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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
10 March 2011 (10.03.2011)

PCT

(10) International Publication Number
WO 2011/028984 A1

(51) International Patent Classification:
G06F 9/38 (2006.01) *G06F 9/50* (2006.01)
G06F 9/46 (2006.01)

(21) International Application Number:
PCT/US2010/047784

(22) International Filing Date:
3 September 2010 (03.09.2010)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/239,730 3 September 2009 (03.09.2009) US
12/616,636 11 November 2009 (11.11.2009) US

(71) Applicant *(for all designated States except US)*: **ADVANCED MICRO DEVICES, INC.** [US/US]; One Amd Place, Sunnyvale, CA 94088 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **SADOWSKI, Greg** [US/US]; 321 Harvard Street #303, Cambridge, MA 02139 (US). **IOURCHA, Konstantine** [US/US]; 7186 Wooded Lake Drive, San Jose, CA 95120 (US). **BROTHERS, John** [US/US]; 1257 Lakeside Drive, Apt. 1226, Sunnyvale, CA 94085 (US).

(74) Agents: **WOOD, Theodore, A.** et al.; Sterne, Kessler, Golstein & Fox P.L.L.C., 1100 New York Avenue, N.W., Washington, DC 20005 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, 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, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, 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, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, 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, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

[Continued on next page]

(54) Title: AN INTERNAL, PROCESSING-UNIT MEMORY FOR GENERAL-PURPOSE USE

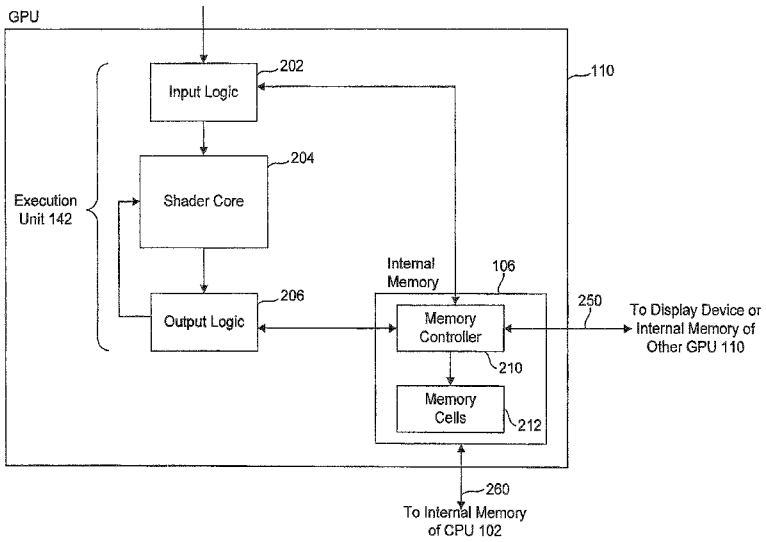

FIG. 2

(57) Abstract: Disclosed herein is a graphics-processing unit (GPU) having an internal memory for general-purpose use and applications thereof. Such a GPU includes a first internal memory, an execution unit coupled to the first internal memory, and an interface configured to couple the first internal memory to a second internal memory of an other processing unit. The first internal memory may comprise a stacked dynamic random access memory (DRAM) or an embedded DRAM. The interface may be further configured to couple the first internal memory to a display device. The GPU may also include another interface configured to couple the first internal memory to a central processing unit. In addition, the GPU may be embodied in software and/or included in a computing system.

WO 2011/028984 A1

Published:

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

AN INTERNAL, PROCESSING-UNIT MEMORY FOR GENERAL-PURPOSE USE

BACKGROUND OF THE INVENTION

Field of the Invention

[0001] The present invention is generally directed to computing devices (e.g., computers, embedded devices, hand-held devices, and the like). More particularly, the present invention is directed to memory used by processing units of such computing devices.

Background Art

[0002] A computing device typically includes one or more processing units, such as a central-processing unit (CPU) and a graphics-processing unit (GPU). The CPU coordinates the activities of the computing device by following a precise set of instructions. The GPU assists the CPU by performing data-parallel computing tasks, such as graphics-processing tasks and/or physics simulations which may be required by an end-user application (e.g., a video-game application). The GPU and CPU may be part of separate devices and/or packages or may be included in the same device and/or package. Further, each processing unit may be included in another larger device. For example, GPUs are frequently integrated into routing or bridge devices such as, for example, Northbridge devices.

[0003] There are several layers of software between the end-user application and the GPU. The end-user application communicates with an application-programming interface (API). An API allows the end-user application to output graphics data and commands in a standardized format, rather than in a format that is dependent on the GPU. Several types of APIs are commercially available — including DirectX® developed by Microsoft Corporation of Redmond, Washington; OpenGL® and OpenCL maintained by Khronos Group. The API communicates with a driver. The driver translates standard code received from the API into a native format of instructions understood by the GPU. The driver is typically written by the manufacturer of the GPU. The GPU then executes the instructions from the driver.

[0004] In a conventional system, the CPU and GPU are each typically coupled to an external memory. The external memory may include instructions to be executed and/or data to be used by the CPU and/or GPU. The external memory may be, for example, a dynamic random-access memory (DRAM). The external memory can be configured to be quite large, thereby providing ample storage capacity to each processing unit to which it's coupled. Unfortunately, accessing the external memory may take several hundred clock cycles. Accordingly, an external memory may not provide memory sufficient bandwidth or fast memory access for high-end GPUs.

[0005] One potential solution for providing sufficient memory bandwidth to a GPU is to provide the GPU with an internal memory. The internal memory may be, for example, an embedded or stacked DRAM. Compared to external memory, an internal memory provides higher bandwidth, faster memory access, and consumes less power. However, the capacity of the internal memory cannot easily be scaled to meet the storage demands of high-end GPUs. For example, a high-end GPU may require more memory than can be included in an internal memory of the GPU.

[0006] Given the foregoing, what is needed is memory, and applications thereof, that provide both sufficient memory capacity (like external memory) and high bandwidth (like embedded memory).

