Running EGEE services and worker nodes using virtual machines

M. Ruda, J. Svec
CESNET

What is MetaCenter

- academic grid infrastructure in Czech republic
- consists of centers at different universities
  - Masaryk University in Brno
  - Charles University in Prague
  - West Bohemian university in Pilsen
  - and at CESNET
- hardware – around 750 CPUs
  - mostly XEON/Opteron SMP clusters
  - SGI Altix servers
  - Opteron 16way servers
- dedicated network between sites
  - 10Gbps Ethernet
  - DWDM optical network
- participating in EGEE/EGEE2 with another 250 CPUs

Motivation

- why virtualization?
  - attempt to create IP layer for grid environments
  - sharing of environment control between users/admins

- could enhance MetaCenter (or any grid) in several ways
  - variety of user requirements => several machines with different OS or Linux flavor on the same machine
  - support for various grid environments => possibility to run different images for for different groups, support different grid middlevare
  - migration => better scheduling, robustness
  - suspend/resume => checkpointing, interactive jobs
  - isolation => provide illusion of dedicated cluster

Current usage of virtual machines

- portability tests, running services in different Linux distributions
  - environment for software development
  - portability tests (EGEE LB service)
  - simulation of distributed environment
  - some software may require specific Linux distribution
- server consolidation
- EGEE/MetaCenter consolidation
- job preemption

