

(51) International Patent Classification:
G06F 9/50 (2006.01)(21) International Application Number:
PCT/EP2010/051929(22) International Filing Date:
16 February 2010 (16.02.2010)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
12/394,669 27 February 2009 (27.02.2009) US(71) Applicant (for all designated States except US): **INTERNATIONAL BUSINESS MACHINES CORPORATION** [US/US]; New Orchard Road, Armonk, New York 10504 (US).(71) Applicant (for MG only): **IBM UNITED KINGDOM LIMITED** [GB/GB]; P.O. Box 41, North Harbour, Portsmouth, Hampshire PO6 3AU (GB).

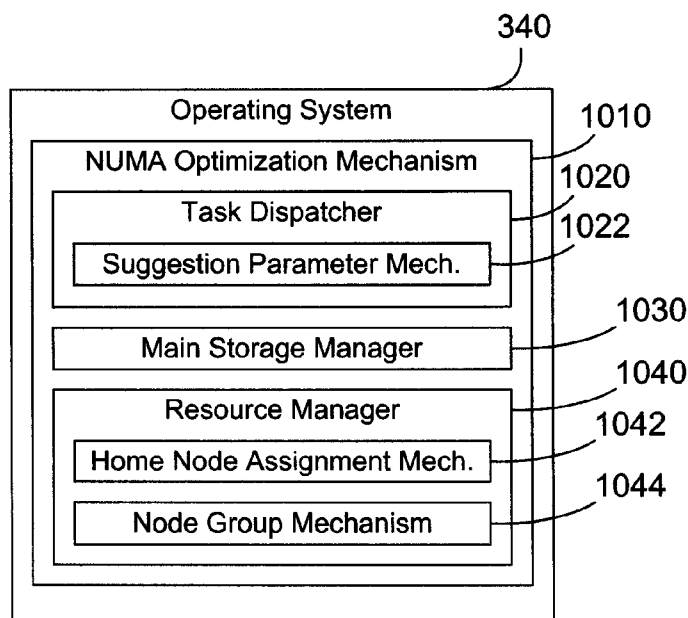
(72) Inventors; and

(75) Inventors/Applicants (for US only): **ANAND, Vijayanthimala** [US/US]; IBM Corporation, 11501 Burnet Road, Austin, Texas 78758-3400 (US). **FUNK, Mark** [US/US]; IBM Corporation, MP HP4A, 3605 Highway 52 North,Rochester, Minnesota 55901-1407 (US). **KUNKEL, Steven** [US/US]; IBM Corporation, MP 50-2, 3605 Highway 52 North, Rochester, Minnesota 55901-1407 (US). **SRINIVAS, Mysore** [US/US]; IBM Corporation, MP 9571, 11501 Burnet Road, Austin, Texas 78758-3400 (US). **SWANBERG, Randal** [US/US]; IBM Corporation, MP 905-4G011, 11501 Burnet Road, Austin, Texas 78758-3400 (US). **YOUNG, Ronald** [US/US]; IBM Corporation, MP HP4, 3605 Highway 52 North, Bldg 006, Floor 2, Office E202, Rochester, Minnesota 55901 (US).(74) Agent: **STRETTON, Peter, John**; IBM United Kingdom Limited, Intellectual Property Law, Hursley Park, Winchester, Hampshire SO21 2JN (GB).

(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.

[Continued on next page]

(54) Title: NON-UNIFORM MEMORY ACCESS (NUMA) ENHANCEMENTS FOR SHARED LOGICAL PARTITIONS

**FIG. 10**

(57) Abstract: In a NUMA-topology computer system that includes multiple nodes and multiple logical partitions, some of which may be dedicated and others of which are shared, NUMA optimizations are enabled in shared logical partitions. This is done by specifying a home node parameter in each virtual processor assigned to a logical partition. When a task is created by an operating system in a shared logical partition, a home node is assigned to the task, and the operating system attempts to assign the task to a virtual processor that has a home node that matches the home node for the task. The partition manager then attempts to assign virtual processors to their corresponding home nodes. If this can be done, NUMA optimizations may be performed without the risk of reducing the performance of the shared logical partition.



(84) **Designated States** (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (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).

Published:

- with international search report (Art. 21(3))
- before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))

NON-UNIFORM MEMORY ACCESS (NUMA) ENHANCEMENTS FOR SHARED LOGICAL PARTITIONS

BACKGROUND

5 Technical Field

This disclosure generally relates to computer systems, and more specifically relates to computer systems that include multiple logical partitions that share resources.

10 Background Art

The combination of hardware and software on a particular computer system defines a computing environment. Different hardware platforms and different operating systems thus provide different computing environments. In recent years, engineers have recognized that it is possible to provide different computing environments on the same physical computer system by logically partitioning the computer system resources to different computing environments. The Power Systems computer system developed by IBM is an example of a computer system that supports logical partitioning. If logical partitioning on a Power Systems computer system is desired, partition manager code (referred to as a “hypervisor” in IBM terminology) is installed that allows defining different computing environments on the same platform. Once the partition manager is installed, logical partitions may be created that define different computing environments. The partition manager manages the logical partitions to assure that they can share needed resources in the computer system while maintaining the separate computing environments defined by the logical partitions.

25 Non-uniform memory access (NUMA) is a computer architecture that provides multiple nodes that each have multiple processors coupled to local memory. While all memory in all nodes may be accessed by any processor in the system, the time to access memory that is local to a processor is much less than the time to access memory that is remote from a processor (e.g., on a different node). In this sense, the memory access times in the system are non-uniform, meaning local memory accesses are faster than remote memory accesses.

30

Logical partitions may be dedicated or shared. A dedicated logical partition has resources that are dedicated to it, which means only the dedicated logical partition may use these resources. A shared logical partition, on the other hand, uses resources that may also be used by other logical partitions. When a computer has a NUMA topology and is logically partitioned, NUMA optimizations may be enabled on all dedicated logical partitions. However, NUMA optimizations are typically disabled for all shared logical partitions because the sharing of the resources reduces the effectiveness of the NUMA optimizations. In addition, if NUMA optimizations were enabled on shared logical partitions, it could lead to a net decrease in performance for the logical partition. As a result, when a NUMA computer system is logically partitioned in the prior art, NUMA optimizations are enabled for dedicated logical partitions but are disabled for shared logical partitions. Without a way to enable NUMA optimizations on shared logical partitions without the possibility of decreased performance, shared logical partitions in a NUMA computer system will not benefit from NUMA optimizations.

SUMMARY OF THE INVENTION

The present invention accordingly provides, in a first aspect, a computer system comprising:
a plurality of logical partitions defined on a plurality of nodes that each have multiple physical processors coupled to a local memory in a non-uniform memory access (NUMA) topology, the plurality of logical partitions including a plurality of shared logical partitions that share a pool of common nodes; an operating system in each of the plurality of shared logical partitions that assigns a home node to a task when the task is created, assigns the task to a first virtual processor corresponding to the home node when the first virtual processor is available, and assigns the task to a second virtual processor corresponding to a node other than the home node with a suggestion parameter that specifies the home node when the first virtual processor is not available; and a partition manager that runs each virtual processor on one of the physical processors, the partition manager running the first virtual processor on a physical processor corresponding to the home node when the home node has an available physical processor, the partition manager running the first virtual processor on a physical processor corresponding to a node different than the home node when the home node has no available physical processor.

Preferably, the partition manager reads the suggestion parameter in the second virtual processor, and when the home node has an available physical processor, the partition manager runs the second virtual processor on a physical processor corresponding to the suggestion parameter. Preferably, when the home node does not have an available physical processor, the partition manager runs the second virtual processor on a physical processor corresponding to the node other than the home node. Preferably, each of the plurality of nodes is in a node group, and wherein the partition manager runs the first virtual processor on a physical processor corresponding to a node different than the home node and in a same node group as the home node when the home node has no available physical processor. Preferably, each virtual processor specifies one of the plurality of nodes as a home node.

Preferably, each virtual processor comprises a suggestion parameter that specifies a suggested node, wherein the suggestion parameter is written by an operating system that dispatches tasks to the virtual processor.

In a second aspect, there is provided a method for performing non-uniform memory access (NUMA) optimizations in a computer system that includes a plurality of logical partitions defined on a plurality of nodes that each have multiple physical processors coupled to a local memory, the plurality of logical partitions including a plurality of shared logical partitions that share a pool of common nodes in the computer system, the method comprising the steps of: (A) enabling NUMA optimizations in a plurality of operating systems in a plurality of shared logical partitions; (B) each operating system in the plurality of shared logical partitions performing the steps of: assigning a home node to a task when the task is created; assigning the task to a first virtual processor corresponding to the home node when the first virtual processor is available; and assigning the task to a second virtual processor corresponding to a node other than the home node with a suggestion parameter that specifies the home node when the first virtual processor is not available; (C) a partition manager that runs each virtual processor on one of the physical processors by performing the steps of: running the first virtual processor on a physical processor corresponding to the home node when the home node has an available physical processor; and running the first virtual processor on a physical processor corresponding to a node different than the home node when the home node has no available physical processor.

The method may further comprise the steps of: reading the suggestion parameter in the second virtual processor; and when the home node has an available physical processor, running the second virtual processor on a physical processor corresponding to the suggestion parameter. The method may further comprise the step of when the home node does not have an available physical processor, running the second virtual processor on a physical processor corresponding to the node other than the home node. Preferably, each of the plurality of nodes is assigned to a node group, and further comprising the step of: the partition manager running the first virtual processor on a physical processor corresponding to a node different than the home node and in the a node group as the home node when the home node has no available physical processor. Preferably, each virtual processor specifies one of the plurality of nodes as a home node for the virtual processor. Preferably, each virtual processor further specifies a suggested node written by an operating system corresponding to the virtual processor.

In a third aspect there is provided a computer program comprising computer program code to, when loaded into a computer system and executed thereon, cause said computer system to perform all the steps of the method of the second aspect.

The computer program may be implemented in the form of an article of manufacture comprising: (A) an operating system for each of a plurality of shared logical partitions that share a pool of common nodes, each node having multiple physical processors coupled to a local memory in a non-uniform memory access (NUMA) topology, the operating system assigning a home node to a task when the task is created, assigning the task to a first virtual processor corresponding to the home node when the first virtual processor is available, and assigning the task to a second virtual processor corresponding to a node other than the home node with a suggestion parameter that specifies the home node when the first virtual processor is not available; (B) a partition manager that runs each virtual processor on one of the physical processors, the partition manager running the first virtual processor on a physical processor corresponding to the home node when the home node has an available physical processor, the partition manager running the first virtual processor on a physical processor corresponding to a node different than the home node when the home node has no available physical processor; and (C) recordable media bearing (A) and (B).

