), EEPROMs (Electrically Erasable Programmable Read Only Memories), magnetic or optical cards, flash memory, or other type of media/machine-readable medium suitable for storing machine-executable instructions.

Moreover, embodiments may be downloaded as a computer program product, wherein the program may be transferred from a remote computer (e.g., a server) to a requesting computer (e.g., a client) by way of one or more data signals embodied in and/or modulated by a carrier wave or other propagation medium via a communication link (e.g., a modem and/or network connection).

The following clauses and/or examples pertain to specific embodiments or examples thereof. Specifics in the examples may be used anywhere in one or more embodiments. The various features of the different embodiments or examples may be variously combined with some features included and others excluded to suit a variety of different applications. Examples may include subject matter such as a method, means for performing acts of the method, at least one machine-readable medium including instructions that, when performed by a machine cause

the machine to performs acts of the method, or of an apparatus or system according to embodiments and examples described herein. Various components can be a means for performing the operations or functions described.

Those skilled in the art will appreciate from the foregoing description that the broad techniques of the embodiments can be implemented in a variety of forms. Therefore, while the embodiments have been described in connection with particular examples thereof, the true scope of the embodiments should not be so limited since other modifications will become apparent to the skilled practitioner upon a study of the drawings, specification, and following claims.

CLAIMS

What is claimed is:

1. A processor including logic to accelerate convolutional neural network processing, the processor including:

first logic to apply a convolutional layer to an image to generate a first convolution result; and

second logic to apply a look-up convolutional layer to the first convolution result to generate a second convolution result, the second convolution result associated with a location of the first convolution result within a global filter kernel.

2. The processor as in claim 1, wherein the global filter kernel is significantly larger than the first convolution result and the second convolution result is an enlarged feature map.

3. The processor as in claim 2, wherein the global filter kernel is associated with an environment filter map for use in camera pose estimation.

4. The processor as in claim 3, wherein the environmental filter map is represented by a global filter kernel having a shape defined by a view angle of an imaging module used to capture the image.

5. The processor as in claim 3, additionally including third logic to process the second convolution result to estimate a camera pose for a camera associated with the image.

6. The processor as in claim 5, the third logic is configured to apply one or more fully-connected neural network layers to the enlarged feature map to generate a value associated with a camera pose estimate.

7. The processor as in claim 5, the third logic is configured to apply one or more fully-connected neural network layers to a subsample of the enlarged feature map to generate a value associated with a camera pose estimate.

8. The processor as in claim 5, the third logic to estimate the camera pose for a camera associated with an autonomous robot.

9. The processor as in claim 5, the third logic estimate the camera pose for a camera associated with an autonomous vehicle.

10. The processor as in claim 1, additionally including graphics or image processor core to include the first or second logic and additionally including fourth logic to receive one or more convolutional kernels associated with the convolutional layer and at least one look-up convolutional kernel associated with the look-up convolution layer.

11. The processor as in claim 10, the fourth logic associated with a graphics processor computational application programming interface.

12. A computer processing method executed by one or more processors, the method comprising:

applying a convolutional layer of a convolutional neural network to an image to generate a first feature map;

applying a look-up convolution layer to the first feature map to generate a second feature map, the second feature map associated with a location of the first feature map within a look-up convolution kernel; and

estimating a position and orientation of a camera associated with the image.

13. The method as in claim 12, wherein the image is a multi-dimensional image of an environment and the look-up convolution kernel is a global feature kernel associated with a localization environment map.

14. The method as in claim 13, wherein the global feature kernel is selected based on a camera configuration of an imaging unit associated with a localization module of an autonomous vehicle or autonomous robot.

15. A data processing system including means for performing a method as in any one of claims 12-14.

16. A non-transitory computer-readable medium storing instructions which, when executed by one or more processors, cause the one or more processors to perform a method as in any one of claims 12-14.

17. Machine-readable storage media storing data which, when read by one or more machines, cause the one or more machines to manufacture an integrated circuit to perform a method comprising:

applying a convolutional layer of a convolutional neural network to an image to generate a first feature map;

applying a look-up convolution layer to the first feature map to generate a second feature map, the second feature map associated with a location of the first feature map within a look-up convolution kernel; and

estimating a position and orientation of a camera associated with the image.

18. The machine readable storage media as in claim 17, wherein the image is a multi-dimensional image of an environment and the look-up convolution kernel is a global feature kernel associated with a localization environment map.

19. The machine readable storage media as in claim 18, wherein the global feature kernel is selected based on a camera configuration of an imaging unit associated with a localization module of an autonomous vehicle or autonomous robot.

20. The machine readable storage media as in claim 19, wherein the global feature kernel is associated with a cylindrical environment map.

21. The machine readable storage media as in claim 19, wherein the global feature kernel is associated with a spherical environment map.

22. The machine readable storage media as in claim 17, wherein applying a look-up convolution layer to the first feature map includes applying multiple look-up convolution kernels to the image.

23. The method as in claim 22, wherein the image is a multi-dimensional image having one or more of a red, green, and blue channel.

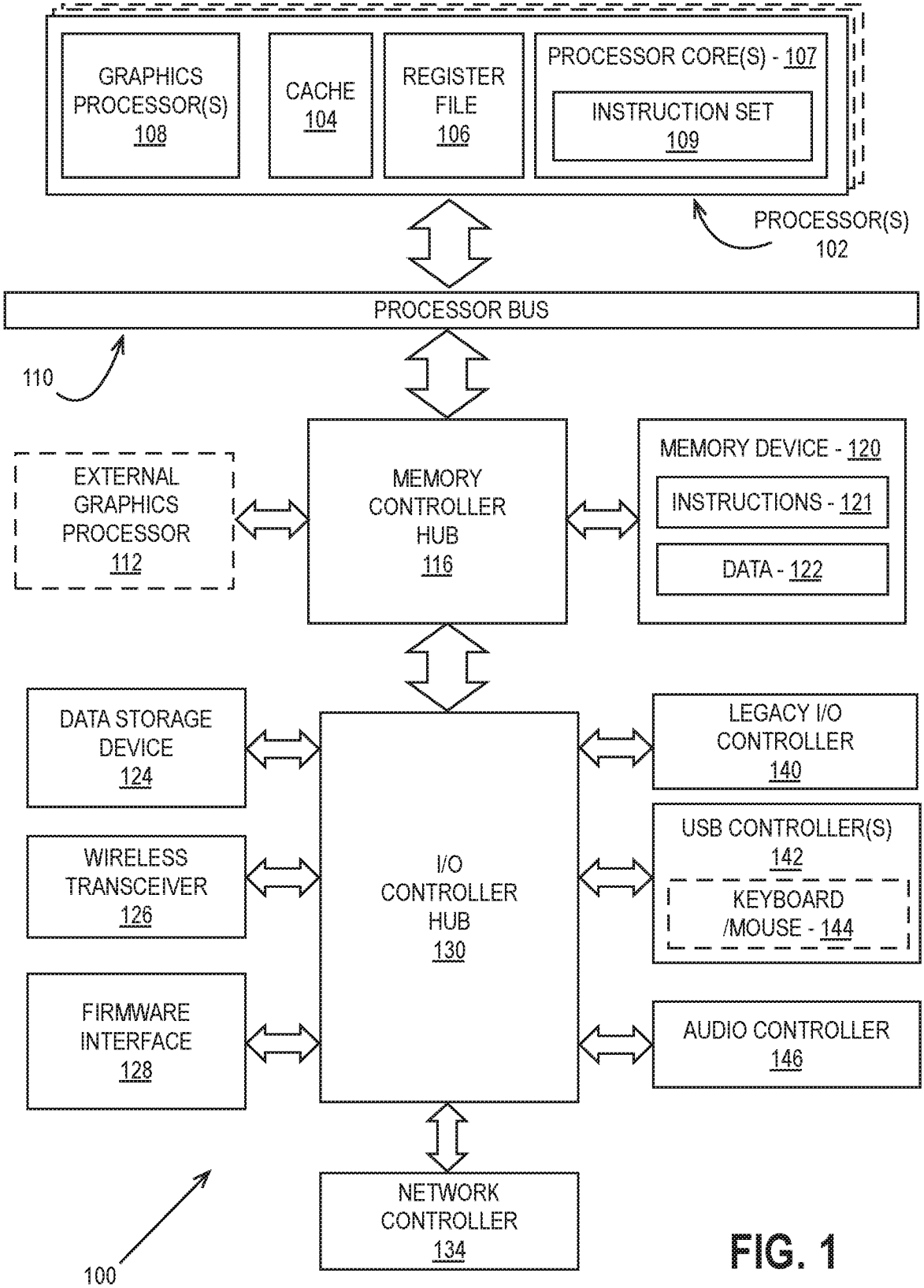

GRAPHICS PROCESSOR(S) 108
CACHE 104
REGISTER FILE 106
PROCESSOR CORE(S) - 107
INSTRUCTION SET 109
PROCESSOR(S) 102
PROCESSOR BUS
110
EXTERNAL GRAPHICS PROCESSOR 112
MEMORY CONTROLLER HUB 116
MEMORY DEVICE - 120
INSTRUCTIONS - 121
DATA - 122
DATA STORAGE DEVICE 124
WIRELESS TRANSCEIVER 126
FIRMWARE INTERFACE 128
I/O CONTROLLER HUB 130
LEGACY I/O CONTROLLER 140
USB CONTROLLER(S) 142
KEYBOARD /MOUSE - 144
AUDIO CONTROLLER 146
NETWORK CONTROLLER 134
100

FIG. 1

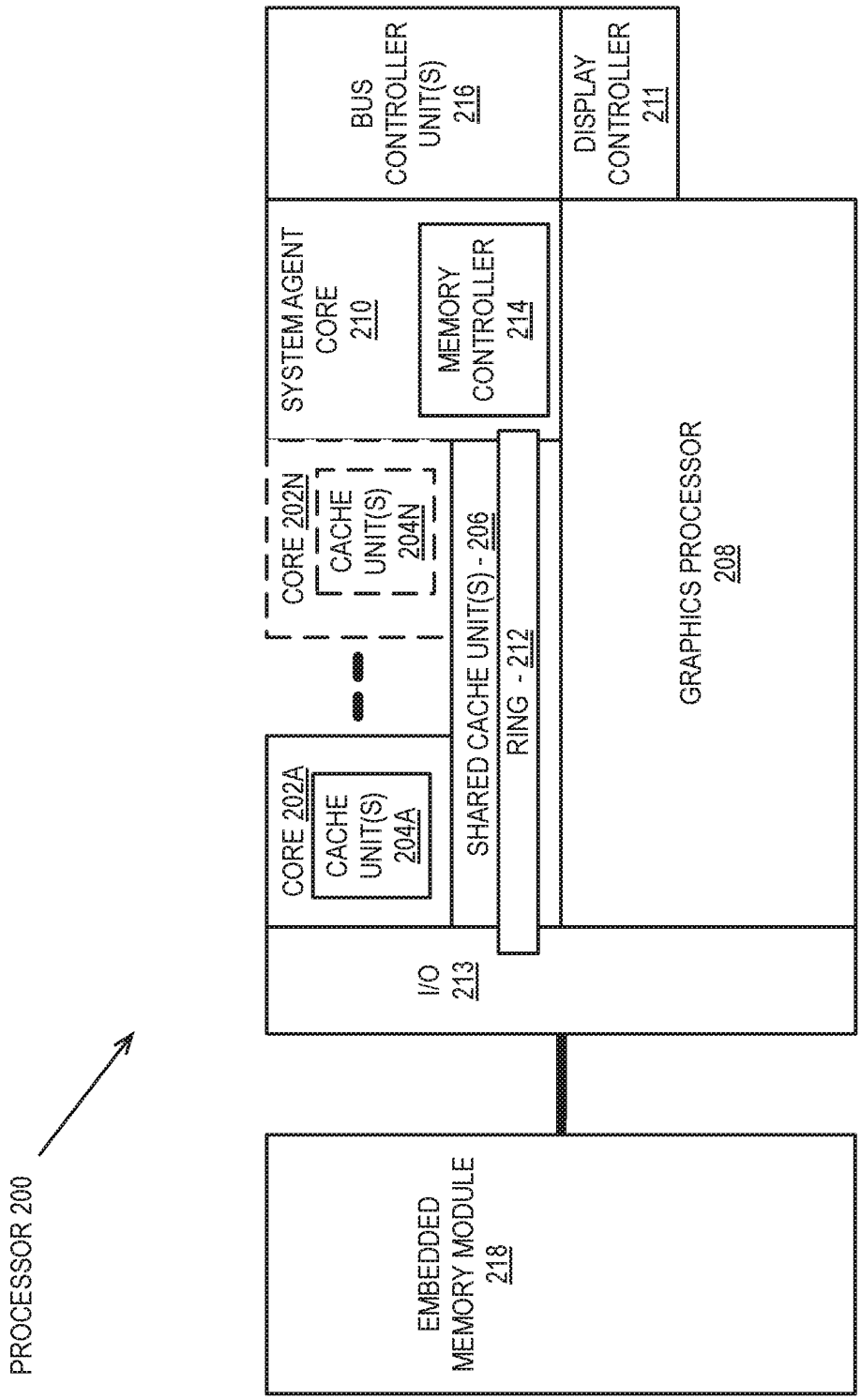

PROCESSOR 200
EMBEDDED MEMORY MODULE 218
I/O 213
CORE 202A
CACHE UNIT(S) 204A
CORE 202N
CACHE UNIT(S) 204N
SHARED CACHE UNIT(S) - 206
RING - 212
SYSTEM AGENT CORE 210
MEMORY CONTROLLER 214
GRAPHICS PROCESSOR 208
BUS CONTROLLER UNIT(S) 216
DISPLAY CONTROLLER 211

