Improving Best-Effort Job Management in Grids

Jérôme Gallard, Adrien Lèbre, Oana Goga
Paris Project
April 2009
The idea

- Improving "transparent dynamicity" in Grid usage by using latest VM functionalities

- Major issues to design a Grid VM management framework?
Context

- Grid level: Heterogeneous system
  (architecture / environment / network)

- Multiple users
  (developers, administrators, users)

- Several kinds of applications
  (business, scientific, service)
Context

- 'OAR BestEffort' like mode
  - Many jobs are executed in this mode
  - Low level of priority
  - Could be withdrawn from resources at any time
- Important loss of computations
Context

<table>
<thead>
<tr>
<th></th>
<th>Best-Effort Jobs</th>
<th>Killed Best-Effort Jobs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nb</td>
<td>578438</td>
<td>132504</td>
</tr>
<tr>
<td>Days</td>
<td>262907</td>
<td>113353</td>
</tr>
<tr>
<td>Time %</td>
<td></td>
<td>43.12</td>
</tr>
</tbody>
</table>

Grid5000 School 2009 April 2009
Context

<table>
<thead>
<tr>
<th>Best-Effort Jobs</th>
<th>Killed Best-Effort Jobs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nb</td>
<td>Days</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>Total</td>
<td>578438</td>
</tr>
</tbody>
</table>

43,12%
Objectives

➔ Efficiency
   (try to minimize loss of computation)

➔ Performance
   (try to optimize execution time of applications)

➔ Transparency
   (simplicity of use and administration)
   ➔ How virtualization can help us ?
Virtualization

- Isolation
- Server consolidation
- Suspend/Restart
- Migration
Virtualization at grid level

New issues:

→ VM image creation (initial setup)
→ VMM deployment
→ VM image repository
→ VM deployment / job submission
→ VM / job life cycle (migration, suspend/restart ...)
→ Job completion and VM shutdown
Virtualization at grid level

New issues:

- VM image creation (initial setup)
- VMM deployment
- VM image repository
- VM deployment / job submission
- VM / job life cycle (migration, suspend/restart ...)
- Job completion and VM shutdown
Virtualization at grid level

New issues:

- VM image creation (initial setup)
- VMM deployment
- VM image repository
- VM deployment / job submission
- VM / job life cycle (migration, suspend/restart ...)
- Job completion and VM shutdown
VM Storage Management

- Initial deployment
- VM's snapshot management
VM Storage Management

➔ Initial deployment
VM Storage Management

- Initial deployment
- Diskless approach
- Local VM image
- Efficient copy
VM Storage Management

- Initial deployment
- Diskless Approach
Diskless Approach: NFS

VM Storage Management

Frontend + Storage image nodes (Working Environment) + Storage image VMs

Compute nodes

N0 N1 N2

N3 N4 N5
Diskless Approach: NFS

VM Storage Management

Frontend
+ Storage image nodes (Working Environment)
+ Storage image VMs

Compute nodes

Grid5000 School 2009
April 2009
Diskless Approach: NFS

VM Storage Management

- Frontend
  - Storage image nodes (Working Environment)
  - Storage image VMs

Compute nodes

Copy

Grid5000 School 2009

April 2009
Diskless Approach: NFS

VM Storage Management

Frontend

+ Storage image nodes
  (Working Environment)
  + Storage image VMs

Compute nodes

N0

N1

N2

N3

N4

N5

Frontend

Storage image nodes

(Working Environment)

Storage image VMs

Copy

VMM

VMM

VMM

VMM

VMM

Grid5000 School 2009
April 2009
Diskless Approach: NFS

VM Storage Management

Frontend

+ Storage image nodes (Working Environment)

+ Storage image VMs

Compute nodes

Storage space issue?

Bottleneck at server side: scalability issue at grid level!
VM Storage Management

- Initial deployment
- Local VM Image
Local VM Image: all VM Image are already on the compute nodes

VM Storage Management

Frontend + Storage image nodes (Working Environment) + Storage image VMs

Compute nodes
Local VM Image: all VM Image are already on the compute nodes

Current G5K approach?

Frontend
+ Storage image nodes (Working Environment)
  + Storage image VMs

Compute nodes

Grid5000 School 2009

April 2009
VM Storage Management

→ Initial deployment
→ Efficient copy (Taktuk)
VM Storage Management

Frontend
+
Storage image nodes (Working Environment)
+
Storage image VMs

Compute nodes

Efficient Copy
Efficient Copy

VM Storage Management

Frontend
+ Storage image nodes (Working Environment)
+ Storage image VMs

Compute nodes

Grid5000 School 2009
April 2009
VM Storage Management

Efficient Copy

Frontend

+ Storage image nodes (Working Environment)

+ Storage image VMs

Compute nodes

Copy

Grid5000 School 2009

April 2009
VM Storage Management

Frontend

+ Storage image nodes

(Working Environment)

+ Storage image VMs
Preliminary Experiments

Use of TakTuk with a VM image of 330MB

<table>
<thead>
<tr>
<th>nb of nodes</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypervisor deployment</td>
<td>10:31</td>
<td>09:41</td>
<td>12:36</td>
<td>10:27</td>
</tr>
<tr>
<td>VM copy</td>
<td>00:28</td>
<td>00:29</td>
<td>00:35</td>
<td>00:37</td>
</tr>
<tr>
<td>VM configuration</td>
<td>01:31</td>
<td>01:32</td>
<td>01:51</td>
<td>01:59</td>
</tr>
<tr>
<td>Network traffic</td>
<td>347</td>
<td>348</td>
<td>349</td>
<td>351</td>
</tr>
</tbody>
</table>
Preliminary Experiments

