

Implementing the SUSE Linux Enterprise High Availability Extension on System z

Mike Friesenegger

SUSE Technical Specialist

mikef@novell.com but soon to be mikef@suse.com



Agenda

- SUSE Linux Enterprise Server for System z
- What is a high availability (HA) cluster?
- What is required to build a HA cluster using SLES?
- Demoing the features of SLE HAE
 - Managing a cluster with the GUI and CLI
 - Resources primitives and resource groups
 - Resource Constraints
 - STONITH
 - cLVM and OCFS2
- Call to Action



SUSE Linux Enterprise Server for System z

SUSE® Linux Enterprise Server for System z 10 years on the Mainframe



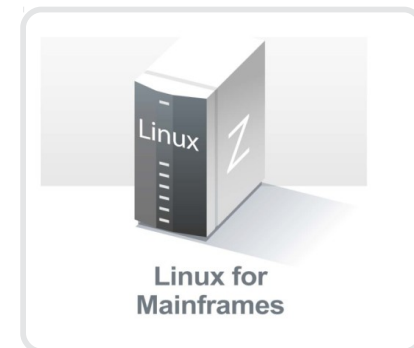
http://www.novell.com/partners/ibm/mainframe/img/timeline_lores.pdf

- The first deployments on Linux for the mainframe were file and print servers.
- The first piece of software that became popular was Samba.
- The first large commercial customer for SUSE Linux Enterprise Server for S/390 was Telia, the largest telecommunications company in Sweden.
- Today, companies are running their mission-critical workloads on top of SUSE Linux Enterprise Server for System z.

Why Customers Prefer SUSE® Linux Enterprise Server for System z

The optimized version for IBM System z:

- **SUSE Linux Enterprise Server is #1**
 - in mainframe Linux market (80%+ share)
 - in SAP-on-Linux market (75% share)
 - in High Performance Computing (6 of top 10)
- **SUSE Linux Enterprise Server for System z:**
 - Fully supported by IBM – supports all benefits of the mainframe
 - 10 years of expertise (available five years ahead of competition)
 - Ideal for workload consolidation, providing major cost savings
 - New features specific to System z
 - Hosting of Subscription Management Tool on System z
 - More than 1,700 certified applications available