Preferably, the partition manager reads the suggestion parameter in the second virtual processor, and when the home node has an available physical processor, the partition manager runs the second virtual processor on a physical processor corresponding to the suggestion parameter. Preferably, when the home node does not have an available physical processor, the partition manager runs the second virtual processor on a physical processor corresponding to the node other than the home node. Preferably, each of the plurality of nodes is assigned to a node group, and wherein the partition manager runs the first virtual processor on a physical processor corresponding to a node different than the home node and in a same node group as the home node when the home node has no available physical processor. Preferably, each virtual processor specifies one of the plurality of nodes as a home node. Preferably, each virtual processor comprises a suggestion parameter that specifies a suggested node, wherein the suggestion parameter is written by an operating system that dispatches tasks to the virtual processor.

In a NUMA-topology computer system that includes multiple nodes and multiple logical partitions, some of which may be dedicated and others of which are shared, NUMA optimizations are enabled in shared logical partitions. This is done by specifying a home node parameter in each virtual processor assigned to a logical partition. When a task is created by an operating system in a shared logical partition, a home node is assigned to the task, and the operating system attempts to assign the task to a virtual processor that has a home node that matches the home node for the task. The partition manager then attempts to assign virtual processors to their corresponding home nodes. If this can be done, NUMA optimizations may be performed in shared logical partitions as well as dedicated logical partitions.

BRIEF DESCRIPTION OF THE DRAWING(S)

A preferred embodiment of the present invention will now be described, by way of example only, with reference to the accompanying drawings, in which:

FIG. 1 is a block diagram of one specific prior art implementation for a node that may be used in a NUMA-topology computer system;

FIG. 2 is a block diagram of a sample prior art computer system that includes four of the nodes shown in FIG. 1;

FIG. 3 is a block diagram of a computer system similar to the computer system in FIG. 2 that is logically partitioned, but with operating systems that allow NUMA optimizations in shared logical partitions;

FIG. 4 is a flow diagram of a prior art method for a partition manager to run virtual processors in a logically partitioned computer system;

FIG. 5 is a block diagram of a prior art operating system that includes a NUMA optimization mechanism;

FIG. 6 is a flow diagram of a prior art method for enabling NUMA optimizations for dedicated logical partitions while disabling NUMA optimizations for shared logical partitions;

FIG. 7 is a flow diagram of a prior art method for a partition manager to run virtual processors in a way that allows NUMA optimizations for dedicated logical partitions;

FIG. 8 is a flow diagram of a prior art method for allowing NUMA optimizations in dedicated logical partitions;

FIG. 9 is a flow diagram of a method for enabling NUMA optimizations for both dedicated and shared logical partitions;

FIG. 10 is a block diagram of an operating system for logical partitions that allows NUMA optimizations in both dedicated and shared logical partitions;

FIG. 11 is a block diagram showing parameters in a virtual processor that allow NUMA optimizations in shared logical partitions;

FIG. 12 is a flow diagram of a method for an operating system in a logical partition to dispatch tasks in a way that supports NUMA optimizations in both dedicated and shared logical partitions;

FIG. 13 is a flow diagram of a method for a partition manager to run a virtual processor in a way that supports NUMA optimizations in both dedicated and shared logical partitions;

FIG. 14 is a diagram showing a first specific example that illustrates steps in an operating system in a shared logical partition dispatching a task to a virtual processor, followed by the partition manager assigning the virtual processor to a processor on a node;

FIG. 15 is a diagram showing a second specific example that illustrates steps in an operating system in a shared logical partition dispatching a task to a virtual processor, followed by the partition manager assigning the virtual processor to a processor on a node;

FIG. 16 is a diagram showing a third specific example that illustrates steps in an operating system in a shared logical partition dispatching a task to a virtual processor, followed by the partition manager assigning the virtual processor to a processor on a node; and

FIG. 17 is a diagram showing a fourth specific example that illustrates steps in an operating system in a shared logical partition dispatching a task to a virtual processor, followed by the partition manager assigning the virtual processor to a processor on a node.

DETAILED DESCRIPTION

The disclosure and claims herein relate to a computer system with a NUMA topology that is logically partitioned into multiple logical partitions that share resources, and how NUMA optimizations may be accomplished in these shared logical partitions.

FIG. 1 shows a sample prior art node 100 in a NUMA-topology computer system. The specific node 100 in FIG. 1 includes a processor chip 102 with four physical processors,

shown in FIG. 1 as 110A, 110B, 110C and 110D. The physical processors on chip 102 each typically include on-board cache memory, and are all interconnected via a common bus. This bus is also connected to a memory 120 external to the processor chip 102. Each physical processor may access the memory 120 with the same latency. In NUMA
5 terminology, the memory 120 is termed “local” to the physical processors on chip 102, while memory that is local to processors on other nodes is “remote” to the physical processors on chip 102.

A sample prior art computer system 200 is shown in FIG. 2 to include four nodes shown in
10 FIG. 1, which are labeled in FIG. 2 as 100A, 100B, 100C and 100D. These nodes are interconnected with various busses that allow any processor in any node to access any memory on any node. Because the physical processors can access their local memory (within the same node) much more quickly than accessing remote memory (on a different node), the time required to access different memories in such a system is non-uniform,
15 giving rise to the term Non-Uniform Memory Access (NUMA).

FIG. 3 shows a computer system 300 similar to the prior art computer system 200 in FIG. 2 that has been logically partitioned. Note the individual nodes are shown at the bottom of FIG. 3, and represent physical hardware that is allocated to the logical partitions. Computer
20 system 300 is logically partitioned into three separate logical partitions, shown in FIG. 3 as logical partitions 310A, 310B and 310C. Each logical partition specifies a memory size that represents a desired amount of memory for each logical partition that is allocated by partition manager 350 from the nodes. Each logical partition also specifies a computer capacity which corresponds to a number of physical processors that are allocated to the
25 logical partition. Note the compute capacity may be specified in any suitable fraction of a processor, and need not be specified in whole numbers. For the specific example in FIG. 3, logical partition 310A includes a memory size 315A of 20 GB and a compute capacity of 4 processors 320A; logical partition 310B includes a memory size 315B of 12 GB and a compute capacity of 4 processors 320B; and logical partition 310C includes a memory size
30 315C of 32 GB and a compute capacity of 8 processors 320C. The allocation of memory size and compute capacity could vary from the specific example shown in FIG. 3.

Each logical partition also includes virtual processors. Because the computer system 300 is logically partitioned, the operating systems 340A, 340B and 340C do not dispatch a task to a physical processor. Instead, the operating systems dispatch tasks to virtual processors. After a task is dispatched to a virtual processor, the partition manager 350 finds a suitable physical processor to run the virtual processor. This allows the operating system to function the same way it does when operating in a stand-alone computer system, dispatching tasks to virtual processors in a logical partition in the same way it dispatches tasks to physical processors in a stand-alone computer system. For the specific example in FIG. 3, logical partition 310A includes four virtual processors, shown as VP1-1, VP2-1, VP3-1 and VP4-1. In similar fashion, logical partition 310B includes four virtual processors, namely VP1-2, VP2-2, VP3-2 and VP4-2. Logical partition 310C includes eight virtual processors, namely VP1-3, VP2-3, VP3-3, VP4-3, VP5-3, VP6-3, VP7-3 and VP8-3. Operating system 340A dispatches tasks to one of the virtual processors VP1-1, VP2-1, VP3-1 and VP4-1 in logical partition 310A. Operating system 340B dispatches tasks to one of the virtual processors VP1-2, VP2-2, VP3-2 and VP4-2 in logical partition 310B. Operating system 340C dispatches tasks to one of the virtual processors VP1-3, VP2-3, VP3-3, VP4-3, VP5-3, VP6-3, VP7-3 and VP8-3 in logical partition 310C. As explained above, once an operating system dispatches a task to a virtual processor, the partition manager 350 then finds a suitable physical processor to run the virtual processor. In the specific example in FIG. 3, the number of virtual processors is equal to the computer capacity for each logical partition. Note, however, this need not be the case. In fact, the number of virtual processors could far exceed the compute capacity for a given logical partition.

In the specific configuration shown in FIG. 3, the first logical partition 310A is a dedicated logical partition, which means a single node 100A with four processors is dedicated to executed the four virtual processors VP1-1, VP2-1, VP3-1 and VP4-1 in the logical partition. Because logical partition 310A is a dedicated partition, each of the virtual processors may be assigned to a corresponding physical processor in node A 100A. The fact that logical partition 310A is a dedicated logical partition allows NUMA optimizations to be performed for logical partition 310A according to known methods. The second logical partition 310B and the third logical partition 310C are shown as shared logical partitions, which means they share a pool of resources shown at the bottom of FIG. 3 to include Node B 100B, Node C

100C and Node D 100D. This means the partition manager 350 may run any of the four virtual processors VP1-2, VP2-2, VP3-2 and VP4-2 in logical partition 310B on any physical processor in any of the three pooled nodes 100B, 100C and 100D. In similar fashion, the partition manager 350 may run any of the eight virtual processors VP1-3, VP2-3, VP3-3, VP4-3, VP5-3, VP6-3, VP7-3 and VP8-3 in logical partition 310C on any physical processor
5 in any of the three pooled nodes 100B, 100C and 100D.

Computer system 300 may include other features of computer systems that are not shown in FIG. 3 but are well-known in the art. For example, computer system 300 may include a display interface, a network interface, and a mass storage interface to an external direct access storage device (DASD). The display interface is used to directly connect one or more displays to computer system 300. These displays, which may be non-intelligent (i.e., dumb) terminals or fully programmable workstations, are used to provide system administrators and users the ability to communicate with computer system 300. Note, however, that while a display interface is provided to support communication with one or more displays, computer system 300 does not necessarily require a display, because all needed interaction with users and other processes may occur via the network interface.
10
15

The network interface is used to connect the computer system 300 to multiple other computer systems via a network. The network interface and network broadly represent any suitable way to interconnect electronic devices, regardless of whether the network comprises present-day analog and/or digital techniques or via some networking mechanism of the future. In addition, many different network protocols can be used to implement a network. These protocols are specialized computer programs that allow computers to communicate across a network. TCP/IP (Transmission Control Protocol/Internet Protocol) is an example of a suitable network protocol.
20
25

The mass storage interface is used to connect mass storage devices, such as a direct access storage device, to computer system 300. One specific type of direct access storage device is a readable and writable CD-RW drive, which may store data to and read data from a CD-RW.
30

At this point, note that while the description above is in the context of a fully functional computer system, those skilled in the art will appreciate that the operating system and partition manager disclosed herein may be distributed as an article of manufacture in a variety of forms, and the claims extend to all suitable types of computer-readable media used to actually carry out the distribution, including recordable media such as floppy disks and CD-RW.