Use of TakTuk with a VM image of 330MB

<table>
<thead>
<tr>
<th>nb of nodes node</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypervisor deployment</td>
<td>10:31</td>
<td>09:41</td>
<td>12:36</td>
<td>10:27</td>
</tr>
<tr>
<td>VM copy</td>
<td>00:28</td>
<td>00:29</td>
<td>00:35</td>
<td>00:37</td>
</tr>
<tr>
<td>VM configuration</td>
<td>01:31</td>
<td>01:32</td>
<td>01:51</td>
<td>01:59</td>
</tr>
<tr>
<td>Network traffic</td>
<td>347</td>
<td>348</td>
<td>349</td>
<td>351</td>
</tr>
</tbody>
</table>

VM copy scalability issue seems to be solved!
VM Storage Management

- VM's snapshot management
- Copy-On-Write
VM Storage Management

→ VM's snapshot management
→ Copy-On-Write

later...
VM Storage Management

- VM's snapshot management
- Copy-On-Write

later...

diff file
to save
VM Storage Management

- VM's snapshot management
  - Using a distributed file system
  - Saving snapshot locally
  - Copy from all nodes to a server repository
VM Storage Management

- VM's snapshot management
- Using a distributed file system
VM Storage Management

Frontend
+
Storage image nodes
(Working Environment)
+
Storage image VMs

VM's snapshot: NFS

Compute nodes
VM Storage Management

VM's snapshot: NFS

Frontend

+ Storage image nodes (Working Environment)

+ Storage image VMs

Compute nodes

VMM

VMM

VMM

VMM

VMM

VMM

VMM

VMM

Grid5000 School 2009

April 2009
VM Storage Management

Frontend

+ Storage image nodes (Working Environment)
+ Storage image VMs

Compute nodes

Bottleneck at server side: scalability issue at grid level!

VM's snapshot: NFS

Storage space issue
VM Storage Management

- VM's snapshot management
- Saving Snapshot locally
VM's snapshot: NFS

VM Storage Management

Frontend

+ Storage image nodes (working environment)

+ Storage image VMs

Compute nodes
VM's snapshot: NFS

VM Storage Management

Compute nodes

Frontend

+ Storage image nodes (working environment)

+ Storage image VMs

Grid5000 School 2009  April 2009
VM's snapshot: NFS

VM Storage Management

Frontend
+ Storage image nodes (working environment)
+ Storage image VMs

Compute nodes

Grid5000 School 2009
April 2009
VM Storage Management

Frontend
+ Storage image nodes (working environment)
+ Storage image VMs

Compute nodes

VM's snapshot: NFS
VM's snapshot: NFS

VM Storage Management

Frontend
+ Storage image nodes (working environment) + Storage image VMs

Compute nodes
VM Storage Management

Frontend
+ Storage image nodes (Working Environment)
+ Storage image VMs

Compute nodes

VM's snapshot: NFS
VM storage Management

- VM's snapshot management
- Copy from all nodes to a server repository
From all nodes to one node

VM Storage Management

Frontend

+ Storage image nodes (Working Environment)

+ Storage image VMs

Compute nodes

Storage space issue?

Bottleneck at server side: scalability issue at grid level!
Preliminary Experiments

Use of kaget with a VM image of 330MB

<table>
<thead>
<tr>
<th>nb of nodes</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Snapshot Time</td>
<td>02:01</td>
<td>02:59</td>
<td>04:09</td>
<td>07:34</td>
</tr>
<tr>
<td>Network traffic</td>
<td>674</td>
<td>1387</td>
<td>2779</td>
<td>5557</td>
</tr>
</tbody>
</table>
Preliminary Experiments

Use of kaget with a VM image of 330MB

The more VM images we have, the longer is the required time.
Network Configuration and Mobility

- VM network card configuration
- VM migration at Grid level
Network Configuration and Mobility

→ VM network card configuration
   → Central DHCP Server
   → Hybrid DHCP Server
Network Configuration and Mobility

- VM network card configuration
- Central DHCP Server
Seems to be technically impossible to solve:
- MAC and IP address issues
Network Configuration and Mobility

- VM network card configuration
- Hybrid DHCP Server
Network Configuration and Mobility

- VM network card configuration
- Hybrid DHCP Server
Network Configuration and Mobility

- VM network card configuration
- Hybrid DHCP Server
Network Configuration and Mobility

- VM network card configuration
- Hybrid DHCP Server

Node1 (with VMM)

Switch for physical network

SubNet

Master Node2

eth0

ethvirt0
Network Configuration and Mobility

- VM network card configuration
- Hybrid DHCP Server

Node1
- eth0
- ethvirt0

VM1
- eth0
- eth1

(Node1 running on Node1)

Switch for physical network

Master Node2
- eth0
- ethvirt0

SubNet

VM1 (running on Node1) ➔ Hybrid DHCP Server ➔ Network Configuration and Mobility

Grid5000 School 2009

April 2009
Network Configuration and Mobility

- VM network card configuration
- Hybrid DHCP Server
Network Configuration and Mobility

- VM network card configuration
- Hybrid DHCP Server

- Node1 (with VMM)
  - eth0
  - ethvirt0
- VM1 (running on Node1)
  - eth0
  - eth1
- Switch for physical network
- SubNet
- Master Node2
  - eth0
  - ethvirt0
Network Configuration and Mobility

- VM network card configuration
- Hybrid DHCP Server
Network Configuration and Mobility

- VM migration at Grid level
  - Mobile IP
    - Add-ons with IPV4 or natively with IPV6
Network Configuration and Mobility

- VM migration at Grid level
- VLAN

  Complete level isolation at level 2

  Kavlan status? (talk tomorrow morning :-) )
Conclusion

By using snapshoting and migration capabilities of VM, users can exploit the grid in a more transparent and dynamic way.

However several issues have still to be solved...