

# 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

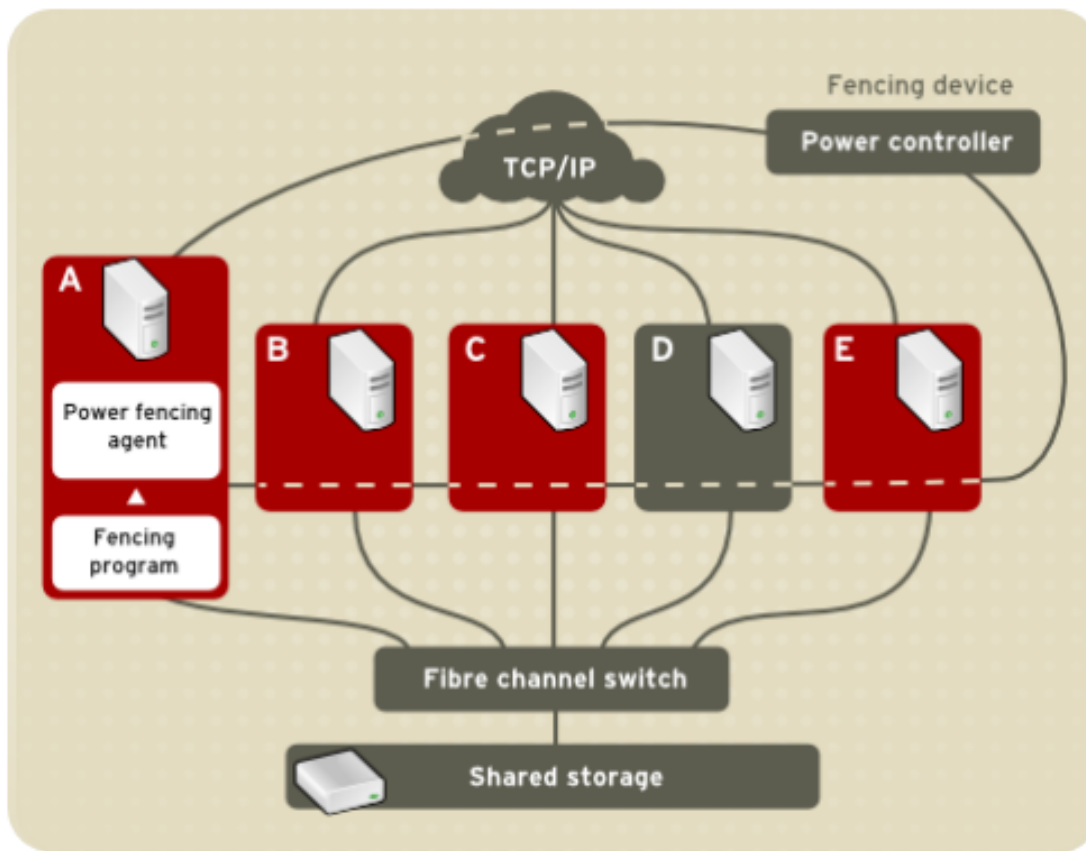
# Failover Management

- 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:

