



(86) Date de dépôt PCT/PCT Filing Date: 2007/08/07
(87) Date publication PCT/PCT Publication Date: 2008/02/14
(85) Entrée phase nationale/National Entry: 2009/02/06
(86) N° demande PCT/PCT Application No.: US 2007/075407
(87) N° publication PCT/PCT Publication No.: 2008/019380
(30) Priorités/Priorities: 2006/08/07 (US60/821,673);
2007/08/07 (US11/835,307)

(51) Cl.Int./Int.Cl. *G06F 9/455* (2006.01)
(71) Demandeur/Applicant:
BEA SYSTEMS, INC., US
(72) Inventeur/Inventor:
DAHLSTEDT, JOAKIM, SE
(74) Agent: GOWLING LAFLEUR HENDERSON LLP

(54) Titre : SYSTEME ET PROCEDE DE VIRTUALISATION DE MATERIEL DANS UN ENVIRONNEMENT DE MACHINE VIRTUELLE
(54) Title: SYSTEM AND METHOD FOR PROVIDING HARDWARE VIRTUALIZATION IN A VIRTUAL MACHINE ENVIRONMENT

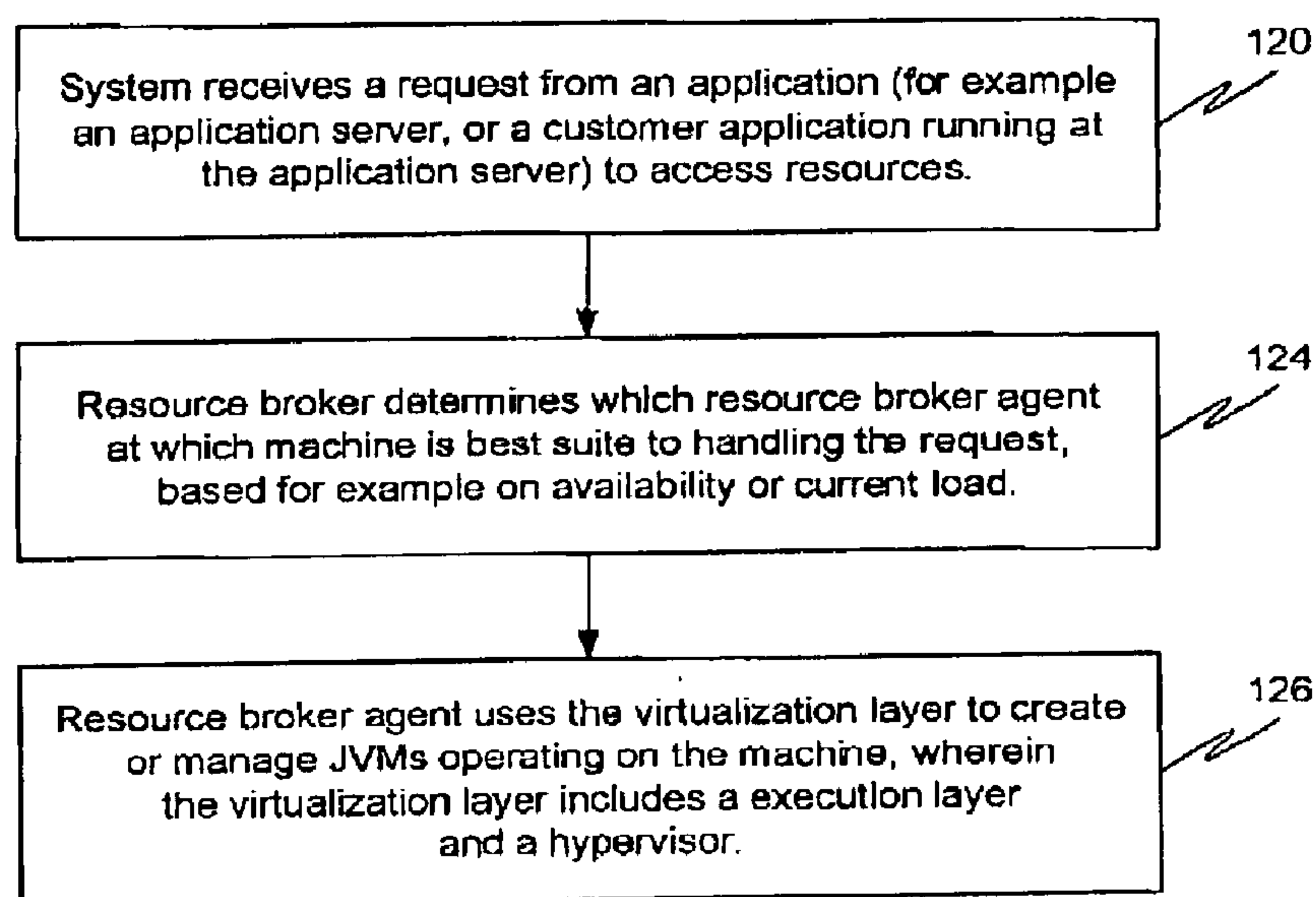


Figure 8

(57) **Abrégé/Abstract:**

A system and method for providing hardware virtualization and resource management in a virtual machine environment. In accordance with an embodiment, an application server environment includes a computer system, application server, and virtual machine (for example, a Java Virtual Machine or JVM). In accordance with an embodiment, a virtualization layer is provided at each physical machine, including a hypervisor for partitioning virtual machines over the machine. An execution layer runs a single Java-based application, focusing on running that application as efficiently as possible. In accordance with another embodiment, the system comprises two main components: a lower-level, execution environment that replaces the usual operating system; and a higher-level, resource broker that is responsible for handing out new virtual machines to the above layers, and for monitoring the current resource usage of the running virtual machines.



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

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
14 February 2008 (14.02.2008)

PCT

(10) International Publication Number
WO 2008/019380 A3

(51) International Patent Classification:

G06F 9/455 (2006.01) **G06F 9/46** (2006.01)

(21) International Application Number:

PCT/US2007/075407

(22) International Filing Date: 7 August 2007 (07.08.2007)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

60/821,673 7 August 2006 (07.08.2006) US
11/835,307 7 August 2007 (07.08.2007) US

(71) Applicant (for all designated States except US): **BEA SYSTEMS, INC.** [US/US]; 2315 North First Street, San Jose, California 95131 (US).

(72) Inventor; and

(75) Inventor/Applicant (for US only): **DAHLSTEDT, Joakim** [SE/SE]; P.O. Box 4213, S-10265 Stockholm (SE).

(74) Agents: **MEYER, Sheldon, R.** et al.; Fliesler Meyer LLP, 650 California Street, Fourteenth Floor, San Francisco, California 94108 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM,

AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, 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, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, 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, 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, HU, IE, IS, IT, LT, LU, LV, MC, MT, 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

(88) Date of publication of the international search report:
18 December 2008

(54) Title: SYSTEM AND METHOD FOR PROVIDING HARDWARE VIRTUALIZATION IN A VIRTUAL MACHINE ENVIRONMENT

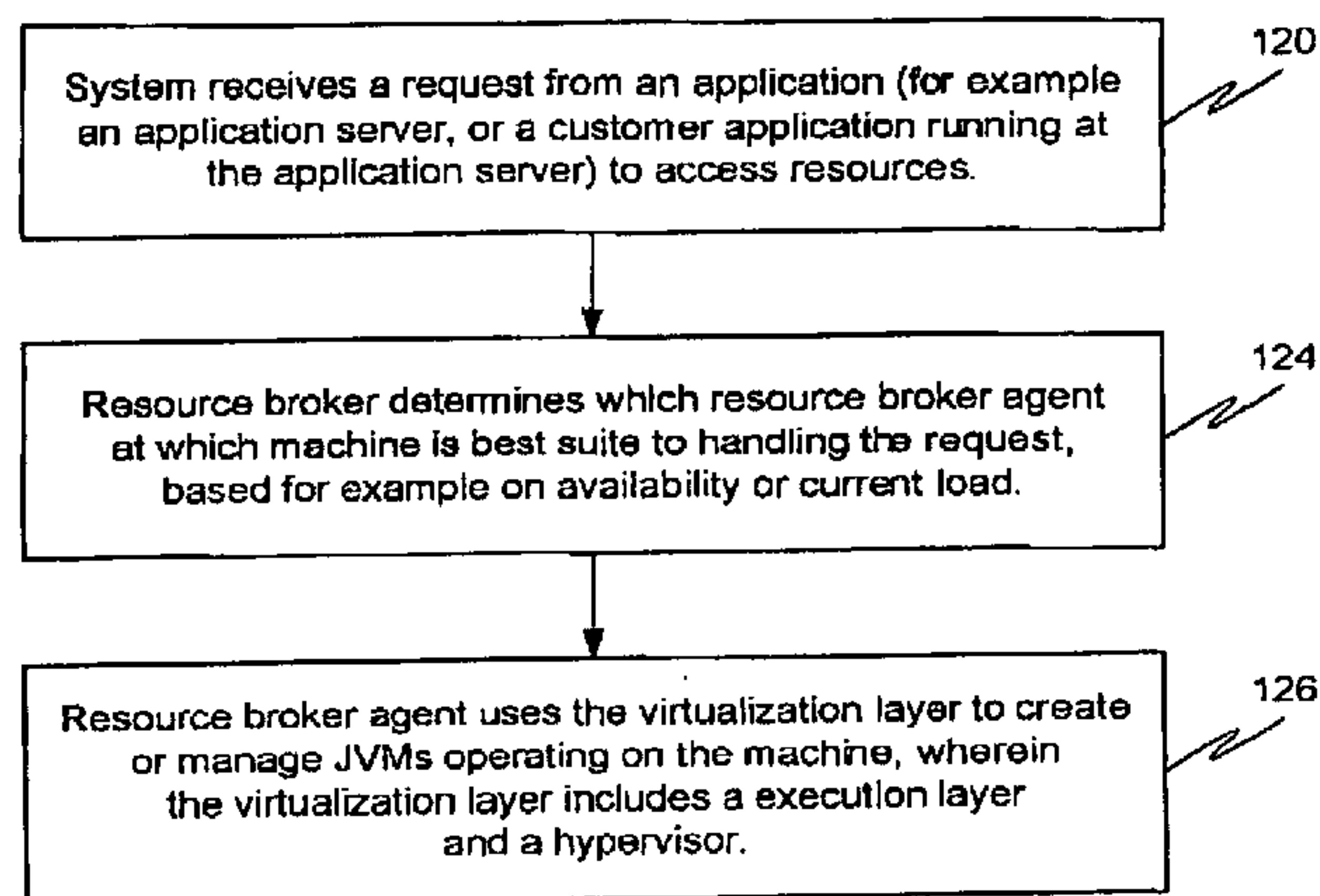


Figure 8

(57) Abstract: A system and method for providing hardware virtualization and resource management in a virtual machine environment. In accordance with an embodiment, an application server environment includes a computer system, application server, and virtual machine (for example, a Java Virtual Machine or JVM). In accordance with an embodiment, a virtualization layer is provided at each physical machine, including a hypervisor for partitioning virtual machines over the machine. An execution layer runs a single Java-based application, focusing on running that application as efficiently as possible. In accordance with another embodiment, the system comprises two main components: a lower-level, execution environment that replaces the usual operating system; and a higher-level, resource broker that is responsible for handing out new virtual machines to the above layers, and for monitoring the current resource usage of the running virtual machines.

