



- (51) **International Patent Classification:**
G06F 9/50 (2006.01) *G06F 15/177* (2006.01)
- (21) **International Application Number:**
PCT/MY2012/000080
- (22) **International Filing Date:**
12 April 2012 (12.04.2012)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
PI2011001637 12 April 2011 (12.04.2011) MY
- (71) **Applicant (for all designated States except US):** **MIMOS BERHAD** [MY/MY]; Technology Park of Malaysia, Bukit Jalil, 57000 Kuala Lumpur (MY).
- (72) **Inventors; and**
- (75) **Inventors/Applicants (for US only):** **MOHD NAZIR, Mohd Amril Nurman** [MY/MY]; c/o Mimos Berhad, Technology Park of Malaysia, Bukit Jalil, 57000 Kuala Lumpur (MY). **ABD KARIM, Mohd Bazli** [MY/MY]; c/o Mimos Berhad, Technology Park of Malaysia, Bukit Jalil,

57000 Kuala Lumpur (MY). **SALLEH, Mohd Sidek** [MY/MY]; c/o Mimos Berhad, Technology Park of Malaysia, Bukit Jalil, 57000 Kuala Lumpur (MY). **NG, Kwang Ming** [MY/MY]; c/o Mimos Berhad, Technology Park of Malaysia, Bukit Jalil, 57000 Kuala Lumpur (MY).

(74) **Agent:** **MOHAN, K.**; Adastra Intellectual Property Sdn Bhd, A-28-10, Menara UOA Bangsar, No. 5, Jalan Bangsar Utama 1, 59000 Kuala Lumpur (MY).

(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, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) **Designated States** (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH,

[Continued on next page]

(54) **Title:** METHOD AND SYSTEM FOR AUTOMATIC DEPLOYMENT OF GRID COMPUTE NODES

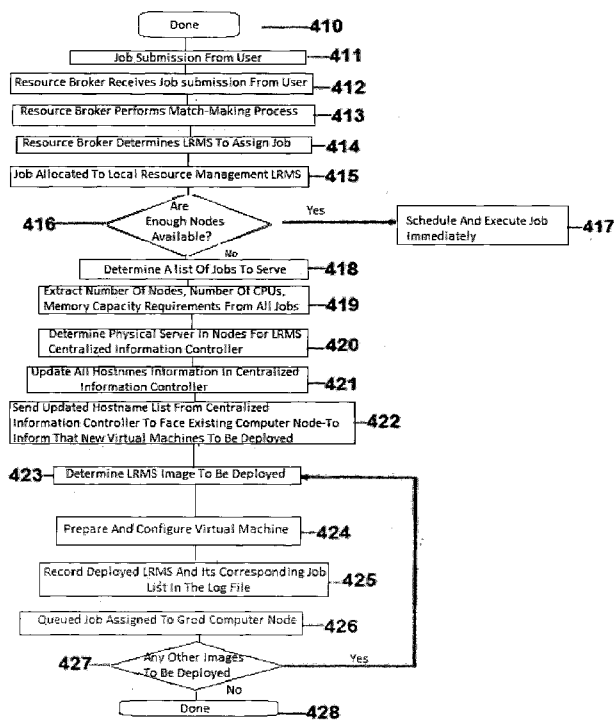


FIGURE 5

(57) **Abstract:** A system and method for automatic creation and deployment of new grid compute nodes by deploying virtual machines over physical machines to serve long waiting, high priority, urgent, and/or communication-intensive jobs.



GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, 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))*

METHOD AND SYSTEM FOR AUTOMATIC DEPLOYMENT OF GRID COMPUTE**NODES****TECHNICAL FIELD**

5

The present invention relates to resource management in a cluster, grid environment, and more particularly to system and method for automatic creation and deployment of new grid compute nodes by deploying virtual machines over physical machines to serve long waiting, high priority, urgent, and/or communication-intensive jobs.

10

BACKGROUND ART

15

A computational grid enables the aggregation of multiple clusters for dynamic sharing, selection, and management of resources. The grid middleware provides protocols and functionalities to enable these dynamic mechanisms of sharing resources. Existing cluster Management Systems such as Condor, SGE, LSF, Torque, and PBS Pro cluster systems are widely deployed in academic and commercial environments to manage the computational needs for the execution of sequential and parallel applications.

20

However, current grid and resource management systems have limitations. One of the limitations is long waiting time under scarce competitions of resources. Arriving jobs that request resources tend to cause fragmentations to other waiting jobs in queue causing long waiting time for most jobs. Hence, the user job is forced to wait

25

in the queue until suitable resources are released by other running jobs that have completed their execution. The current grid and resource management system cannot accommodate urgent/high priority jobs immediately/cannot accommodate jobs submitted by high priority users immediately. Currently, there is no build in
5 support to obtain the resources when user job needs them at burst or sudden spikes in demand. In most cases, at high load, it is often difficult to find physical resources that exactly match with the resource specification requested by the user.

On system where network connections do not resemble a flat all-to-all topology,
10 resource placement may impact performance as communication intensive parallel jobs. If latencies and network bandwidth between any two or more compute nodes vary significantly, the node allocation policy should attempt to allocate compute nodes of a given job as close to each other as possible to minimize impact of bandwidth and latency differences. Currently there is no support for resolving the
15 bandwidth and latency issues.

Patent US6353898 and US2010162259 disclosed resource management in clustered computer system, and virtualization-based resource management apparatus and method and computing system for virtualization-based resource
20 management respectively. US6353898 is distinctive over the present invention wherein the main focus is to provide a failure-proof system. In order to achieve this, failures and possible failures by node software, hardware and interconnections can be detected. Compensation is made for preventing failures, such as emergency communications and sharable resource allocation with minimal locking. Furthermore,

virtualization technology is not involved. It is observed that the steps involved are distinctive from that of the present invention.

