Linux High Availability on IBM System z

Neale Ferguson
Sine Nomine Associates
High Availability

Neale Ferguson
Sine Nomine Associates
Tuesday 13 August, 2013
13857
Agenda

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

• Four types
  • Storage
  • High Availability
  • High Performance
  • Load Balancing – may be incorporated with previous two cluster types
High Availability

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

• 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

- 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

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

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

- The resource manager (rgmanager) manages and provides failover capabilities for collections of cluster resources called services, resource groups, or resource trees
- Allows administrators to define, configure, and monitor cluster services
- In the event of a node failure, rgmanager will relocate the clustered service to another node with minimal service disruption
Failover Management

- 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
- Resource Actions - 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

- 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

- 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:

```
<table>
<thead>
<tr>
<th></th>
<th>Column 66</th>
<th>Column 131</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td>V</td>
<td>V</td>
<td></td>
</tr>
<tr>
<td>XXXXXXXXX</td>
<td>ALL</td>
<td>IMAGE_OPERATIONS</td>
</tr>
</tbody>
</table>
```
# z/VM Power Fencing

<table>
<thead>
<tr>
<th>A</th>
<th>SMAPI Srv</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Node B fails

CP

Complete your sessions evaluation online at SHARE.org/BostonEval
z/VM Power Fencing

<table>
<thead>
<tr>
<th>A</th>
<th>SMAPI Srv</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node A detects node B is down</td>
<td>Node B fails</td>
<td></td>
</tr>
</tbody>
</table>

CP
z/VM Power Fencing

Node A detects node B is down
Uses SMAPI to recycle

Node B fails

CP
Node A detects node B is down. Node B fails. SMAPI forces Node B to recycle.
Node A detects node B is down
Uses SMAPI to recycle

SMAPI forces Node B Waits

Node B fails
Gets forced off

CP
z/VM Power Fencing

Node A detects node B is down
Uses SMAPI to recycle

SMAPI forces Node B
Waits Autologs Node B

Node B fails
Gets forced off

CP
z/VM Power Fencing

Node A detects node B is down
Uses SMAPI to recycle

Node B fails
Gets forced off
Recreated

CP
Lock Management

• 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

- 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

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 COS10C MR
MDISK 151 3390 6286 3338 COS109 MR
MDISK 153 3390 0001 3338 COS20E MW
MDISK 200 3390 3007 0020 COS10F 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 COS204 MR
MDISK 151 3390 4281 3338 COS107 MR
<?xml version="1.0"?>
<cluster config_version="52" 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="c13f0s" authuser="CTS6XCN1" name="SMAPITCP" smapiserver="vm.devlab.sinenomine.net"/>
  </fencedevices>
  <cman expected_votes="3"/>
</cluster>
...Sample Configuration

```xml
<rm>
  <resources>
    <apache config_file="conf/httpd.conf" name="SNA_WebServer" server_root="/etc/httpd" shutdown_wait="0"/>
    <clusterfs device="/dev/mapper/vg_snatest-gfs2" fsid="35269" fstype="gfs2" mountpoint="/var/www/html" name="SNA_GFS2"/>
    <ip address="172.17.16.185/24" sleeptime="3"/>
  </resources>
  <failoverdomains>
    <failoverdomain name="SNA_Failover">
      <failoverdomainnode name="cts6xcn2.devlab.sinenomine.net"/>
    </failoverdomain>
  </failoverdomains>
  <service domain="SNA_Failover" name="GFS2SERVICE" recovery="relocate">
    <clusterfs ref="SNA_GFS2"/>
    <ip ref="172.17.16.185/24"/>
    <apache ref="SNA_WebServer"/>
  </service>
</rm>
<brmont label="QDISK"/>
<logging>
  <logging_daemon debug="on" logfile="/var/log/cluster/qdiskd.log" logfile_priority="debug" name="qdiskd"/>
</logging>
<fence_daemon post_fail_delay="10"/>
</cluster>
```
Configuration using luci
...Configuration using luci...
...Configuration using luci...