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

# z/VM Power Fencing

A

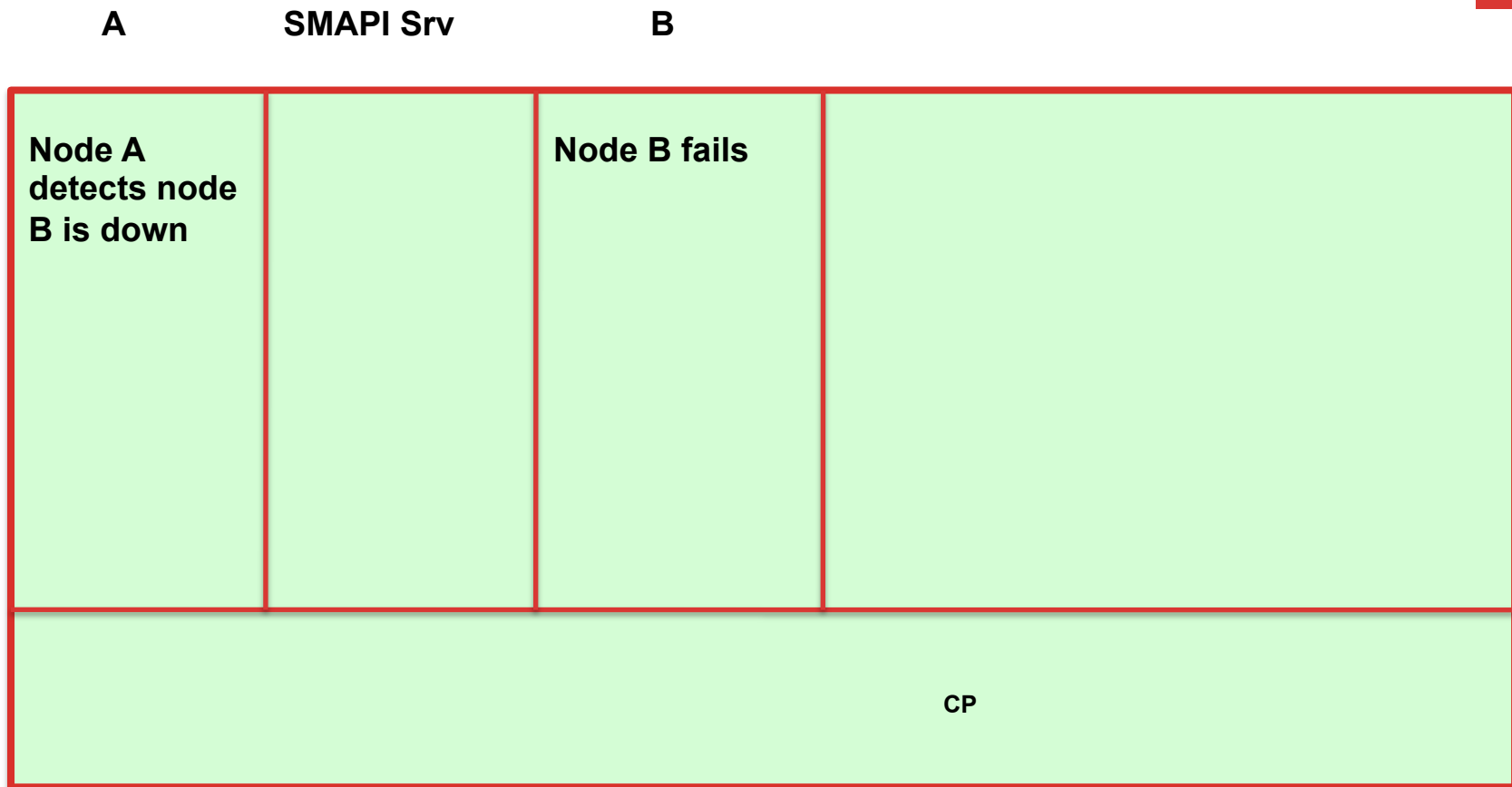
SMAPI Srv

B

Node B fails

