

La haute disponibilité dans la vraie vie

Arnaud Gomes-do-Vale

Le 2 août 2010

Sommaire

- 1 Généralités
- 2 Problématique
- 3 Outils
- 4 Déploiement : le présent
- 5 Déploiement : l'avenir

Sommaire

1 Généralités

2 Problématique

3 Outils

4 Déploiement : le présent

5 Déploiement : l'avenir

Généralités

- services
- point critique (SPOF)
 - ▶ machine
 - ▶ logiciel
 - ▶ réseau
 - ▶ stockage
 - ▶ électricité
 - ▶ bâtiment

La règle des 9

99%	3,6 jours
99,9%	8,7 heures
99,99%	52 minutes
99,999%	5 minutes
99,9999%	31 secondes

Sommaire

1 Généralités

2 Problématique

3 Outils

4 Déploiement : le présent

5 Déploiement : l'avenir

L'Ircam

Présentation générale

- recherche
- production musicale
- enseignement
- beaucoup de machines auto-administrées
- pas d'hébergement «lourd»

L'Ircam

Quelques chiffres

- 200-300 utilisateurs
- 400 machines
- 40 serveurs
- 160 machines virtuelles

L'environnement

- un local réseau par bâtiment
- une salle machines neuve
- deux circuits électriques indépendants dont un secouru de bonne qualité ou pas
- budget réduit

- Linux (principalement CentOS)
- Xen, Linux VServer
- pas ou peu de stockage centralisé
- beaucoup de services...
 - ▶ essentiels : DNS, LDAP, SMTP, POP/IMAP, DHCP, homes
 - ▶ importants : MySQL, CUPS, «gros» sites web
 - ▶ peu importants : VPN, sites web divers...

Les problèmes

- pour une machine qui tombe, 20 ou 30 services arrêtés
- impossibilité de rebooter certains serveurs à moins de fermer le bureau à clé
- utilisateur != client

Sommaire

- 1 Généralités
- 2 Problématique
- 3 Outils**
- 4 Déploiement : le présent
- 5 Déploiement : l'avenir

La base

- nagios
- puppet

Redondance native

- DNS (autorité)

```
@    IN    NS maelzel.ircam.fr.  
     IN    NS nadia.ircam.fr.  
     IN    NS vaslav.ircam.fr.  
  
     IN    NS ns0.pasteur.fr.  
     IN    NS ns1.pasteur.fr.
```

- résolveur DNS

```
nameserver 129.102.2.10  
nameserver 129.102.2.11  
  
options timeout:1 rotate
```

Redondance native

- réplication LDAP

```
host ldap1.ircam.fr ldap2.ircam.fr
```

```
AuthLDAPUrl ldap://ldap1 ldap2/dc=ircam,dc=fr
```

- SMTP

```
@    IN    MX 10 mx1.ircam.fr.  
      IN    MX 20 mx2.ircam.fr.  
      IN    MX 100 mx3.ircam.fr.
```

- + mise en œuvre facile

- service dégradé quand une instance tombe

Cluster HA

Concepts

- nœud, service
- données, stockage partagé
- split brain
- quorum
- loi de l'information de Dunn
Quand on ne sait pas, on ne sait pas.
Et on ne peut rien inventer.
- fencing, STONITH (shoot the other node in the head)

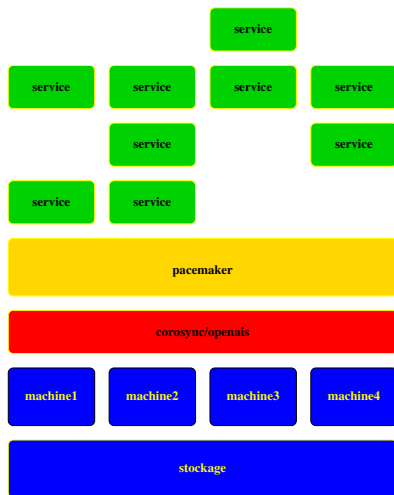
Cluster HA

Stockage partagé

- SAN, NAS
- OCFS2, GFS
- réplication
- DRBD
 - ▶ réplication de données au niveau bloc
 - ▶ 2 ou 3 machines
 - ▶ précautions à prendre en cas de split brain

Heartbeat & co

Architecture



Heartbeat & co

Configuration de corosync

```
totem {  
  
    .../...  
  
    interface {  
        ringnumber: 0  
        bindnetaddr: 129.102.2.0  
        mcastaddr: 239.102.2.10  
        mcastport: 694  
    }  
  
    .../...  
  
}
```

Heartbeat & co

Multicast

- switches Cisco + IGMP snooping + multicast non routé = attention !

