High Availability for RHEL on System z

David Boyes
Sine Nomine Associates

Agenda

- Clustering
- High Availability
- Cluster Management
- Failover
- Fencing
- Lock Management
- GFS2

Clustering

SINE NOMINE ASSOCIATES

- Four types
  - Storage
  - High Availability
  - Load Balancing
  - High Performance

High Availability

- Eliminate Single Points of Failure
- Failover
- Simultaneous Read/Write
- Node failures invisible outside the cluster
- rgmanager is the core software

High Availability

SINE NOMINE ASSOCIATES

- Major Components
  - Cluster infrastructure — Provides fundamental functions for nodes to work together as a cluster
    - Configuration-file management, membership management, lock management, and fencing
  - High availability Service Management — Provides failover of services from one cluster node to another in case a node becomes inoperative
  - Cluster administration tools — Configuration and management tools for setting up, configuring, and managing the High Availability Implementation

High Availability

SINE NOMINE ASSOCIATES

- Other Components
  - Red Hat GFS2 (Global File System 2) — Provides a cluster file system for use with the High Availability Add-On. GFS2 allows multiple nodes to share storage at a block level as if the storage were connected locally to each cluster node
  - Cluster Logical Volume Manager (CLVM) — Provides volume management of cluster storage
  - Load Balancer — Routing software that provides IP-Load-balancing

Cluster Infrastructure

SINE NOMINE ASSOCIATES

- Cluster management
- Lock management
- Fencing
- Cluster configuration management

Cluster Management

SINE NOMINE ASSOCIATES

- CMAN
  - Manages quorum and cluster membership
  - Distributed manager that runs in each node
  - Tracks membership and notifies other nodes

Failover Management

SINE NOMINE ASSOCIATES

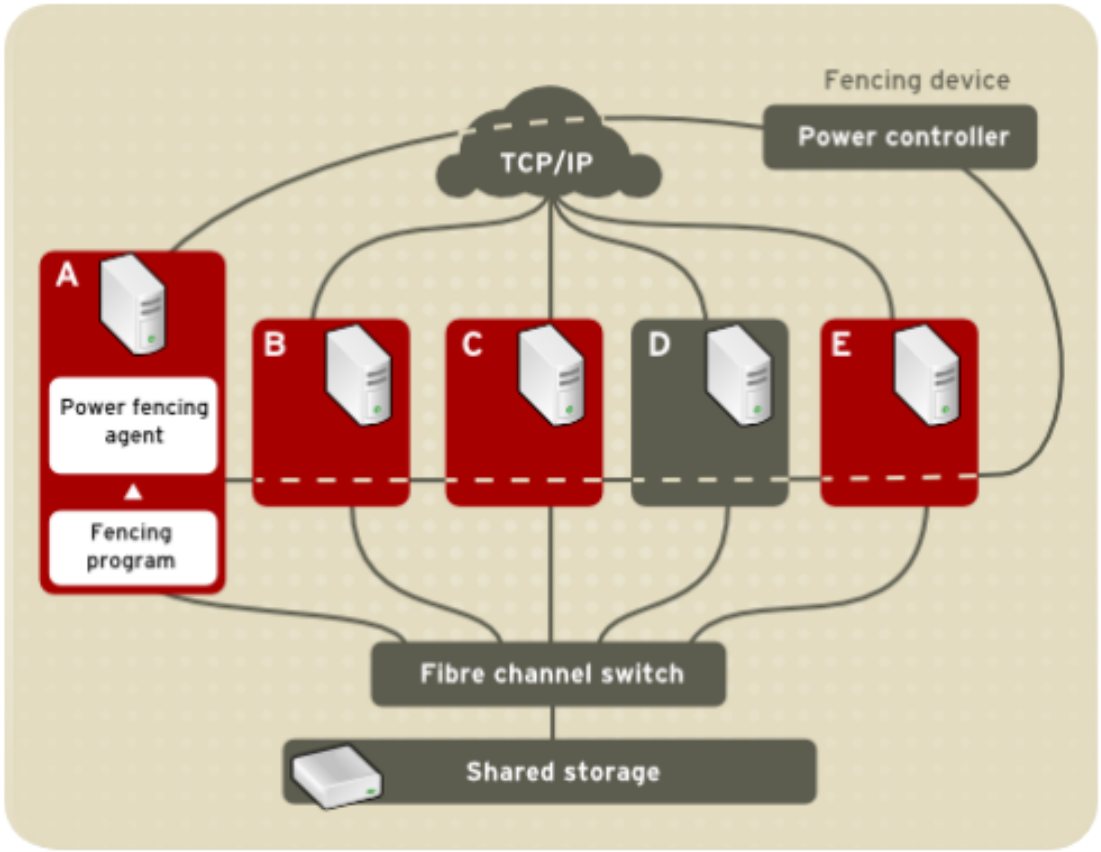
- Failover Domains - How the RGManager failover domain system work
- Service Policies - rgmanager's service startup and recovery policies
- Resource Trees - How rgmanager's resource trees work, including start/stop orders and inheritance
- Service Operational Behaviors - How rgmanager's operations work and what states mean
- Virtual Machine Behaviors - Special things to remember when running VMs in a rgmanager cluster
- ResourceActions - The agent actions RGManager uses and how to customize their behavior from the cluster.conf file.
- Event Scripting - If rgmanager's failover and recovery policies do not fit in your environment, you can customize your own using this scripting subsystem.

