



SUSE® Linux Enterprise 11

High Availability Extension

Best Practice and Project Examples

Lars Pinne
SUSE Linux GmbH



Agenda

- Overview SUSE Linux Enterprise 11 High Availability
- Data Replication Options
- Best Practices
- Project Examples
- Miscellaneous
- Appendix

SUSE Linux Enterprise 11 High Availability

SUSE Linux Enterprise 11

High Availability

- The extension addresses customer needs to run mission critical workloads with High Availability, Mission Critical applications, Cluster File System, Load-Balancing
- SUSE Linux Enterprise stack has matured over years
- Local Cluster and Metro Cluster (max. ~30km*)
- Supplied as an add-on product (own iso image)
- Integration with SLES (i586, ia64, x86_64, ppc, s390x)
- Available on SLES product release

* see release notes



SUSE® Linux Enterprise 11

HA Components

Linux HA stack, incl. pacemaker, corosync, hawk, and YaST module

- resource-agents – monitor availability of resources
- stonith – fencing support (also Xen, KVM, VMware VMs)
- pacemaker – cluster resource manager
- Corosync - cluster communication layer
- hawk – web konsole for cluster resource and dependencies editing
- crm – command line to interact with the CIB: editing, prepare multiple changes & commit once, syntax validation, etc.
- YaST module – easy basic installation

What Is Pacemaker?

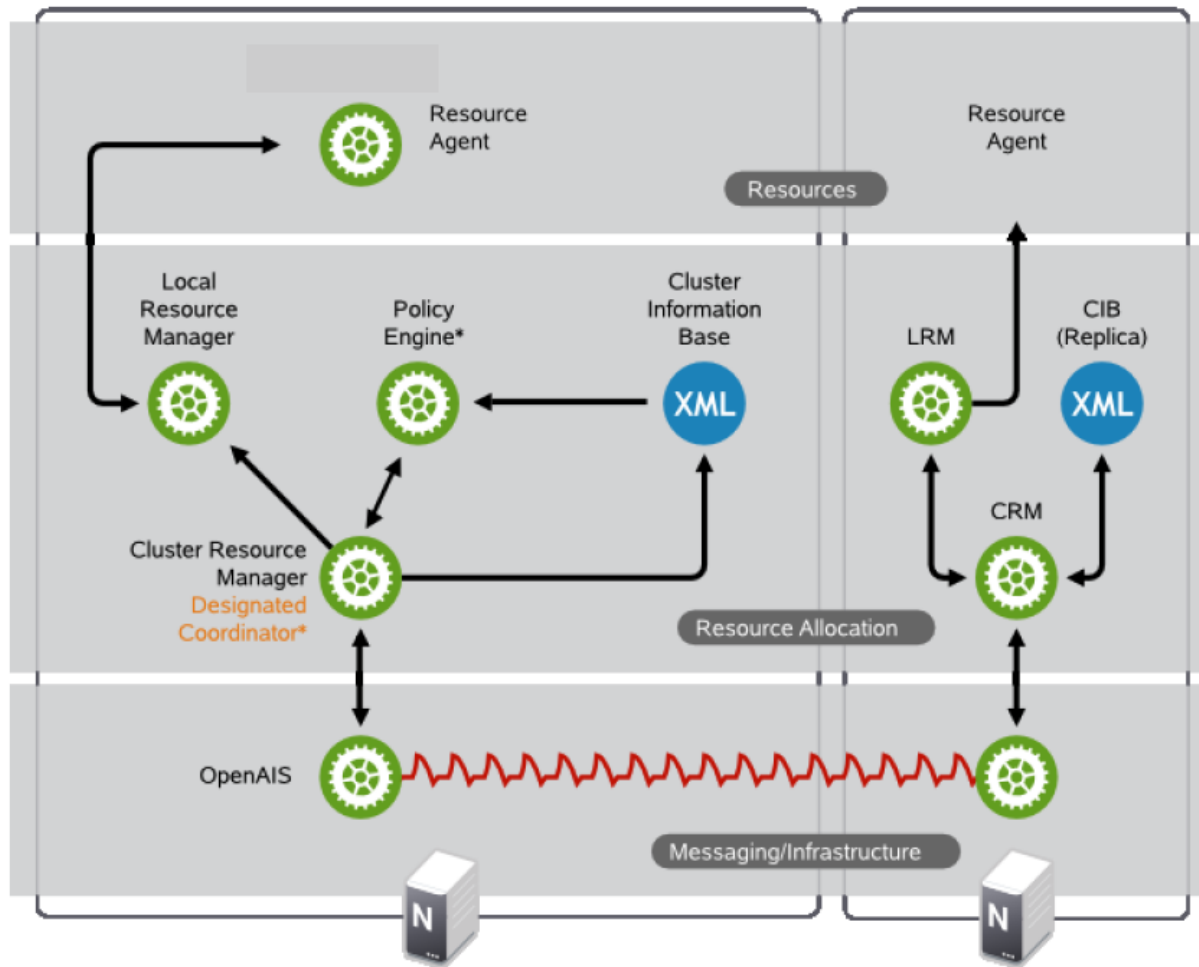


The brain

Non-cluster-aware
components

Cluster communication

SUSE® Linux Enterprise 11 HA Architecture




- Cluster resources are probed on all nodes during Cluster startup, node joining or resource reconfiguration.
- Cluster information base is replicated on all nodes.
- Cluster coordinator is active on one node.
- Cluster infrastructure is active on all nodes.

























































