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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
9 July 2015 (09.07.2015)

WIPO | PCT

(10) International Publication Number
WO 2015/102797 A1

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

(21) International Application Number:
PCT/US2014/068352

(22) International Filing Date:
3 December 2014 (03.12.2014)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
14/145,915 31 December 2013 (31.12.2013) US

(71) Applicant: **MICROSOFT CORPORATION** [US/US]; One Microsoft Way, Redmond, WA 98052-6399 (US).

(72) Inventors: **GARDEN, Euan, Peter**; c/o Microsoft Corporation, One Microsoft Way, LCA - International Patents (8/1172), Redmond, WA 98052-6399 (US). **JUSTICE, John, Raymond**; c/o Microsoft Corporation, One Microsoft Way, LCA - International Patents (8/1172), Redmond, WA 98052-6399 (US). **SHARMA, Madhumitra**; c/o Microsoft Corporation, One Microsoft Way, LCA - International Patents (8/1172), Redmond, WA 98052-6399 (US).

(74) Agents: **HOWARD, Jason, O.** et al.; (USOC - Shook, Hardy & Bacon), One Microsoft Way, Microsoft Corporation, LCA - International Patents (8/1172), Redmond, WA 98052-6399 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, 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).

Declarations under Rule 4.17:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

[Continued on next page]

(54) Title: MULTIMODE GAMING SERVER

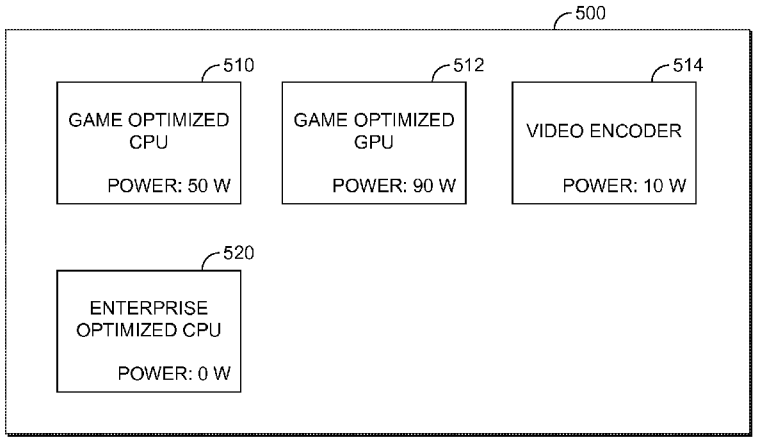

FIG. 5.

(57) Abstract: Aspects of the present invention relate to a multimode gaming server with different types of computing resources provided within the server. The different computing resources can be optimized for different computing tasks. For example, a first type of resource can be optimized for producing high definition graphics and a second type of resource for enterprise computing. Each resource may be activated or deactivated as demand for different computing tasks change throughout the day. In one aspect, the resources are different chip sets in different mother board sockets. In one aspect, provisioning of the other components (e.g., cooling, power supply, network bandwidth) in the multimode server is not adequate for both computing resources to run simultaneously.

WO 2015/102797 A1

— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:

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

MULTIMODE GAMING SERVER

BACKGROUND OF THE INVENTION

Generally, the servers selected for deployment in a data center can perform a wide range of computing tasks, but may not perform some specialized computing tasks very efficiently. For example, video intensive compute projects are better performed on a server with powerful GPU and video encoders. An enterprise server may be able to perform some video related work through the CPU, but the work may be inefficient. On the other hand, a server designed for video related work may perform general computing projects inefficiently.

SUMMARY OF THE INVENTION

This summary is provided to introduce a selection of concepts in a simplified form that are further described below in the detailed description. This summary is not intended to identify key features or essential features of the claimed subject matter, nor is it intended to be used in isolation as an aid in determining the scope of the claimed subject matter.

Aspects of the present invention relate to a multimode gaming server with different types of computing resources provided within the server. The different computing resources can be optimized for different computing tasks. For example, a first type of resource can be optimized for producing high definition graphics and a second type of resource for enterprise computing. Each resource may be activated or deactivated as demand for different computing tasks change throughout the day. In one aspect, the resources are different chip sets in different mother board sockets. In one aspect, provisioning of the other components (e.g., cooling, power supply, network bandwidth) in the multimode server is not adequate for both computing resources to run simultaneously.

In one aspect, the cooling capacity for the server is intentionally sized to be incapable of providing enough cooling to maintain an acceptable operational temperature for the multimode server, if all of the computing resources in the server are simultaneously in an active processing mode. The data center's control fabric can maintain an acceptable operating temperature within the server by assigning workloads to only one type of computing resources within the multimode server at a given point in time. For example, at any given

time only the gaming optimized compute resource in a server may be assigned a workload and be in an active processing state. The remaining compute resources in the servers are set to a low-power state.

In one aspect, different computing resources within the multimode server have a similar maximum power usage when in an active processing mode. For example, a gaming optimized resource (e.g., a GPU and specialty CPU) having a maximum power usage of 150 W can be deployed in the same multimode server as an enterprise CPU having a 150 W capacity.

In one aspect, computing resources are selected for inclusion in a multimode server according to anticipated peak periods of usage. Resources designed for specialized workload having a peak period of usage that differs from each other can be included in a multimode gaming server. For example, a first type of computing resources associated with a specialized workload with peak hours from 4 PM to 12 PM can be matched with a second type of computing resource having a specialized workload with peak hours from 6 AM to 2 PM. In other words, during a given period either a first type or a second type of computing resource within a server will be in a high demand.

BRIEF DESCRIPTION OF THE DRAWING

Aspects of the invention are described in detail below with reference to the attached drawing figures, wherein:

FIG. 1 is a block diagram of an exemplary computing environment suitable for implementing aspects of the invention;

FIG. 2 is a diagram depicting a gaming environment, in accordance with an aspect of the present invention;

FIG. 3 is a diagram depicting a remote gaming environment having one or more data centers with an nonhomogeneous arrangement of gaming servers and general purpose servers, in accordance with an aspect of the present invention;

FIG. 4 is a diagram depicting an arrangement of multimode gaming servers in various modes, in accordance with an aspect of the present invention;

FIG. 5 is a diagram depicting a mother board with active gaming resources, in accordance with an aspect of the present invention;

FIG. 6 is a diagram depicting a mother board with active general purpose resources, in accordance with an aspect of the present invention; and

FIG. 7 is a diagram depicting a method for managing workloads within a data center, in accordance with an aspect of the present invention.

DETAILED DESCRIPTION OF THE INVENTION

The subject matter of aspects of the invention is described with specificity herein to meet statutory requirements. However, the description itself is not intended to limit the scope of this patent. Rather, the inventors have contemplated that the claimed subject matter might also be embodied in other ways, to include different steps or combinations of steps similar to the ones described in this document, in conjunction with other present or future technologies. Moreover, although the terms "step" and/or "block" may be used herein to connote different elements of methods employed, the terms should not be interpreted as implying any particular order among or between various steps herein disclosed unless and except when the order of individual steps is explicitly described.

Aspects of the present invention relate to a multimode gaming server with different types of computing resources provided within the server. The different computing resources can be optimized for different computing tasks. For example, a first type of resource can be optimized for producing high definition graphics and a second type of resource for enterprise computing. Each resource may be activated or deactivated as demand for different computing tasks change throughout the day. In one aspect, the resources are different chip sets in different mother board sockets. In one aspect, provisioning of the other components (e.g., cooling, power supply, network bandwidth) in the multimode server is not adequate for both computing resources to run simultaneously.