In the prior art, dedicated logical partitions and shared logical partitions are treated differently in a NUMA environment. Referring to FIG. 4, a prior art method 400 determines if a partition is dedicated (step 410). If so (step 410=YES), the partition manager assigns each virtual processor to a corresponding physical processor (step 420). The operating system in the dedicated partition then dispatches tasks to the virtual processors (step 430). In response, the partition manager runs each virtual processor on its corresponding physical processor (step 440). If the partition is a shared logical partition (step 410=NO), the partition manager associates each virtual processor to a pool of physical processors (step 450). Thus, for the computer system 300 shown in FIG. 3, each virtual processor in both logical partitions 310B and 310C are associated with the pool of processors in nodes 100B, 100C and 100D. The operating system in the shared partition dispatches tasks to the virtual processors (step 460). In response, the partition manager runs the virtual processor on any one of the physical processors in the pool (step 470).

A prior art operating system used in a NUMA environment is shown in FIG. 5. Operating system 500 includes a NUMA optimization mechanism 510 that changes the function of the task dispatcher 520, main storage manager 530 and resource manager 540 to allow NUMA optimization in stand-alone computer systems and in dedicated logical partitions. The task dispatcher 520, main storage manager 530 and resource manager 540 also include other known functions that have nothing to do with NUMA optimizations, and which are not shown in FIG. 5. Part of prior art NUMA optimization mechanism 510 deals with allocating memory to tasks in a location where the task is most likely to execute, which is not discussed in detail herein.

The NUMA optimization mechanism 510 typically does not perform NUMA optimizations

in shared logical partitions. This is shown in prior art method 600 shown in FIG. 6. NUMA optimizations are enabled in the operating systems in all dedicated logical partitions (step 610). NUMA optimizations are disabled in the operating systems in all shared logical partitions (step 620). Note the enabling in step 610 and disabling in step 620 may be done in any suitable way. For example, if there is a flag in the operating system that enables NUMA optimizations, the flag would be set for the dedicated logical partitions in step 610 and would not be set for the shared logical partitions in step 620.

Some NUMA optimizations have been performed in dedicated logical partitions, as shown by prior art method 700 in FIG. 7. First, the partition manager assigns each virtual processor to a corresponding node (step 710). This differs from step 420 in FIG. 4 because the virtual processor is assigned to a node that contains multiple processors, and is not assigned to a specific physical processor. The operating system then specifies a home node for a task when the task is created, and dispatches that task with the specification of the home node to a virtual processor that has a home node that corresponds to the task's home node (step 720). The partition manager then tries to run the virtual processor on one of the physical processor in the home node (step 730). Method 700 recognizes that memory access time for all processors on a node to the local memory is equal, so the partition manager need not assign a virtual processor to a specific physical processor, but may assign a virtual processor to a node, allowing the virtual processor to run on any of the physical processors on that node. Note, however, method 700 only applies to dedicated logical partitions. In fact, steps 710, 720 and 730 could replace steps 420, 430 and 440 in FIG. 4, which changes the steps for the dedicated partition (when step 410=YES), but does not change the steps for shared logical partitions (when step 410=NO).

Referring to FIG. 8, a prior art method 800 is performed when NUMA optimizations are enabled for dedicated partitions (step 810). The operating system in the dedicated partition assigns a "home node" for each task as the task is created (step 820). The operating system then looks for a virtual processor that matches the home node parameter in the task, and dispatches the task with the specified home node to a virtual processor that corresponds to the home node (step 830). If a physical processor on the home node is available (step 840=YES), the partition manager runs the virtual processor on a physical processor on the

home node (step 850). If a physical processor on the home node is not available (step 840=NO), the partition manager runs the virtual processor on a physical processor on a different node (step 860). Note there are memory allocation issues that would also be performed to accomplish NUMA optimizations that are not shown in FIG. 8 or discussed in detail herein.

A simple example will illustrate. Let's assume a dedicated partition had two nodes X and Y, each with four processors as shown in FIG. 1. Let's assume the operating system assigns X as the home node for a new task (step 820), then dispatches the task to a virtual processor corresponding to node X (step 830). If the partition manager determines a physical processor on node X is available (step 840=YES), the partition manager runs the virtual processor on a physical processor on node X (step 850). If, however, the partition manager determines there is no physical processor on node X available (step 840=NO), the partition manager may still run the virtual processor on a physical processor on node Y (step 860). Because the partition manager will try to run a virtual processor on its corresponding home node, method 800 supports NUMA optimizations in dedicated logical partitions.

In the prior art, NUMA optimizations may be performed in dedicated logical partitions but are not performed in shared logical partitions. The disclosure and claims herein overcome this limitation by allowing NUMA optimizations to be performed in shared logical partitions. Referring to FIG. 9, a method begins by enabling NUMA optimizations by operating systems in all dedicated logical partitions (step 610). Note this step is preferably the same as step 610 in prior art method 600 in FIG. 6. Next, NUMA optimizations are enabled by operating systems in all shared logical partitions (step 920). By enabling NUMA optimizations in shared logical partitions, method 900 provides enhanced performance when compared to method 600 in FIG. 6 that disables NUMA optimizations in shared logical partitions.

Referring to FIG. 10, an operating system 340 is representative of the operating systems 340A, 340B and 340C shown in FIG. 3. Operating system 340 includes a NUMA optimization mechanism 1010 that allows NUMA optimizations to be performed in shared logical partitions. A task dispatcher 1020 includes a suggestion parameter mechanism 1022

that can specify a suggested node for a virtual processor even when the virtual processor has a home node that is different than the suggested node. The main storage manager 1030 allocates memory to a task in a way that attempts to allocate memory local to a processor where a virtual processor for the task is likely to execute. The resource manager 1040 includes a home node assignment mechanism 1042 that assigns a home node to each task as the task is created. Resource manager 1040 attempts to achieve a balance between the workloads of different virtual processors, and therefore determines when a task is created which home node to assign to the task depending on current conditions when the task is created. The resource manager 1040 may also optionally include a node group mechanism 1044 that allows grouping nodes into node groups. Note the various mechanisms shown in FIG. 10 are one suitable implementation in accordance with the disclosure and claims herein. However, the suggestion parameter mechanism 1022, home node assignment mechanism 1042 and node group mechanism 1044 could be implemented in any suitable form or location within operating system 340. In addition, these could also be implemented in code external to the operating system 340 that is accessed by the operating system 340.

Each virtual processor includes two parameters, as shown in virtual processor X 1100 shown in FIG. 11. The first is a home node parameter 1110 that is assigned by the partition manager to relate a virtual processor to a particular node in the hardware. The second is an operating system suggested node parameter 1120 that is written to by the operating system to suggest a node to the partition manager even if the operating system dispatches a task to a virtual processor that does not have a home node that corresponds to the home node of the task. This allows the partition manager to determine on-the-fly whether a physical processor in the home node becomes available, and to run the virtual processor that has a different home node on the suggested node, which preferably corresponds to the home node of the task. In this manner a partition manager has information from the operating system that allows the partition manager to override the home node of the virtual processor and assign the virtual processor to a different node instead, which is the home node of the task that was dispatched to the virtual processor.

Referring to FIG. 12, a method 1200 begins when NUMA optimizations are enabled for both dedicated and shared logical partitions (step 1210). An operating system assigns a home

node for each task as the task is created, and optionally determines a node group for each task (step 1220). The disclosure and claims recognize that a NUMA topology may include many different levels of memory hierarchy. This means nodes at a similar level of memory hierarchy may be grouped. If a physical processor on a home node is not available, but a
5 physical processor on a different node on the same group is available, the partition manager may run the virtual processor on the different node, thereby increasing system performance when compared to running the virtual processor on a node outside the node group.

If a virtual processor is available that has a home node parameter 1110 that corresponds to
10 the home node for the task assigned in step 1220 (step 1230=YES), the operating system dispatches the task with the specified home node to a virtual processor that corresponds to the specified home node (step 1240). If a virtual processor corresponding to the home node is not available (step 1230=NO), method 1200 next determines whether a virtual processor
15 corresponding to the node group of the home node is available (step 1250). If so (step 1250=YES), the operating system dispatches the task with the specified home node to a virtual processor corresponding to the node group (step 1260). If no virtual processor corresponding to the node group is available (step 1250=NO), the operating system
20 dispatches the task with the specified home node to a virtual processor corresponding to a node that is not the home node and is not in the node group, with a suggestion parameter (e.g., 1120 in FIG. 11) that suggests the home node or node group (step 1270).

While method 1200 in FIG. 12 shows steps in an operating system dispatching tasks to virtual processors, method 1300 in FIG. 13 shows steps for a partition manager to run a virtual processor on a physical processor. Method 1300 begins when the partition manager
25 needs to run a virtual processor that includes home node or node group parameters (step 1310), such as home node parameter 1110 in FIG. 11. If there is a physical processor available that corresponds to the home node parameter specified in the virtual processor (step 1320=YES), the partition manager runs the virtual processor on a physical processor on the home node (step 1330). If there is no physical processor available that corresponds to
30 the home node parameter (step 1320=NO), but a physical processor outside of the home node but within the node group is available (step 1340=YES), the partition manager runs the virtual processor on a physical processor outside of the home node with within the node

group (step 1350). If there is no physical processor outside of the home node but within the node group that is available (step 1340=NO), the partition manager runs the virtual processor on a physical processor corresponding to the home node parameter in the virtual processor (step 1360), which is a physical processor on a node different than the home node parameter for the task.

Some simple examples are now presented to illustrate how NUMA optimizations are enabled in shared logical partitions. Referring to FIG. 14, we assume for this example the operating system in partition 3, namely operating system 340C in FIG. 3, begins by creating a task, shown at 1410 in FIG. 14. The operating system assigns a home node to the task, as shown at 1420. The operating system may optionally determine a node group for the task. We assume for this example nodes B and C in FIG. 3 are in Group 2, so we indicate at 1430 that node C is in group 2. We next assume VP1-3 and VP2-3 are not available at 1440, but VP3-3 is available at 1450. The home node for VP3-3 is Node C 100C shown in FIG. 3. The operating system then dispatches the task to VP3-3 at 1460. Note the OS suggested node parameter 1120 shown in FIG. 11 is preferably set to NULL, meaning the operating system does not make any suggestion for a node other than the home node that corresponds to the virtual processor. At this point the virtual processor needs to be run by the partition manager. The partition manager reads the home node parameter 1110 from VP3-3 at 1470 and determines Node C is the home node, then determines a physical processor on Node C 100C is available at 1480. The partition manager then runs the virtual processor on a physical processor in the home node, Node C 100C, at 1490. This example illustrates that when a virtual processor corresponding to the home node of the task is available, and a physical processor corresponding to the home node is also available, the virtual processor is run on a physical processor in the specified home node.