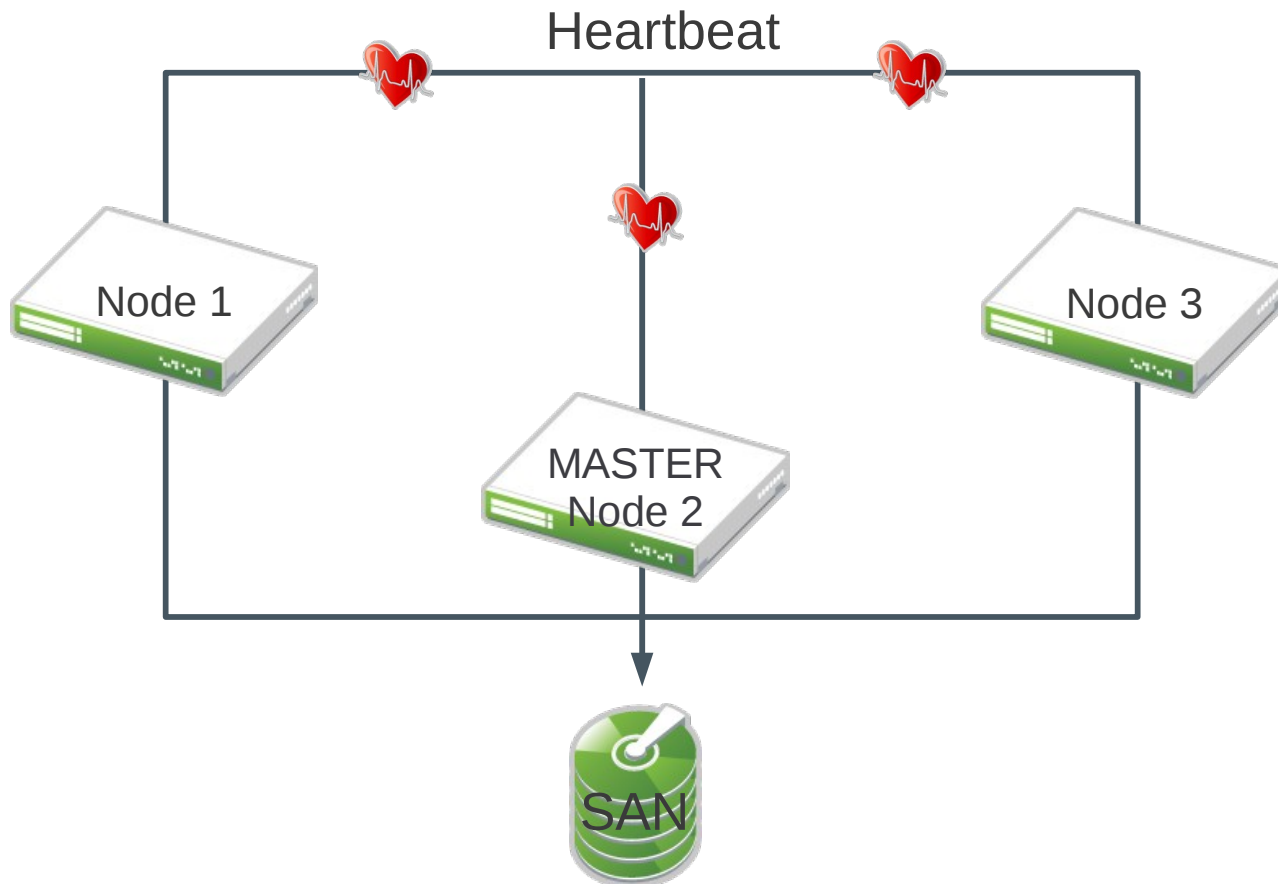


Differentiators - Unique tools for SUSE Linux Enterprise Server for System z

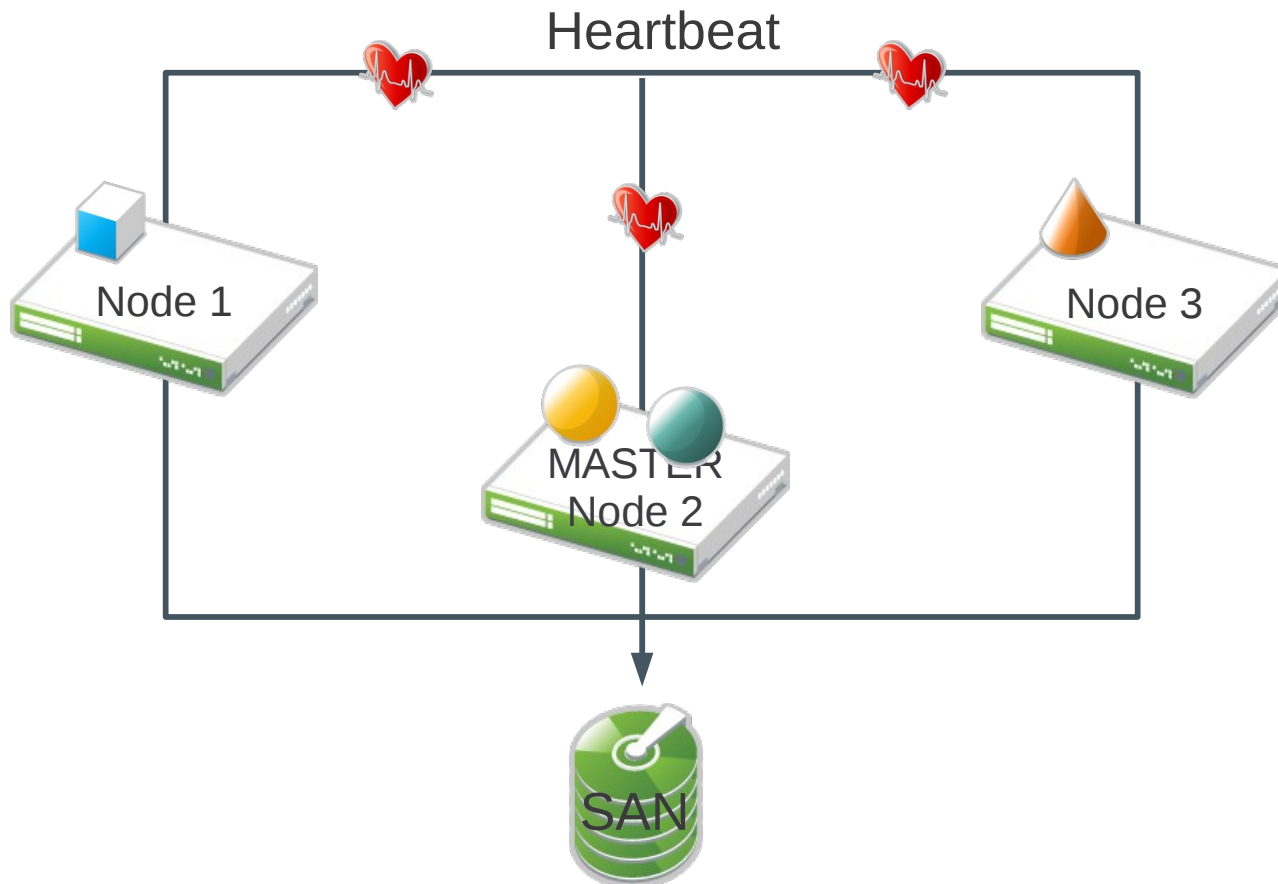
- **Yast and Integrated Systems Management**
Configure every aspect of the server
- **Subscription Management Tool Hosting**
Subscription and patch management made easy
- **High Availability Extension for SLES**
Included in SLES for System z
- **.NET Applications on Linux: Mono**
Migrate existing .NET applications to Linux without having to rewrite code
- **Starter System for System z**
Starter System for System z is a pre-built installation server