WO 2008/019380 A3

SYSTEM AND METHOD FOR PROVIDING HARDWARE VIRTUALIZATION IN A VIRTUAL MACHINE ENVIRONMENT

5 Inventor: Joakim Dahlstedt

Copyright Notice

10 A portion of the disclosure of this patent document contains material which is subject to copyright protection. The copyright owner has no objection to the facsimile reproduction by anyone of the patent document or the patent disclosure, as it appears in the Patent and Trademark Office patent file or records, but otherwise reserves all
15 copyright rights whatsoever.

Claim of Priority:

[0001] This application claims the benefit of provisional patent application "SYSTEM AND METHOD FOR PROVIDING HARDWARE VIRTUALIZATION IN A
20 VIRTUAL MACHINE ENVIRONMENT", U.S. Application No. 60/821,673, filed August 7, 2006, and patent application "SYSTEM AND METHOD FOR PROVIDING HARDWARE VIRTUALIZATION IN A VIRTUAL MACHINE ENVIRONMENT", U.S. Application No. 11/835,307, filed August 7, 2007 which are incorporated herein by reference.

Field of the Invention:

[0002] The invention relates generally to operating systems, application servers, and virtual machines, and more particularly to a system and method for providing hardware virtualization and resource management in a virtual machine environment.

Background:

[0003] Many of today's mission-critical business applications are developed and deployed on Java-based application servers, (including, for example, application servers that are compliant with the Java 2 Enterprise Edition or J2EE specification). These applications have traditionally been deployed on single-vendor RISC computers and
35 servers, and other expensive hardware platforms. The readily available and lower-cost PC-based servers had not been considered a suitable alternative because of the absence of a competitive, high performing Java Virtual Machine (JVM) that could be

used with these architectures. The latest virtual machine environments, such as the JRockit product from BEA Systems, Inc., tackle this problem in that these virtual machines can be optimized for use in a variety of environments, including PC-based ones, enabling Java applications to run with increased reliability and performance on
5 lower cost, standards-based platforms. The JRockit virtual machine is an example of a virtual machine that is particularly well-designed to power demanding server-side Java applications, delivering superior performance, manageability, and reliability.

[0004] However, while current virtual machine product offerings are adept at providing software virtualization, little has been done to date to support virtualization at
10 the underlying hardware or operating system level, or to manage resources at the virtual machine level, while still allowing the application server to control or to participate in the virtualization process.

Summary:

[0005] Disclosed herein is a system and method for providing hardware
15 virtualization
and resource management in a virtual machine environment. In accordance with an embodiment, an application server environment that includes a computer system, application server, and virtual machine (for example, a Java Virtual Machine or JVM) is
20 extended to provide an interface from the higher layers in the application server environment to the system's actual processing power, such that the allocation of processing can be made in a machine-independent way. A layer of software is placed between the virtual machine and the hardware that improves the efficiency of the virtualized Java execution system. In accordance with an embodiment, the system
25 comprises two main components: a first, lower-level, execution environment that replaces the usual operating system; and a second, higher-level, resource broker that is responsible for handing out new virtual machines to the above layers, and for monitoring the current resource usage of the running virtual machines. In this way, the underlying physical machine can be partitioned to support multiple JVMs. Additionally, goals can
30 be set within the system as to CPU and other resource allocation, and the system can allocate the JVMs appropriately.

Brief Description of the Figures:

[0006] Figure 1 shows an illustration of an application server environment that includes a virtual machine in accordance with an embodiment of the invention.

[0007] Figure 2 shows an illustration of a layered virtual machine architecture in accordance with an embodiment of the invention.

[0008] Figure 3 shows an illustration of a multiple layered domain environment in accordance with an embodiment of the invention.

[0009] Figure 4 shows an illustration of how an embodiment of the virtualization layer can be used in a utility computing environment in accordance with an embodiment of the invention.

[0010] Figure 5 shows an illustration of bare metal domains together with resource brokers in accordance with an embodiment of the invention.

[0011] Figure 6 shows an illustration of bare metal domains together with JVM proxies in accordance with an embodiment of the invention.

[0012] Figure 7 shows a logical diagram of the virtual machine environment components in accordance with an embodiment of the invention.

[0013] Figure 8 shows a flowchart of a method for using virtualization in accordance with an embodiment of the invention.

Detailed Description:

[0014] Disclosed herein is a system and method for providing hardware virtualization and resource management in a virtual machine environment. In accordance with an embodiment, an application server environment that includes a computer system, application server, and virtual machine (for example, a Java Virtual Machine or JVM) is extended to provide an interface from the higher layers in the application server environment to the system's actual processing power, such that the allocation of processing can be made in a machine-independent way. A layer of software is placed between the virtual machine and the hardware that improves the efficiency of the virtualized Java execution system. In accordance with an embodiment, the system comprises two main components: a first, lower-level, execution environment that replaces the usual operating system; and a second, higher-level, resource broker that is responsible for handing out new virtual machines to the above layers, and for

monitoring the current resource usage of the running virtual machines. In this way, the underlying physical machine can be partitioned to support multiple JVM's. Additionally, goals can be set within the system as to CPU and other resource allocation, and the system can allocate the JVMs appropriately.

5

Glossary

[0015] The following terms are used herein.

[0016] Domain: An operating system instance running on top of a hypervisor layer. Multiple domains can run on the same machine unaware of each other. A domain
10 may also be referred to herein as a Guest operating system (OS), Guest, or VM.

[0017] Bare Metal Domain: A domain running a virtualization layer, virtual machine, and an application in accordance with an embodiment of the invention. (For example, in one embodiment the bare metal domain comprises the bare metal operating system (OS) layer plus a JVM such as the JRockit JVM, plus an application). A bare
15 metal domain may also be referred to herein as a BM Domain.

[0018] Node: A physical machine or a physical computer in the network.

[0019] Grid Computing: A set of multiple nodes working together to solve a single computational problem.

[0020] Utility Computing: A system or environment which allows many
20 applications to be run efficiently over many physical machines. The utility computing environment allows high-level policies to be specified on the applications to help the system decide which particular application(s) to prioritize. The utility computing environment also allows a user to deploy new applications to the system; provide high-level policies and/or guarantees for each application; monitor the status the
25 applications and the machines; handle rapid increases in resource utilization for a specific application; provision system upgrades; and provide high availability.

[0021] **Figure 1** shows an illustration of an application server environment that includes a virtual machine, and which can be used with an embodiment of the invention.

30 As shown in Figure 1, the application server environment 10 includes an application server 14, (for example, the WebLogic server product developed by BEA Systems, Inc., or another type of application server), together with a virtual machine 16, (for example,

the JRockit JVM also developed by BEA Systems, Inc., or another type of virtual machine). Figure 1 also illustrates optional components that are often used in an application server environment, for example a development environment 18, user integration component 20, or process integration component 22. Examples of these latter optional components include the WebLogic Workshop, WebLogic Portal, and WebLogic Integration products respectively, each of which are also developed by BEA Systems, Inc. Additional components can be added to the environment as necessary.

[0022] Figure 2 shows a more detailed illustration of a virtual machine architecture in accordance with an embodiment. As shown in Figure 2, a typical virtual machine 16, such as the JRockit JVM, comprises a threads management component 18 to manage the running threads in the system; a memory management component 20 to take care of procedures such as garbage collection; a code generation component 22; and a Java model component 24. A typical virtual machine also provides external interfaces 28 for use in managing the virtual machine. Although Figure 2 represents the JRockit product, it will be evident that other virtual machines in addition to the JRockit implementation can be used within the spirit and scope of the invention.

[0023] To enable efficient application virtualization, a correspondingly efficient hardware resource virtualization is required. This makes it possible to provide functionality such as suspension, live provisioning, and live migration. In accordance with an embodiment of the invention, an extra layer of abstraction is added. This enables the virtual machine to be moved relatively closer to the underlying hardware.

[0024] Figure 3 shows an illustration of a multiple layered domain environment in accordance with an embodiment of the invention. As shown in Figure 3, the computing environment 30 comprises a physical computer hardware or a computer machine 32. The machine can be a physical server, computer machine, network device, or an equivalent type of computing or processing device. The computing environment includes a virtualization layer 34 that sits logically on top of the processing device in the place of the typical operating system. In accordance with an embodiment, the virtualization layer comprises an execution layer component 38 (also referred to herein as the "bare metal layer"), and a hypervisor layer component 40 or hypervisor layer. Examples of currently available hypervisor layer components include the Xen hypervisor, and the VMWare hypervisor. Other hypervisor layers and hypervisor layer components

can be used in accordance with other embodiments.

[0025] The hypervisor layer is inserted between the server's hardware and the operating system. This provides an abstraction layer that allows each physical server to run one or more virtual servers, and effectively decouples the operating system and its applications from the underlying physical server. Once a virtual server image has been created it can then be run on any server. On the same principle, multiple virtual servers can simultaneously share a single physical server or machine. The net effect is to increase the overall server utilization, while at the same time receiving a resource guarantee that ensures that any particular application layer performance criteria is met.

[0026] As further illustrated in Figure 3, the virtual machine 16 is logically located above the virtualization layer. In accordance with an embodiment a resource broker 36, or a set of resource broker services, is then provided at a layer above the virtual machine. The resource broker is used by the application server 14, and by other applications, to use the services provided by the virtualization layer.