SUSE® Linux Enterprise 11 HA Resource Agents

- Compliant to Open Cluster Framework (OCF).
- **Start** - Service must be functional after start returns.
- **Stop** - After this operation has completed, service must be fully stopped!
- **Monitor** - Return the status of the service:
 - 0 (\$OCF_SUCCESS): Service running fine
 - 7 (\$OCF_NOT_RUNNING): Service NOT running at all
 - 1 (\$OCF_ERR_GENERIC): Considered active but failed
- Provide meta-data about service for configuration.
- Live in /usr/lib/ocf/resource.d/<provider>/ .
 - f.e. “oracle”, “oralsnr” for Oracle databases
 - f.e. “exportfs” for NFS servers
 - f.e. “Filesystem” for filesystems like OCFS2 or Ext3
- STONITH RAs are special, because crucial.

SUSE Linux Enterprise 11 HA

HA Web Konsole

 Cluster Status

Cluster Status			Inactive Resources
hex-0: Online 	hex-7: Online 	hex-9: Online 	
Clone Set: base-clone			
dln:0: Started 	dln:1: Started 	dln:2: Started 	
o2cb:0: Started 	o2cb:1: Started 	o2cb:2: Started 	
clvm:0: Started 	clvm:1: Started 	clvm:2: Started 	
cmirrord:0: Started 	cmirrord:1: Started 	cmirrord:2: Started 	
vg1:0: Started 	vg1:1: Started 	vg1:2: Started 	
ocfs2-1:0: Started 	ocfs2-1:1: Started 	ocfs2-1:2: Started 	
fencing-sbd: Started 	vm-07: Pending 	vm-23: Started 	vm-00: Stopped 
vm-08: Started 	vm-09: Pending 	vm-24: Started 	dummy1: Stopped 
vm-10: Started 	vm-11: Pending 	vm-25: Started 	Delay1: Stopped 
vm-12: Started 	vm-13: Pending 	vm-26: Started 	vm-01: Stopped 
vm-14: Started 	vm-15: Pending 	vm-27: Started 	vm-02: Stopped 
vm-16: Started 	vm-17: Pending 	vm-28: Started 	vm-03: Stopped 
vm-18: Started 	vm-19: Pending 	vm-29: Started 	vm-04: Stopped 
vm-20: Started 		vm-30: Started 	vm-05: Stopped 
vm-22: Started 			vm-06: Stopped 
			vm-31: Stopped 
			vm-32: Stopped 
			vm-33: Stopped 

Start

Stop

Move...

Drop Relocation Rule

Clean Up

View Recent Events...

Fencing options

- Server fencing with STONITH +
Storage fencing with SFEX
 - needs LAN access to remote management board (or virtualisation host)
 - critical resources
 - proven
- Server fencing with SBD + watchdog
 - needs separate disk, ideally on 3rd site
 - very critical resource
 - One, two, or three SBD devices
- Storage fencing with SCSI3 persistent reservation
 - needs procedure to unlock storage if node dies
 - needs device-specific cluster integration
 - implemented in projects, currently no roadmap
- Storage fencing with storage based mirror
 - additional hardware features needed
 - needs device-specific cluster integration

Fencing - SFEX

SFEX prevents the cluster from corrupting the data.

SFEX should be placed close to the data it should protect.

A single SFEX is a single point of failure - data integrity is always on cost of service availability.

To reduce the impact:

- Configure the SAN boxes identically on both sites, having the exact same LUNs and partitions.
- Initialise the SFEX in exact the same way on both sites.
- Only use one SFEX for the cluster. If you have one preferred site, use the SFEX there.

This gives:

- In LAN split scenario, the SFEX device is not affected, the service could still run on either site.
- In SAN split scenario, the site where the SFEX device is used, will win. If the service was running there, nothing happens. If the service was on the other site, it will be migrated.
- In complete split brain scenario, the site where the SFEX device is used, will win.
- In disaster scenario, either the site with the SFEX device will stay alive. or the other. If the site without SFEX device has survived, you simply have to replace the SFEX device in the RA configuration and to start the service.
- If the SFEX device fails for whatever reason, you have to replace the SFEX device in the RA configuration and to restart the service.

Data integrity is given in **all** described **scenarios**.



Fencing - SBD

- goal: split-brain protection
- uses shared SAN storage device as medium for sending fencing requests (poison pills)
- all nodes must be able to regularly check the SBD device for fencing requests
- if a node loses I/O connection to the SBD device, it must go down immediately
- watchdog resets server when timer expires
- SBD daemon (user space, RT prio) regularly resets watchdog timer if SBD device is accessible
- 1, 2, or 3 SBD device.
With 2 or 3 SBD devices: If one device goes down, nodes stay.