BRIEF SUMMARY OF EMBODIMENTS OF THE INVENTION

[0007] Embodiments of the present invention meets the above-described needs by providing an internal, processing-unit memory for general-purpose use and applications thereof. The internal, processing-unit memory of embodiments of the present invention provides high bandwidth because it is embedded within a processing unit. It also provides sufficient storage capacity because a plurality of processing-unit memories may be combined into a sufficiently large memory pool.

[0008] For example, an embodiment of the present invention provides a GPU. The GPU includes a first internal memory, an execution unit coupled to the first internal memory, and an interface configured to couple the first internal memory to a second internal memory of another processing unit. In an embodiment, the GPU is embodied in software. In another embodiment, the GPU is included in a system. The system may comprise, for example, a supercomputer, a desktop computer, a laptop computer,

a video-game console, an embedded device, a handheld device (e.g., a mobile telephone, smart phone, MP3 player, a camera, a GPS device, or the like), or another system that includes or is configured to include a GPU.

[0009] Further features and advantages of the invention, as well as the structure and operation of various embodiments of the invention, are described in detail below with reference to the accompanying drawings. It is noted that the invention is not limited to the specific embodiments described herein. Such embodiments are presented herein for illustrative purposes only. Additional embodiments will be apparent to persons skilled in the relevant art(s) based on the teachings contained herein.

BRIEF DESCRIPTION OF THE DRAWINGS/FIGURES

[0010] The accompanying drawings, which are incorporated herein and form part of the specification, illustrate the present invention and, together with the description, further serve to explain the principles of the invention and to enable a person skilled in the relevant art(s) to make and use the invention.

[0011] FIGS. 1A and 1B illustrate example systems that include internal, processing-unit memories for general-purpose use in accordance with an embodiment of the present invention.

[0012] FIG. 2 illustrates details of an example GPU having an internal memory for general-purpose use in accordance with an embodiment of the present invention.

[0013] FIG. 3 illustrates an example stacked memory that may be included in a processing element in accordance with an embodiment of the present invention.

[0014] FIG. 4 illustrates an example method implemented by the GPU of FIG. 2 in accordance with an embodiment of the present invention.

[0015] The features and advantages of the present invention will become more apparent from the detailed description set forth below when taken in conjunction with the drawings, in which like reference characters identify corresponding elements throughout. In the drawings, like reference numbers generally indicate identical, functionally similar, and/or structurally similar elements. The drawing in which an element first appears is indicated by the leftmost digit(s) in the corresponding reference number.

DETAILED DESCRIPTION OF EMBODIMENTS OF THE INVENTION

I. Overview

[0016] The present invention provides an internal, GPU memory for general-purpose use and applications thereof. In the detailed description that follows, references to "one embodiment," "an embodiment," "an example embodiment," etc., indicate that the embodiment described may include a particular feature, structure, or characteristic, but every embodiment may not necessarily include the particular feature, structure, or characteristic. Moreover, such phrases are not necessarily referring to the same embodiment. Further, when a particular feature, structure, or characteristic is described in connection with an embodiment, it is submitted that it is within the knowledge of one skilled in the art to affect such feature, structure, or characteristic in connection with other embodiments whether or not explicitly described.

[0017] In accordance with an embodiment, a GPU includes an internal memory (e.g., an embedded or stacked DRAM) that is configured to be used by one or more other processing units. The GPU includes an interface and implements a protocol, allowing the one or more other GPUs access to its internal memory. The interface may provide each other GPUs dedicated access to the internal memory or may provide the other GPUs shared access to the internal memory. Access to the internal memory of the GPU may be controlled by the GPU itself or each other GPUs.

[0018] In an embodiment, the interface and protocol allows the internal memory to be combined with external memories, forming a larger memory pool accessible to the GPU. The external memories may be included in other GPUs. In an embodiment, for example, a computing device includes a plurality of GPUs, wherein each GPU includes an internal memory that is configured to be shared with the other GPUs. In this embodiment, the internal memory of each GPU is combined into a unified memory pool. The size of the memory pool scales with the number of participating GPUs. Any participating GPU may use the memory pool for its storage needs.

[0019] Further details of an example GPU in accordance with an embodiment of the present invention are described below. Before providing these details, however, it is

helpful to describe an example computing device in which such GPUs may be implemented.

II. An Example Computing System

[0020] FIGS. 1A and 1B illustrate an example computing system 100 having a plurality of GPUs, each including an internal memory configured for general-purpose use in accordance with embodiments of the present invention. Compared to external memories, the internal memories provide each GPU higher bandwidth access to data. In addition, the internal memories of each GPU can be combined into a larger memory pool accessible by each GPU, thereby providing sufficient storage capacity to each GPU.

[0021] In the embodiment of FIG. 1A, each GPU is given dedicated access to the internal memory of another GPU. In the embodiment of FIG. 1B, each GPU has shared access to the internal memories of the other GPU via a shared interface. In embodiments, computing system 100 may comprise a supercomputer, a desktop computer, a laptop computer, a video-game console, an embedded device, a handheld device (e.g., a mobile telephone, smart phone, MP3 player, a camera, a GPS device, or the like), or some other device that includes or is configured to include a CPU and/or GPU.

[0022] Referring to FIGS. 1A and 1B, computing device 100 includes a CPU 102, a first GPU 110A, and a second GPU 110B. CPU 102 executes instructions to control the functionality of computing device 100. GPUs 110 assist CPU 102 by performing data-parallel processing tasks (such as, for example, graphics-processing tasks and/or general-compute tasks). Based on their design, GPUs 110 can typically perform data-parallel processing tasks faster than CPU 102 could perform them in software.

[0023] First GPU 110A and second GPU 110B each include their own internal memory and execution unit. Specifically, first GPU 106A includes an internal memory 106A and an execution unit 142A; and second GPU 106B includes an internal memory 106B and an execution unit 142B. Similarly, CPU 102 includes a cache memory 130 and an execution unit 132. Internal memories 106 (and optionally cache memory 130) are available to GPUs 110 to provide faster access and higher bandwidth to certain data than would be possible if the data were externally stored

(e.g., if the data were stored in a system memory 104). The internal memories 106 may comprise, for example, embedded or stacked DRAM.