[0027] In some embodiments, the above-described system can then be used as part of an overall utility computing environment, by providing a hardware virtualization service for Java applications within a local network. When used in the context of a utility computing environment, services can be provided that allow a networked application to interface with the virtual machine. In a utility computing or multiple virtual machine environment, examples of services that can be provided include: the ability to create a new virtual machine or JVM somewhere within the networked environment; resource guarantees, such as minimum and maximum CPU, memory, or network bandwidth usages; resource metering and system thermostat functions; the ability to dynamically increase or decrease resources when an application needs more or less of those resources; features such as suspend to disk or resume from disk a running application (referred to herein as "freeze-dry"); the ability to clone a "freeze-dried" application (similar to a Unix fork-like functionality, i.e. instantiate the cloned application with new instance specific values); and the ability to move a running instance to another computer machine for planned system maintenance.

[0028] **Figure 4** shows an illustration of how an embodiment of the virtualization layer can be used within a utility computing environment 50. As shown in Figure 4, applications can use application virtualization features 52, including for example the

resource broker services or utility computing services described above, to access the underlying hardware virtualization layers 54 via an application server cloud 56. In accordance with an embodiment, the application server cloud can be a WebLogic server environment, which may in turn provide other advanced server features such as enterprise clustering.

[0029] Figure 5 shows an illustration of bare metal domains together with resource brokers in accordance with an embodiment of the invention. In accordance with an embodiment, the resource broker provides subservices to the application server and to the other software applications running thereon. These subservices can include for example: obtain a new JVM; provision the JVMs running on the server; meter the resources of a particular JVM; and control the resources of a JVM. In accordance with an embodiment, the main component for JVM creation is the resource broker 68 (as described above) which in this instance tracks which of the machines 60, 62 are available to create new virtualization layer instances on. In accordance with an embodiment, a resource broker agent (RBA) 64, 66 is located on each machine. In accordance with other embodiments, there need not be any one-to-one mapping between physical machine and agent, for example a single agent can cover multiple physical machines. The resource broker agent is responsible for the local creation of the JVM instance 70, 72, 74, i.e. for creation and management of JVM's operating on the same physical machine as the agent. All outside communication, from a first machine to a second machine, or throughout a cluster, travels first through the resource broker, which then directs the request to the responding resource broker agent. In this manner, the resource broker acts like a pool of agents, and keeps track of which particular agent controls which particular machine that in turn has the necessary available power to fulfill a request from an application.

[0030] In accordance with an embodiment, the resource broker is a general hardware abstraction layer. It can potentially be used with a large number of different machine types. If the machine type supports partitioning, then the resource broker can use that functionality to partition the machine as needed. Depending on which features a certain machine type supports, different pieces of the high-end functionality can then be made available. At a minimum the functionality should allow new JVMs to be created on the machine with approximate resource usage targets (for example at least limited

guarantees, limited sharing, no suspension and no migration).

[0031] It should be noted that the resource broker does not itself require the execution layer (i.e. the bare metal layer) for it to perform many of its operations, although when the resource broker is combined with the execution layer then the system will run better, and will support a greater range of features.

[0032] It should be noted that normal, i.e. non-virtualized machines can also be added to the resource broker, even though functionality may be reduced in these instances. In accordance with other embodiments, the resource broker can also include support for third-party systems such as Azul, Solaris Containers and other hypervisor-stacks. In accordance with an embodiment, Xen domains and VMWare guests are also able to support native bare metal resource guarantees, resource sharing, suspension, migration, hot-pluggability, and virtualized file systems.

JVM Proxy

[0033] **Figure 6** shows an illustration of bare metal domains together with optional JVM proxies in accordance with an embodiment of the invention. As described above, in some embodiments the resource broker services can act as a stand-alone entity. However, in these instances one will not obtain the same resource sharing and the same resource guarantees as offered by the combination of all of the elements. As shown in **Figure 6**, a physical server machine 80 can have many JVM's running thereon. In some instances JNI code execution and real device driver access are handled by the normal operating system (OS) 90. When it is run on top of a hypervisor layer 96, the execution layer 92, 94 is built to then optimize the execution of any JVMs 86, 88 operating thereon.

In particular the hypervisor layer / execution layer combination supports isolation and copy-on-write sharing of data to reduce the footprint of the multiple JVMs running on the same physical machine. In accordance with an embodiment a JVM proxy 82, 82 above the OS layer is provided to allow functions and services in the execution layer to be invoked via the OS. In accordance with other embodiments the bare metal execution layer can be used without the need for JVM proxies.

Performance Monitoring Framework - CPU Power and CPU Share

[0034] As described above, one of the useful features of the system is its ability to allocate available resources to respond to application requests. In accordance with an embodiment, CPU Power is an approximation of how much CPU time the application is consuming or is guaranteed to have. This information is then used to allocate JVM resources over the available machines or other physical hardware. In accordance with an embodiment, all units of measurements are considered estimates. The measurements are then used as relative measurements (i.e. 200 CPU Power units will be roughly twice as much computing power as 100 CPU Power units). It should also always be true that the number of units one physical machine says it can produce should be exactly the sum of all of the virtual machines current usage plus the idle units. On a particular machine the relative measurements are likely to work better, and to make more sense, than if considered across several different machines. In particular the more different two machines are the less exact the relative measurements will likely be. The goal is typically that the CPU Power should be comparable to the furthest extent possible between different machines.

System Interfaces and Functions

[0035] Table 1 lists some of the interfaces and functionality that are provided by the system in accordance with one embodiment. The functions allow a high-level application to create JVMs and allocate resources over several machines, including within a cluster. It will be evident that in other embodiments, additional interfaces and functions can be provided.

Creation Interface	
Request:	Response:
Give me a new JVM that follows this resource specification; and assign this instance a machine, but don't start it yet.	JVM-Handle.
Give me a new JVM; assign this instance a machine; use the specified image to initiate the instance; but don't start the instance yet.	JVM-Handle/
Remove this JVM.	OK / Fail.

Initiation Interface	
Request:	Response:
Mount the directory labeled X in your view; and mount it in (read-only, read-writable, exclusive)-mode.	OK / Fail.
Set the command-line to start with.	OK / Fail.
Start the JVM.	OK / Fail.
Suspend/Resume Interface	
Request:	Response:
Suspend this running JVM.	OK / Fail.
Store a suspend JVM image to a specified location. the image can then be used in a later call to create a new JVM.	OK / Fail.
Resume; resume a previously suspended JVM; any active network connections will be disconnected.	OK / Fail.
Resource Interface	
Request:	Response:
Report how much resources you are using.	Resources in use.
Request a change (increase or decrease) in available resources.	OK / Fail.

Table 1

[0036] In accordance with an embodiment, the system can request and measure any or all of four types of resources, which it can then use to allocate resources over the appropriate JVMs. These can include: a CPU Power measurement (as described above); Java Heap Memory; Network Bandwidth; and Disk I/O. Additional embodiments can provide additional types of resource measurement options.

[0037] When requesting CPU Power, the above layers can set a minimum (i.e. what must be guaranteed to be available) and a weight. For the other three resource types only a minimum can be specified. Potentially overcapacity is provided on a first-come first-served basis as guided by the weights. This design choice simplifies the optimization space fundamentally, and makes placement of servers simpler. The

underlying reason for this approach is because most applications tend to be processing or memory bound (i.e. memory bound as in waiting for a load, not as in needing more memory).

[0038] In accordance with an embodiment the memory heap is designed expand on demand to the needs of the application, so a maximum heap size should not generally be specified. An OutOfMemory error is thrown only when the system has run out of address space (which is completely unfixable) or run out of physical memory (which may be resolved by moving to a box with more memory, moving a VM on the same machine to another machine, or by reducing the memory in use by other VMS).

Benefits of using a Hypervisor

[0039] The benefits of using a hypervisor include that, typically, hypervisors are good at isolation, whereas operating system (OS) processes are not as good. Especially requirements on a minimum amount of CPU are hard to guarantee at the process level. Features such as Suspend and Resume Saving the complete state of the system is also difficult at any layer above the OS layer. The problem is that this is the only layer that has complete information of the state of the system (e.g. the JVM may be able to keep track of open files, mapped memory, threads in native code etc, but it is impossible for the JVM to completely restore the state upon Resume). The Live Migration feature is simpler: since the hypervisor abstracts away the hardware the system has the ability to migrate a running system to another box. By cooperation between the JVM-layer, across the OS-layer, to the hypervisor-layer this can be performed more efficiently. Using the hypervisor can also ensures that, for example, a certain amount of CPU time can be made available to the system, and that the system can meter how much CPU time each container is using.

Benefits of using a Special OS Layer

[0040] In some embodiments, a special operating system (OS) layer is used. Memory sharing can be performed at many levels (application, app. server, JVM, OS). But some types of sharing are more efficiently performed at certain layers of the stack. In particular, the combination of the support from the combination of the JVM and the hypervisor allows the system to share memory between different JVMs in an

inverse-fork-like manner. This means that the system can "after the fact" tell the hypervisor that certain pages are identical between different JVMs. The hypervisor can then let these pages be mapped to the same physical pages with a copy-on-write strategy.

5 **[0041]** In addition, a special OS-layer enables some functionality that was previously not possible, for example: the OS will only handle one process the JVM process that runs the application; and the OS understands the JVM and there is smart cooperation / communication between the JVM and the OS that enables powerful optimizations. Some features are impossible without OS-level support, for example:

- 10 • Pauseless compaction by disabling some pages for mutator threads while the garbage collector is compacting in those areas;
- Smart thread scheduling that for example instead of spinning on a lock schedules the thread that is holding that lock.

15 **[0042]** The special OS layer also creates some efficiencies through OS-layer removal. Removing the OS-layer leads to lower pause times, better behavior during high contention and low memory.

[0043] Adding a hypervisor adds an extra level of abstraction which also quite naturally slows down the system. By removing the operating system the system can
20 compensate for this slowdown and perhaps even improve over it. By using a hypervisor and having an ordinary operating system installed as one of the guest OSes, Bare Metal can avoid writing device drivers for different devices and just one simple one that talks through the hypervisor to the device drivers on the guest OS with the operating system.

25 **System Design**

[0044] In accordance with an embodiment, the system includes a Resource Management Service that includes a Java interface as its main interface. Users of the interface can include application servers, applications, and enterprise-wide applications, for example clustering applications. The actual Java APIs for one embodiment are
30 illustrated below as Appendix A.

