

Installation de Redhat Cluster Suite

sur CentOS 5.4 avec cluster actif/passif d'Oracle 11gR2

1	PREREQUIS.....	3
1.1	OPERATING SYSTEM	3
1.2	MATERIEL	3
1.3	RESEAU	3
2	INSTALLATION DES PACKAGES ET CONFIGURATION SYSTEME.....	4
2.1	PARAMETRAGE DU FICHIER HOSTS ET/OU DU DNS.....	4
2.2	CREER UNE INTERFACE BOND0 AGREGEANT DEUX INTERFACES ETHERNET.....	4
2.3	CONFIGURATION DU DEPOT RPM.....	5
2.4	INSTALLER/CONFIGURER LES OUTILS CLUSTERS SUR LES DEUX NOEUDS	5
2.5	CREATION ET CONFIGURATION DU CLUSTER	6
	2.5.1 <i>Création du cluster</i>	6
	2.5.2 <i>Ajout des logs</i>	7
	2.5.3 <i>Ajout du fence_vmware</i>	8
	2.5.4 <i>Création du quorum disk</i>	9
	2.5.5 <i>Ajout d'une heuristique</i>	9
3	MISE EN PLACE D'UN VOLUME GFS	10
3.1	PROBLEMATIQUE	10
3.2	CREATION DU VOLUME	11
3.3	MONTAGE DU VOLUME	11
4	INSTALLATION D'UNE BASE ORACLE.....	12
4.1	PREREQUIS	12
4.2	INSTALLATION DES BINAIRES ORACLE	12
4.3	CONFIGURATION ORACLE	13
	4.3.1 <i>Configuration listener.ora et tnsnames.ora</i>	13
	4.3.2 <i>Création de l'instance</i>	13
	4.3.3 <i>Création des scripts d'arrêt relance</i>	13
	4.3.4 <i>Mise en conformité du 2ème nœud</i>	14
4.4	AUTRES ELEMENTS IMPORTANTS	14
5	CREATION DU SERVICE ORACLE DANS LE CLUSTER	15
5.1	CREATION D'UNE RESSOURCE IP VIRTUELLE.....	15
5.2	CREATION D'UN DOMAINE DE FAILOVER.....	15
5.3	CREATION D'UN SERVICE CLUSTER	15
6	ANNEXE	17
6.1	TEST DE FONCTIONNEMENT.....	17
	6.1.1 <i>Démarrage du cluster</i>	17
	6.1.2 <i>Arrêt du nœud actif</i>	17
	6.1.3 <i>Reboot du nœud actif via luci</i>	17
	6.1.4 <i>Arrêt / démarrage du service via luci</i>	17
	6.1.5 <i>Arrêt manuel du service</i>	18
	6.1.6 <i>Relocate du service via luci</i>	18
	6.1.7 <i>Blocage temporaire d'un nœud</i>	18
	6.1.8 <i>Arrêt brutal non planifié d'un nœud</i>	18
	6.1.9 <i>Fencing via luci</i>	18
	6.1.10 <i>Coupure d'une interface réseau</i>	18
	6.1.11 <i>Coupure totale du réseau</i>	18
	6.1.12 <i>Perte du quorum disk sur un nœud</i>	19
6.2	SOURCES	19
	6.2.1 <i>Documentation officielle Redhat</i>	19
	6.2.2 <i>Autre</i>	19

1 Prérequis

1.1 Operating System

Pour réaliser les tests, il est préférable d'utiliser CentOS 5.4 version 64bits, qui permet de disposer du logiciel Redhat Cluster Suite dans son intégralité sans avoir besoin de s'acquitter des droits de licences de Redhat et des composants additionnels nécessaires.

Un environnement de production devra lui être installé avec Redhat Enterprise Linux sous licences disposant des modules additionnels nécessaires.

1.2 Matériel

La plateforme de développement a été créé à partir de machine virtuelles VMware. Ces machines disposent chacune :

- D'un disque virtuel d'au moins 20Go
- De plus d'1.1 Go de RAM (prérequis Oracle 11gR2)
- D'une interface réseau ethernet 1Gbit/s
- D'un accès à deux LUN partagés par les deux nœuds publiés depuis le SAN sur le serveur ESX (un de 100 Mo et un de 20 Go)

Sur la plateforme de production, il est nécessaire de disposer en plus :

- D'une seconde interface réseau pour agréger l'interface principale, de manière à prévenir les pannes réseau
- D'un accès direct au SAN (carte HBA) pour accéder au LUN, si possible avec de la redondance de fibre (volumes présentés en mode « multipath »)
- D'un « fencing device » ayant pour but d'empêcher un serveur considéré comme défaillant d'écrire sur le SAN (alimentation pilotée à distance, switch fibre piloté à distance, port console ILO)

1.3 Réseau

Le but du cluster étant de permettre de réduire voire faire disparaître les temps d'inaccessibilité aux services (ici Oracle), il est nécessaire de disposer d'une adresse IP virtuelle supplémentaire, depuis laquelle on attaquera le service, quel que soit le nœud actif

```
Nœud 1 :  
Hostname.....tstclstr-node1  
IP ..... 192.168.0.100
```

Nœud 2 :
Hostname.....tstclstr-node2
IP 192.168.0.101

Cluster :
Virtual hostname tstclstr
IP virtuelle du cluster..... 192.168.0.200

2 Installation des packages et configuration système

2.1 Paramétrage du fichier hosts et/ou du DNS

Vérifier que le fichier /etc/hosts est bien configuré avec toutes les IP sur les deux nœuds. Attention, l'entrée « localhost » ne doit surtout pas contenir le hostname, car cela peut générer des conflits avec la suite logicielle.