Fencing

SINE NOMINE ASSOCIATES

- The disconnection of a node from the cluster's shared storage. Fencing cuts off I/O from shared storage, thus ensuring data integrity
- The cluster infrastructure performs fencing through the fence daemon: fenced
- CMAN determines that a node has failed and communicates to other cluster-infrastructure components that the node has failed
- fenced, when notified of the failure, fences the failed node

Power Fencing



z/VM Power Fencing

SINE NOMINE ASSOCIATES

- Two choices of SMAPI-based fence devices
  - IUCV-based
  - TCP/IP
- Uses image_recycle API to fence a node
- Requires SMAPI configuration update to AUTHLIST:

```
Column 1              Column 66              Column 131
       |                     |                     |
       V                     V                     V
       XXXXXXXX              ALL                   IMAGE_OPERATIONS
```

Lock Management

SINE NOMINE ASSOCIATES

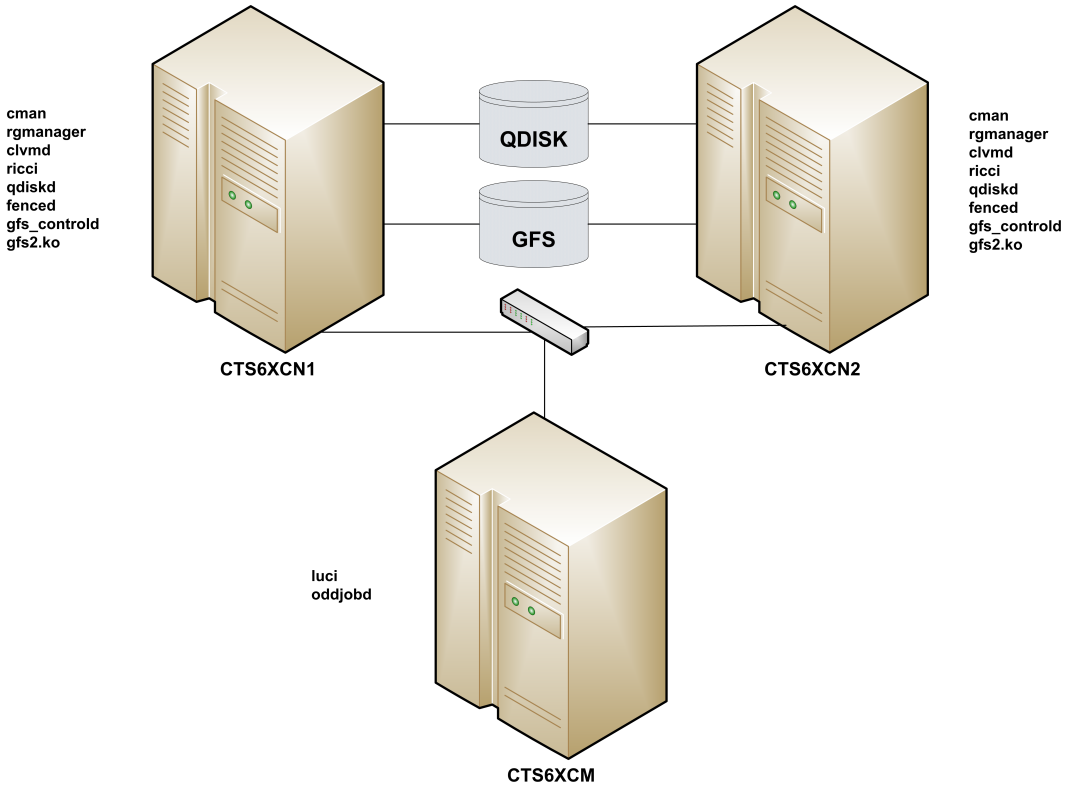
- Provides a mechanism for other cluster infrastructure components to synchronize their access to shared resources
- DLM – Distributed Lock Manager used in RHEL systems
- Lock management is distributed across all nodes in the cluster. GFS2 and CLVM use locks from the lock manager
- GFS2 uses locks from the lock manager to synchronize access to file system metadata (on shared storage)
- CLVM uses locks from the lock manager to synchronize updates to LVM volumes and volume groups (also on shared storage)
- rgmanager uses DLM to synchronize service states.

