

S H A R E

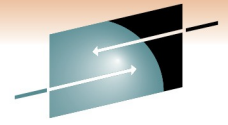
Technology • Connections • Results

Implementing the SUSE Linux Enterprise High Availability Extension on System z

Mike Friesenegger
Novell

Monday, February 28, 2011
Session Number: 8474





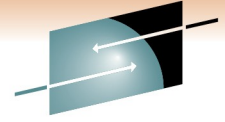
Agenda

- What is a high availability (HA) cluster?
- What is required to build an HA cluster using SLES?
- Demoing the details
 - Managing a cluster with the GUI and CLI
 - Resources primitives and resource groups
 - Resource Constraints
 - STONITH
 - cLVM and OCFS2
- Wrap-up/Questions

If you attended this session at SHARE Boston in 2010...



- Thank you for coming back!
- Did anyone setup a HA cluster?
 - How did it go?
- I have revamped the presentation and the demos
 - So no one should be bored!!
 - My goal is that everyone will learn new things about SLE HAE
 - Hopefully some attendees are interested in setting up SLE HAE on SLES when you get back!!
- If you want hands-on then attend the “SuSE Linux High Availability Extensions Hands-on Workshop”
 - Monday 3pm – 6pm (two sessions)
 - Hosted by Richard Lewis of IBM!

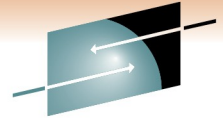


SHARE

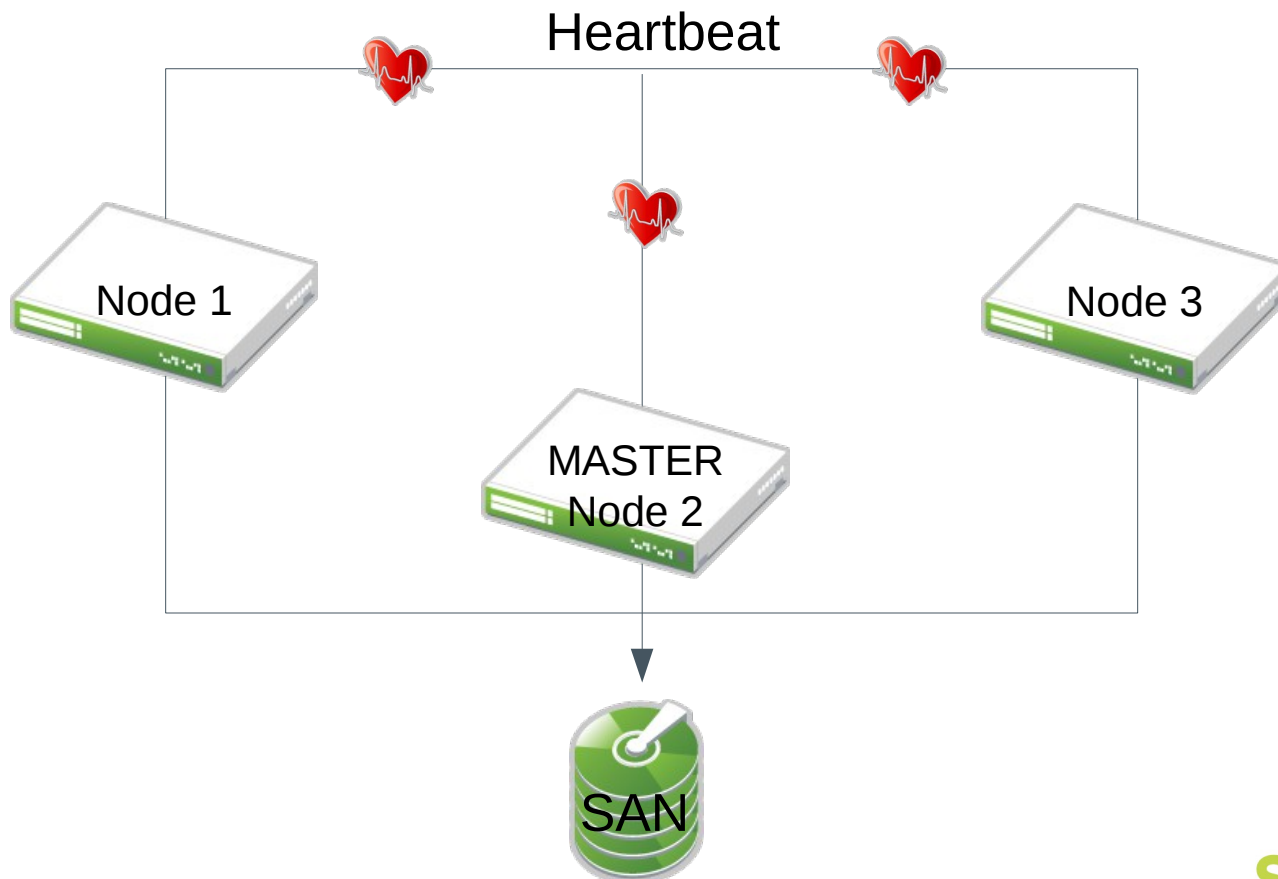
Technology • Connections • Results

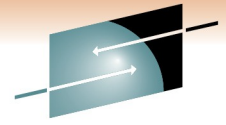
What is a high availability (HA) cluster?