[0024] Internal memories 106A, 106B (and optionally cache memory 130) may be combined into a larger memory pool to provide substantial storage capacity (e.g., more than 4 GB), while also providing fast, high bandwidth memory access. Although conventional external memories may provide sufficient storage capacity (e.g., more than 4 GB), conventional external memories provide insufficient bandwidth for certain high-end uses. Similarly, although conventional embedded memories may provide sufficient bandwidth for these high-end uses, conventional embedded memories provide insufficient storage capacity (e.g., fewer than 4 GB) for these high-end uses. Unlike conventional external memories and/or conventional embedded memories, embodiments of the present invention not only provide sufficient storage capacity (e.g., more than 4 GB), but also provide high bandwidth by providing GPUs that include internal memories which are available to other GPUs for general-purpose use.

[0025] For example, a frame buffer (i.e., a buffer that stores a complete frame of data to be displayed on a display device) of a high-end GPU may require high bandwidth access to a substantially large memory (e.g., more than 4 gigabytes (GB)). In embodiments, first GPU 110A may use internal memories 106A, B and optionally cache memory 130 of CPU 102 to define the frame buffer of first GPU 110A. Similarly, second GPU 110B may also use internal memories 106A, B and optionally cache memory 130 of CPU 102 to define the frame buffer of second GPU 110B. In this way, unlike conventional external or embedded memories, the frame buffer defined in accordance with embodiments of the present invention provides high bandwidth access to a substantially large memory (e.g., more than 4 GB).

[0026] In the embodiment of FIG. 1A, each GPU 110 is given dedicated access to internal memory 106 of another processing unit, as alluded to above. Specifically, a first interface 101 provides first GPU 110A dedicated access to internal memory 106B of second GPU 110B and provides second GPU 110B dedicated access to internal memory 106A of first GPU 110A. Data may be written to or retrieved from either internal memory 106A or internal memory 106B based on an address range of the data. For example, internal memory 106A may be assigned a first address range (e.g.,

less than a first predetermined address *A* and greater than or equal to a second predetermined address *B*), and internal memory 106B may be assigned a second address range (e.g., all addresses not within the first address range). It is to be appreciated, however, that other schemes for writing data to and retrieving data from internal memory 106A and/or internal memory 106B may be implemented without deviating from the spirit and scope of the present invention, provided first GPU 110A and second GPU 110B can each have access to internal memory 106A of first GPU 110A and internal memory 106B of second GPU 110B.

[0027] In an embodiment, first interface 101 comprises a display controller interface. The display controller interface provides a display device 140 access to the frame buffer of a GPU. By incorporating the display controller interface into first interface 101, first interface 101 can be provided on a standard pin that is already included in conventional GPU designs.

[0028] In addition to first interface 101, a second interface 103 provides CPU 102 dedicated access to internal memory 106B of second GPU 110B and provides second GPU 110B dedicated access to cache memory 130 of CPU 102. In this way, second GPU 110B and CPU 102 can each have access to internal memory 106B of second GPU 110B and cache memory 130 of CPU 102. Likewise, a third interface 105 provides first GPU 110A dedicated access to cache memory 130 of CPU 102 and provides CPU 102 dedicated access to internal memory 106A of first GPU 110A. In this way, first GPU 110A and CPU 102 can each have access to internal memory 106A of first GPU 110A and cache memory 130 of CPU 102.

[0029] In the embodiment of FIG. 1B, each processing unit has shared access to the internal memories of the other processing units via a shared interface 164. Shared interface 164 provides each processing unit (e.g., first GPU 110A, second GPU 110B, and CPU 102) high bandwidth access to the internal memory of the other processing units. Data may be written to or retrieved from internal memory 106A, internal memory 106B, or cache memory 130 based on an address range of the data. For example, internal memory 106A may be assigned a first address range; internal memory 106B may be assigned a second address range; and cache memory 130 may be assigned a third address range. It is to be appreciated, however, that other schemes for writing data to and retrieving data from internal memory 106A, internal memory

106B, and/or cache memory 130 may be implemented without deviating from the spirit and scope of the present invention, provided first GPU 110A, second GPU 110B, and CPU 102 can each have access to internal memory 106A of first GPU 110A, internal memory 106B of second GPU 110B, and cache memory 130 of CPU 102.

[0030] In embodiments, computing device 100 also includes a system memory 104, a secondary memory 120, an input-output (I/O) interface 116, and/or display device 140. System memory 104 stores information that is frequently accessed by programs running on CPU 102. System memory 104 typically comprises volatile memory, meaning that data stored in system memory 104 are lost when power to computing device 100 is turned off. Secondary memory 120 stores data and/or applications used by computing device 100. Secondary memory 120 typically has much larger storage capacity compared to system memory 104 and typically comprises non-volatile (persistent) memory, meaning that data stored in secondary memory 120 persists even when power to computing device 100 is turned off. I/O interface 116 allows computing device 100 to be coupled an external device 116 (such as, an external display device, an external storage device (e.g., video-game cartridge, CD, DVD, flash drive, or the like), a network card, or some other type of external device). Display device 140 displays content of computing device 100. Display device may comprise a cathode ray tube, a liquid-crystal display (LCD), a plasma screen, or some other type of display device whether now known or later developed.

[0031] GPUs 110 and CPU 102 communicate with each other and system memory 104, secondary memory 120, and I/O interface 116 over a bus 114. Bus 114 may be any type of bus used in computing devices, including a peripheral component interface (PCI) bus, an accelerated graphics port (AGP) bus, a PCI Express (PCIE) bus, or another type of bus whether presently available or developed in the future.

[0032] In embodiments, computing device 100 may include a video processing unit (VPU) in lieu of or in addition to GPU 110. For example, in an embodiment, computing device 100 includes GPU 110A, CPU 102; and in lieu of GPU, 110B illustrated in FIGS. 1A and 1B, computing device 100 includes a VPU. In this way, CPU 102 can perform general processing functions, GPU 110A can perform graphics-processing functions, and the VPU can perform video-processing functions.