SUSE Linux Enterprise High Availability - Overview

SLE-HA in VMware virtual machines

- . Main benefit is application monitoring
 - see <http://www.emc.com/solutions/application-environment/sap/virtualization.htm>
- . SLE-HA is supported on VMware
 - for SAP and other workloads
 - application support depends on ISVs
 - the cluster acts as on physical machines
 - VMware placing policy to place cluster nodes on different physical hosts
- . SUSE Best Practices can be implemented
 - cluster timings have to be aligned with virtualisation infrastructure
 - network bandwidth has to be sufficient for vMotion to avoid timeouts
- . Shared Storage is used
 - for SBD fencing and Cluster Resources
 - no need for Shared SCSI Bus
 - if VCB backup is not available, ReaR might be an option
 - Shared Storage could be VMFS file or RDM, even with vMotion *

* see <http://kb.vmware.com/kb/1034165>

Note: If no shared storage or iSCSI is used at all, fencing could be done via vSphere API, see links in appendix.



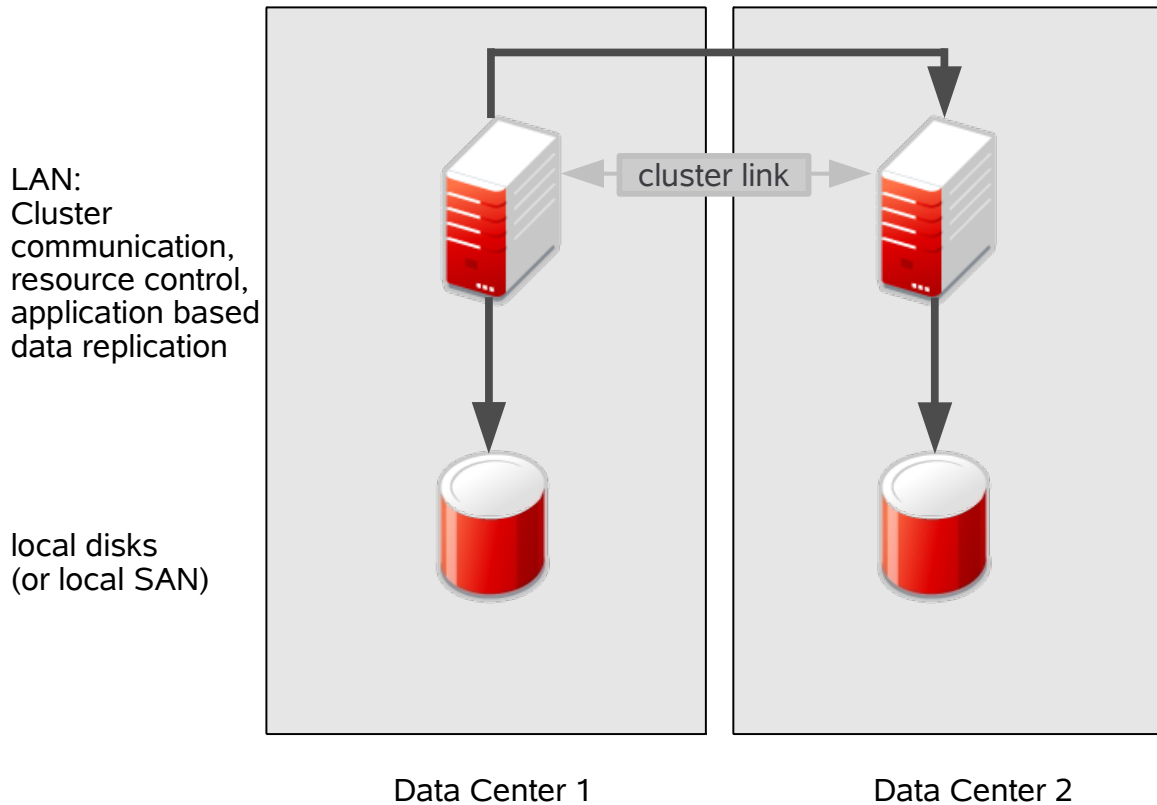
Data Replication Options

Data Replication Options

- Application based mirror
 - f.e. Oracle data guard, DB2 HADR
 - on top of operating system SLES
- Network based mirror
 - Distributed Replicated Block Device (DRBD)
 - part of SLE11-HA
- Host based mirror
 - active/active: cLVM, cmirrord
 - active/passive: MD-RAID
 - part of SLE11-HA
- Storage based mirror
 - f.e. EMC SRDF, HP CLX EVA, NetApp Metro Cluster
 - beneath operating system SLES