- ▶ solution 1 :

```
ip multicast-routing
interface Vlan102
    ip pim sparse-dense-mode
```

- ▶ solution 2 :

```
interface Vlan102
    ip igmp snooping querier
```

- Ne pas oublier d'ouvrir les firewalls !

```
iptables -A INPUT -p udp -d 239.102.2.10 \
    --dport 694 -j ACCEPT
iptables -A INPUT -p igmp -j ACCEPT
```

Heartbeat & co

Configuration de pacemaker

```
<cib validate-with="pacemaker-1.0" crm_feature_set="3.0.1"
  have-quorum="1" admin_epoch="0" epoch="871"
  dc-uuid="dns-hal.ircam.fr" num_updates="28">
<configuration>
  .../...
<resources>
  <primitive class="ocf" id="ip_dede" provider="heartbeat"
    type="IPaddr">
    <instance_attributes id="ip_dede-instance_attributes">
      <nvpair id="ip_dede-instance_attributes-ip" name="ip"
        value="129.102.2.10"/>
      <nvpair id="ip_dede-instance_attributes-cidr_netmask"
        name="cidr_netmask" value="32"/>
    </instance_attributes>
```

Heartbeat & co

Configuration de pacemaker

```
crm configure property stonith-enabled="false"
crm configure property no-quorum-policy="ignore"

crm configure primitive ip_dede ocf:heartbeat:IPaddr \
  params ip=129.102.2.10 cidr_netmask=32 \
  op monitor interval=30s
crm configure primitive ip_nenesse ocf:heartbeat:IPaddr \
  params ip=129.102.2.11 cidr_netmask=32 \
  op monitor interval=30s
crm configure colocation ip_dede-away-from-ip_nenesse \
  -500: ip_dede ip_nenesse

crm configure rsc_defaults resource-stickiness=100
```

Heartbeat & co

Calcul des scores

- décider quel nœud fait tourner quel service
- pour une ressource donnée, calcul d'un score pour chaque nœud
- à chaque changement de situation
- contraintes
 - ▶ localisation
 - ▶ ordre
 - ▶ colocalisation
- Exemple :

```
crm configure colocation ip_dede-away-from-ip_nenesse \  
-500: ip_dede ip_nenesse
```

```
crm configure rsc_defaults resource-stickiness=100
```

Heartbeat & co

Clusters de deux machines

- pas de quorum

```
crm configure property no-quorum-policy="ignore"
```

- attention au fencing

```
crm configure property stonith-enabled="false"
```

- un bon cluster de 2 machines, c'est un cluster de 3 machines

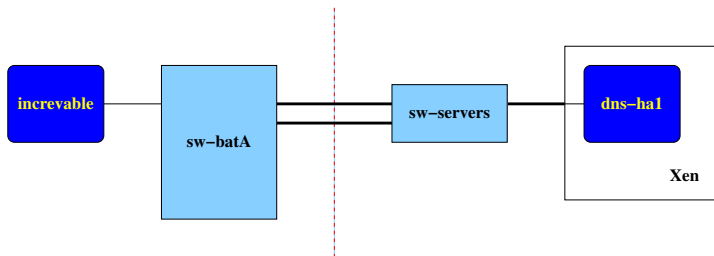
```
crm node standby machine3
```

Sommaire

- 1 Généralités
- 2 Problématique
- 3 Outils
- 4 Déploiement : le présent**
- 5 Déploiement : l'avenir

Le cluster en production

- increvable.ircam.fr : vieux serveur 32 bits
- dns-ha1.ircam.fr : invité Xen 64 bits
- CentOS 5
- dépôt RPM Clusterlabs <http://www.clusterlabs.org/rpm/>



Résolveur DNS

Avant

- 2 vservers indépendants sur 2 hôtes différents
- BIND
- resolv.conf adapté sur les clients

```
nameserver 129.102.2.10  
nameserver 129.102.2.11  
options timeout:1 rotate
```

- problème sur les machines clientes non administrées

Résolveur DNS

Après

- Unbound (géré par puppet)
- 2 adresses IPv4 réparties sur les 2 machines

```
crm configure primitive ip_dede ocf:heartbeat:IPaddr \  
  params ip=129.102.2.10 cidr_netmask=32 \  
  op monitor interval=30s
```

```
crm configure primitive ip_nenesse ocf:heartbeat:IPaddr \  
  params ip=129.102.2.11 cidr_netmask=32 \  
  op monitor interval=30s
```