In one aspect, the cooling capacity for the server is intentionally sized to be incapable of providing enough cooling to maintain an acceptable operational temperature for the server, if all of the computing resources in the server are simultaneously in an active processing mode. The data center's control fabric can maintain an acceptable operating temperature within the server by assigning workloads to only one type of computing resources within the multimode server at a given point in time. For example, at any given time only the gaming optimized compute resource in a server may be assigned a workload

and be in an active processing state. The remaining compute resources in the servers are set to a low-power state.

In one aspect, different computing resources within the multimode server have a similar maximum power usage when in an active processing mode. For example, a gaming optimized resource (e.g., a GPU and specialty CPU) having a maximum power usage of 150 W can be deployed in the same multimode server as an enterprise CPU having a 150 W capacity. Even though the total power usage of the two types of computing resources is similar, the distribution of power usage in the resource can differ greatly. For example, the gaming optimized resource may have a graphics processing unit ("GPU") that uses 100 W and a central processing unit ("CPU") that uses 50 W. The enterprise resource may not have a GPU, but could have a more powerful CPU that consumes 150 W.

In one aspect, computing resources are selected for inclusion in a multimode server according to anticipated peak periods of usage. Resources designed for specialized workload having a peak period of usage that differs from each other can be included in a multimode gaming server. For example, a first type of computing resources associated with a specialized workload with peak hours from 4 PM to 12 PM can be matched with a second type of computing resource having a specialized workload with peak hours from 6 AM to 2 PM. In other words, during a given period either a first type or a second type of computing resource within a server will be in a high demand.

As used herein, a "gaming optimized computing resource" is adapted to output a rendered video game image to a client device, such as a game console. The video game image may be rendered as a streaming video communicated to the client. In order to render a high quality video game image, a gaming optimized computing resource can have a graphics processing unit that is more powerful than a graphics processing unit, if any, found in a general-purpose computing resource. The gaming optimized computing resource may also have dedicated video encoding capabilities.

Power consumption can be used as a proxy for a processor's capabilities. In one aspect, a gaming optimized computing resource can be defined by the inclusion of a GPU that consumes greater than a threshold percentage of power used by the game optimized computing resource during peak power consumption. In one aspect, the threshold percentage of power is greater than 40% of peak power, for example greater than 50%, for example greater than 60%, for example greater than 70%, or for example greater than 80%. For

example, a GPU in a gaming optimized computing resource could use 100 W, with a total peak power usage (e.g., GPU and CPU) of 150 W in the gaming optimized server resource.

As used herein, the terms "general-purpose computing resource" or "general processing optimized computing resource" describes a resource designed to emphasize computing process typically associated with a central processing unit. General-purpose computing resources can be capable of performing specialized computing processes, but may not be optimized for that purpose. For example, a CPU can perform graphics processing less efficiently than the same or similar tasks can be performed by a GPU.

Aspects of the present invention may transition various types of computing resources between different power modes or states. As used herein, the term "low-power mode" means a resource is presently operating at less than 20% of the resource's maximum rate of power. As an example, a resource in low-power mode may be shut off at the motherboard socket, but able to respond to a power-on command.

As used herein, the phrase "active processing mode" means a computing resource is actively processing a computing workload. A computing resource in active processing mode can be using greater than 20% of the resource's maximum rated power.

Having briefly described an overview of aspects of the invention, an exemplary operating environment suitable for use in implementing aspects of the invention is described below.

Exemplary Operating Environment

Referring to the drawings in general, and initially to FIG. 1 in particular, an exemplary operating environment for implementing embodiments of the invention is shown and designated generally as computing device 100. Computing device 100 is but one example of a suitable computing environment and is not intended to suggest any limitation as to the scope of use or functionality of the invention. Neither should the computing device 100 be interpreted as having any dependency or requirement relating to any one or combination of components illustrated.

The invention may be described in the general context of computer code or machine-useable instructions, including computer-executable instructions such as program components, being executed by a computer or other machine, such as a personal data assistant or other handheld device. Generally, program components, including routines, programs, objects, components, data structures, and the like, refer to code that performs

particular tasks or implements particular abstract data types. Embodiments of the invention may be practiced in a variety of system configurations, including handheld devices, consumer electronics, general-purpose computers, specialty computing devices, etc. Embodiments of the invention may also be practiced in distributed computing environments where tasks are performed by remote-processing devices that are linked through a communications network.

With continued reference to FIG. 1, computing device 100 includes a bus 110 that directly or indirectly couples the following devices: memory 112, one or more processors 114, one or more presentation components 116, input/output (I/O) ports 118, I/O components 120, and an illustrative power supply 122. Bus 110 represents what may be one or more busses (such as an address bus, data bus, or combination thereof). Although the various blocks of FIG. 1 are shown with lines for the sake of clarity, in reality, delineating various components is not so clear, and metaphorically, the lines would more accurately be grey and fuzzy. For example, one may consider a presentation component such as a display device to be an I/O component 120. Also, processors have memory. The inventors hereof recognize that such is the nature of the art, and reiterate that the diagram of FIG. 1 is merely illustrative of an exemplary computing device that can be used in connection with one or more embodiments of the invention. Distinction is not made between such categories as "workstation," "server," "laptop," "handheld device," etc., as all are contemplated within the scope of FIG. 1 and refer to "computer" or "computing device."

Computing device 100 typically includes a variety of computer-readable media. Computer-readable media can be any available media that can be accessed by computing device 100 and includes both volatile and nonvolatile media, removable and non-removable media. By way of example, and not limitation, computer-readable media may comprise computer storage media and communication media. Computer storage media includes both volatile and nonvolatile, removable and non-removable media implemented in any method or technology for storage of information such as computer-readable instructions, data structures, program modules or other data.

Computer storage media includes RAM, ROM, EEPROM, flash memory or other memory technology, CD-ROM, digital versatile disks (DVD) or other optical disk storage, magnetic cassettes, magnetic tape, magnetic disk storage or other magnetic storage devices. Computer storage media does not comprise a propagated data signal.

Communication media typically embodies computer-readable instructions, data structures, program modules or other data in a modulated data signal such as a carrier

wave or other transport mechanism and includes any information delivery media. The term "modulated data signal" means a signal that has one or more of its characteristics set or changed in such a manner as to encode information in the signal. By way of example, and not limitation, communication media includes wired media such as a wired network or direct-wired connection, and wireless media such as acoustic, RF, infrared and other wireless media. Combinations of any of the above should also be included within the scope of computer-readable media.

Memory 112 includes computer-storage media in the form of volatile and/or nonvolatile memory. The memory 112 may be removable, nonremovable, or a combination thereof. Exemplary memory includes solid-state memory, hard drives, optical-disc drives, etc. Computing device 100 includes one or more processors 114 that read data from various entities such as bus 110, memory 112 or I/O components 120. Presentation component(s) 116 present data indications to a user or other device. Exemplary presentation components 116 include a display device, speaker, printing component, vibrating component, etc. I/O ports 118 allow computing device 100 to be logically coupled to other devices including I/O components 120, some of which may be built in. Illustrative I/O components 120 include a microphone, joystick, game pad, satellite dish, scanner, printer, wireless device, etc.