Pour fonctionner correctement, il est également nécessaire d'affecter un FQDN aux serveurs utilisés. A défaut, il est possible d'utiliser « .localdomain »

```
127.0.0.1      localhost.localdomain localhost  
::1           localhost6.localdomain6 localhost6  
192.168.0.100  tstclstr-node1 tstclstr-node1.localdomain  
192.168.0.101  tstclstr-node2 tstclstr-node2.localdomain  
192.168.0.200  tstclstr
```

2.2 Créer une interface bond0 agrégeant deux interfaces ethernet

Pour réduire au maximum le nombre de SPF (single point of failure), on utilise fréquemment deux interfaces réseaux, si possible sur des cartes physiques différentes, et on les agrège en une seule interface. L'intérêt est double puisqu'en plus de prévenir une panne matérielle, on peut aussi répartir le trafic sur les deux interfaces et ainsi multiplier le débit ethernet par deux.

Créer le fichier /etc/sysconfig/network-scripts/ifcfg-bond0

```
DEVICE=bond0  
IPADDR=192.168.0.10X #0 OU 1 POUR NODE1 ET NODE2  
NETMASK=255.255.255.0  
NETWORK=192.168.0.0  
BROADCAST=192.168.0.255  
ONBOOT=yes  
BOOTPROTO=none  
USERCTL=no  
BONDING_OPTS='miimon=100 mode=0'  
GATEWAY=192.168.0.254
```

TYPE=Ethernet

Modifier les fichiers /etc/sysconfig/network-scripts/ifcfg-eth0 et ...eth1

```
DEVICE=eth0 (ou eth1)
USERCTL=no
BOOTPROTO=none
MASTER=bond0
SLAVE=yes
HWADDR=[laisser l'adresse MAC telle quelle]
ONBOOT=yes
TYPE=Ethernet
```

Ajouter la ligne suivante au fichier /etc/modprobe.conf

```
alias bond0 bonding
```

2.3 Configuration du dépôt RPM

Dans le cas où les dépôts Internet ne sont pas disponibles, il est possible d'utiliser le DVD ROM d'installation de CentOS, en désactivant tous les repositories dans /etc/yum.repos.d sauf CentOS-Media.repo

2.4 Installer/configurer les outils clusters sur les deux noeuds

Pour faciliter l'installation, il est possible d'installer les groupes Clustering et Cluster Storage qui contiennent toutes les dépendances nécessaire pour la mise en place d'un cluster

```
yum -y groupinstall Clustering
yum -y groupinstall "Cluster Storage"
```

Pour piloter de manière simple le cluster, RHCS propose deux agents qui doivent être installés et configurés sur les serveurs du cluster. Luci est le démon qui gère le serveur web permettant d'interagir avec le cluster et ricci est le client permettant de transmettre les ordres aux nœuds. La présence de ricci est donc obligatoire sur tous les serveurs du nœud, et luci n'est obligatoire que sur au moins un nœud.

Configurer luci sur le nœud d'administration

```
luci_admin init
#entrer un mot de passe de l'administrateur « admin »
chkconfig luci on
service luci start
```

Configurer le niveau de sécurité du serveur pour qu'il puisse fonctionner. Le niveau de sécurité maximum de SELinux est Permissif (ou Désactivé) pour un cluster Linux/Oracle, et des règles doivent être ajoutées. Pour IPTables, il faut autoriser tous les flux susceptibles de passer entre les deux serveurs ainsi que ceux de l'interface d'administration luci.

system-config-securitylevel

Configurer le lancement automatique des outils clusters sur les deux nœuds. Le démon cman est le gestionnaire principal des services clusters, et le démon gfs2 permet d'utiliser le système de fichier du même nom (à adapter si on utilise gfs ou clvmd plutôt que gfs2). Seul ricci peut démarrer avant que le cluster ait été configuré sur les deux nœuds.

```
chkconfig ricci on
chkconfig gfs2 on
chkconfig cman on
service ricci start
```

2.5 Création et configuration du cluster

2.5.1 Création du cluster

L'interface d'administration luci est disponible à l'adresse suivante

```
https://tstclstr-node1:8084/
ou
https://192.168.0.100:8084/
```

Aller dans l'onglet cluster pour en créer un nouveau dénommé "tstclstr". La création du cluster permet la génération du fichier de configuration, nécessaire avant de pouvoir démarrer cman.