Data replication

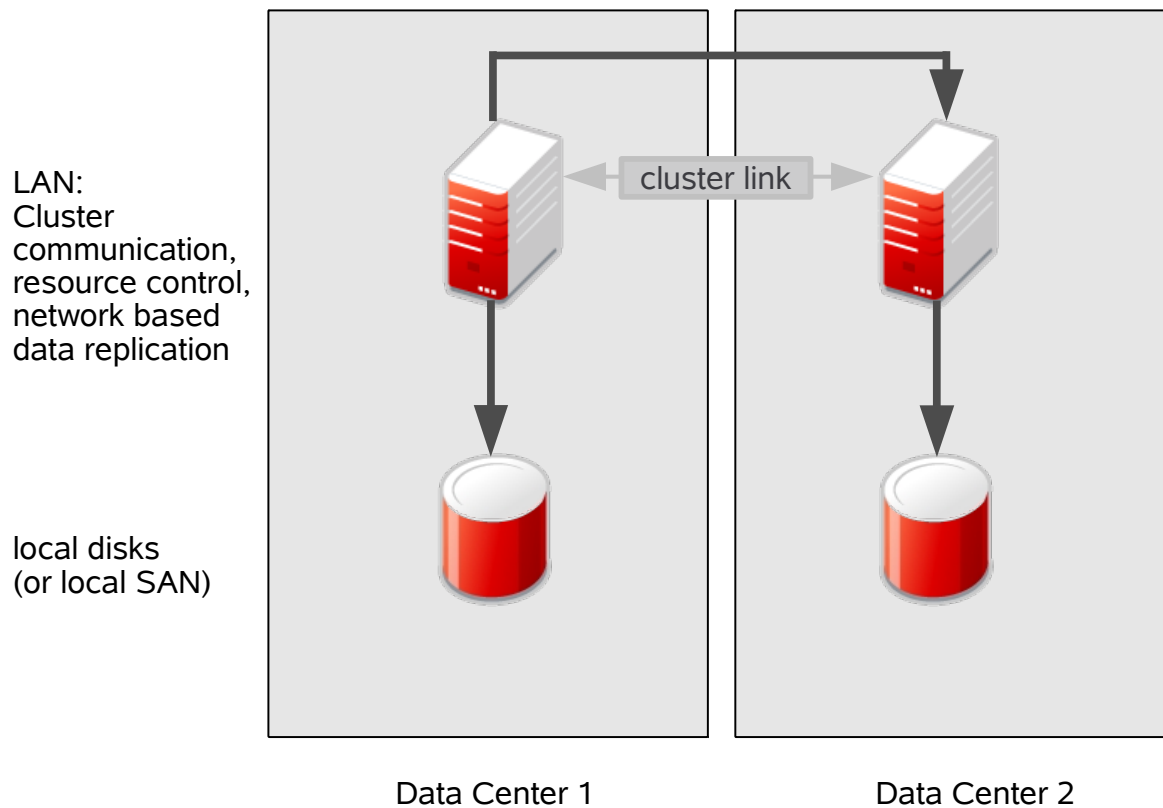
Application based mirror



- + long distance possible
- + no SAN needed
- + application aware
- LAN bandwidth needed
- specific application only

Data replication

Network based mirror



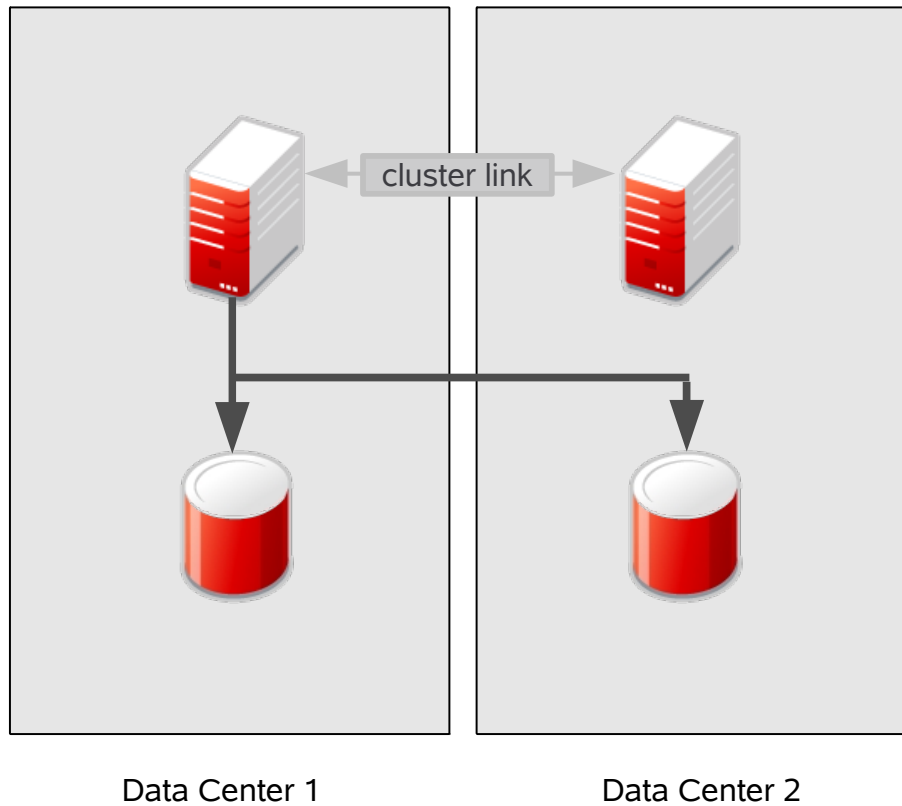
- + long distance possible
- + no SAN needed
- + part of SLE11 HA
- + vendor independent
- LAN bandwidth needed