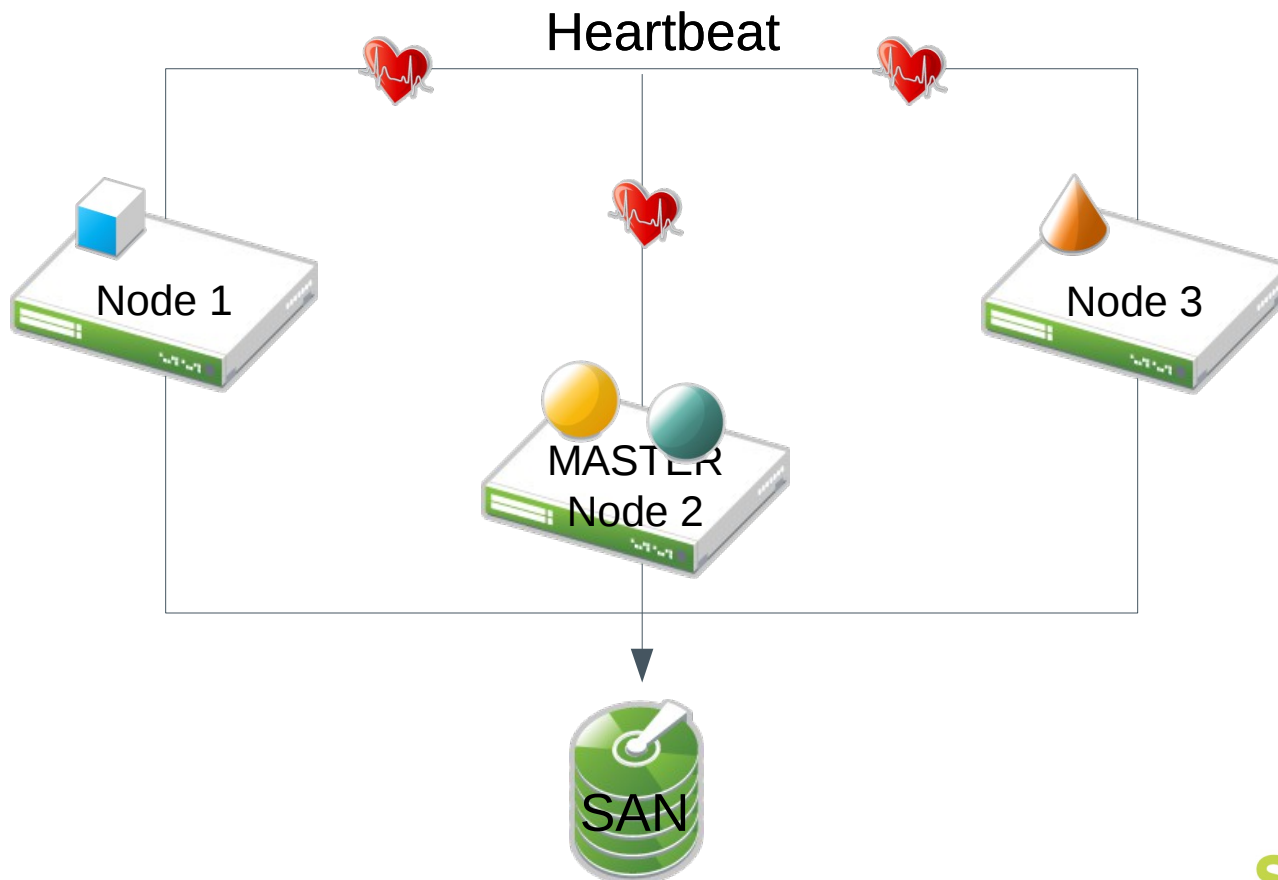


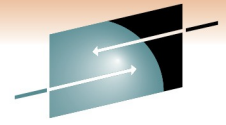
A Simple HA Cluster



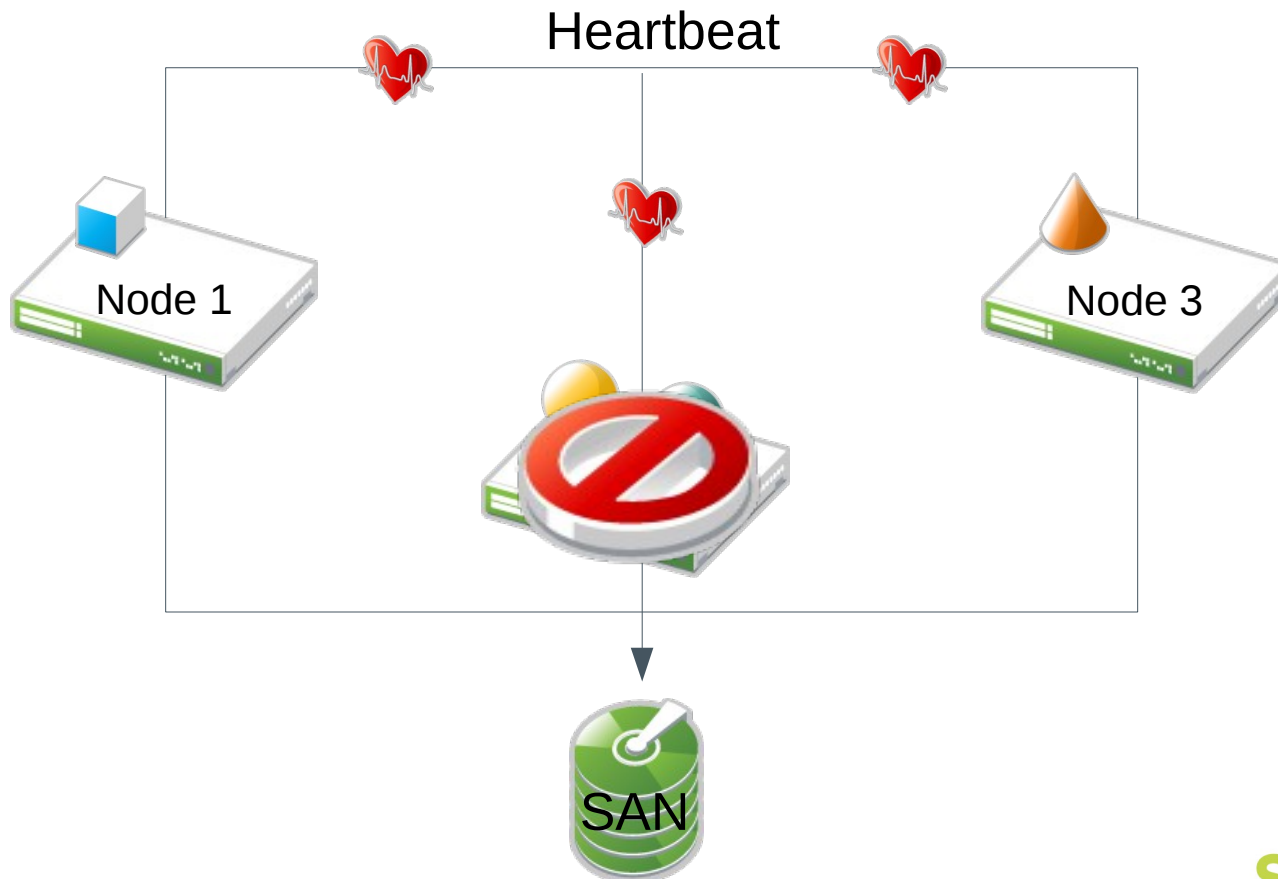


Resources Running in the Cluster

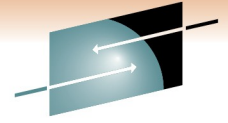




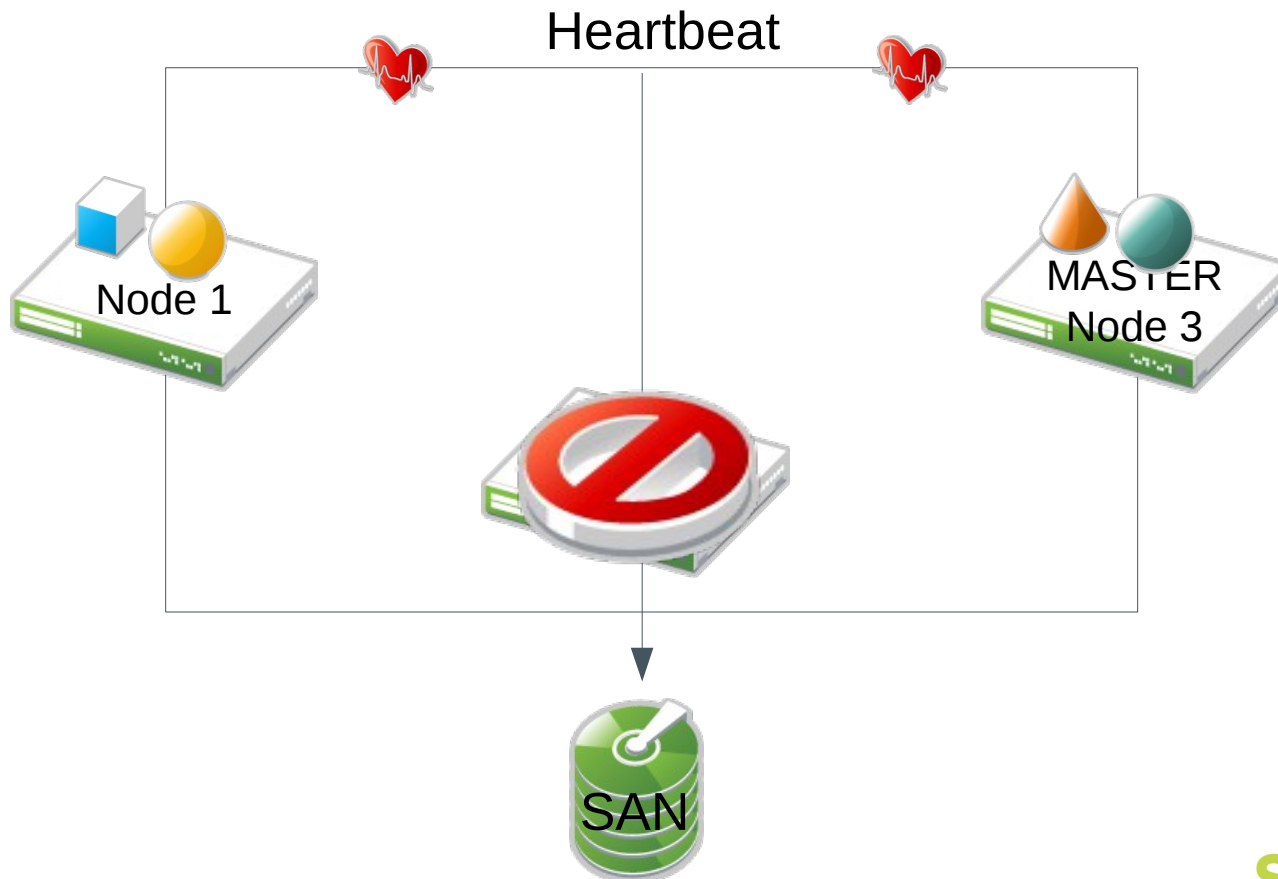
Node Failure in the Cluster

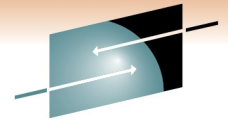


Services brought up on other nodes the Cluster



SHARE
Technology • Connections • Results





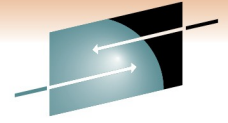
SHARE

Technology • Connections • Results

What is required to build an HA cluster using SLES?



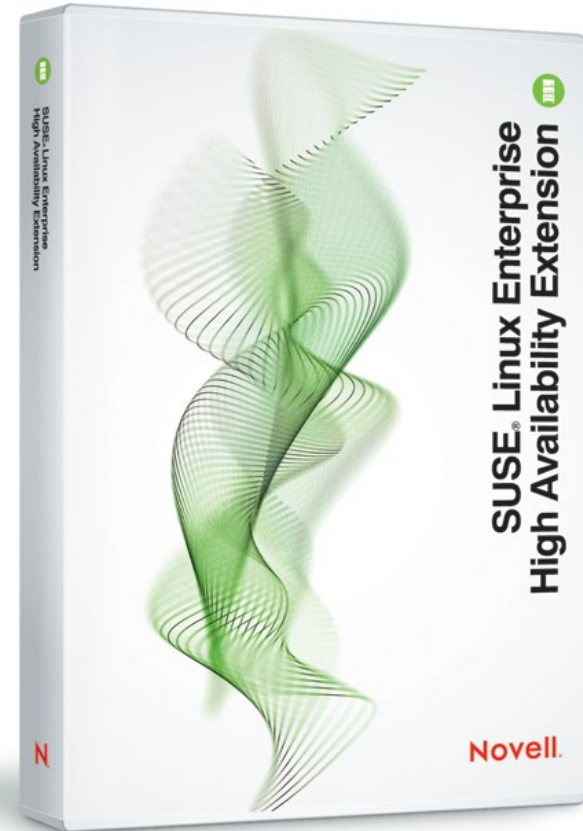
SUSE® Linux Enterprise High Availability Extension



SHARE
Technology • Connections • Results

An affordable, integrated suite of robust open source clustering technologies that enables you to implement highly available physical and virtual Linux clusters.