US2010162259 is distinguished over the present invention whereby it covers method
5 for resource management comprising a plurality of virtual machines which monitors physical machines, in term of the resources utilized by each of the plurality of physical machines and time costs of the plurality of virtual machines, and to perform a resource reallocation and resource reclamation. However, the role and steps of the virtual machines are distinctive from that of the present invention, as the present
10 invention is focused on how to deploy compute node based on queued jobs and not to monitor only physical machines.

Effectively, there is a need for mechanism to deploy resources on demand when they are needed most, meaning those resources have to be made available
15 immediately with very little advance notice. This requirement is a must for urgent and/or high priority and bandwidth and latency sensitive jobs.

DISCLOSURE OF THE INVENTION

20 The present invention aims to provide a system and method for automatic creation and deployment of new grid compute nodes by deploying virtual machines over physical machines to serve long waiting, high priority, urgent, and/or communication-intensive jobs.

In a preferred embodiment of the present invention, a method for automatic deployment of grid compute nodes comprising the steps of determining a list of queued jobs to serve from queue of LRMS, extracting information on number of nodes, number of required CPUs, and minimum memory capacity requirements
5 from all queued jobs, determining quantity and physical server nodes to deploy virtual machine, integrating all newly virtual machine hostnames and IP addresses to be deployed in centralized information controller, sending updated hostname list from centralized information controller to each existing grid compute node to notify existence of new virtual machines to all grid compute nodes, determining virtual
10 memory image to be deployed, preparing and configuring virtual machine, recording number of deployed virtual machines and their corresponding job unique identifiers in the log file and assigning queued jobs to grid compute nodes.

In another preferred embodiment of the present invention, the method for preparing
15 and configuring the virtual machine images further comprising the steps of preparing images and configuration files for new virtual machines, mounting virtual machine images to physical server nodes, configuring host files for the virtual machine images, configuring network interface for virtual machine image, configuring name server for virtual machine image, sending updated host name lists from Centralized
20 Information Controller to new image, unmounting virtual machine image from physical server node, notifying LRMS of the addition of new virtual machine image, registering new virtual machine image as Grid Compute Node at LRMS and deploying new Grid Compute Node.

In another preferred embodiment of the present invention, configuration files include information on MAC address, VM name, number of and virtual CPUs.

In another preferred embodiment of the present invention, preparing image includes
5 duplicating a pre-defined virtual image for the creation of new images.

In another preferred embodiment of the present invention, the new images are updated with latest hostname list retrieved from centralized information controller.

10 In another preferred embodiment of the present invention, system for automatic deployment of grid compute nodes comprising a user interface, a virtual image repository, a local resource management and grid compute nodes.

The present invention consists of features and a combination of parts hereinafter
15 fully described and illustrated in the accompanying drawings, it being understood that various changes in the details may be made without departing from the scope of the invention or sacrificing any of the advantages of the present invention.

BRIEF DESCRIPTION OF THE ACCOMPANYING DRAWINGS

20

To further clarify various aspects of some embodiments of the present invention, a more particular description of the invention will be rendered by references to specific embodiments thereof, which are illustrated, in the appended drawings. It is appreciated that these drawings depict only typical embodiments of the invention
25 and are therefore not to be considered limiting of its scope. The invention will be

described and explained with additional specificity and detail through the accompanying drawings in which:

FIG.1 illustrates flow diagram of job submission to Local Resource Management System (LRMS) of a prior art.

FIG.2 illustrates flow of job execution based on priority on Local Resource Management System (LRMS) of a prior art.

FIG.3 illustrates prior art of Job 4, Job 5 and Job 6 which are queued by LRMS due to insufficient resources.

FIG.4 illustrates the utilization of centralized information controller for synchronization of all network information.

FIG.5 illustrates method to serve job of the present invention.

FIG.6 illustrates steps to configure virtual machine image for deployment

FIG. 7 illustrates steps to prepare pre-configured virtual machine images to support automatic deployment and connection with LRMS server at runtime.

FIG. 8 illustrates preparing pre-configured virtual machine image for grid compute node

DETAILED DESCRIPTION OF THE ACCOMPANYING DRAWINGS

Now referring to FIG. 1, the figure illustrates flow diagram of job submission to Local Resource Management System (LRMS) of one of the prior arts (100). The job will be submitted to the LRMS (10) and determine whether sufficient nodes are available (11) to submit to compute nodes (14). When there are no free compute nodes available to execute jobs (13), the jobs will be queued until sufficient compute nodes are available and the process ends (15).

Meanwhile, as according to the present invention, FIG. 2 describes flow of job execution based on priority on Local Resource Management System (LRMS). Job is submitted to LRMS (210). The LRMS will check whether there are sufficient nodes available (211). If the priority of the job is low (113), the system will wait until there is available nodes (117). When high priority jobs are allocated at queue, low priority jobs that are executed will be suspended (114) and queued jobs will be released (115) and executed (116). During determining whether there s sufficient nodes available (112), if there is suffiecient nodes, the system will directly schedule the job to the LRMS (118) and conclude the job (119).

Now referring to FIG.3, the figure illustrates a prior art where Job 4, Job 5 and Job 6 are queued by LRMS (210) due to insufficient resources. Meanwhile, the present invention in FIG. 3b illustrates the present invention through automatic deployment of grid compute nodes. The solution employs modern virtualization technology to configure and deploy virtual machines to accommodate and satisfy the requirements of queue and/or pending jobs. The solution is through present steps and methods to

configure, and deploy virtual machines on physical server nodes registered in a grid system. New methods and system are disclosed on the provisioning, and deploying, and managing jobs on a pool of grid compute nodes.

