Virtualisation et Haute disponibilité

Fabien Muller – Hubert Hollender

OpenVz, Pacemaker

Plan de l'exposé

- Présentation
- Problématique de départ
- OpenVZ
- Pacemaker
- Solutions mises en oeuvre
- Bilan

Présentation

- Institut de Physique et de Chimie des Matériaux de Strasbourg (IPCMS)
  - unité mixte CNRS/UDS
  - spécialisé dans le domaine des nanosciences
  - regroupe des physiciens et des chimistes
  - étudier et élaborer de nouveaux matériaux
  - 5 départements de recherche, 230 personnes

Présentation

- Ressources informatiques
  - des serveurs de calcul (environ 1200 coeurs)
  - des serveurs de services
  - des postes instrumentaux et de travail
  - environ 800 systèmes interconnectés en réseau
    - infrastructure systèmes et réseaux centralisée
    - s’appuyant sur des technologies de haute-disponibilité
    - gérées par un service de 5 personnes

Plan de l’exposé

- Présentation
- Problématique de départ
- OpenVZ
- Pacemaker
- Solutions mises en oeuvre
- Bilan

Problématique de départ

- Renouvellement des solutions existantes
  - Mécanismes de haute disponibilité
  - Virtualisation
- Sécuriser les services de la DMZ
  - Messagerie (SMTP) et agendas (OBM)
  - Serveurs web, ftp et accès
- Sécuriser les services de la zone locale
  - Partage de fichiers (Samba, NFS), CUPS
  - DNS, DHCP, LDAP, MYSQL, POSTGRESQL

Plan de l’exposé

- Présentation
- Problématique de départ
- OpenVZ
- Pacemaker
- Solutions mises en oeuvre
- Bilan

OpenVZ : principes

- C'est une virtualisation au niveau noyau en réalisant un partitionnement des ressources systèmes
- C'est le noyau du système d'exploitation qui fait une isolation entre des machines virtuelles et permet d'exécuter des applications dans des contextes différents
- Chaque contexte d'exécution est une machine virtuelle (VPS) se partageant le même noyau.
- Un processus d'une machine virtuelle ne peut pas faire de déni de service en dehors de sa machine virtuelle
- Il n'y a pas d'« émulation » à proprement parler comme dans d'autres système de virtualisation.

OpenVZ : fonctionnalités

- Chaque « Virtual Private Servers » VPS est un système indépendant
- Les VPS sont des systèmes Linux normaux (file system, scripts, programmes) : aucune spécificité openVZ
- Les VPS sont totalement isolés les uns des autres (mémoire, file system, communication inter-processus (IPC))
- Chaque VPS a sa propre adresse réseau, les adresses multiples par VPS sont permises. Le trafic réseau de chaque VPS est isolé (pas de snooping possible).

OpenVZ : fonctionnalités

- Migration à chaud : Sauvegarde sur disque de l'état du serveur virtuel via un mécanisme de snapshot. Ce fichier peut alors être transféré sur une autre machine et restauré en état de marche en quelques secondes.
- Les « beancounters » : Ensemble de paramètres (une vingtaine) pouvant être attribué à chaque serveur virtuel pour imposer des limites et préserver les ressources de la machine hôte.
- Gestion des ressources : Outils permettant de contrôler l'utilisation des ressources de l'hôte par machine virtuelle (mémoire résidente et virtuelle, quotas d'utilisation CPU et disque, priorités d'accès CPU et disque)
- OS template : « file system » d'une distribution Linux permettant de « peupler » les VPS, disponibles sous forme de paquetages ou pouvant être créés.