GFS2

SINE NOMINE ASSOCIATES

- A shared disk file system for Linux computer clusters
- GFS2 differs from distributed file systems (such as AFS, Coda, or InterMezzo) because it allows all nodes to have direct concurrent access to the same shared block storage
- GFS2 can also be used as a local filesystem.
- GFS has no disconnected operating-mode, and no client or server roles: All nodes in a GFS cluster function as peers
- Requires hardware to allow access to the shared storage, and a lock manager to control access to the storage
- GFS2 is a journaling file system

Sample Configuration



Sample Configuration

SINE NOMINE ASSOCIATES

```
USER CTS6XCN1 XXXXXXXX 768M 2G G
*FL= N
  ACCOUNT 99999999 GENERAL
  MACHINE ESA
  *AC= 99999999
  COMMAND SET VSWITCH VSWITCH2 GRANT &USERID
  COMMAND COUPLE C600 TO SYSTEM VSWITCH2
  IUCV VSMREQIU
  IPL CMS PARM AUTOCR FILEPOOL USER01
  CONSOLE   0009 3215 T OPERATOR
  SPOOL 00C 2540 READER *
  SPOOL 00D 2540 PUNCH A
  SPOOL 00E 1403 A
  LINK MAINT 190 190 RR
  LINK MAINT 19E 19E RR
  NICDEF C600 TYPE QDIO DEVICES 3
  MDISK 150 3390 3116 3338 CO510C MR
  MDISK 151 3390 6286 3338 CO5109 MR
  MDISK 153 3390 0001 3338 CO520E MW
  MDISK 200 3390 3007 0020 CO510F MW
```

```
USER CTS6XCN2 XXXXXXXX 768M 2G G 64
*FL= N
  ACCOUNT 99999999 LINUX
  MACHINE ESA
  *AC= 99999999
  COMMAND SET VSWITCH VSWITCH2 GRANT &USERID
  COMMAND COUPLE C600 TO SYSTEM VSWITCH2
  IUCV VSMREQIU
  IPL CMS PARM AUTOCR FILEPOOL USER01
  CONSOLE   0009 3215 T OPERATOR
  SPOOL 00C 2540 READER *
  SPOOL 00D 2540 PUNCH A
  SPOOL 00E 1403 A
  LINK MAINT 190 190 RR
  LINK MAINT 19E 19E RR
  LINK CTS6XCN1 153 152 MW
  LINK CTS6XCN1 200 200 MW
  NICDEF C600 TYPE QDIO DEVICES 3
  MDISK 150 3390 0001 3338 CO5204 MR
  MDISK 151 3390 4281 3338 CO5107 MR
```

Sample Configuration

SINE NOMINE ASSOCIATES

```
<?xml version="1.0"?>
<cluster config_version="23" name="SNATEST">
        <clusternodes>
                <clusternode name="cts6xcn1.devlab.sinenomine.net" nodeid="1">
                        <fence>
                                <method name="SMAPITCP">
                                        <device name="SMAPITCP" target="CTS6XCN1"/>
                                </method>
                        </fence>
                </clusternode>
                <clusternode name="cts6xcn2.devlab.sinenomine.net" nodeid="2">
                        <fence>
                                <method name="SMAPITCP">
                                        <device name="SMAPITCP" target="CTS6XCN2"/>
                                </method>
                        </fence>
                </clusternode>
        </clusternodes>
        <fencedevices>
                <fencedevice agent="fence_zvm" name="ZVMSMAPI" smapiserver="VSMREQIU"/>
                <fencedevice agent="fence_zvmip" authpass=”xxxxxx" authuser="CTS6XCN1" name="SMAPITCP" smapiserver="vm.devlab.sinenomine.net"/>
        </fencedevices>
        <cman expected_votes="3"/>
        <rm>
                <resources>
                        <clusterfs device="/dev/mapper/vg_snatest-gfs2" fsid="35269" fstype="gfs2" mountpoint="/mnt/gfs2" name="GFS2TEST"/>
                </resources>
                <service name="GFS2SERVICE" recovery="relocate">
                        <clusterfs ref="GFS2TEST"/>
                </service>
        </rm>
        <quorumd label="QDISK"/>
        <logging>
                <logging_daemon debug="on" logfile="/var/log/cluster/qdiskd.log" logfile_priority="debug" name="qdiskd"/>
        </logging>
</cluster>
```

Questions


SINE NOMINE
ASSOCIATES