Data replication

Host based mirror

LAN:
cluster
communication,
resource control

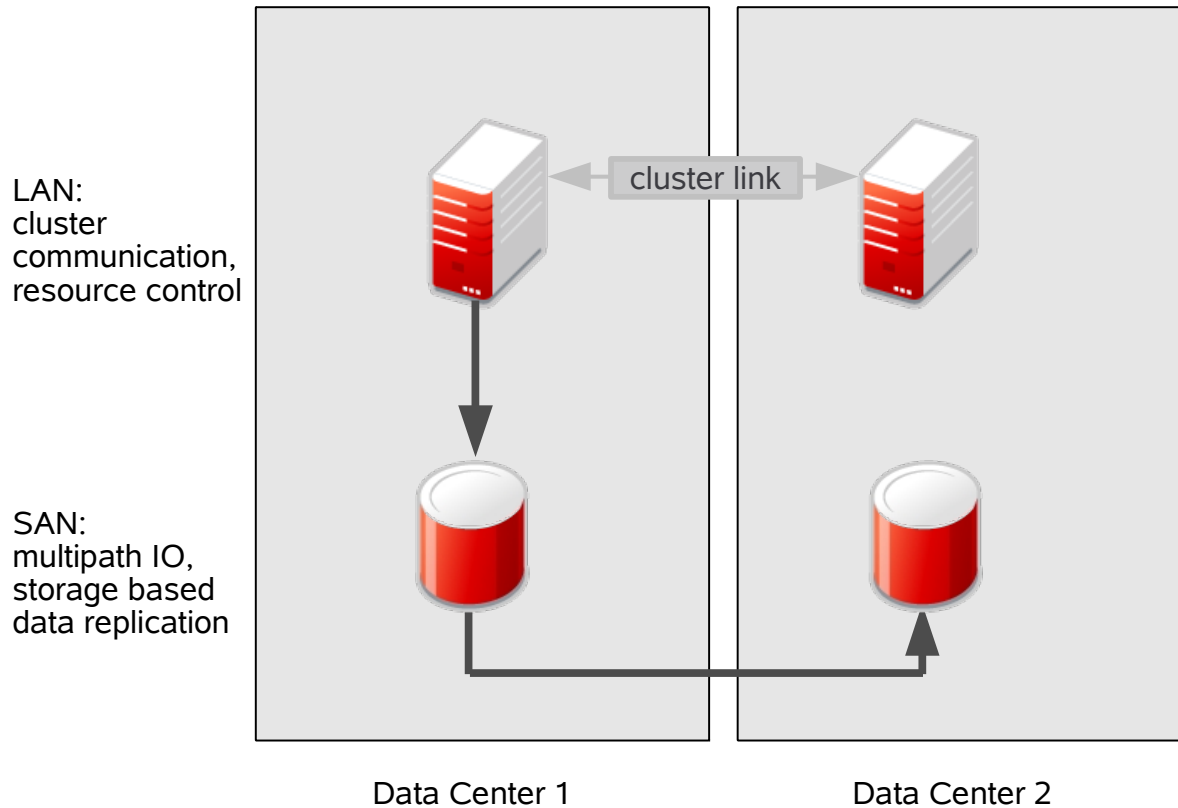
SAN:
multipath IO,
host based
data replication



- + proven solution
- + part of SLE11 HA
- + vendor independent
- limited distance

Data replication

Storage based mirror



- + long distance support by hardware vendors
- + reduced load on local SAN
- + simplified Linux IO stack
- vendor / hardware specific

Best Practices

The DOs

It's good to do this:

- + host based mirror with MD-Raid for high availability under cluster control
- + storage based mirror for disaster recovery, manual switch
- + MD-Raid to protect shared LUNs against forbidden access
- + split large filesystems
- + SFEX to protect filesystem on shared LUNs, if LAN-based stonith is used
- + two independent LAN links for cluster communication
- + check blade servers for independent LAN links
- + make CIB simple, f.e. few groups instead of many constraints
- + resource naming schema, f.e. prefixes rsc_, msl_, grp_, ord_, loc_, col_
- + set up cluster step-by-step
- + use crm
- + always issue crm unmigrate after migration has completed
- + prepare custom scripts for admin tasks
- + define and perform tests for all failure scenarios



The DON'Ts

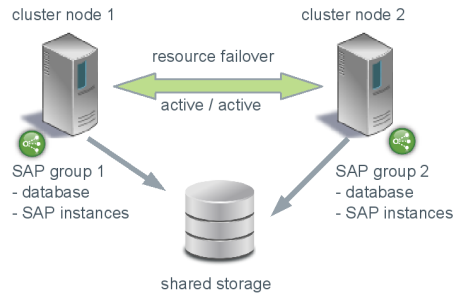
It's good NOT to do this:

- directly re-use concepts from other cluster solutions
- storage based mirror for high availability under cluster control
- storage based mirror with inactive pathes visible
- multipath user friendly names, but not all names pre-defined
- cluster resource, STONITH, and SBD timings shorter than SAN timings
- both corosync rings in same subnet
- OCFS2 if no concurrent access is needed
- OCFS2 on MD-RAID
- host based mirror with cLVM without extended careful testing
- no stonith at all
- other software use the watchdog in parallel to SBD
- go live without tests planned and done