OpenVZ : avantages

- Pas besoin d'image disque de machine. Il suffit de copier un file system pour installer une machine virtuelle
- Consommation mémoire légère (la mémoire est mutualisée entre le serveur hôte et les VPS et la mémoire demandée à l'hôte est celle réellement utilisée par les processus du VPS)
- Cela peut simplement être vu comme un chroot du file system amélioré par une isolation des processus
- Tous les processus des machines virtuelles font les appels système à un seul noyau. Les E/S sont donc plus efficaces que sur un système qui tournerait à travers une émulation
- Intégré dans la distribution Debian en standard
- Gestion fine des ressources

OpenVZ : Configuration

- Récupération d'un template d'OS :
  - cd /vz/template/cache
  - wget http://download.openvz.org/template/precreated/debian-5.0-x86_64.tar.gz
- Création et configuration du VE :
  - vzctl create 101 --ostemplate debian-5.0-x86_64.tar.gz
  - vzctl set 101 --onboot yes –save
  - vzctl set 101 --hostname xstra –save
  - vzctl set 101 --ipadd 172.16.0.101 –save
  - vzctl set 101 --nameserver 130.79.200.200 –save
  - vzctl set 101 –-diskspace 80G:80G –save
  - vzctl set 100 –-privvmpages 250M:250M –save
  - vzctl start 101
  - vzctl enter 101

OpenVZ : Configuration

- Autres commandes utiles :
  - vzlist
  - vzctl stop 101
  - vzctl destroy 101
  - vzctl exec 101 ps ax ou vztop
  - vzmemcheck, vzcpucheck, vzcalc
- Sauvegarde et restauration :
  - Par défaut dans /var/lib/vz/dump
  - vzdump –compress –dumpdir /mnt/xstra/ --stop 101 –mailto sem-xstra@ipcms.u-strasbg.fr
  - vzdump –restore /mnt/xstra/xstra.tgz 101
- Migration à chaud :
  - Fonctionne avec rsync et ssh (automatisation via clé ssh)
  - Vzmigrate –r no –online –v 172.16.0.102 101

OpenVZ : Configuration

- Fichier de configuration générale :
  - /etc/vz/vz.conf
- Fichiers de configuration des VPS :
  - /etc/vz/conf/<ctid>.conf
  - /etc/vz/conf/<ctid>.mount
  - /etc/vz/conf/<ctid>.umount
  - Configuration lue au démarrage
  - Modifiable à chaud via l'utilitaire : vzctl set
- Files system des VPS :
  - /var/lib/vz
    - dump pour les sauvegardes
      - private pour les files systems
      - template pour les templates

