Antibrouillards dans le Cloud

Jonathan Schaeffer
jonathan.schaeffer@univ-brest.fr

9 Juin 2011

Cluster Shared Disk

Definition

Une grappe de machines ayant un accès total à l'ensemble des disques hébergeant les données du cluster.

Dans le cadre d'un cloud, les Machines Virtuelles sont toutes indépendantes du serveur physique

Objectifs et Contraintes

Objectifs

- Virtualiser les services
- Supporter une panne sur les serveurs physiques
- Interventions transparentes sur le serveur physique

Objectifs et Contraintes

Objectifs

- Virtualiser les services
- Supporter une panne sur les serveurs physiques
- Interventions transparentes sur le serveur physique

Contraintes

- Budget pour un seul serveur
- Second serveur existant (DeLL R710)

Prérequis pour un cloud

- Quantité de mémoire vive suffisante pour l'ensemble des VMs (ksm depuis kernel 2.6.32)
- Accès direct simultané au même stockage
- Performance en accès disques
- Type de processeur physique identique sur tous les serveurs physiques

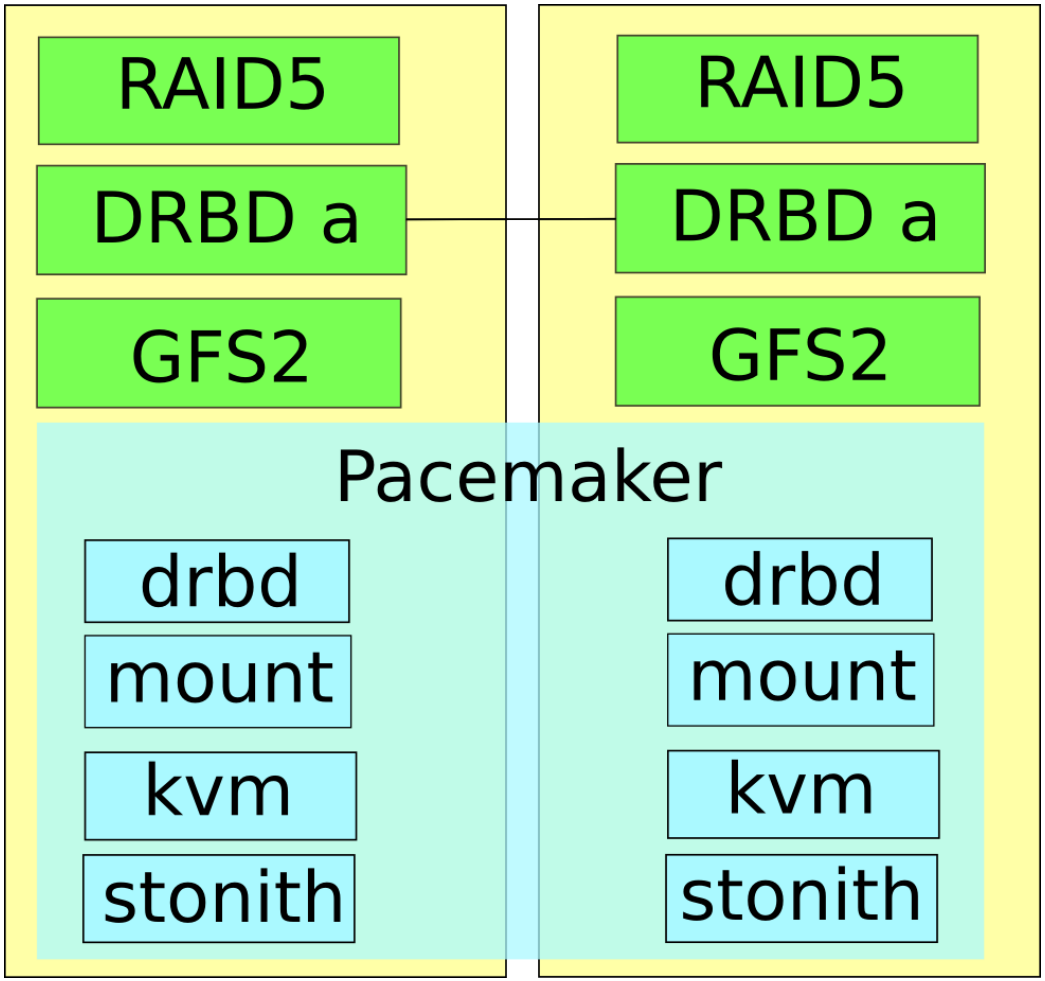
Solution matérielle

- 2x Serveurs DeLL R710
- Interface réseau directe *bonded*
- 2x 3 disques de 500G 15kt/m
- 2x 32G RAM