What is a high availability (HA) cluster?

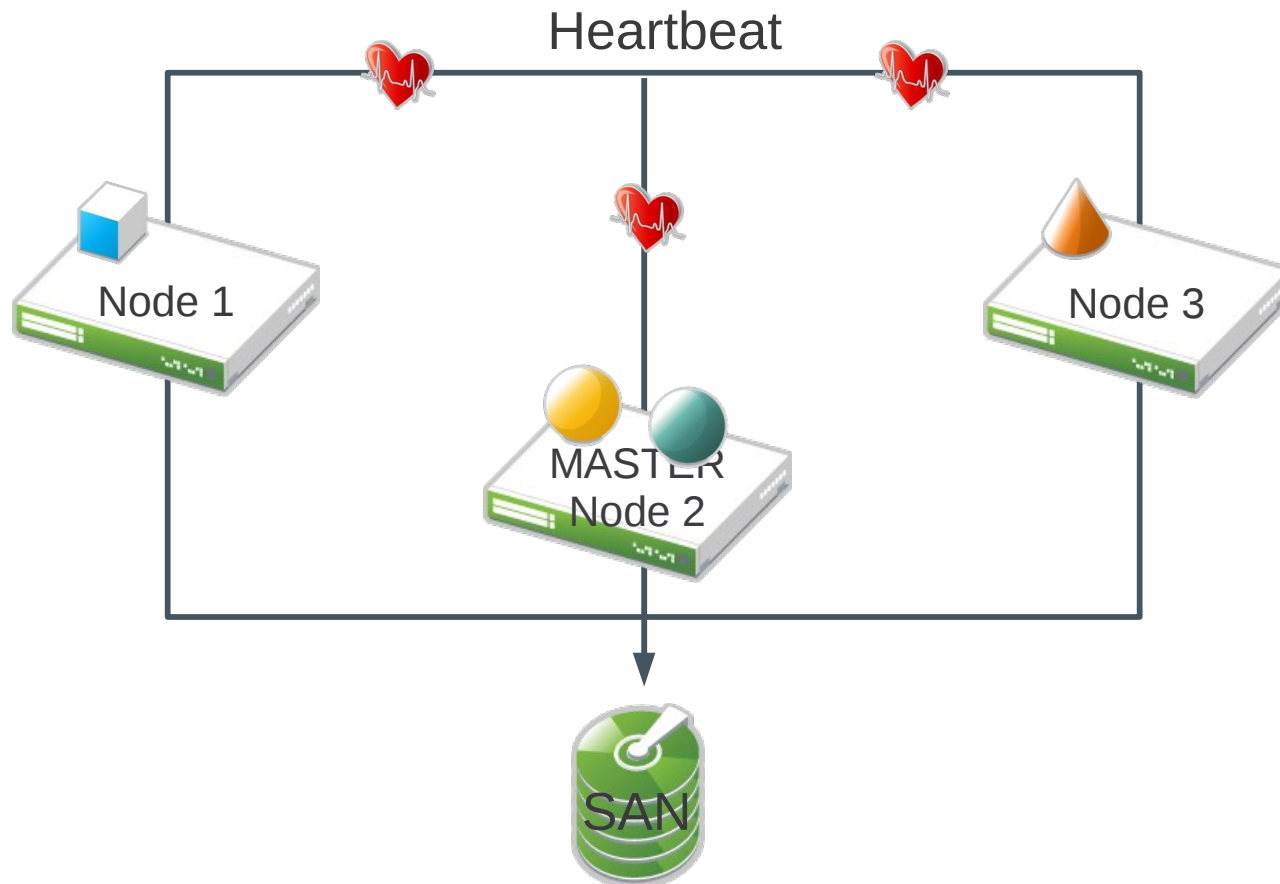
A Simple HA Cluster



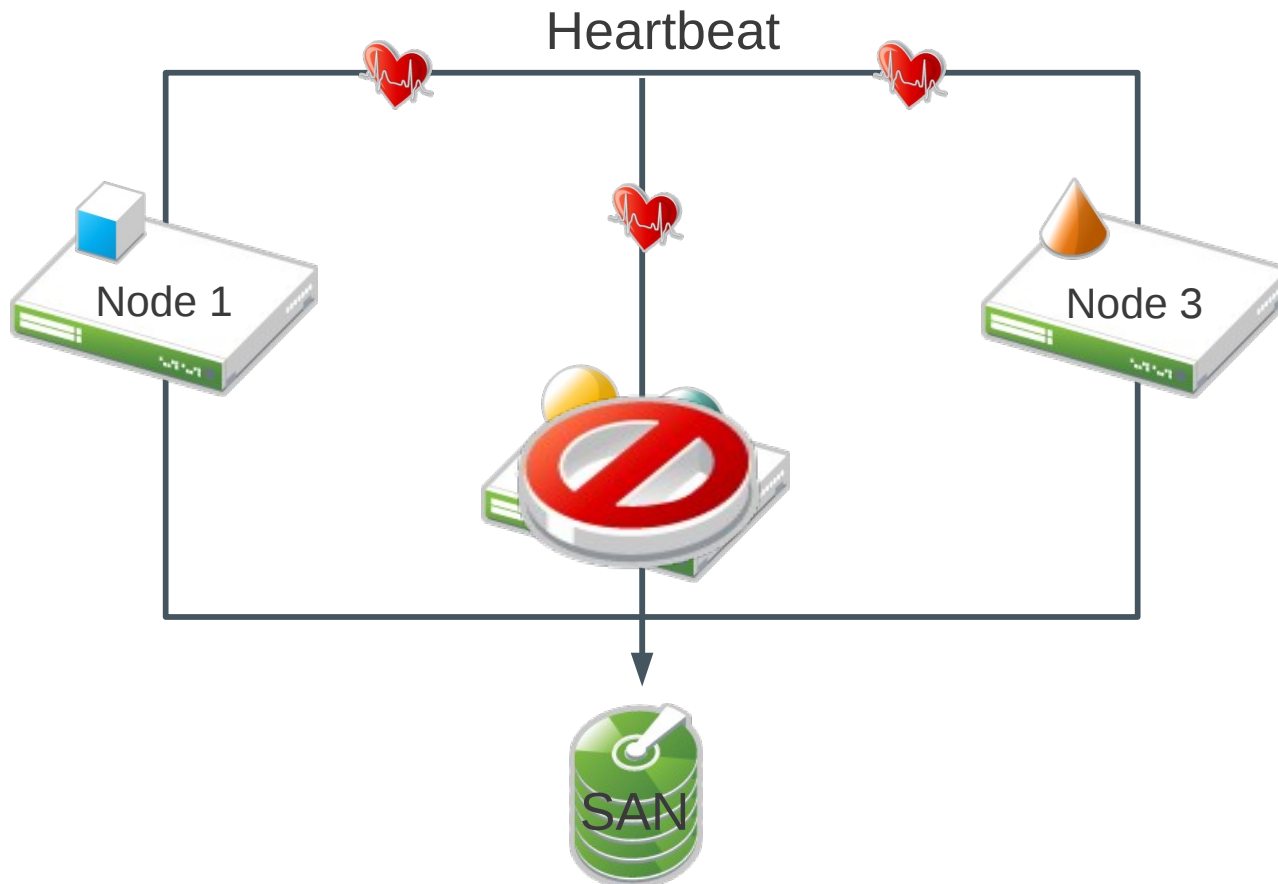
Resources Running in the Cluster