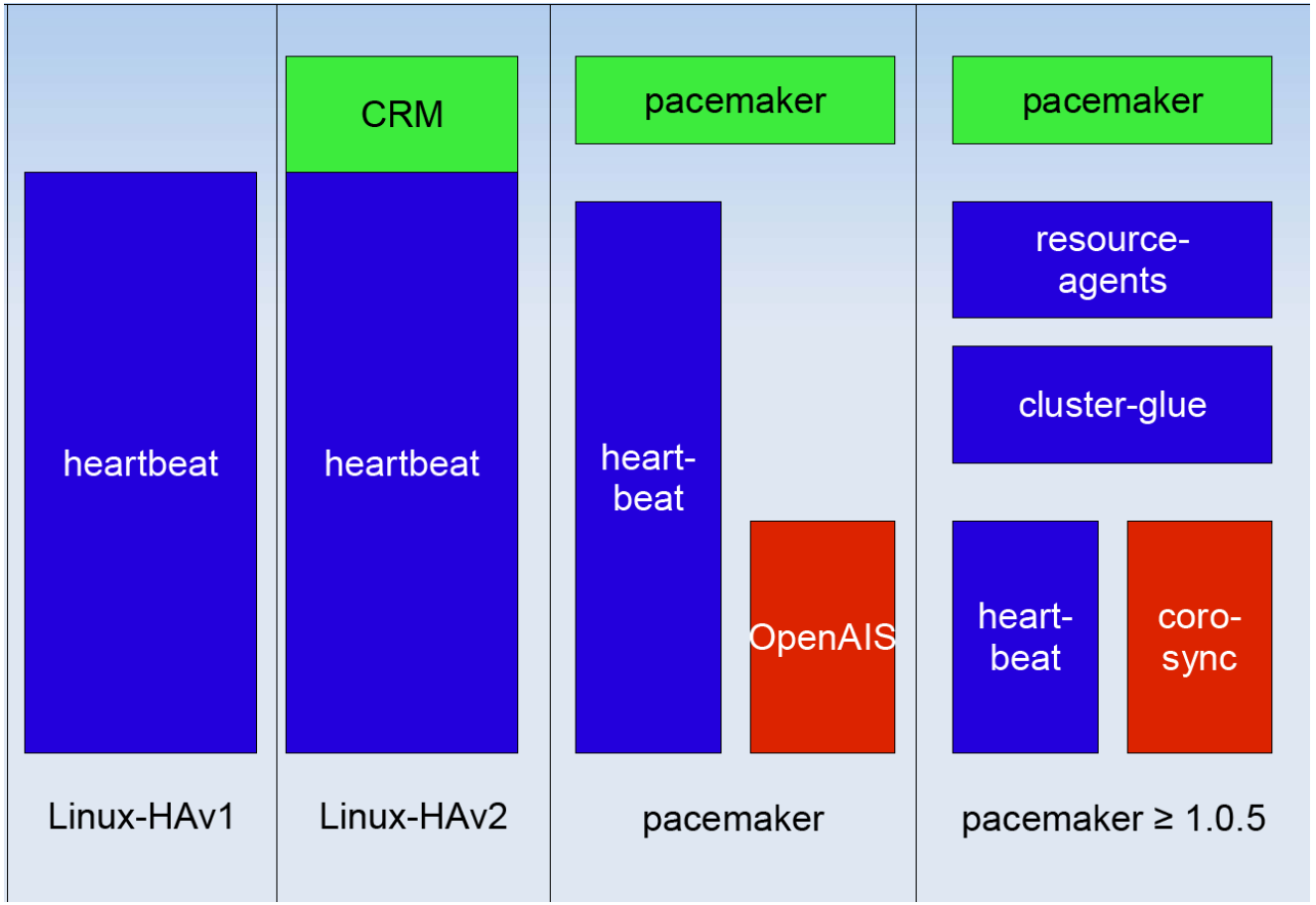
Plan de l'exposé

- Présentation
- Problématique de départ
- OpenVZ
- Pacemaker
- Solutions mises en oeuvre
- Bilan

Pacemaker

- Solution OpenSource de clustering
- Gestion de la haute disponibilité des services et données
- Evolution de heartbeat (différentes branches, version 3)
- Gestion de cluster jusqu'à 16 noeuds
- Gestion de différents modes : Actif/Passif, Actif/Actif
- Depuis la version 2 gestion intégrée des ressources
- Déplacement dynamique/manuel des ressources
- Co-location et ordre de lancement des ressources

Pacemaker (évolutions)