FIG. 2

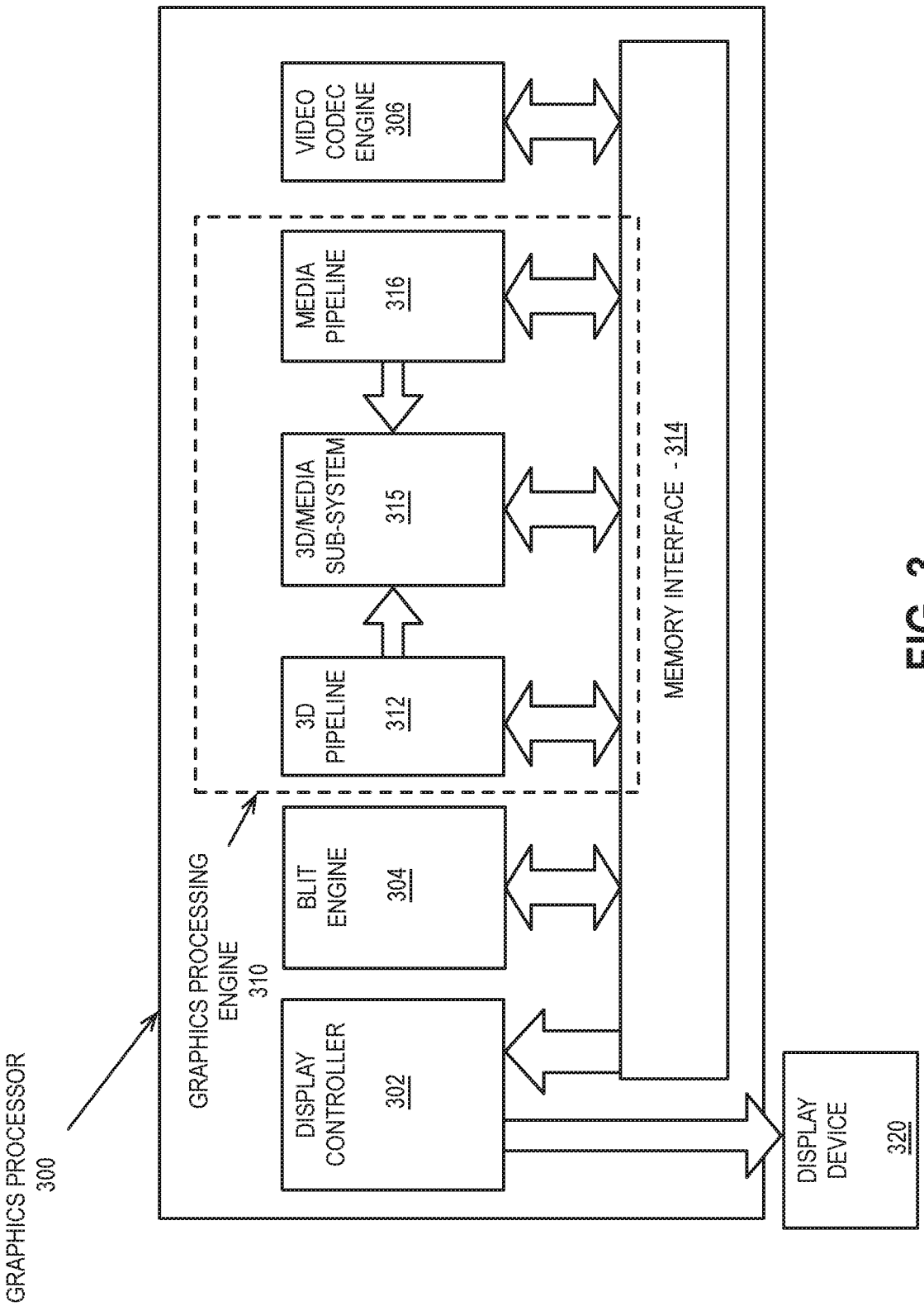

GRAPHICS PROCESSOR
300
GRAPHICS PROCESSING ENGINE
310
DISPLAY CONTROLLER
302
BLIT ENGINE
304
3D PIPELINE
312
3D/MEDIA SUB-SYSTEM
315
MEDIA PIPELINE
316
VIDEO CODEC ENGINE
306
MEMORY INTERFACE - 314
DISPLAY DEVICE
320

FIG. 3

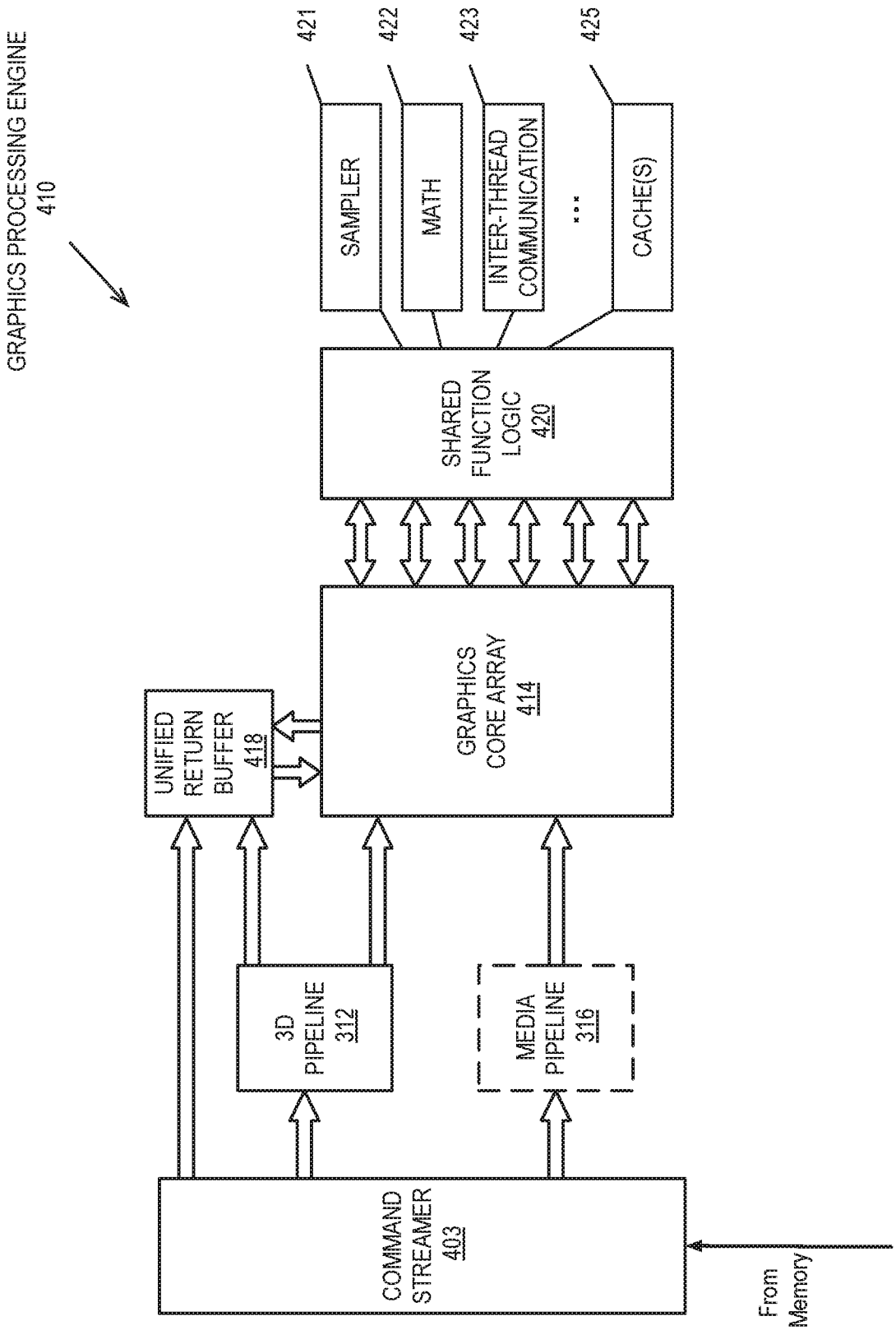

GRAPHICS PROCESSING ENGINE
410
421
422
423
425
SAMPLER
MATH
INTER-THREAD COMMUNICATION
...
CACHE(S)
SHARED FUNCTION LOGIC 420
GRAPHICS CORE ARRAY 414
UNIFIED RETURN BUFFER 418
3D PIPELINE 312
MEDIA PIPELINE 316
COMMAND STREAMER 403
From Memory