Exemplary Online Gaming Environment

Turning now to FIG. 2, an online gaming environment 200 in which a multimode gaming servers may be deployed within a data center is shown, in accordance with an embodiment of the present invention. The online gaming environment 200 comprises various game clients connected through a network 220 to a game service 230. Exemplary game clients include a game console 210, a tablet 212, and a personal computer 214. Use of other game clients, such as smart phones, are also possible. The game console 210 may have one or more game controllers communicatively coupled to it. In one embodiment, the tablet 212 may act as an input device for a game console 210 or a personal computer 214. In another embodiment, the tablet 212 is a stand-alone game client. Network 220 may be a wide area network, such as the Internet.

Game service 230 comprises multiple computing devices communicatively coupled to each other. In one embodiment, the game service 230 is implemented using one or more data centers that comprise multimode gaming servers. The data centers may be spread out across various geographic regions including cities throughout

the world. In this scenario, the game clients may connect to the closest data centers. Embodiments of the present invention are not limited to this setup.

The game service 230 allows the game to be executed within the computing devices provided by the game service 230. A communication session between the game service and game clients carries input traffic to the game service 230 and returns a rendered game image. In this embodiment, a computing device that is part of the game service executes the video game code using a control stream generated by input devices associated with the various game clients. The rendered video game is then communicated over the network to the game client where the rendered game is output for display.

The game service 230 may be provided by a data center that uses a combination of gaming optimized servers to execute the game and render a video game image. The gaming optimized servers may be deployed with multimode gaming servers. When a suitable task is not available for the multimode gaming servers, the gaming CPU and GPU within the multimode gaming servers can be placed in a low-power mode and the non-gaming optimized CPU activated.

Exemplary Game Service

Turning now to FIG. 3, an exemplary remote gaming environment 300 is shown, in accordance with an embodiment of the present invention. The gaming environment 300 includes a game client 310 that is shown communicatively coupled to a game service 340 through a network 330. The gaming service may use one or more multimode gaming servers to accommodate peak gaming demand.

In one embodiment, the network may be the Internet. The game client 310 is connected to a first game input device 312, a second game input device 314, and a display 316. Exemplary game input devices include game pads, keyboards, a mouse, a touch pad, a touch screen, a microphone for receiving voice commands, a depth camera, a video camera, a keyboard, and a trackball. Embodiments of the present invention are not limited to these input devices. The display device 316 is capable of displaying video game content. For example, the display 316 may be a television or computer screen. In another embodiment, the display 316 is a touch screen integrated with the game client 310.

The game client 310 is a computing device that is able to execute video games. The game client 310 could be a tablet or a laptop computer. In another embodiment, the game client 310 is a game console and the display 316 is a remote display

communicatively coupled to the game console. The game client 310 includes an operating environment 320, a game execution environment 322, a game data store 324, a game service client 326, and a player profile data store 328.

The operating environment 320 may be provided by an operating system that manages the hardware and provides services to application running on the game client 310. The operating environment may allocate client resources to different applications as part of the game migration. For example, the operating environment may give control of the display to the game execution environment 322 once game play is migrated to the game client 310.

The game execution environment 322 comprises the gaming resources on the client 310 required to execute instances of a game or a game preview. The game execution environment 322 comprises active memory along with computing and video processing. The game execution environment 322 receives gaming controls and causes the game to be manipulated and progressed according to its programming. In one embodiment, the game execution environment 322 outputs a rendered video stream that is communicated to the display 316.

The game data store 324 stores downloaded games, game previews, and partially downloaded games.

The game service client 326 is a client application that displays rendered video game images received from the game service 340. The game service client 326 may also process game input and change it into an easily uploadable format that is communicated to the game service 340. The game service client 326 may also scale the rendered video game images received from the service 340 to a size optimized for display 316.

The player profile data store 328 stores player profile information for individual games. The player profile information may also save tombstones or game-saved data for individual games, including previews. Both the game-save file and the tombstone record game progress. The game execution environment 322 then reads the game-saved data to start the game where the player left off on the server. The opposite scenario is also possible where the game-saved data and player profile information is uploaded from the game client 310 to the game service 340 to initiate the game.

The game service 340 comprises a connection manager 342, a player profile data store 344, a game execution environment 348, and a game data store 350. Though depicted as a single box, the game service 340 could be implemented in a data center that comprises numerous machines, or even several data centers.

The connection manager 342 builds a connection between the client 310 and the service 340. The connection manager 342 may also provide various authentication mechanisms to make sure that the user is authorized to access the game service 340. The connection manager 342 may also analyze the bandwidth available within a connection and throttle the download of a game during game play to make sure that game play is not degraded.

The player profile data store 344 may work in conjunction with the connection manager 342 to build and store player information. Part of the player profile may comprise demographic and financial information such as a player's name, address and credit card information or other mechanism for paying for or purchasing games and experiences provided by the game service.

In addition, the player profile data store 344 may store a player's progress within an individual game. As a player progresses through a game or game preview, the player's score and access to game levels may be stored. Further, the player profile data store 344 may store information about individual player preferences such as language preferences. Information regarding a player's game client and speed of the network connection may also be stored and utilized to optimize the gaming experience. For example, in one embodiment, when a geographically proximate data center is busy, players with higher latency Internet connections may be preferentially connected to proximate data centers while players with lower latency connections may be connected to data centers that are further away. In this way, the players with the network connections that are best able to handle the additional latency are connected to data centers that create additional latency because of their location.

The player profile data store 344 may also store a usage history for the individual player. A player's history of purchasing games, sampling games, or playing games through a game service that does not require the purchase of the games may be stored. The usage information may be analyzed to suggest games of interest to an individual player. In one embodiment, the purchase history may include games that are not purchased through the game service. For example, the purchase history may be augmented by the player entering in a key from a game purchased in a retail store. In some embodiments, the player may then have access to that game both on their game client 310 and through the game service when they are no longer at their game client.

The game execution environment 348 comprises the gaming resources required to execute instances of a game. These are the resources described previously that

are managed by the game manager 352 and other components. The game execution environment 348 comprises active memory along with computing and video processing. The game execution environment 348 receives gaming controls through an I/O channel and causes the game to be manipulated and progressed according to its programming. In one embodiment, the game execution environment 348 outputs a rendered video stream that is communicated to the game client. In other embodiments, the game execution environment 348 outputs game geometry, or other representations, which may be combined with local objects on the gaming client to render the gaming video.

The game data store 350 stores available games. The games may be retrieved from the data store and activated through an active memory. The game data store 350 may be described as passive or secondary memory. In general, games may not be played off of the game data store 350. However, in some embodiments, the secondary memory may be utilized as virtual memory, in which case portions of the game data store 350 may also serve as active memory. This illustrates that active memory is not necessarily defined by a particular hardware component, but is defined by the ability of the game resources to actively manipulate and access objects within the memory to execute the game.

Turning now to FIG. 4, an arrangement of multimode gaming servers within a data center 400 is shown, in accordance with an aspect of the present invention. The arrangement comprises rack 410, rack 412, rack 414, and rack 416. Four racks are shown for the sake of simplicity; an actual implementation could include tens, hundreds, or thousands of racks deployed within a data center. Each rack can comprise a quantity of servers, power distribution equipment, and networking equipment. In one arrangement, a networking cable is run to a router/switch within the rack. Each server in the rack then connects to the router. Similarly, power may be run to a power distribution station associated with the rack. Each server is then coupled to the power distribution station.

