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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
26 May 2006 (26.05.2006)

PCT

(10) International Publication Number
WO 2006/055590 A1

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

(21) International Application Number:
PCT/US2005/041447

(22) International Filing Date:
15 November 2005 (15.11.2005)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

60/628,420	15 November 2004 (15.11.2004)	US
11/271,681	10 November 2005 (10.11.2005)	US

(71) Applicant *(for all designated States except US)*: **PLATFORM SOLUTIONS, INC.** [US/US]; 501 Macara Avenue, Sunnyvale, CA 94085 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **JUNG-IK, Lee** [KR/US]; 2892 Coventry Lane Sw, #2511, Tumwater, WA 98512 (US). **KETCHAM, Daniel G.** [US/US]; 20711 Dunbar Drive, Cupertino, CA 95014 (US). **HILTON, Ronald N.** [US/US]; 20861 Fargo Drive, Cupertino, CA 95014 (US).

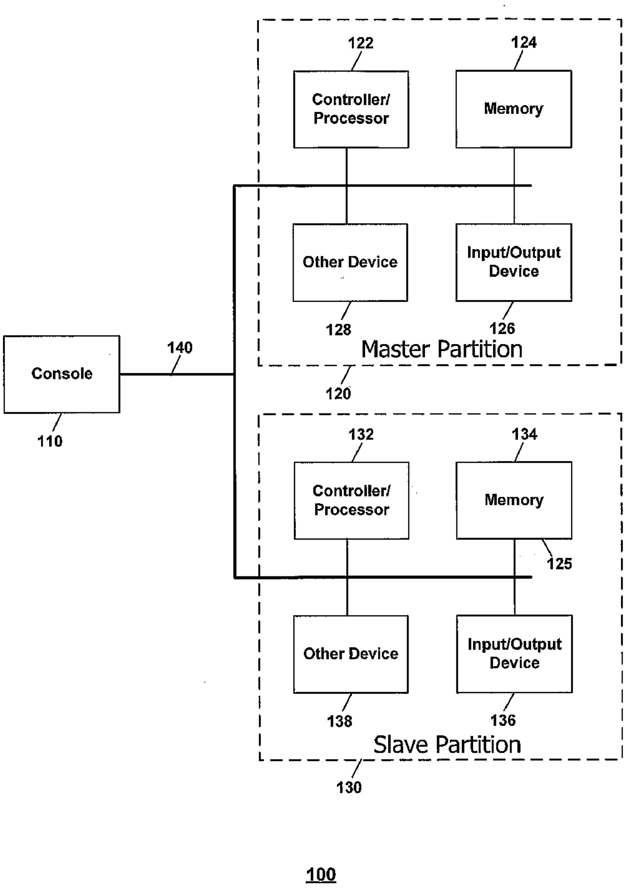
(74) Agents: **NEAL, Stephen, T.** et al.; KENYON & KENYON, 333 W. San Carlos St., Suite 600, San Jose, CA 95110 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(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),

[Continued on next page]

(54) Title: PEER-BASED PARTITIONING METHOD FOR SYSTEM RESOURCE SHARING



(57) **Abstract:** A method and system for partitioning a computer system into multiple virtual machines is disclosed. The system may include multiple logical partitions, each with their own set of physical resources controlled by a partition processor. The system may also include an external processor that maintains a map of what resources belong to what partitions. The logical partitions within the system may maintain a peer-to-peer relationship.

European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

PEER-BASED PARTITIONING METHOD FOR SYSTEM RESOURCE SHARING

BACKGROUND

[0001] The present invention relates to the virtual memory of a computer system and, in particular, to a peer-based partitioning method for system resource sharing.

[0002] The physical resources of a system can be divided into multiple logical partitions where each logical subset of the physical machine acts as a virtual machine. Partitioning of a computer system into multiple virtual machines is a well-established concept that was originally developed in an era when enterprise-class computer hardware was very expensive. Therefore, taking physical/logical resources and virtualizing them (i.e. creating separate virtual machines) were desirable so that the expensive hardware could be shared.