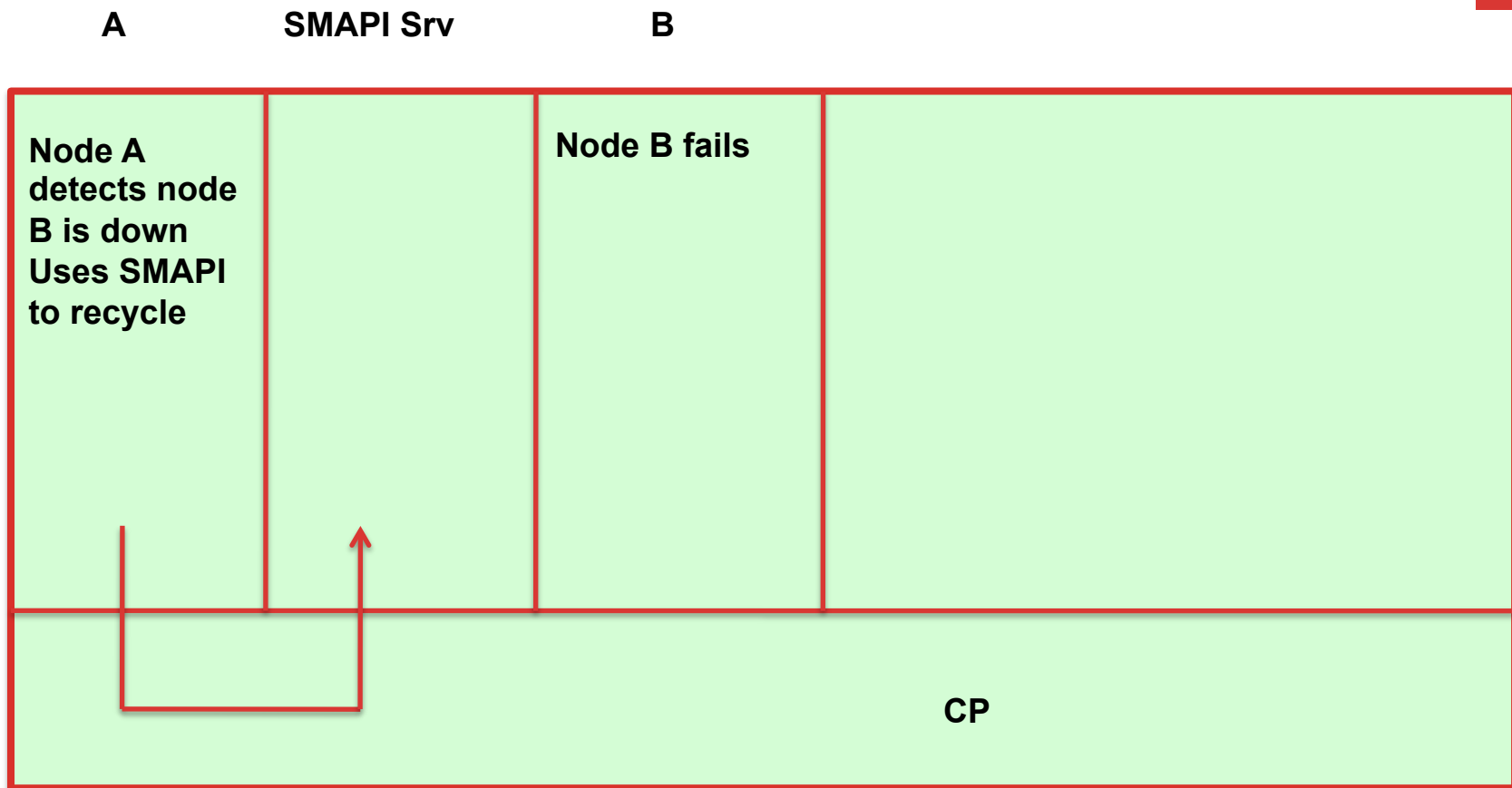
CP

# z/VM Power Fencing

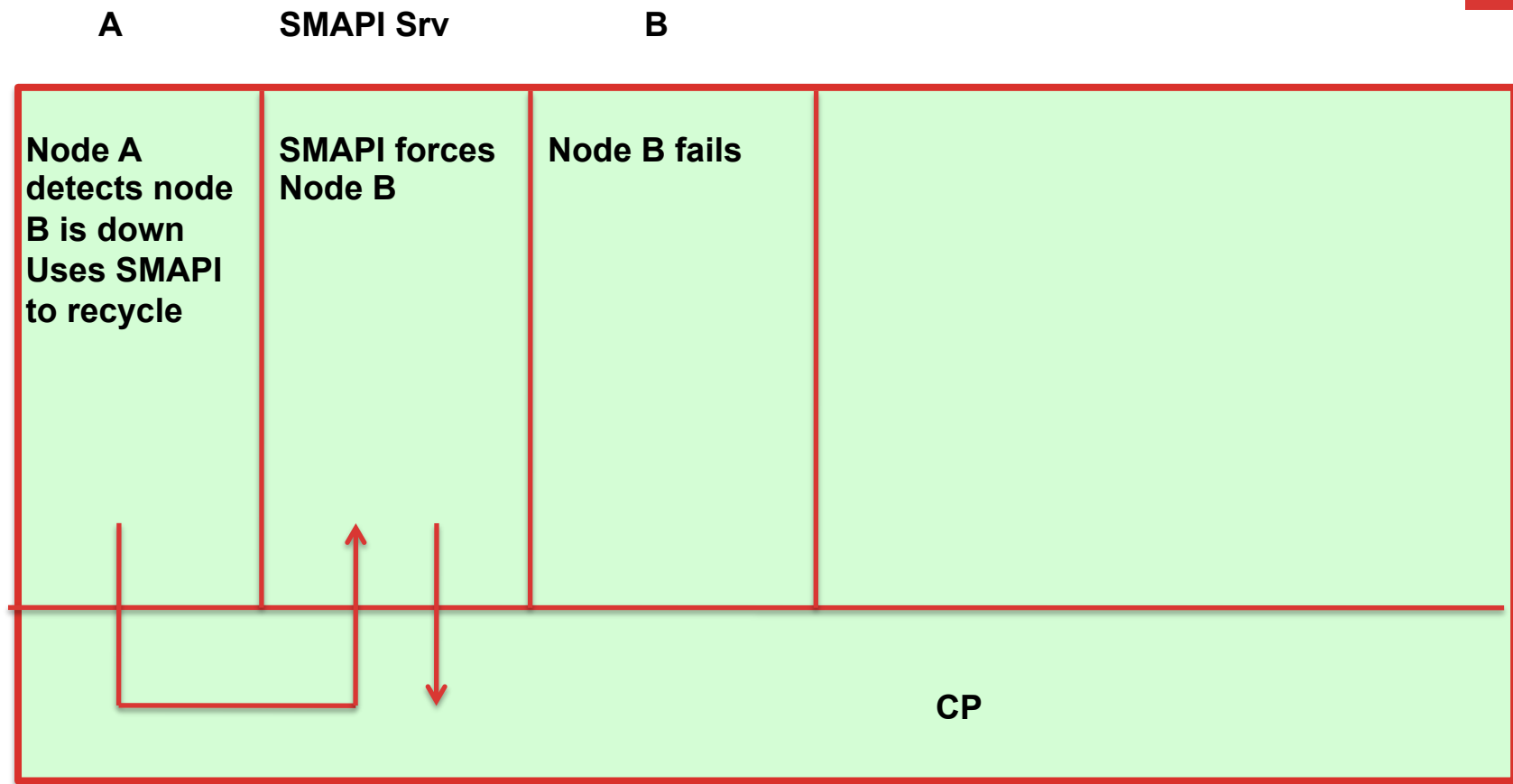




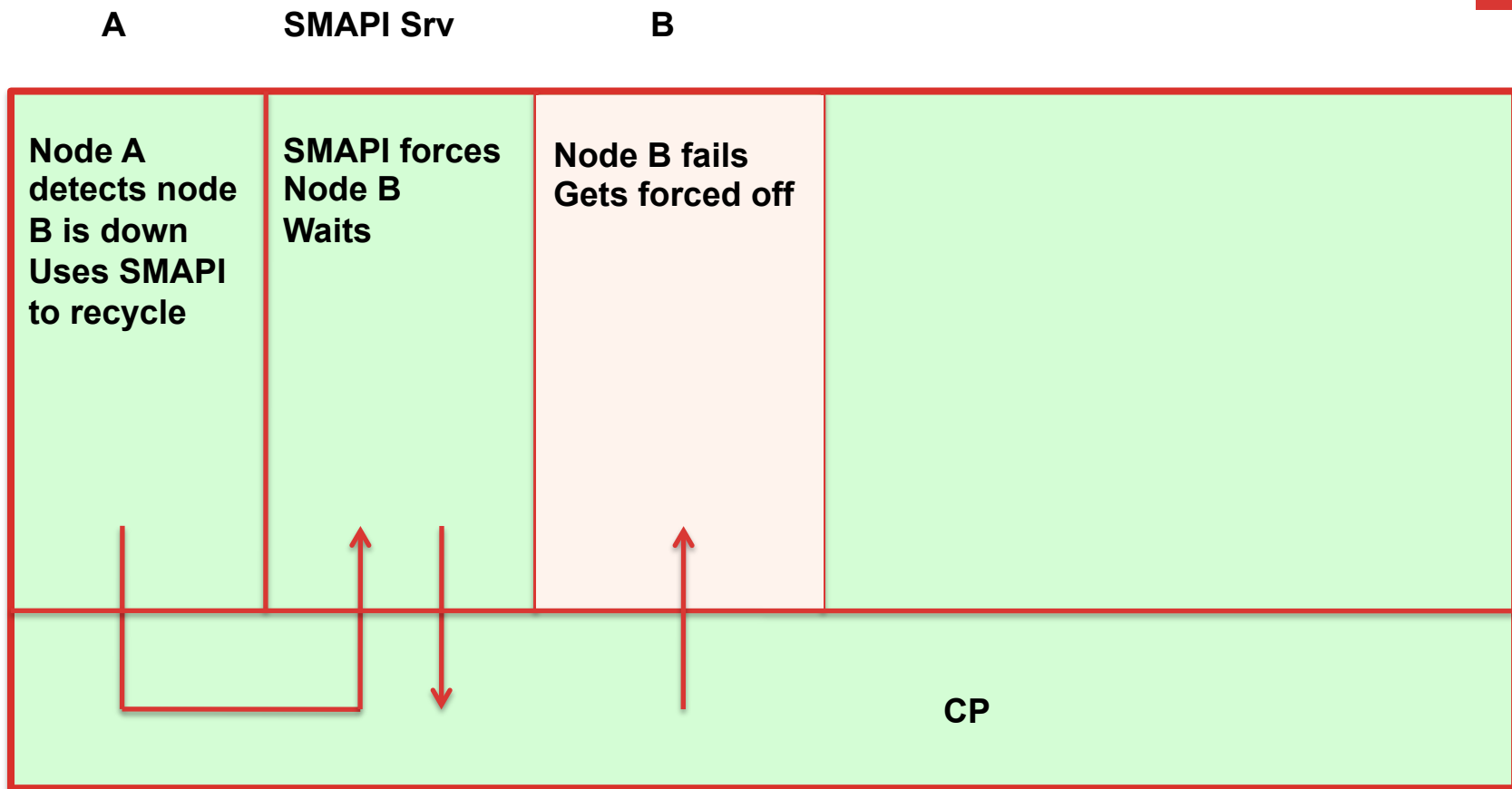
# z/VM Power Fencing