FIG. 4

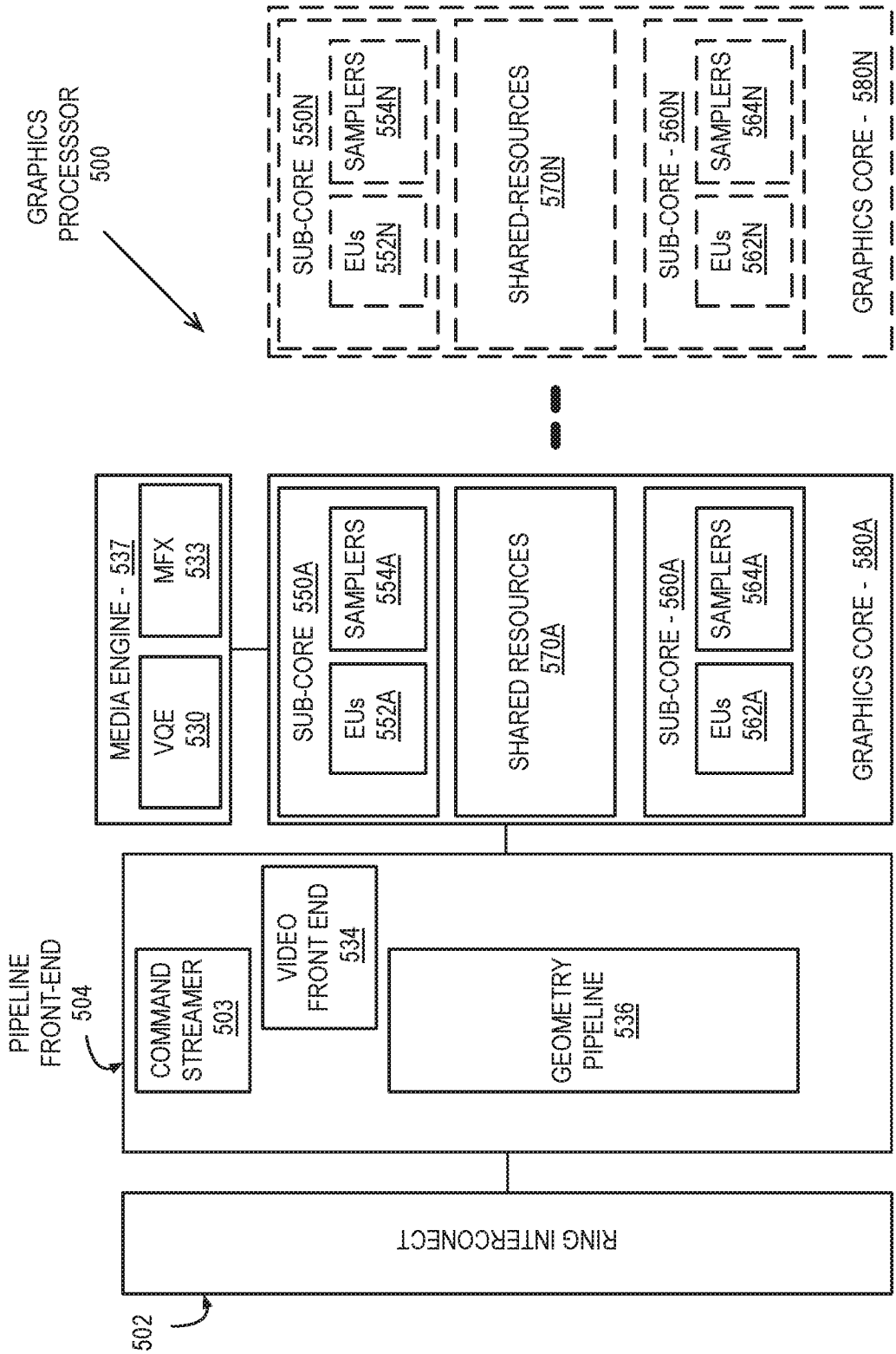

GRAPHICS PROCESSSOR 500
MEDIA ENGINE - 537
VQE 530
MFX 533
SUB-CORE 550A
EUs 552A
SAMPLERS 554A
SHARED RESOURCES 570A
SUB-CORE - 560A
EUs 562A
SAMPLERS 564A
GRAPHICS CORE - 580A
SUB-CORE 550N
EUs 552N
SAMPLERS 554N
SHARED-RESOURCES 570N
SUB-CORE - 560N
EUs 562N
SAMPLERS 564N
GRAPHICS CORE - 580N
PIPELINE FRONT-END 504
COMMAND STREAMER 503
VIDEO FRONT END 534
GEOMETRY PIPELINE 536
RING INTERCONECT
502

FIG. 5

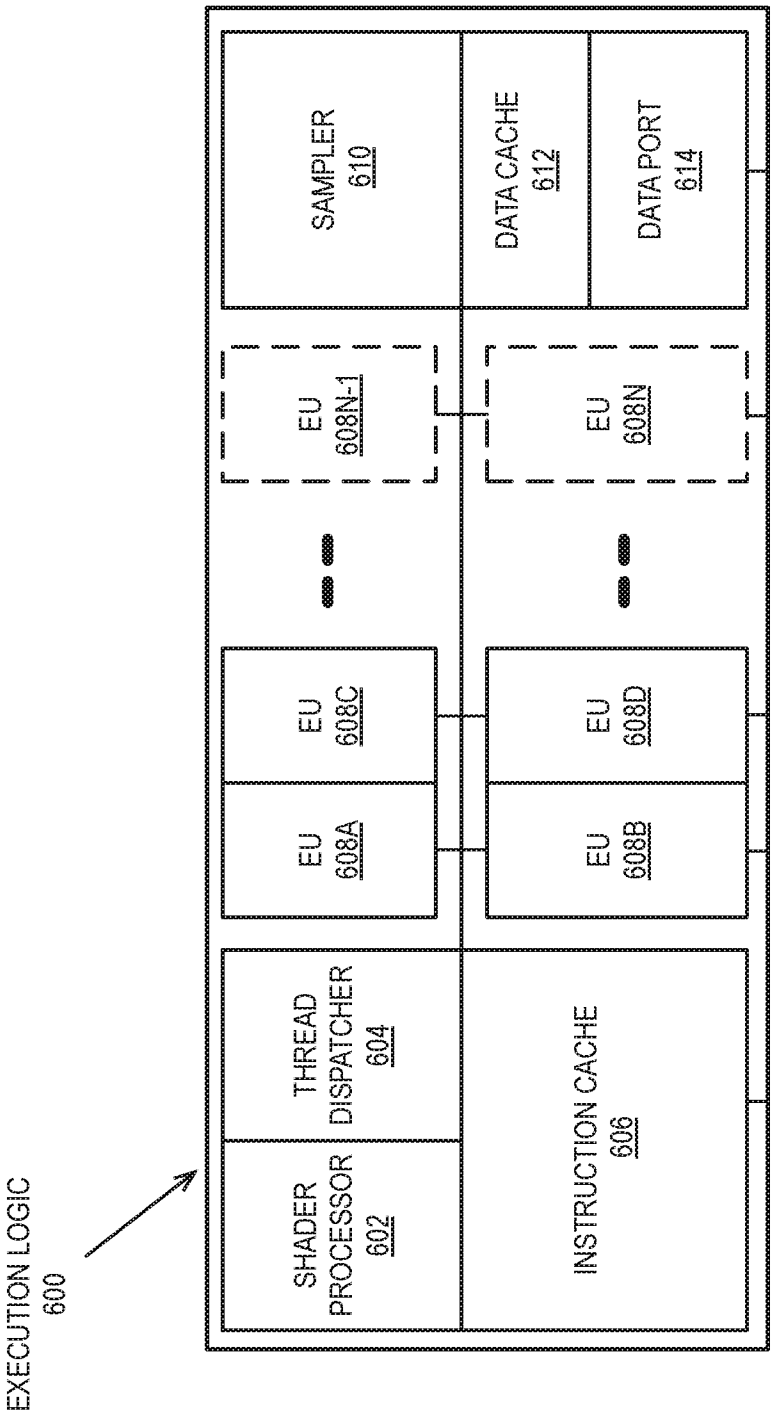

EXECUTION LOGIC
600
SHADER PROCESSOR
602
THREAD DISPATCHER
604
INSTRUCTION CACHE
606
EU
608A
EU
608C
EU
608N-1
EU
608B
EU
608D
EU
608N
SAMPLER
610
DATA CACHE
612
DATA PORT
614

FIG. 6

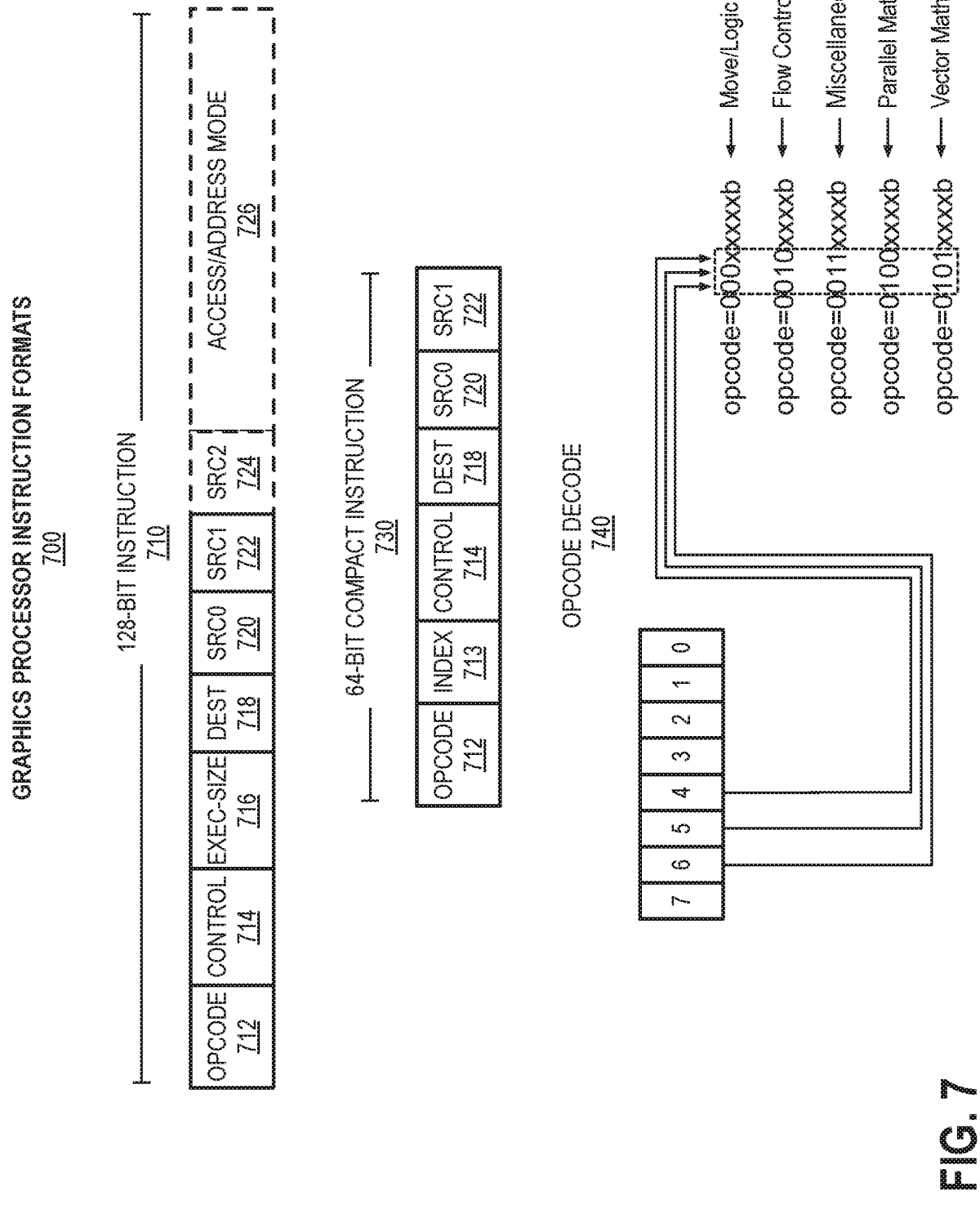

GRAPHICS PROCESSOR INSTRUCTION FORMATS
700
128-BIT INSTRUCTION
710
OPCODE 712
CONTROL 714
EXEC-SIZE 716
DEST 718
SRC0 720
SRC1 722
SRC2 724
ACCESS/ADDRESS MODE 726
64-BIT COMPACT INSTRUCTION
730
OPCODE 712
INDEX 713
CONTROL 714
DEST 718
SRC0 720
SRC1 722
OPCODE DECODE
740
7 6 5 4 3 2 1 0
opcode=0000xxxxb
opcode=0010xxxxb
opcode=0011xxxxb
opcode=0100xxxxb
opcode=0101xxxxb
Move/Logic - 742
Flow Control - 744
Miscellaneous - 746
Parallel Math - 748
Vector Math - 750