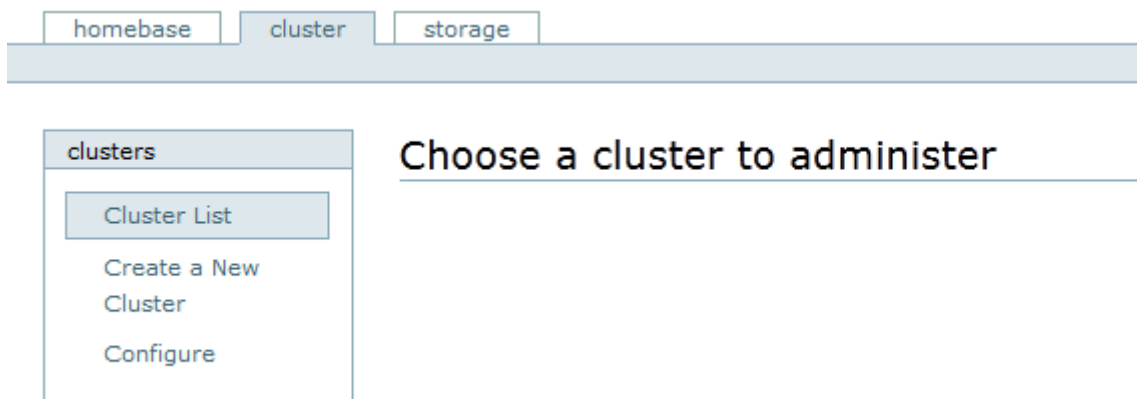


Figure 1 : Créer un cluster via luci

clusters

Cluster List

Create a New Cluster

Configure

Create a new cluster

Cluster Name

Node Hostname	Root Password	Key ID
<input type="text" value="192.168.0.100"/>	<input type="password" value="....."/>	
<input type="text" value="192.168.0.101"/>	<input type="password" value="....."/>	

Download packages
 Use locally installed packages.

Enable Shared Storage Support
 Reboot nodes before joining cluster
 Check if node passwords are identical.

Figure 2 : Ajouter des nœuds au cluster via luci

Create a new cluster

Cluster Name

Node Hostname	Root Password	Key ID
<input type="text" value="192.168.0.100"/>	<input type="password" value="....."/>	
<input type="text" value="192.168.0.101"/>	<input type="password" value="....."/>	

no key fingerprint available

Figure 3 : En cas d'absence de ricci, l'outil de création de cluster affiche cette erreur

2.5.2 Ajout des logs

Par défaut, les logs du cluster sont situés dans `/var/log/messages`. Cependant, ces logs sont assez peu fournis, et peuvent se perdre dans la masse d'information disponible dans le fichier `/var/log/messages`.

Ajouter les lignes suivantes à la fin du fichier `/etc/syslog.conf`:

```
# Red Hat Cluster
local4.* /var/log/rgmanager
```

Ajouter `/var/log/rgmanager` dans la liste des fichiers concernés par la politique de rotation `logrotate` de `syslog` dans `/etc/logrotate.d/syslog`

```

/var/log/messages /var/log/secure /var/log/maillog /var/log/spooler
/var/log/boot.log /var/log/cron /var/log/rgmanager {
    sharedscripts
    postrotate
        /bin/kill -HUP `cat /var/run/syslogd.pid 2> /dev/null` 2> /dev/null
    || true
        /bin/kill -HUP `cat /var/run/rsyslogd.pid 2> /dev/null` 2> /dev/null
    || true
    endscript
}

```

Modifier le `<rm>` pour ajouter le logging dans `/etc/cluster/cluster.conf` et inc la version

```
<rm log_facility="local4" log_level="5">
```

Propager manuellement la configuration sur les nœuds avec la commande suivante

```
ccs_tool update /etc/cluster/cluster.conf
```

2.5.3 Ajout du fence vmware

La plateforme de développement n'étant pas une machine physique, il est nécessaire d'ajouter un composant logiciel pour pouvoir disposer du fencing. Ce fencing se fait par le biais de l'API perl VMware ainsi que de l'outil fence_vmware de Redhat Cluster Suite

```

yum -y install openssl-devel perl-URI
tar xzf VMware-vSphere-SDK-for-Perl-4.0.0-161974.x86_64.tar.gz
cd vmware-vmware-cli-distrib
./vmware-install.pl

```

ATTENTION : la commande fence_vmware fournie dans les repositories CentOS 5.4 du DVD est incompatible avec la version 4 du serveur ESX. Il est nécessaire d'obtenir une version plus récente via Internet ou de l'adapter à la main ce script de la façon suivante :

```

vi /sbin/fence_vmware
23 #la commande timeout trop vite
24 SHELL_TIMEOUT=6
[...]
33 #Il faut changer l'ordre des arguments, le -v doit etre situe en debut
de commande avec ESX4
34 cmd_line=VMWARE_COMMAND+" -H "+options["-A"]+" -U
"+options["-L"]+" -P "+options["-P"]+" "+options["-n"]+"""
35 #if options.has_key("-A"):
36 # cmd_line+=" -v"

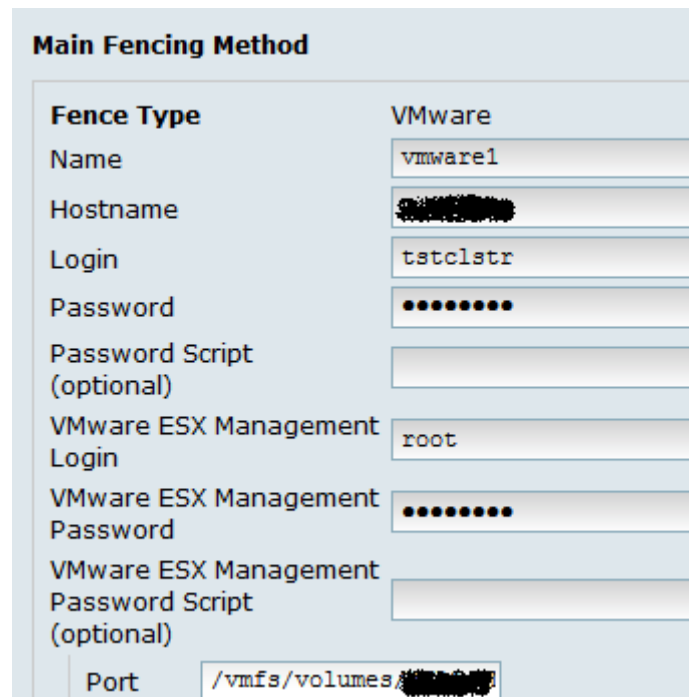
```