Additionally, each rack can include cooling equipment, such as fans. In one arrangement, a fan wall is provided behind the servers to draw air through the servers. In a vertical cooling arrangement, one or more fans are located above or below the rack to facilitate airflow to the servers within the rack. The cooling equipment can also include thermocouples and other sensors that measure temperature, pressure, and air flow throughout the rack. The rack may include one or more fixed or adjustable baffles to distribute air where needed for cooling.

A control fabric 402 is communicatively coupled to the racks and computing devices within the racks. The control fabric 402 manages the state of each multimode gaming server. For example, the control fabric 402 can transition a multimode server between modes by activating a first type of computing resource and deactivating a second type of computing resource. The control fabric 402 can distribute workloads to computing devices. The control fabric 402 can also manage cooling equipment within the racks. For example, the control fabric 402 can lower fan speed within a rack when the servers within the rack are in a low power mode.

Racks 414 and 416 illustrate multimode servers in a mixture of modes. Aspects of the present invention are not limited to mixing modes within a data center unit, such as a rack chassis. In one aspect, all multimodal gaming servers within a data center unit are operating in the same mode. As FIG. 4 illustrates, a mixture of modes within a rack is also possible.

Illustratively, rack 416 includes multimode server 420 in game mode, multimode server 422 in general processing mode, and multimode server 424 in general processing mode. Illustratively, rack 414 includes multimode server 430 in game mode, multimode server 432 in game mode, and multimode server 434 in game mode. In one aspect, the computing devices within a rack have a homogeneous hardware configuration that allows them to switch between a gaming optimized mode and a non-gaming optimized mode.

The multimode servers can be transitioned between a gaming optimization and general-purpose optimization by activating different computing resources within the server. In one aspect, the general-purpose resources and game-optimized resources have peak usage periods that do not significantly overlap. Racks with an arrangement of multimode gaming servers, such as racks 414 and 416, may be deployed within a data center in combination with racks of single mode servers optimized for a single function, such as gaming. The quantity of single mode servers may be specified to accommodate base demand for the computing service provided by an optimized server. The deployment of single mode servers to meet base demand allows the single mode servers to be active above a threshold amount of time on average. For example, an amount of single mode servers deployed may be limited to those able to be active, on average, 80% of a day. Multimode servers can be used during peak usage periods to accommodate demand in combination with single mode servers optimized for the work load, such as gaming.

Turning now to FIG. 5, different types of resources within a multimode server are shown, in accordance with an aspect of the present invention. The multipurpose server includes motherboard 500. The motherboard 500 includes a first type of computing resource optimized for gaming. The first type of computing resource comprises a game optimized CPU 510, a game-optimized GPU 512, and a video encoder 514. This illustrates that a computing resource, as used herein, can comprise multiple hardware items. Also, though not shown, the computing resources may include memory and other components that support the resource. For example, a CPU can have dedicated DRAM memory. Aspects of the invention are not limited to use with a separate video encoder. A video encoder could be part of a CPU or GPU.

In one aspect, the game optimized CPU 510 and game-optimized GPU 512 are the same chips as found in a commercially available game console. Exemplary game consoles include the Xbox 360, Sony's PlayStation® family, Xbox One, and Nintendo's Wii™, and such. The hardware configuration associated with the game optimized computing resources can be configured to allow games written for a commercially available game console to run on the multimode server without modification to the game code and to interact with the hardware in the same way the games interact with hardware in the game console. For example, a process performed by a game console's GPU can be performed by the game-optimized computing resource's GPU.

The game optimized CPU 510, the game-optimized GPU 512, and the video encoder 514 can all be coupled to sockets within the motherboard. In one aspect, a computing resource is deactivated for transition to a low power mode by turning off the socket to which the resource is attached.

The general-purpose computing resource on motherboard 500 comprises enterprise-optimized CPU 520. In game mode, as shown in FIG. 5, the enterprise-optimized CPU 520 is drawing 0 W. In contrast, the game optimized CPU 510 is drawing 50 W, the game-optimized GPU 512 is drawing 90 W, and the video encoder 514 is drawing 10 W for a total use of 150 W in game mode.

Turning now to FIG. 6, power use in general purpose compute mode is illustrated, in accordance with an aspect of the present invention. Now the game optimized CPU 510, the game-optimized GPU 512 and a video encoder 514 are all drawing 0 W. In contrast, the enterprise-optimized CPU 520 is drawing 150 W. In one aspect, the rated power consumption of computing resources associated with different modes is substantially equal.

In this case, the computing resources associated with the gaming mode draw the same amount of power as the computing resources associated with the enterprise, or general compute mode. In one aspect, the cooling system for a multimode server is only capable of providing a knot cooling for one type of computing resource to be active at a given point in time.

Turning now to FIG. 7, a method 700 for managing workloads within a data center is provided, in accordance with an aspect of the present invention. Method 700 may be performed by a control fabric that manages workloads within a data center.

At step 710, substantially all of a first type of computing resource within a plurality of multimode servers are set to a low powered mode during a first time period. Each multimode server has multiple computing resources optimized for different types of work. The multiple computing resources comprise at least the first type of computing resource and a second type of computing resource. In one aspect, the first period of time corresponds to a low demand period for a workload the first type of computing resource is optimized to process. For example, a low demand period for a gaming workload may occur during the day and gaming optimized servers may be set to a low power mode during this period of time.

At step 720 substantially all of the second type of computing resource within the plurality of multimode servers are set to the low power mode at a second time period. The first and second time periods do not substantially overlap in one aspect. The nonoverlapping time periods allow the first type of computing resource and a second type of computing resource to satisfy peak demand for the computing loads they are optimized to handle.

The type of computing resource that is active within an individual multimode server can be adjusted based on workload demand. As the supply of a first type of workload increases, resources adapted to process the first type of workload can be activated. As demand for different types of workload changes, a mixture of states may be present in the plurality of multimode servers. For example, all the multimode servers could be set to process a first type of workload, a third of the multimode servers could be set to process a first type of workload, half of the multimode servers could be set to process a first type of workload, or none of the multimode servers could be set to process a first type of workload. Multimode servers that are not set to handle the first type of workload could be set to use a different type of workload.

Aspects of the invention have been described to be illustrative rather than restrictive. It will be understood that certain features and subcombinations are of utility and

may be employed without reference to other features and subcombinations. This is contemplated by and is within the scope of the claims.

CLAIMS

What is claimed is:

1. A multimode server comprising a gaming optimized computing resource having a first hardware configuration and a general processing optimized computing resource having a second hardware configuration that is different from the first hardware configuration, wherein the gaming optimized computing resource and the general processing optimized computing resource are power balanced to use a substantially equal amount of power at peak power, and wherein the multimode server comprises a control to activate either the gaming optimized computing resource or the general processing optimized computing resource at a given point in time, but not both simultaneously.

2. The multimode server of claim 1, wherein the gaming optimized computing resource is designed for a first workload with a peak usage during a first time period and the general processing optimized computing resource is designed for a second workload with a peak usage during a second time period that does not substantially overlap with the first time period.