[0003] A layer of software, called a "hypervisor," controls the interaction between the various virtual machines. Any time a virtual machine wants to access a shared, physical resource, it does so through the hypervisor. The hypervisor acts as a master-control program. Anytime a logical partition wants access to a physical resource, it must go through the hypervisor. The hypervisor has to monitor and be actively involved with each logical partition. While significant progress has been made in both software and hardware technology for reducing the overhead of partitioning, a hypervisor overhead of 10-15% is still typical. Reducing the 10-15% overhead is desirable.

SUMMARY OF THE INVENTION

A method and system for partitioning a computer system into multiple virtual machines is disclosed. The system may include multiple logical partitions, each with their own set of physical resources controlled by a partition processor. The system may also include an external processor that maintains a map of what resources belong to what partitions. The logical partitions within the system may maintain a peer-to-peer relationship.

BRIEF DESCRIPTION OF THE DRAWINGS

[0004] **Figure 1** is a diagram of a possible configuration of a computer system to execute the present invention.

[0005] **Figure 2** is a diagram of one embodiment of a partition resource map.

[0006] **Figure 3** is a flowchart one embodiment of a method of logically partitioning a computer system into multiple virtual machines.

DETAILED DESCRIPTION

[0007] A method and system for logically partitioning a computer system into multiple virtual machines is disclosed. A set of physical resources of a computer system are partitioned into a set of logical partitions. A logical partition may contain one or more physical processors. One of the processors may be a controller of the partition. Each partition processor may then control its logical partition. The partitions may have a master slave configuration.

[0008] **Figure 1** illustrates a possible configuration of a computer system 100 to execute the present invention. The physical resources of the computer system 100 may be separated into a set of logical partitions. The computer system 100 may have an external processor called the console 110 to maintain a map of resources to the logical partitions. A partition controller may act as a console 110 or may control multiple partitions. One embodiment may have a master partition 120/slave partition 130 configuration. The master partition 120 and slave partition 130 may each have a set of physical resources, such as processors 122 and 132, memories 124 and 134, I/O cards/ports 126 and 136, and other devices 128 and 138. Partitions may not be configured symmetrically. For example, one partition may not have any I/O ports of its own, using another partition.

[0009] In a peer based system, each partition has its own domain of physical resources, accesses the resources directly, and communicates directly with other partitions rather than going through a hypervisor. The console 110 takes those physical resources and maps them to different logical partitions. The console 110 may be involved when a partition needs to access the physical resources of another partition. As long as system resources are not being redistributed from one partition to another, the console 110 need not be involved in the actual tasks being performed. Each partition may monitor itself.

[0010] **Figure 2** illustrates in a block diagram one embodiment of a partition resource map stored in memory. A console 110 may use a partition resource map 210 to map physical components in a partitioned environment, and may use a partition resource map 210 to provide system management and configuration functionality. A partition resource map 210 may be used to map physical components such as processors 212, memory pages 214, and I/O ports 216 among the various partitions by listing a logical and physical partition for each physical resource. A partition resource map may further be used for monitoring the reliability, availability, serviceability, and configuration functionality (RASC functionality) of the system's physical resources. For example, in one embodiment of the present invention, a partition resource map may be used to monitor the characteristics of the system's physical resources, and therefore, certain physical resources may not be used because they are off-line, unplugged, or are not functioning correctly.

[0011] In the embodiment presented in Figure 2, a partition resource map 210 may contain information relating to the physical location of a resource (PHYS), the virtual configuration of the resource (LOGL), the partition that a physical resource is associated with (PART), and additional information (STAT). Additional information may include physical status, logical status, partition status, or RASC functionality. Control of these resources may then be passed to the partition processors themselves within each partition. A console 110 may allow a first partition processor within a first partition to access a physical component in a second partition. Or, when a partition processor within one partition wants to access the physical resources of another partition, it can do so through the console 110, which acts as a conduit. This approach may be used with other forms of partitioning as well, such as hard partitioning and soft partitioning, to provide a whole range of options in terms of performance, granularity, and isolation. Soft partitioning may be either software based, firmware based, or kernel based. Furthermore, the peer-based approach need not preclude the selective sharing of resources between peers.