Solution logicielle

- Système ubuntu LTS 10.04.1
- DRBD Active/Active
- Système de fichier distribué Clustered LVM + GFS2
- Pacemaker
- KVM + libvirt

Architecture initiale



configuration de pacemaker

- Primitives
  - DRBD
  - GFS2 (lock daemon et montage)
  - pour chaque VM
- Caractéristiques
  - Ordre de démarrage et dépendance
  - fréquence de test
  - action en cas d'échec
  - localisation du service

Exemples de primitives

```
primitive drbd-data ocf:linbit:drbd \
    params drbd_resource="orque-raid" \
    op monitor interval="60s" \
    op start interval="0" timeout="240s" \
    op stop interval="0" timeout="100s"
ms drbd-data-clone drbd-data \
    meta master-max="2" master-node-max="1" \
    clone-max="2" clone-node-max="1" notify="true"
```

Exemples de primitives

```
orque ~ # crm status
==============
Last updated: Thu Mar  3 15:56:19 2011
Stack: openais
Current DC: orque - partition with quorum
Version: 1.0.8-042548a451fce8400660f6031f4da6f0223dd5dd
2 Nodes configured, 2 expected votes
7 Resources configured.
==============

Online: [ orque2 orque ]

 Master/Slave Set: drbd-data-clone
     Masters: [ orque2 orque ]
```

Exemples de primitives

```
primitive dlm ocf:pacemaker:controld \
        op monitor interval="120s" \
        op start interval="0" timeout="90s" \
        op stop interval="0" timeout="100s"
primitive gfs-control ocf:pacemaker:controld \
        params daemon="gfs_controld.pcmk" args="-g 0" \
        op monitor interval="120s" \
        op start interval="0" timeout="90s" \
        op stop interval="0" timeout="100s"
clone dlm-clone dlm \
        meta interleave="true" target-role="Started"
clone gfs-clone gfs-control \
        meta interleave="true" target-role="Started"
colocation gfs-with-dlm inf: gfs-clone dlm-clone
order gfs-after-dlm inf: dlm-clone gfs-clone
```

Exemples de primitives

```
orque2 ~ # crm status
==============
Last updated: Thu Mar  3 16:07:03 2011
Stack: openais
Current DC: orque - partition with quorum
Version: 1.0.8-042548a451fce8400660f6031f4da6f0223dd5dd
2 Nodes configured, 2 expected votes
7 Resources configured.
==============

Online: [ orque2 orque ]

 Master/Slave Set: drbd-data-clone
     Masters: [ orque2 orque ]
 Clone Set: dlm-clone
     Started: [ orque2 orque ]
 Clone Set: gfs-clone
     Started: [ orque2 orque ]
```

STONITH

Prévenir les catastrophes de "Split Brain"
Méthodes de STONITH :

- shutdown par SSH
- coupure electrique par ipmi
- opérateur manuel "meatware"

Quorum

Déterminer quelle machine est HS

- Défi (monter un filesystem, atteindre un serveur tierce)
- Ajouter un troisième serveur dans le cluster

difficultés de Pacemaker

- Avec ou sans STONITH ?
- Sensibilité du paramétrage
- Bugs dans des couches diverses (GFS2, libvirt)

Bonus : synchronisation de libvirt par GIT

Dépot GIT : /etc/libvirt/qemu

```
root@orque:/etc/libvirt/qemu# git remote show origin
* remote origin
  Fetch URL: orque2:/etc/libvirt/qemu
  Push  URL: orque2:/etc/libvirt/qemu
```

```
root@orque2:/etc/libvirt/qemu# git remote show origin
* remote origin
  Fetch URL: orque:/etc/libvirt/qemu
  Push  URL: orque:/etc/libvirt/qemu
  HEAD branch: master
```