3. The multimode server of claim 1, wherein the gaming optimized computing resource has a graphics processing unit ("GPU") and a central processing unit ("CPU"), and wherein a maximum power usage of the GPU comprises more than 40% of the gaming optimized computing resource's maximum power usage.

4. The multimode server of claim 1, wherein a power source provided to the multimode server does not supply enough power to run the gaming optimized computing resource and the general processing optimized computing resource simultaneously.

5. The multimode server of claim 1, wherein at least part of the gaming optimized computing resource is connected to a first socket on a mother board and at least part of the general processing optimized computing resource is connected to a second socket on the mother board.

6. A method for managing workloads within a data center, the method comprising: during a first time period, setting substantially all of a first type of computing resource within a plurality of multimode servers to a low power mode, each multimode server having multiple computing resources optimized for different types of work, the multiple computing resources comprising at least the first type of computing resource and a second type of computing resource; and during a second time period, setting substantially all of the second type of computing resource within the plurality of multimode servers to the low power mode, wherein the second time period does not substantially overlap with the first time period.

7. The method of claim 6, wherein said setting the first type of computing resource to the low power mode comprises deactivating one or more motherboard sockets to which the first type of computing resource is attached.

8. The method of claim 6, wherein the first type of computing resource is designed for a first workload with a peak usage during the first time period and the second type of computing resource is designed for a second workload with a peak usage during the second time period.

9. The method of claim 6, wherein the first type of computing resource and the second type of computing resource generate a substantially similar amount of heat when in use.

10. The method of claim 6, wherein the first type of computing resource outputs a rendered video game image over a wide area network to a remotely located gaming device.

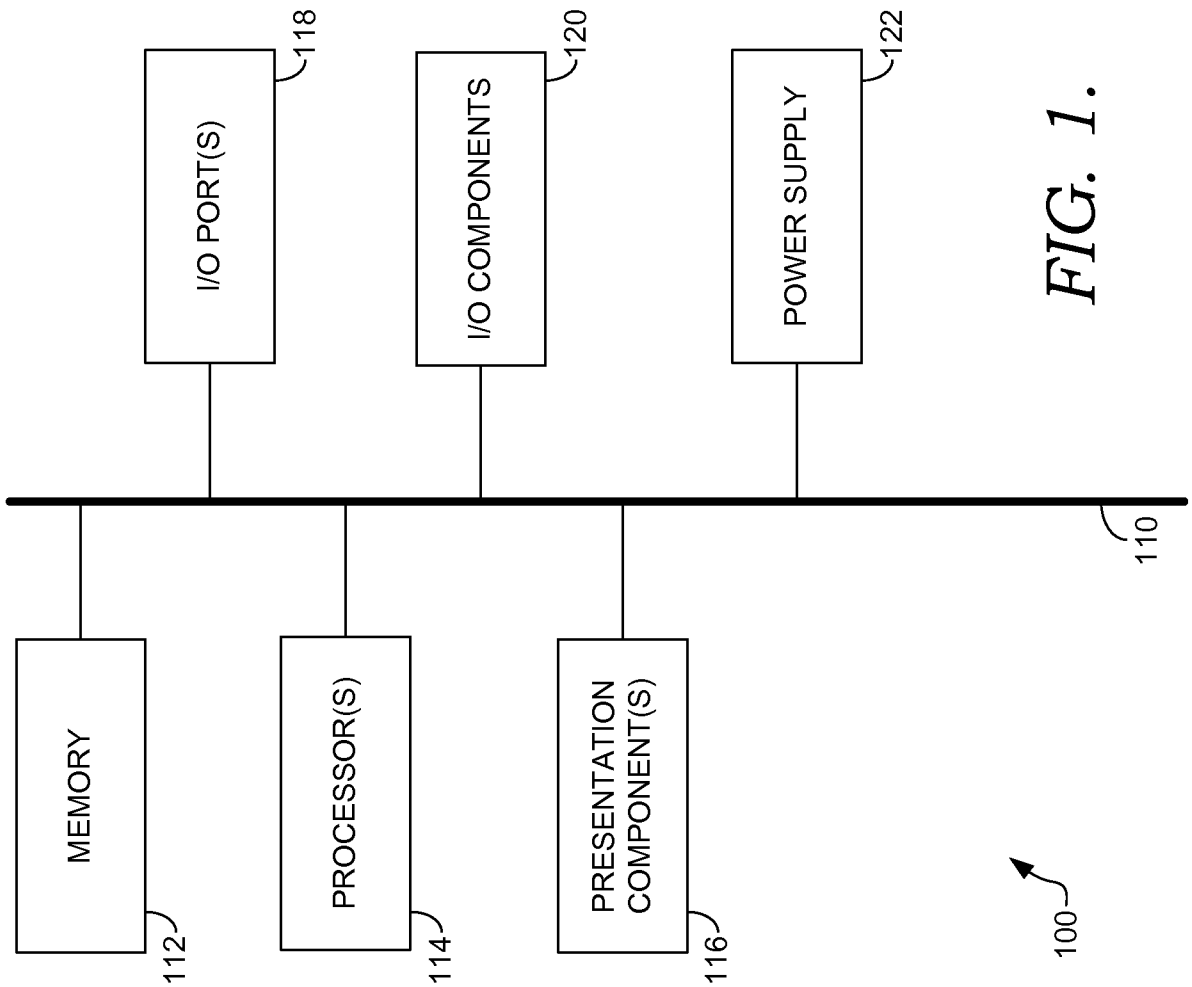
I/O PORT(S)
118
I/O COMPONENTS
120
POWER SUPPLY
122
110
MEMORY
112
PROCESSOR(S)
114
PRESENTATION COMPONENT(S)
116
100

FIG. 1.