Used with SUSE Linux Enterprise Server, it helps you maintain business continuity, protect data integrity, and reduce unplanned downtime for your mission critical Linux workloads



SHARE
in Anaheim
2011

SUSE® Linux Enterprise High Availability Extension



Features

Flexible, policy-driven clustering solution

–OpenAIS

- » *Open Source initiative's certified implementation of the Service Availability Forum Application Interface Specification*
- » *Leading standards-based communication protocol for server and storage clustering*
- » *Messaging and membership layer*

–Pacemaker

- » *Cluster resource manager to continuously monitor resource health, manage dependencies, and automatically stop and start services*
- » *Configurable engine that uses rules and policies*
- » *Metro Area Cluster up to 20 miles*
- » *Clustered SAMBA (CIFS)*

SUSE® Linux Enterprise High Availability Extension



Features cont.

Resource Agents

- For popular third-party applications included at no extra charge
 - » *SAP Instance and Database, IBM WebSphere Application Server, DB2, and Informix, Oracle and VMware*
- For popular open source applications included at no extra charge
 - » *Apache, Ipv4 and IPv6, LVM, RAID, Pure-FTPd, Route, ServeRAID, Squid, VIPArip, Xen, Xinted, DRBD, Novell eDirectory™, iscsi, mysql, nfsserver, and postgres, sfex, tomcat, filesystems*
- For the most up to date list of resource agents, visit:
www.novell.com/products/highavailability

SUSE® Linux Enterprise High Availability Extension

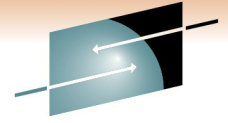


Features cont.