NB : Une version plus récente de la distribution corrige probablement ce problème.

Avant d'ajouter le fencing device dans luci, il faut d'abord désactiver le démon acpid sur chaque nœud, car il provoquera des conflits avec fenced.

```
service acpid stop  
chkconfig --del acpid
```

Depuis luci, et sur chaque nœud du cluster, ajouter un « fencing device ». Pour les machines virtuelles VMware, la valeur à entrer dans « port » est le chemin absolu vers le fichier de définition de la machine virtuelle « .vmx ».



Main Fencing Method	
Fence Type	VMware
Name	vmware1
Hostname	[REDACTED]
Login	tstclstr
Password	[REDACTED]
Password Script (optional)	
VMware ESX Management Login	root
VMware ESX Management Password	[REDACTED]
VMware ESX Management Password Script (optional)	
Port	/vmfs/volumes/[REDACTED]

Figure 4 : Ajout d'un module fence_vmware pour un nœud

2.5.4 Création du quorum disk

Le quorum disk est une partition sur lesquels tous les nœuds peuvent écrire leur état à tout moment et ainsi vérifier celui des autres. Cette partition permet également d'arbitrer quel nœud doit prendre le relais dans des cas complexes (perte du réseau uniquement entre les deux serveurs du cluster par exemple)

```
fdisk -l #la partition doit être vue par les deux serveurs  
fdisk /dev/sdb # Créer sdb1 en partition principale sur un seul nœud  
mkqdisk -c /dev/sdb1 -l tstclstr # Formater la partition en qdisk sur un nœud  
partprobe # à lancer sur le second nœud pour prise en compte du label  
mkqdisk -L #verifier qu'on a bien le label « tstclstr » vu par les deux noeuds  
service qdiskd start #démarrer le service sur chaque nœud  
chkconfig qdiskd on #configurer le service au démarrage de chaque nœud
```

2.5.5 Ajout d'une heuristique

En plus du quorum disque, on peut ajouter une ou plusieurs règles qui définissent des conditions dans lesquelles le nœud doit être considéré comme n'appartenant plus au cluster (interface réseau KO par exemple). Le script

suivant vérifie régulièrement que l'interface choisie (eth0 ou bond0 par exemple) est bien opérationnels.

```
vi /usr/share/cluster/check_eth_link.sh
#!/bin/sh
#Network link status checker
ethtool $1 | grep -q "Link detected.*yes"
exit $?

chmod +x /usr/share/cluster/check_eth_link.sh
```

Redhat recommande l'ajout d'une heuristique qui vérifie que la passerelle réseau est toujours accessible. On peut la mettre à la place de l'heuristique ci-dessus, ou en parallèle, au choix.

```
ping -c3 -t2 [@IP_passerelle]
```

Une fois ces opérations terminées, mettre à disposition le quorum et son heuristique dans luci.

The screenshot shows the 'Quorum Partition Configuration' page in the luci interface. The 'Use a Quorum Partition' option is selected. The configuration includes the following values:

- Interval: 2
- Votes: 1
- TKO: 5
- Minimum Score: 1
- Label: tstclstr

The 'Heuristics' section contains a table with the following data:

Path to Program	Interval	Score
/usr/share/cluster/check_eth_link.	2	1

Figure 5 : Ajout d'un qdisk et d'une heuristique

3 Mise en place d'un volume GFS

3.1 Problématique

Un des problèmes soulevés par la mise en place d'un cluster est de pouvoir reposer sur un stockage partagé qui puisse supporter les écritures concurrentes dans le cas où plusieurs nœuds écriraient sur le même stockage.

Pour pallier à ce problème, il est possible d'utiliser des gestionnaires de volumes ou des filesystems qui sont capable de gérer l'accès concurrentiel. Redhat

préconise l'utilisation de GFS/GFS2, son propre mécanisme de filesystem dédié au clustering, mais on peut également utiliser la version dite « cluster » de LVM (via le démon CLVMd), le OCFS d'Oracle et même NFS via des mécanismes de contournement des sécurités fournis par Redhat Cluster Suite.

Dans le cas présent, le problème ne se pose pas vraiment, puisque le stockage « partagé » n'est normalement utilisé que par un seul nœud à la fois (le nœud actif). On peut donc théoriquement utiliser un volume LVM simple distribué par le SAN. Pour la plateforme de développement, le stockage GFS a été testé.