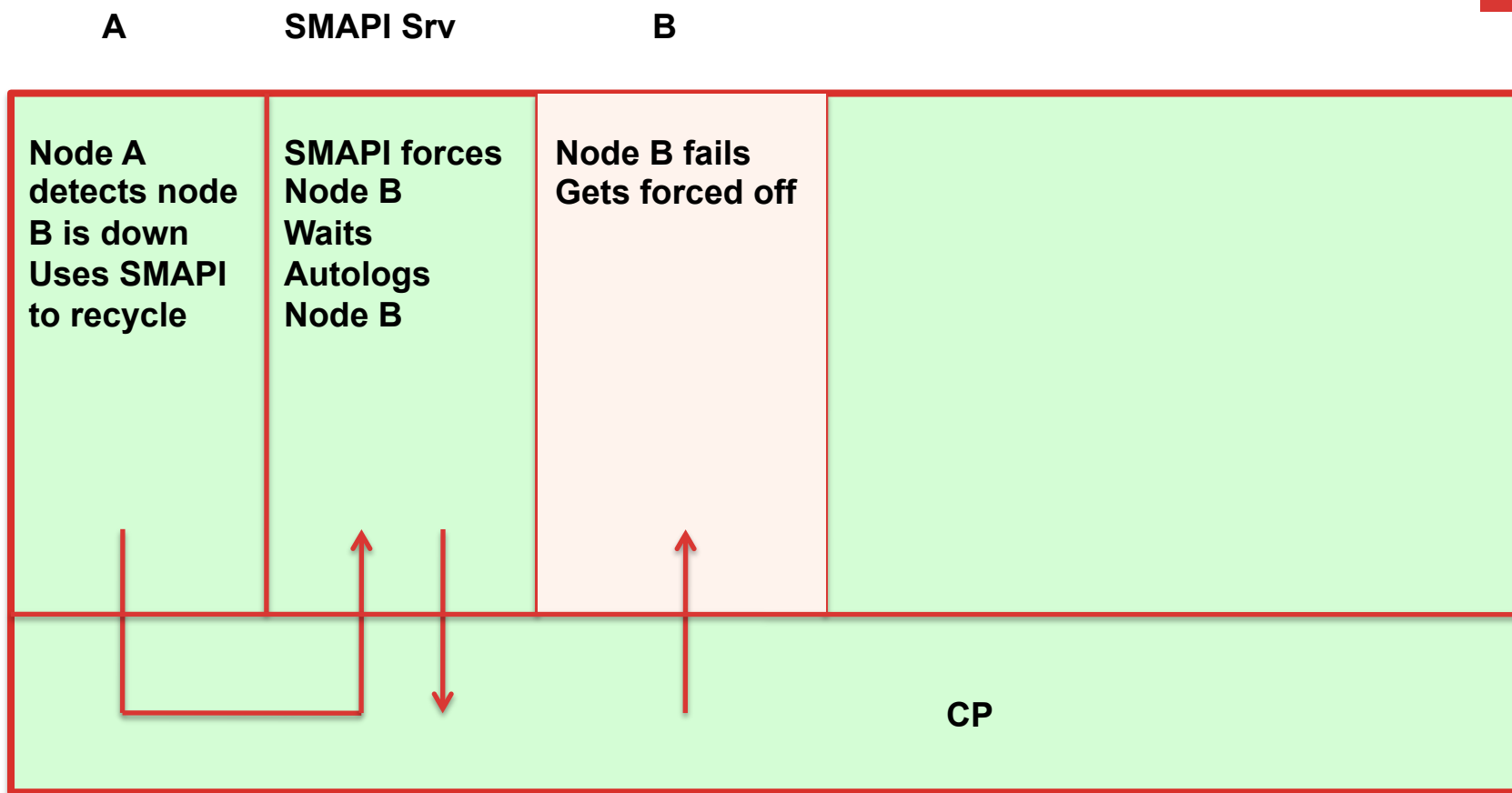
# z/VM Power Fencing



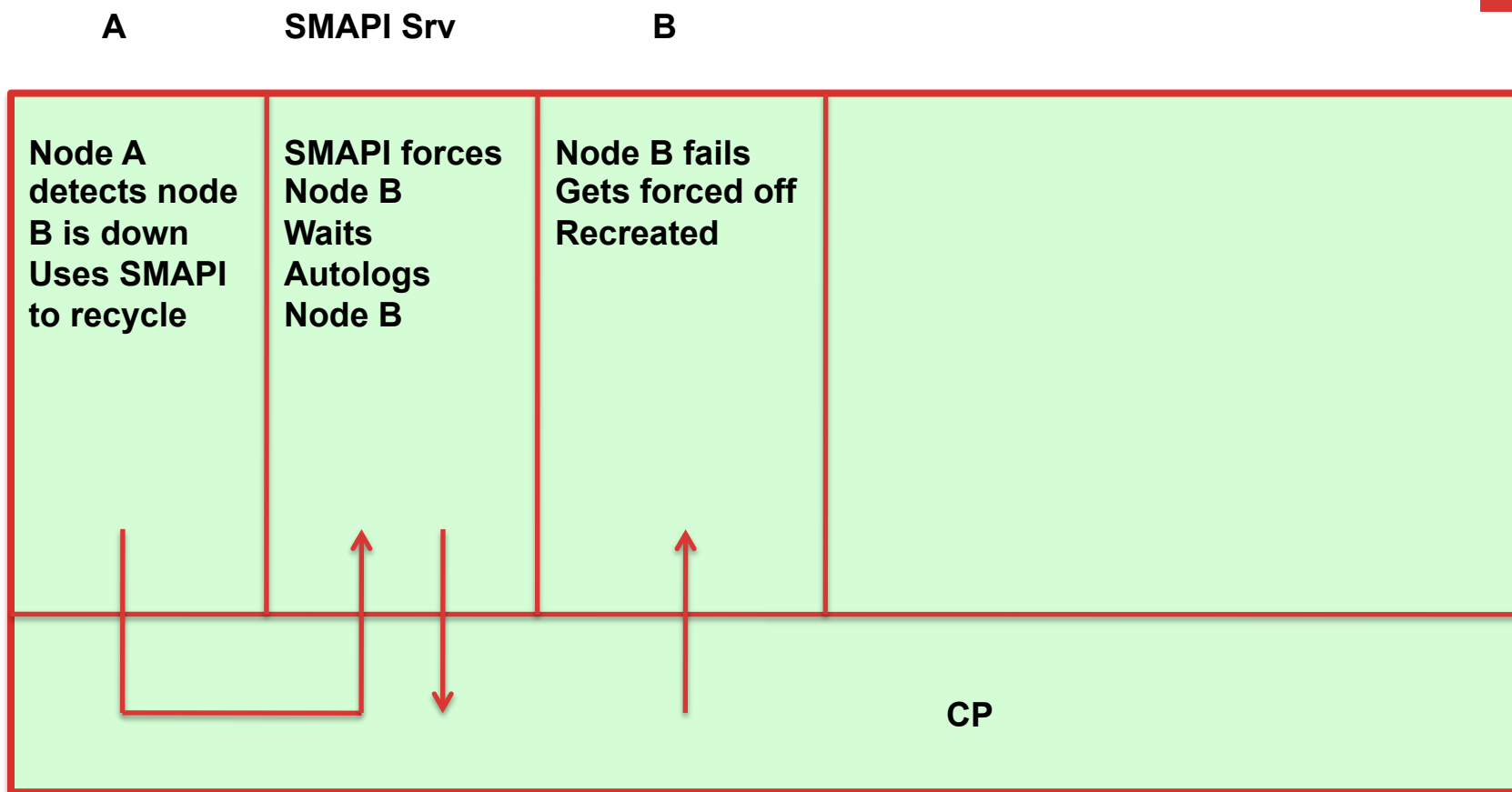
# z/VM Power Fencing



# z/VM Power Fencing



# z/VM Power