[0012] **Figure 3** illustrates in a flowchart one embodiment 300 of a method of partitioning a computer system into multiple virtual machines. The console 110 may partition a physical resource set (PRS) into logical partitions (Block 310). The console 110 may map a partition processor to each logical partition, so that a different partition processor is mapped to each logical partition (Block 320). Each partition processor would control its logical partition (Block 330). A partition processor may be mapped to control more than one partition. The console 110 may share memory between the logical partitions to implement high-speed communication between the logical partitions (Block 340). The console 110 may selectively share resources between the logical partitions (Block 350). One logical partition may virtualize an I/O subsystem to be used by other logical partition (Block 360).

[0013] Several embodiments of the present invention are specifically illustrated and described herein. However, it will be appreciated that modifications and variations of the present invention are covered by the above teachings and within the purview of the appended claims without departing from the spirit and intended scope of the invention.

WHAT IS CLAIMED IS:

1. A method comprising:

partitioning a set of physical resources of a computer system into a set of logical partitions;

mapping a first partition processor to a first logical partition of the set of logical partitions;

controlling the first logical partition with the first partition processor;

mapping a second partition processor to a second logical partition of the set of logical partitions; and

controlling the second logical partition with the second partition processor.

2. The method of claim 1, wherein the first logical partition and the second logical partition have a master/slave configuration.

3. The method of claim 1, further comprising sharing memory to implement high-speed communication between the first logical partition and the second logical partition.

4. The method of claim 1 further comprising selectively sharing resources between the first logical partition and the second logical partition.

5. The method of claim 1, further comprising virtualizing an I/O subsystem with the first logical partition for use by the second logical partition.

6. The method of claim 1 wherein the partitioning is hard partitioning.

7. The method of claim 1 wherein the partitioning is soft partitioning.

8. The method of claim 1, further comprising maintaining information related to a status of a physical resource of the set of physical resources.

9. The method of claim 1, further comprising maintaining the partitioning and mapping information in a partition resource map.

10. A set of instructions residing in a storage medium, said set of instructions capable of being executed by a storage controller to implement a method for partitioning a computer system into more than one virtual machine, the method comprising:

partitioning a set of physical resources of a computer system into a set of logical partitions;

mapping a first partition processor to a first logical partition of the set of logical partitions;

controlling the first logical partition with the first partition processor;

mapping a second partition processor to a second logical partition of the set of logical partitions; and

controlling the second logical partition with the second partition processor.

11. The method of claim 10, wherein the first logical partition and the second logical partition have a master/slave configuration.

12. The method of claim 10, further comprising sharing memory to be used to implement high-speed communication between the first and the second logical partition.

13. The method of claim 10, further comprising selectively sharing resources between the first logical partition and the second logical partition.

14. The method of claim 10, further comprising virtualizing an I/O subsystem with the first logical partition for use by the second logical partition.

15. The method of claim 10, further comprising maintaining the partitioning and mapping information in a partition resource map.

16. A computer system comprising:

a first logical partition of a set of physical resources controlled by a first partition processor by maintaining a partition resource map listing physical resources associated with the first logical partition; and

a second logical partition of the set of physical resources controlled by a second partition processor.

17. The system of claim 16, wherein the first logical partition and the second logical partition have a master/slave configuration.

18. The system of claim 16, further comprising resources selectively shared between the first logical partition and the second logical partition.

19. The system of claim 16, wherein the first logical partition virtualizes an I/O subsystem for use by the second logical partition.

20. The system of claim 16, wherein the partition resource map maintains information related to a status of a physical resource of the set of physical resources.