Referring to FIG. 15, we assume the same steps are performed by the operating system as shown in FIG. 14, resulting in a task 1410 with a home node 1420 of Node C and a group 1430 of Group 2. We assume for this example none of the virtual processors that have a home node of Node C are available at 1510. We further assume that two of the virtual processors that have a home node of node D are not available at 1520, but we assume VP7-3 that has a home node of Node D is available at 1530. The operating system dispatches the

task with the OS Suggested Node 1120 of Node C to VP7-3. Note the home node of VP7-3 is Node D, but the operating system suggested node 1120 is Node C. This suggestion is provided to allow the partition manager to run the virtual processor on Node C if there is a physical processor on Node C available. The partition manager reads data from VP7-3 at 1550, and determines its home node is Node D. We assume the physical processors on Node C are not available at 1555, but a physical processor on Node B is available at 1560. The partition manager runs the virtual processor on Node B at 1570 due to the OS Suggested Node of Node C, and because Node B is in the same group as Node C. Thus, even though the task cannot be dispatched to a virtual processor that corresponds to the home node of the task, the virtual processor may be run on a node in the same group, thereby improving system performance.

Referring to FIG. 16, this example assumes all the same steps are taken by the operating system above the dotted line, as discussed in detail above. In response, the partition manager reads data from VP7-3 at 1610, and determines the home node of VP7-3 is Node D while the OS suggested node 1120 is Node C. We assume for this example a physical processor on Node C becomes available at 1620, before the partition manager runs the virtual processor VP7-3. Because a physical processor on Node C is available at 1630, the partition manager runs the virtual partition on Node C due to the OS suggested node parameter 1120 at 1640. This example shows how the partition manager can use the OS suggested node parameter 1120 to run a virtual processor on a node that is different than the home node for the virtual processor, but on a node that matches the home node parameter 1420 of the task. The partition manager may thus override the home node specified in the virtual processor, and choose instead to execute the task on the node specified in the OS suggested node parameter 1120.

Referring to FIG. 17, this example again assumes the operating system performs all the same steps shown in FIGS. 15 and 16 above the dotted line, as discussed in detail above. In response, the partition manager reads data from VP7-3 at 1710, and determines the home node of VP7-3 is Node D while the OS suggested node 1120 is Node C. We assume the physical processors on Node C are unavailable at 1720, the physical processors on Node B are unavailable at 1730, but a physical processor for Node D is available at 1740, as also

shown at 1750. As a result, the partition manager runs the virtual processor on Node D because there are no processors available on Node C (which corresponds to the OS Suggested Node) and there are no processor available on Node B (which is in the same node group, Group 2, as Node C). The partition manager thus runs the virtual processor on a physical processor on Node D, which is the home node for VP7-3.

The disclosure and claims herein enable NUMA optimizations on shared logical partitions. Each virtual processor has a home node parameter and an operating system suggested node parameter. When an operating system creates a task, it assigns a home node parameter for the task. The operating system then dispatches the task to a virtual processor, giving preference first to a virtual processor corresponding to the home node specified in the task, giving second preference to a virtual processor that is in the same node group as the home node specified in the task, and if neither of these preferences may be met, dispatching the task to a virtual processor that has a home node parameter different than the home node of the task. When the partition manager is called to run the virtual processor, the partition manager first determines whether the OS suggested node parameter in the virtual processor specifies a node other than the home node of the virtual processor. If so, the partition manager first attempts to run the virtual processor on a physical processor on the node specified in the OS suggested node parameter. If no physical processors on the node specified in the OS suggested node are available, the partition manager determines whether there is a node in the same node group as the OS suggested node, and if so, the task is dispatched to a processor on a different node but in the same node group. If there is no physical processor available in either the node specified in the OS suggested node parameter or in its node group, the partition manager runs the virtual processor on a physical processor in the node corresponding to the home node parameter of the virtual processor. In this manner NUMA optimizations may be enabled in shared logical partitions, thereby increasing the performance of this type of logically-partitioned computer system.

One skilled in the art will appreciate that many variations are possible within the scope of the claims.

CLAIMS

1) A computer system comprising:

a plurality of logical partitions defined on a plurality of nodes that each have multiple physical processors coupled to a local memory in a non-uniform memory access (NUMA) topology, the plurality of logical partitions including a plurality of shared logical partitions that share a pool of common nodes;

an operating system in each of the plurality of shared logical partitions that assigns a home node to a task when the task is created, assigns the task to a first virtual processor corresponding to the home node when the first virtual processor is available, and assigns the task to a second virtual processor corresponding to a node other than the home node with a suggestion parameter that specifies the home node when the first virtual processor is not available; and

a partition manager that runs each virtual processor on one of the physical processors, the partition manager running the first virtual processor on a physical processor corresponding to the home node when the home node has an available physical processor, the partition manager running the first virtual processor on a physical processor corresponding to a node different than the home node when the home node has no available physical processor.

2) The computer system of claim 1 wherein the partition manager reads the suggestion parameter in the second virtual processor, and when the home node has an available physical processor, the partition manager runs the second virtual processor on a physical processor corresponding to the suggestion parameter.

3) The computer system of claim 2 wherein, when the home node does not have an available physical processor, the partition manager runs the second virtual processor on a physical processor corresponding to the node other than the home node.

4) The computer system of claim 1 wherein each of the plurality of nodes is in a node group, and wherein the partition manager runs the first virtual processor on a physical processor corresponding to a node different than the home node and in a same node group as

the home node when the home node has no available physical processor.

5) The computer system of claim 1 wherein each virtual processor specifies one of the plurality of nodes as a home node.

6) The computer system of claim 5 wherein each virtual processor comprises a suggestion parameter that specifies a suggested node, wherein the suggestion parameter is written by an operating system that dispatches tasks to the virtual processor.

7) A method for performing non-uniform memory access (NUMA) optimizations in a computer system that includes a plurality of logical partitions defined on a plurality of nodes that each have multiple physical processors coupled to a local memory, the plurality of logical partitions including a plurality of shared logical partitions that share a pool of common nodes, the method comprising the steps of:

(A) enabling NUMA optimizations in a plurality of operating systems in a plurality of shared logical partitions;

(B) each operating system in the plurality of shared logical partitions performing the steps of:

assigning a home node to a task when the task is created;

assigning the task to a first virtual processor corresponding to the home node when the first virtual processor is available; and

assigning the task to a second virtual processor corresponding to a node other than the home node with a suggestion parameter that specifies the home node when the first virtual processor is not available;

(C) a partition manager that runs each virtual processor on one of the physical processors by performing the steps of:

running the first virtual processor on a physical processor corresponding to the home node when the home node has an available physical processor; and

running the first virtual processor on a physical processor corresponding to a node different than the home node when the home node has no available physical processor.

8) The method of claim 7, further comprising the steps of:

reading the suggestion parameter in the second virtual processor; and
when the home node has an available physical processor, running the second virtual processor on a physical processor corresponding to the suggestion parameter.

5 9) The method of claim 8, further comprising the step of:

when the home node does not have an available physical processor, running the second virtual processor on a physical processor corresponding to the node other than the home node.

10 10) The method of claim 9, wherein each of the plurality of nodes is assigned to a node group, and further comprising the step of:

the partition manager running the first virtual processor on a physical processor corresponding to a node different than the home node and in the a node group as the home node when the home node has no available physical processor.

15

11) The method of claim 7, wherein each virtual processor specifies one of the plurality of nodes as a home node for the virtual processor.

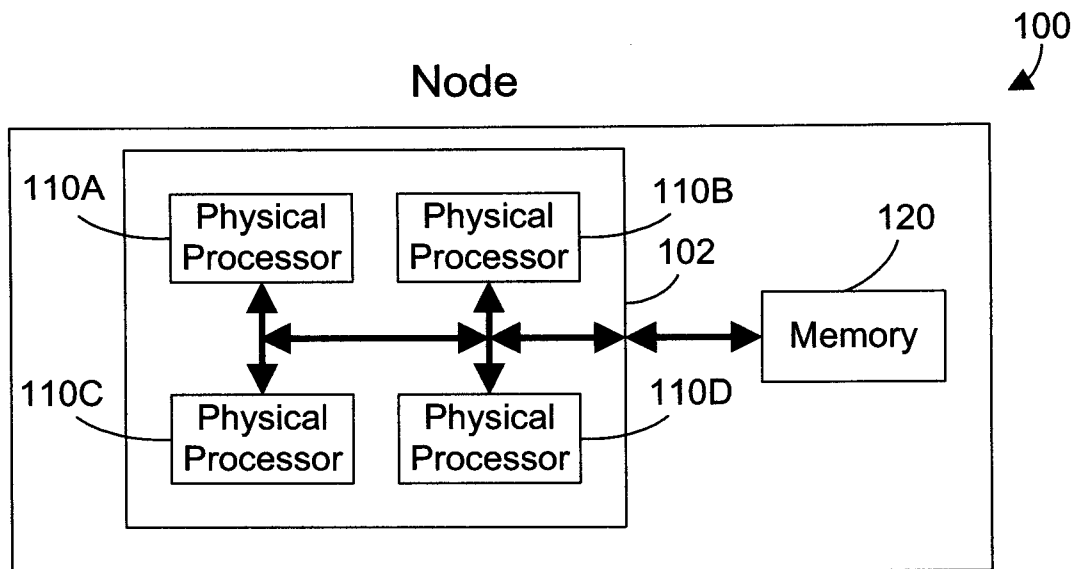
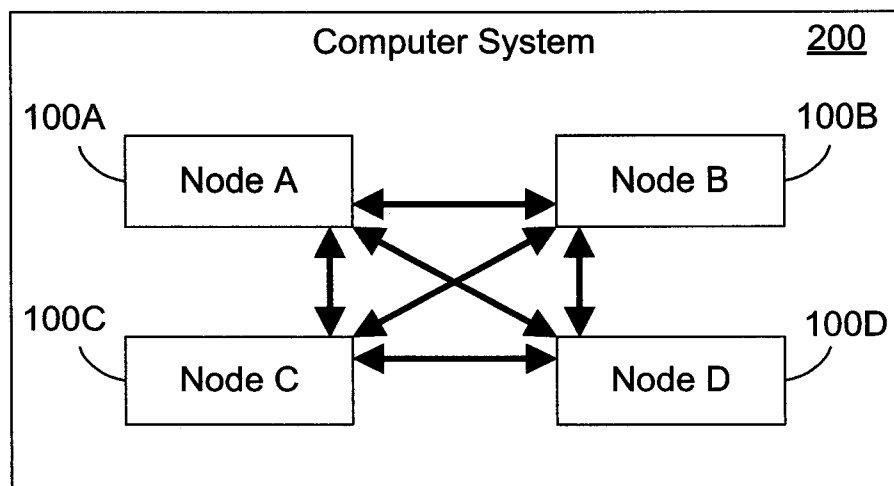
20

12) The method of claim 11, wherein each virtual processor further specifies a suggested node written by an operating system corresponding to the virtual processor.