[0045] **Figure 7** shows a logical diagram 100 of the virtual machine environment components in accordance with an embodiment of the invention. The virtualization layer

should ideally be as small as possible. As shown in Figure 7, in accordance with an embodiment this reduces to a thin layer of hypervisor (for example, Xen or VMWare, or another hypervisor), a thin layer of operating environment (for example, the bare metal execution layer), and a fatter blob of execution engine (for example, JRockit or another JVM), all running a very fat application (for example, the WebLogic Server or another application server, plus customer applications).

[0046] In some embodiments, it is assumed that the OS layer is built assuming a para-virtualized hypervisor (i.e. there is cooperation between the OS and the hypervisor, the OS is hypervisor-aware); however other embodiments may use cooperative OS and hypervisors.

[0047] **Figure 8** shows a flowchart of a method for using virtualization in accordance with an embodiment of the invention. As shown in Figure 8, in step 120, the system receives a request from an application (for example an application server, or a customer application running at the application server) to access resources, which must be provided by a JVM. In step 122, the resource broker determines which resource broker agent at which machine is best suited to handling the request, based for example on availability or current load. In step 124 the resource broker agent uses the virtualization layer to create or manage JVMs operating on the machine, wherein the virtualization layer includes an execution layer and a hypervisor.

Appendix A

[0048] The following code defines an actual Java API for use with an embodiment of the current invention. It will be evident that other API's may be defined in other embodiments within the spirit and scope of the invention.

Resource Broker API
package com.bea.jvmbroker;

public interface JVMBroker

{

/**

* Create a new JVM instance. The JVM instance is bound to a
* specific machine and has a UUID and can be populated hereafter.

* The JVM is not using any resource yet though

*/

```

public JVM createJVM(JVMRequest rc);

/**
 * Create a new JVM instance.
5  * @note This is a convenience method see createJVM for more information.
 */
public JVM createJVM();

/**
10  * Create a new JVM instance from an image. The image is a previously
 * suspended JVM. When started the JVM will be started in a suspended
 * state and a call to resume() will be required to resume execution.
 *
 * @note This is a convenience method see createJVM for more information.
15  */
public JVM createJVMFromImage(JVMRequest rc, File image);

}

20  /**
 * This interface represents a JVM handle from the JVM Resource Brokers
 * perspective.
 *
 * This is the interface used to configure a JVM. At the point you have
25  * one of these the JVM is bound to a machine, but it may not be started
 * yet.
 */
public interface JVM
{
30  /**
 * Get the IP addresses of the JVM. This is typically one IP address.
 *
 */
public InetAddress[] getAddresses();
35  /**
 * Set the startup command to the JVM.
 *
 * @param name - the name of the main class to start
40  * @param args - the arguments passed to the main method of the main class
 */
public void setStartCommand(String name, String[] args) throws IllegalStateException;

/**
45  * Mount the predefined mount point called path. Path must be a valid
 * predefined mount point on the machine in question. Path is a UNIX-like
 * path.

```

```
*
* i.e. mount("/wls") or mount("/wls90");
*/
public void mount(String path);
5
/**
 * Start the JVM. This method can only be called once.
 */
public void start() throws IllegalStateException;
10
/**
 * Suspends a running JVM.
 */
public void suspend();
15
/**
 * Resumes the execution of a suspended JVM.
 */
public void resume();
20
/**
 * Called to increase or decrease the agreed upon resource
 * usage for this JVM
 */
25 public boolean changeRequirements(JVMRequirements rc);

/**
 * Get an estimate of the current resource usage.
 *
30 * This info can be used to make clustering decisions at
 * a higher level.
 *
 * e.g. if you are using equal or more
 * power units than you requested, this application is
35 * probably power-unit-limited (similar to CPU/memory bound)
 * so it makes sense to request more PUs locally if possible.
 *
 * Alternatively, if the application is NOT using all it's power
 * units the bottleneck is probably elsewhere and requesting more
40 * power or starting more instances will not help.
 */
public JVMResources getResourceUsage();

/**
45 * Get the physical machine this JVM is placed on. Physical here
 * means what the Resource Broker thinks is a physical machine.
 * Some other layer may of course have fooled us into believing
```

```

    * that a virtual machine is a physical machine
    */
    public Machine getMachine();

5    /**
    * Will save the suspended image to the specified file (the file
    * is specified as a local path from the caller of storeImage.)
    * If the JVM is on another machine than the caller, this operation
    * may take numerous seconds to complete as the data will have to
10    * be moved from that server to the caller (or whatever fileserver
    * the caller is referring to)
    *
    * This method can only be called when the JVM is suspended.
    */
15    public void storeImage(File f) throws IllegalStateException;

    /**
    * Register a listener that is interested in JVM events
    */
20    public void registerListener(JVMListener jl);
}

/**
    * This is the interface for specifying lan-wide requirements on a JVM before
25    * it is started both in terms of how much resources it may use, as well
    * as more global requirements such as where it may be located in relation
    * to other JVMs
    *
    * For the JVM Local requirements also see JVMRequirements.
30    *
    * @see com.bea.jvmbroker.JVMRequirements
    * @author Joakim Dahlstedt (joakim.dahlstedt@bea.com)
    */
    public interface JVMRequest {
35        /**
        * Get the JVM resource requirements
        */
        public abstract JVMRequirements getJVMRequirements();

40        /**
        * Set the JVM resource requirements for this JVM
        */
        public abstract void setJVMRequirements(JVMRequirements r);

45        /**
        * Get the group of JVMs this JVM should belong to. A group is normally used
        * to inform the resource broker that this set of JVMs serve the same application

```

```
    */
    public abstract JVMGroup getJVMGroup();

    /**
5      * Set the group of JVMs this JVM should belong to. A group is normally used
      * to inform the resource broker that this set of JVMs serve the same application
      */
    public abstract void setJVMGroup(JVMGroup jg);

10    /**
      * Get the specific machine this JVM must run on
      */
    public abstract Machine getMachine();

15    /**
      * Require this JVM to run on a specific machine.
      *
      * This is not recommended behavior. It will strongly limit the
      * resource brokers abilities to schedule resource to take full
20    * advantage of the machines.
      *
      * Cases where setMachine may be used: when a special machine is the only
      * one that can access a certain service (say it's the only one that has
      * the connection to the CICS system or to a REUTERS terminal).
25    *
      * @note this is not supported functionality in the pilot
      * @note the idea is to support this in v1.0
      */
    public abstract void setMachine(Machine m);

30    /**
      * Get the auto start value
      */
    public abstract int getAutoRestart();

35    /**
      * Specify whether to restart if the JVM shutdown with a non-zero exit code
      *
      * The number specifies how many times to trying restart before giving up.
40    *
      * 0 = no auto restart (default)
      */
    public abstract void setAutoRestart(int ntimes);
45  }

    /**
```

```

* This class represent local resource constraints on a JVM. Local in this case
* local to the JVM itself, if you want to put requirements on the dependency
* between different JVMs or want a certain JVM to execute on a certain machine
* or something like that, you should look at JVMRequest.
5  *
* @see com.bea.jvmbroker.JVMRequest
*/
public class JVMRequirements {
    private static final int DEFAULT_WEIGHT = 100;
10  private static final long MIN_MEMORY = 16; // jrockit supports no less
    private static final long MIN_PU = 1;
    private static final long MIN_NBW = 50; // 50kBit ~ modem speed
    private static final long MIN_DS = 500; // 0.5GB
    /**
15  * The preferred constructor for JVMRequirements. Only specify how much
    * power you want. Power is not merely CPU-MHz, its actually bogo-mips used
    * for the application. This means JVM administration such as garbage
    * collection and compilation are not included in the measurement, and to
    * get more Power this system may for example give you more memory instead
20  * of more CPUs.
    */
    public JVMRequirements(long minPU);

    /**
25  * A simple constructor for JVMRequirements. Sets the default values
    * for all requirements.
    */
    public JVMRequirements();

30  public JVMRequirements(long minPU, long minMem, long minNB);

    public JVMRequirements(long minPU, long minMem, long minNB, long minDS);

35  }

public class JVMResources implements Serializable {
    /**
    * Return the number of CPU Power in use (this is an average computed over
    * the last 3 seconds of how many "java operations" this JVM was
40  * performing). The thinking is that a twice as fast machine should have the
    * capacity to perform twice as may CPU Power, it is unclear if this will
    * be the case.
    */
    public long getCPU Power() {
45  return CPU Power;
    }
}

```

```
/**
 * This is all the memory the JVM is using (including native memory and heap
 * memory)
 */
5 public long getMemory() {
    return memory;
}

/**
10 * Get the amount of local disk space currently used by this JVM
 */
public long getDiskUsage() {
    return diskUsage;
}

15 /**
 * Get the network bandwidth (this is an average computed over the last 3
 * seconds)
 */
20 public long getNetworkBandwidth() {
    return networkBandwidth;
}
}

25 public interface JVMLListener
{
    public void aboutToSuspend();

    public void resumed();
30

    /**
     * Called after a JVM has exited (shut down). The value passed in is the
     * exit code for the JVM (as returned by the process...
     * i.e. System.exit(value)
35 */
    public void exited(int exitCode);
}

public interface ResourceBroker
40 {
    /**
     * Will return an array of the machines associated
     * with the ResourceBroker.
     */
45 Machine[] getMachines();

    /**
```

Attorney Docket No.: BEAS-01936WO0

```

    * Will return an array of JVMs that the ResourceBroker
    * thinks are currently running within its "domain".
    */
JVM[] getJVMs();
5
/**
 * A request is made from external source to remove a machine
 * from the pool (probably to put in another pool.).
 */
10 public Machine relinquishMachine();

    public void addMachine(Machine m);
}

15 /**
 * This interface represents something the jvm resource broker believes to be
 * an actual physical machine. If there is a hypervisor installed on the
 * machine the machine can be divided into multiple virtual machines on
 * demand by the resource broker.
20 */
public interface Machine
{
    /**
    * Get the IP address of the machine in this LAN. All machines should
25 * have exactly one IP-address in this LAN. (Then they may of course
    * be running multiple domains on the box where each may have a local
    * IP of its own.)
    */
    public InetAddress getAddress();
30

    /**
    * Get a snapshot of how much of the machine resources currently
    * are in use. The machine resources are the big 4: CPU/Memory/
    * Disk/Network
35 */
    public ResourceUsage getResourceUsage();

    /**
    * Return the total amount of memory on the machine in MB.
40 */
    public long getTotalMemory();

    /**
    * Return the total amount of disk space on the machine in MB.
45 */
    public long getTotalDiskSpace();
}
```

```

/**
 * Check whether the current machine has the support of a hypervisor
 * (and hence consequently can partition the box into multiple boxes
5  * with resource guarantees, suspend/resume/migrate capabilities)
 */
public boolean hasHypervisor();
}

```

10 **[0049]** The present invention may be conveniently implemented using a conventional general purpose or a specialized digital computer or microprocessor programmed according to the teachings of the present disclosure. Appropriate software coding can readily be prepared by skilled programmers based on the teachings of the present disclosure, as will be apparent to those skilled in the software art.