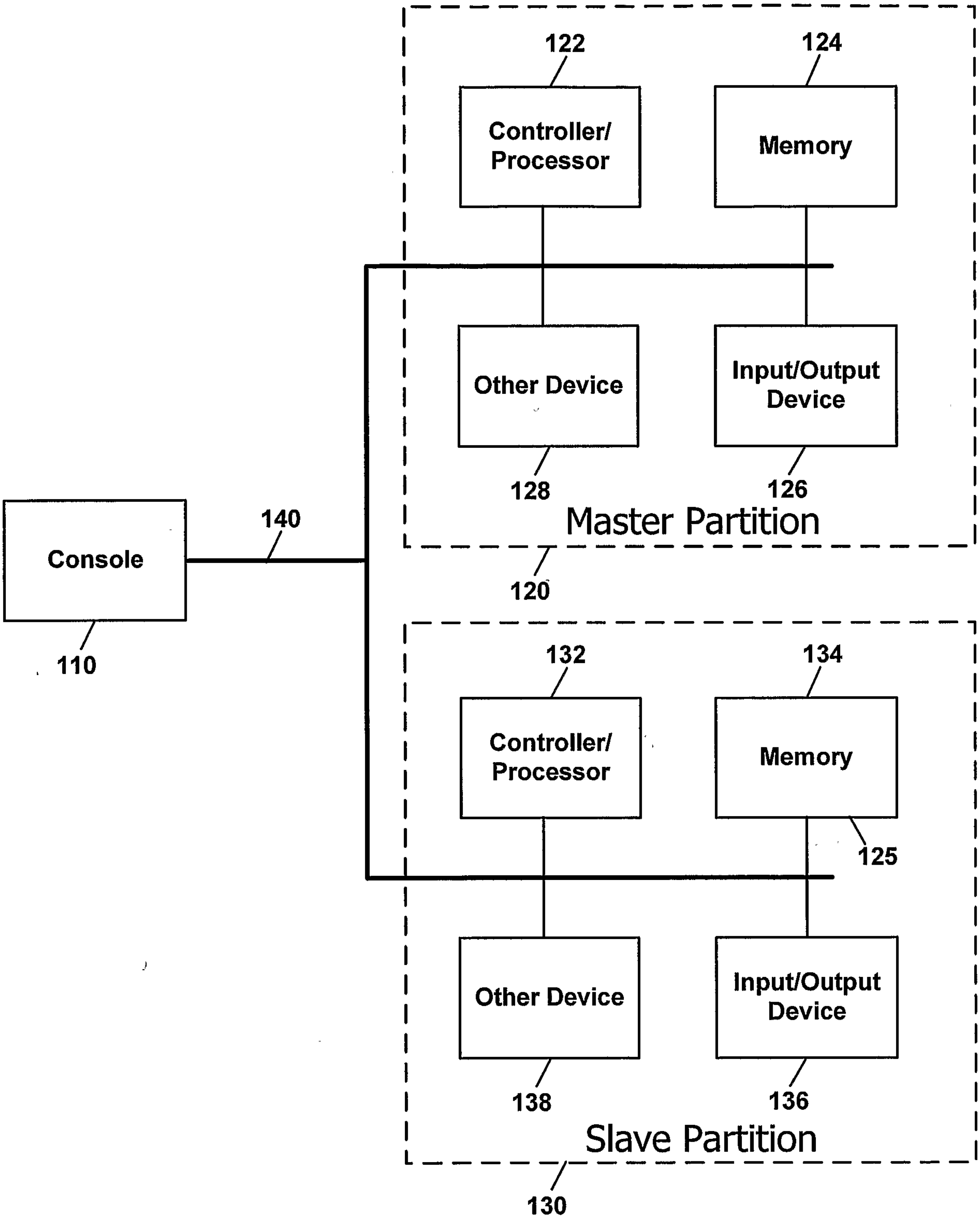