FIG. 7

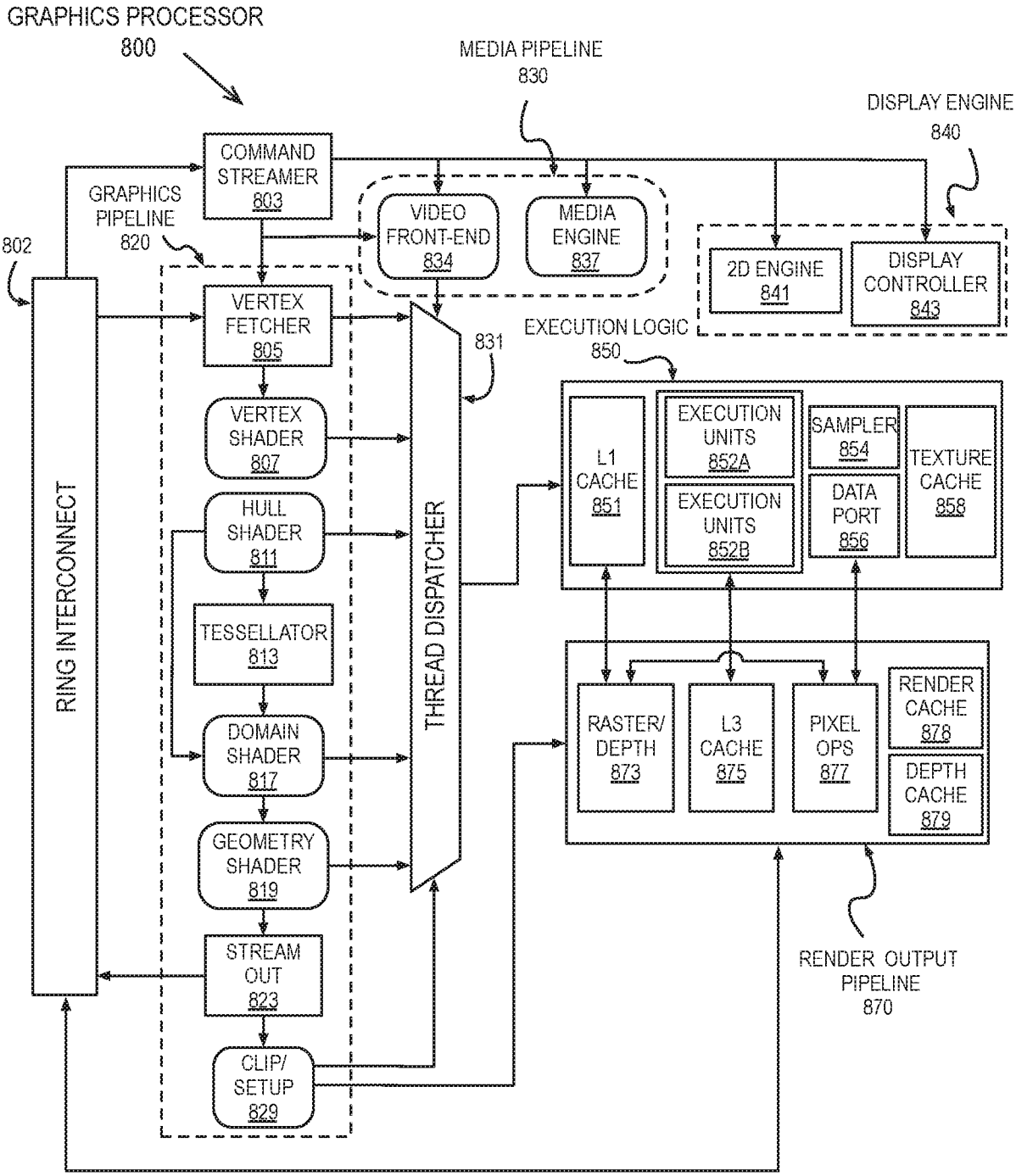
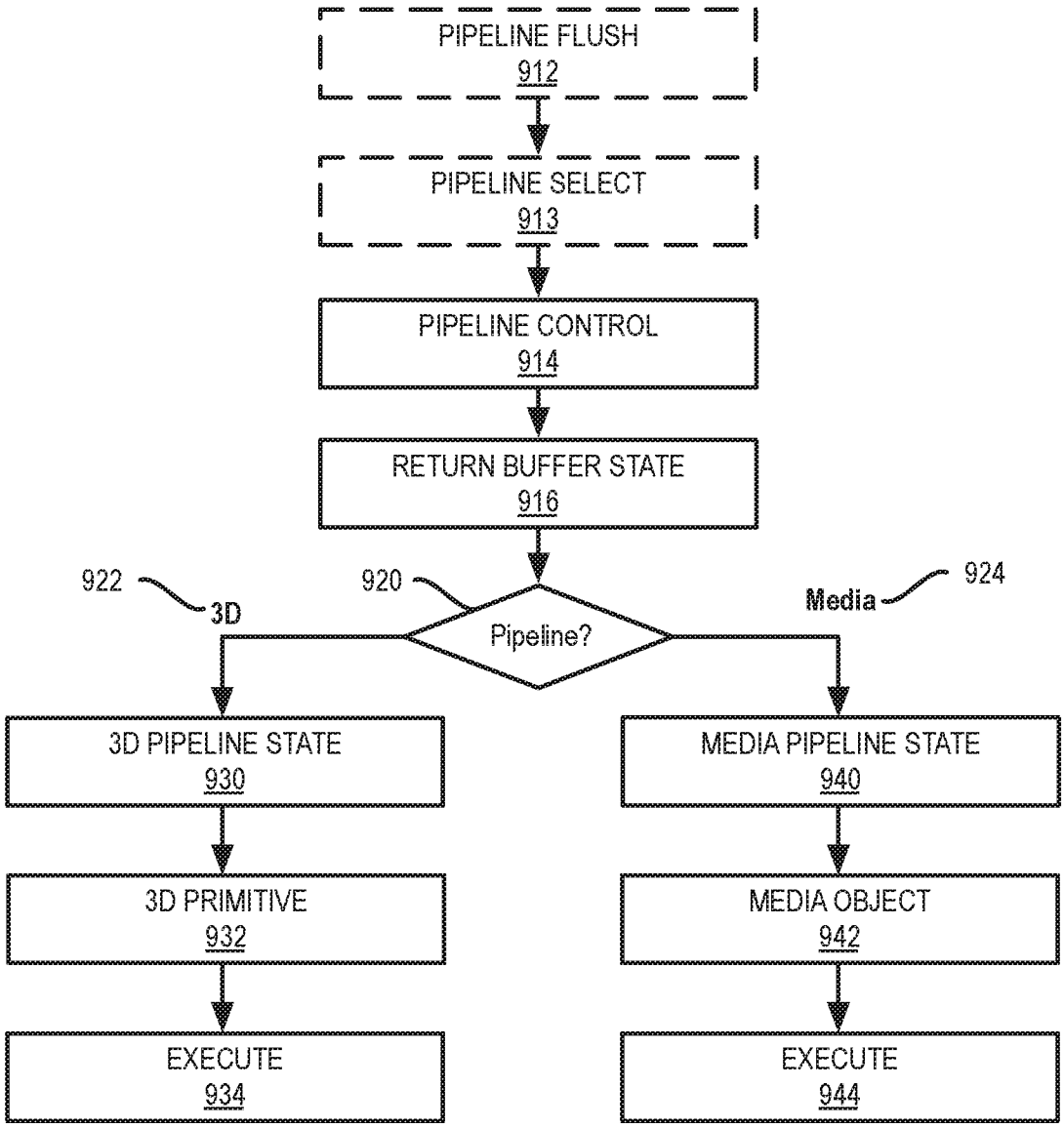

GRAPHICS PROCESSOR
800
MEDIA PIPELINE
830
DISPLAY ENGINE
840
COMMAND STREAMER
803
GRAPHICS PIPELINE
820
802
VIDEO FRONT-END
834
MEDIA ENGINE
837
2D ENGINE
841
DISPLAY CONTROLLER
843
VERTEX FETCHER
805
831
EXECUTION LOGIC
850
VERTEX SHADER
807
HULL SHADER
811
TESSELLATOR
813
DOMAIN SHADER
817
GEOMETRY SHADER
819
STREAM OUT
823
CLIP/ SETUP
829
RING INTERCONNECT
THREAD DISPATCHER
L1 CACHE
851
EXECUTION UNITS
852A
EXECUTION UNITS
852B
SAMPLER
854
DATA PORT
856
TEXTURE CACHE
858
RASTER/ DEPTH
873
L3 CACHE
875
PIXEL OPS
877
RENDER CACHE
878
DEPTH CACHE
879
RENDER OUTPUT PIPELINE
870

FIG. 8

FIG. 9A GRAPHICS PROCESSOR COMMAND FORMAT
900

FIG. 9B GRAPHICS PROCESSOR COMMAND SEQUENCE
910



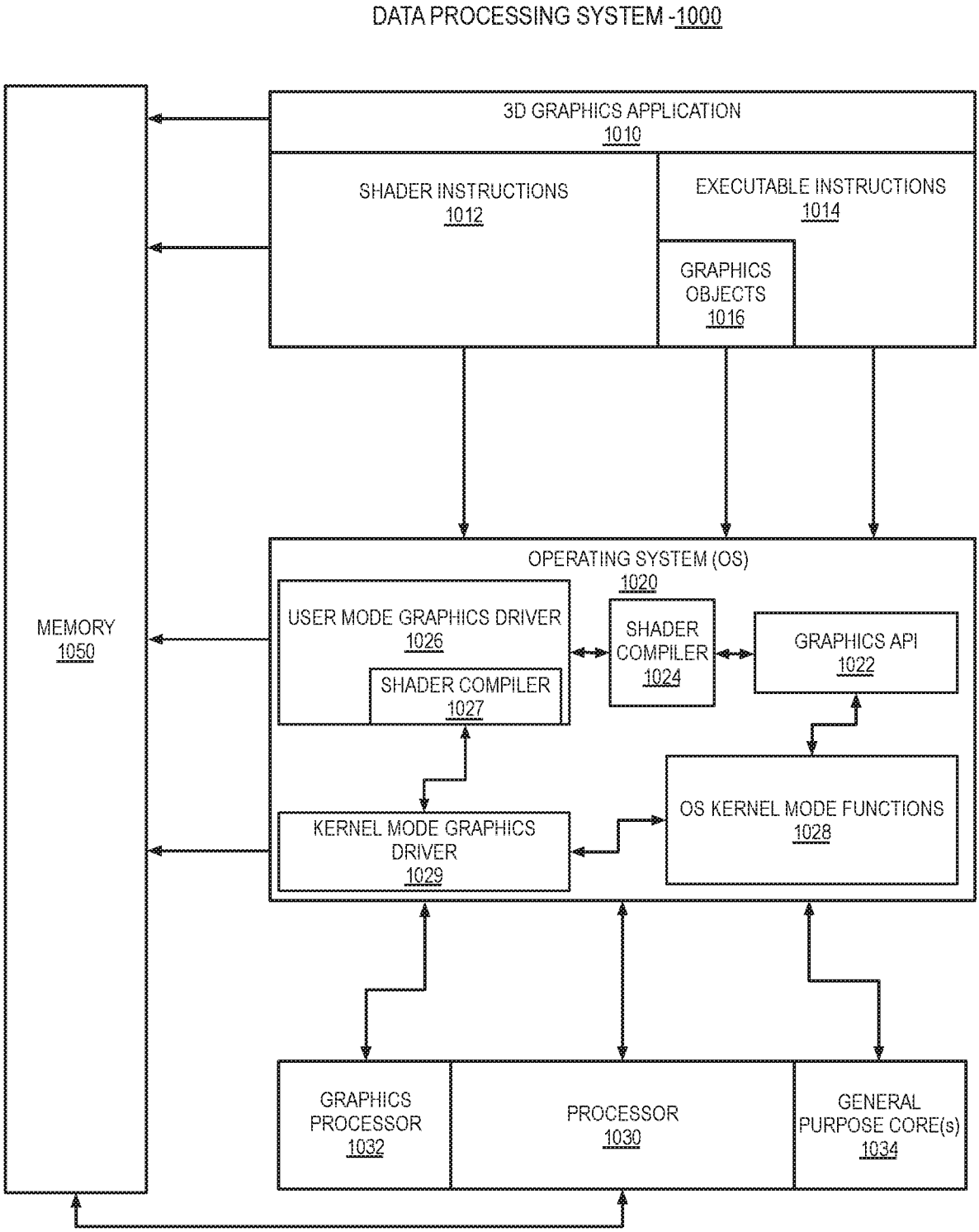

DATA PROCESSING SYSTEM -1000
3D GRAPHICS APPLICATION
1010
SHADER INSTRUCTIONS
1012
EXECUTABLE INSTRUCTIONS
1014
GRAPHICS OBJECTS
1016
MEMORY
1050
OPERATING SYSTEM (OS)
1020
USER MODE GRAPHICS DRIVER
1026
SHADER COMPILER
1027
SHADER COMPILER
1024
GRAPHICS API
1022
KERNEL MODE GRAPHICS DRIVER
1029
OS KERNEL MODE FUNCTIONS
1028
GRAPHICS PROCESSOR
1032
PROCESSOR
1030
GENERAL PURPOSE CORE(s)
1034

