

Magrathea – Scheduling Virtual Grids with Preemption

Jiří Denemark and Miroslav Ruda

CESNET and Masaryk University, Czech Republic



Motivation for virtualization of Grid

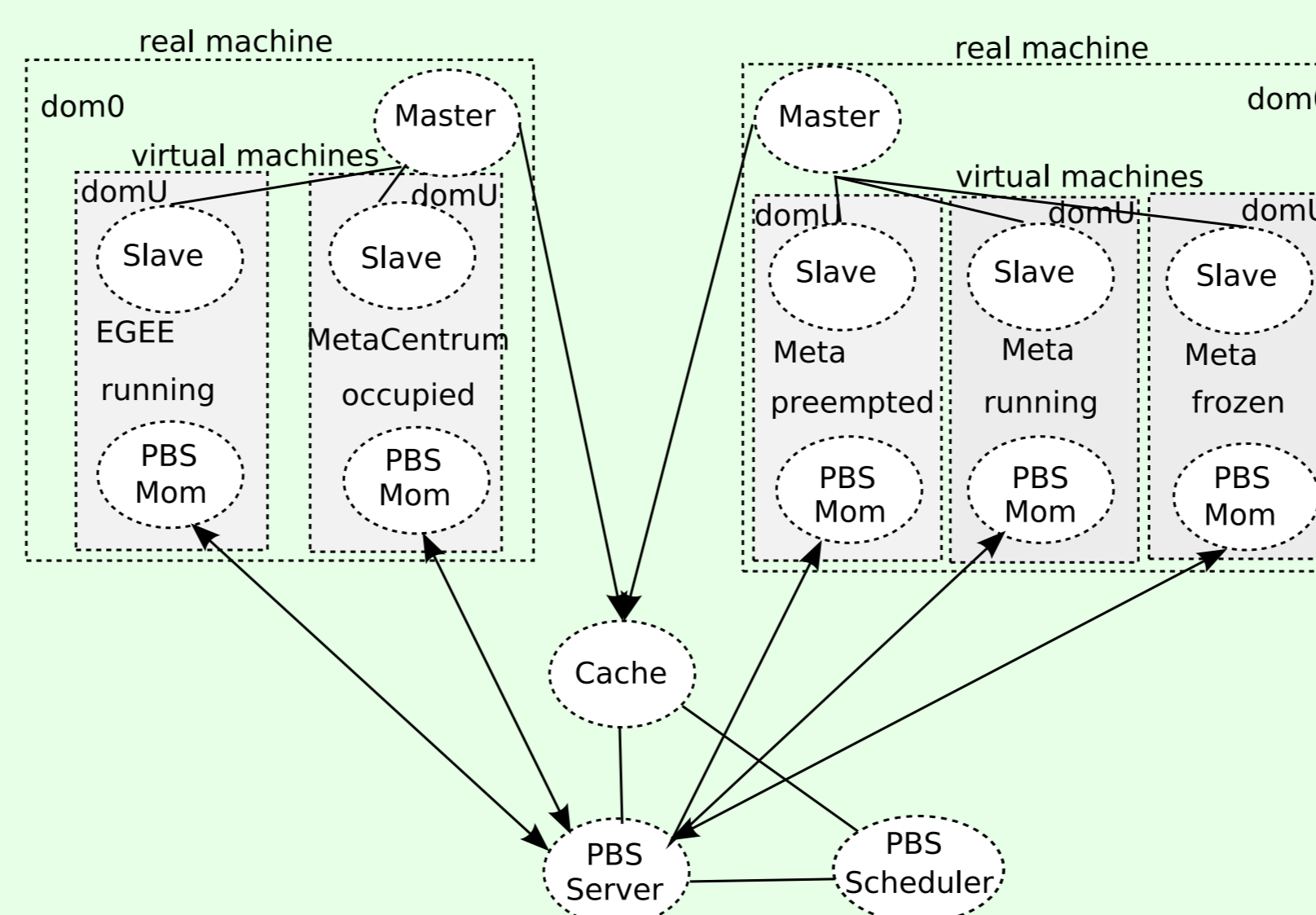
- Different production environment (strict requirements of projects)
 - ▶ flexible setup, where each application can run in its tailored environment
- Coexistence of traditional long-running batch jobs with
 - ▶ fast turn-around of short jobs
 - ▶ interactive jobs
 - ▶ service whose use varies over time
- Encapsulation used for
 - ▶ advance scheduling techniques (migration, preemption)
 - ▶ security improvements – separation of jobs