Pacemaker : Configuration

- Via les commandes du gestionnaire de ressources :
  - crm : Exemple

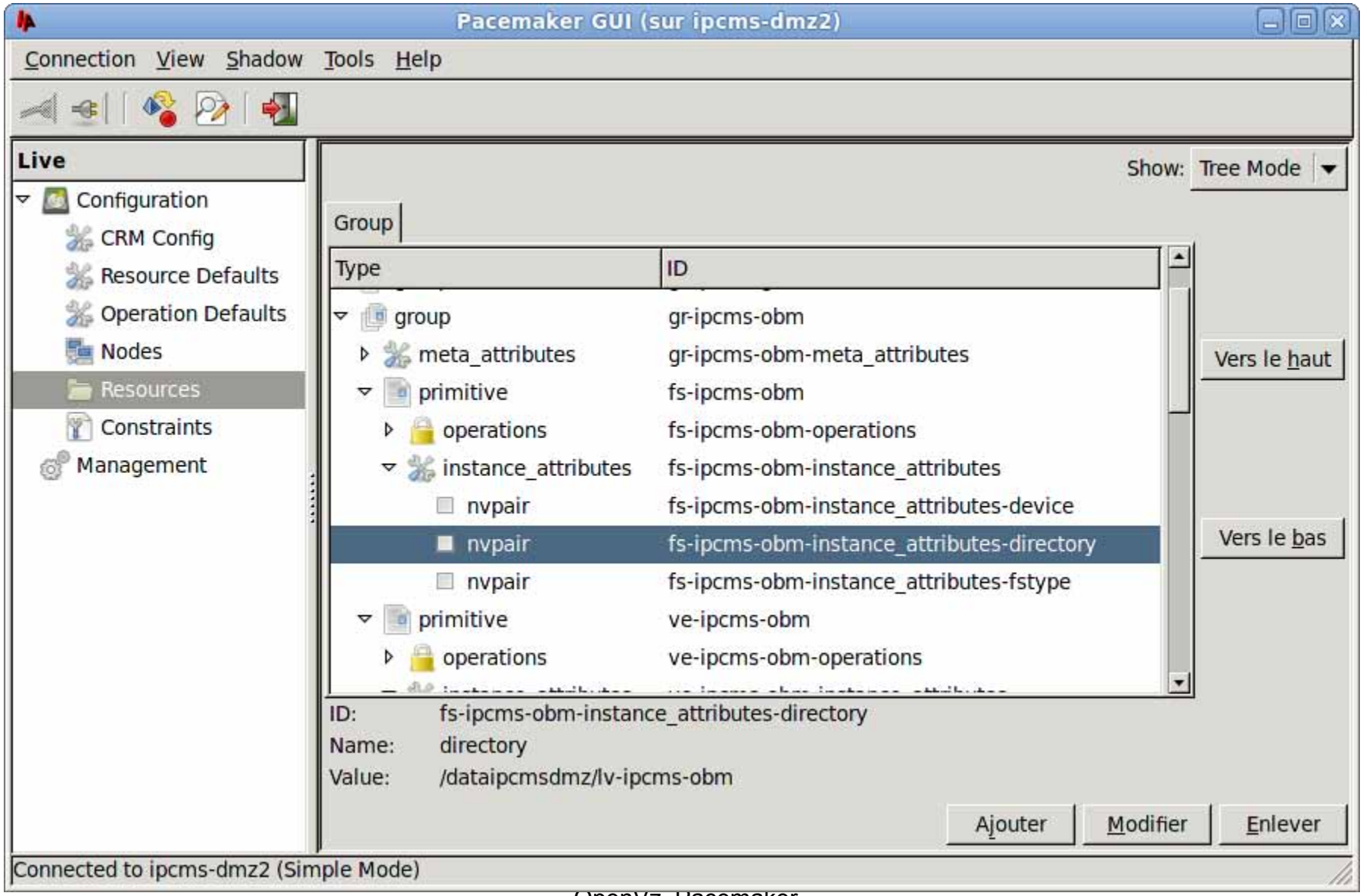
```
primitive fs-serveur-web-01 ocf:heartbeat:Filesystem \
operations $id="fs-serveur-web-01-operations" op \
monitor interval="20" timeout="40" \
params device="/dev/vg-dataipcmsdmz/lv-serveur-web-01" \
directory="/dataipcmsdmz/lv-serveur-web-01" \
fstype="ext3"

primitive ve-serveur-web-01 ocf:heartbeat:ManageVE \
operations $id="ve-serveur-web-01-operations" op \
monitor interval="10" timeout="10" \
params veid="2000"

group gr-serveur-web-01 fs-serveur-web-01 ve-serveur-web-01\
meta target-role="Started"
```