5 FIG. 4 shows the present invention makes use of the centralized information controller (314) to synchronize all network information of each existing and new Grid compute node. The centralized information controller (314) compiled all the jobs from LRMS (313) prior to distribution to the physical servers (311, 312). Without a centralized information controller (314), it is not possible for grid compute node to
10 recognize one another. The image repository (310) stores the image from the system.

According to FIG. 5, the method integrates virtualization technology to support execution requirement of queued, pending, urgent and/or high priority jobs. It
15 provides method and system allowing critical execution of queued and/or pending, urgent and high priority jobs in the times when current available resources do not have the capabilities to match the requirements of the job.

The system starts (410) by receiving job submission directly from the user (410),
20 from which resource broker performs match-making process (413) to determine Local Resource Management System (LRMS) to assign the job (414). If job requirements can be met, the job is allocated to LRMS (415) and will be scheduled immediately (417) when there are sufficient nodes (416). However, if the job requirement cannot exactly be met, the present invention provides method to serve
25 the job by determining a list of queued jobs (418) to serve from the queue of the

LRMS. Thereafter, it will extract information on the number of nodes, number of required CPUs, and minimum memory capacity requirements from all queued jobs (420).

5 For a list of current queued jobs, the system determines how many and which physical server nodes to deploy virtual machine (420). It includes all newly virtual machine hostnames and IP addresses to be deployed in centralized information controller (421). Thereafter, updated hostname list from centralized information controller is sent to each existing grid compute nodes as to notify the existence of
10 new virtual machines to all grid compute nodes (422). Later, virtual machine image is determined to be deployed (423). The preparation of virtual machine can be seen from FIG.6.

According to FIG. 6, virtual machines is later prepared and configured. The steps to
15 configure virtual machines are by preparing images and the configuration files for the new virtual machines. Configuration files include information on MAC address, VM name, number of and virtual CPUs, preparing image involves in duplicating a pre-defined virtual image for the creation of new images (510). The new images are updated with the latest hostname list retrieved from centralized information
20 controller.

The VM image is then mounted to the physical server nodes and host files for the VM image are configured (511, 512, 513 and 514). Thereafter, updated hostname lists from Centralized Information Controller are delivered to new image (515) and
25 VM image from physical server node is unmounted (516). LRMS is notified by

addition of new VM image (517). New VM image is registered as Grid Compute Node at LRMS (518). Finally, new Grid Compute Node is deployed (519).

According to FIG. 7, it is also necessary to prepare the pre-configured VM images to support automatic deployment and connection with the LRMS server at runtime. The method starts (610) with configuring the LRMS server to support auto-deployment and connection with grid compute nodes (611). At this stage, it is assumed that LRMS has been installed successfully. The LRMS is configured to have two network interfaces for internal and external networks. The internal network allows communication between the LRMS and grid compute nodes. The external network allows communication between LRMS to public network. Network Address Translator (NAT) is configured on LRMS (612). NAT configuration allows communication of grid compute nodes to public network. IP forwarding at LRMS is enabled (613) at LRMS to allow communication of grid compute nodes to public network. Routing from internal IP address (gateway) to external IP address is enabled (614) to allow communication of grid compute nodes to public network. The method ends (615) once the steps are completed.

According to FIG. 8, the pre-configured virtual machine image is prepared for grid compute nodes. The method is executed (710) through creating a new disk image (711) and deploying and installing operating system on newly created disk image (712). Thereafter, LRMS client is configured to establish communication with LRMS server (713). The LRMS server information is included in LRMS client configuration files to allow auto deployment of grid compute node. A user account from LRMS server to LRMS client is mapped (714) to enable job execution at LRMS client.

Passwordless communication between LRMS server and LRMS client is then configured for all user accounts (715). The steps end (716) once the sequence is completed.

- 5 In the foregoing specification, specific embodiments of the present invention have been described. However, one of ordinary skill in the art will appreciate that various modifications and changes can be made without departing from the scope of the present invention as set forth in the various embodiments discussed above and the claims that follow. Accordingly, the specification and figures are to be regarded in an
- 10 illustrative rather than a restrictive sense, and all such modifications are intended to be included within the scope of present invention. The benefits, advantages, solutions to problems, and any element(s) that may cause any benefit, advantage, or solution to occur or become more pronounced are not to be construed as a critical, required, or essential features or elements as described herein.

CLAIMS

1. A method for automatic deployment of grid compute nodes comprising the
5 steps of:

determining a list of queued jobs to serve from queue of LRMS (418);

extracting information on number of nodes, number of required CPUs, and minimum
memory capacity requirements from all queued jobs (419);

determining quantity and physical server nodes to deploy virtual machine (420);

10 integrating all newly virtual machine hostnames and IP addresses to be deployed in
centralized information controller (421);

sending updated hostname list from centralized information controller to each
existing grid compute node to notify existence of new virtual machines to all grid
compute nodes (422);

15 determining virtual memory image to be deployed (423);

preparing and configuring virtual machine (424);

recording number of deployed virtual machines and their corresponding job unique
identifiers in the log file (425); and

assigning queued jobs to grid compute nodes (426).

20

2. The method for preparing and configuring the virtual machine images as
according to claim 1 further comprising the steps of

preparing images and configuration files for new virtual machines (510);

mounting virtual machine images to physical server nodes (511);

25 configuring host files for the virtual machine images (512);

configuring network interface for virtual machine image (513);

configuring name server for virtual machine image (514);

sending updated host name lists from Centralized Information Controller to new image (515);

5 unmounting virtual machine image from physical server node (516);

notifying LRMS of the addition of new virtual machine image (517);

registering new virtual machine image as Grid Compute Node at LRMS (518); and

deploying new Grid Compute Node (519).

10 3. The method according to claim 2 wherein configuration files includes information on MAC address, VM name, number of and virtual CPUs.

