Virtual Clusters as a New Service of MetaCentrum, the Czech NGI

M. Ruda, Z. Sustr, J. Sitera, D. Antos, L. Hejtmanek, P. Holub

Cesnet
Czech Republic

Krakow, 2009

Outline

- Motivation for virtualization
- Virtual clusters
- Implementation
- Current status, plans

META Centrum

META Centrum (http://meta.cesnet.cz)

- anyone remembers term metacomputing?
- Czech national grid infrastructure
  - under umbrella of Cesnet
- computational resources
  - mostly clusters
  - installed across country, centrally managed
- the same team involved in EGEE
  - computing site, user and VO support, gLite development
- virtualization as one of key research focus

Virtualization

Virtualization of worker nodes:

- run applications with different requirements on the same node
- manage resources given to application
- suspend, preempt, migrate virtual machine
- isolation – provide illusion of dedicated hardware

New scheduling strategies and tools are needed:

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

Magrathea design principles

- virtual machine support in PBSPro
- more than one virtual machines on real machine
  - more that one online in most cases
  - PBS scheduler understands that resources are shared between virtual machinesoverloading real machine
- Magrathea
  - assigns resources (CPU/memory) to active domains according job requirements
  - computes magrathea status of domain, this information is used by scheduler
- minimal changes to resource management system
  - PBS Scheduler (and Server) respects magrathea status
- independent on virtual machine implementation
  - Xen, Vserver
- independent on image management systems
  - domains are managed by external service

Magrathea – Use cases

Development originally concentrated on four basic use cases

1. two static domains
   - only one is allowed to run jobs in any particular time
   - useful for two incompatible images
2. preemption
   - high priority jobs submitted to privileged domain
   - standard domain preempted, but still online
3. more than one active domain
   - CPUs assigned according job requirements
4. domains dedicated to services
   - suspended when service inactive
   - can be reactivated on user request

Virtual cluster

Virtual cluster:

- consists of virtual machines instead of real machines
- hides topology of physical nodes (different clusters)

New use-cases – "cloud inspiration"

5. run job in environment preferred by applications
    - even deploy user supplied environment (OS image)
6. build semi-permanent cluster from such nodes
    - even with it’s own job/account management

But integrated to standard environment

- the same user interface (PBS command line + web)
- cooperation with standard jobs
- the same scheduling policies

On demand node instalation

Run job in preferred environment

- different Linux flavors required by different groups
  - WLCG SLC4/5/6, Debian, RedHat/Suse for commercial applications
- `qsub -l nodes=1:xeon:debian`
- nodes are not pre-installed, instalation on the fly
- extended Magrathea and PBS - bootable nodes
  - down, but Magrathea is able to install different images
- image repository with different Linux images

User supplied images

- Magrathea extended to support "foreign" images
  - nothing installed in OS running inside of virtual machines
- security implications
  - user has root privileges
  - no guarantee for all security patches installed
- inside private network

On demand node instalation

Run job in preferred environment

- different Linux flavors required by different groups
  - WLCG SLC4/5/6, Debian, RedHat/Suse for commercial applications
- `qsub -l nodes=1:xeon:debian`
- nodes are not pre-installed, instalation on the fly
- extended Magrathea and PBS - bootable nodes
  - down, but Magrathea is able to install different images
- image repository with different Linux images

User supplied images

- Magrathea extended to support "foreign" images
  - nothing installed in OS running inside of virtual machines
- security implications
  - user has root privileges
  - no guarantee for all security patches installed
- inside private network

Virtual clusters

Clusters

- `qsub -l cluster=NAME -l nodes=2:debian+4:slc5`
- scheduling very similar to standard parallel job
- when started, user can
  - use ssh to nodes
  - run it's own job management system inside
  - use central PBS instalation to submit jobs into cluster
- both from standard or user supplied images

Authorization (planned)

- user can reboot virtual machine running his job
- user may define group of users allowed to use his cluster (departmental clusters)

Clusters within VPN

Private network

- `qsub -l cluster=NAME,net=private`
- added one service node with DHCP and VPN servers
  - DHCP configuration according user specification
  - authorization on VPN server
  - possible gateways for our services (filesystem)
- VPLS VLANs across Cesnet backbone
- standard openvpn setup, allows
  - NAT mode
  - "tunelling of cluster into user network"

Status and Conclusions

Current status:

- prototype implemented
- in process of deployment for early adopters
- group authorization not implemented
- work done in PBSPro, port for Torque later this year

Next steps:

- production setup available on complete NGI
- integration with cloud interfaces - Globus Workspaces