- Via l'interface graphique:
  - hb_gui

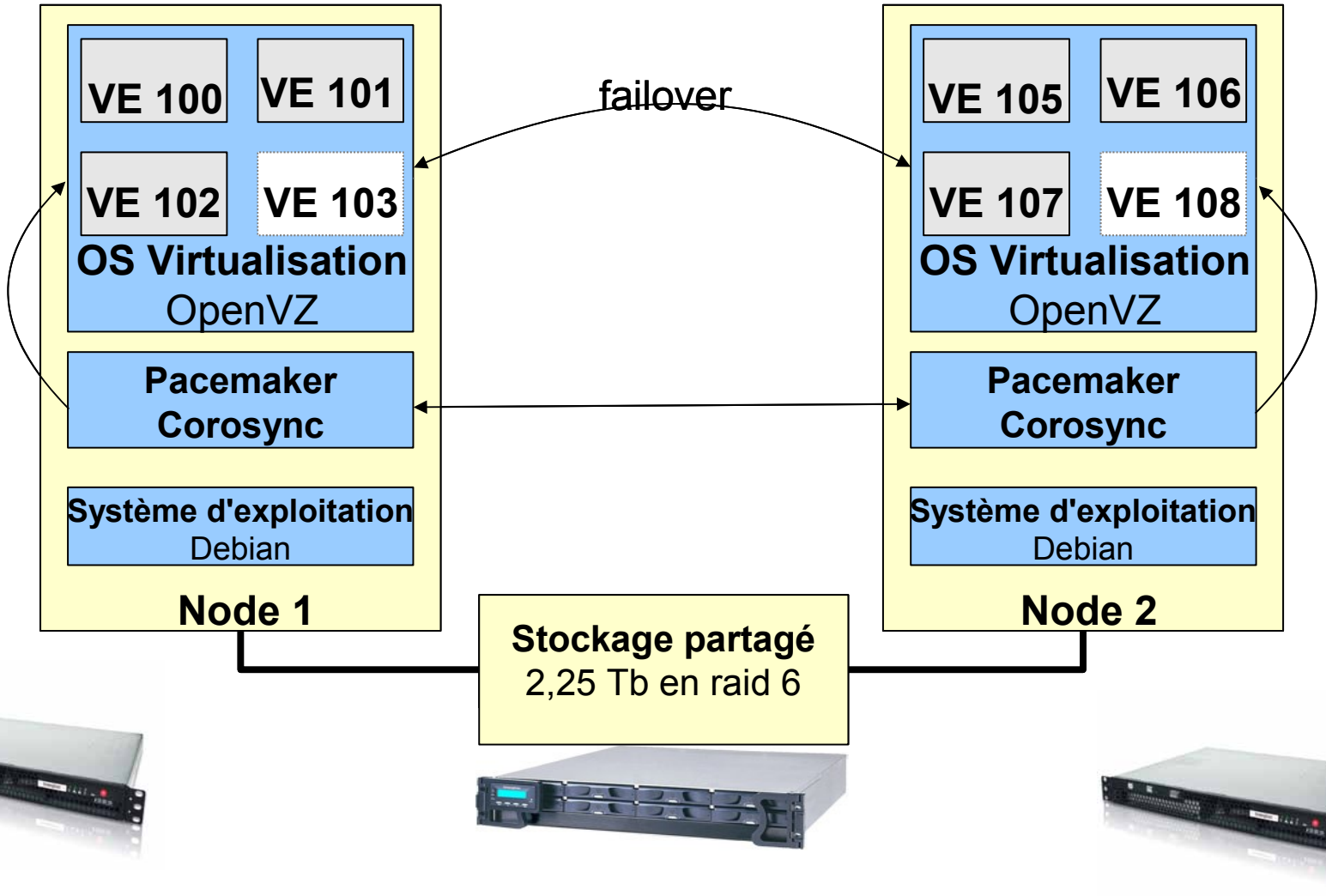
Pacemaker (GUI)