4. The method according to claim 2 wherein preparing image (510) includes duplicating a pre-defined virtual image for the creation of new images.

15

5. The method according to claim 4 wherein the new images are updated with latest hostname list retrieved from centralized information controller.

6. The method according to claim 1 includes system for automatic deployment of
20 grid compute nodes comprising a user interface, a virtual image repository, a local resource management and grid compute nodes.

1/8

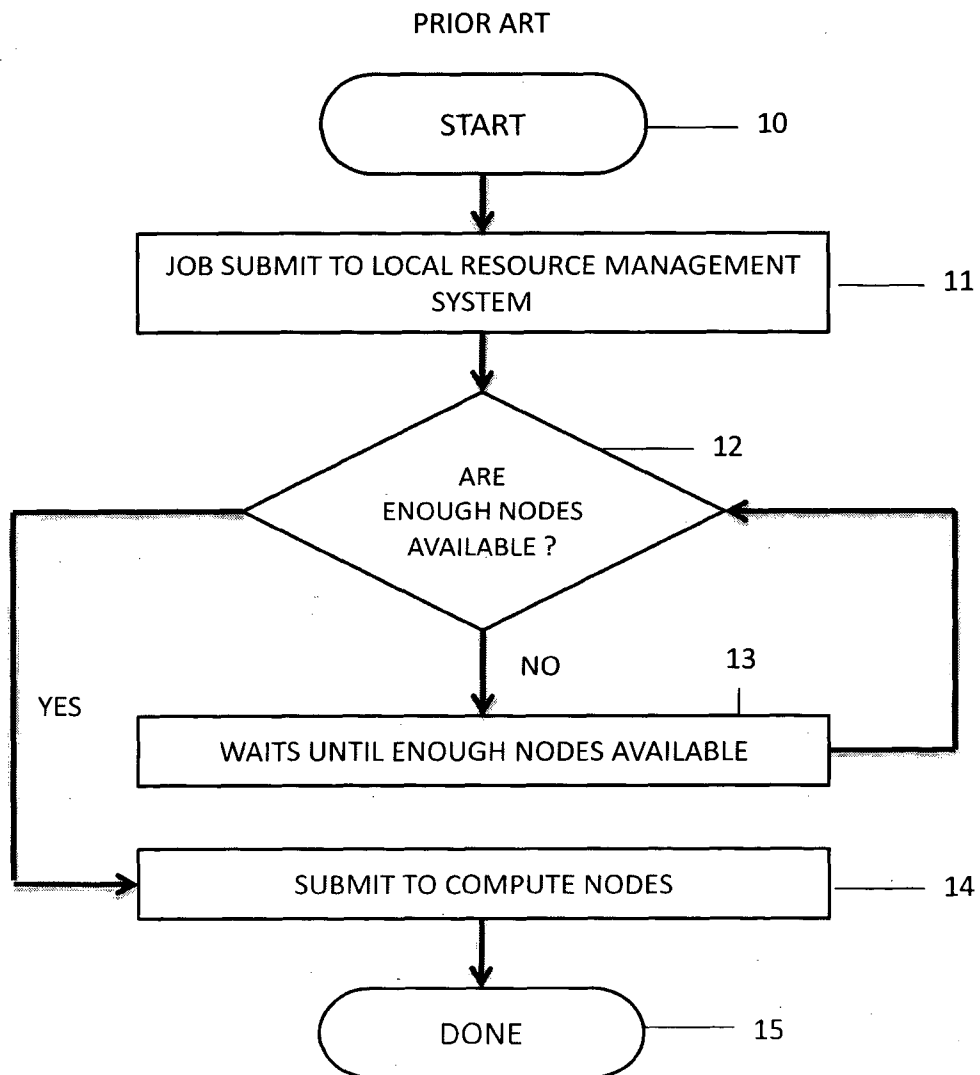
100

FIGURE 1

2/8

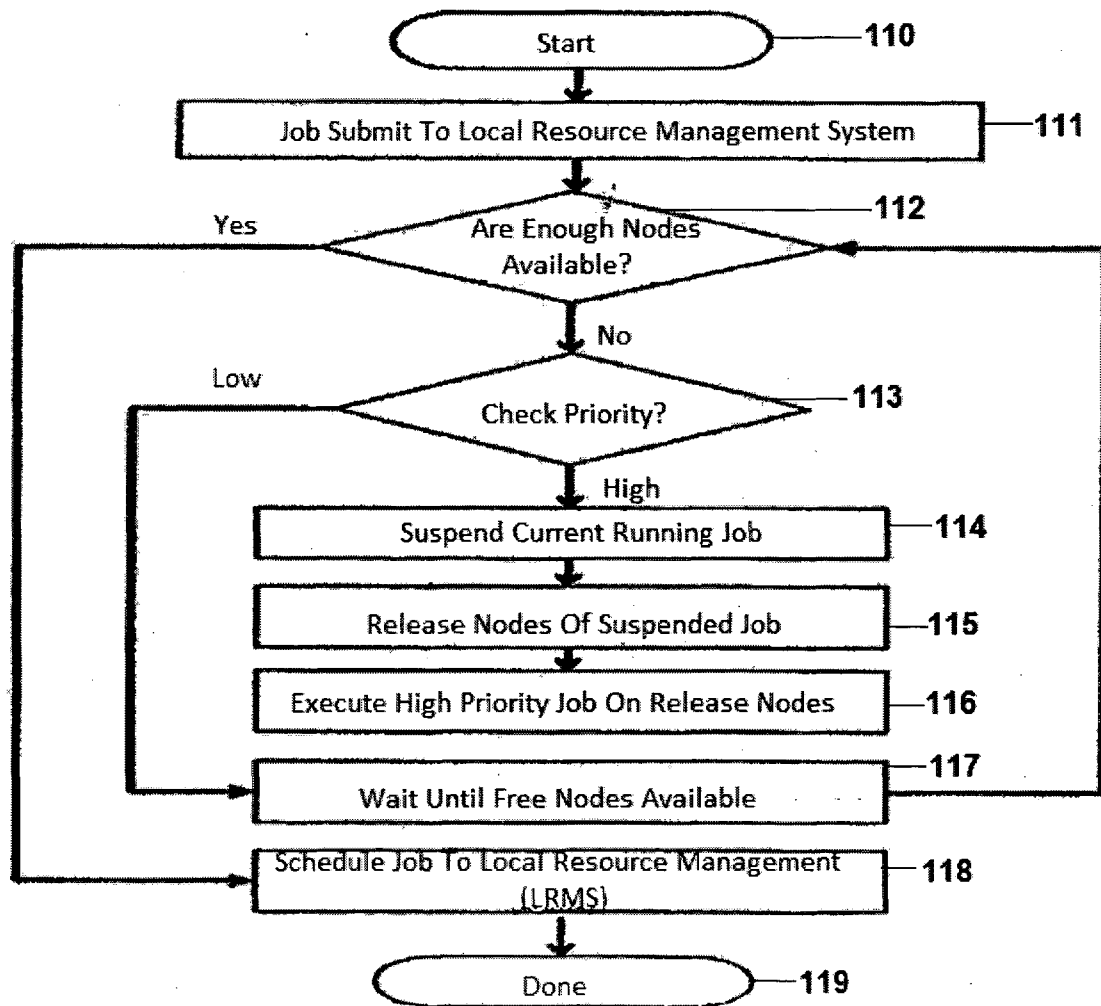
200

FIGURE 2

3/8

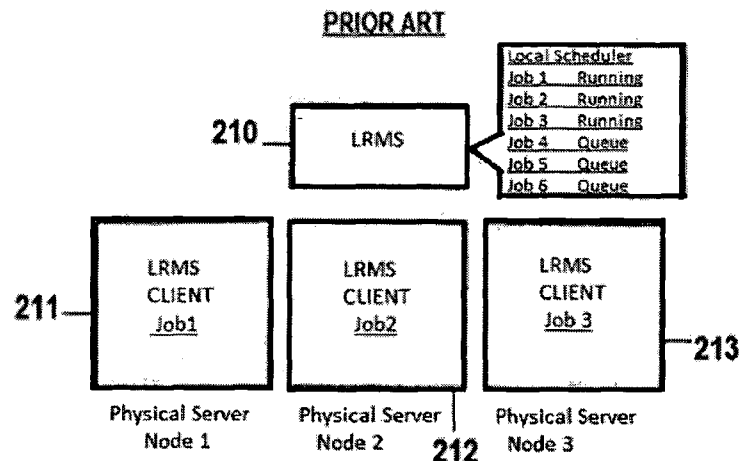
300

FIGURE 3a

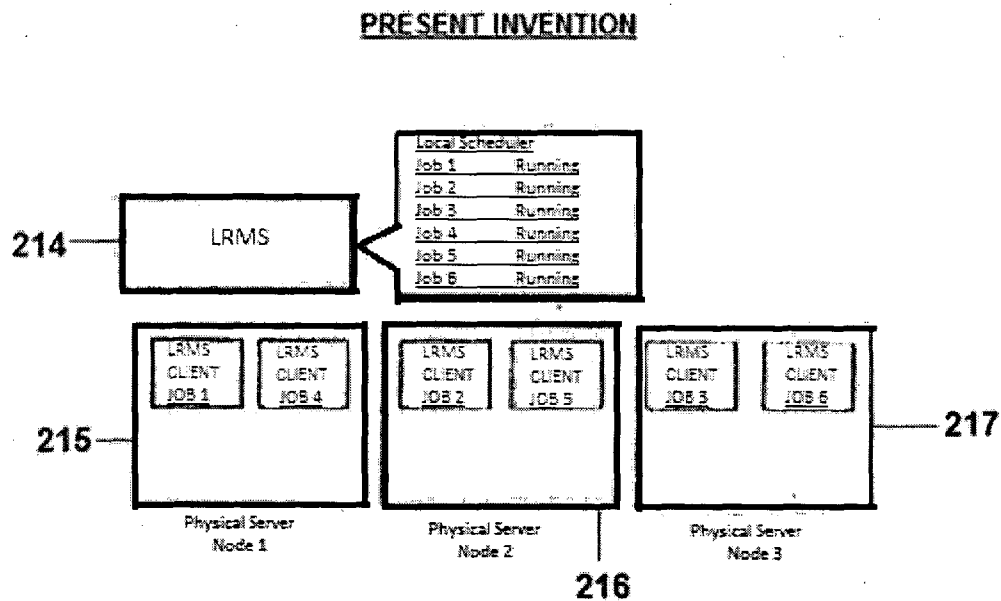


FIGURE 3b

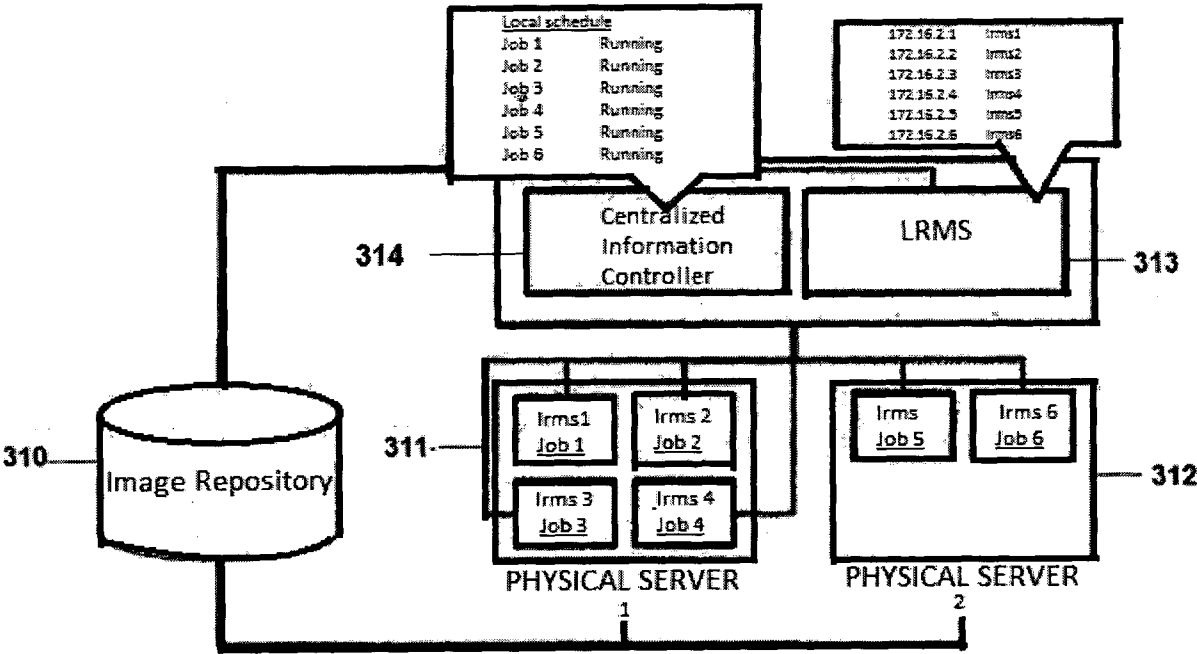


FIGURE 4

5/8

500

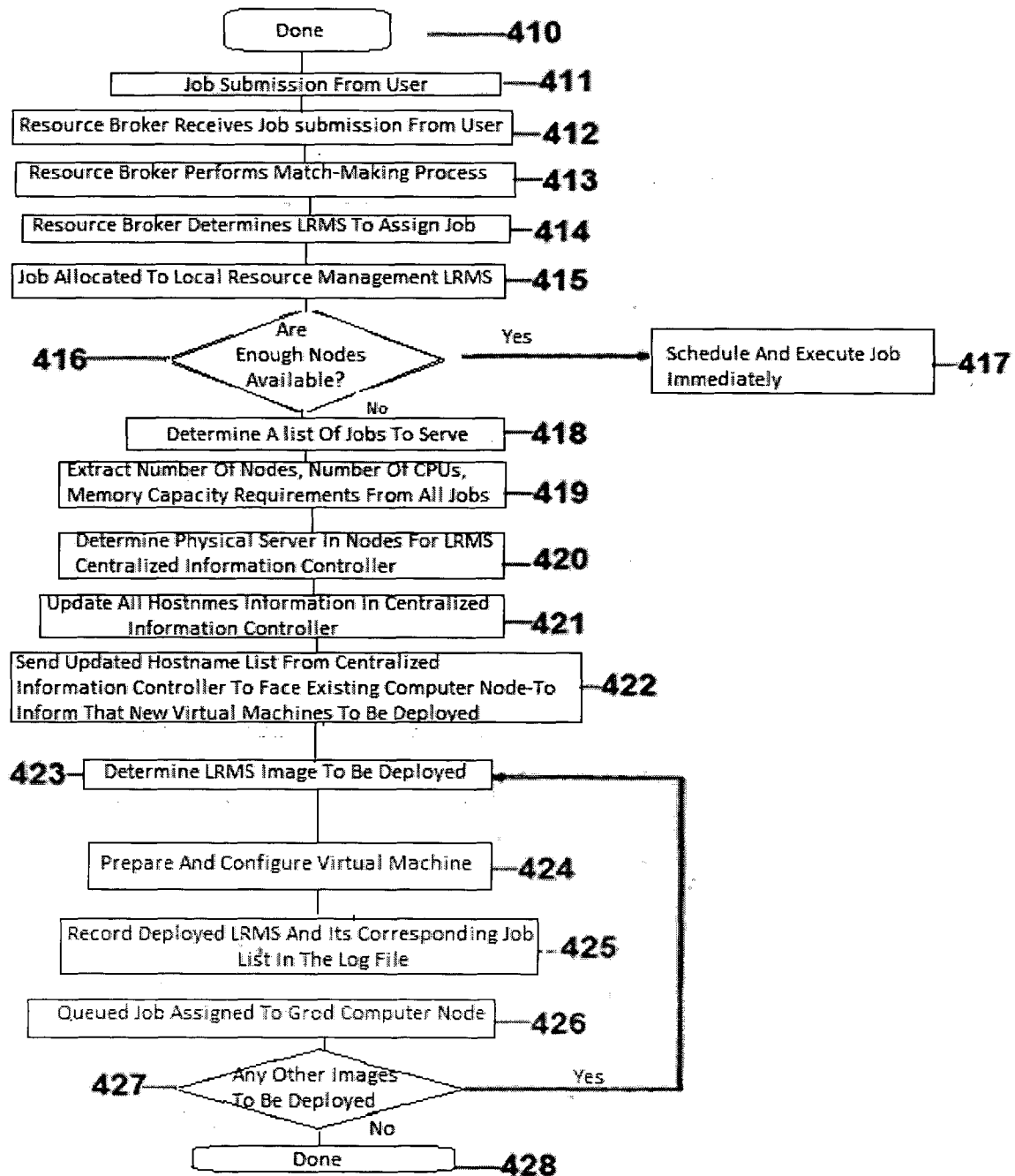


FIGURE 5

6/8