Cluster aware file system and volume manager

- OCFS2 (Oracle Cluster File System)
 - » *Shared-disk POSIX-compliant generic cluster file system*
 - » *Cluster-aware POSIX locking*
 - » *Parallel I/O*
 - » *Dual node use with DRBD*
- CLVM2 (Clustered Logical Volume Manager)
 - » *Convenient, single, cluster-wide view of storage*
 - » *Clustering extensions to the standard LVM2 toolset*
 - » *Eliminates need to learn a new set of tools*

SUSE® Linux Enterprise High Availability Extension



SHARE
Technology • Connections • Results

Features cont.

Host Based Continuous data replication

- DRBD8 (Distributed Replicated Block Device)
 - » *Leading open source networked disk management tool*
 - » *Build single partitions from multiple disks that mirror each other*
 - » *Fast data resynchronization capabilities*
 - » *Supports both synchronous and asynchronous mirroring*
 - » *Provides replicated storage area network (SAN) semantics, allowing cluster-aware file systems to be used without additional SANs*

SUSE® Linux Enterprise High Availability Extension



Features cont.

User-friendly tools

-Unified command line interface

- » *Powerful tool for installing, configuring and managing Linux clusters*
- » *For more experienced IT professionals*

-Graphical user interface

- » *Simple tool for monitoring and administering clustered environment*
- » *Does not require in-depth knowledge*
- » *Web Interface included*

-YaST modules

- » *DRBD*
- » *OpenAIS*
- » *Multipath*
- » *IP load balancer*

SUSE® Linux Enterprise High Availability Extension

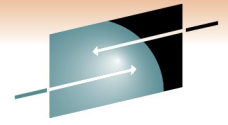


x86 and x86_64

- Additional cost per year, per server
- Support level inherited by base SUSE Linux Enterprise Server

System z, Power, Itanium

- Bundled with base SUSE Linux Enterprise Server at no additional charge
- Support level inherited by base SUSE Linux Enterprise Server



S H A R E

Technology • Connections • Results

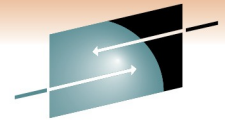
Demoing the details:

Managing a cluster with the GUI and CLI

NOTE: Its hard to show a demo in slides so come to the session if you want to see the live demo!
Several screenshots are provided to help visualize the demoed topic.

SHARE
in Anaheim
2011





Start the GUI with `crm_gui`

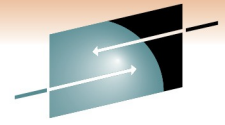
The screenshot shows the Pacemaker GUI interface. The window title is "Pacemaker GUI". The menu bar includes "Connection", "View", "Shadow", "Tools", and "Help". The left sidebar shows a tree view under "Live" with categories: Configuration (CRM Config, Resource Defaults, Operation Defaults), Nodes, Resources, Constraints, and Management (selected). The main area displays a table of cluster components:

Name	Status	Details
Cluster	● have quorum	Openais & Pacemaker
s390vmi03	● online	
s390vm36	● online (dc)	
Resources	●	
stonith_sbd	● running on [s390vmi03]	stonith::external/sbd
nfs_res	● running on [s390vm36]	ocf::heartbeat:Filesystem
ip-apache_group	● group	
ip_res	● running on [s390vm36]	ocf::heartbeat:IPAddr2
apache_res	● running on [s390vm36]	lsb::apache2
htdocs_res	● running on [s390vm36]	ocf::heartbeat:Filesystem

Below the table, the following status information is displayed:

Validate With: pacemaker-1.2
Epoch: 376
Num Updates: 23
CRM Feature Set: 3.0.2
Have Quorum: 1
DC UUID: s390vm36

At the bottom of the window, it says "Connected to 127.0.0.1 (Simple Mode)".



SHARE

Technology • Connections • Results

Use crm_mon and crm for the CLI

```
File Edit View Terminal Help
=====
Last updated: Tue Feb 15 14:21:27 2011
Stack: openais
Current DC: s390vm36 - partition with quorum
Version: 1.1.2-ecb1e2ea172ba2551f0bd763e557fccde68c849b
2 Nodes configured, 2 expected votes
4 Resources configured.
=====

Online: [ s390vmi03 s390vm36 ]

stonith_sbd (stonith:external/sbd): Started s390vmi03
nfs_res (ocf::heartbeat:Filesystem): Started s390vm36
Resource Group: ip-apache_group
  ip_res (ocf::heartbeat:IPAddr2): Started s390vm36
  apache_res (lsb:apache2): Started s390vm36
htdocs_res (ocf::heartbeat:Filesystem): Started s390vm36

```

This is the CRM command line interface program.