15 **[0050]** In some embodiments, the present invention includes a computer program product which is a storage medium (media) having instructions stored thereon/in which can be used to program a computer to perform any of the processes of the present invention. The storage medium can include, but is not limited to, any type of disk including floppy disks, optical discs, DVD, CD-ROMs, microdrive, and magneto-optical
20 disks, ROMs, RAMs, EPROMs, EEPROMs, DRAMs, VRAMs, flash memory devices, magnetic or optical cards, nanosystems (including molecular memory ICs), or any type of media or device suitable for storing instructions and/or data.

[0051] The foregoing description of the present invention has been provided for the purposes of illustration and description. It is not intended to be exhaustive or to limit
25 the invention to the precise forms disclosed. Many modifications and variations will be apparent to the practitioner skilled in the art. Particularly, while the embodiments of the systems and methods described above are described in the context of a WebLogic application server, and the JRockit JVM product, it will be evident that the system and methods may be used with other types of application server, JVM and virtual machine
30 products. Similarly, while the embodiments described above are described in the context of the Xen hypervisor, other types of hypervisor can be used. The embodiments were chosen and described in order to best explain the principles of the invention and its practical application, thereby enabling others skilled in the art to understand the invention for various embodiments and with various modifications that are suited to the

particular use contemplated. It is intended that the scope of the invention be defined by the following claims and their equivalence.

Claims:

What is claimed is:

- 5 1. A system for providing hardware virtualization in a virtual machine environment, comprising:
 - a machine or computer, further comprising a virtualization layer, the virtualization layer including a hypervisor for partitioning virtual machines over the machine;
 - an execution layer adapted to run a single Java-based application; and
 - 10 a Java Virtual Machine that runs on the execution layer, for use in running Java-based applications.
2. The system of claim 1, further comprising a plurality of machines and a virtualization layer located at each machine.
- 15 3. The system of claim 1, wherein the execution layer is adapted to focus on running only the single Java-based application at a particular time, without being able to run any other applications.
- 20 4. The system of claim 3, wherein the execution layer is adapted to run the single Java-based application as efficiently as possible.
5. The system of claim 1, further comprising a plurality of execution layers, each of which execution layer is adapted to run only a single Java-based application at a time,
- 25 as efficiently as possible.
6. The system of claim 5, wherein each of the plurality of execution layers is adapted to focus on the single Java-based application, without being able to run any other applications.
- 30 7. The system of claim 5 wherein a plurality of execution layers are located on the same combination of hypervisor, virtualization layer, and machine.

8. A system for providing hardware virtualization in a virtual machine environment, comprising:

a machine including a virtualization layer, the virtualization layer including a
5 hypervisor for partitioning virtual machines over the physical machine;

a plurality of Java Virtual Machines that run on the machine; and

a plurality of execution layers, including a separate execution layer for each of the
plurality of Java Virtual Machines, wherein each of the execution layers provides for
efficient operation of its own Java Virtual Machine, and a Java application running
10 thereon.

9. The system of claim 8 wherein each of the execution layers provides for efficient
operation of its own Java Virtual Machine, and of a single Java application running
thereon, without being able to run any other applications.

10. The system of claim 8 wherein each execution layer is provided as an operating
system (OS) that only handles, as a single process, the JVM process that runs the Java
application.

11. The system of claim 8 wherein the hypervisor allows memory to be shared
between two or more virtualization layers, and the Java Virtual Machines running
thereon.

12. The system of claim 8 wherein the execution layer for each Java Virtual Machine
25 informs the hypervisor layer of the presence of identical pages shared between different
Java Virtual Machines.

13. The system of claim 12 wherein the hypervisor allows the identical pages to be
mapped to the same physical pages using copy-on-write.

14. The system of claim 8 wherein the hypervisor is a Xen hypervisor.

15. The method of claim 8 wherein the hypervisor is a VMware hypervisor.

16. A system for providing hardware virtualization in a virtual machine environment, comprising:

5 one or a plurality of machines or computers, further comprising a plurality of virtual machines operating on the one or a plurality of machines;

a virtualization layer located at each machine, wherein the virtualization layer includes a hypervisor for partitioning virtual machines over that machine; and

a resource broker service that

10 receives a request from an application to access resources to be provided by a virtual machine,

determines which of the plurality of virtual machines is best suited to handling the request, and

15 communicates the request to the chosen virtual machine via the virtualization layer.

17. The system of claim 16 wherein the system further comprises a plurality of resource broker agents that are responsible for communicating requests to the best suite virtual machine.

20

18. The system of claim 16 wherein the virtualization layer and the resource broker service allocates requests according to predefined guarantees for the application making the request.

25 19. The system of claim 16 wherein the hypervisor is a Xen hypervisor.

20. The system of claim 16 wherein the hypervisor is a VMware hypervisor.

21. A system for providing hardware virtualization in a virtual machine environment, comprising:

30

a plurality of virtual machines operating on one or more physical machines or computers;

a virtualization layer at each machine, including a hypervisor and an execution layer for partitioning virtual machines over the physical machine; and

a resource broker service that receives a request from an application to access resources to be provided by a virtual machine, determines which of the plurality of virtual machines is best suited to handling the request, and communicates the request to the chosen virtual machine via the virtualization layer.

22. The system of claim 21 wherein the system further comprises a plurality of resource broker agents including an agent that is responsible for communicating requests to the best suited virtual machine.

23. The system of claim 21 wherein the virtualization layer and the resource broker service allocates requests according to predefined guarantees for the application making the request.

24. The system of claim 21 wherein the hypervisor is a Xen hypervisor.

25. The system of claim 21 wherein the hypervisor is a VMware hypervisor.

26. The system of claim 21 wherein the system further comprises a virtual machine proxy that allows an application to tunnel requests through the operating system at a first physical machine to the virtualization layer and a virtual machine at a second physical machine.

27. A method for providing hardware virtualization in a virtual machine environment, comprising the steps of:

providing a machine or computer, further comprising a virtualization layer, the virtualization layer including a hypervisor for partitioning virtual machines over the machine;

providing an execution layer adapted to run a single Java-based application; and providing a Java Virtual Machine that runs on the execution layer, for use in running Java-based applications.

28. The method of claim 27, further comprising providing a plurality of machines and a virtualization layer located at each machine.

5 29. The method of claim 27, wherein the execution layer is adapted to focus on running only the single Java-based application at a particular time, without being able to run any other applications.

30. The method of claim 29, wherein the execution layer is adapted to run the single
10 Java-based application as efficiently as possible.

31. A method for providing hardware virtualization in a virtual machine environment, comprising the steps of:

15 receiving a request from an application to access resources to be provided by a virtual machine;

determining which of a plurality of virtual machines or a plurality of physical machines or computers is best suited to handling the request;

20 communicating the request using a virtualization layer to create or manage virtual machines operating on the chosen physical machine or virtual machine, wherein the virtualization layer includes a execution layer and a hypervisor; and

processing the request at the chosen virtual machine.

32. The method of claim 31 wherein the method further comprises providing a plurality of resource broker agents including an agent that is responsible for
25 communicating requests to the best suited virtual machine.

33. The method of claim 31 wherein the virtualization layer and the resource broker service allocates requests according to predefined guarantees for the application making the request.