Plan de l’exposé

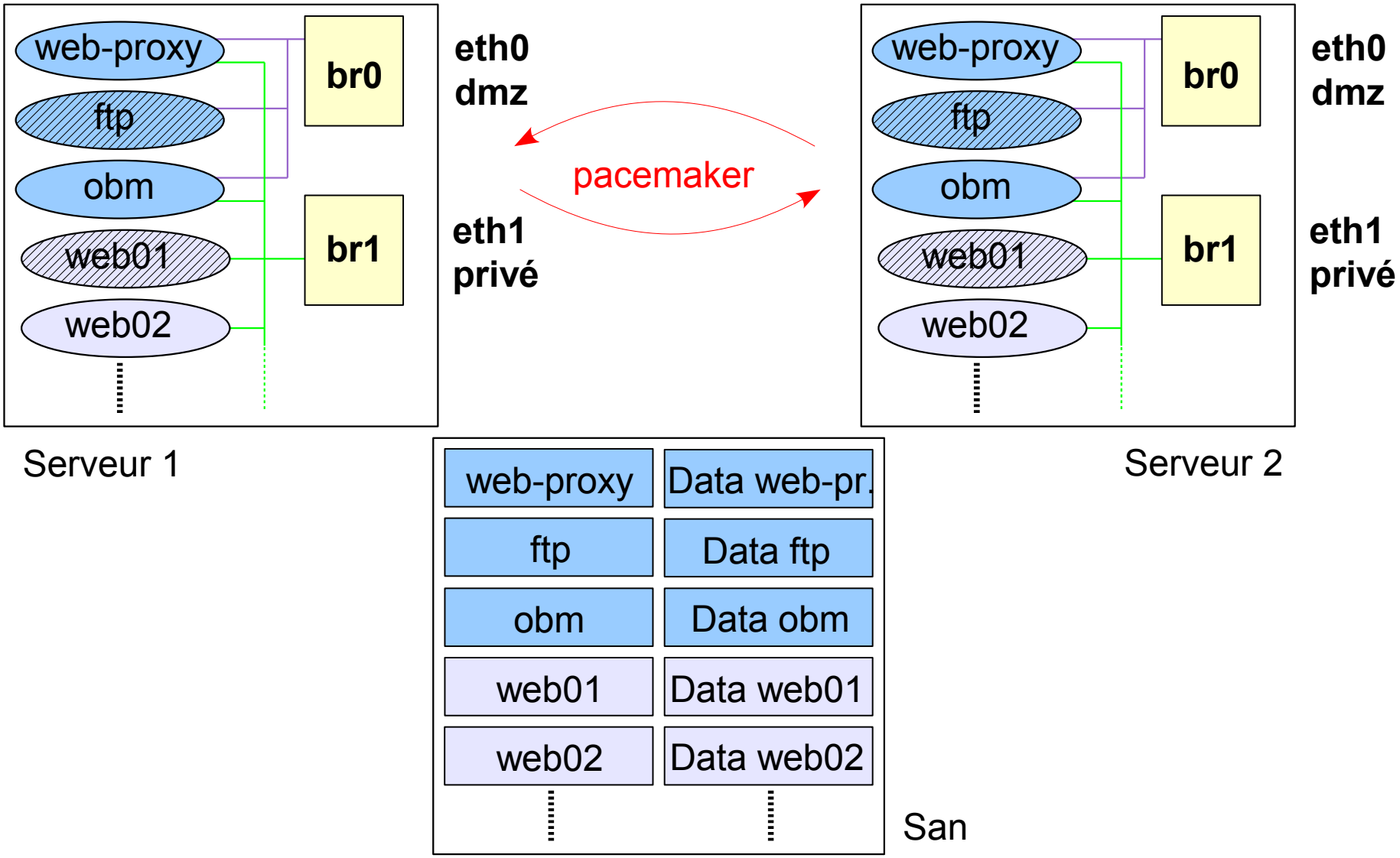
- Présentation
- Problématique de départ
- OpenVZ
- Pacemaker
- Solutions mises en œuvre
- Bilan