Après une modif sur orque2 :

```
root@orque2:/etc/libvirt/qemu#  ssh orque \
    "cd /etc/libvirt/qemu && git pull"
```

Avantages et Inconvénients

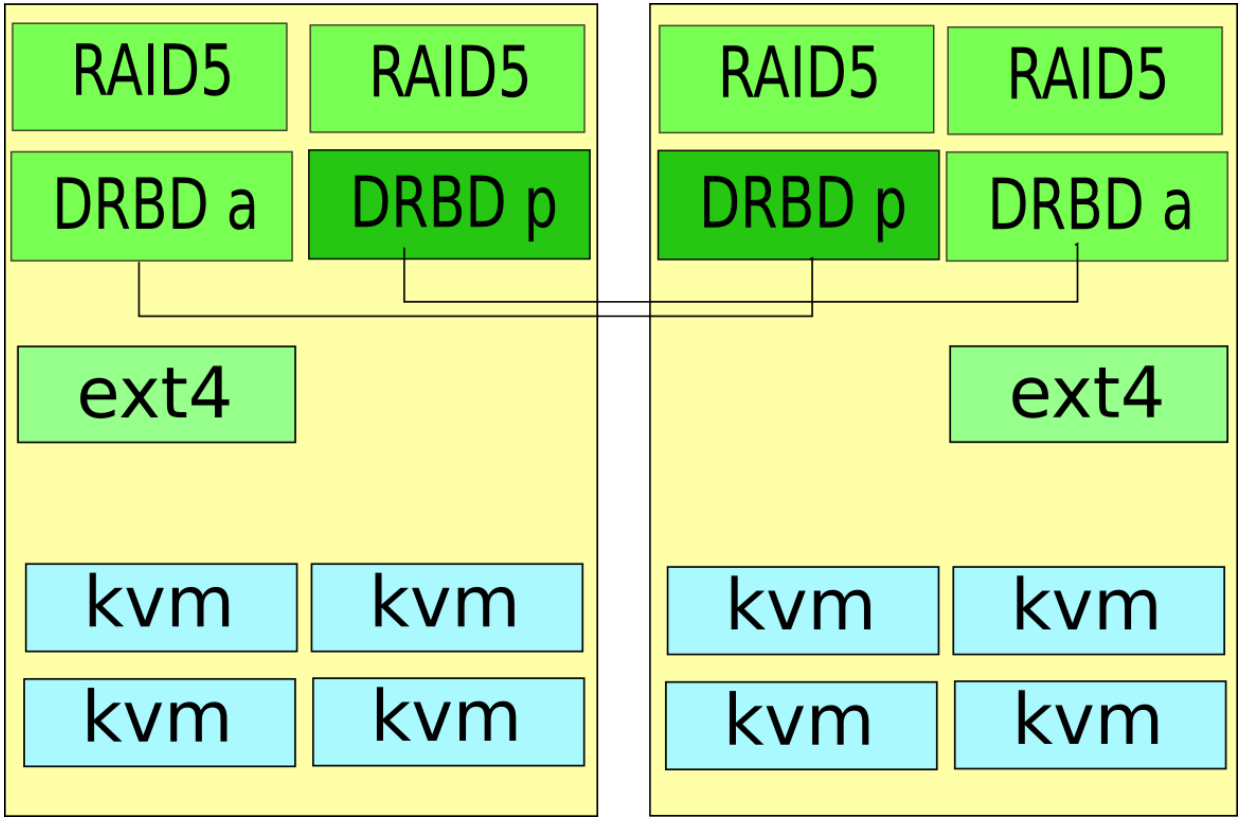
Avantages

- Sécurité des données (RAID 5+0)
- Souplesse de l'architecture (migration à chaud)
- Redondance de tous les points de fragilité

Inconvénients

- Instabilité logicielle
- Configuration très délicate

Environnement de production



- Scientific Linux 6
- Autre système de fichier ? (Gluster, Ceph, ...)

Références

- Linux HA : www.linux-ha.org/
- Guide pacemaker http://www.clusterlabs.org/doc/en-US/Pacemaker/1.0/html/Pacemaker_Explained/
- Cluster From Scratch : http://www.clusterlabs.org/doc/Cluster_from_Scratch.pdf