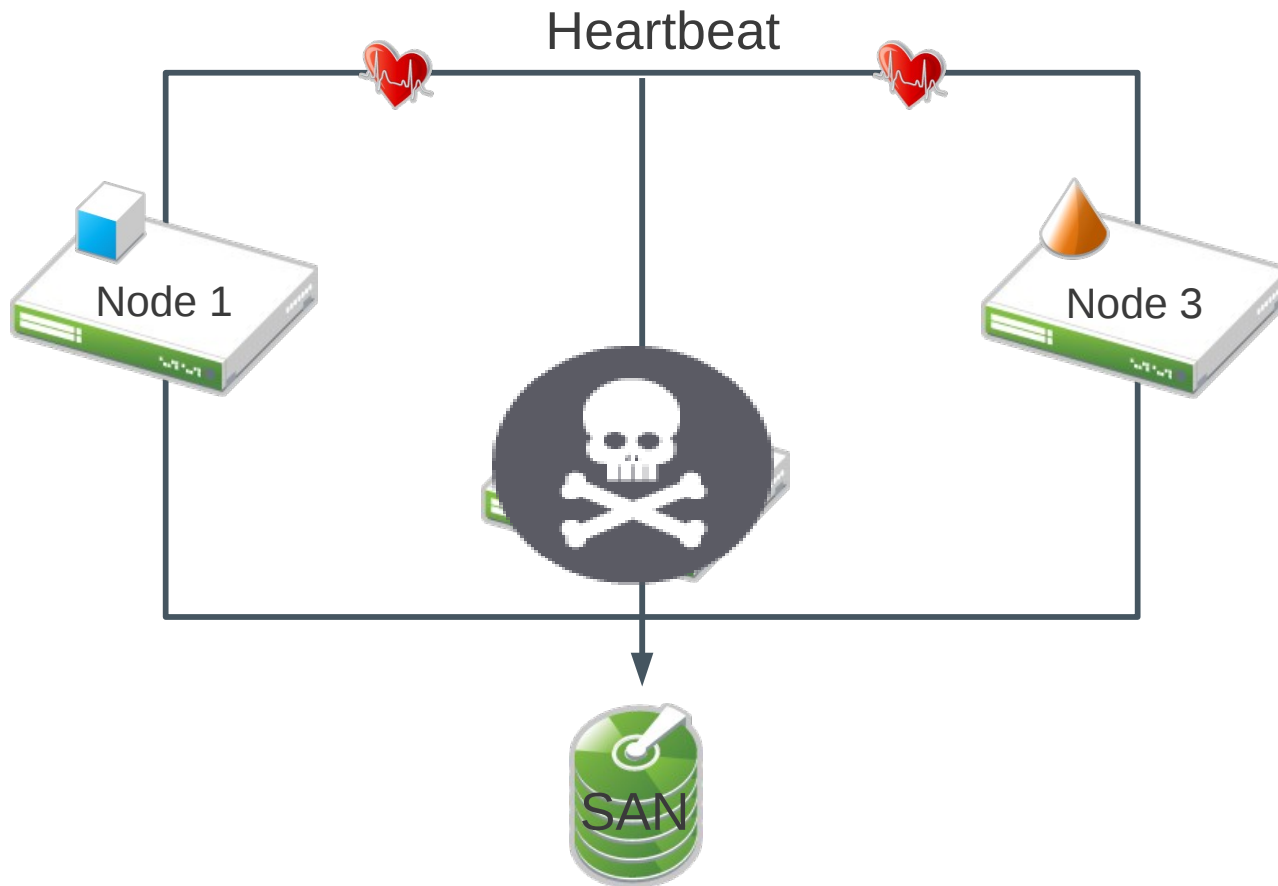
Migrating a Resource



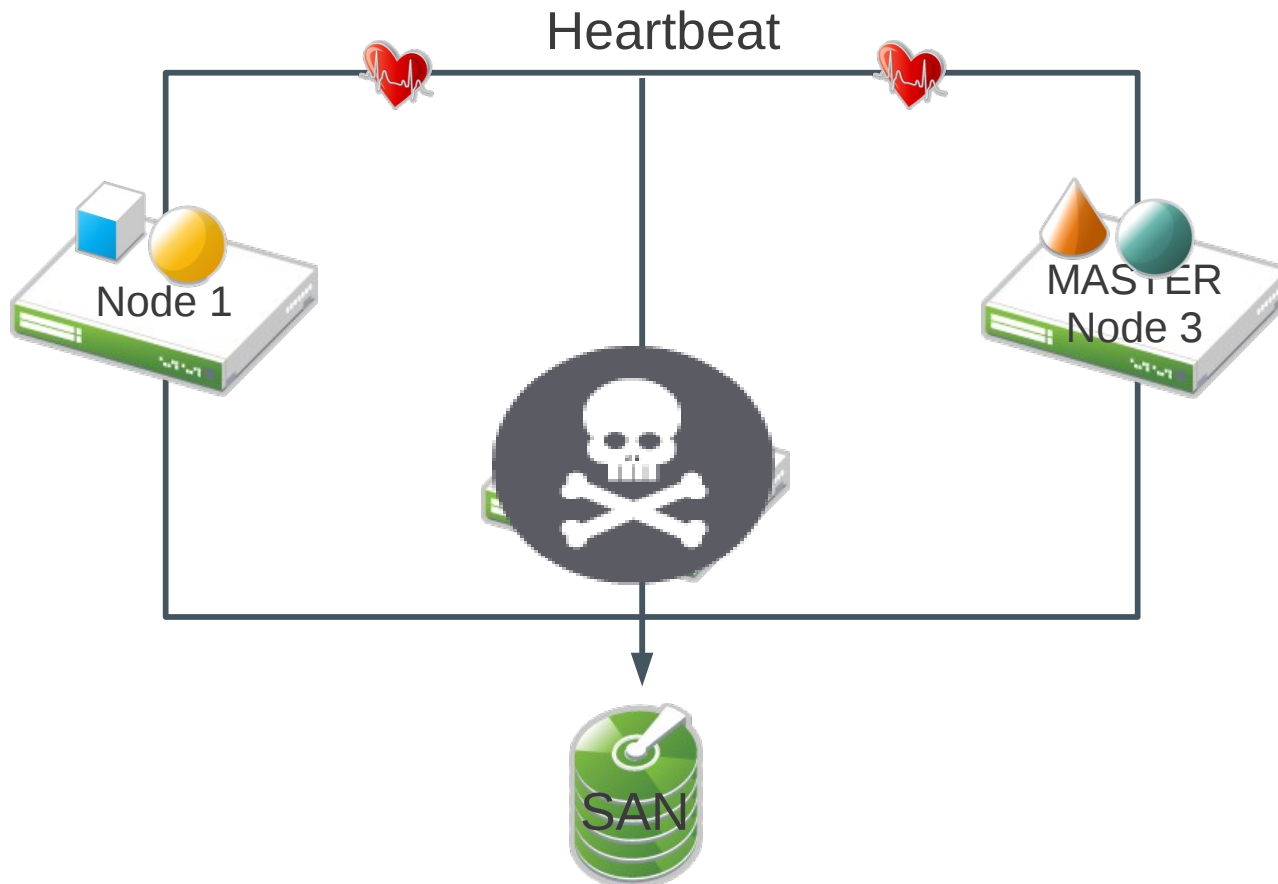
Node Failure in the Cluster