FIG. 10

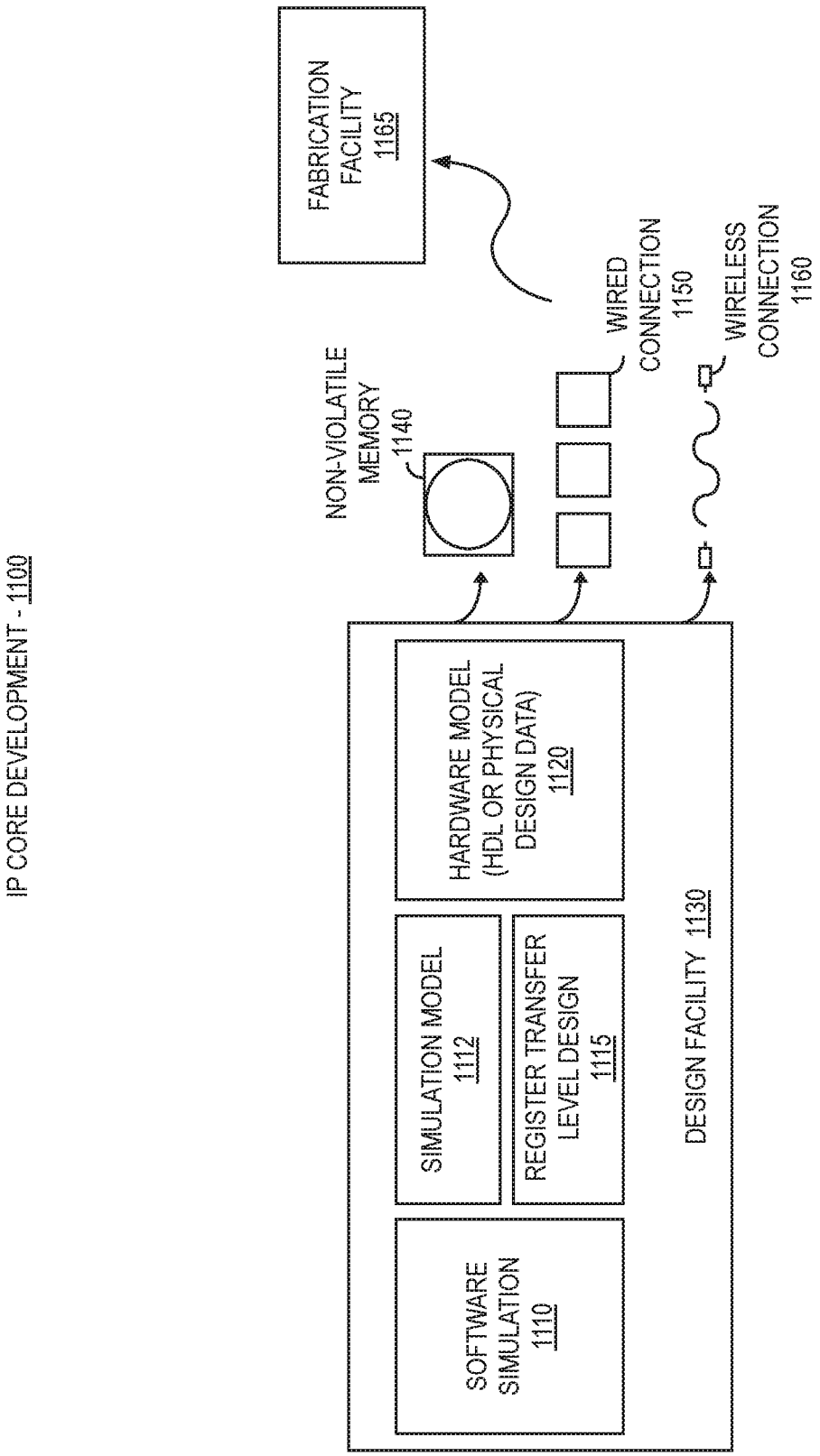

IP CORE DEVELOPMENT - 1100
FABRICATION FACILITY 1165
NON-VIOLATILE MEMORY 1140
WIRED CONNECTION 1150
WIRELESS CONNECTION 1160
SOFTWARE SIMULATION 1110
SIMULATION MODEL 1112
REGISTER TRANSFER LEVEL DESIGN 1115
HARDWARE MODEL (HDL OR PHYSICAL DESIGN DATA) 1120
DESIGN FACILITY 1130

FIG. 11

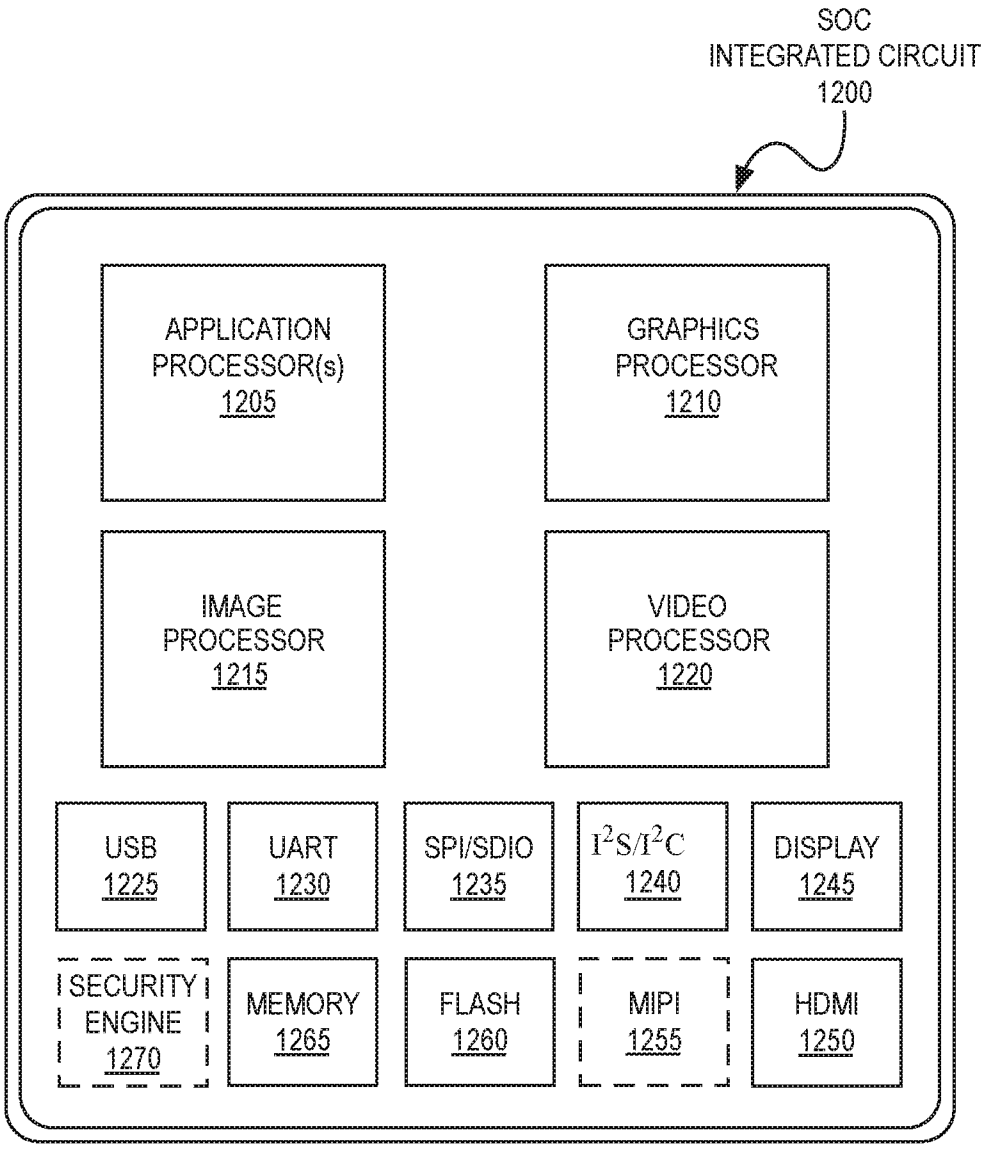

SOC
INTEGRATED CIRCUIT
1200
APPLICATION PROCESSOR(s) 1205
GRAPHICS PROCESSOR 1210
IMAGE PROCESSOR 1215
VIDEO PROCESSOR 1220
USB 1225
UART 1230
SPI/SDIO 1235
I2S/I2C 1240
DISPLAY 1245
SECURITY ENGINE 1270
MEMORY 1265
FLASH 1260
MIPI 1255
HDMI 1250

FIG. 12

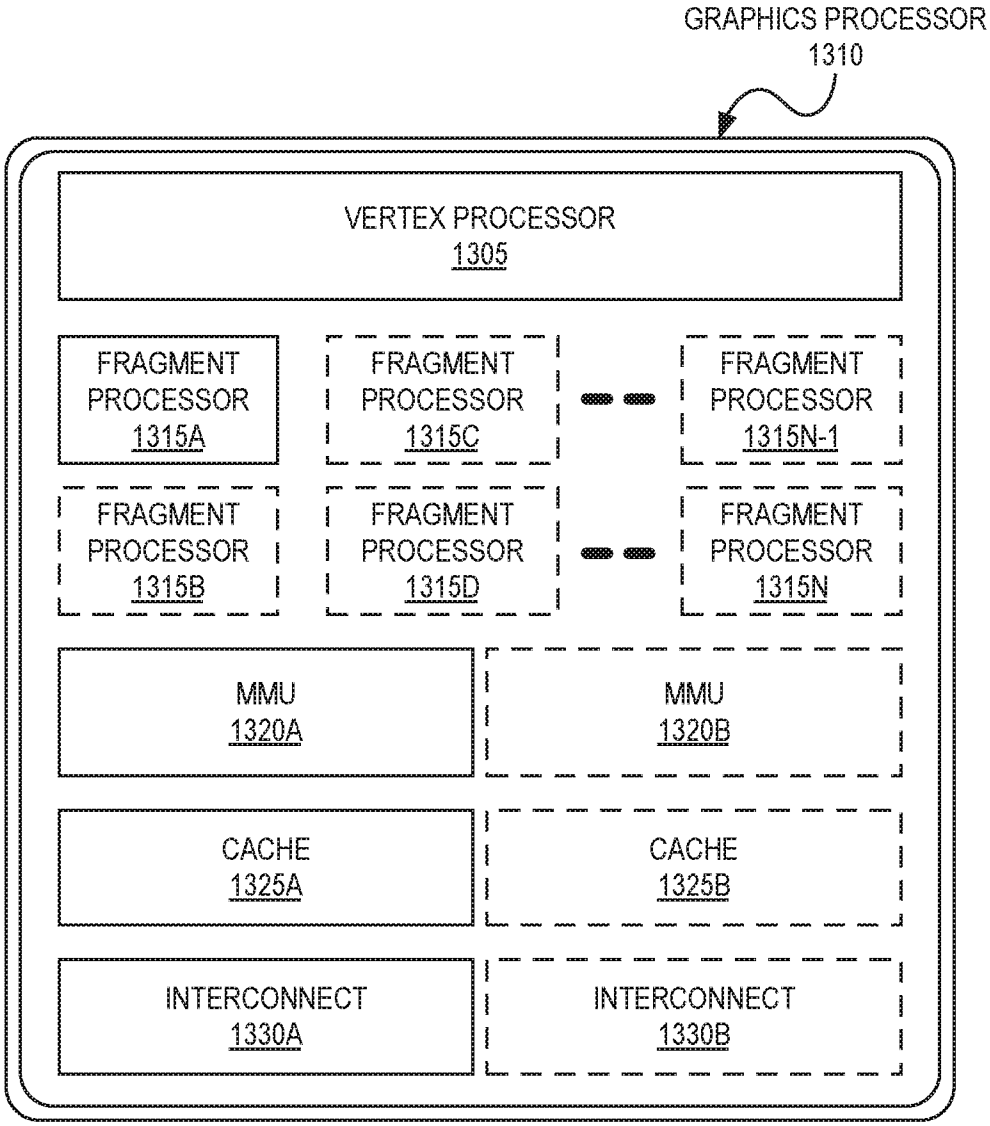

GRAPHICS PROCESSOR
1310
VERTEX PROCESSOR
1305
FRAGMENT PROCESSOR
1315A
FRAGMENT PROCESSOR
1315C
FRAGMENT PROCESSOR
1315N-1
FRAGMENT PROCESSOR
1315B
FRAGMENT PROCESSOR
1315D
FRAGMENT PROCESSOR
1315N
MMU
1320A
MMU
1320B
CACHE
1325A
CACHE
1325B
INTERCONNECT
1330A
INTERCONNECT
1330B