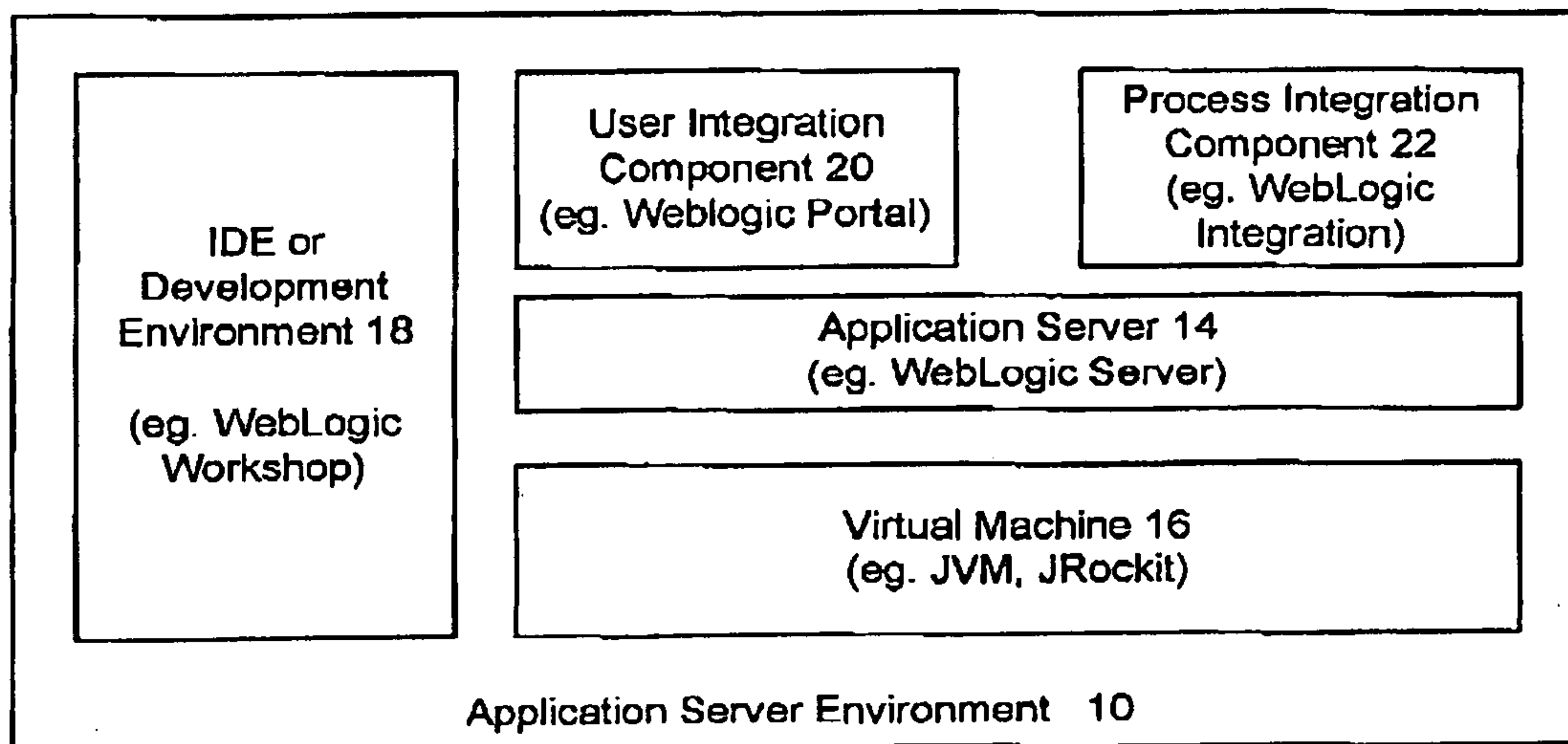
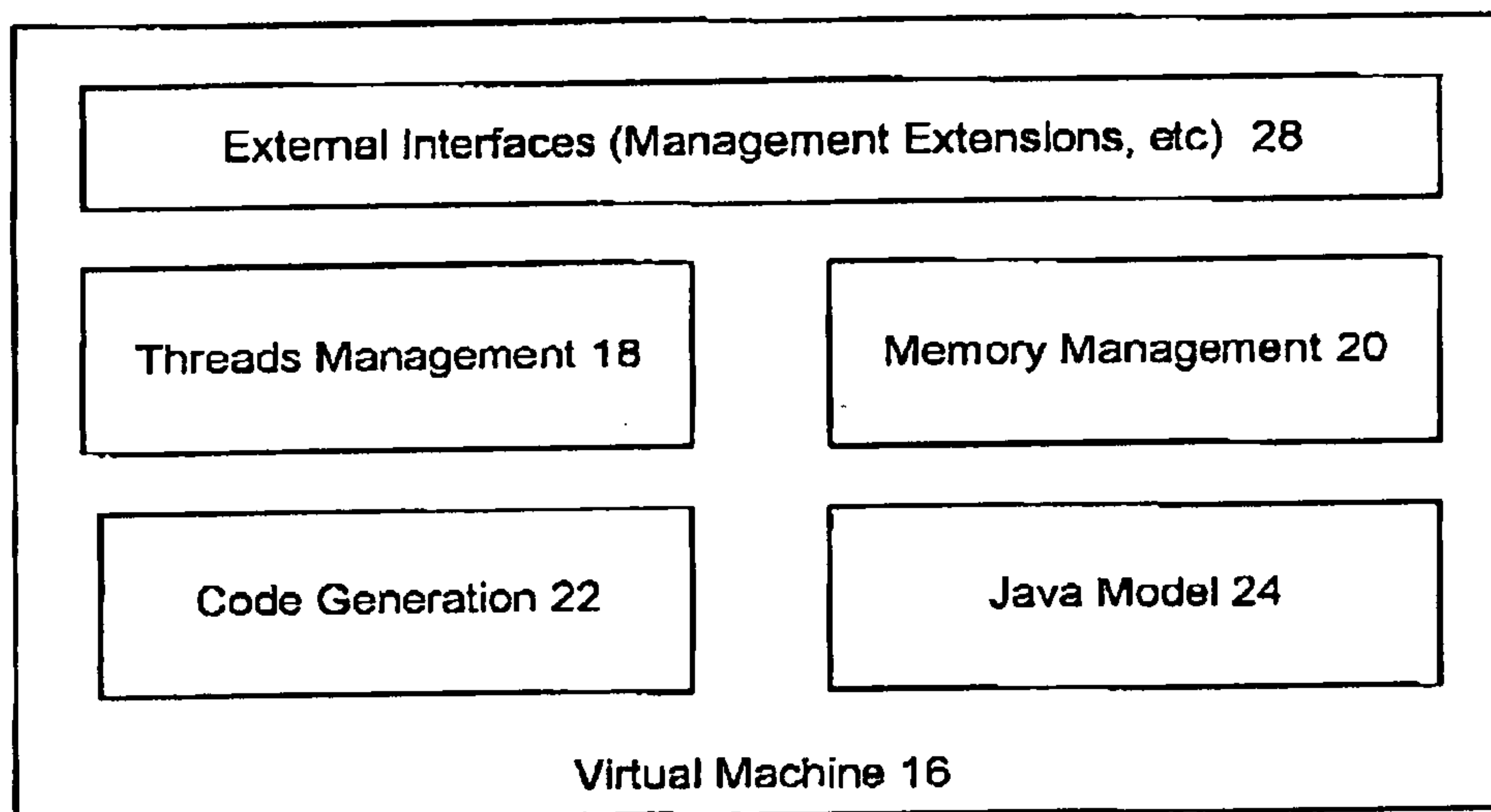
30

34. The method of claim 31 wherein the hypervisor is a Xen hypervisor.

35. The method of claim 31 wherein the hypervisor is a VMware hypervisor.

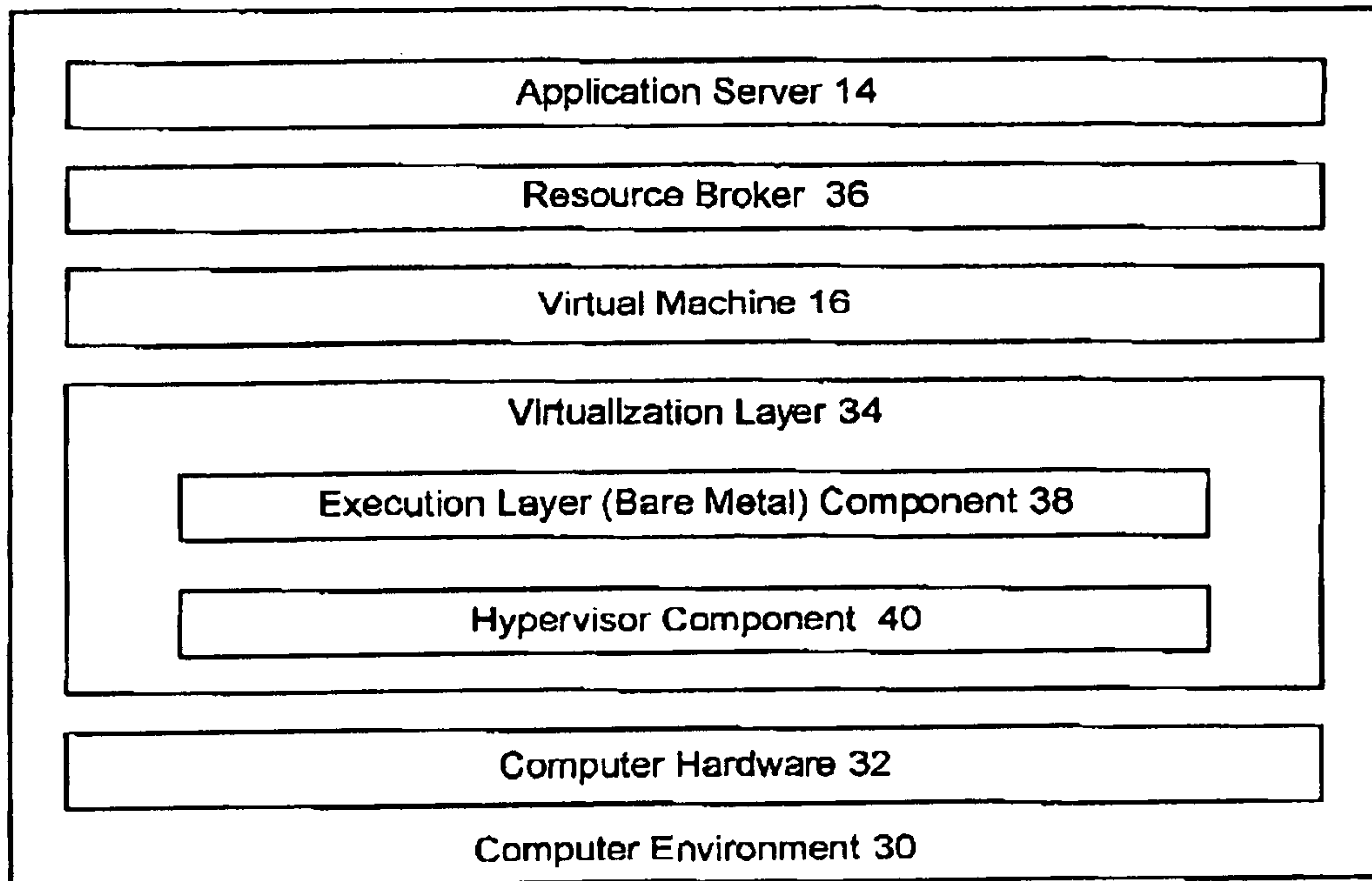
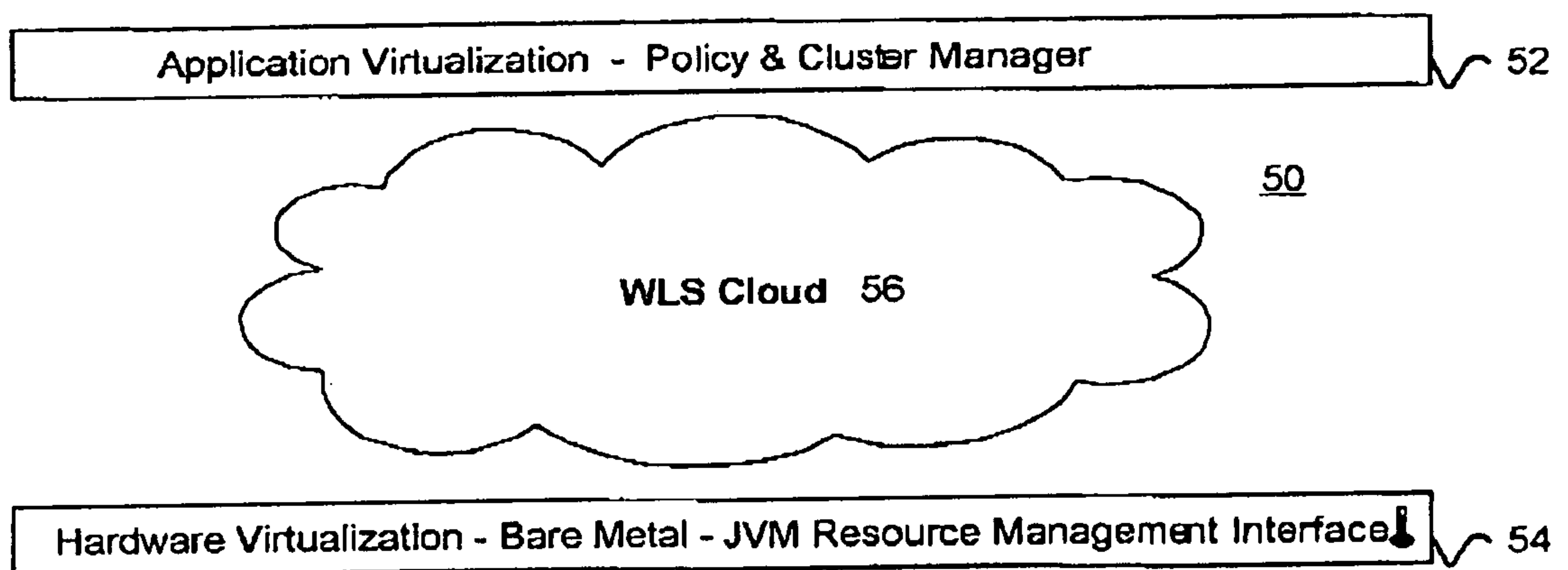
Attorney Docket: BEAS-01936WO0