III. An Example GPU

[0033] FIG. 2 illustrates example details of GPU 110 having an internal memory 106. In accordance with an embodiment of the present invention, internal memory 106 can be used by another GPU, or a CPU, to increase overall system performance by combining the graphics processing power based on an augmented memory footprint size.

[0034] As mentioned above, GPU 110 includes execution unit 142 and internal memory 106. Referring to FIG. 2, execution unit 142 includes input logic 202, a shader core 204, and output logic 206. Internal memory 106 includes a memory controller 210 and memory cells 212. Memory controller 210 controls access to memory cells 212. Memory cells 212 store data.

[0035] In an embodiment, internal memory 106 comprises an embedded, dynamic random access memory (DRAM). An embedded DRAM is a memory encapsulated in a common package with a processing unit. In another embodiment, internal memory 106 comprises a stacked DRAM, as illustrated in FIG. 3. A stacked memory includes a plurality of memory elements stacked on top of each other in a three-dimensional structure.

[0036] Internal memory 106 is coupled to execution unit 142 via both input logic 202 and output logic 206. In particular, input logic 202 can retrieve data from internal memory 106, and output logic 206 can send data to internal memory 106 to be stored in memory cells 212.

[0037] Internal memory 106 may also be coupled to the internal memory of another GPU via a first interface 250. Coupling internal memory 106 to the internal memory of another GPU can increase the total memory pool available to execution unit 142. In an embodiment, first interface 250 provides dedicated access between internal memory 106 of GPU 110 and an internal memory of another GPU, as illustrated by interface 101 of FIG. 1A. In this embodiment, first interface 250 is a provided on a standard pin of a conventional GPU. For example, first interface 250 may comprise a display-controller interface, which provides a display device access to a local frame buffer included in internal memory 106. In another embodiment, first interface 250 provides shared access between internal memory 106 of GPU 110 and internal memories of other processing units, as illustrated by interface 164 of FIG. 1B.

[0038] Internal memory 106 may also be coupled to cache memory 130 of CPU 102 via a second interface 260. Accordingly, the combination of internal memory 106 and cache memory 130 can increase the memory pool available to GPU 110. In an embodiment, second interface 260 provides a dedicated connection between internal memory 106 of GPU 110 and cache memory 130 of CPU 102, such as connection 103 or connection 105 of FIG. 1A. In another embodiment, second interface 260 provides a connection that is shared by only GPU 110 and CPU 102, such as connection 164 of FIG. 1B. In a further embodiment, second interface couples GPU 110 to CPU 102 on a common bus, such as bus 114 of FIGS. 1A and 1B.

IV. Example Operation of GPU 110

[0039] FIG. 4 illustrates an example method 400 implemented by GPU 110 in accordance with an embodiment of the present invention. Method 400 is described below with reference to FIGS. 3 and 4.

[0040] Method 400 begins at a step 402 in which instruction are received. In an embodiment, input logic 202 receives instructions to be executed by GPU 110. The instructions may comprise, for example, a graphics-processing task or a data-parallel processing task provided by an end-user application running on CPU 102 of system 100.

[0041] In a step 404, a location of data associated with an instruction is identified. In one example, the data may be included with a received instruction. Such data is commonly referred to as immediate data. In another example, the instruction provides the location of the data. For instance, the instruction may include an address within which the data is stored. In a further example, the instruction includes information from which input logic 202 computes the address within which the data is stored. The data may be stored in either internal memory 106, an internal memory of another GPU to which internal memory 106 is coupled, or cache memory 130 of CPU 102.

[0042] In a step 406, the data is retrieved. If the data is immediate data, input logic 202 simply extracts the immediate data from the instruction. If the data is stored in internal memory 106 or a memory to which internal memory 106 is coupled, input logic 202 sends a request to memory controller 210 for access to the data. If, on the one hand, the data is stored in memory cells 212, the data is retrieved and provided to

input logic 202. If, on the other hand, the data is stored in another memory coupled to internal memory 106, the request from input logic 202 is forwarded to the other memory via interface 250 or interface 260. The data is then retrieved from the other memory and provided to input logic 202.

[0043] In a step 408, the instruction is executed. Shader core 204 executes the instruction based on the data obtained by input logic 202 in step 406.

[0044] In a step 410, results of the instruction execution are provided to output logic 206. Output logic 206 determines whether further processing is required on these results, as indicated in decision step 412. Results provides to output logic 206 may have a flag or some other indicia to indicate whether additional processing is necessary. If in decision step 412 output logic 206 determines that further processing is necessary, then output logic 206 forwards the results back to shader core 204 and steps 408 and 410 of method 400 are repeated. If, on the other hand, output logic 206 determines, in decision step 412, that no further processing is necessary, then output logic 206 provides the results to internal memory 106, as indicated in step 414.

[0045] The results may then be written to internal memory 106 or to a memory coupled to internal memory 106, depending on the address to which the results are to be written. If the results are to be written to internal memory 106, memory controller 210 provides access to the appropriate address in memory cells 212 and the results are stored therein. If, on the other hand, the results are to be written to a memory coupled to internal memory 106, then memory controller 210 forwards the results to the other memory via interface 250 or interface 260 and the results are stored in memory cells of the other memory.

V. Example Software Implementations

[0046] In addition to hardware implementations of GPU 110, such GPUs may also be embodied in software disposed, for example, in a computer-readable medium configured to store the software (*e.g.*, a computer-readable program code). The computer-readable program code enables embodiments of the present invention, including the following embodiments: (i) the functions of the systems and techniques disclosed herein (such as, providing tasks to GPU 110, scheduling tasks in GPU 110, executing tasks in GPU 110, or the like); (ii) the fabrication of the systems and

techniques disclosed herein (such as, the fabrication of GPU 110); or (iii) a combination of the functions and fabrication of the systems and techniques disclosed herein.