Available commands:

- | | |
|---------------|------------------------------------|
| cib | manage shadow CIBs |
| resource | resources management |
| node | nodes management |
| options | user preferences |
| configure | CRM cluster configuration |
| ra | resource agents information center |
| status | show cluster status |
| quit,bye,exit | exit the program |
| help | show help |
| end,cd,up | go back one level |

crm(live)#

```
File Edit View Terminal Help
s390vm36:/ # crm
crm(live)# help

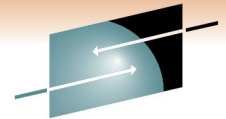
This is the CRM command line interface program.

Available commands:

cib                manage shadow CIBs
resource           resources management
node               nodes management
options            user preferences
configure          CRM cluster configuration
ra                 resource agents information center
status             show cluster status
quit,bye,exit      exit the program
help               show help
end,cd,up          go back one level

crm(live)#

s390vm36:/ # crm resource show
stonith_sbd (stonith:external/sbd) Started
nfs_res (ocf::heartbeat:Filesystem) Started
Resource Group: ip-apache_group
  ip_res (ocf::heartbeat:IPAddr2) Started
  apache_res (lsb:apache2) Started
htdocs_res (ocf::heartbeat:Filesystem) Started
s390vm36:/ #
```



S H A R E

Technology • Connections • Results

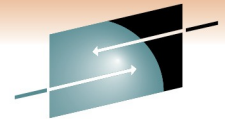
Demoing the details:

Resource primitives and resource groups

**NOTE: Its hard to show a demo in slides so come to the session if you want to see the live demo!
Several screenshots are provided to help visualize the demoed topic.**

SHARE
in Anaheim
2011





A resource primitive

The screenshot displays the Pacemaker GUI with the 'Edit Primitive' dialog box open. The main window shows a list of resource primitives, and the dialog provides detailed configuration for the selected 'nfs_res' primitive.

Pacemaker GUI - Primitive List

ID	Class	Provider	Type	Description
stonith_sbd	stonith		external/sbd	
nfs_res	ocf	heartbeat	Filesystem	
htdocs_res	ocf	heartbeat	Filesystem	

Edit Primitive - Required Fields

ID: nfs_res
Class: ocf
Provider: heartbeat
Type: Filesystem

Optional

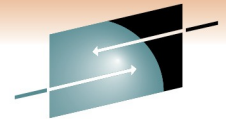
Description
Manages filesystem mounts.
Resource script for Filesystem. It manages a Filesystem on a shared storage medium

Instance Attributes

Name	Value
device	10.10.0.100:/dist
directory	/nfsmnt
fstype	

ID: nfs_res-instance_attributes-device
Name: device
Value: 10.10.0.100:/dist

Buttons: Add, Edit, Remove, Cancel, Reset, OK

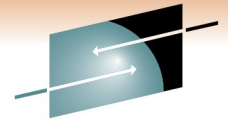


A resource group

The screenshot displays the Pacemaker GUI interface. On the left is a navigation tree under 'Live' with categories: Configuration, CRM Config, Resource Defaults, Operation Defaults, Nodes, Resources (selected), Constraints, and Management. The main area shows a 'Group' tab for 'ip-apache_group'. An 'Edit Group' dialog is open, showing the 'Optional' section with a table of meta-attributes:

ID	Class	Provider	Type	Description
ip_res	ocf	heartbeat	IPAddr2	
apache_res	lsb		apache2	

Below the table, the selected primitive 'ip_res' is detailed with its properties: ID: ip_res, Class: ocf, Provider: heartbeat, Type: IPAddr2. The dialog includes 'Up' and 'Down' buttons for the table, and 'Add', 'Edit', 'Remove', 'Cancel', 'Reset', and 'OK' buttons at the bottom.



S H A R E

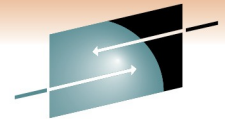
Technology • Connections • Results

Demoing the details: Resource Constraints

NOTE: Its hard to show a demo in slides so come to the session if you want to see the live demo!
Several screenshots are provided to help visualize the demoed topic.

SHARE
in Anaheim
2011





A resource location constraint

Pacemaker GUI

Connection View Shadow Tools Help

Live

- Configuration
 - CRM Config
 - Resource Defaults
 - Operation Defaults
 - Nodes
 - Resources
 - Constraints**
 - Management

Show: List Mode

ID	Resource	Score	Node
nfs_res-loc	nfs_res	INFINITY	s390vm36

Edit Resource Location

Show: List Mode

Required

ID: nfs_res-loc

Resource: nfs_res

Score: INFINITY

Node: s390vm36

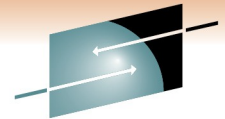
+ Add Edit - Remove

Cancel Reset OK

ID: nfs_res-loc
Resource: nfs_res
Score: INFINITY
Node: s390vm36

+ Add Edit - Remove

Connected to 127.0.0.1 (Simple Mode)



A resource colocation constraint

The screenshot displays the Pacemaker GUI with the 'Edit Resource Colocation' dialog box open. The dialog shows the configuration for a resource colocation constraint with the following details:

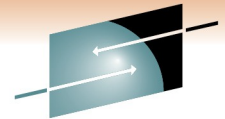
- Required:**
 - ID: htdocs_ip-apache_colo
 - Resource: ip-apache_group
 - With Resource: htdocs_res
- Optional:**
 - Score: INFINITY
 - Score Attribute: (empty)
 - Score Attribute Mangle: (empty)
 - Node Attribute: (empty)
 - Resource Role: (empty)
 - With Resource Role: (empty)
- Description:**
 - * Make ip-apache_group on the same node as htdocs_res (ip-apache_group according to htdocs_res)
 - * If htdocs_res cannot be on any node, then ip-apache_group won't be anywhere
 - * If ip-apache_group cannot be on any node, htdocs_res won't be affected

The background window shows the 'Resource Colocation' tab with a table listing the constraint:

ID	Score	Score Attribute	Score Mangle
htdocs_ip-apache_colo	INFINITY		

At the bottom of the background window, the constraint details are repeated:

ID: htdocs_ip-apache_colo
Score: INFINITY
Resource: ip-apache_group
With Resource: htdocs_res



A resource order constraint

The screenshot displays the Pacemaker GUI interface. On the left, a navigation pane shows a tree structure under 'Live' with categories: Configuration (CRM Config, Resource Defaults, Operation Defaults), Nodes, Resources, Constraints (selected), and Management. The main window has tabs for 'Resource Location', 'Resource Colocation', and 'Resource Order'. The 'Resource Order' tab contains a table with columns: ID, Symmetrical, Score, Kind, First, Then, First Action, and Then. A single entry is visible: ID: htdocs_ip-apache_order, Symmetrical: (checkbox), Score: (input), Kind: (input), First: htdocs_res, Then: ip-apache_group, First Action: (input), Then: (input). Below the table, a summary shows: ID: htdocs_ip-apache_order, First: htdocs_res, Then: ip-apache_group. An 'Edit Resource Order' dialog box is open, showing the configuration for the selected constraint. It includes fields for ID, First, and Then, and a 'Description' section with the following text: '* Start htdocs_res before start ip-apache_group', '* If cannot start htdocs_res, do not start ip-apache_group', '* Stop ip-apache_group before stop htdocs_res', '* If cannot stop ip-apache_group, do not stop htdocs_res'. The dialog also features buttons for '+ Add', 'Edit', '- Remove', 'Cancel', 'Reset', and 'OK'.

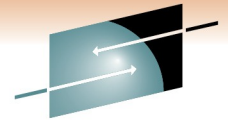
ID	Symmetrical	Score	Kind	First	Then	First Action	Then
htdocs_ip-apache_order				htdocs_res	ip-apache_group		

ID: htdocs_ip-apache_order
First: htdocs_res
Then: ip-apache_group

Required
ID: htdocs_ip-apache_order
First: htdocs_res
Then: ip-apache_group

Optional

Description
* Start htdocs_res before start ip-apache_group
* If cannot start htdocs_res, do not start ip-apache_group
* Stop ip-apache_group before stop htdocs_res
* If cannot stop ip-apache_group, do not stop htdocs_res



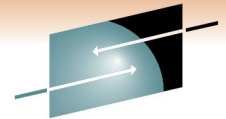
SHARE
Technology • Connections • Results

Demoing the details:

STONITH – Shoot The Other Node In The Head

NOTE: Its hard to show a demo in slides so come to the session if you want to see the live demo!
Several screenshots are provided to help visualize the demoed topic.

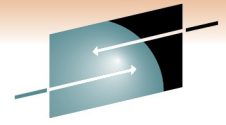




What is STONITH?

- Simple concept
 - A machine in the cluster wants to make sure another machine in the cluster is dead
 - STONITH is used to remotely power down a node in the cluster
 - Simple and reliable, albeit admittedly brutal
- Fencing is another term but not as graphic!
- Modular and extensible
 - 33 STONITH modules included in SLE11 SP1 HAE
- SLE HAE requires a STONITH device by default
 - Recommended practice to have one configured!

A Split Brain Detector (SBD) STONITH resource



SHARE
Technology • Connections • Results

The screenshot displays the Pacemaker GUI interface. The main window shows a list of resources under the 'Resources' tab. The 'stonith_sbd' resource is selected, and an 'Edit Primitive' dialog box is open, showing the configuration for this resource.

Pacemaker GUI

Connection View Shadow Tools Help

Live

Configuration

- CRM Config
- Resource Defaults
- Operation Defaults
- Nodes
- Resources**
- Constraints

Primitive Group

ID	Class	Provider	Type	Description
stonith_sbd	stonith		external/sbd	
nfs_res	ocf	heartbeat	Filesystem	
htdocs_res	ocf	heartbeat	Filesystem	

Show: List Mode

Up

Edit Primitive

Show: List Mode

Meta Attributes Instance Attributes Operations

Name	Value
sbd_device	/dev/disk/by-id/scsi-1IBM_2105_71526069-par

Up

Down

ID: stonith_sbd
Class: stonith
Provider:
Type: external/sbd

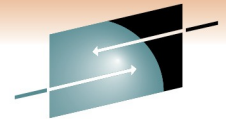
Optional

Description
Shared storage STONITH device
.
sbd uses a shared storage device as a medium to communicate