3.2 Création du volume

Une fois le LUN créé et publié sur le serveur VMware, il faut l'affecter aux machines virtuelles tel que cela été fait lors de la création du quorum disk. Ce disque peut être affecté au même contrôleur SCSI (paramétré en option « physique » pour permettre le partage).

```
fdisk -l #la partition doit être vue par les deux serveurs
fdisk /dev/sdc #créer le disque en partition principale sur un seul nœud
mkfs -t gfs -p lock_dlm -t tstclstr:lv_data -j 2 /dev/sdc1
```

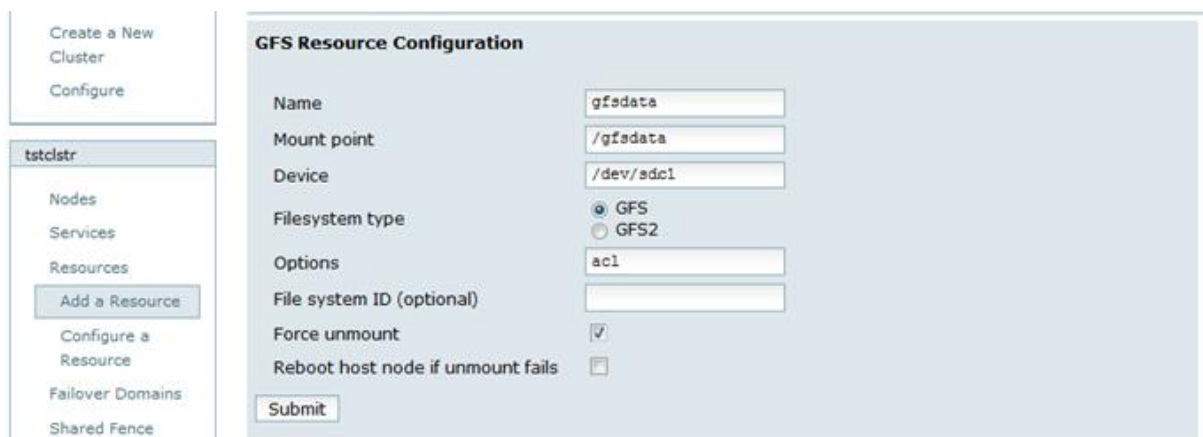
Chaque volume GFS nécessite au moins un journal par nœud qui utilise ce volume. Dans le cadre d'un volume uniquement monté sur le nœud actif, on peut se contenter d'un seul journal. Cependant, il est préférable d'affecter plus de journaux que nécessaire car il est nécessaire d'agrandir l'espace total du volume si on souhaite augmenter le nombre de journaux et donc de nœud simultanés sur le volume GFS.

3.3 Montage du volume

Lors du montage, il est important de ne pas oublier l'option "-o acl", sinon les ACL ne peuvent qu'être lues et non modifiées.

```
mkdir -p /gfsdata
mount -o acl -t gfs /dev/sdc1 /gfsdata
```

Le FS doit être ajouté en tant que ressource dans luci



The screenshot shows the Luci web interface for configuring a GFS resource. On the left, there is a navigation menu with options like 'Nodes', 'Services', 'Resources', and 'Add a Resource'. The main area is titled 'GFS Resource Configuration' and contains the following fields:

- Name: gfsdata
- Mount point: /gfsdata
- Device: /dev/sdc1
- Filesystem type: GFS (selected), GFS2
- Options: acl
- File system ID (optional):
- Force unmount:
- Reboot host node if unmount fails:
- Submit button

Figure 6 : ajout d'un FS GFS en tant que ressource dans luci

4 Installation d'une base Oracle

4.1 Prérequis

```
yum -y install compat-libstdc++-33 elfutils-libelf-devel gcc-c++ libaio-  
devel libstdc++-devel sysstat unixODBC unixODBC-devel pdksh  
  
mkdir /appli/oracle  
  
groupadd dba  
  
groupadd oinstall  
  
useradd -g dba -d /appli/oracle oracle  
  
chown oracle:dba /appli/oracle  
  
passwd oracle  
  
passwd -x -1
```

4.2 Installation des binaires Oracle

Se connecter en temps qu'oracle

```
export DISPLAY=a.b.c.d:0.0 (pour les sessions X distantes)  
  
/appli/distrib/oracle/database/runInstaller #Installer le logiciel de bdd  
uniquement
```