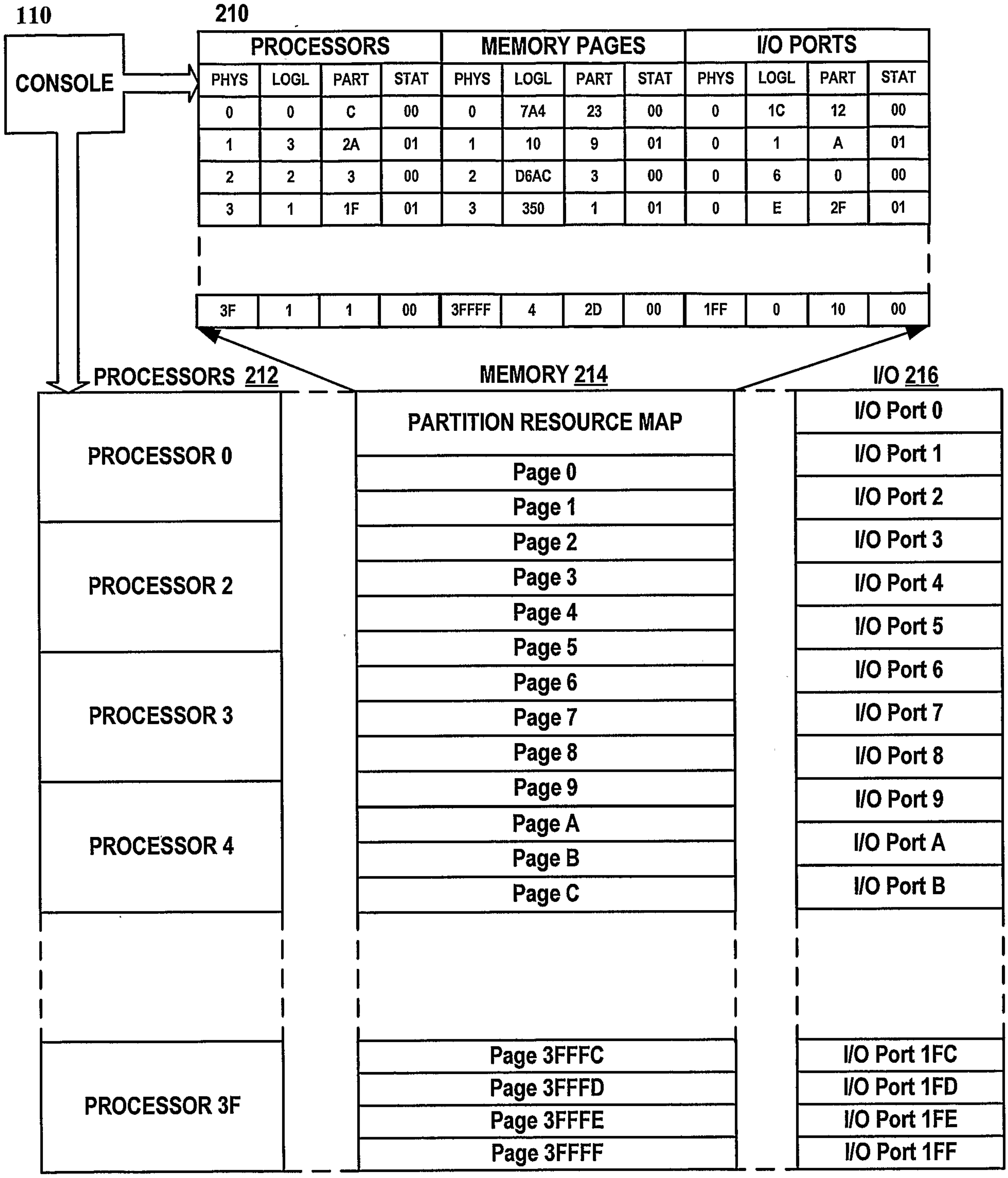