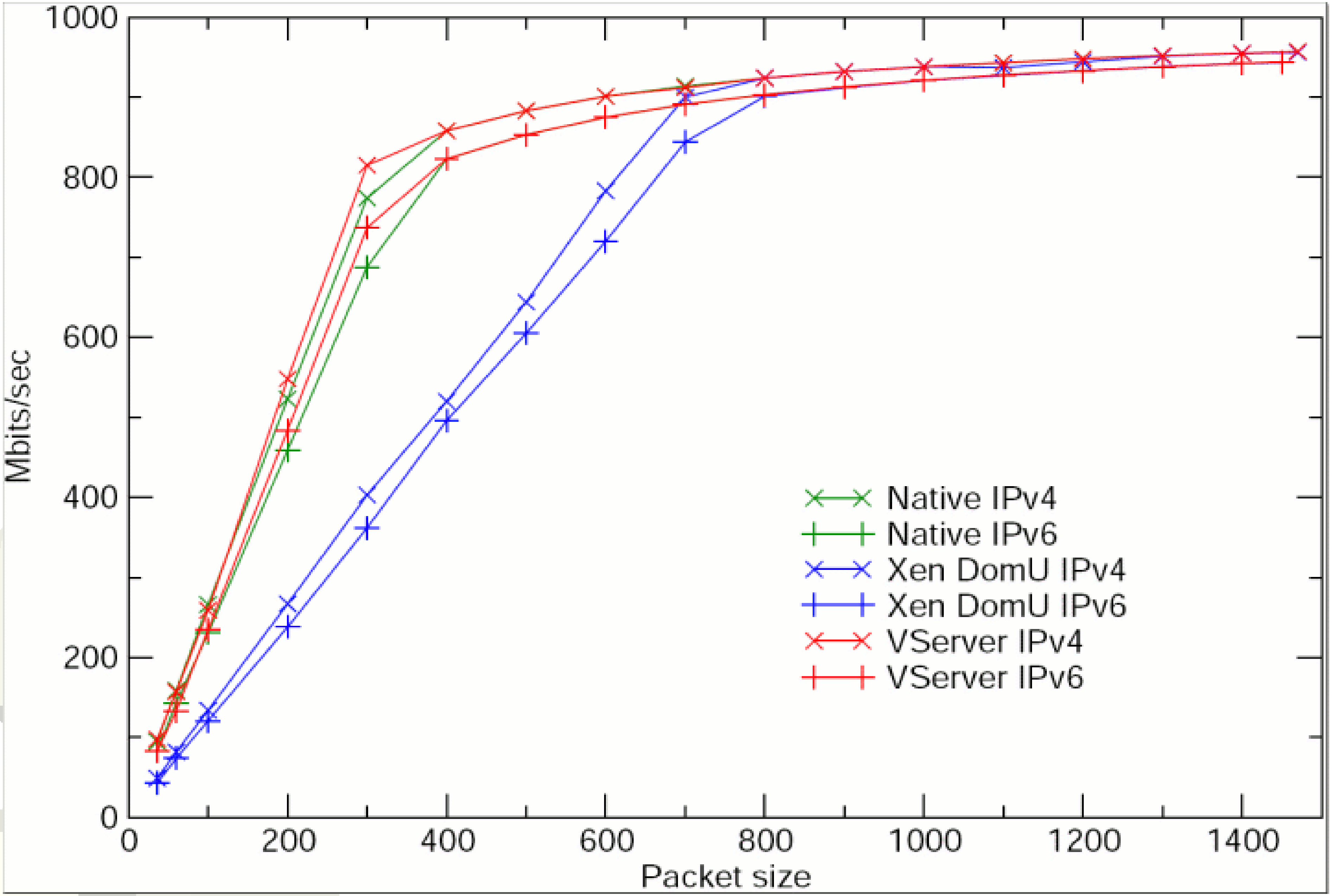
Xen vs. Vserver

- Xen
  - para virtualization due to performance
  - useful for complete encapsulation (user supplied images)
  - support for complete linux distributions
  - perfect solution for service consolidation
  - currently used for EGEE/MetaCenter consolidation
- Vserver
  - one kernel space
  - no performance penalty
  - better memory management, system daemons/services running only once
  - may be better solution for preemption (two domains of the same flavor)

Xen performance results / issues

- good results on small SMP machines / minimal delay for CPU, memory, disk intensive applications
- bad results for fast networks – one CPU is required bridging on full speed 1Gb ethernet
- initial tests with HVM not encouraging
- good results for Infiniband – driver runs only in domU
- lack of IP addresses => IPv6

Network bridging



Xen overhead

- active use of memory
  - dom0
  - every running domU needs at least 100MB of memory
- disk partitions dedicated to different VMs
  - not easy (read only) sharing of root filesystems
  - required splitting of scratch partitions

Server consolidation

- primary motivation – efficient use of hardware
  - EGEE in a box
  - 12 domains running all EGEE services in different VMs (WMS, LB, MyProxy, VOMS, CE, SE, UI...)
  - used for production (prague_cesnet) and pre-production testbed (prague_cesnet_pps), development and testing
  - also used for production WMS for VOCE VO
- DELL PE1950, 2x 3GHz quadcore Xeons, 16GB
- Xen is perfect solution, overhead is minimal
  - all services running all the time, statical splitting of memory is OK
  - root filesystem is different for different domains