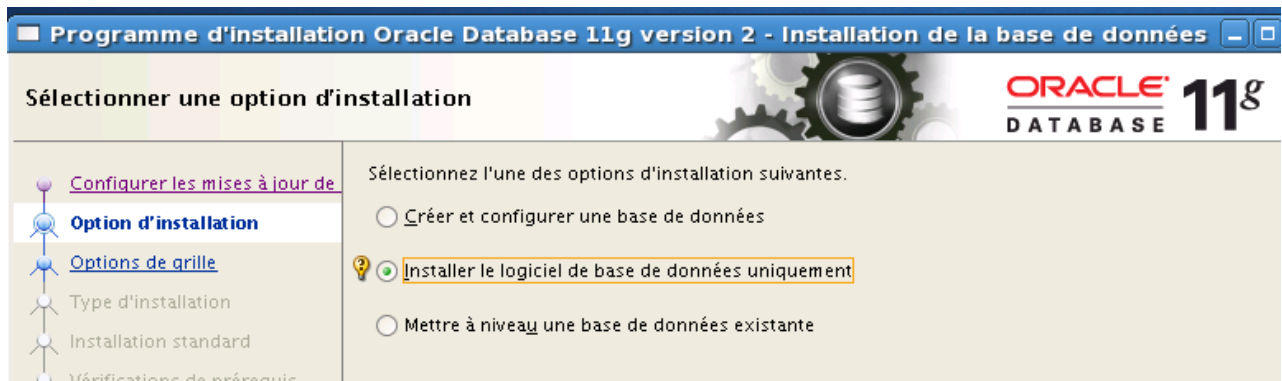


Figure 7 : Installation des binaires Oracle uniquement

Avant de commencer l'installation à proprement parler, Oracle va vérifier les prérequis matériels et sur l'OS. Il pourra être nécessaire d'exécuter un script en tant que « root » pour corriger les éventuels problèmes.

Attention aux chemins définis : toujours utiliser /appli/oracle et vérifier que le logiciel d'installation ne descend pas plus bas dans l'arborescence pour éviter d'avoir des chemins absolus trop longs. L'idéal est d'avoir le \$ORACLE_HOME à /appli/oracle/11.2.0

A la fin de l'installation, Oracle demande d'exécuter deux scripts en tant que « root »

```
/appli/oracle/oralInventory/orainstRoot.sh  
/appli/oracle/11.2.0/root.sh
```

4.3 Configuration Oracle

4.3.1 Configuration listener.ora et tnsnames.ora

Pour pouvoir fonctionner correctement, il est nécessaire d'affecter au listener sur chaque instance Oracle l'IP virtuelle du cluster.

netca #créer un listener via l'outil graphique d'Oracle

On souhaite obtenir un fichier listener.ora et tnsnames.ora qui ressemblent aux fichiers suivants :

```
-bash-3.2$ cat listener.ora
LISTENER_CLUSTER =
  (DESCRIPTION_LIST =
    (DESCRIPTION =
      (ADDRESS = (PROTOCOL = TCP)(HOST = TSTCLSTR)(PORT = 1521))
    )
  )
```

```
ADR_BASE_LISTENER_CLUSTER = /appli/oracle
```

```
-bash-3.2$ cat tnsnames.ora
CLUSTER =
  (DESCRIPTION =
    (ADDRESS = (PROTOCOL = TCP)(HOST = TSTCLSTR)(PORT = 1521))
    (CONNECT_DATA =
      (SERVER = DEDICATED)
      (SERVICE_NAME = CLUSTER.LOCAL)
    )
  )
```

4.3.2 Création de l'instance

La création de l'instance n'est bien entendu à faire qu'une seule fois (sur un seul nœud), à l'aide de la GUI Oracle dbca

Créer une base de données de type « BD généraliste ou traitement transactionnel ».

Ne pas oublier que tous les fichiers data de la base de données (control files, ...) doivent tous être situés sur l'emplacement commun. Si ces fichiers ne sont pas au bon endroit, la base ne pourra pas être montée lors de la bascule

4.3.3 Création des scripts d'arrêt relance

Placer le contenu des fichiers ci-dessous respectivement dans les scripts /etc/init.d/oracle, \$ORACLE_HOME/startdb et \$ORACLE_HOME/stopdb. Tous les scripts doivent disposer du droit d'exécution et les scripts startdb et stopdb doivent appartenir au user/groupe « oracle:dba ».



Une fois créés sur le premier nœud, vérifier que tout fonctionne et les propager sur le second nœud via scp

Enfin, le script /etc/init.d/oracle doit être ajouté en tant que ressource du cluster. Ce script permet aux démons du cluster d'organiser l'arrêt et le lancement du service en fonction du nœud qui est censé être le nœud actif, mais aussi de déterminer si le service est bien fonctionnel, notamment à l'aide de la commande **/etc/init.d/oracle status**.

Script Resource Configuration

Name

Full path to script file

Figure 8 : Ajout d'un script en tant que ressource du cluster

4.3.4 Mise en conformité du 2ème nœud

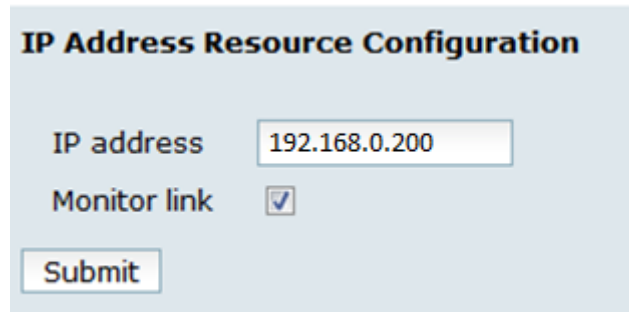
```
scp /appli/oracle/.profile /appli/oracle/startdb /appli/oracle/stopdb  
oracle@TSTCLSTR-NODE2:/appli/oracle/  
scp /etc/init.d/oracle TSTCLSTR-NODE2:/etc/init.d/  
scp /appli/oracle/11.2.0/dbs/* oracle@TSTCLSTR-  
NODE2:/appli/oracle/11.2.0/dbs  
scp /appli/oracle/11.2.0/network/admin/*.ora TSTCLSTR-  
NODE2:/appli/oracle/11.2.0/network/admin
```

4.4 Autres éléments importants

Pour garantir qu'une éventuelle bascule se passe bien, il est nécessaire de mettre en place une mise à jour périodique du fichier spfile depuis le nœud actif vers le nœud passif.