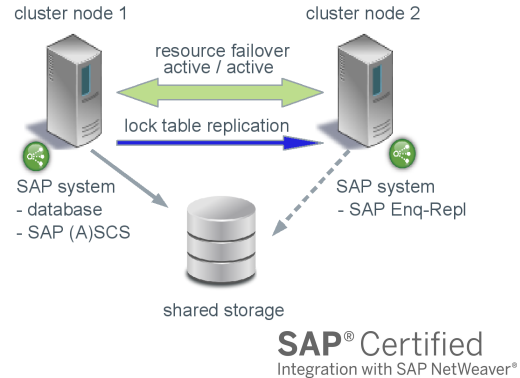


Best Practice Examples

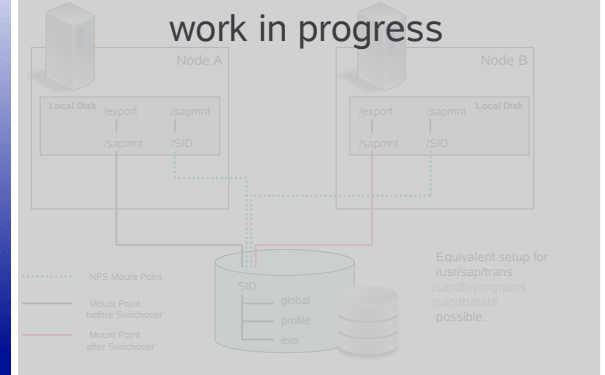
Simple Stack HA



Enqueue Replication



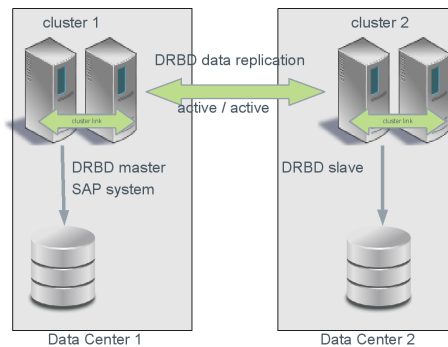
NFS and SAP in one Cluster



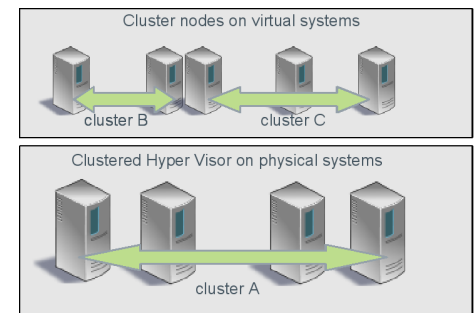
HP CLX SAN Based Mirror



DRBD Data Replication



HA in Virtualized Environments



See also "Protection of business critical applications in VMware vSphere4 virtualized SLES environments demonstrated with SAP NetWeaver"

SLE-HA for SAP

General Available Solutions 1/2

- Simple Stack SAP and Database
 - Netweaver: **ABAP**[°], Java, Double Stack
 - **MaxDB**[°], Oracle, DB/2, Sybase
- Enqueue Replication
 - Netweaver: **ABAP**[°] or Java
 - **SAP Certified**[°]
- Simple Stack Database
 - Oracle*, DB/2

[°] whitepaper published, * training or workshop available



SLE-HA for SAP

General Available Solutions 2/2

- External NFS server
 - SAN shared storage
 - **DRBD**[°]
- Storage-based mirror
 - if completely transparent to OS and cluster
- Cluster in virtual machine
 - **VMware**[°], **KVM**[^], Xen

[^] under evaluation



SLE-HA for SAP

Custom Solutions

- Enqueue Replication - misc.
 - specific configuration
- SAP Web Dispatcher
- SAP HANA System Replication
- Database - misc.
 - Oracle RAC, Dataguard
 - Storage-based mirror that needs cluster integration (SRDF, CLXEVA, ...)
 - Simple Stack MaxDB, Sybase
 - specific configuration
- Three or more nodes in one cluster

SLE-HA for SAP

Currently NO Solutions

- SAP and NFS server in same cluster ^
- Enqueue Replication Double Stack with one single common IP address
 - not recommended by SAP anymore
 - f.e. Solution Manager or PI in “traditional” setup
- Quorum Server
 - alternative option: 3rd SBD on iSCSI

^ under evaluation

Project Examples

Sample: HA Stack for SAP

The screenshot shows the Pacemaker GUI (on node0101) with the following configuration and status:

- Configuration:**
 - CRM Config
 - Resource Defaults
 - Operation Defaults
 - Nodes
 - Resources
 - Constraints
 - ACLs
 - Management
- Cluster Status:**
 - have quorum (green circle)
 - node0102 (green circle)
 - node0101 (green circle)
- Resources:**
 - rsc_stonith_sbd (green circle)
 - grp_sap_NA1 (green circle)
 - rsc_ip_NA1_sapna1ci (green circle)
 - rsc_ip_NA1_sapna1db (green circle)
 - rsc_ip_NA1_sapna1as (green circle)
 - rsc_md_NA1_md0 (green circle)
 - rsc_lvm_NA1_sapvg (green circle)
 - rsc_fs_NA1_usrsap (green circle)
 - rsc_fs_NA1_sapmnt (green circle)
 - rsc_fs_NA1_sapdb (green circle)
 - rsc_sapdb_NA1 (green circle)
 - rsc_sapinst_NA1_ASCS00_sapna1as (green circle)
 - rsc_sapinst_NA1_DVEBMGS01_sapna1ci (green circle)
- Migration Threshold:** 5
- Call ID Table:**

Call ID	Operation	Interval	Return Code	Status	Last Run	Exec Time	Queue Time
35	start		ok (rc=0)	complete	Tue Nov 1 19:53:02 2011	16100ms	0ms
36	monitor	180000ms	ok (rc=0)	complete		360ms	0ms

Connected to hacluster@127.0.0.1 (Simple Mode)

- SLE11-HA Cluster add-on for SLES11
- Resource agents for SAP Instance and SAP Database are part of SLE11-HA
- Supported databases
 - DB/2
 - MaxDB
 - Sybase
 - Oracle
- Host based mirror
 - MD-RAID
 - LVM2
- SBD to better handle split-brain scenarios
 - Stonith + SFEX alternatively

Sample: HA SAP Enqueue Replication

The screenshot displays the SAP Cluster Status web interface. The browser address bar shows the URL `https://ls3198v7.wdf.sap.corp:7630/main/status`. The page title is "Cluster Status". On the left, there is a sidebar with icons for various cluster management functions. The main content area is divided into three columns. The first column, "Online", contains a list of resources: `ci2n01: Online`, `ci2n02: Online`, `Master/Slave Set: msl_sap_enqrepl_HA0`, `rsc_sap_HA0_ASCS00:0: Master`, `stonith-sbd: Started`, `rsc_ip_HA0_sapha0as: Started`, `rsc_ip_HA0_sapha0ci: Started`, `rsc_fs_HA0_dvebmgs01: Started`, `rsc_sap_HA0_DVEBMGS01: Started`, `rsc_ip_HA0_sapha0d2: Started`, `rsc_fs_HA0_d02: Started`, and `rsc_sap_HA0_D02: Started`. The second column, "Inactive Resources", is currently empty.

- SLE11-HA Cluster add-on for SLES11
- Resource agent for SAP Instance is part of SLE11-HA
- Enqueue and Replicated Enqueue as Master/Slave resource
- SLE11-HA certified by SAP

Sample: SAP Two Enq. Replication

The screenshot shows the Pacemaker GUI (on cl3n01) with a tree view of the cluster configuration. The 'Live' section is expanded, showing the following resources and their status:

Name	Status	Details
Cluster	have quorum	Openais & Pacemaker
cl3n01	online (dc)	
cl3n02	online	
Resources		
rsc_stonith_sbd	running on [cl3n01]	stonith::external/sbd
grp_sapdb_so1	group	
rsc_ip_SO1_sapso1db	running on [cl3n01]	ocf::heartbeat:IPAddr2
rsc_md_SO1_md0	running on [cl3n01]	ocf::heartbeat:Raid1
rsc_lvm_SO1_sapso1db	running on [cl3n01]	ocf::heartbeat:LVM
rsc_fs_SO1_sapdb	running on [cl3n01]	ocf::heartbeat:Filesystem
rsc_sapdb_SO1	running on [cl3n01]	ocf::heartbeat:SAPDatabase
grp_sapci_so1	group	
rsc_ip_SO1_sapso1ci	running on [cl3n01]	ocf::heartbeat:IPAddr2
rsc_sapinst_SO1_DVEBMGS20_sapso1ci	running on [cl3n01]	ocf::heartbeat:SAPInstance
msl_sapas_SO1	master	
rsc_sapinst_SO1_ASCS00_sapso1as:0	running (Master) on [cl3n02]	ocf::heartbeat:SAPInstance
rsc_sapinst_SO1_ASCS00_sapso1as:1	running (Slave) on [cl3n01]	ocf::heartbeat:SAPInstance
msl_sapcs_SO1	master	
rsc_sapinst_SO1_SCS01_sapso1cs:0	running (Slave) on [cl3n02]	ocf::heartbeat:SAPInstance
rsc_sapinst_SO1_SCS01_sapso1cs:1	running (Master) on [cl3n01]	ocf::heartbeat:SAPInstance
rsc_ip_SO1_sapso1er	running on [cl3n01]	ocf::heartbeat:IPAddr2
rsc_ip_SO1_sapso1e2	running on [cl3n02]	ocf::heartbeat:IPAddr2
rsc_ip_SO1_sapso1cs	running on [cl3n01]	ocf::heartbeat:IPAddr2
rsc_ip_SO1_sapso1as	running on [cl3n02]	ocf::heartbeat:IPAddr2

Connected to hacluster@127.0.0.1 (Simple Mode)