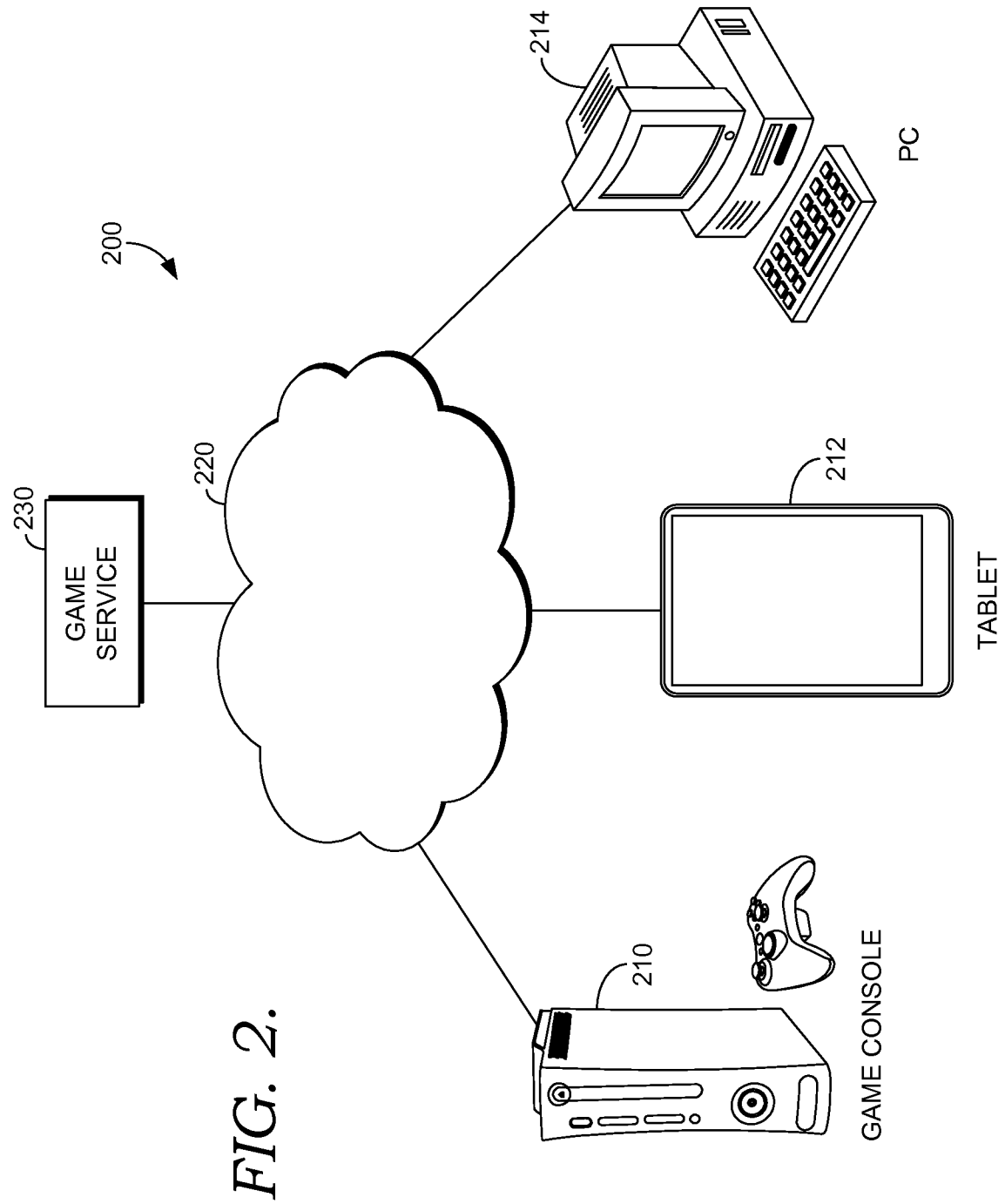

FIG. 2.

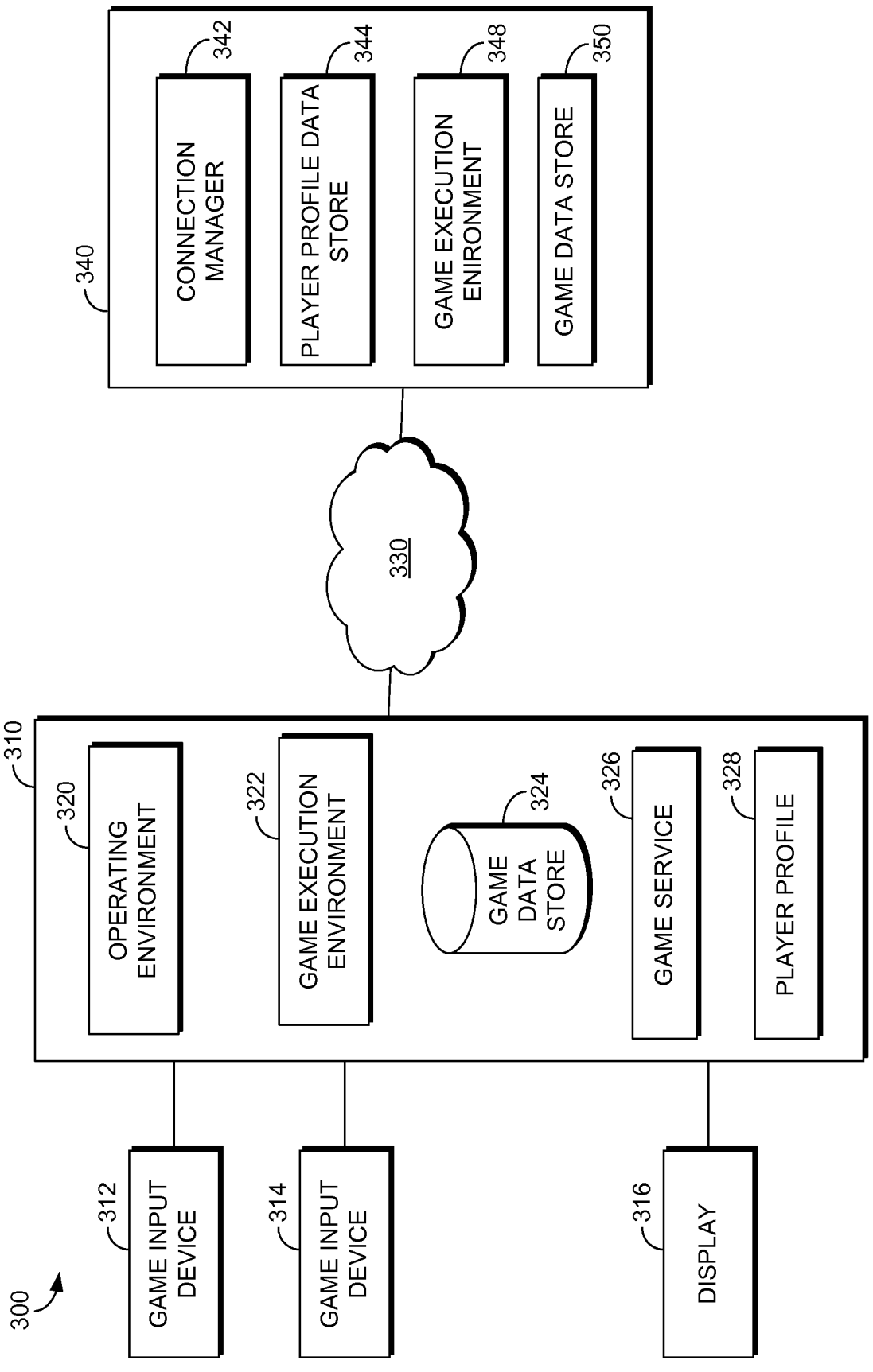
300
310
312
GAME INPUT DEVICE
314
GAME INPUT DEVICE
316
DISPLAY
320
OPERATING ENVIRONMENT
322
GAME EXECUTION ENVIRONMENT
324
GAME DATA STORE
326
GAME SERVICE
328
PLAYER PROFILE
330
340
342
CONNECTION MANAGER
344
PLAYER PROFILE DATA STORE
348
GAME EXECUTION ENIRONMENT
350
GAME DATA STORE

FIG. 3.