Fencing



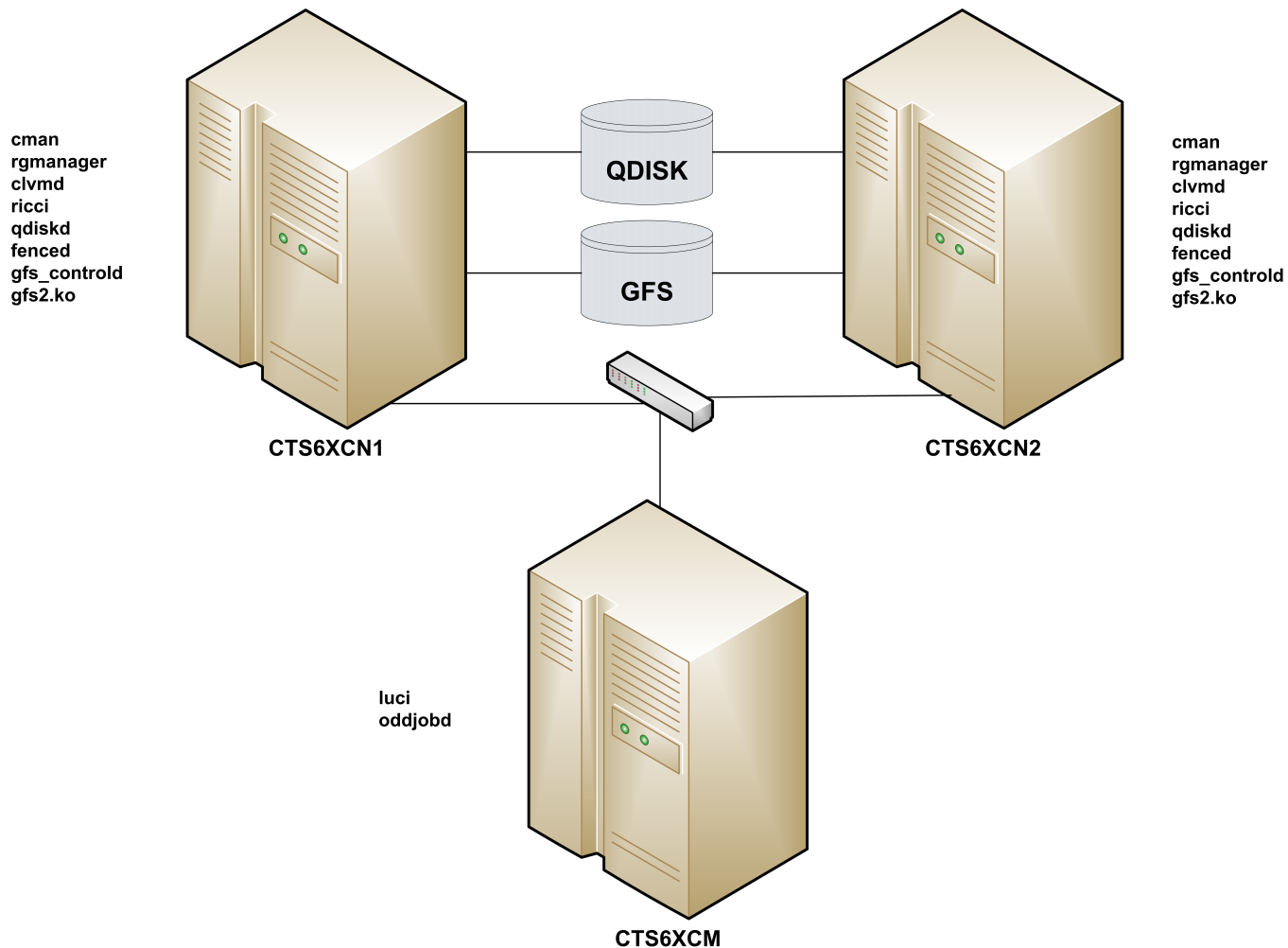
# Lock Management

- 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 AUTO CR 