[0047] This can be accomplished, for example, through the use of general-programming languages (such as C or C++), hardware-description languages (HDL) including Verilog HDL, VHDL, Altera HDL (AHDL) and so on, or other available programming and/or schematic-capture tools (such as circuit-capture tools). The computer-readable program code can be disposed in any known computer-readable medium including semiconductor, magnetic disk, or optical disk (such as CD-ROM, DVD-ROM). As such, the computer-readable program code can be transmitted over communication networks including the Internet and internets. It is understood that the functions accomplished and/or structure provided by the systems and techniques described above can be represented in a core (such as a shader core) that is embodied in computer-readable program code and may be transformed to hardware as part of the production of integrated circuits.

VI. Conclusion

[0048] Described above is an internal, GPU memory for general-purpose use and applications thereof. It is to be appreciated that the Detailed Description section, and not the Summary and Abstract sections, is intended to be used to interpret the claims. The Summary and Abstract sections may set forth one or more but not all exemplary embodiments of the present invention as contemplated by the inventor(s), and thus, are not intended to limit the present invention and the appended claims in any way.

WHAT IS CLAIMED IS:

1. A graphics processing unit (GPU), comprising:

a first internal memory;

an execution unit coupled to the first internal memory; and

an interface configured to couple the first internal memory to a second internal memory of an other processing unit.

2. The processing unit of claim 1, wherein the other processing unit comprises a GPU.

3. The processing unit of claim 1, wherein the other processing unit comprises a central processing unit.

4. The processing unit of claim 1, wherein the first internal memory comprises a stacked dynamic random-access memory.

5. The processing unit of claim 1, wherein the first internal memory comprises an embedded dynamic random-access memory.

6. The processing unit of claim 1, wherein the interface is further configured to couple the first internal memory to a display device.

7. A computer-program product comprising a computer-readable storage medium containing instructions that, if executed on a computing device, define a graphics-processing unit (GPU), wherein the GPU comprises:

a first internal memory;

an execution unit coupled to the first internal memory; and

an interface configured to couple the first internal memory to a second internal memory of another processing unit.

8. The computer-program product of claim 7, wherein the other processing unit comprises a GPU.

9. The computer-program product of claim 7, wherein the other processing unit comprises a central processing unit.

10. The computer-program product of claim 7, wherein the first internal memory of the GPU comprises a stacked dynamic random-access memory.

11. The computer-program product of claim 7, wherein the first internal memory of the GPU comprises an embedded dynamic random-access memory.

12. The computer-program product of claim 7, wherein the GPU embodied in hardware description language software.

13. The computer-program product of claim 7, wherein the GPU is embodied in one of Verilog hardware description language software, Verilog-A hardware description language software, and VHDL hardware description language software.

14. A system, comprising:

a first graphics-processing unit (GPU) comprising a first internal memory, a first execution unit coupled to the first internal memory, and a first interface configured to couple the first internal memory to an internal memory of another GPU; and

a second GPU comprising a second internal memory, a second execution unit coupled to the second internal memory, and a second interface configured to couple the second internal memory to an internal memory of another GPU;

wherein the first internal memory and the second internal memory are coupled together, enabling the first execution unit of the first GPU to access the second internal memory of the second GPU and enabling the second execution unit of the second GPU to access the first internal memory of the first GPU.

15. The system of claim 14, wherein the first internal memory comprises a stacked dynamic random-access memory.

16. The system of claim 14, wherein the first internal memory comprises an embedded dynamic random-access memory.

17. The system of claim 16, wherein:

the first interface is further configured to couple the first internal memory to the display device; and

the second interface is further configured to couple the second internal memory to the display device.

18. The system of claim 14, further comprising:

an external memory;

a central-processing unit (CPU) comprising a cache memory; and

a bus coupled between the external memory and the CPU.

19. The system of claim 18, wherein the first GPU further comprises an other interface configured to couple the first internal memory to the cache memory of the CPU.

20. The system of claim 18, wherein the second GPU further comprises an other interface configured to couple the second internal memory to the cache memory of the CPU.