FIG. 13

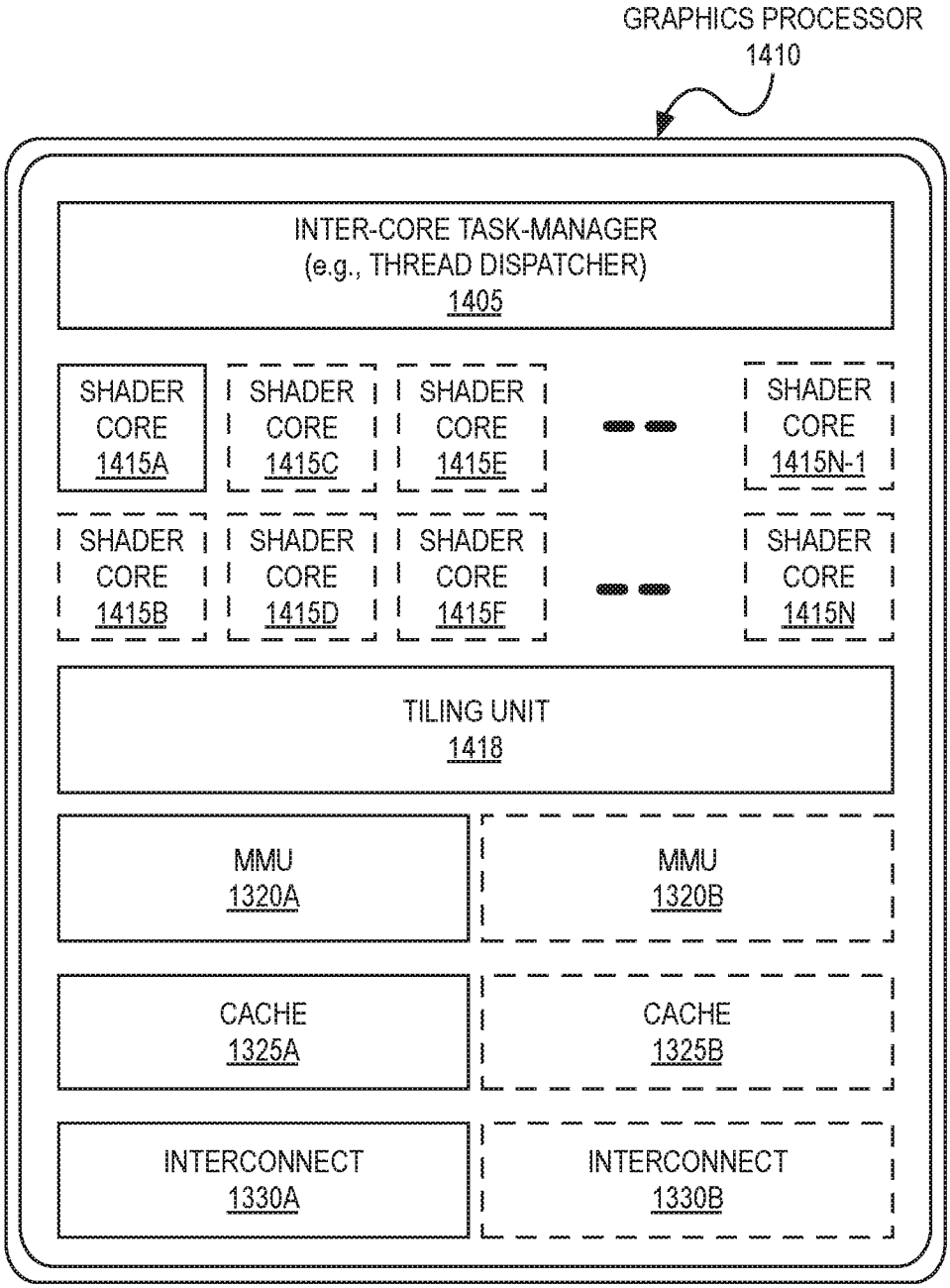

GRAPHICS PROCESSOR
1410
INTER-CORE TASK-MANAGER
(e.g., THREAD DISPATCHER)
1405
SHADER CORE 1415A
SHADER CORE 1415C
SHADER CORE 1415E
SHADER CORE 1415N-1
SHADER CORE 1415B
SHADER CORE 1415D
SHADER CORE 1415F
SHADER CORE 1415N
TILING UNIT
1418
MMU
1320A
MMU
1320B
CACHE
1325A
CACHE
1325B
INTERCONNECT
1330A
INTERCONNECT
1330B

FIG. 14

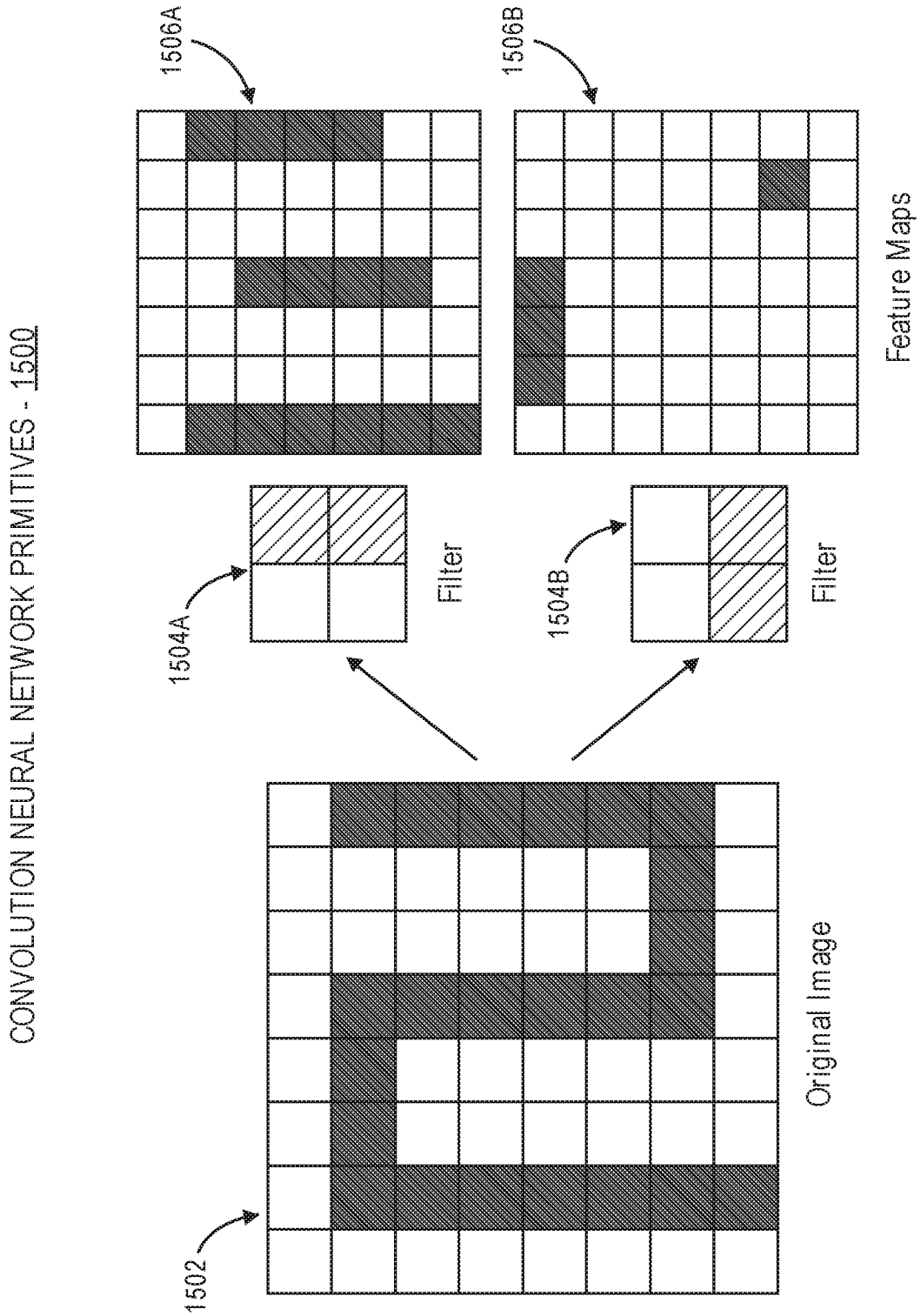

CONVOLUTION NEURAL NETWORK PRIMITIVES - 1500
1506A
1506B
Feature Maps
1504A
Filter
1504B
Filter
1502
Original Image