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 AUTO CR 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
```

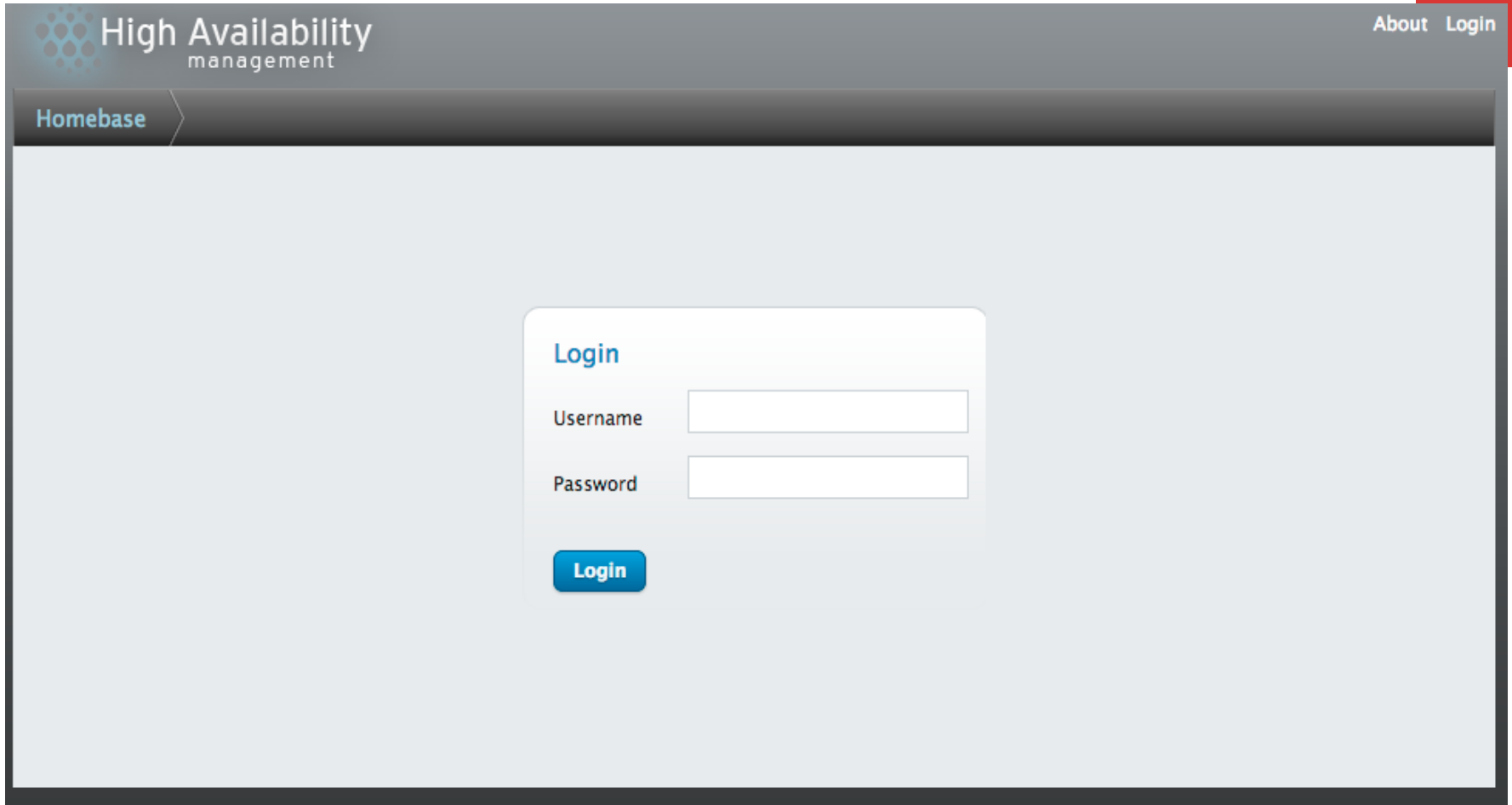
# Sample Configuration...

