MySQL High Availability and Geographical Disaster Recovery with Percona Replication Manager

Yves Trudeau
November 2013
Agenda

- Geo-DR problems and challenges
- Introduction to Corosync
- Introduction to Pacemaker
- Introduction to Booth
- Introduction to PRM
- PRM + geo-DR
- Demo
Introduction to Corosync

- Corosync is a communication service
- Totem protocol (google it...)
- Token ring with membership and quorum
- Not tied to Pacemaker
- UDP and multicast but can do unicast
- Redundant rings
- Encryption
Corosync sample config

totem {
    version: 2

    # Enable encryption
    secauth: on
    token: 2000

    # This specifies the mode of redundant ring, which may be none, active, or passive.
    rrp_mode: none

    interface {
        # The following values need to be set based on your environment
        ringnumber: 0
        bindnetaddr: 172.30.212.0
        netmask: 255.255.255.0
        mcastaddr: 239.255.42.1
        mcastport: 5405
    }
}
Introduction to Pacemaker

- Distributed resource management framework
- The logic, the brain, pengine on the DC
- Score based framework
- Control the resources: start, stop, monitor, promote, demote and notify
- Generate events and synchronize them
- Concepts of group, order, location, colocation, master-slave, clone, etc.
- Configuration, CIB, is synced between the nodes by corosync
Pacemaker sample config

node pacemaker-1
node pacemaker-2
node pacemaker-3
primitive writer_vip ocf:heartbeat:IPaddr2 \ 
  params ip="172.30.212.100" nic="eth1" \ 
  op monitor interval="10s"
location loc_writer_vip_prefer_pacemaker-2 writer_vip \ 
  rule $id="rule_loc_writer_vip_prefer_pacemaker-2" 10: #uname eq pacemaker-2
property $id="cib-bootstrap-options" \ 
  dc-version="1.1.7-ee0730e13d124c3d58f00016c3376a1de5323cfff" \ 
  cluster-infrastructure="openais" \ 
  expected-quorum-votes="3" \ 
  stonith-enabled="false"
Pacemaker cluster status

==============

Last updated: Tue Nov  5 11:47:45 2013
Last change: Tue Nov  5 11:44:12 2013 via cibadmin on pacemaker-1
Stack: openais
Current DC: pacemaker-2 - partition with quorum
Version: 1.1.7-ee0730e13d124c3d58f00016c3376a1de5323cffe
3 Nodes configured, 3 expected votes
1 Resources configured.
==============

Online: [ pacemaker-1 pacemaker-2 pacemaker-3 ]

writer_vip     (ocf::heartbeat:IPaddr2):      Started pacemaker-2
Pacemaker resource agents

- These are scripts, most of them are written in bash
- Implement: start/stop/monitor/meta-data and optionally promote/demote/notify
- Pacemaker passes parameters as environment variables
- Pacemaker reacts to return values
- Pacemaker reacts to timeouts
Introduction to booth

- Why? To detect network partitions on WAN
- An implementation of the Paxos protocol [link]
- A token distribution framework with timeout
- Multiple phases: Prepare, Promise, Accept Request, Accepted
- Very lightweight
- Similar in principle to DHCP, the IP being the token
- Needs 3 sites, the arbitrator is very light
Booth life cycle - 1

Initial state:
- All connected
- Token granted to Node 1
Network is cut between nodes: Nothing happens
Booth life cycle – 3

Network is cut between the node and the arbitrator:
- Nothing happens
Booth life cycle - 4

Node 1 is isolated:
- token will timeout
- arbitrator and node 2 have quorum
- Node 2 gets the token
Booth and Pacemaker

- Booth daemon run as a Pacemaker resource
- Booth daemon sets an attribute in CIB
- Logic can be built around the attribute value
PRM : basics

- Starts and stops MySQL
- Manages replication → change master
- Monitors MySQL, adjust reader VIPs for slaves
- If master crash, try to restart in place once
- Master promotion is based on slave lag, most up to date slave is promoted
- Can be used with semi-sync replication
PRM : basics-2

- Upcoming (within 6 months)
  - master score based on executed master pos
  - Support for 5.6 gtid
  - Slaves consistency with md5 hash of last relay log trx
PRM: Initial state
PRM: Node 1 promotion