Solution mise en œuvre : DMZ



OpenVz, Pacemaker

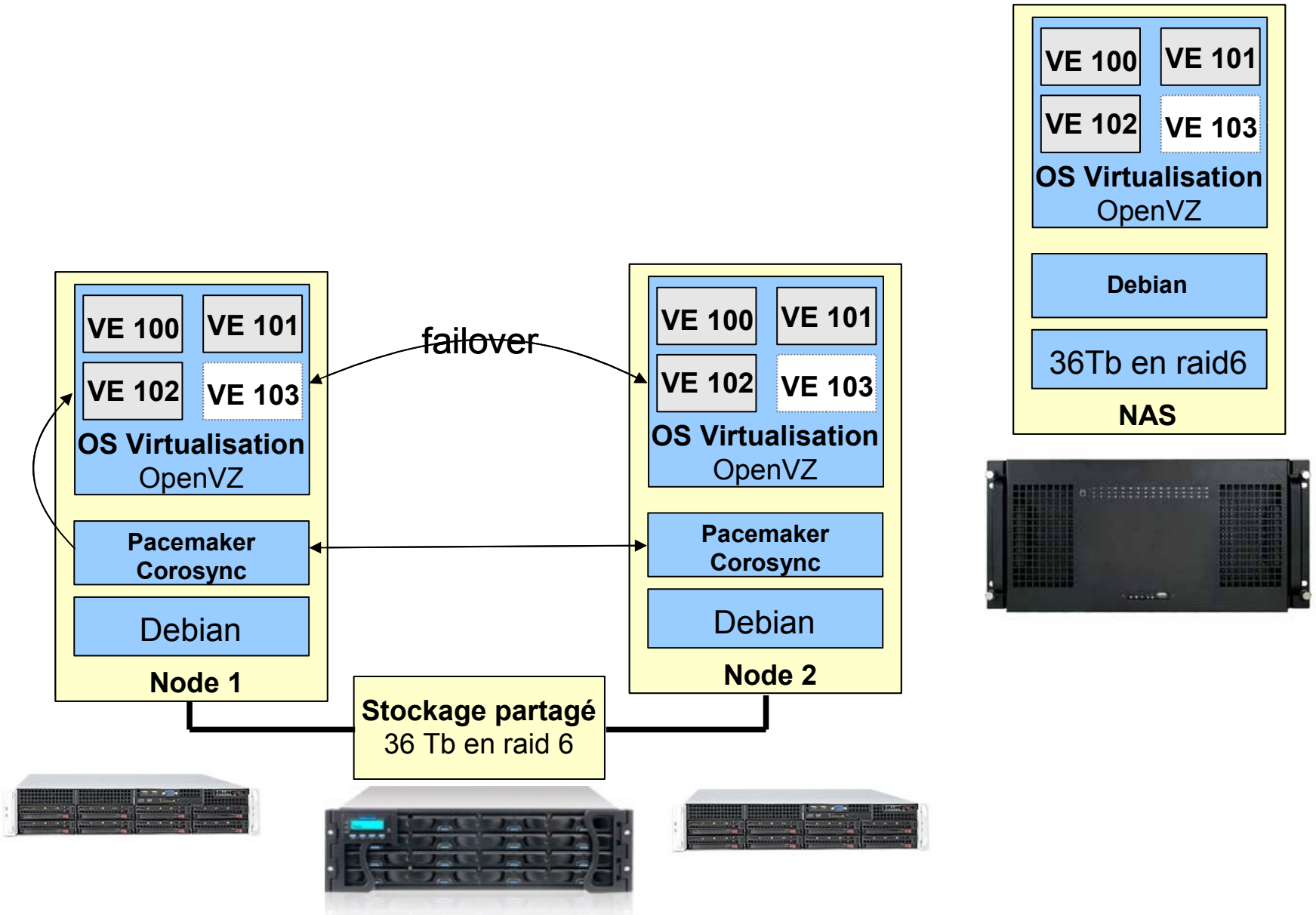
Solution mise en œuvre : DMZ



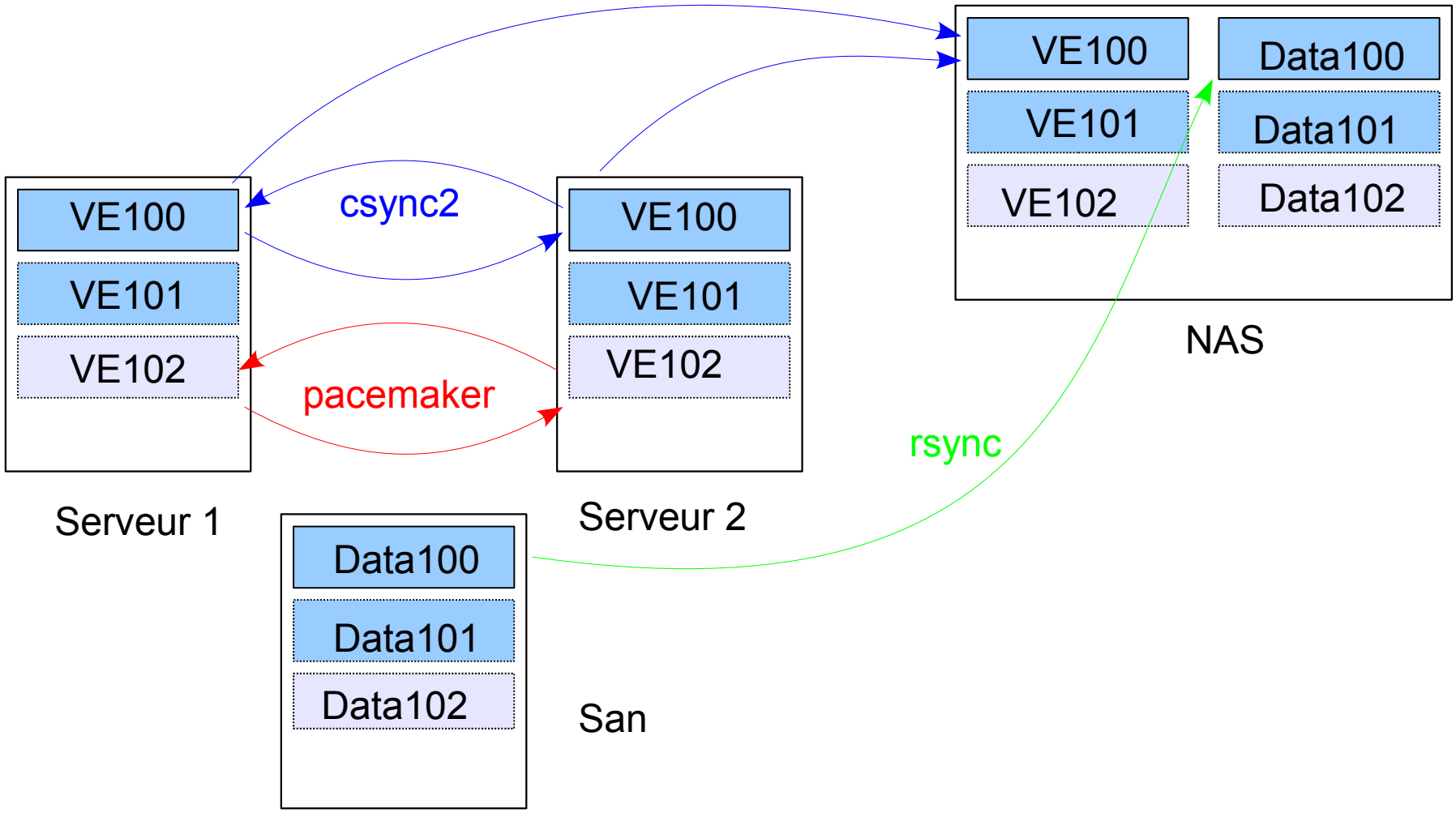
Solution mise en œuvre : DMZ

- Baie Provigo 610 SAS :
  - 8 disques de 500 Go SATA à 7200 tr/min
  - 1 disque Hot Spare
  - 1 Raid Group (7 disques) en Raid 6 (2,5 To utile)
  - 1 contrôleur, cache 1 Go, unité batterie
  - alimentation redondante
  - double connexions SAS vers les 2 serveurs
- 2 Serveurs Calleo 121
  - Quad Core
  - 8 Go de mémoire, 160 Go (en mirroir), 2 x 1 Gb Ethernet + IPMI
- Cluster
  - Actif/Actif
  - SMTP, OBM, SSH, HTTP, HTTPS, FTP

Solution mise en œuvre : Zone locale



Solution mise en œuvre : Zone locale



Solution mise en œuvre : Zone locale

- Baie Infortrend 10 Gb ISCSI :
  - 22 disques de 2 To Go SATA Nearline à 7200 tr/min
  - 2 disques Hot Spare
  - 2 Raids Group (11 disques) en Raid 6 (36 To utile)
  - 2 contrôleurs, cache 1 Go, 2 unités batterie
  - alimentation redondante
  - double connexions fibre 10 Gb vers les 2 serveurs
- 2 Serveurs Supermicro
  - Biprocesseurs Quad Core
  - 24 Go de mémoire, 300 Go (en mirroir)
  - 2 x 10 Gb Ethernet + IPMI, 2 x 10 Gb ISCSI
- Cluster
  - Actif/Actif
  - SAMBA, NFS, CUPS, CYRUS, DNS, DHCP, LDAP, SQL

Plan de l'exposé

- Présentation
- Problématique de départ
- OpenVZ
- Pacemaker
- Solutions mises en œuvre
- Bilan

Bilan

- En service depuis 2 ans (DMZ) et 6 mois (locale)
- Performante, bon niveau de sécurité
- Peu consommatrice en ressources
- Faible coût d'acquisition (6 K€ et 35 K€)
- Investissement pour maitriser la technologie
- Evolutions futures :
  - Proxmox, OCFS2
  - Fusion des deux systèmes de stockage
  - Reprise de tous les services sur le NAS