```
<?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="VSMREQUIU"/>
    <fencedevice agent="fence_zvmip" authpass="c13f0s" authuser="CTS6XCN1" name="SMAPITCP" smapiserver="vm.devlab.sinenomine.net"/>
  </fencedevices>
  <cman expected_votes="3"/>
</cluster>
```

# ...Sample Configuration

```
<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>
<quorumd 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



The screenshot displays the 'High Availability management' web interface. The header includes a logo on the left and 'About Login' links on the right. A 'Homebase' tab is visible on the left side of the main content area. In the center, there is a 'Login' form with the following elements:

- Login** (title)
- Username** label next to a text input field.
- Password** label next to a text input field.
- Login** button.

# ...Configuration using luci...

Homebase	CLUSTER SUMMARY		
Manage Clusters	Name	Status	Nodes Joined
	SNATEST	Quorate	2 of 2

# ...Configuration using luci...

Nodes Fence Devices Failover Domains Resources Service Groups Configure						
+ Add ⚙ Reboot 🔗 Join Cluster ⚙ Leave Cluster ✕ Delete						
!	Node Name	Node ID	Votes	Status	Uptime	Hostname
<input type="checkbox"/>	cts6xcn1.devlab.sinenomine.net	1	1	Cluster Member	00:00:59:54	cts6xcn1.devlab.sinenomine.net
<input type="checkbox"/>	cts6xcn2.devlab.sinenomine.net	2	1	Cluster Member	00:23:01:56	cts6xcn2.devlab.sinenomine.net
Select an item to view details						

# ...Configuration using luci...

cts6xcn1.devlab.sinenomine.net

Status Cluster Member

Properties

Update Properties

Number of votes

1

ricci host

cts6xcn1.devlab.siner

ricci port

11111

Services

GFS2SERVICE

Failover Domains

Priority

Fence Devices

Method

SMAPITCP

Remove

Name

Type/Values

SMAPITCP

IBM z/VM – SMAPI using TCP/IP  
target : CTS6XCN1

Add Fence Instance

Add Fence Method

Cluster Daemons

Status

cman

Running

rgmanager

Running

ricci

Running

modclusterd

Running

clvmd

Running

# ...Configuration using luci...

## Cluster Daemons

	Status
cman	Running
rgmanager	Running
ricci	Running
modclusterd	Running
clvmd	Running



# ...Configuration using luci...

Nodes	Fence Devices	Failover Domains	Resources	Service Groups	Configure
<div><div>+</div> Add <div>×</div> Delete</div>					
Name		Fence Type	Nodes Using		
<input type="checkbox"/>	ZVMSMAPI	IBM z/VM – SSI	0		
<input type="checkbox"/>	SMAPITCP	IBM z/VM – SMAPI using TCP/IP	2		

# ...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

IBM z/VM – SMAPI using TCP/IP

SMAPITCP

vm.devlab.sinenomine.net

CTS6XCN1

\*\*\*\*\*

Apply

Nodes

! Node Name

Status

cts6xcn1.devlab.sinenomine.net

OK

cts6xcn2.devlab.sinenomine.net

OK

# ...Configuration using luci...

+ Add ✕ Delete

Name	Prioritized	Restricted
<input type="checkbox"/> SNA_Failover	No	No

### SNA\_Failover

✕

Update Properties

☐ Prioritized

Order the nodes to which services failover.

☐ Restricted

Service can run only on nodes specified.

☐ No Failback

Do not send service back to 1st priority node when it becomes available again.

#### Services

	GFS2SERVICE
--	-------------

#### Members

Update Settings

Member	Priority
cts6xcn1.devlab.sinenomine.net	<input type="checkbox"/>
cts6xcn2.devlab.sinenomine.net	<input checked="" type="checkbox"/>

# ...Configuration using luci...

Nodes

Fence Devices

Failover Domains

Resources

Service Groups

Configure

+ Add

x Delete

	Name/IP	Type	In Use
<input type="checkbox"/>	SNA_WebServer	Apache Server	✓
<input type="checkbox"/>	SNA_GFS2	GFS2	✓
<input type="checkbox"/>	172.17.16.185/24	IP Address	✓

# ...Configuration using luci...

## SNA\_WebServer

---

### Apache

Name	SNA_WebServer
Server Root	/etc/httpd
Config File	conf/httpd.conf
httpd Options	
Shutdown Wait (seconds)	0