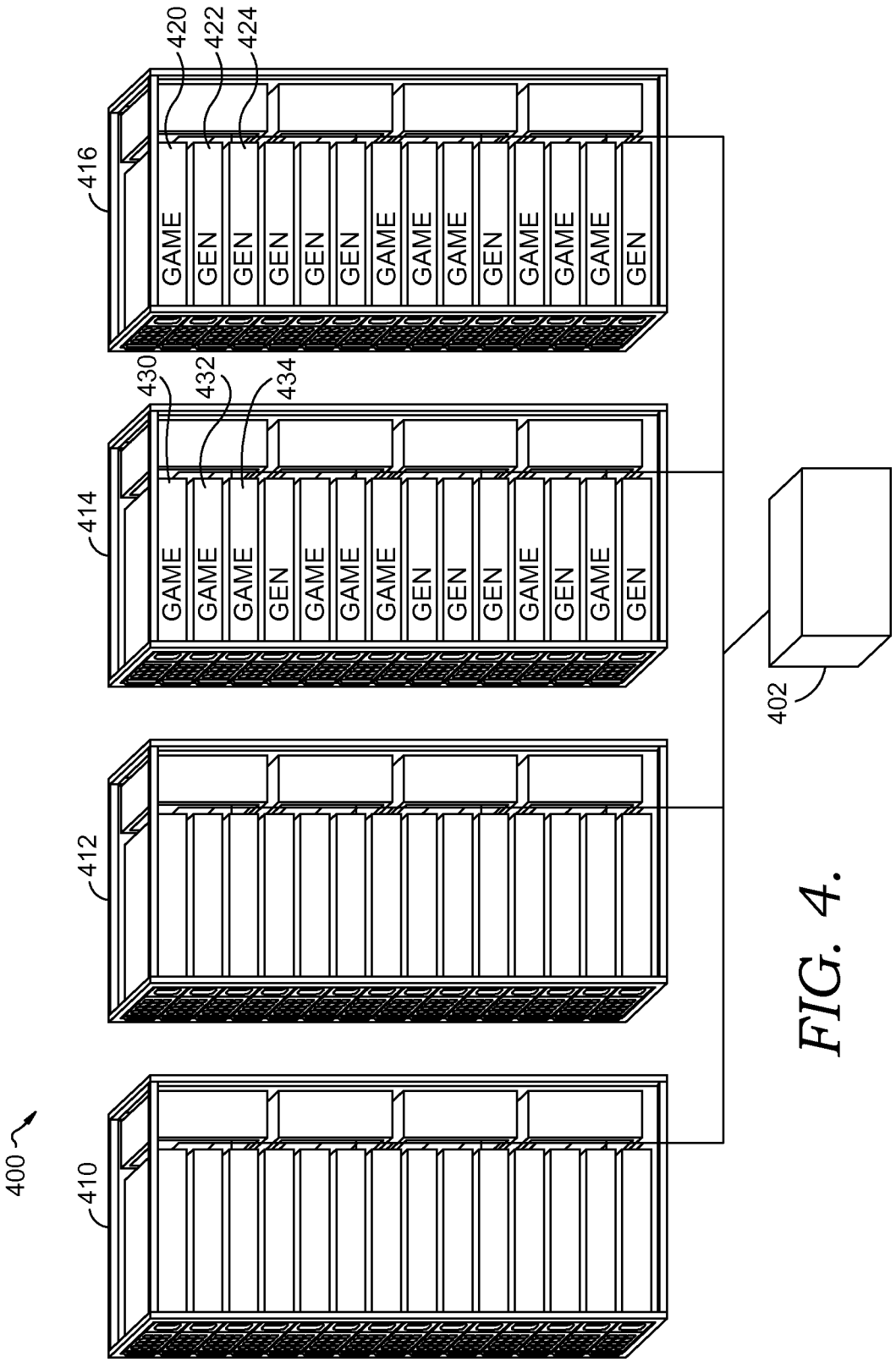

FIG. 4.

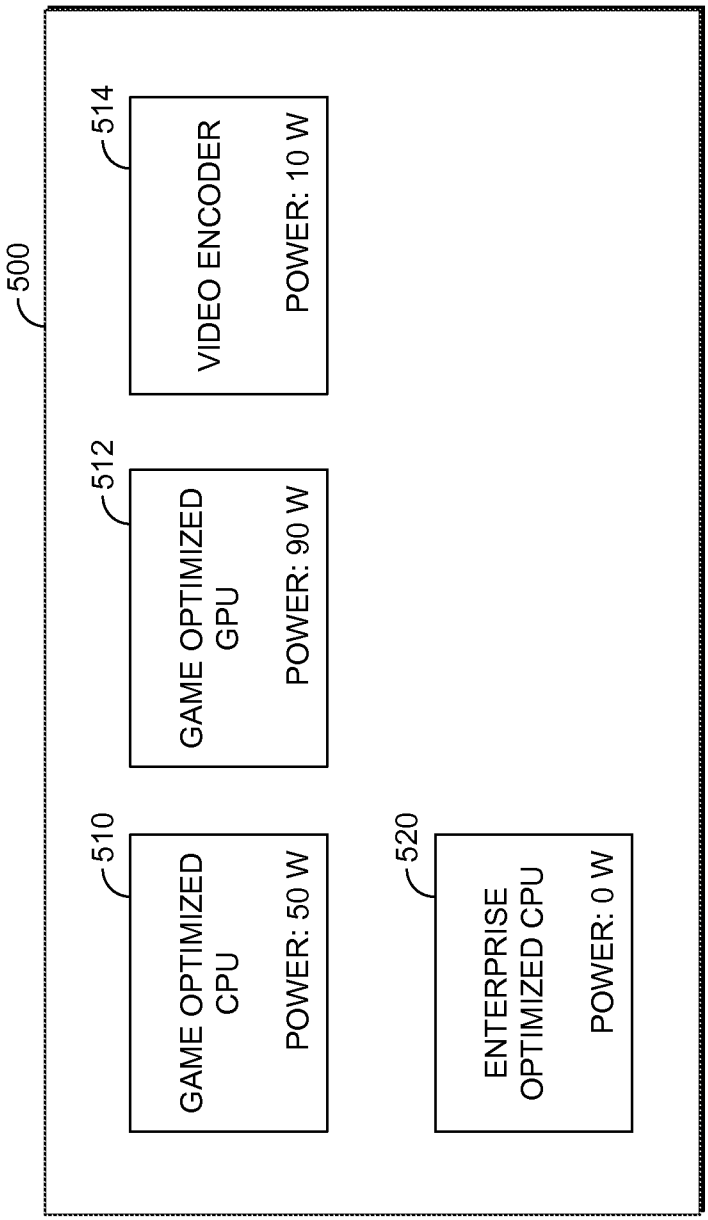
500
514
VIDEO ENCODER
POWER: 10 W
512
GAME OPTIMIZED GPU
POWER: 90 W
510
GAME OPTIMIZED CPU
POWER: 50 W
520
ENTERPRISE OPTIMIZED CPU
POWER: 0 W

FIG. 5.