13) A computer program comprising computer program code to, when loaded into a computer system and executed thereon, cause said computer system to perform all the steps of the method of any of claims 7 to 12.

25

1/12

FIG. 1 Prior ArtFIG. 2 Prior Art

2/12

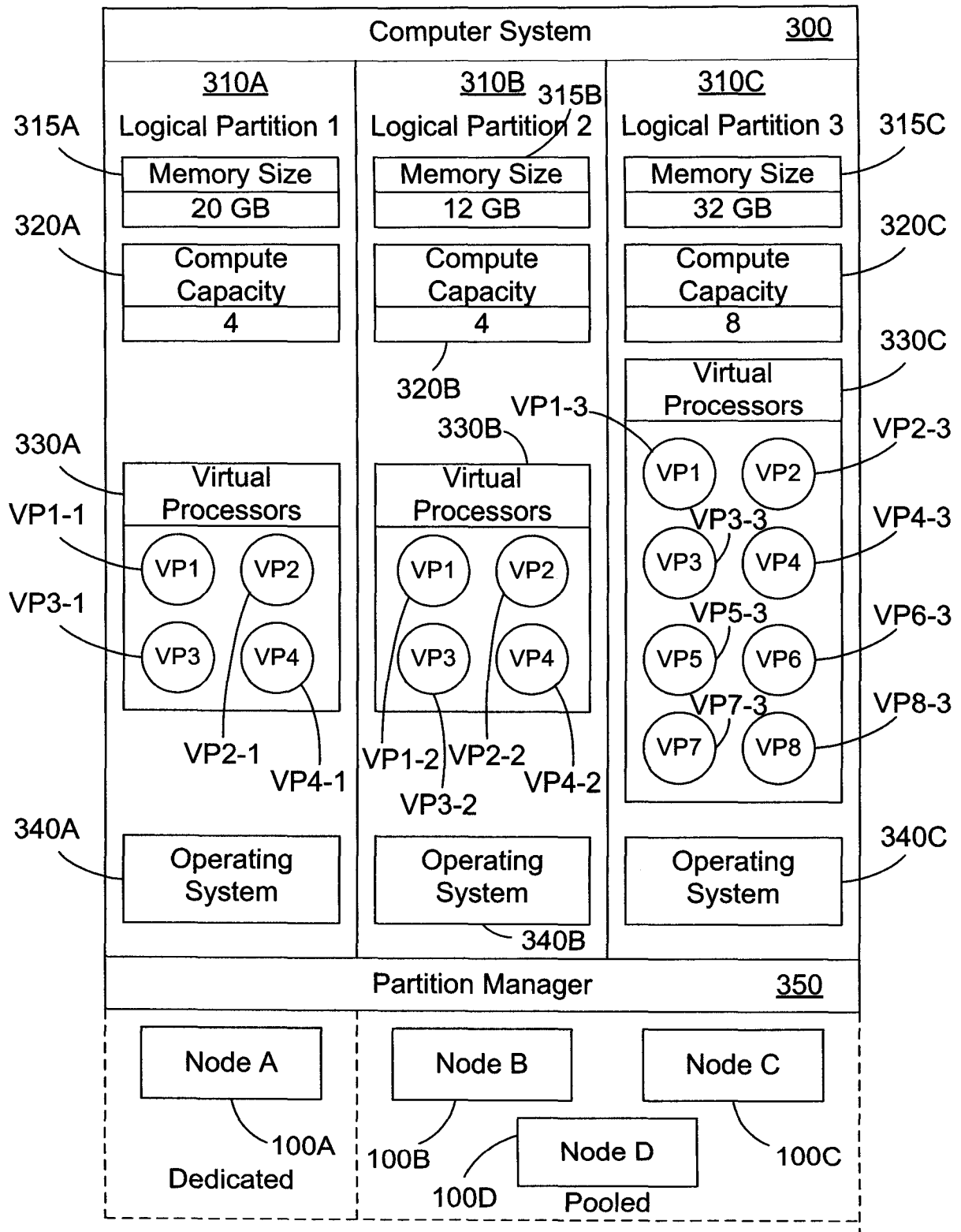


FIG. 3

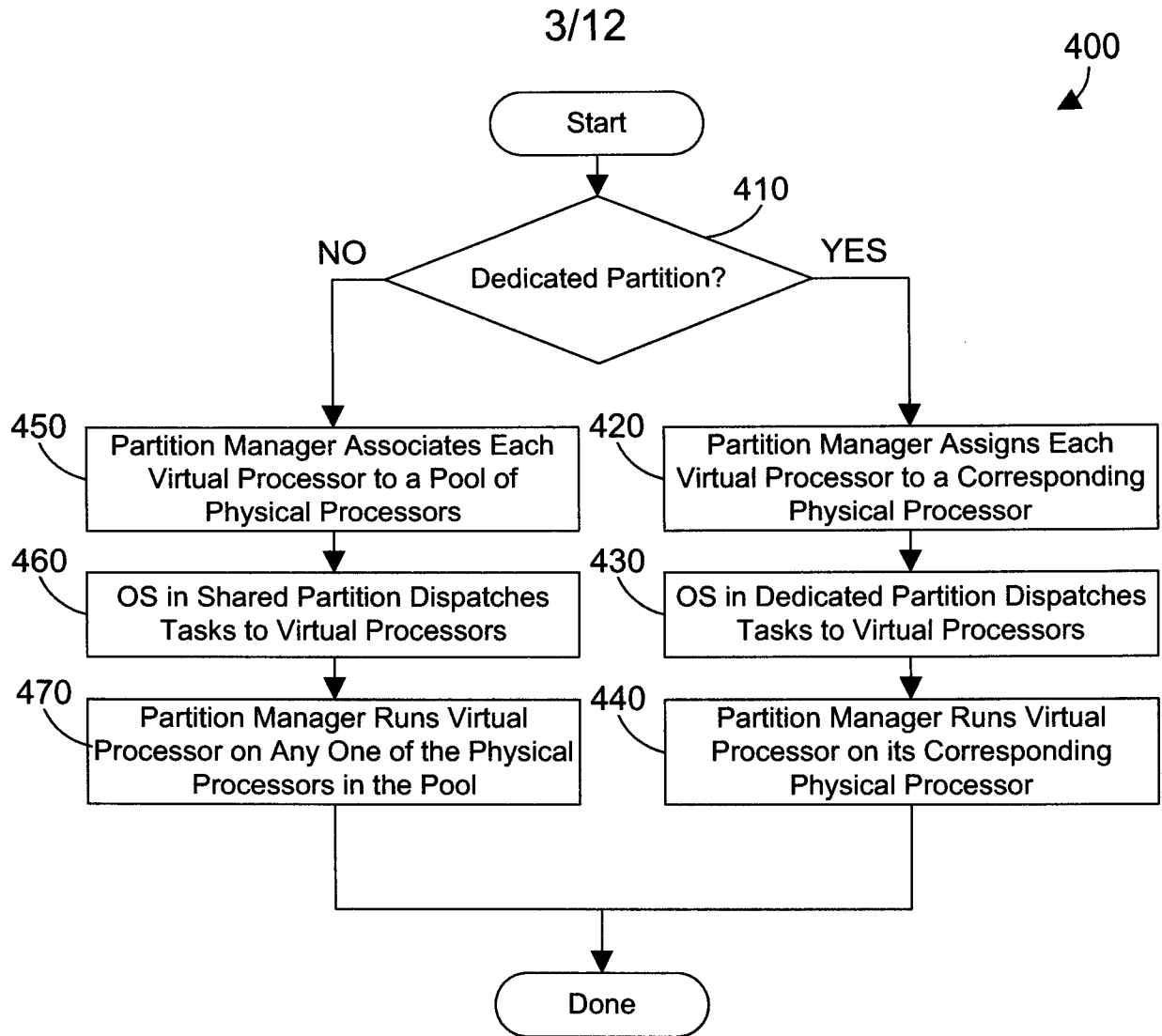


FIG. 4

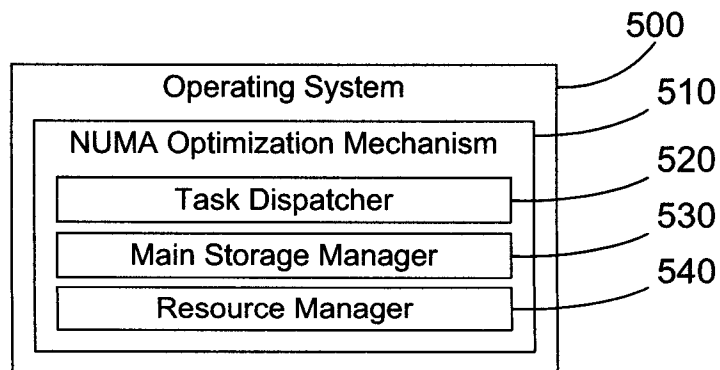
Prior Art

