LinuxConf Europe 2007

Ganeti
an open source multi-node high-availability cluster based on Xen

Roman Marxer
Content

• Traditional vs. virtualization-based high-availability clusters
• Ganeti overview and administration
• Ganeti disk details and recovery
• Design goals and principles
• Ganeti usage in Google
• Ganeti code
• Roadmap
Traditional high-availability cluster

application failover
(data, IP)

application data, SW

system 1

OS disk

heartbeat

system 1

OS disk

application
Virtualization-based multi-node high-availability cluster

master: tests and control

virtual system failover
Ganeti overview (1/2)

• Xen cluster manager
• n-node high-availability cluster (future)
• software used
  ▪ virtualization: Xen
  ▪ disk management: LVM / DRBD / MD
  ▪ language and RPC: Python / Twisted
Ganeti overview (2/2)

Ganeti cluster

failover

Xen dom0 = node
Xen domU = instance

virt. system 1

virt. system 1'

system 1 (dom0)

system 2 (dom0)

system 3 (dom0)

Ganeti master

... more

Ganeti cluster
Ganeti administration (1/3)

• Ganeti master (special role of a node)
• gnt-node: add / remove / list cluster nodes
• gnt-instance:
  ▪ add / remove instance
  ▪ failover instance, change secondary
  ▪ stop / start instance, change parameters
• gnt-os: instance OS definitions
• gnt-cluster: cluster commands
Ganeti administration (2/3)

- `gnt-instance add` *(instance creation)*
  - parameters: name, disk size, RAM size, OS type, disk layout, primary / secondary node

- OS creation script
  - script to install an OS in a partition
    - e.g. debootstrap and additional scripts
  - set IP address and hostname of the image
  - copy image and change
# gnt-instance list

<table>
<thead>
<tr>
<th>Instance</th>
<th>OS</th>
<th>Primary_node</th>
<th>Autostart</th>
<th>Status</th>
<th>Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>instance1.example.com</td>
<td>etch</td>
<td>node1.example.com</td>
<td>yes</td>
<td>running</td>
<td>128</td>
</tr>
<tr>
<td>instance2.example.com</td>
<td>etch</td>
<td>node3.example.com</td>
<td>yes</td>
<td>running</td>
<td>512</td>
</tr>
<tr>
<td>instance3.example.com</td>
<td>etch</td>
<td>node3.example.com</td>
<td>yes</td>
<td>running</td>
<td>1024</td>
</tr>
<tr>
<td>instance4.example.com</td>
<td>etch</td>
<td>node2.example.com</td>
<td>yes</td>
<td>running</td>
<td>128</td>
</tr>
<tr>
<td>instance5.example.com</td>
<td>etch</td>
<td>node4.example.com</td>
<td>yes</td>
<td>running</td>
<td>512</td>
</tr>
</tbody>
</table>

# gnt-node list

<table>
<thead>
<tr>
<th>Node</th>
<th>DTotal</th>
<th>DFree</th>
<th>MTotal</th>
<th>MNode</th>
<th>MFree</th>
<th>Pinst</th>
<th>Sinst</th>
</tr>
</thead>
<tbody>
<tr>
<td>node1.example.com</td>
<td>858240</td>
<td>442752</td>
<td>4095</td>
<td>511</td>
<td>3456</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>node2.example.com</td>
<td>572160</td>
<td>567296</td>
<td>4095</td>
<td>511</td>
<td>3456</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>node3.example.com</td>
<td>858240</td>
<td>858240</td>
<td>4095</td>
<td>511</td>
<td>2048</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>node4.example.com</td>
<td>356032</td>
<td>356032</td>
<td>4095</td>
<td>511</td>
<td>3072</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Ganeti disk details

- disk types
  - plain
  - local_raid1
  - remote_raid1

remote_raid1 details

Instance disk

MD device

DRBD device

LVM logical volume

Physical disks

Node 1

Node 2
remote_raid1 failover

1. dark blue DRDB set serves data
2. node fails in dark blue DRDB set
3. admin: gnt-instance replace-disks
4. light blue DRDB set gets added and is synchronized
5. dark blue DRDB set gets removed
Design goals and principles

• goals
  ▪ increase availability
  ▪ reduce hardware cost
  ▪ increase flexibility
  ▪ transparency

• principles
  ▪ not dependent on specific hardware (e.g. SAN)
  ▪ scales linearly with the number of systems
Ganeti usage in Google

- 20-node Ganeti cluster
- 64-bit node OS
- 80 virtual instances
- used for internal systems
- not used for google.com
- good for non-resource intensive systems
Ganeti code

• developed at Google
• license: GPL v2
• code location: http://code.google.com/p/ganeti/
• August 2007
  ▪ beta release and open source
• November 2007
  ▪ release v.1.2
  ▪ development contributions possible
Roadmap

- HVM integration (Windows support)
- automatic instance failover / node allocation
- master node election
- manager GUI / instance allocator

![Diagram showing 20-node Ganeti cluster in location A and 30-node Ganeti cluster in location B, connected to a manager and allocator.](image-url)
Thank You!

Q&A