1/5

**Figure 1****Figure 2**

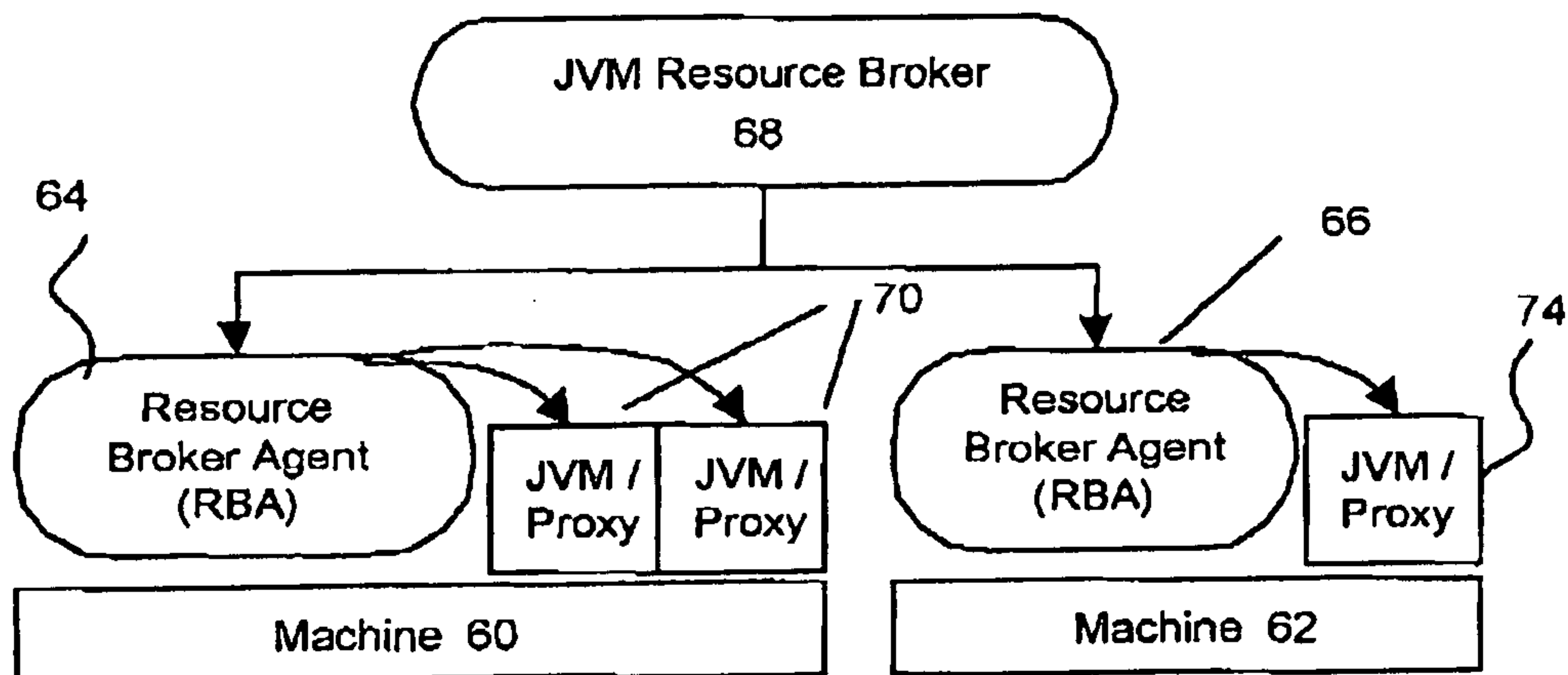
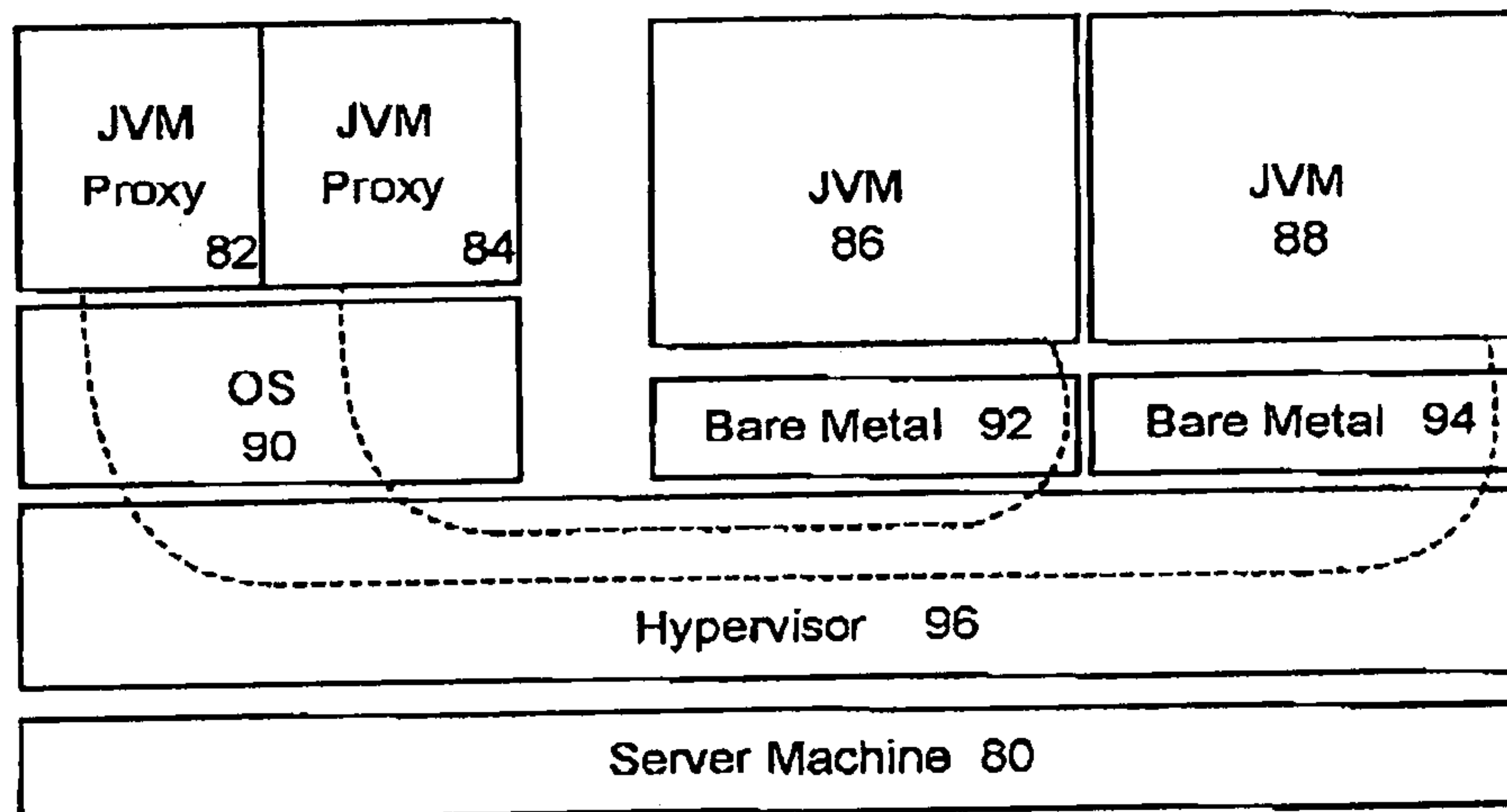
Attorney Docket: BEAS-01936WO0

