Heart attack Virtualization
An HA virtualization cluster based on Pacemaker, Corosync, Xen and DRBD

The Challenge

Servers deployment should be decoupled from hardware setup and operative system installation. It should be possible to manage startup/shutdown/restart operations of building a new server farm for the Computing Center of the Faculty of Engineers.

Virtualization platform requirements

- Hardware access: virtual instances must have direct access to hardware (CPU, IO, NIC, etc.)
- Flexibility: it must be possible to assign different resources to each virtual instance.
- Reliability: virtual servers should reach near-metal performances.
- Security: the virtual platform should implement high availability technologies.

The Computing Center of the Faculty of Engineer manages the network. The area of the Faculty of Engineer's campus network, its buildings, its departments, and the resources of the faculty's server farm are a unique challenge. Our solution is managed by the Corosync/Pacemaker cluster technology. This solution is managed by the Corosync/Pacemaker cluster technology.

Virtualization platform solutions

- Xen virtual machine are plain old Xen instances, and are still managed by xen tools even after they are configured as cluster resources;
- DRBD volumes are builded on custom logical volumes, but are fully manageable with dbrd tools;

The main benefit of this solution is that, even if every component is fully customizable, it is still standard:

- No lock-in and no constraints

The 2009 server farm

The 2009 server farm consists of 16 nodes: each node has two quad-core AMD Opteron 2390 boxes equipped with 16 GB of RAM and 500 GB of storage, running two Linux systems: one for the data center, the other for the application server.

The Solution

Materials

Hardware

- vnodes: Dual GigaTron X8 - 16 Gbyte RAM
- Dual Xeon 3.2GHz 64bit Linux
- Dual Quadcore Opten - 64 Gbyte RAM
- 1 x Quadcore Opten - 32 Gbyte RAM
- 4 x 500 Gbyte SATA 7.2k rpm HDD

Software

- Debian (inside Xen)
- Linux Kernel 2.6.13 (inside Xen)
- Xen 3.2.1
- Corosync 1.1.3 (cluster engine)
- Pacemaker 1.1.8 (cluster resource manager)
- DRBD 3.7.1
- LVM 2

Cluster Network Layout

We set up three layers for the cluster:

- The network layer: this is the server network, shared by all the virtual instances.
- The DRBD layer: dedicated to DRBD synchronization.
- The XCS layer: dedicated to XCS devices access.

Cluster Storage

On each server we created two LVM volume groups where to put logical volumes for virtual instances. On vnodes we replicated every logical volume that belongs to virtual instances.

The make and after完整的XCS target that each node can subscribe is:

- DRBD 8.3.7
- Xen 3.2.1
- LVM 2

Cluster Logical Layout

The cluster communication is based on Corosync (the red box) on top of the node. Pacemaker (the blue box) manages the cluster configuration and communication layer is between the Corosync layer and the virtual instances. Virtual instances are configured on a single node, and display virtual instance configurations in the form displayed on a single node.

Virtual instances are plain old virtual machines. The main benefit of this solution is that it is a fully customizable, but it is still standard.

Conclusions

The new serverfarm reached all the desired target:

- hardware usage consolidation
- decoupling server deployment
- simplicitiy servers management

The 2009 server farm presents some drawbacks:

- Unibiquitous management for free
- Pacemaker configuration and management is powerful and simple, but best of all it does not even require a server, since each node runs the cluster engine management daemon.

Simpliciy servers deployment

We can setup a new instances in minutes using predefined OS images.

The 2009 server farm provides a lot of redundancy for virtual instances, and we have four main nodes running twenty virtualized instances:

- we can manage and monitor each server from one coherent console;
- we can start an instance if the DRBD device is already primary on demand, and it also prevent split brain refusing to start an instance if the DRBD device is already primary on another physical node.

Cloud Computing: Thans, But No Thanks

Cloud Computing solutions are becoming more and more familiar in the academia. The on demand delivery of computing resources paradigm is easy to reach at a community of users well satisfied from the application systems. This is not the case for some scenarios where such a computing model is not adequate.

The question is if it is also a solid answer for the services that provide any resource.

References

- Davide Vaghetti davide.vaghetti@ing.unipi.it
- Marco Tomassetti m.tomassetti@ing.unipi.it
- Sergio Pavesi Sergio.Pavesi@ing.unipi.it
- Davide Mandelli davide.mandelli@ing.unipi.it

The new serverfarm reached all the desired target:

- hardware usage consolidation
- decoupling server deployment
- simplicitiy servers management

The 2009 server farm presents some drawbacks:

- Unibiquitous management for free
- Pacemaker configuration and management is powerful and simple, but best of all it does not even require a server, since each node runs the cluster engine management daemon.

Simpliciy servers deployment

We can setup a new instances in minutes using predefined OS images.

The 2009 server farm provides a lot of redundancy for virtual instances, and we have four main nodes running twenty virtualized instances:

- we can manage and monitor each server from one coherent console;
- we can start an instance if the DRBD device is already primary on demand, and it also prevent split brain refusing to start an instance if the DRBD device is already primary on another physical node.

Cloud Computing: Thans, But No Thanks

Cloud Computing solutions are becoming more and more familiar in the academia. The on demand delivery of computing resources paradigm is easy to reach at a community of users well satisfied from the application systems. This is not the case for some scenarios where such a computing model is not adequate.

The question is if it is also a solid answer for the services that provide any resource.

References

- Davide Vaghetti davide.vaghetti@ing.unipi.it
- Marco Tomassetti m.tomassetti@ing.unipi.it
- Sergio Pavesi Sergio.Pavesi@ing.unipi.it
- Davide Mandelli davide.mandelli@ing.unipi.it

The new serverfarm reached all the desired target:

- hardware usage consolidation
- decoupling server deployment
- simplicitiy servers management

The 2009 server farm presents some drawbacks:

- Unibiquitous management for free
- Pacemaker configuration and management is powerful and simple, but best of all it does not even require a server, since each node runs the cluster engine management daemon.

Simpliciy servers deployment

We can setup a new instances in minutes using predefined OS images.

The 2009 server farm provides a lot of redundancy for virtual instances, and we have four main nodes running twenty virtualized instances:

- we can manage and monitor each server from one coherent console;
- we can start an instance if the DRBD device is already primary on demand, and it also prevent split brain refusing to start an instance if the DRBD device is already primary on another physical node.

Cloud Computing: Thans, But No Thanks

Cloud Computing solutions are becoming more and more familiar in the academia. The on demand delivery of computing resources paradigm is easy to reach at a community of users well satisfied from the application systems. This is not the case for some scenarios where such a computing model is not adequate.

The question is if it is also a solid answer for the services that provide any resource.

References

- Davide Vaghetti davide.vaghetti@ing.unipi.it
- Marco Tomassetti m.tomassetti@ing.unipi.it
- Sergio Pavesi Sergio.Pavesi@ing.unipi.it
- Davide Mandelli davide.mandelli@ing.unipi.it

The new serverfarm reached all the desired target:

- hardware usage consolidation
- decoupling server deployment
- simplicitiy servers management

The 2009 server farm presents some drawbacks:

- Unibiquitous management for free
- Pacemaker configuration and management is powerful and simple, but best of all it does not even require a server, since each node runs the cluster engine management daemon.

Simpliciy servers deployment

We can setup a new instances in minutes using predefined OS images.

The 2009 server farm provides a lot of redundancy for virtual instances, and we have four main nodes running twenty virtualized instances:

- we can manage and monitor each server from one coherent console;
- we can start an instance if the DRBD device is already primary on demand, and it also prevent split brain refusing to start an instance if the DRBD device is already primary on another physical node.