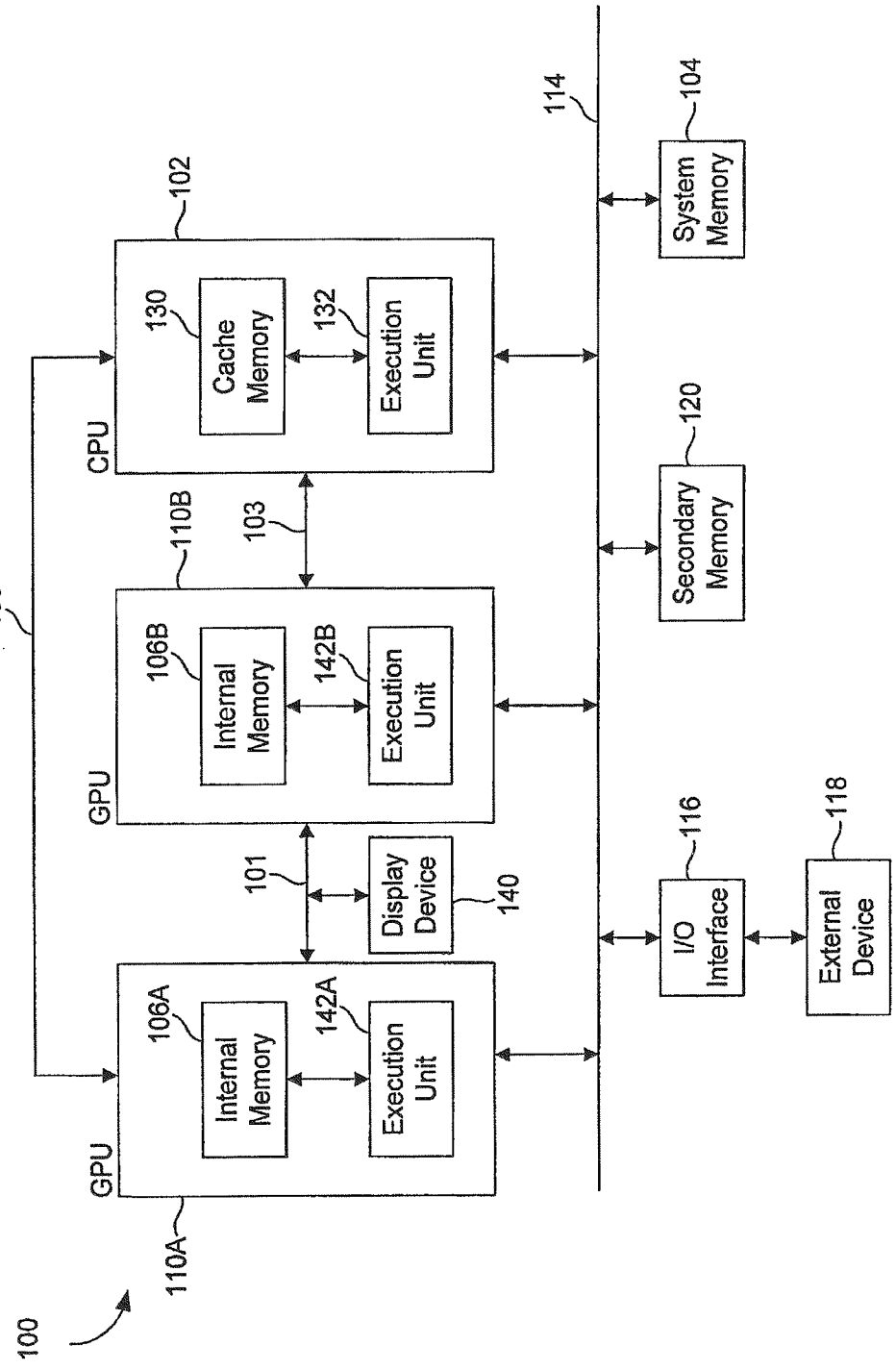
100
110A
GPU
106A
Internal Memory
142A
Execution Unit
101
Display Device
140
105
GPU
106B
Internal Memory
142B
Execution Unit
110B
103
CPU
102
130
Cache Memory
132
Execution Unit
114
104
System Memory
120
Secondary Memory
116
I/O Interface
118
External Device

FIG. 1A

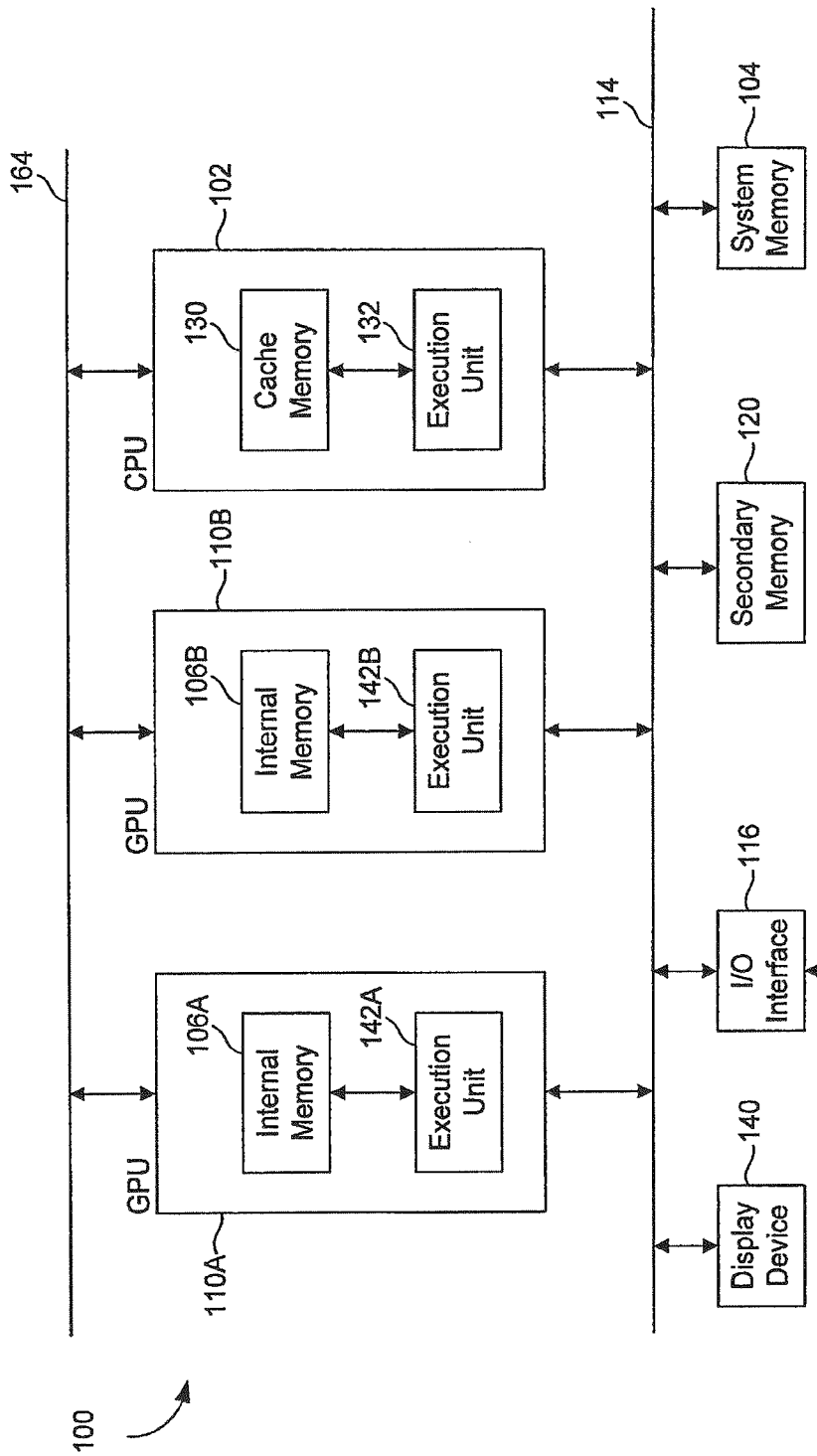
100
164
114
CPU
102
130
Cache Memory
132
Execution Unit
GPU
110B
106B
Internal Memory
142B
Execution Unit
GPU
110A
106A
Internal Memory
142A
Execution Unit
System Memory
104
Secondary Memory
120
I/O Interface
116
External Device
118
Display Device
140