2/5

**Figure 3****Figure 4**

Attorney Docket: BEAS-01936WO0

3/5

**Figure 5****Figure 6**

Attorney Docket: BEAS-01936WO0

4/5

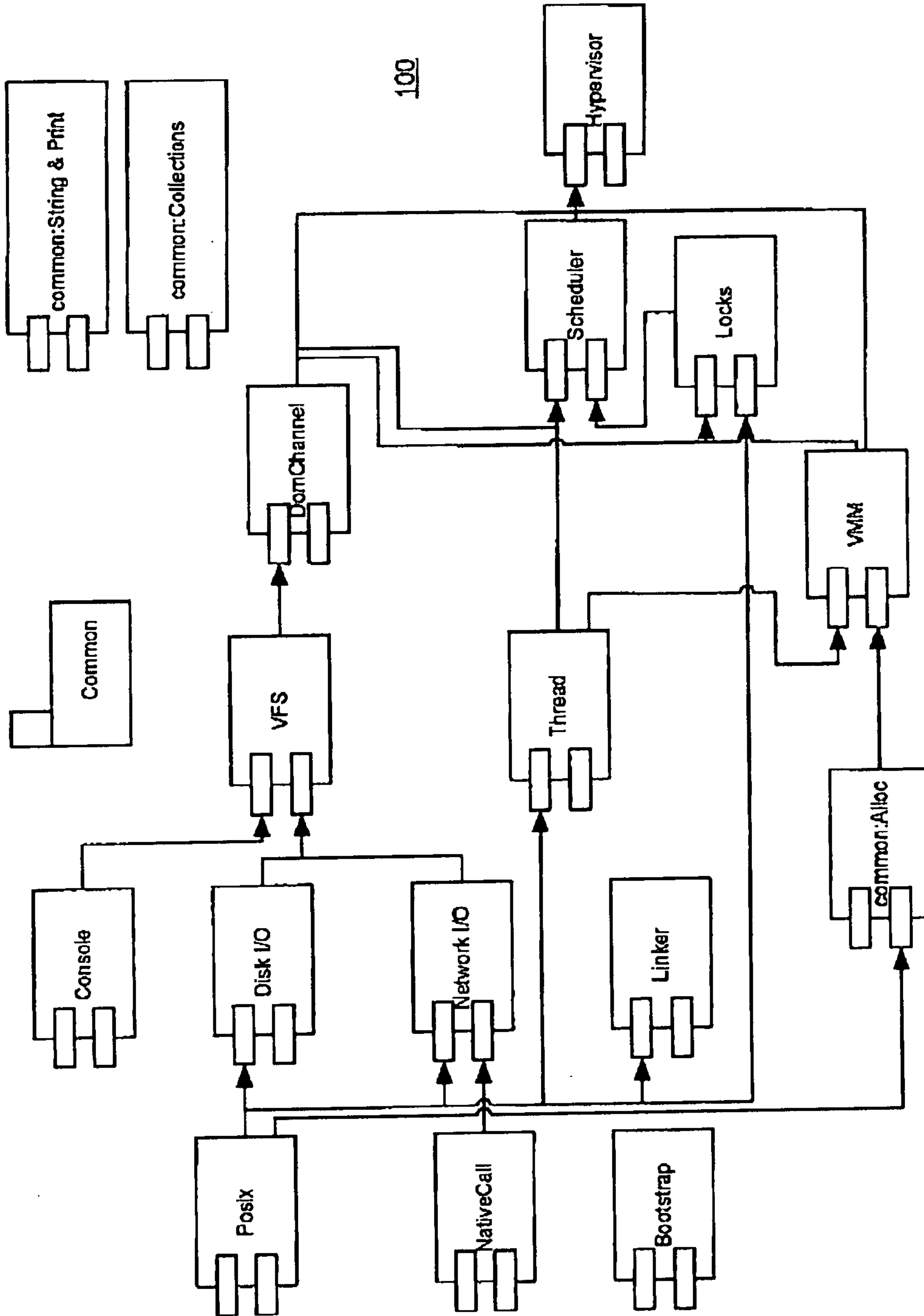
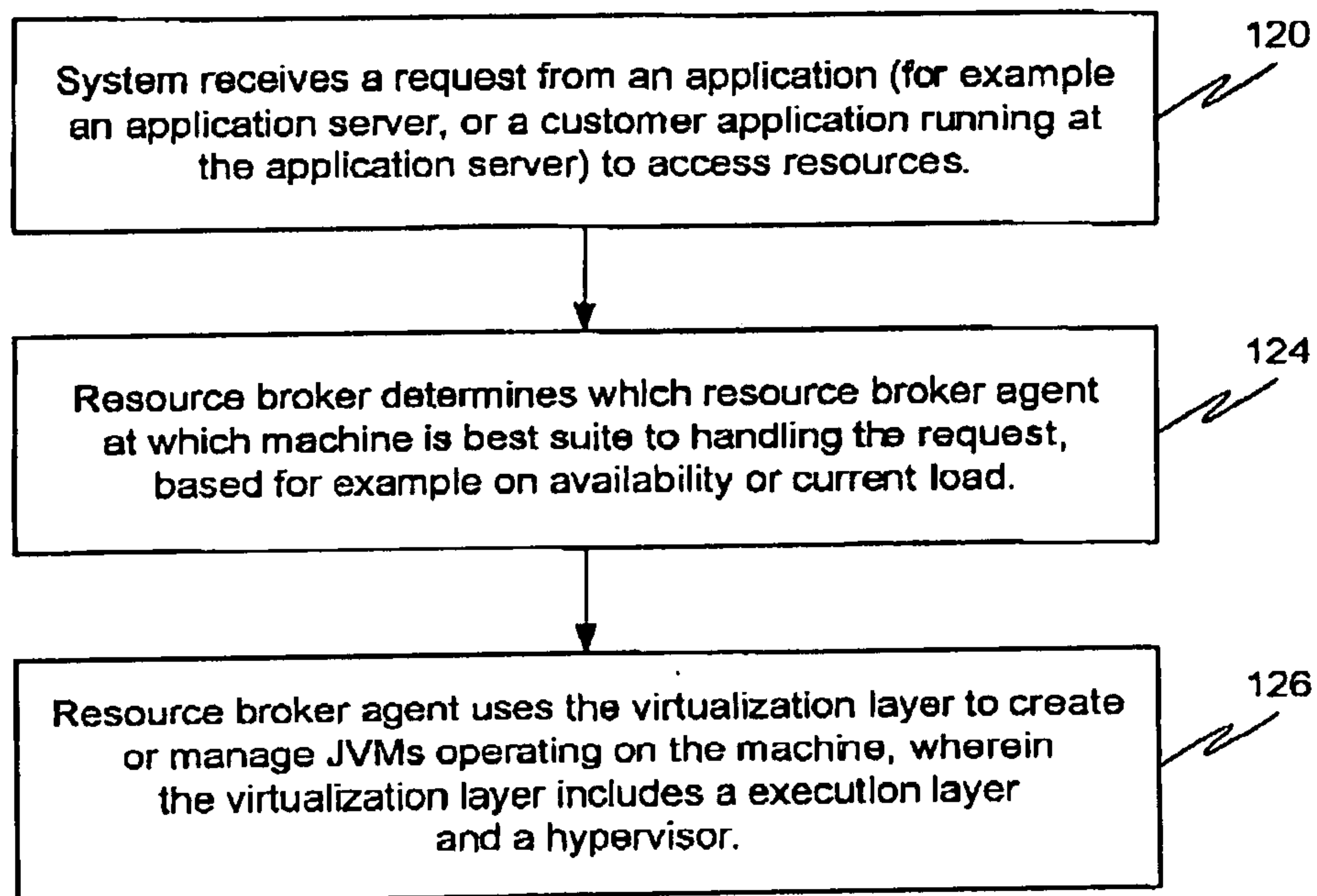


Figure 7

Attorney Docket: BEAS-01936WO0

5/5

**Figure 8**

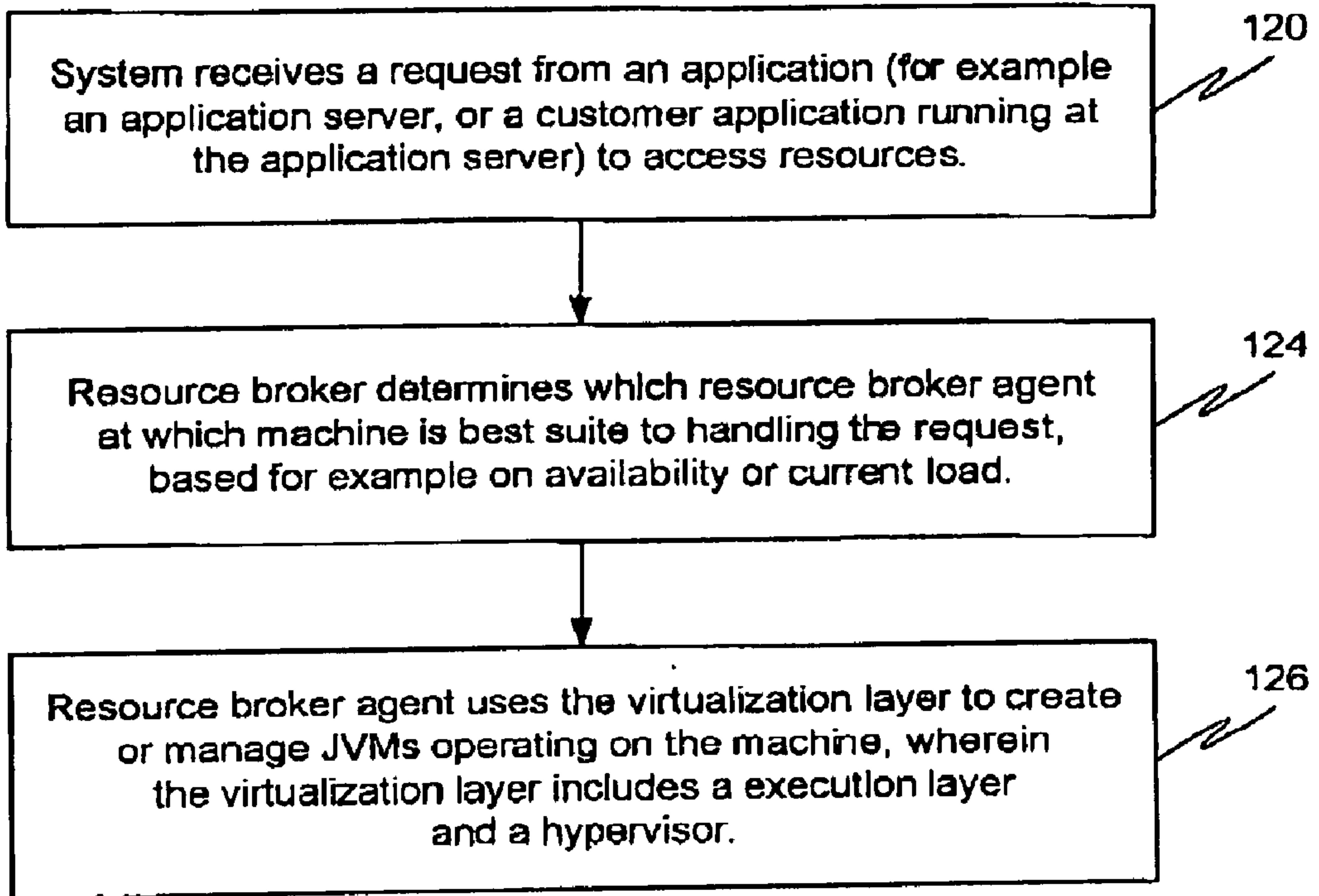


Figure 8