ID: stonith_sbd-instance_attributes-sbd_device
Name: sbd_device
Value: /dev/disk/by-id/scsi-1IBM_2105_71526069-part1

+ Add Edit Remove

Cancel Reset OK



S H A R E

Technology • Connections • Results

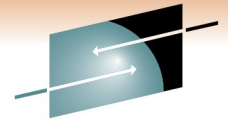
Demoing the details:

cLVM and OCFS2

**NOTE: Its hard to show a demo in slides so come to the session if you want to see the live demo!
Several screenshots are provided to help visualize the demoed topic.**

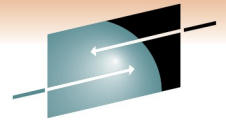
SHARE
in Anaheim
2011





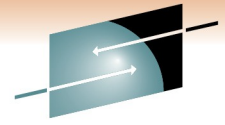
Understanding the definitions of cLVM and OCFS2 in the HA cluster

- cLVM
 - Cluster-aware logical volume manager uses the same LVM management tools to manage PVs, VGs and LVs
- OCFS2
 - Oracle Clustered File System v2
- dlm
 - Distributed Lock Manager manages locking within the cluster
- o2cb
 - OCFS2 cluster software stack
- Cloned resource
 - a resource or resource group that runs on all nodes in the cluster



Understanding the configuration of cLVM and OCFS2 in the HA cluster

- Four resource primitives in a cloned resource group (primitive names are arbitrary)
 - dlm
 - o2cb
 - clvm
 - ocfs2-clusterlv
- Resource primitive start order is important
- The last resource primitive mounts the clustered filesystem on all nodes in the cluster



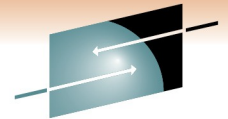
The cLVM and OCFS2 configuration

The screenshot shows the Pacemaker GUI interface. The left sidebar contains a tree view with categories: Configuration, Nodes, Resources, Constraints, and Management (selected). The main area displays a table of resources and their status.

Name	Status	Details
stonith_sbd	running on [s390vmi03]	stonith::external/sbd
nfs_res	running on [s390vm36]	ocf::heartbeat:Filesystem
ip-apache_group	group	
htdocs_res	running on [s390vmi03]	ocf::heartbeat:Filesystem
ocfs2_clone	clone	
ocfs2_group:0	group	
dlm:0	running on [s390vm36]	ocf::pacemaker:controld
o2cb:0	running on [s390vm36]	ocf::ocfs2:o2cb
clvm:0	running on [s390vm36]	ocf::lvm2:clvmd
ocfs2-clusterlv:0	running on [s390vm36]	ocf::heartbeat:Filesystem
ocfs2_group:1	group	
dlm:1	running on [s390vmi03]	ocf::pacemaker:controld
o2cb:1	running on [s390vmi03]	ocf::ocfs2:o2cb
clvm:1	running on [s390vmi03]	ocf::lvm2:clvmd
ocfs2-clusterlv:1	running on [s390vmi03]	ocf::heartbeat:Filesystem

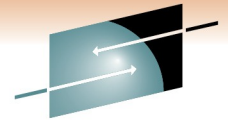
Validate With: pacemaker-1.2
Epoch: 388
Num Updates: 5
CRM Feature Set: 3.0.2
Have Quorum: 1
DC UUID: s390vm36

Connected to 127.0.0.1 (Simple Mode)



The mounted OCFS2 filesystem

```
File Edit View Terminal Help
s390vmi03:~ # mount
/dev/dasda2 on / type ext3 (rw,acl,user_xattr)
proc on /proc type proc (rw)
sysfs on /sys type sysfs (rw)
debugfs on /sys/kernel/debug type debugfs (rw)
devtmpfs on /dev type devtmpfs (rw,mode=0755)
tmpfs on /dev/shm type tmpfs (rw,mode=1777)
devpts on /dev/pts type devpts (rw,mode=0620,gid=5)
fusectl on /sys/fs/fuse/connections type fusectl (rw)
securityfs on /sys/kernel/security type securityfs (rw)
gvfs-fuse-daemon on /root/.gvfs type fuse.gvfs-fuse-daemon (rw,nosuid,nodev)
/dev/sdb2 on /media/disk-10 type ext3 (rw,nosuid,nodev)
/dev/sdb2 on /srv/www/htdocs type ext3 (rw)
none on /sys/kernel/config type configfs (rw)
/dev/mapper/clustervg-clusterlv on /ocfs2mnt type ocfs2 (rw,_netdev,acl,cluster_
stack=pcmk)
s390vmi03:~ #
```



S H A R E

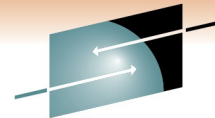
Technology • Connections • Results

Racing the space...

Grow the OCFS2 filesystem while online and being written to by all the clustered nodes!

NOTE: This demo is for attendees only!!





SHARE

Technology • Connections • Results

THANK YOU FOR ATTENDING!!

SHARE
in Anaheim
2011