STONITH the Failed Node Out of the Cluster



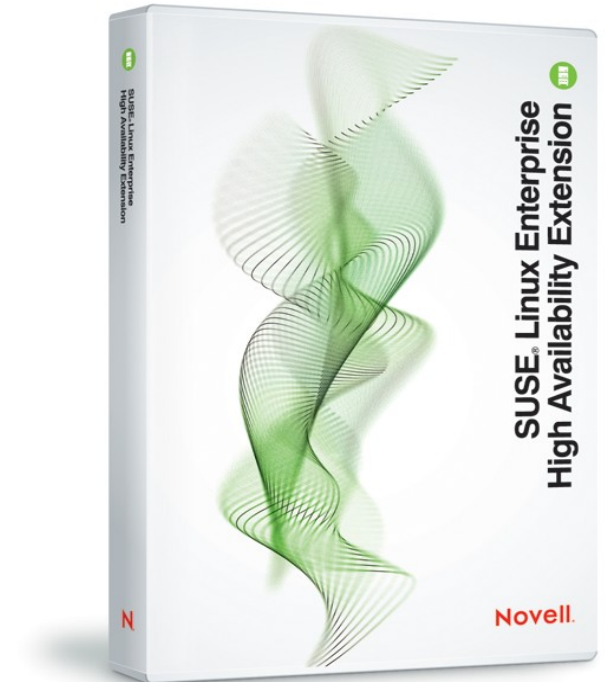
Resources brought up on other nodes the Cluster



What is required to build a
HA cluster using SLES?