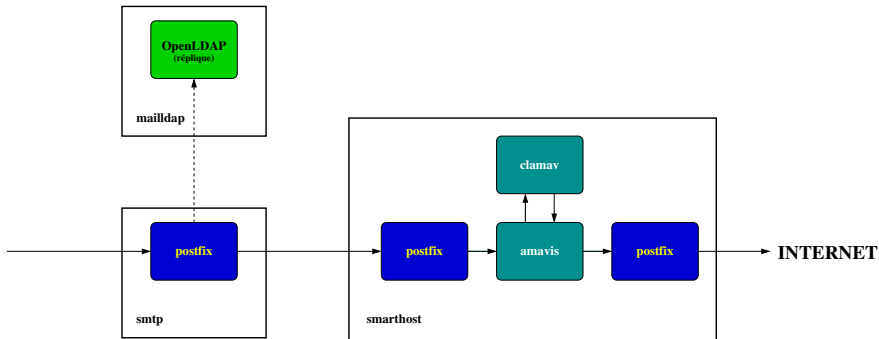
```
crm configure colocation ip_dede-away-from-ip_nenesse \  
  -500: ip_dede ip_nenesse
```

- en cas de coupure entre les switches, les IP restent visibles des deux côtés (mais le routage entre les VLAN ne se fait que d'un côté)

SMTP sortant

Avant

- postfix, OpenLDAP, amavis, clamav
- 1 vserver *smtp* et 1 vserver *mailldap* pour la réécriture
- 1 vserver *smarthost* pour la distribution



SMTP sortant

Après

- postfix, OpenLDAP, amavis, clamav (gérés par puppet)
- 1 adresse IPv4 *smtp*

```
crm configure primitive ip_smtp ocf:heartbeat:IPaddr \  
    params ip=129.102.2.82 cidr_netmask=32 \  
    op monitor interval=30s
```

- chaque machine du cluster a son propre OpenLDAP local
- 1 MX *smarthost* (2 vservers) pour la distribution

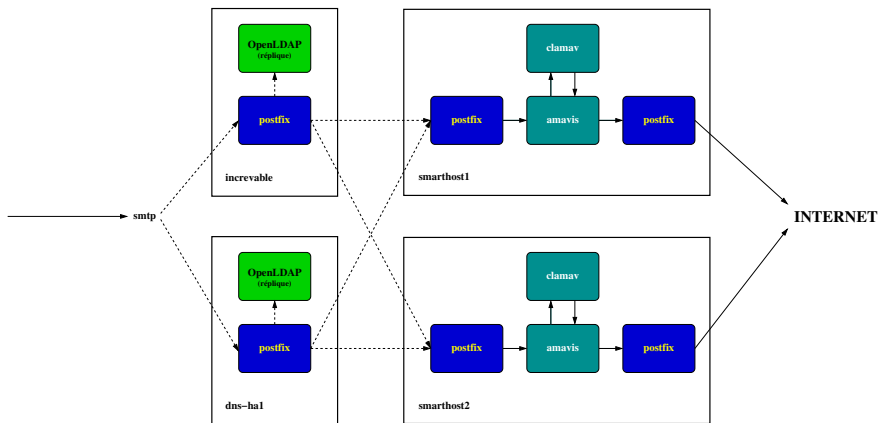
```
smarthost      IN      MX 10 smarthost1  
               IN      MX 10 smarthost2
```

- configuration Postfix adaptée sur les machines du cluster

```
relayhost = smarthost.ircam.fr
```

SMTP sortant

Après



Problèmes rencontrés

- IPv6 (ressource IPv6Addr absente des paquets RPM de Clusterlabs)

Sommaire

- 1 Généralités
- 2 Problématique
- 3 Outils
- 4 Déploiement : le présent
- 5 Déploiement : l'avenir**

DHCP

DHCP Failover

- (plus ou moins) normalisé : draft-ietf-dhc-failover-12.txt, RFC3074
- 2 serveurs (primaire et secondaire)
- chacun son IP
- adresses statiques en double
- plages dynamiques réparties entre les deux serveurs
- équilibrage de charge : RFC 3074
- PARTNER-DOWN : promotion (manuelle) du survivant
- reconfiguration du service lors du changement d'un des serveurs

- MySQL Cluster ?
- partage de données ?
- actif-passif, SAN, DRBD
- réplication maître-esclave
 - ▶ MySQL 3.23.x : réplication d'instructions
 - ▶ MySQL 5.1.x : réplication de données

MySQL

MySQL et CentOS

- `mysql-5.0.77-4.el5_5.3`
- `http://www.iuscommunity.org`
- `yum-plugin-replace`

```
yum replace mysql --replace-with mysql51
```

```
Error: Missing Dependency: perl-DBD-MySQL is needed by package  
mysql51-server-5.1.48-2.ius.el5.i386 (ius)
```

```
You could try using --skip-broken to work around the problem
```

```
You could try running: package-cleanup --problems
```

```
package-cleanup --dupes
```

```
rpm -Va --nofiles --nodigest
```

Questions ?