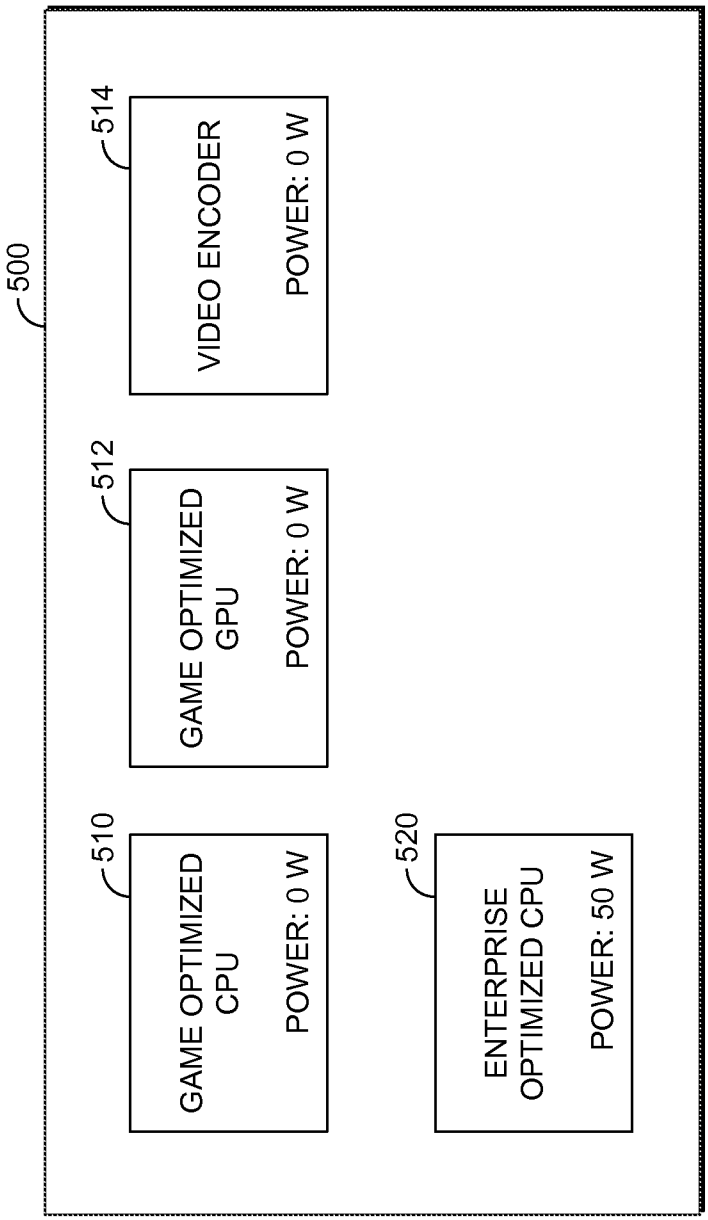
500
514
VIDEO ENCODER
POWER: 0 W
512
GAME OPTIMIZED GPU
POWER: 0 W
510
GAME OPTIMIZED CPU
POWER: 0 W
520
ENTERPRISE OPTIMIZED CPU
POWER: 50 W

FIG. 6.

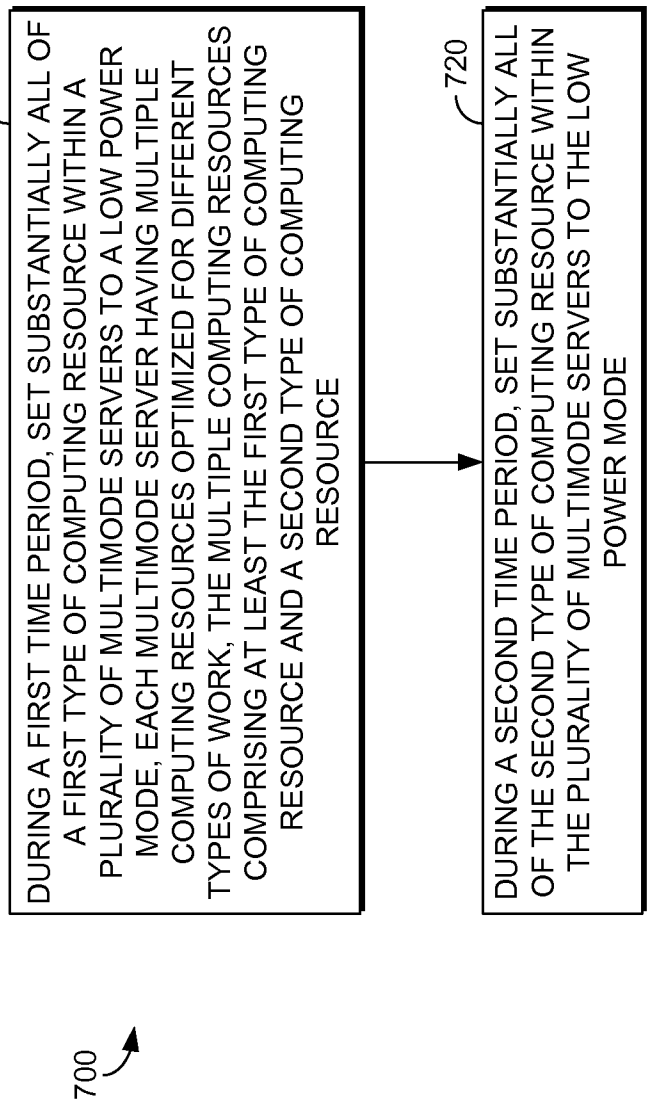
700
710
DURING A FIRST TIME PERIOD, SET SUBSTANTIALLY ALL OF A FIRST TYPE OF COMPUTING RESOURCE WITHIN A PLURALITY OF MULTIMODE SERVERS TO A LOW POWER MODE, EACH MULTIMODE SERVER HAVING MULTIPLE COMPUTING RESOURCES OPTIMIZED FOR DIFFERENT TYPES OF WORK, THE MULTIPLE COMPUTING RESOURCES COMPRISING AT LEAST THE FIRST TYPE OF COMPUTING RESOURCE AND A SECOND TYPE OF COMPUTING RESOURCE
720
DURING A SECOND TIME PERIOD, SET SUBSTANTIALLY ALL OF THE SECOND TYPE OF COMPUTING RESOURCE WITHIN THE PLURALITY OF MULTIMODE SERVERS TO THE LOW POWER MODE

FIG. 7.

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2014/068352

A. CLASSIFICATION OF SUBJECT MATTER
INV. 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 practicable, search terms used)
EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2010/318818 A1 (MANGIONE-SMITH WILLIAM HENRY [US]) 16 December 2010 (2010-12-16) abstract; figures 1,46,52 paragraphs [0087], [0096], [0212]	1-5
A	US 2012/016528 A1 (RAMAN MADHUSUDAN [US] ET AL) 19 January 2012 (2012-01-19) paragraphs [0001], [0027], [0028], [0030], [0033], [0077], [0079], [0094]	1-5

☐ 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
20 February 2015	27/02/2015

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	Kingma, Ype

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2014/068352

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2010318818 A1	16-12-2010	NONE	
US 2012016528 A1	19-01-2012	US 2012016528 A1	19-01-2012
		US 2013103221 A1	25-04-2013

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2014/068352

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☒ Claims Nos.: 6-10
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

see FURTHER INFORMATION sheet PCT/ISA/210

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

Continuation of Box II.2

Claims Nos.: 6-10

Due to unclarities in the claims the unity between the group of claims 1-5 and
the group of claims 6-10 cannot be established.

The reasons are given under Box VIII and Box III in the separate sheet of the written opinion.

The applicant's attention is drawn to the fact that claims relating to inventions in respect of which no international search report has been established need not be the subject of an international preliminary examination (Rule 66.1(e) PCT). The applicant is advised that the EPO policy when acting as an International Preliminary Examining Authority is normally not to carry out a preliminary examination on matter which has not been searched. This is the case irrespective of whether or not the claims are amended following receipt of the search report or during any Chapter II procedure. If the application proceeds into the regional phase before the EPO, the applicant is reminded that a search may be carried out during examination before the EPO (see EPO Guidelines C-IV, 7.2), should the problems which led to the Article 17(2) declaration be overcome.