FIG. 5

Prior Art

4/12

600

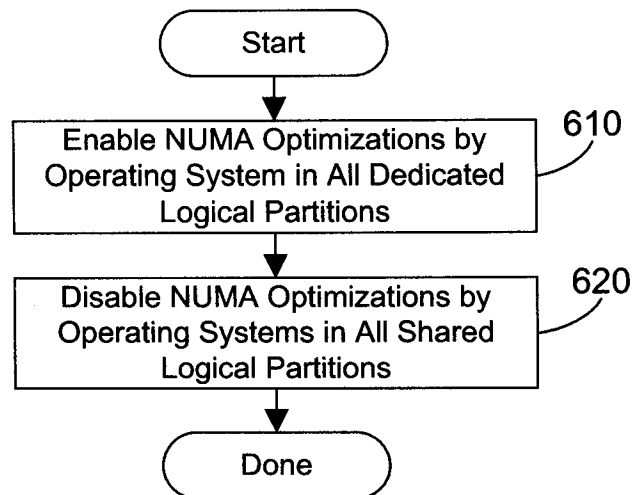


FIG. 6

Prior Art

700

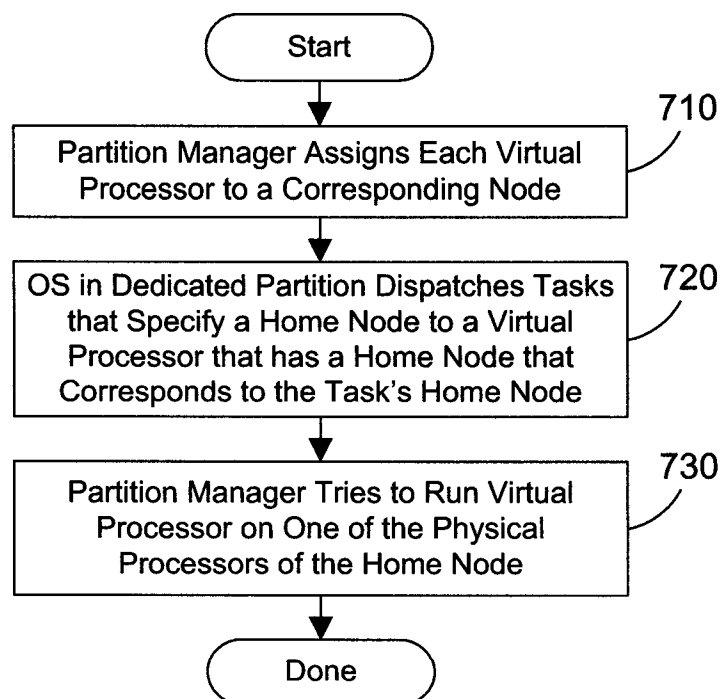


FIG. 7

Prior Art

5/12

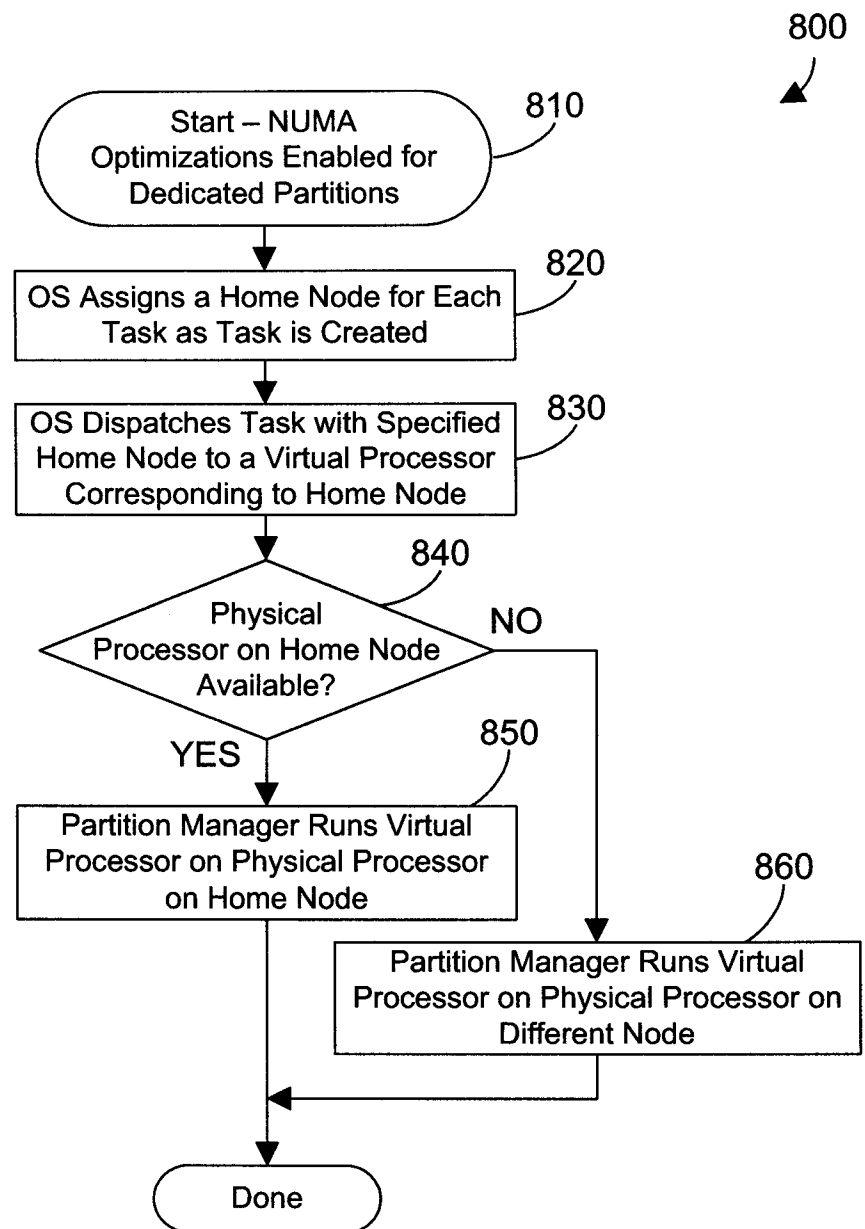


FIG. 8

Prior Art

6/12

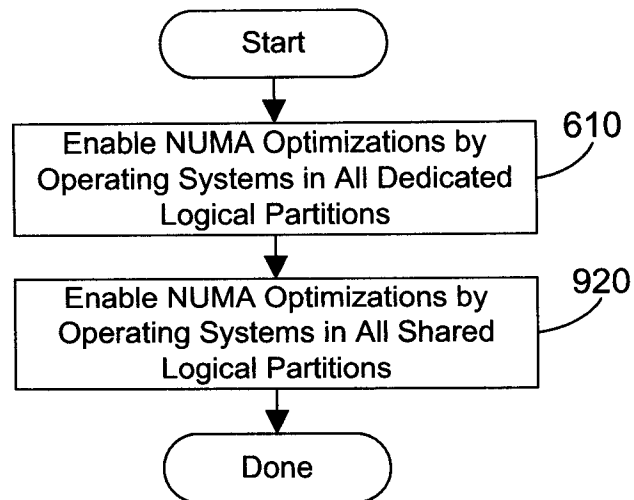


FIG. 9

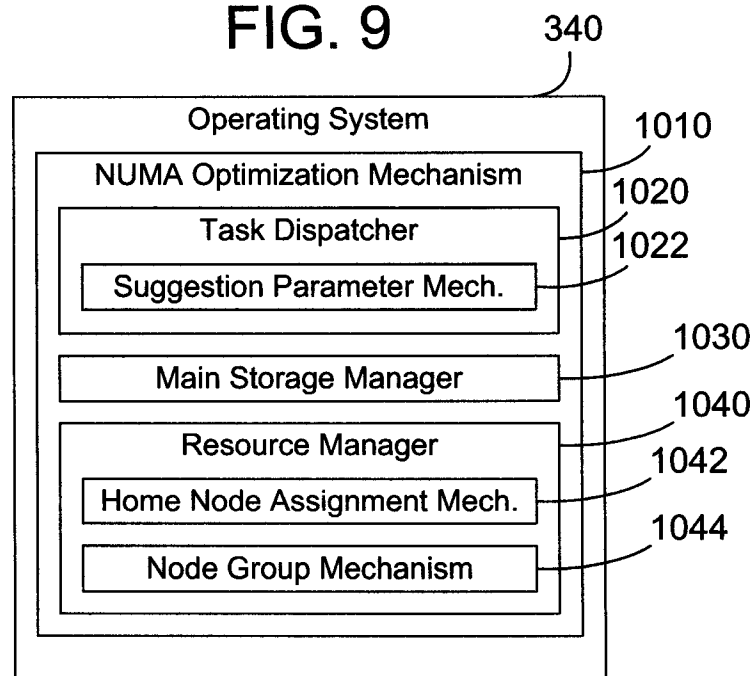


FIG. 10

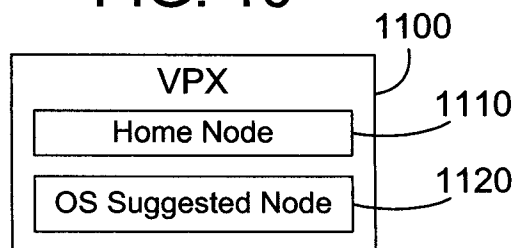
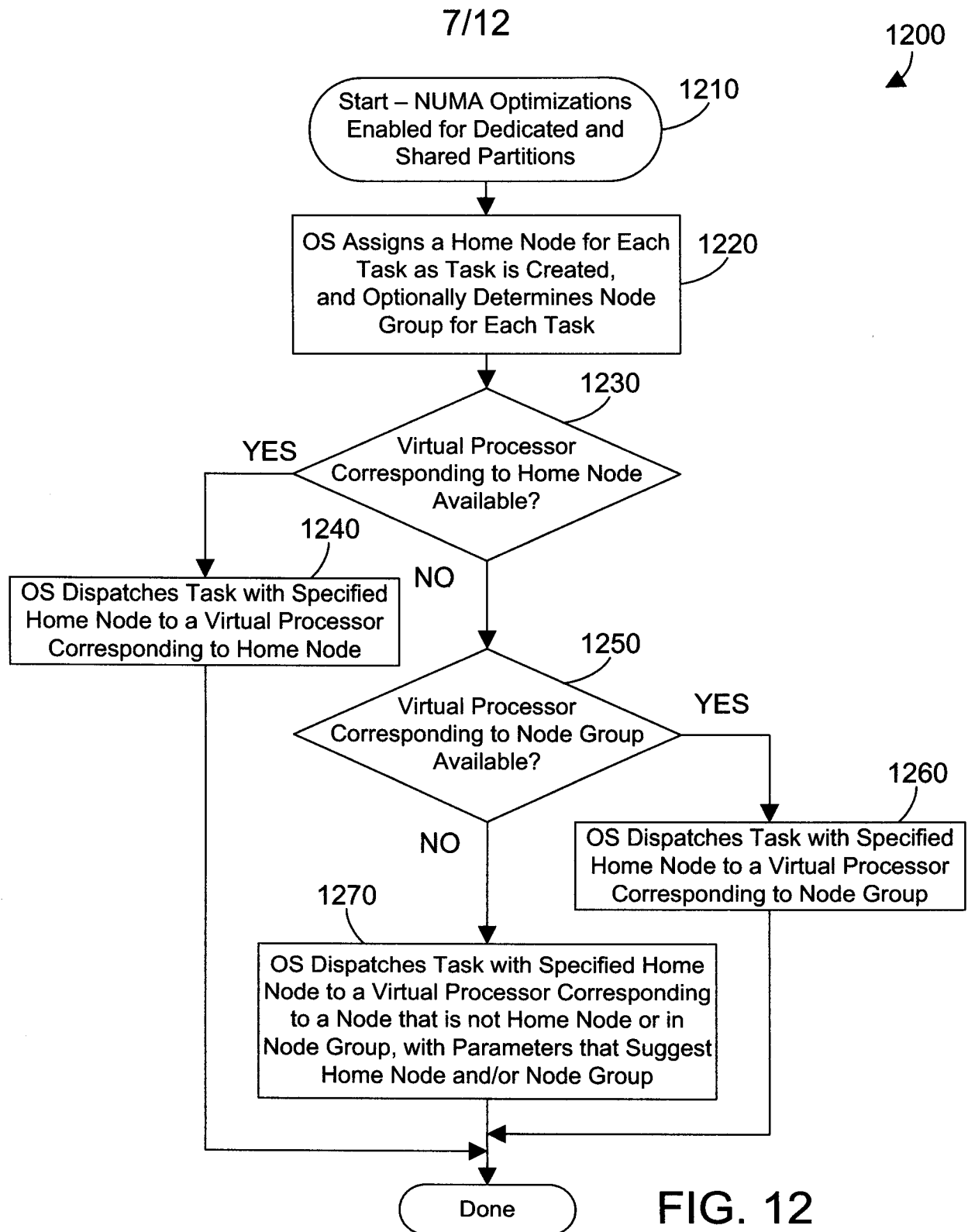