SUSE® Linux Enterprise High Availability Extension

- An affordable, integrated suite of robust open source clustering technologies
- Used with SUSE Linux Enterprise Server
- Implement a cluster using physical or virtual Linux servers
- Benefits
 - Cost effectively meet your service-level agreements
 - Ensure continuous access to your mission-critical systems and data
 - Maintain data integrity
 - Increase resource utilization



SUSE® Linux Enterprise High Availability Extension 11 SP1

Key Capabilities Today

- All components use open source technologies
- Flexible, policy-driven clustering solution
 - Metro area cluster up to 20 miles
 - Clustered SAMBA (CIFS)
 - Includes 82 Resource Agents like Oracle, DB2, SAP, Apache, MySQL, PostgreSQL
- Cluster aware file system and volume manager
 - OCFS2 shared-disk POSIX-compliant generic purpose cluster file system
 - Clustering extensions to the standard LVM2 toolset
- Host Based Continuous Data Replication
- Disaster Recovery Integration
- User Friendly Tools
 - No matter if you prefer the CLI or GUI



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



SUSE® Linux Enterprise High Availability Extension

Key Capabilities Future

- **Usability and Management**
 - Full web GUI
 - Improved access control
- **Ease of Use**
 - Guided and automated configuration
 - Prepackages applications
 - Preloaded clusters
- **Clusters Functionality**
 - Wide area clusters
 - Improved data replication
 - Unix cluster stack leadership
- **Backup and Disaster Recovery**
 - Backup integration
 - DR automation



Demoing the features:

Managing a cluster with the GUI and CLI

NOTE: Screenshots are provided to help visualize the demoed features during the session.

Start the GUI with crm_gui

The screenshot displays 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 shows 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 configuration details are shown:

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)".