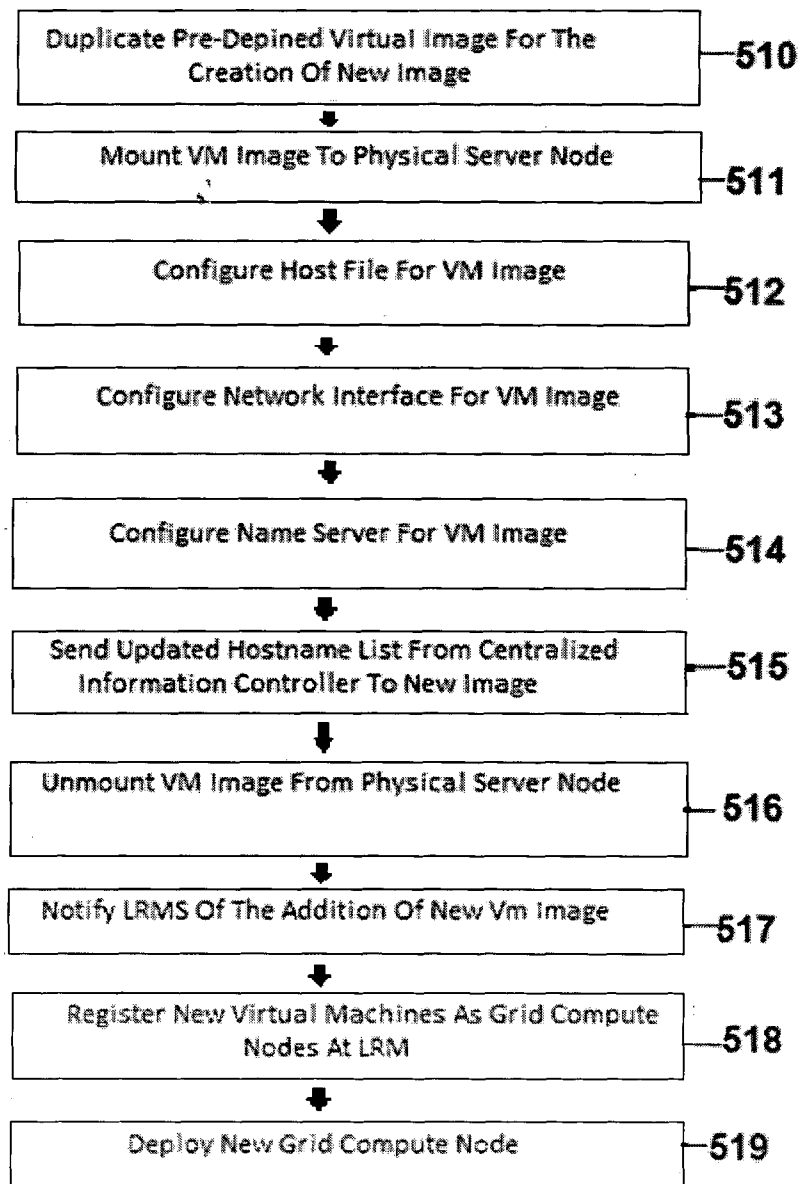
600

FIGURE 6

7/8

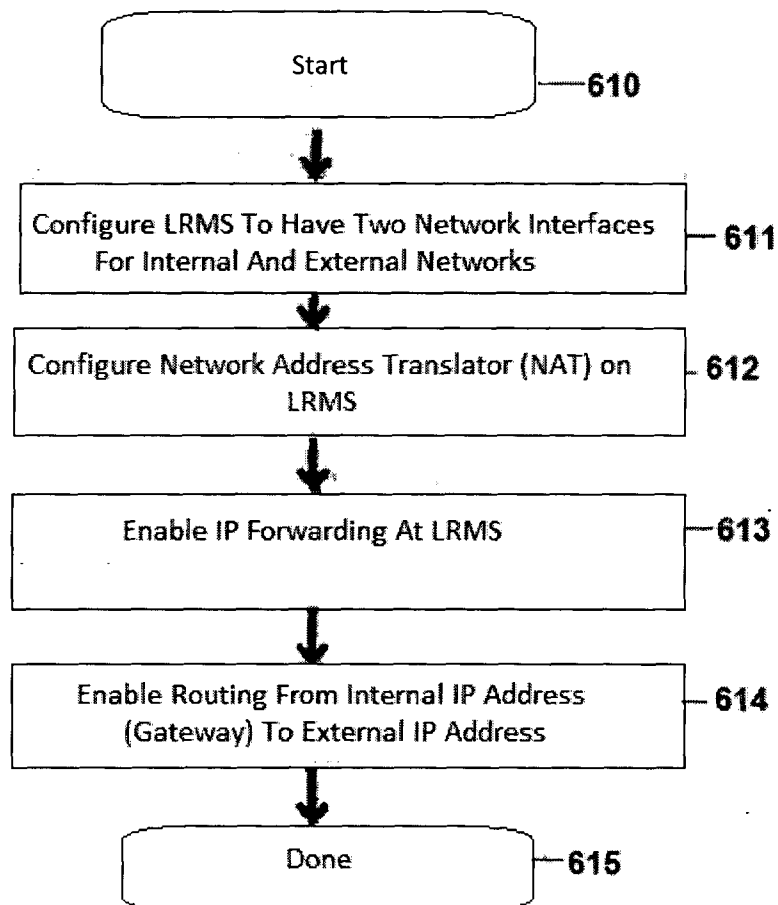
700

FIGURE 7

8/8

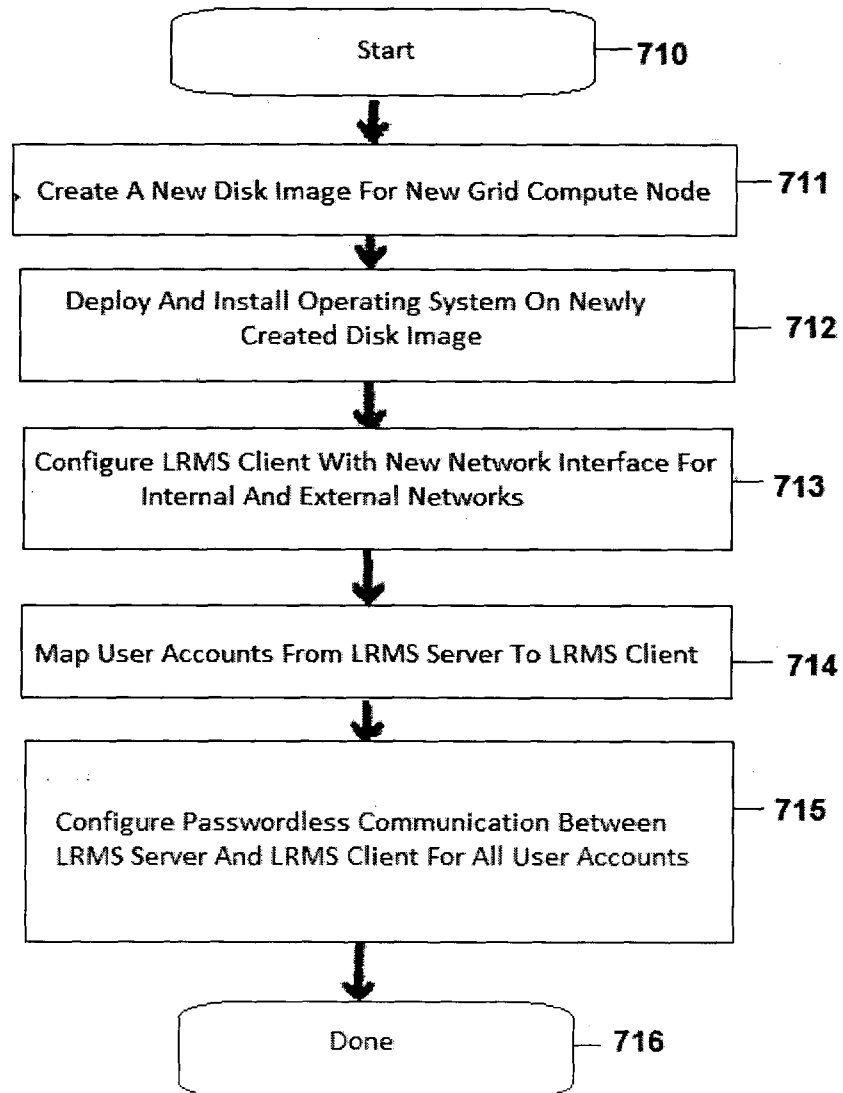
800

FIGURE 8

INTERNATIONAL SEARCH REPORT

International application No.

PCT/MY2012/000080

A. CLASSIFICATION OF SUBJECT MATTER

G06F 9/50 (2006.01) G06F 15/177 (2006.01)

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)

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

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

WPI, Google Scholar, IEEE Xplore; Keywords: (grid, cluster, virtual, machine, appliance, vm, deploy, configure, create, new, instantiate, generate, build, additional, increase, schedule, work flow, job, task, queue, list, demand, memory, cpu, processor, node, QOS, require, minimum, parameter, number, quantity, notify, inform, alert, advice, announce, image) + similar keywords