# ...Configuration using luci...

SNA\_GFS2

## GFS2

Name

SNA\_GFS2

Mount Point

/var/www/html

Device, FS Label, or UUID

/dev/mapper/vg\_snatest-gfs2

Filesystem Type

GFS2

Mount Options

Filesystem ID (optional)

35269

Force Unmount

☐

Reboot Host Node if Unmount Fails

☐

# ...Configuration using luci...

172.17.16.185/24

## IP Address

IP Address

Netmask Bits (optional)

Monitor Link



Disable Updates to Static Routes



Number of Seconds to Sleep After Removing an IP Address

# ...Configuration using luci...

Nodes					Fence Devices					Failover Domains					Resources					Service Groups					Configure				
+ Add					▶ Start					↺ Restart					■ Disable					✕ Delete									
! Name					Status										Autostart					Failover Domain									
<input type="checkbox"/>					GFS2SERVICE					Running on cts6xcn1.devlab.sinenomine.net										<input checked="" type="checkbox"/>					SNA_Failover				



# ...Configuration using luci...

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

[+ Add](#) [▶ Start](#) [↺ Restart](#) [■ Disable](#) [✕ Delete](#)

!	Name	Status	Autostart	Failover Domain
<input type="checkbox"/>	GFS2SERVICE	Running on cts6xcn1.devlab.sinenomine.net	<input checked="" type="checkbox"/>	SNA_Failover

**GFS2SERVICE**

Status Running on cts6xcn1.devlab.sinenomine.net

[▶](#) [↺](#) [■](#) [✕](#)

# ...Configuration using luci...

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

General

Fence Daemon

Network

Redundant Ring

QDisk

Logging

## General Properties

Cluster Name

SNATEST

Configuration Version

52

**Apply**

# ...Configuration using luci...

General

Fence Daemon

Network

Redundant Ring

QDisk

Logging

## Fence Daemon Properties

Post Fail Delay (seconds)

10

Post Join Delay (seconds)

3

# ...Configuration using luci...

General

Fence Daemon

Network

Redundant Ring

QDisk

Logging

## Network Configuration

Network Transport Type

☒ UDP Multicast and Let Cluster Choose the Multicast Address

☐ UDP Multicast and Specify the Multicast Address Manually

☐ UDP Unicast (UDPU)

Multicast Address

# ...Configuration using luci...

General

Fence Daemon

Network

Redundant Ring

QDisk

Logging

## 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...

General

Fence Daemon

Network

Redundant Ring

QDisk

Logging

### Quorum Disk Configuration

☐ Do Not Use a Quorum Disk

☒ Use a Quorum Disk

#### Specify Physical Device

☒ By Device Label

☐ By Filesystem Path to Device (deprecated)

#### Heuristics

Path to Program	Interval	Score	TKO
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>

Minimum Total Score

# ...Configuration using luci...

General

Fence Daemon

Network

Redundant Ring

QDisk

Logging

## Logging Configuration

### Global Settings

Log Debugging Messages ☐

### Syslog

Log Messages to Syslog ☒

Syslog Message Facility 

daemon

Syslog Message Priority 

info

### Log File

Log Messages to Log File ☒

Log File Path

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

rgmanager Syslog Message Priority

### Log File

Log rgmanager Messages to Log File ☒

rgmanager Log File Path

rgmanager Log File Message Priority

► qdiskd



# Failover...

```
Aug 07 15:26:02 rgmanager [apache] Checking Existence Of File /var/run/cluster/apache/
apache:SNA_WebServer.pid [apache:SNA_WebServer] > Failed
Aug 07 15:26:05 rgmanager [apache] Monitoring Service apache:SNA_WebServer > Service Is Not
Running
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 15:26:08 rgmanager [apache] Verifying Configuration Of apache:SNA_WebServer
Aug 07 15:26:11 rgmanager [apache] Checking Syntax Of The File /etc/httpd/conf/httpd.conf
Aug 07 15:26:14 rgmanager [apache] Checking Syntax Of The File /etc/httpd/conf/httpd.conf >
Succeed
Aug 07 15:26:17 rgmanager [apache] Stopping Service apache:SNA_WebServer
Aug 07 15:26:21 rgmanager [apache] Checking Existence Of File /var/run/cluster/apache/
apache:SNA_WebServer.pid [apache:SNA_WebServer] > Failed - File DoAug 07 15:26:23 rgmanager
[apache] Stopping Service apache:SNA_WebServer > Succeed
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
```

# Failover...

```
Aug 07 15:26:33 rgmanager Recovering failed service service:GFS2SERVICE
Aug 07 15:26:41 rgmanager [clusterfs] mounting /dev/dm-6 on /var/www/html
Aug 07 15:26:44 rgmanager [clusterfs] mount -t gfs2 /dev/dm-6 /var/www/html
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 15:27:18 rgmanager [apache] Verifying Configuration Of apache:SNA_WebServer
:
Aug 07 15:27:37 rgmanager [apache] Starting Service apache:SNA_WebServer
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
Aug 07 15:27:59 rgmanager [apache] Checking: SHA1 checksum > succeed
Aug 07 15:28:04 rgmanager [apache] Generating New Config File /etc/cluster/apache/
apache:SNA_WebServer/httpd.conf From /etc/httpd/conf/httpd.conf
Aug 07 15:28:12 rgmanager [apache] Generating New Config File /etc/cluster/apache/
apache:SNA_WebServer/httpd.conf From /etc/httpd/conf/httpd.conf > SuccAug 07 15:28:18 rgmanager
[apache] Starting Service apache:SNA_WebServer > Succeed
Aug 07 15:28:20 rgmanager Service service:GFS2SERVICE started
```