FIG. 15

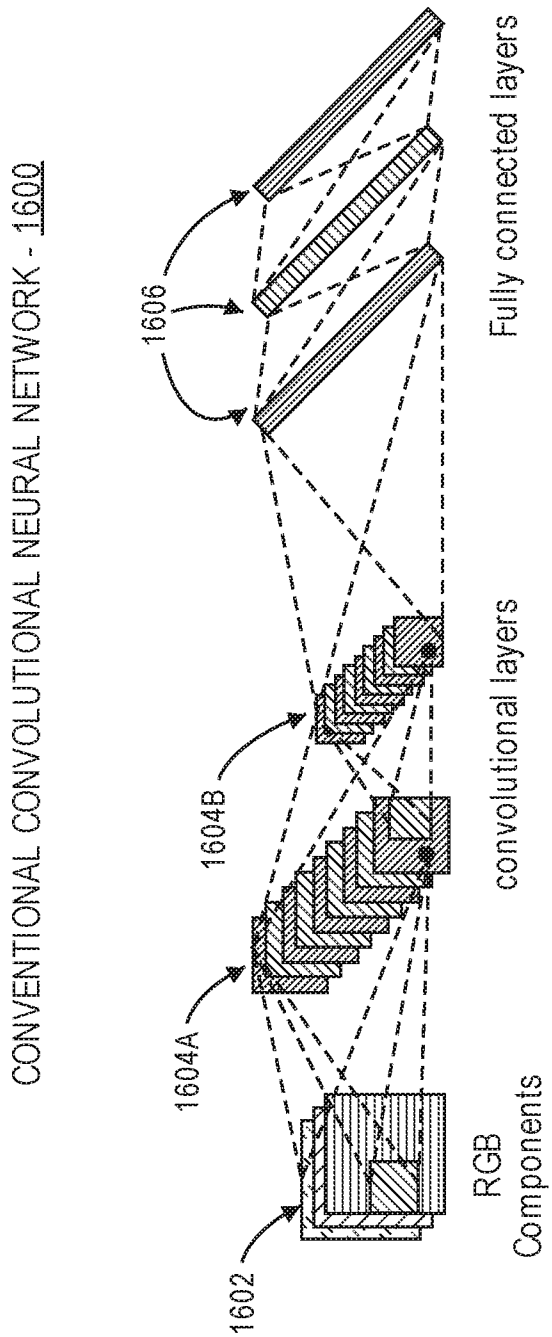


FIG. 16A

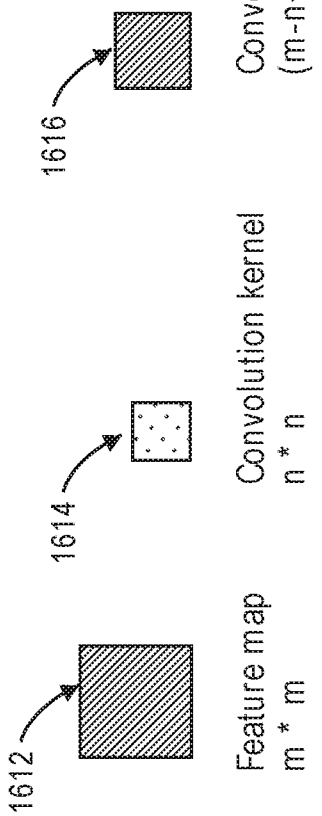


FIG. 16B

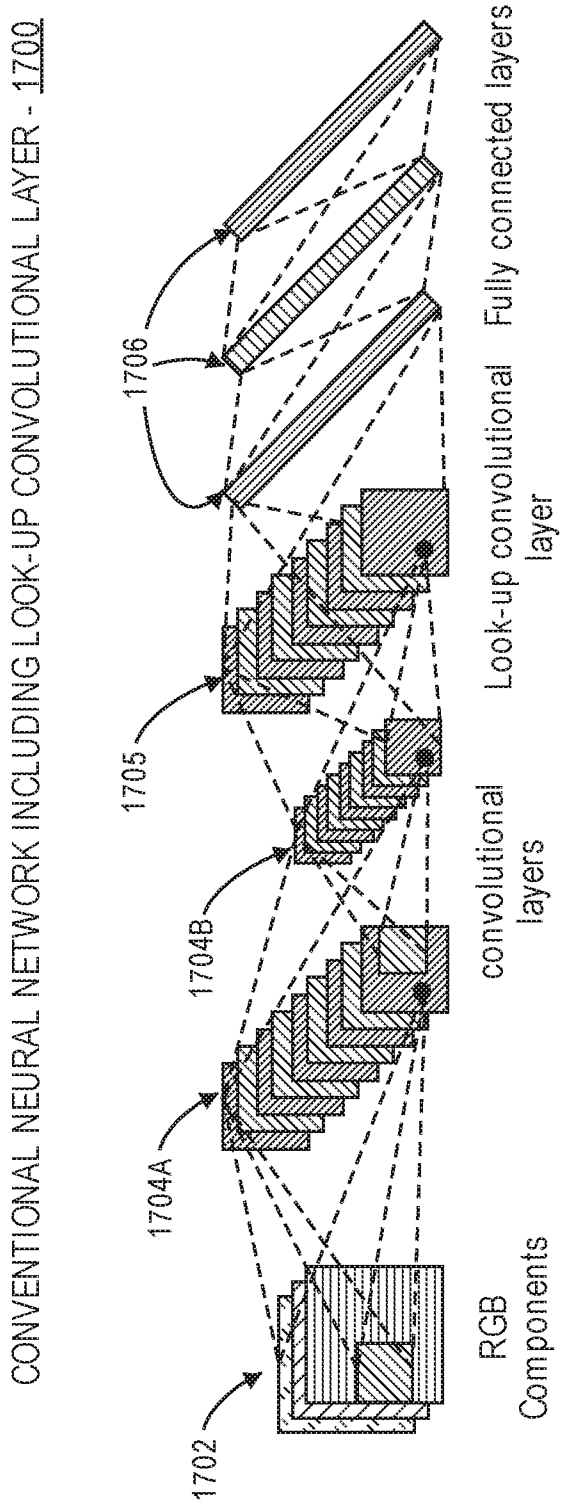


FIG. 17A

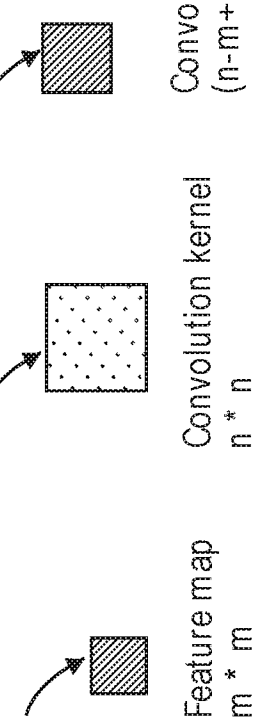


FIG. 17B

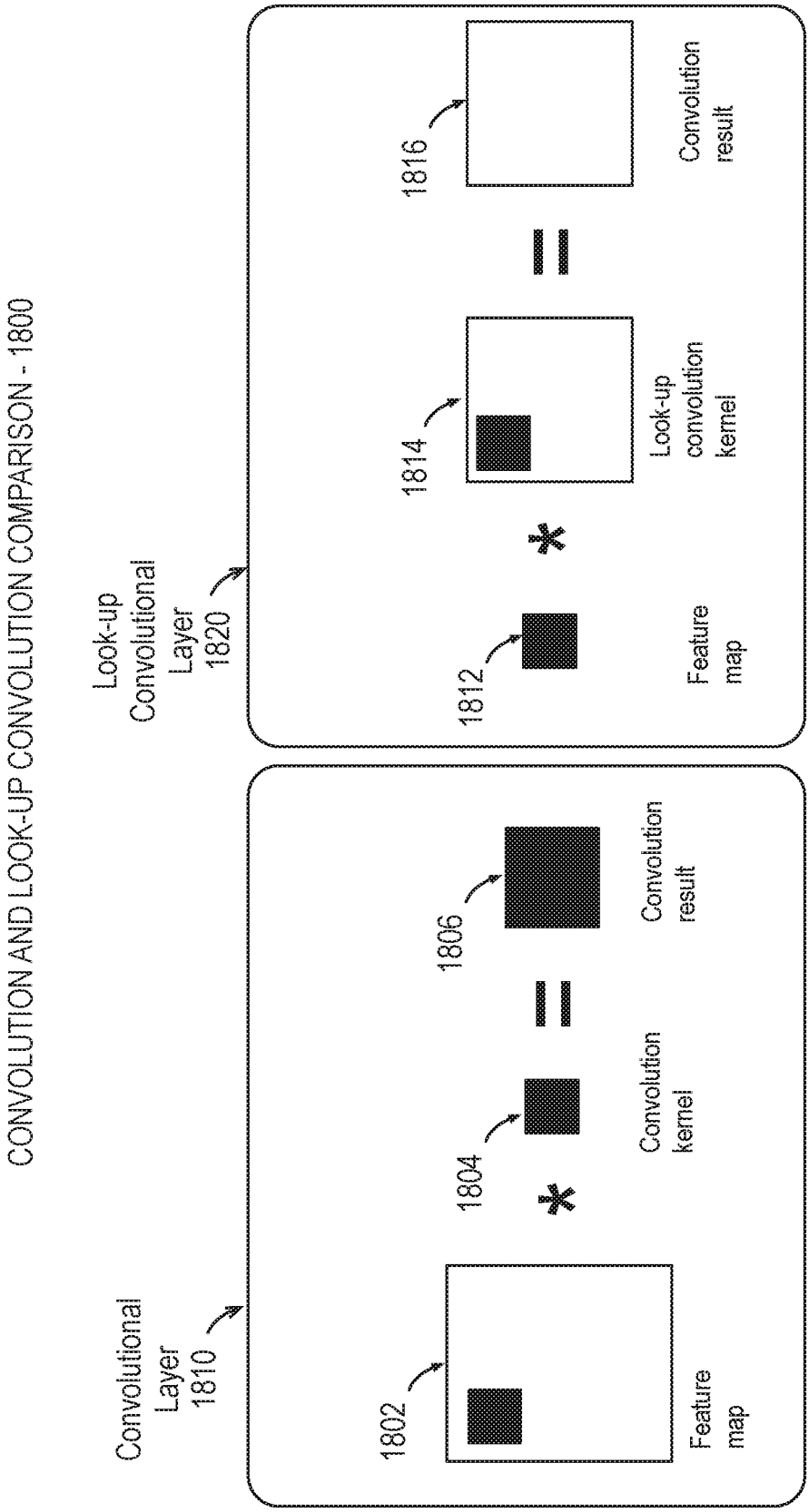

CONVOLUTION AND LOOK-UP CONVOLUTION COMPARISON - 1800
Convolutional Layer 1810
1802
Feature map
1804
Convolution kernel
1806
Convolution result
Look-up Convolutional Layer 1820
1812
Feature map
1814
Look-up convolution kernel
1816
Convolution result

FIG. 18

FIG. 19A

FIG. 19B

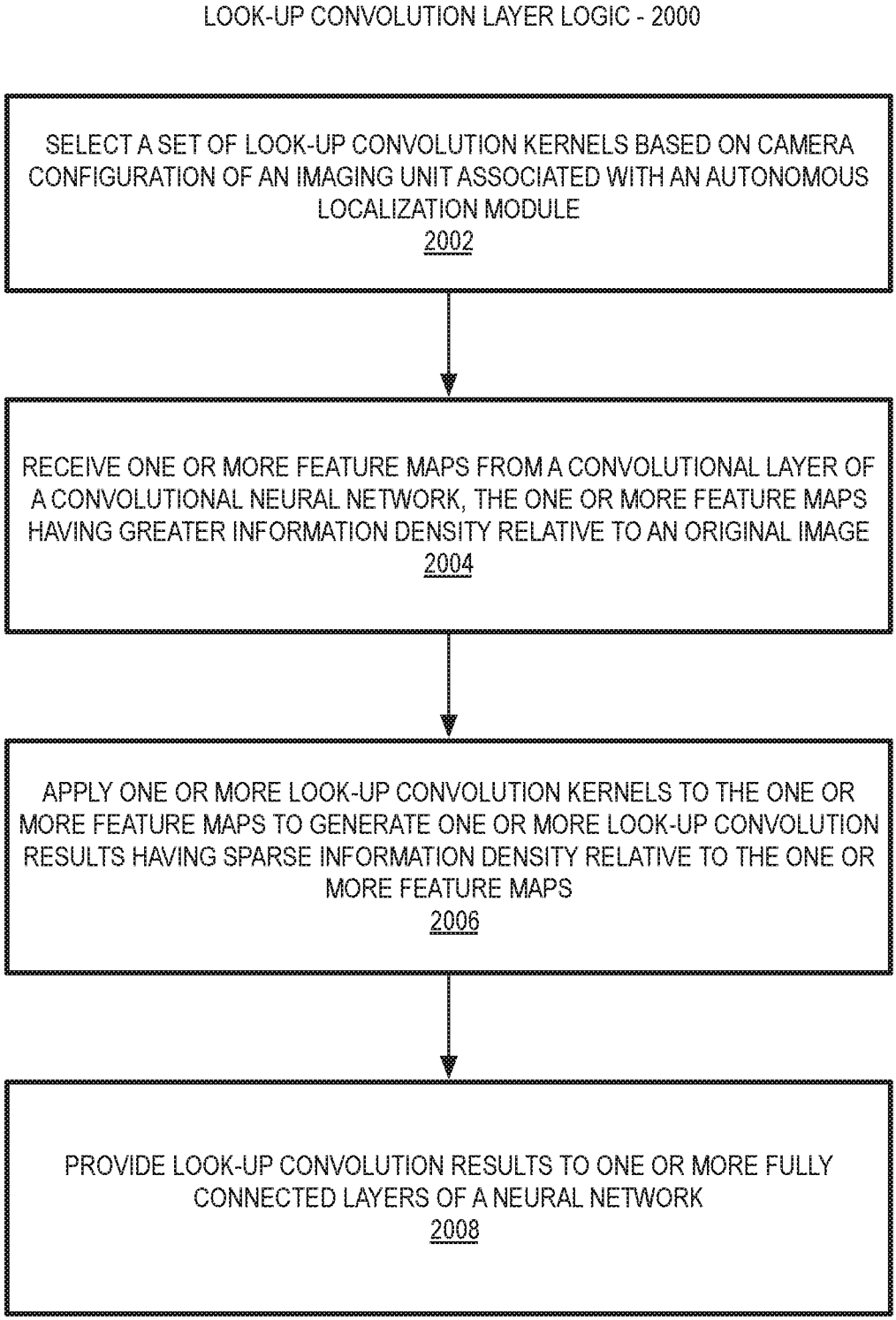

LOOK-UP CONVOLUTION LAYER LOGIC - 2000
SELECT A SET OF LOOK-UP CONVOLUTION KERNELS BASED ON CAMERA CONFIGURATION OF AN IMAGING UNIT ASSOCIATED WITH AN AUTONOMOUS LOCALIZATION MODULE
2002
RECEIVE ONE OR MORE FEATURE MAPS FROM A CONVOLUTIONAL LAYER OF A CONVOLUTIONAL NEURAL NETWORK, THE ONE OR MORE FEATURE MAPS HAVING GREATER INFORMATION DENSITY RELATIVE TO AN ORIGINAL IMAGE
2004
APPLY ONE OR MORE LOOK-UP CONVOLUTION KERNELS TO THE ONE OR MORE FEATURE MAPS TO GENERATE ONE OR MORE LOOK-UP CONVOLUTION RESULTS HAVING SPARSE INFORMATION DENSITY RELATIVE TO THE ONE OR MORE FEATURE MAPS
2006
PROVIDE LOOK-UP CONVOLUTION RESULTS TO ONE OR MORE FULLY CONNECTED LAYERS OF A NEURAL NETWORK
2008