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
	Documents are listed in the continuation of Box C	



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	"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
"E" earlier application or patent but published on or after the international filing date	"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
"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)	"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
"O" document referring to an oral disclosure, use, exhibition or other means	"&" document member of the same patent family
"P" document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search
10 August 2012

Date of mailing of the international search report
13 August 2012

Name and mailing address of the ISA/AU

AUSTRALIAN PATENT OFFICE
PO BOX 200, WODEN ACT 2606, AUSTRALIA
Email address: pct@ipaustalia.gov.au
Facsimile No.: +61 2 6283 7999

Authorized officer

Olutope Omogbenigun
AUSTRALIAN PATENT OFFICE
(ISO 9001 Quality Certified Service)
Telephone No. 0262832876

INTERNATIONAL SEARCH REPORT		International application No.
C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		PCT/MY2012/000080
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2005/0060704 A1 (BULSON et al.) 17 March 2005 See whole document especially abstract, paragraphs [0003],[0006],[0020],[0027]-[0029],[0032]-[0034],[0036],[0037],[0043],[0048]	1-6
Y	MOORE J. et al., "Managing Mixed-Use Clusters with Cluster-on-Demand", Technical Report, Computer Science Department, Duke University, November 2002, pages 1-12. [online], [retrieved on 7 August 2012]. Retrieved from the Internet <URL: http://www.cs.duke.edu/nicl/pub/papers/cod.pdf > See whole document especially abstract, Sections 1-5,7	1-6

INTERNATIONAL SEARCH REPORT Information on patent family members		International application No. PCT/MY2012/000080	
This Annex lists known patent family members relating to the patent documents cited in the above-mentioned international search report. The Australian Patent Office is in no way liable for these particulars which are merely given for the purpose of information.			
Patent Document/s Cited in Search Report		Patent Family Member/s	
Publication Number	Publication Date	Publication Number	Publication Date
US 2005/0060704 A1	17 Mar 2005	CN 1604039 A	06 Apr 2005
		EP 1665047 A2	07 Jun 2006
		JP 2007506169 A	15 Mar 2007
		KR 20060069464 A	21 Jun 2006
		US 2005060704 A1	17 Mar 2005
		WO 2005026947 A2	24 Mar 2005
End of Annex			
Due to data integration issues this family listing may not include 10 digit Australian applications filed since May 2001. Form PCT/ISA/210 (Family Annex)(July 2009)			