5 Création du service Oracle dans le cluster

5.1 Création d'une ressource IP virtuelle



IP Address Resource Configuration

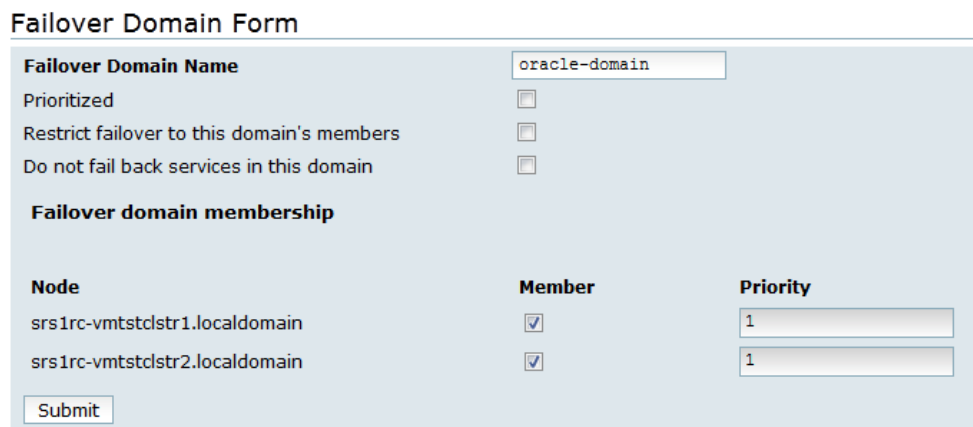
IP address

Monitor link

Figure 9 : Ajout d'une IP virtuelle en tant que ressource pour le cluster

5.2 Création d'un domaine de failover

Dans le cadre d'un cluster de machine important, on peut définir, pour un service donné, un liste de serveurs qui ont spécifiquement le droit de prendre le relai dans le cas où le service ne serait plus desservi par le nœud actif



Failover Domain Form

Failover Domain Name

Prioritized

Restrict failover to this domain's members

Do not fail back services in this domain

Failover domain membership

Node	Member	Priority
srs1rc-vmtstdclstr1.localdomain	<input checked="" type="checkbox"/>	<input type="text" value="1"/>
srs1rc-vmtstdclstr2.localdomain	<input checked="" type="checkbox"/>	<input type="text" value="1"/>

Figure 10 : Ajout d'un failover domain pour Oracle dans luci

5.3 Création d'un service cluster

Un service est représenté par un ensemble de ressources qui sont affectées au nœud actif. Pour Oracle, les 3 ressources qui sont nécessaires sont

- Un script d'arrêt relance
- Le stockage partagé
- L'IP virtuelle affectée uniquement au nœud réellement actif

Automatically start this service	<input checked="" type="checkbox"/>
Enable NFS lock workarounds	<input type="checkbox"/>
Run exclusive	<input checked="" type="checkbox"/>
Failover Domain	oracle-domain ▾
Recovery policy	Restart ▾
Maximum number of restart failures before relocating	0
Length of time in seconds after which to forget a restart	0

Figure 11 : Ajout du service Oracle pour le cluster

La toute dernière étape consiste à ajouter les ressources précédemment créées au service, de manière à le rendre fonctionnel.

Script Resource Configuration

Name

Full path to script file

This resource is an independent subtree

GFS Resource Configuration

Name

Mount point

Device

Filesystem type GFS GFS2

Options

File system ID (optional)

Force unmount

Reboot host node if unmount fails

This resource is an independent subtree

IP Address Resource Configuration

IP address

Monitor link

This resource is an independent subtree

Figure 12 : Ensemble des ressources à ajouter pour fournir le service Oracle en cluster

6 Annexe

6.1 Test de fonctionnement

6.1.1 Démarrage du cluster

Au boot des serveurs, si le service a été configuré de cette façon, les démons cluster doivent être opérationnels et le service Oracle doit se lancer seul. On peut vérifier que tout fonctionne à l'aide de la commande suivante :

```
# clustat
Cluster Status for tstclstr @ Thu Feb 23 13:48:58 2012
Member Status: Quorate

Member Name                               ID Status
-----
tstclstr-node1.localdomain                1 Online, Local,
rgmanager
tstclstr-node2.localdomain                2 Online, rgmanager
/dev/disk/by-id/scsi-3600508b400106b6c0000800003030000-part1  0
Online, Quorum Disk

Service Name                               Owner (Last)
State
-----
service:oracle_cluster                    tstclstr-node1.localdomain
started
```

6.1.2 Arrêt du nœud actif

Si le serveur sur lequel est hébergé le service oracle est éteint normalement, le service doit se couper et le nœud passif doit prendre en charge seul le service oracle. On peut vérifier que la bascule s'est bien effectuée à l'aide de la commande clustat.