FIG. 1B

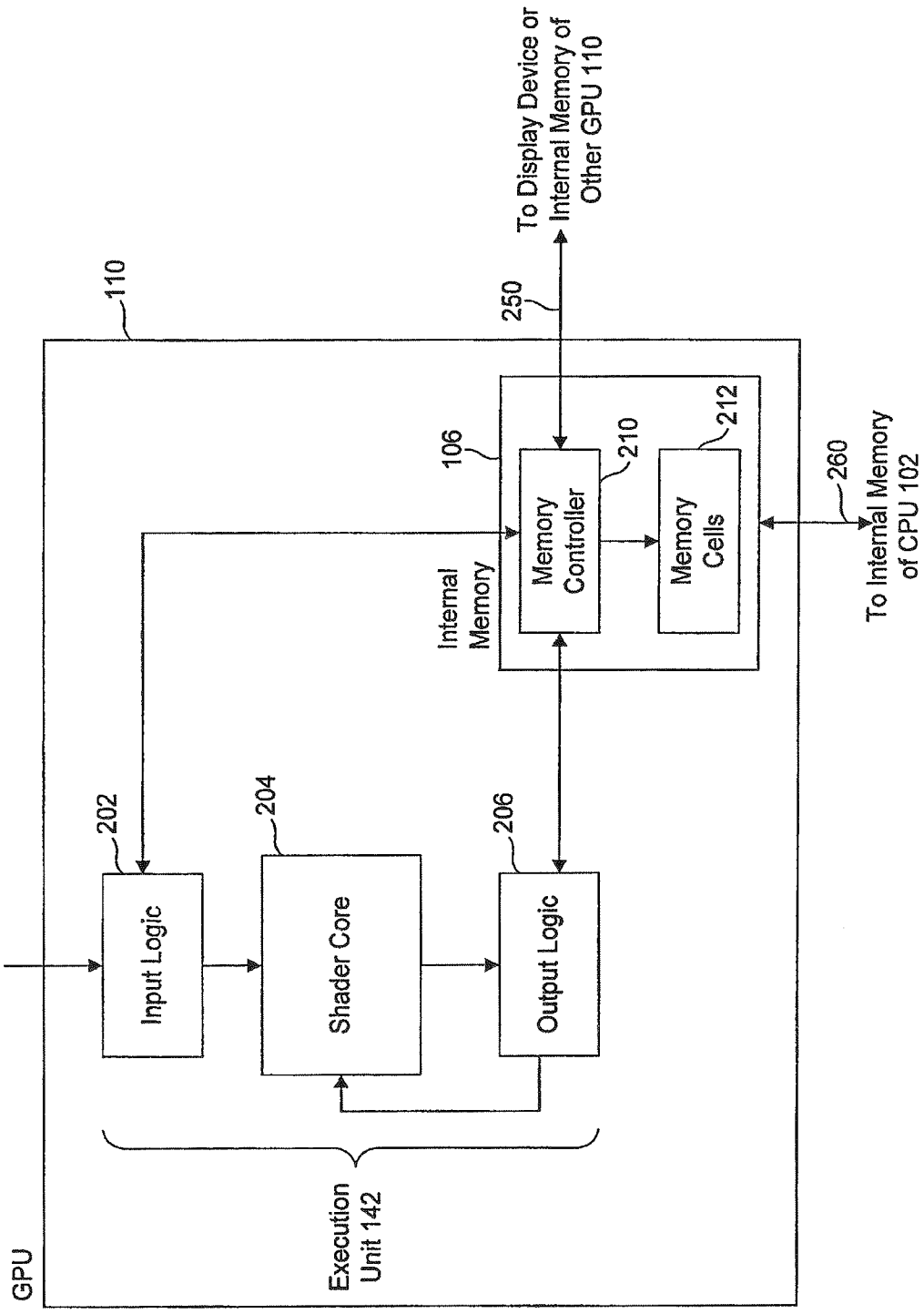
GPU
110
Execution Unit 142
Input Logic
202
Shader Core
204
Output Logic
206
Internal Memory
106
Memory Controller
210
Memory Cells
212
250
To Display Device or Internal Memory of Other GPU 110
260
To Internal Memory of CPU 102