FIG. 20

CAMERA POSE ESTIMATION LOGIC - 2100

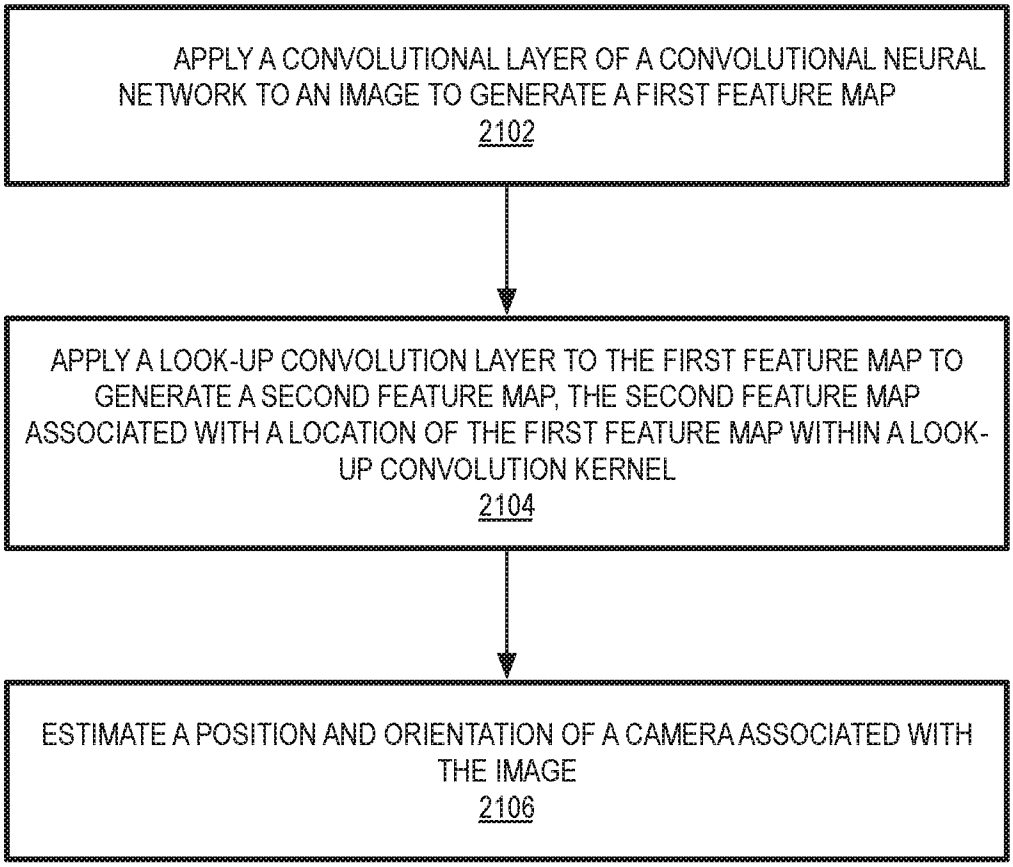


FIG. 21

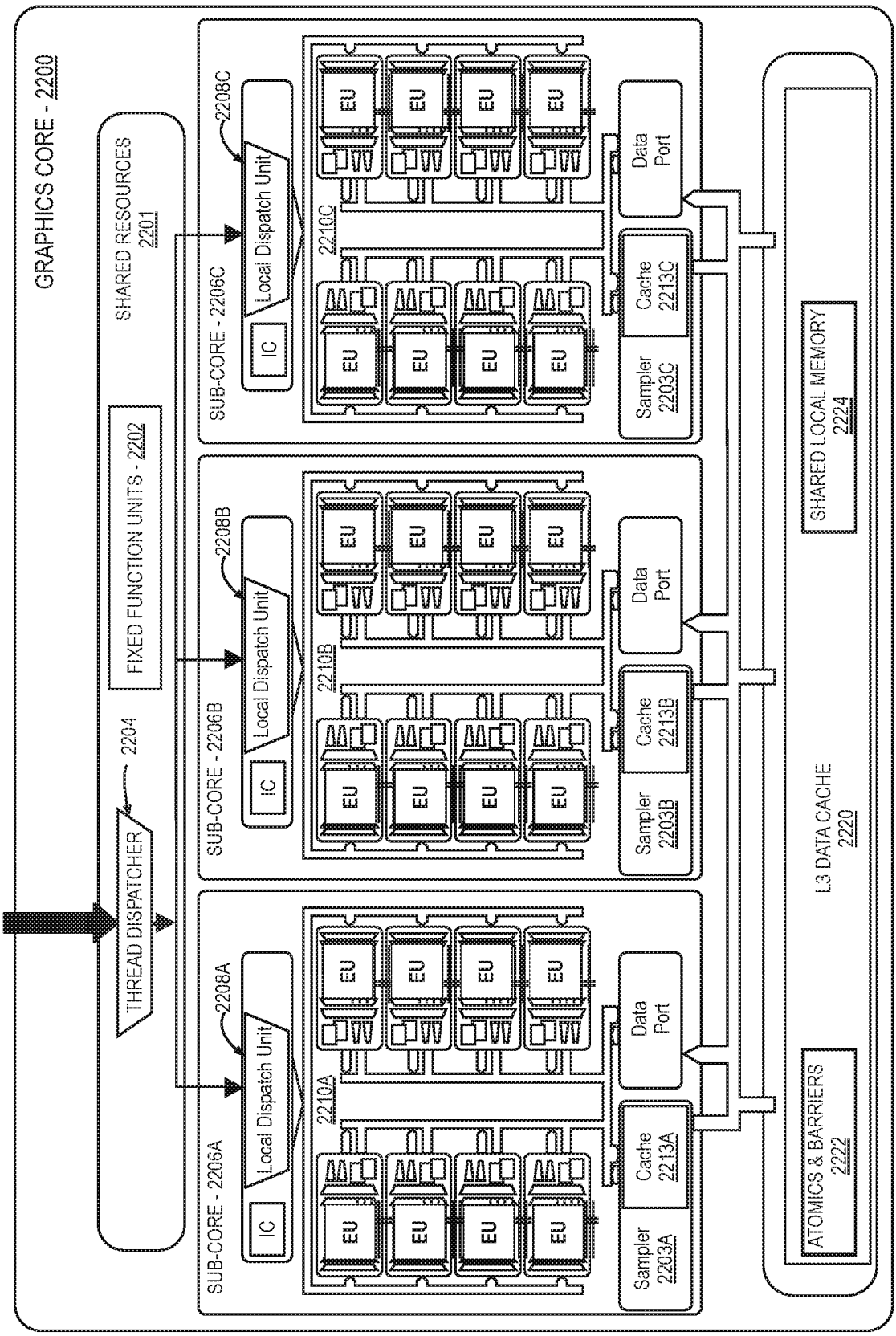

GRAPHICS CORE - 2200
SHARED RESOURCES 2201
FIXED FUNCTION UNITS - 2202
2204
THREAD DISPATCHER
SUB-CORE - 2206A
2208A
Local Dispatch Unit
IC
2210A
EU
Sampler 2203A
Cache 2213A
Data Port
SUB-CORE - 2206B
2208B
Local Dispatch Unit
IC
2210B
EU
Sampler 2203B
Cache 2213B
Data Port
SUB-CORE - 2206C
2208C
Local Dispatch Unit
IC
2210C
EU
Sampler 2203C
Cache 2213C
Data Port
L3 DATA CACHE 2220
SHARED LOCAL MEMORY 2224
ATOMICS & BARRIERS 2222

FIG. 22

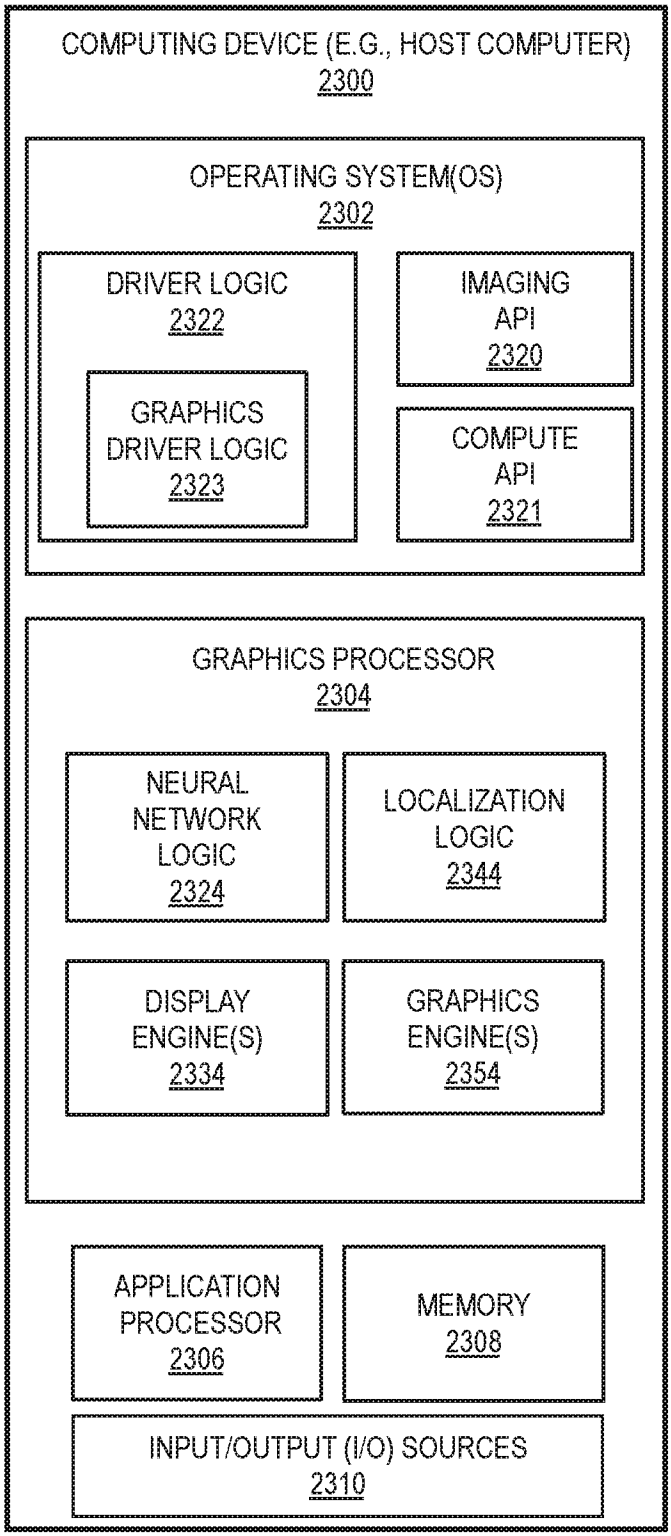

COMPUTING DEVICE (E.G., HOST COMPUTER)
2300
OPERATING SYSTEM(OS)
2302
DRIVER LOGIC
2322
GRAPHICS DRIVER LOGIC
2323
IMAGING API
2320
COMPUTE API
2321
GRAPHICS PROCESSOR
2304
NEURAL NETWORK LOGIC
2324
LOCALIZATION LOGIC
2344
DISPLAY ENGINE(S)
2334
GRAPHICS ENGINE(S)
2354
APPLICATION PROCESSOR
2306
MEMORY
2308
INPUT/OUTPUT (I/O) SOURCES
2310

FIG. 23

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2016/084621

A. CLASSIFICATION OF SUBJECT MATTER

G06K 9/66(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)

G06K G06F H04W H04L H04M H04N

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)

CNKI, CNPAT, IEEE, WPI, EPODOC: feature map, convolution kernel, environment panorama frame, camera frame, size, angle, pose, convolutional neural network, CNN, location, cylindrical, look-up layer, sphere, camera, global filter kernel

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2016148078 A1 (ADOBE SYSTEMS INCORPORAED) 26 May 2016 (2016-05-26) the whole document	1-23
A	CN 104573731 A (XIAMEN UNIVERSITY) 29 April 2015 (2015-04-29) the whole document	1-23
A	CN 104616032 A (ZHEJIANG GONGSHANG UNIVERSITY) 13 May 2015 (2015-05-13) the whole document	1-23
A	US 2007086655 A1 (MICROSOFT CORPORATION) 19 April 2007 (2007-04-19) the whole document	1-23
A	JI, Shuiwang et al. "3D Convolutional Neural Networks for Human Action Recognition" *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE,* Vol. 35, No. 1, 31 January 2013 (2013-01-31), the whole document	1-23

☐ 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	Date of mailing of the international search report
05 December 2016	**26 January 2017**

Name and mailing address of the ISA/CN	Authorized officer
STATE INTELLECTUAL PROPERTY OFFICE OF THE P.R.CHINA **6, Xitucheng Rd., Jimen Bridge, Haidian District, Beijing 100088** **China**	**CHAI,Hua**
Facsimile No. **(86-10)62019451**	Telephone No. **(86-10)62413256**

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2016/084621

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
US	2016148078	A1	26 May 2016	None	
CN	104573731	A	29 April 2015	None	
CN	104616032	A	13 May 2015	None	
US	2007086655	A1	19 April 2007	None	