<table>
<thead>
<tr>
<th>Nodes</th>
<th>Fence Devices</th>
<th>Failover Domains</th>
<th>Resources</th>
<th>Service Groups</th>
<th>Configure</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="add" alt="Add" /></td>
<td><img src="reboot" alt="Reboot" /></td>
<td><img src="join" alt="Join Cluster" /></td>
<td><img src="leave" alt="Leave Cluster" /></td>
<td><img src="delete" alt="Delete" /></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Node Name</th>
<th>Node ID</th>
<th>Votes</th>
<th>Status</th>
<th>Uptime</th>
<th>Hostname</th>
</tr>
</thead>
<tbody>
<tr>
<td>cts6xcn1.devlab.sinenomine.net</td>
<td>1</td>
<td>1</td>
<td>Cluster Member</td>
<td>00:00:59:54</td>
<td>cts6xcn1.devlab.sinenomine.net</td>
</tr>
<tr>
<td>cts6xcn2.devlab.sinenomine.net</td>
<td>2</td>
<td>1</td>
<td>Cluster Member</td>
<td>00:23:01:56</td>
<td>cts6xcn2.devlab.sinenomine.net</td>
</tr>
</tbody>
</table>

Select an item to view details
...Configuration using luci...
...Configuration using luci...

<table>
<thead>
<tr>
<th>Cluster Daemons</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>cman</td>
<td>Running</td>
</tr>
<tr>
<td>rgmanager</td>
<td>Running</td>
</tr>
<tr>
<td>ricci</td>
<td>Running</td>
</tr>
<tr>
<td>modclusterd</td>
<td>Running</td>
</tr>
<tr>
<td>clvmd</td>
<td>Running</td>
</tr>
</tbody>
</table>
...Configuration using luci...

<table>
<thead>
<tr>
<th>Name</th>
<th>Fence Type</th>
<th>Nodes Using</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZVMSMAPI</td>
<td>IBM z/VM – SSI</td>
<td>0</td>
</tr>
<tr>
<td>SMAPITCP</td>
<td>IBM z/VM – SMAPI using TCP/IP</td>
<td>2</td>
</tr>
</tbody>
</table>
Configuration using luci...

SMAPITCP
Type  IBM z/VM – SMAPI using TCP/IP

Fence Type
Name
SMAPI Server Virtual Machine Host Name
SMAPI Authorized User Name
SMAPI Authorized User Password

Apply

Nodes
<table>
<thead>
<tr>
<th>Node Name</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>cts6xcrn1.devlab.sinenomine.net</td>
<td>OK</td>
</tr>
<tr>
<td>cts6xcrn2.devlab.sinenomine.net</td>
<td>OK</td>
</tr>
</tbody>
</table>
…Configuration using luci…

### SNA_Failover

- **Prioritized**: Order the nodes to which services failover.
- **Restricted**: Service can run only on nodes specified.
- **No Fallback**: Do not send service back to 1st priority node when it becomes available again.

#### Services

- **GFS2 SERVICE**

#### Members

<table>
<thead>
<tr>
<th>Member</th>
<th>Priority</th>
</tr>
</thead>
<tbody>
<tr>
<td>cts6xcf1.devlab.sinenomine.net</td>
<td></td>
</tr>
<tr>
<td>cts6xcf2.devlab.sinenomine.net</td>
<td>✔️</td>
</tr>
</tbody>
</table>

**Update Settings**
...Configuration using luci...
...Configuration using luci...

SNA_WebServer

Apache

Name
Server Root
Config File
httpd Options
Shutdown Wait (seconds)

<table>
<thead>
<tr>
<th>SNA_WebServer</th>
</tr>
</thead>
<tbody>
<tr>
<td>/etc/httpd</td>
</tr>
<tr>
<td>conf/httpd.conf</td>
</tr>
<tr>
<td>0</td>
</tr>
</tbody>
</table>
Configuration using luci...

### SNA_GFS2

**GFS2**

<table>
<thead>
<tr>
<th>Name</th>
<th>SNA_GFS2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mount Point</td>
<td>/var/www/html</td>
</tr>
<tr>
<td>Device, FS Label, or UUID</td>
<td>/dev/mapper/vg_snatest-gfs2</td>
</tr>
<tr>
<td>Filesystem Type</td>
<td></td>
</tr>
<tr>
<td>Mount Options</td>
<td></td>
</tr>
<tr>
<td>Filesystem ID (optional)</td>
<td>35269</td>
</tr>
<tr>
<td>Force Unmount</td>
<td></td>
</tr>
<tr>
<td>Reboot Host Node if Unmount Fails</td>
<td></td>
</tr>
</tbody>
</table>
...Configuration using luci...

<table>
<thead>
<tr>
<th>IP Address</th>
<th>172.17.16.185/24</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP Address</td>
<td>172.17.16.185</td>
</tr>
<tr>
<td>Netmask Bits (optional)</td>
<td>24</td>
</tr>
<tr>
<td>Monitor Link</td>
<td>✓</td>
</tr>
<tr>
<td>Disable Updates to Static Routes</td>
<td></td>
</tr>
<tr>
<td>Number of Seconds to Sleep After Removing an IP Address</td>
<td>3</td>
</tr>
</tbody>
</table>
...Configuration using luci...
...Configuration using luci...