FIG. 2

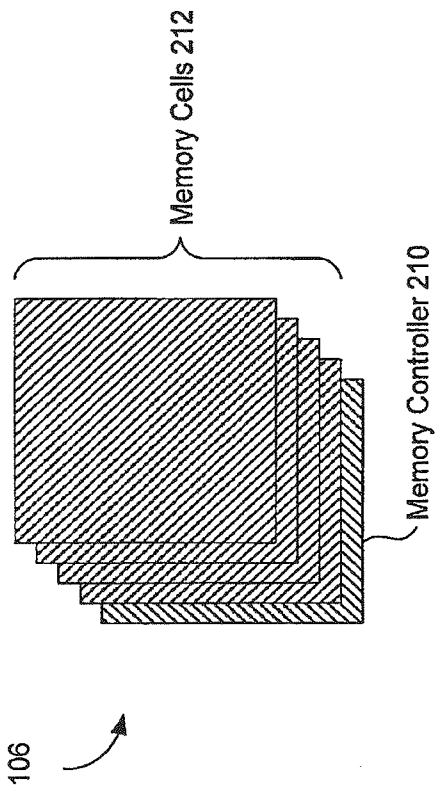

FIG. 3

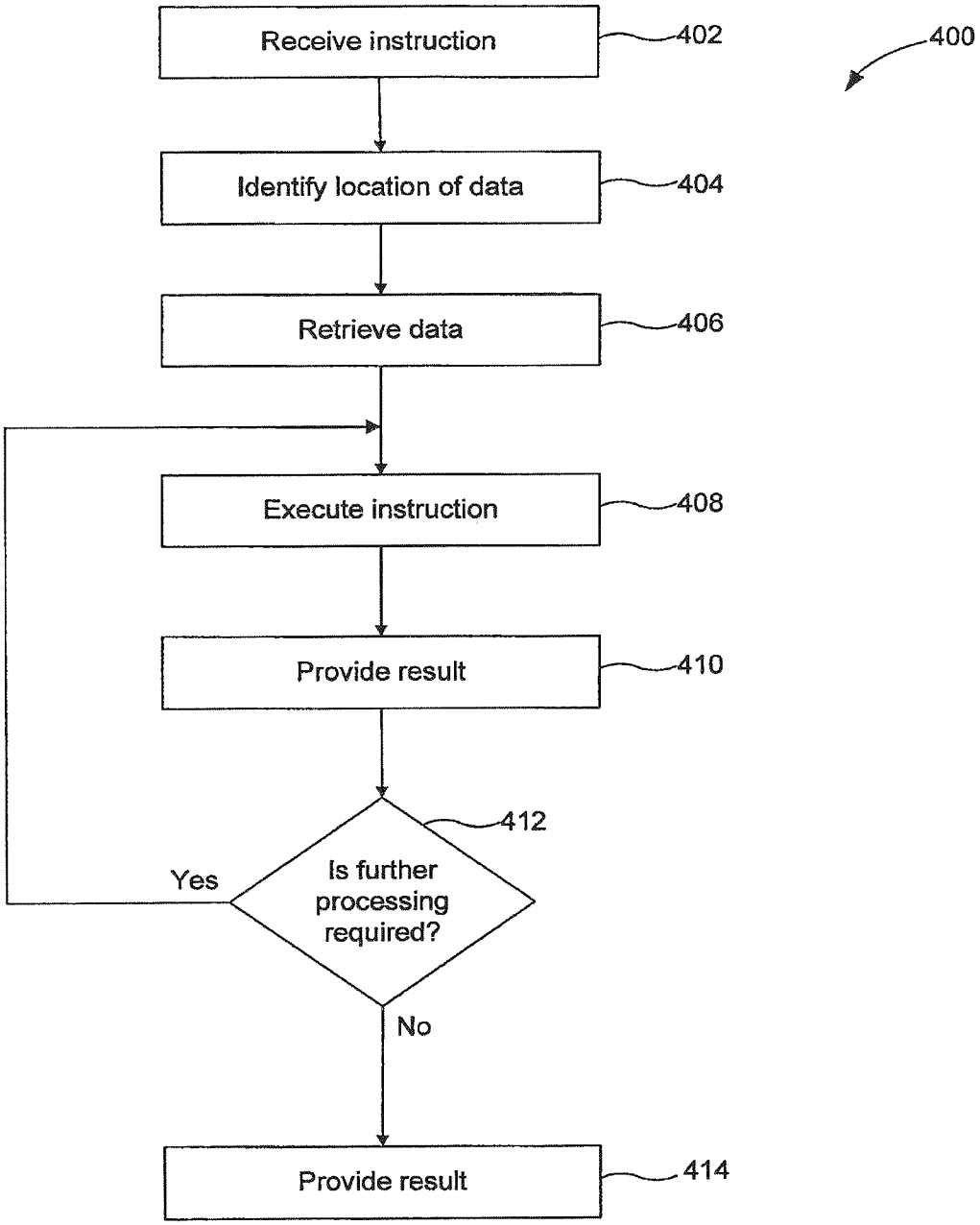
Receive instruction
402
400
Identify location of data
404
Retrieve data
406
Execute instruction
408
Provide result
410
412
Is further processing required?
Yes
No
Provide result
414

FIG. 4

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2010/047784

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F9/38 G06F9/46 G06F9/50
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F

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

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, INSPEC, IBM-TDB, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2007/294696 A1 (PAPAKIPOS MATTHEW N [US] ET AL) 20 December 2007 (2007-12-20) paragraphs [0031], [0091], [0153], [0383], [0386], [0389], [0438]; figures 7D,10	1-20
Y	EP 1 557 755 A1 (THOMSON LICENSING SA [FR]) 27 July 2005 (2005-07-27) paragraphs [0006], [0034] - [0043]	1-20
A	US 2007/074221 A1 (STENSON RICHARD B [US] ET AL) 29 March 2007 (2007-03-29) paragraphs [0006], [0008], [0022], [0024], [0030]	1-20
A	EP 0 442 041 A2 (NAT SEMICONDUCTOR CORP [US]) 21 August 1991 (1991-08-21) page 5, line 10 - line 24; figure 3	1-20

-/--

[X] Further documents are listed in the continuation of Box C.

[X] 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 document 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
15 November 2010	29/11/2010

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Thibaudeau, Jean

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2010/047784

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2004/160449 A1 (GOSSALIA ANUJ B [US] ET AL) 19 August 2004 (2004-08-19) paragraph [0040]; figure 2	6,17

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2010/047784

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2007294696 A1	20-12-2007	NONE	
EP 1557755 A1	27-07-2005	CN 1645351 A	27-07-2005
		FR 2865291 A1	22-07-2005
		JP 2005209206 A	04-08-2005
		KR 20050076702 A	26-07-2005
		MX PA05000788 A	29-08-2005
		US 2005172104 A1	04-08-2005
US 2007074221 A1	29-03-2007	EP 1934738 A1	25-06-2008
		JP 2009510612 T	12-03-2009
		US 2009147013 A1	11-06-2009
		US 2010251245 A1	30-09-2010
		WO 2007038456 A1	05-04-2007
EP 0442041 A2	21-08-1991	JP 3227177 A	08-10-1991
		US RE40942 E1	20-10-2009
US 2004160449 A1	19-08-2004	US 2005168472 A1	04-08-2005