122
124
Controller/ Processor
Memory
Other Device
Input/Output Device
128
126
Master Partition
140
Console
120
110
132
134
Controller/ Processor
Memory
125
Other Device
Input/Output Device
138
136
Slave Partition
130

Figure 1
100



PROCESSORS				MEMORY PAGES				I/O PORTS			
PHYS	LOGL	PART	STAT	PHYS	LOGL	PART	STAT	PHYS	LOGL	PART	STAT
0	0	C	00	0	7A4	23	00	0	1C	12	00
1	3	2A	01	1	10	9	01	0	1	A	01
2	2	3	00	2	D6AC	3	00	0	6	0	00
3	1	1F	01	3	350	1	01	0	E	2F	01
3F	1	1	00	3FFFF	4	2D	00	1FF	0	10	00

Figure 2
200

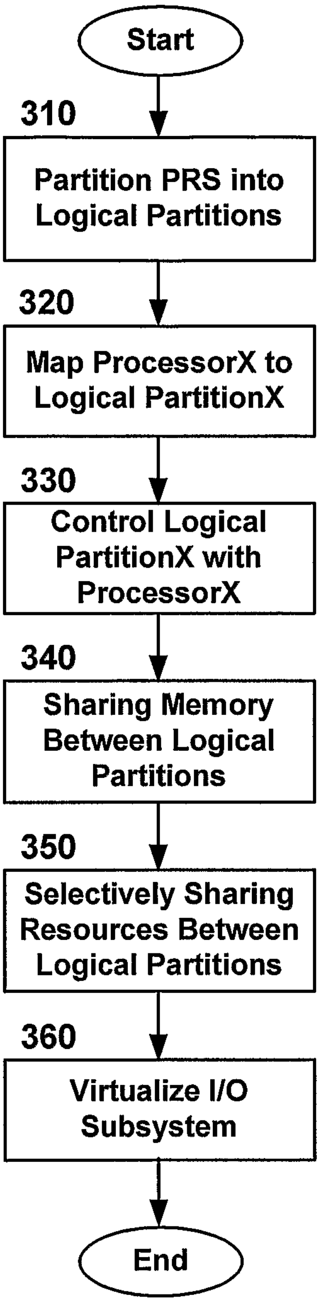

Start
310
Partition PRS into Logical Partitions
320
Map ProcessorX to Logical PartitionX
330
Control Logical PartitionX with ProcessorX
340
Sharing Memory Between Logical Partitions
350
Selectively Sharing Resources Between Logical Partitions
360
Virtualize I/O Subsystem
End

Figure 3
300

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2005/041447

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)

EPO-Internal, PAJ, IBM-TDB, INSPEC, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 02/086698 A (INTERNATIONAL BUSINESS MACHINES CORPORATION) 31 October 2002 (2002-10-31) abstract page 4, line 1 - line 27 page 6, line 1 - page 7, line 12 page 11, line 14 - page 20, line 18; claims 1-12; figure 4 -----	1-20
X	US 6 587 938 B1 (EILERT CATHERINE K ET AL) 1 July 2003 (2003-07-01) abstract column 2, line 23 - line 55 column 4, line 50 - column 25, line 65 ----- -/--	1-20

☒ 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 7 April 2006	Date of mailing of the international search report 25/04/2006
Name and mailing address of the ISA/ European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31–70) 340–2040, Tx. 31 651 epo nl, Fax: (+31–70) 340–3016	Authorized officer Wierzejewski, P

4

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2005/041447

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2004/168170 A1 (MILLER MICAH WILLIAM) 26 August 2004 (2004-08-26) abstract paragraph [0014] - paragraph [0017] paragraph [0026] paragraph [0036] - paragraph [0070]; claims 1-20 -----	1-20
X	US 2004/181625 A1 (ARMSTRONG TROY DAVID ET AL) 16 September 2004 (2004-09-16) abstract paragraph [0009] - paragraph [0010]; figures 1-8 paragraph [0021] - paragraph [0068] -----	1-20
A	"Partitioning for the IBM eserver pSeries 690 System" IBM WHITE PAPER, 2001, page complete12, XP002288349 the whole document -----	1-20
A	EP 0 301 275 A (INTERNATIONAL BUSINESS MACHINES CORPORATION) 1 February 1989 (1989-02-01) abstract claims 1-25 -----	1-20

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2005/041447

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 02086698 A	31-10-2002	EP 1390839 A1	25-02-2004
		US 2002156824 A1	24-10-2002
US 6587938 B1	01-07-2003	JP 3672236 B2	20-07-2005
		JP 2001142858 A	25-05-2001
US 2004168170 A1	26-08-2004	CN 1523498 A	25-08-2004
		JP 2004252988 A	09-09-2004
US 2004181625 A1	16-09-2004	BR PI0408310 A	07-03-2006
		CA 2515450 A1	23-09-2004
		CN 1723440 A	18-01-2006
		EP 1604279 A2	14-12-2005
		WO 2004081699 A2	23-09-2004
EP 0301275 A	01-02-1989	AU 606187 B2	31-01-1991
		AU 2001488 A	02-02-1989
		BR 8803742 A	14-02-1989
		CA 1305799 C	28-07-1992
		DE 3850181 D1	21-07-1994
		DE 3850181 T2	12-01-1995
		JP 1037636 A	08-02-1989
		JP 1945910 C	23-06-1995
		JP 6073108 B	14-09-1994
		US 4843541 A	27-06-1989