- **Nodes**
- **Fence Devices**
- **Failover Domains**
- **Resources**
- **Service Groups**
- **Configure**

<table>
<thead>
<tr>
<th>Name</th>
<th>Status</th>
<th>Autostart</th>
<th>Failover Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>GFS2SERVICE</td>
<td>Running on cts6xcn1.devlab.sinenomine.net</td>
<td>☑</td>
<td>SNA_Failover</td>
</tr>
</tbody>
</table>

GFS2SERVICE

Status: Running on cts6xcn1.devlab.sinenomine.net

Start on node...
...Configuration using luci...
...Configuration using luci...

Fence Daemon Properties
Post Fail Delay (seconds) 10
Post Join Delay (seconds) 3
...Configuration using luci...
...Configuration using luci...

Redundant Ring Protocol Configuration

Alternate Ring Multicast Address
Alternate Ring CMAN Port
Alternate Ring Multicast Packet TTL

Redundant Ring Cluster Node Configuration

Cluster Node
cts6xcn1.devlab.sinenomine.net
cts6xcn2.devlab.sinenomine.net

Alternate Name
...Configuration using luci...

Quorum Disk Configuration

- Do Not Use a Quorum Disk
- Use a Quorum Disk

Specify Physical Device

- By Device Label
  - QDISK
- By Filesystem Path to Device (deprecated)

Heuristics

<table>
<thead>
<tr>
<th>Path to Program</th>
<th>Interval</th>
<th>Score</th>
<th>TKO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Add Another Heuristic

Minimum Total Score

Reset values to defaults
...Configuration using luci...

Logging Configuration

Global Settings

Log Debugging Messages

Syslog

- Log Messages to Syslog: checked
- Syslog Message Facility: daemon
- Syslog Message Priority: info

Log File

- Log Messages to Log File: checked
- Log File Path: blank
- Log File Message Priority: info
Configuration using luci

Daemon-specific Logging Overrides

- rgmanager
  - Log rgmanager Debugging Messages
  - Syslog
    - Log rgmanager Messages to Syslog
    - rgmanager Syslog Message Facility: daemon
    - rgmanager Syslog Message Priority: info
  - Log File
    - Log rgmanager Messages to Log File
    - rgmanager Log File Path
    - rgmanager Log File Message Priority: info

qdiskd
Aug 07 15:26:05 rgmanager status on apache "SNA_WebServer" returned 7 (unspecified)
Aug 07 15:26:05 rgmanager Stopping service service:GFS2SERVICE
Aug 07 15:26:11 rgmanager [apache] Checking Syntax Of The File /etc/httpd/conf/httpd.conf
Aug 07 15:26:27 rgmanager [ip] Removing IPv4 address 172.17.16.154/24 from eth0
Aug 07 15:26:32 rgmanager [clusterfs] Not umounting /dev/dm-3 (clustered file system)
Aug 07 15:26:32 rgmanager Service service:GFS2SERVICE is recovering
Aug 07 15:28:20 rgmanager Service service:GFS2SERVICE is now running on member 1
Aug 07 15:26:33 rgmanager *Recovering failed service service:GFS2SERVICE*
Aug 07 15:26:59 rgmanager [ip] Link for eth0: Detected
Aug 07 15:27:03 rgmanager [ip] Adding IPv4 address 172.17.16.185/24 to eth0
Aug 07 15:27:06 rgmanager [ip] Pinging addr 172.17.16.185 from dev eth0
Aug 07 15:27:11 rgmanager [ip] Sending gratuitous ARP: 172.17.16.185 02:00:00:00:00:15 brd ff:ff:ff:ff:ff:ff
:  
Aug 07 15:27:40 rgmanager [apache] Looking For IP Addresses
Aug 07 15:27:45 rgmanager [apache] 1 IP addresses found for GFS2SERVICE/SNA_WebServer
Aug 07 15:27:49 rgmanager [apache] Looking For IP Addresses > Succeed - IP Addresses Found
Aug 07 15:27:54 rgmanager [apache] Checking: SHA1 checksum of config file /etc/cluster/apache/
apache:SNA_WebServer/httpd.conf
apache:SNA_WebServer/httpd.conf From /etc/httpd/conf/httpd.conf
apache:SNA_WebServer/httpd.conf From /etc/httpd/conf/httpd.conf > Succ
Aug 07 15:28:20 rgmanager **Service service:GFS2SERVICE started**