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

```

```
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)#
```

```
File Edit View Terminal Help
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:/ #
```



Demoing the features:

Resource primitives and resource groups

NOTE: Screenshots are provided to help visualize the demoed features during the session.

A resource primitive

The screenshot displays the Pacemaker GUI interface. The main window shows a list of resource primitives under the 'Primitive' tab. The 'nfs_res' primitive is selected and highlighted in blue. Below the main window, an 'Edit Primitive' dialog box is open, showing the configuration for the 'nfs_res' primitive. The dialog is divided into several sections: 'Required', 'Optional', 'Description', and 'Instance Attributes'. The 'Required' section shows the ID as 'nfs_res', Class as 'ocf', Provider as 'heartbeat', and Type as 'Filesystem'. The 'Description' section contains the text: 'Manages filesystem mounts. Resource script for Filesystem. It manages a Filesystem on a shared storage medium'. The 'Instance Attributes' section is active, showing a table with the following data:

Name	Value
device	10.10.0.100:/dist
directory	/nfsmnt
fstype	

At the bottom of the dialog, there are buttons for '+ Add', 'Edit', '- Remove', 'Cancel', 'Reset', and 'OK'. The status bar at the bottom left of the main window indicates 'Connected to 127.0.0.1'.

A resource group

The screenshot displays the Pacemaker GUI interface. On the left, a sidebar shows a tree view with categories: Live, Configuration (CRM Config, Resource Defaults, Operation Defaults, Nodes, Resources), Constraints, and Management. The 'Resources' category is selected. The main window shows a 'Group' tab with a table listing resource groups. The 'ip-apache_group' is selected and highlighted. An 'Edit Group' dialog box is open, showing the configuration for this group. The dialog has a 'Required' section with ID 'ip-apache_group' and an 'Optional' section. Below the 'Optional' section, there are two tabs: 'Meta Attributes' and 'Primitive'. The 'Primitive' tab is active, showing a table of primitives:

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

Below the table, the details for the selected primitive 'ip_res' are shown:

ID: ip_res
Class: ocf
Provider: heartbeat
Type: IPAddr2

At the bottom of the dialog, there are buttons for '+ Add', 'Edit', 'Remove', 'Cancel', 'Reset', and 'OK'. The main window also has buttons for 'Up' and 'Down' on the right side.

Demoing the features: Resource constraints

NOTE: Screenshots are provided to help visualize the demoed features during the session.

A resource location 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 a menu bar (Connection, View, Shadow, Tools, Help) and a toolbar. Below the menu, there are tabs for 'Resource Location', 'Resource Colocation', and 'Resource Order'. The 'Resource Location' tab is active, showing a table with columns 'ID', 'Resource', 'Score', and 'Node'. A single entry is listed: 'nfs_res-loc', 'nfs_res', 'INFINITY', and 's390vm36'. An 'Edit Resource Location' dialog box is open in the foreground, mirroring the table data. The dialog has a 'Show:' dropdown set to 'List Mode' and fields for 'ID', 'Resource', 'Score', and 'Node'. At the bottom of the dialog are buttons for '+ Add', 'Edit', '- Remove', 'Cancel', 'Reset', and 'OK'. Below the dialog, the current configuration is summarized: ID: nfs_res-loc, Resource: nfs_res, Score: INFINITY, Node: s390vm36. At the bottom of the main window, there are buttons for '+ Add', 'Edit', and '- Remove'. The status bar at the bottom left indicates 'Connected to 127.0.0.1 (Simple Mode)'. The SUSE logo is in the bottom right corner.

Pacemaker GUI

Connection View Shadow Tools Help

Live

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

Show: List Mode

Resource Location Resource Colocation Resource Order

ID	Resource	Score	Node
nfs_res-loc	nfs_res	INFINITY	s390vm36

Up

Down

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 collocation constraint

The screenshot displays the Pacemaker GUI interface. On the left, a navigation pane shows the 'Constraints' section selected. The main window is divided into three tabs: 'Resource Location', 'Resource Collocation', and 'Resource Order'. The 'Resource Collocation' tab is active, showing a table with one entry:

ID	Score	Score Attribute	Score
htdocs_ip-apache_colo	INFINITY		

Below the table, a summary of the selected constraint is shown:

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

An 'Edit Resource Collocation' dialog box is open on the right, showing the configuration for the selected constraint. The 'Required' section is expanded, showing the following fields:

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

The 'Optional' section is collapsed. The 'Score' field is set to INFINITY. Below the dialog, a description is provided:

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

At the bottom of the dialog, there are buttons for '+ Add', 'Edit', 'Remove', 'Cancel', 'Reset', and 'OK'.

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 a menu bar (Connection, View, Shadow, Tools, Help) and a toolbar. The 'Resource Order' tab is active, showing 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). 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 four bullet points. At the bottom of the dialog and the main window, there are buttons for '+ Add', 'Edit', '- Remove', 'Cancel', 'Reset', and 'OK'. The status bar at the bottom indicates 'Connected to 127.0.0.1 (Simple Mode)'.

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

Edit Resource Order

Show: List Mode

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

+ Add Edit - Remove

Cancel Reset OK

+ Add Edit - Remove

Connected to 127.0.0.1 (Simple Mode)

Demoing the features:

STONITH

NOTE: Screenshots are provided to help visualize the demoed features during the session.

What is STONITH?

- **Shoot The Other Node In The Head**
- 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
 - Two of interest for System z: SBD and snIPL
- SLE HAE requires a STONITH device by default
 - Recommended practice to have one configured!



A Split Brain Detector (SBD) STONITH resource

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 to its configuration.

Pacemaker GUI Main Window:

- Menu: Connection, View, Shadow, Tools, Help
- Left Panel: Live, Configuration, CRM Config, Resource Defaults, Operation Defaults, Nodes, Resources (selected), Constraints
- Table:

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

Edit Primitive Dialog:

- Required:
 - ID: stonith_sbd
 - Class: stonith
 - Provider: (empty)
 - Type: external/sbd
- Optional:
 - Description: Shared storage STONITH device
 - sbddescription: sbd uses a shared storage device as a medium to communicate
- Instance Attributes:
 - Table:

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

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

Demoing the features:

cLVM and OCFS2

NOTE: Screenshots are provided to help visualize the demoed features during the session.

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 the following items: Configuration, CRM Config, Resource Defaults, Operation Defaults, Nodes, Resources, Constraints, and Management (highlighted). The main window displays a table of resources with columns for Name, Status, and Details. A red rounded rectangle highlights the OCFS2 resources.

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:~ #
```



Attend the SUSE Linux High Availability Extensions Hands-on Workshop (9348 and 9494).

Monday 3 – 6pm.

Thank you.





Corporate Headquarters
Maxfeldstrasse 5
90409 Nuremberg
Germany

+49 911 740 53 0 (Worldwide)
[+www.suse.com](http://www.suse.com)

Join us on:
www.opensuse.org