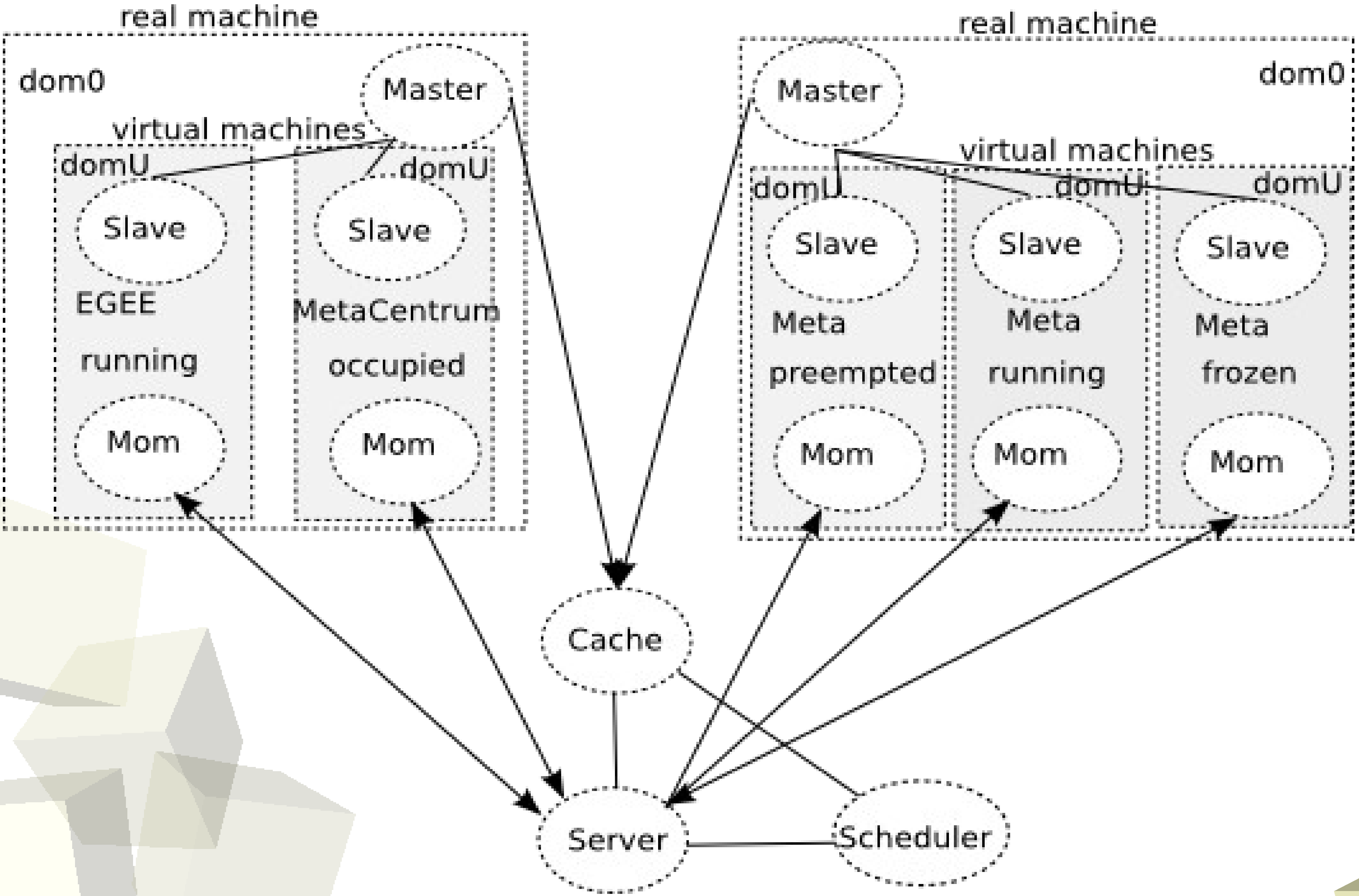
EGEE/MetaCenter Consolidation

- primary motivation – allow coexistence of EGEE and MetaCenter environment
- two images running all the time – Debian/OpenSuse (MetaCenter) and SLC (EGEE)
- EGEE gateway (CE) submits to standard PBS, but to special queue
- dynamic allocation of resources to EGEE and MetaCenter maintained by PBS
- PBS must be aware, that two Vms share the same node, but with minimal changes on PBS side => Magrathea project (more on SC07)
- no changes to EGEE software
- verified on small testbed, just now being deployed on new cluster (10x Altix 310 => 80 cores)

Magrathea

- integrating virtual machines and PBS
  - each node can run several VMs at a time
  - at most one VM on each node is active
  - however, a VM can be activated even if another one is active – preemption
  - active VM is provided with “all” CPU power and memory
- implementation
  - PBS cannot recognize virtual machines from real ones
  - special PBS attribute to distinguish amongst free, running and occupied machines
  - modified PBS scheduler schedules jobs to free machines only
  - current state of VMs is maintained by a daemon running on each physical machine

Magrathea


real machine
dom0
Master
virtual machines
domU
Slave
EGEE
running
Mom
domU
Slave
MetaCentrum
occupied
Mom
real machine
dom0
Master
virtual machines
domU
Slave
Meta
preempted
Mom
domU
Slave
Meta
running
Mom
domU
Slave
Meta
frozen
Mom
Cache
Server
Scheduler

Job Preemption

- first domain available for standard jobs
- second domain available for high priority jobs
- when high priority domain becomes active, almost all CPU/memory resources are given to this domain
- first domain remains alive (PBS monitoring works, no job resubmission)
- jobs in first domain can be suspended by SIGSTOP

- deployed on MetaCentre, cluster of 40 nodes (dual CPU), until the end of 2007 three more clusters (100 nodes) will be deployed too

Scale of deployment

- Current status
  - preemption – 40 nodes
  - Vserver – 2x 16CPU (Opteron)
  - EGEE/Meta consolidation – 10 nodes (2x quad core Xeon each)
  - server consolidation – 2 nodes (=> moving to one 2x quad core Xeon)
- All new clusters will be virtualized
- Experience
  - preemption – since summer 2007
  - server consolidation – more than a year
  - EGEE/Meta consolidation – about a year
  - Vserver – about a year