**Node 1**
- MySQL started
- Promoted
- R/W
- W_VIP
- R_VIP1, R_VIP2

**Node 2**
- MySQL started
- R/O

**Node 3**
- MySQL started
- R/O

CIB repl_info=node1|binlog.123|34567
PRM : slaves notification

- Node 1
  - MySQL started
  - Promoted
  - R/W
  - W_VIP

- Node 2
  - MySQL started
  - Slave of Node 1
  - R/0
  - R_VIP2

- Node 3
  - MySQL started
  - Slave of Node 1
  - R/0
  - R_VIP1

CIB
repl_info=node1|binlog.123|34567
PRM : Node 1 failure

Node 1 - Crashed

CIB repl_info=node1|binlog.123|34567

Node 2
- MySQL started
- Most up to date
- R/O
- R_VIP2

Node 3
- MySQL started
- Was behind Node 2
- R/O
- R_VIP1
PRM : Node 2 promotion

Node 1
- Crashed

CIB
repl_info=node2|binlog.145|45678

Node 2
- MySQL started
- Promoted
- R/W
- W_VIP
- R_VIP1, R_VIP2

Node 3
- MySQL started
- R/O
PRM: slave notification

Node 1
- Crashed

CIB
repl_info=node2|binlog.145|45678

Node 2
- MySQL started
- Promoted
- R/W
- W_VIP
- R_VIP1

Node 3
- MySQL started
- R/0
- R_VIP2

www.percona.com
PRM : Node 1 back

Node 1
- MySQL started
- Slave of Node 2
- R/O
- R_VIP1

CIB
repl_info=node2|binlog.145|45678

Node 2
- MySQL started
- Promoted
- R/W
- W_VIP

Node 3
- MySQL started
- R/O
- R_VIP2
PRM : Geo-DR : Description

- 3 locations: Site 1, Site 2 and Site 3
- Site 1 and Site 2, distinct fully functional Pacemaker/PRM clusters, could be single node
- Site 3: Arbitrator site, very lite
- Priority to local failover
- Only one site writable
PRM : Geo-DR : Normal
PRM : Geo-DR : Local failover
PRM : Geo-DR : Site
PRM : Geo-DR : Our setup

• Site 1

  Network: 10.3.1.0/24
  pacemaker-1-1, database node, IP: 10.3.1.1
  pacemaker-1-2, database node, IP: 10.3.1.2
  booth-vip, 10.3.1.10
  writer-vip, 10.3.1.20
  reader-vips, 10.3.1.21, 10.3.1.22
PRM : Geo-DR : Our setup - 2

• Site 2
  Network: 10.3.2.0/24
  pacemaker-2-1, database node, IP: 10.3.2.1
  pacemaker-2-2, database node, IP: 10.3.2.2
  booth-vip, 10.3.2.10
  writer-vip, 10.3.2.20
  reader-vips, 10.3.2.21, 10.3.2.22

• Site 3
  pacemaker-3-1, arbitrator, IP: 10.3.3.1
PRM : Geo-DR : Prerequisites

- Two working PRM clusters in site 1 and 2
- Identical MySQL dataset
- MySQL grants adjusted for cross-site replication
- WAN open for SSH (22/tcp) and MySQL traffic (3306/tcp) and booth (6666/udp)
- ssh configured
- Booth packages installed
PRM : Geo-DR : SSH

• Root key based (or hacluster group) ssh
  Site-1 ↔ Site-2
• Tune the sshd, must be quick or timeout
• Define timeout in /root/.ssh/config

Host 10.3.*
  ConnectTimeout=2
  StrictHostKeyChecking no
PRM : Geo-DR : Booth packages

• Packages
  – Ubuntu 13.04+, Debian 7+
    apt-get install booth booth-pacemaker
  – EL6, rpm in Percona git repo, link in reference section
    booth-0.1.0-1.21.55ab.dirty.el6.x86_64.rpm
PRM : Geo-DR : Booth config

• /etc/booth/booth.conf
  transport="UDP"
  port="6666"
  arbitrator="10.3.3.1"
  site="10.3.1.10"
  site="10.3.2.10"
  ticket="ticketMaster;120"

• On the arbitrator, make boothd start automatically
PRM : Geo-DR : Pacemaker config

- Setup of the booth resource

```bash
primitive booth ocf:pacemaker:booth-site \
    meta resource-stickiness="INFINITY" \
    target-role="Started" \
    op monitor interval="10s" timeout="20s"
primitive booth-ip ocf:heartbeat:IPaddr2 \
    params ip="10.3.1.10" nic="eth0"
group g-booth booth-ip booth
order order-booth-ms_MySQL inf: g-booth
ms_MySQL:promote
```
PRM : Geo-DR : Pacemaker config - 2

- Addition of the ticket constraint
  - Pacemaker 1.1.10+, crm configure edit
    
    rsc_ticket ms-ticketMaster ticketMaster: \
    ms_MySQL:Master loss-policy=demote
  
  - Pacemaker 1.1.9-, direct xml edit
    <rsc_ticket id="ms-ticketMaster" loss-policy="demote"
    rsc="ms_MySQL" rsc-role="Master" ticket="ticketMaster"/>
• Connection of the 2 sites, parameters to add

```plaintext
geo_remote_IP="10.3.2.10"
booth_master_ticket="ticketMaster"
```
PRM : Geo-DR : Demo

Enough talking, build it!!!
References

- Setup documentation
- Operational Guide
References - 2

- Geo-DR guide
- Mailing lists
  https://groups.google.com/forum/?hl=en#!forum/prm-discuss
  https://groups.google.com/forum/?hl=en#!forum/prm-devel
• Booth RPM for EL6
  https://github.com/percona/percona-pacemaker-agents/raw/master/tools/booth/el6/booth-0.1.0-1.21.55ab.dirty.el6.x86_64.rpm
Questions ?