FIG. 11



8/12

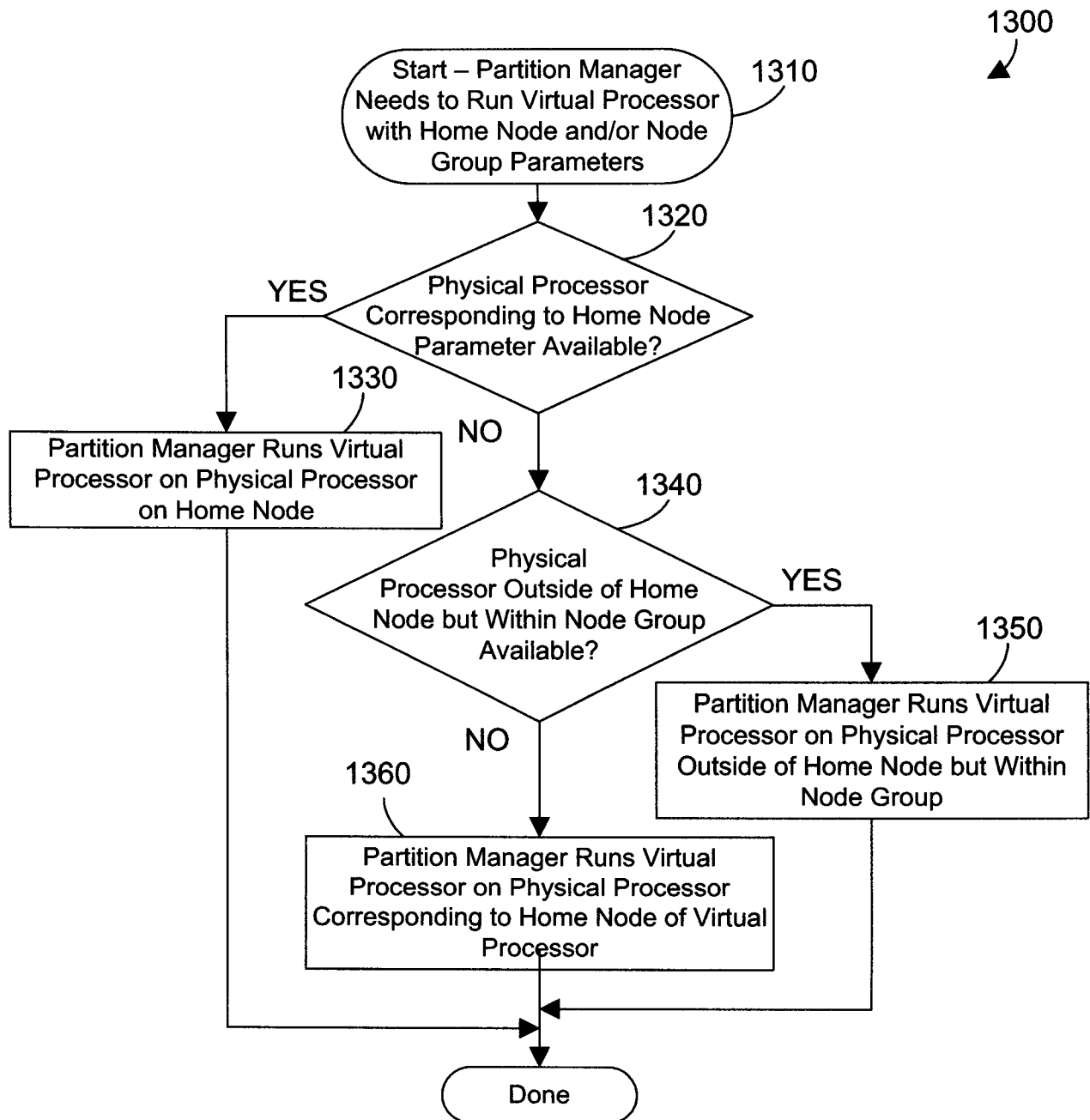


FIG. 13

S1

9/12

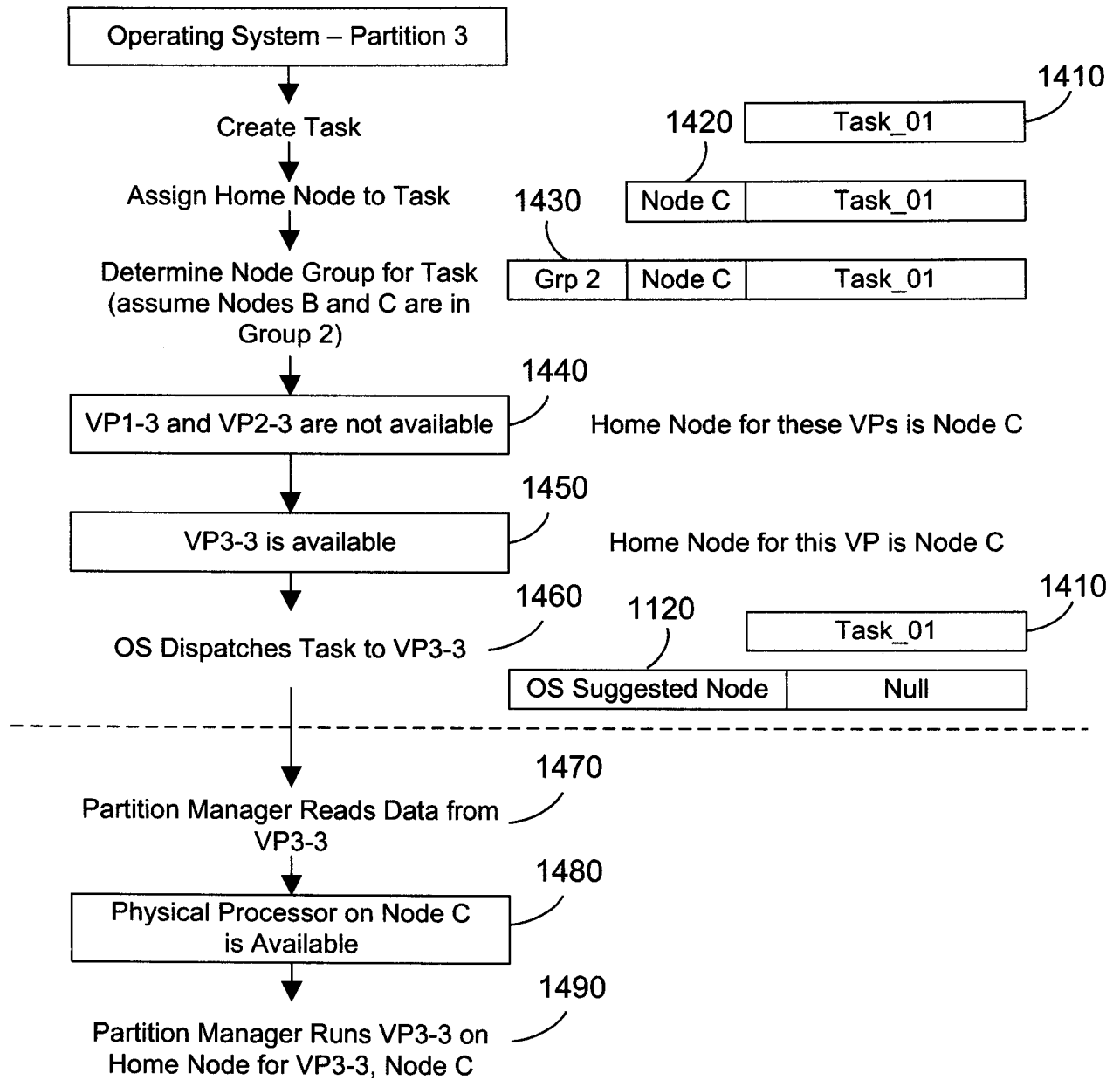


FIG. 14

10/12

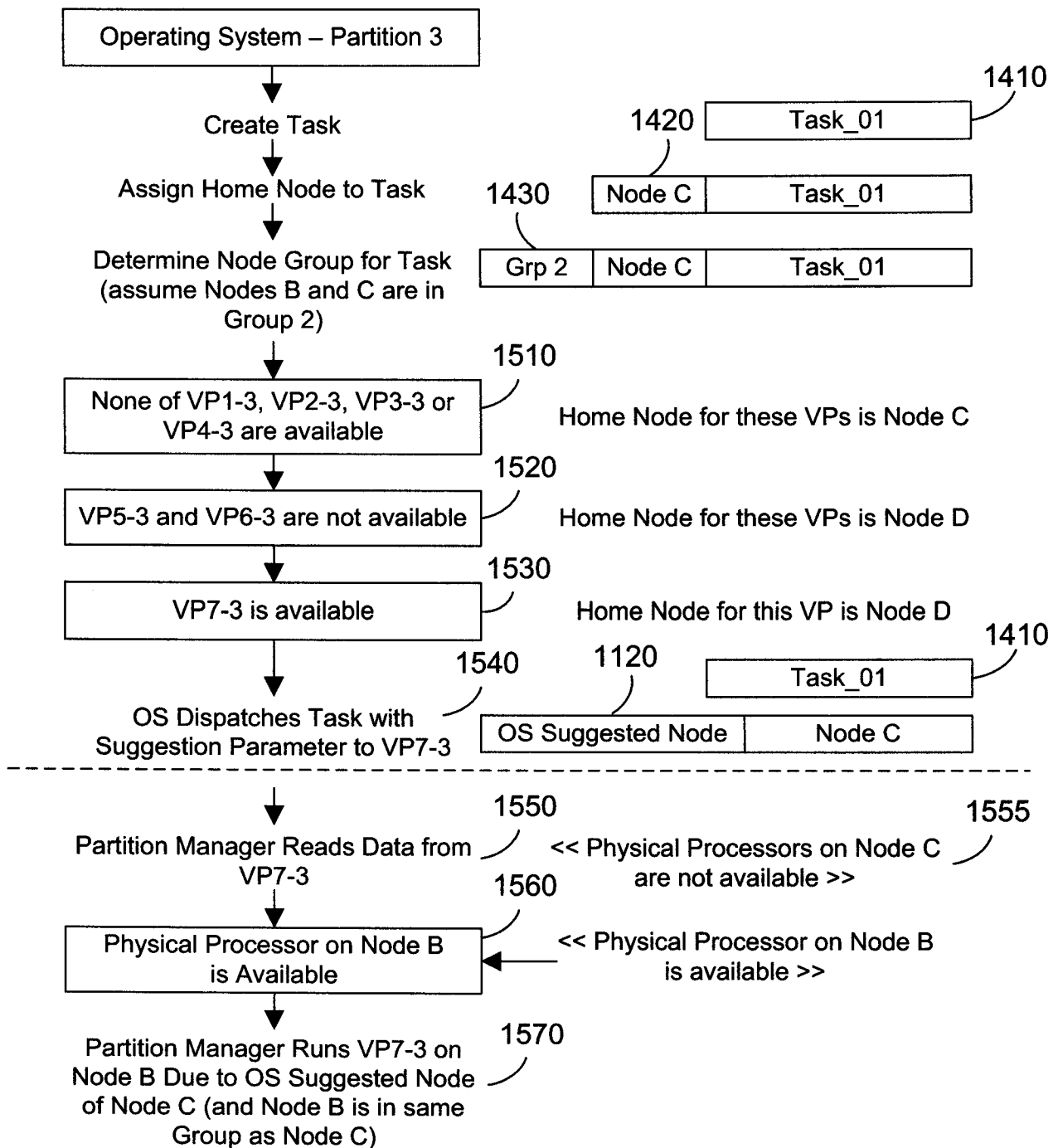


FIG. 15

11/12

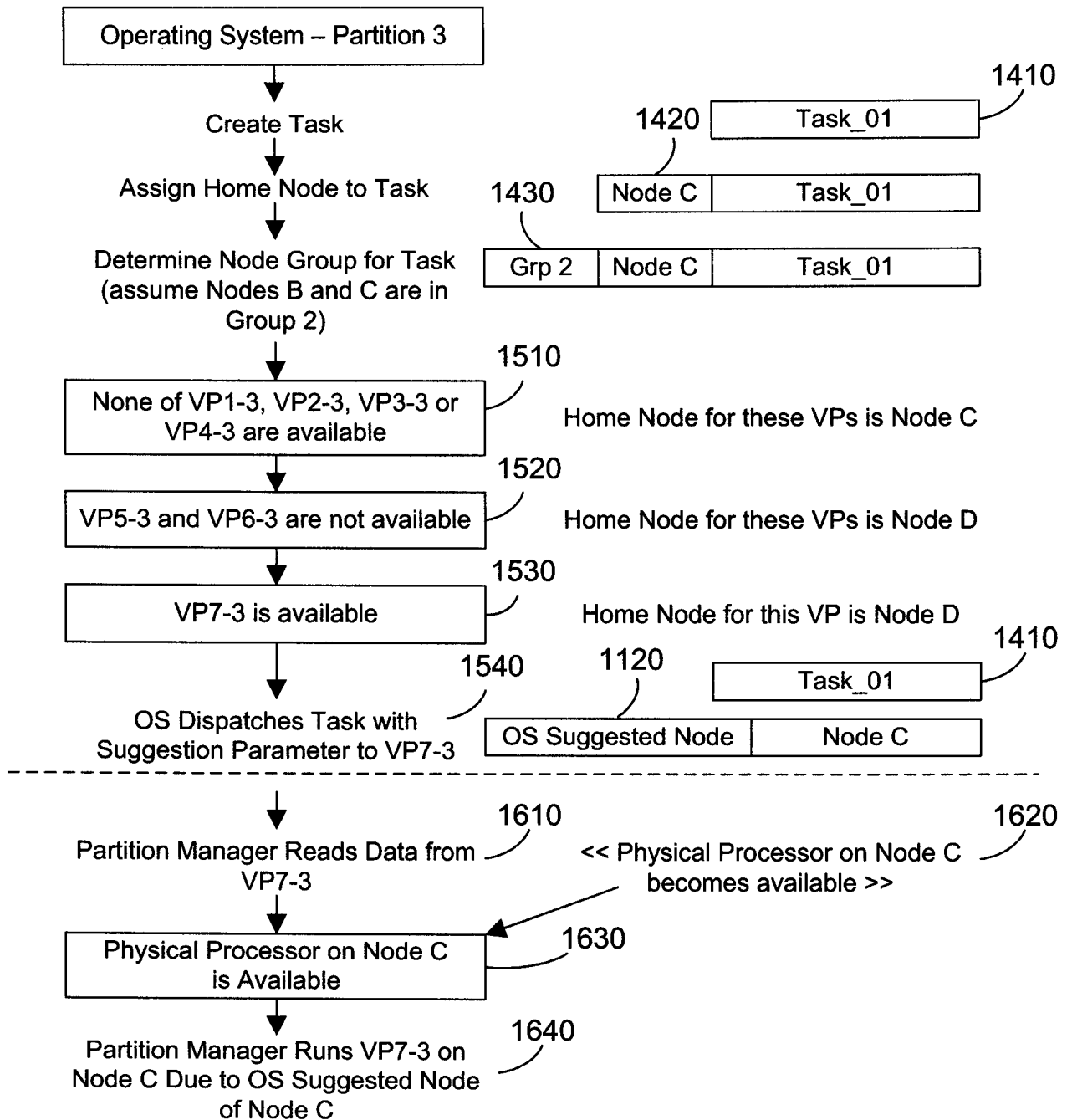


FIG. 16

12/12

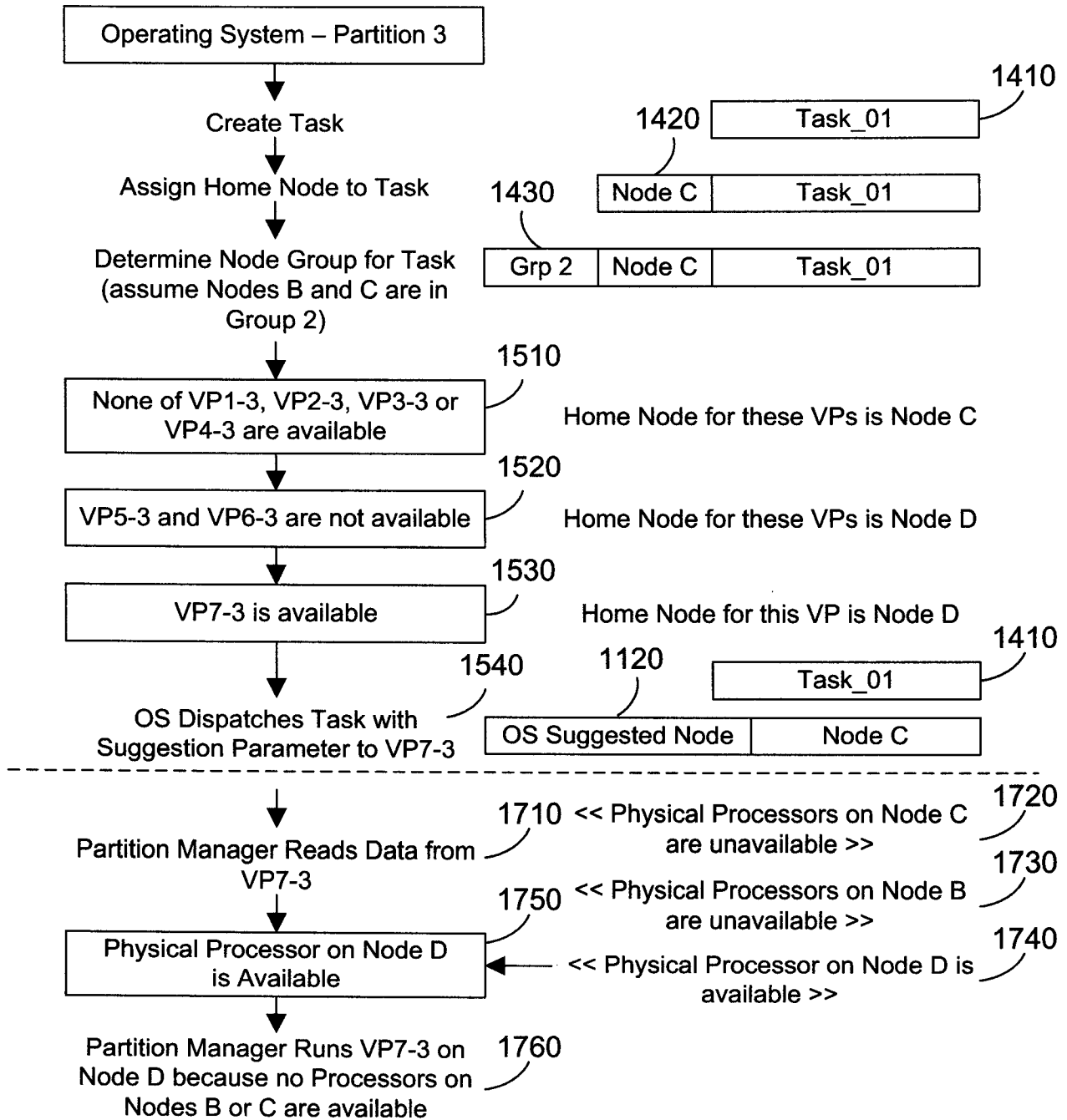


FIG. 17

INTERNATIONAL SEARCH REPORT

International application No

PCT/EP2010/051929

A. CLASSIFICATION OF SUBJECT MATTER
 INV. G06F9/50

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)

EP0-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2004/194098 A1 (CHUNG LYNN KEAT [US] ET AL) 30 September 2004 (2004-09-30) paragraph [0007] - paragraph [0010] paragraphs [0055] - [0061]	1-13
Y	Hales C et al.: "PowerVM Virtualization on IBM System p: Introduction and Configuration, Fourth Edition"[Online] May 2008 (2008-05), pages I-66, XP002586584 Redbooks Retrieved from the Internet: URL: http://www.redbooks.ibm.com/redbooks/pdfs/sg248000.pdf [retrieved on 2010-06-09] pages 5-8 pages 30-33 pages 42,43	1-13

☒ 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 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

10 June 2010

Date of mailing of the international search report

12/08/2010

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
 NL - 2280 HV Rijswijk
 Tel. (+31-70) 340-2040,
 Fax: (+31-70) 340-3016

Authorized officer

de Man, Ronald

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2010/051929

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	EP 1 674 987 A2 (MICROSOFT CORP [US]) 28 June 2006 (2006-06-28) abstract paragraphs [0051], [0052] -----	1-13

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/EP2010/051929

Patent document cited in search report		Publication date	Patent family member(s)		Publication date
US 2004194098	A1	30-09-2004	US 2008022286	A1	24-01-2008
EP 1674987	A2	28-06-2006	CN 1794177	A	28-06-2006
			JP 2006178933	A	06-07-2006
			KR 20060071307	A	26-06-2006
			US 2006136653	A1	22-06-2006