6.1.3 Reboot du nœud actif via luci

En conditions de fonctionnement normal, il est possible d'ordonner depuis luci (il faut donc que ricci soit opérationnel et que le serveur en question réponde) de rebooter le serveur. Si le serveur rebooté est le nœud actif, le service doit être basculé sur le nœud passif.

6.1.4 Arrêt / démarrage du service via luci

Il est possible d'arrêter ou de redémarrer le service Oracle depuis l'interface graphique de luci. Dans les deux cas, la commande doit être prise en compte par

le nœud actif. Le service doit rester affecté au nœud initial et ne doit pas changer.

6.1.5 Arrêt manuel du service

Lors de la mise en place du service de cet exemple, nous avons défini que le cluster devait d'abord tenter une fois de relancer le service avant de le « relocate » sur le nœud passif. Normalement donc, lorsqu'on coupe manuellement le service (kill ou /etc/init.d/oracle stop), le cluster doit tenter de relancer le service sur le même nœud.

6.1.6 Relocate du service via luci

Luci permet de déplacer le service d'un nœud à l'autre. Le service doit se couper proprement sur le nœud actif, puis passer sur le nœud passif. Vérifier avec clustat.

6.1.7 Blocage temporaire d'un nœud

Dans l'hypothèse où le système subirait un blocage total pendant un certain temps, le service doit être relocalisé sur le nœud passif. Un moyen simple de vérifier cette bascule depuis l'environnement de développement sur VMware est de passer la machine virtuelle en mode « suspend ». Une fois le timeout passé sur le cluster, le service doit être relocalisé sur le nœud passif, et le nœud anciennement actif doit passer en grisé dans l'interface luci.

Point très important : si la VM est sortie du mode « suspend », **il est impératif de vérifier que le serveur se reboot immédiatement de lui-même**, car sinon deux serveurs ont le service fonctionnel en même temps.

6.1.8 Arrêt brutal non planifié d'un nœud

En cas d'arrêt brutal du nœud actif, celui-ci doit passer en grisé dans l'interface luci, et le service doit être relocalisé sur le nœud passif. Vérifier avec clustat.

6.1.9 Fencing via luci

Le fencing permet d'éviter qu'une machine instable puisse continuer à écrire sur les données partagées du SAN. Il faut donc s'assurer que le nœud est immédiatement tué (shutdown ou reboot) une fois que l'ordre de fencing a été lancé depuis luci.

6.1.10 Coupure d'une interface réseau

Si tout se passe bien normalement il ne doit rien se passer. Le serveur doit se rendre compte qu'il a perdu une interface réseau et tout faire passer par l'interface toujours fonctionnelle.

6.1.11 Coupure totale du réseau

Lors d'une coupure totale du réseau (bond0, donc eth0 ET eth1), il est probable que le serveur ne soit plus à même de rendre le service voulu. Pour simuler une coupure réseau sur la plateforme de développement, il suffit de déconnecter l'interface réseau depuis la configuration de la machine virtuelle.

Le problème est détecté au bout de deux secondes par l'heuristique, qui détermine que le service doit être basculé sur l'autre nœud. Le nœud rend donc immédiatement la main et reboote (via fencing).

6.1.12 Perte du quorum disk sur un nœud

De la même manière que pour le réseau, pour tester les effets de la perte de la connectivité au quorum depuis un des nœuds, il suffit de déconnecter le disque depuis la page de configuration de la machine virtuelle.

Normalement, la machine virtuelle doit rebooter au bout de quelques secondes, et rendre la main si elle possède le service. Dans le cas où le quorum disk n'est toujours pas disponible, le serveur rejoint quand même le cluster mais il est préférable de s'assurer que le quorum disk est de nouveau visible avant de rebasculer le service.

6.2 Sources

6.2.1 Documentation officielle Redhat

- http://docs.redhat.com/docs/en-US/Red_Hat_Enterprise_Linux/5/html/Cluster_Suite_Overview/index.html
- http://docs.redhat.com/docs/en-US/Red_Hat_Enterprise_Linux/5/html/Configuration_Example_-_Fence_Devices/index.html
- http://docs.redhat.com/docs/en-US/Red_Hat_Enterprise_Linux/5/html/Configuration_Example_-_Oracle_HA_on_Cluster_Suite/index.html
- http://docs.redhat.com/docs/en-US/Red_Hat_Enterprise_Linux/5/html/Global_File_System/index.html

6.2.2 Autre

- <http://linuxdynasty.org/215/howto-setup-gfs2-with-clustering/>
- <http://olex.openlogic.com/wazi/2011/ensure-high-availability-with-centos-6-clustering/>
- <http://www.nxnt.org/2010/09/redhat-cluster-howto/>
- <http://dariodallomo.blogspot.com/2011/08/red-hat-cluster-suite-rhcs.html>
- <http://blog.wains.be/2011/02/17/red-hat-cluster-vmware-esx-fencing/>
- <http://securfox.wordpress.com/2009/08/11/how-to-setup-gfs/>
- http://www.linuxtopia.org/online_books/centos_linux_guides/centos_cluster_configuration_and_management/pt-clumanager.html