Magrathea architecture

Designed to allow Grid job scheduling systems to deal with several virtual machines running on a single computer and to submit jobs correctly into those VMs.

Design Requirements:

- More active (i.e., running) virtual machines than physical resources. The resource management system must schedule jobs to these machines exclusively, not overloading the resources
- As small as possible dependence on actual resource management system
- Minimal changes or modifications of the resource management system (PBSPro in our case)
- Independence on system used for management of virtual machines
- Independence on particular VM implementation (Xen and VServer in current implementation)



Supported scenarios

Exclusive allocation

Exclusive use of the physical resource by one virtual machine at a time while supporting concurrent active “wait” of several virtual machines on the same resource.

Concurrent usage

Sharing one physical machine among several virtual machines running concurrently and assigning of resources (CPUs, memory) to virtual machines according to requirements of jobs running in these virtual machines.

Preemption

Support for preemption of virtual machines, eventually extended with suspension and migration to different physical machine.

Frozen domain

Support for “frozen” services that are repeatedly invoked and suspended on user request.

Domain status

Magrathea status of a virtual node used by the resource management system for decisions made by the scheduler.

PBS scheduler modified to:

- respect Magrathea status – submit only to domains with Magrathea status *free*, *running*, *occupied-would-preempt*, and *running-preemptible*
- sort nodes using Magrathea status (non-preemption first)

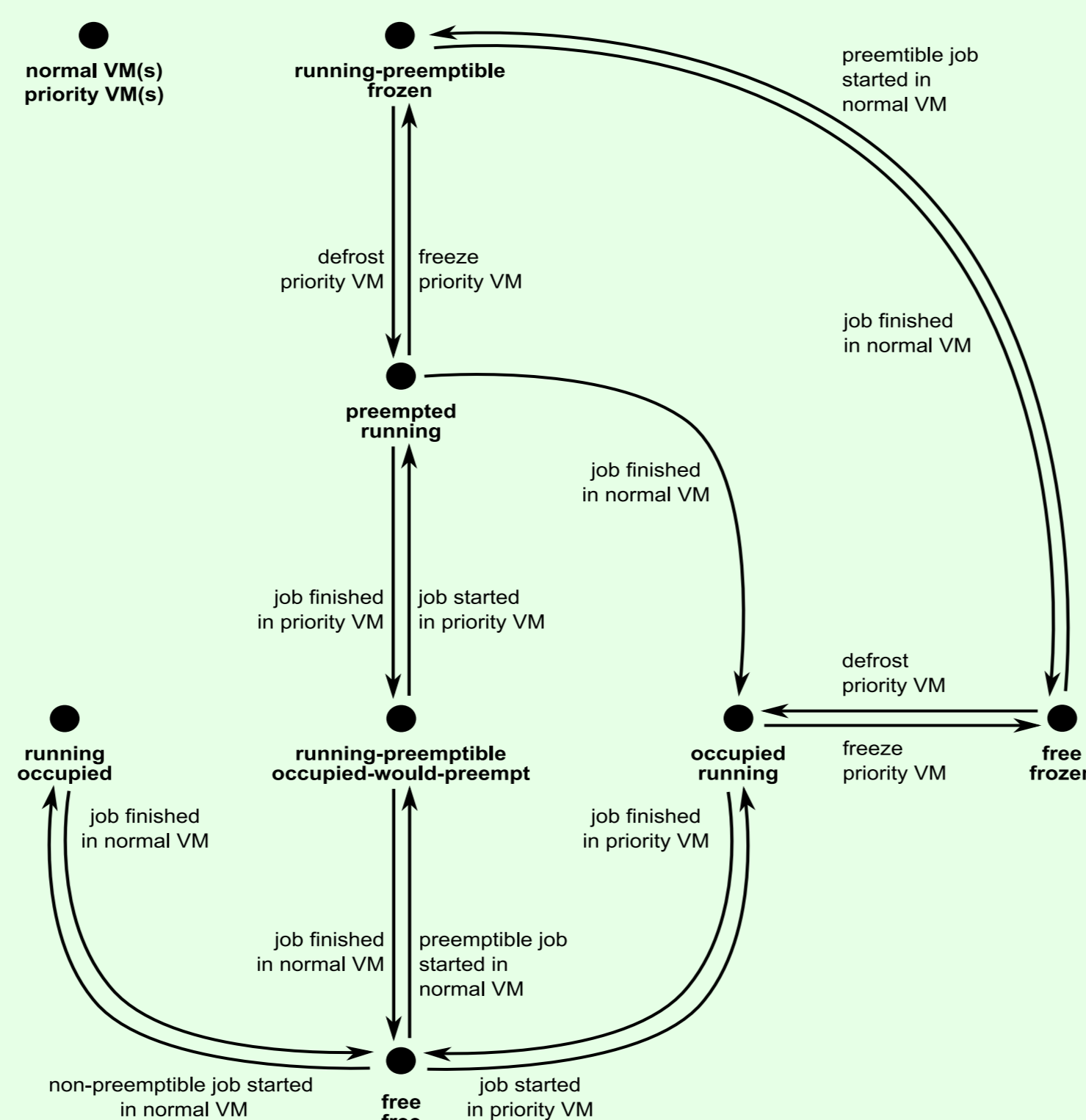
PBS server modified to:

- dedicate queue for high priority jobs
- support frozen domains

Job prologue/epilogue:

- hooks to Magrathea
- called on all nodes (for parallel jobs)

Magrathea status also contains number of CPUs used by the virtual machine and *length of preemption* – number of seconds jobs were preempted aggregated for each virtual node.



Preemption

Preemption techniques

- complete suspend (freezing)
- reducing memory and CPU

Preemption overhead is critical for the second technique, especially for Xen. While ballooning may be sufficient for simple jobs, it is incredibly slow for intensive computations.

Kernel helper module was created utilizing suspend-to-disk functions for freeing memory which is then returned to Xen.

Test scenario: 14 parallel CPU and memory intensive processes consuming 15GB of memory; preemption swaps out 14.7GB; processes are stopped using SIGSTOP before reducing the memory.

disk speed	2m 50s
magrathea	3m 40s
balloon driver	41m 20s
w/o SIGSTOP	hours

Conclusions

Magrathea provides possibility to run different Linux flavors on one cluster node and switch between them dynamically, gives us possibility to preempt sequential jobs and therefore improves support for large parallel jobs on our cluster, supports several active domains concurrently on VServer, and allows for suspending jobs on Xen enabled clusters.

Magrathea is deployed in production environment on computational nodes of MetaCenter, which provides a computational infrastructure for various groups of users with specific requirements and its resources are also provided for European grid infrastructure EGEE.

This project has been supported by research intents “Optical Network of National Research and Its New Applications” and “Parallel and Distributed Systems” (MŠM 6383917201, MŠM 0021622419).

Ongoing work

Integration with a virtual cluster system for enabling seamless coexistence between virtual clusters and normal jobs and allowing a single scheduler to manage both entities at the same time to achieve better resource efficiency.

References

- [1] Miroslav Ruda and Jiří Denemark and Luděk Matyska. Scheduling Virtual Grids: the Magrathea System. VTDC '07: 3rd international workshop on Virtualization technology in distributed computing. Reno, USA, 2007.