- SLE11-HA Cluster add-on for SLES11
- Resource agent for SAP Instance is part of SLE11-HA
- Enqueue and Replicated Enqueue as Master/Slave resource
- SCS and ASCS with Replicated Enqueue, independent from each other
- Similar to SAP certified setup

Sample: HA for SAP, customized

The screenshot shows the Pacemaker GUI with the following structure:

- Live** (selected)
 - Configuration
 - CRM Config
 - Resource Defaults
 - Operation Defaults
 - Nodes
 - Resources
 - Constraints
 - ACLs
 - Management** (highlighted)

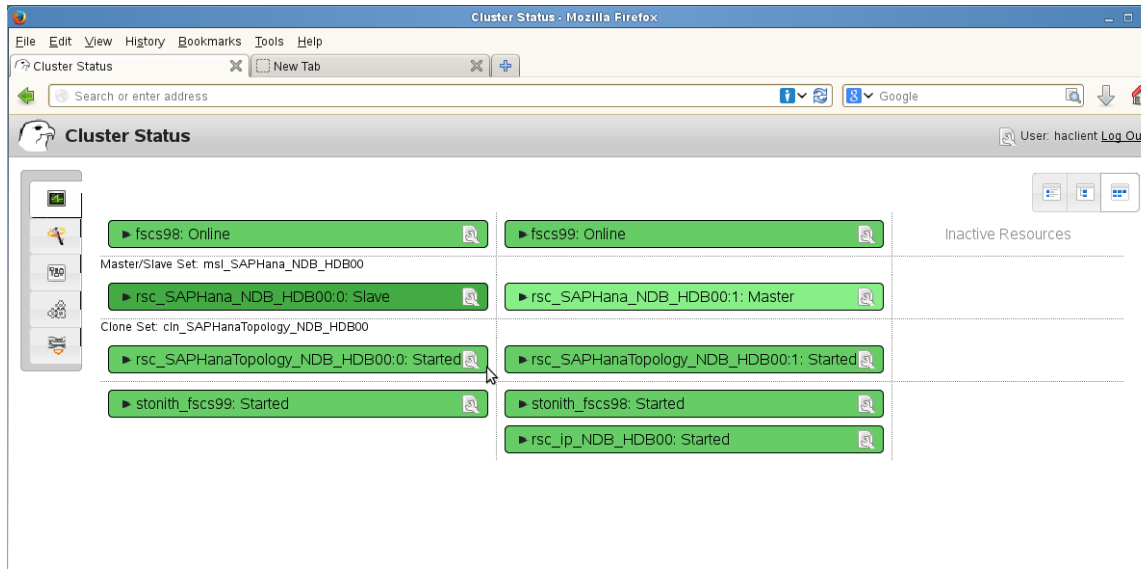
Name	Status	Details
Cluster	● have quorum	Openais & Pacemaker
birne2330	● online	
birne1330	● online (dc)	
Resources	●	
rsc_stonith_sbd	● running on ['birne1330'] stonith::external/sbd	
cln_ethmonitor	● clone	
rsc_ethmonitor:0	● running on ['birne1330'] ocf::cust::ethmonitor	
rsc_ethmonitor:1	● running on ['birne2330'] ocf::cust::ethmonitor	
grp_ciZSA	● group	
rsc_md_ZSA_ZSAci01_md2	● running on ['birne1330'] ocf::heartbeat:Raid1	
rsc_lm_ZSA_ZSAci01	● running on ['birne1330'] ocf::heartbeat:LVM	
rsc_fs_ZSA_DVEBMGS20	● running on ['birne1330'] ocf::heartbeat:Filesystem	
rsc_fs_ZSA_W40	● running on ['birne1330'] ocf::heartbeat:Filesystem	
rsc_fs_ZSA_ASCS25	● running on ['birne1330'] ocf::heartbeat:Filesystem	
rsc_ip_ZSA_zsacineu	● running on ['birne1330'] ocf::heartbeat:IPAddr2	
rsc_app_ZSA_ciZSA	● running on ['birne1330'] ocf::cust::sapresman	
grp_as02ZSA	● group	
grp_dbZSA	● group	

Connected to hacluster@127.0.0.1 (Simple Mode)

- SLE11-HA Cluster add-on for SLES11
- Custom resource agent for Oracle and SAP
- Host based mirrored storage
- Check for network connectivity

SUSE Linux Enterprise High Availability - Examples

Sample: HANA System Replication



- SLE11-HA Cluster add-on for SLES11
- Resource Agents SAPHana and SAPHanaTopology are part of SLES for SAP
- Currently single box system replication (Scale-Up) only
- One network segment (layer 2)
- No other HANA system on cluster nodes

Sample: HA for Oracle, simplified

Pacemaker GUI

Connection View Shadow Tools Help

Live

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

Name	Status	Details
Cluster	have quorum	Openais & Pacemaker
suse11-4d-102	online (dc)	
suse11-4d-101	online	
Resources		
rsc_stonith_sbd	running on ['suse11-4d-102'] stonith::external/sbd	
grp_DemoDB	group	
rsc_md_DemoDB_md0	running on ['suse11-4d-102'] ocf::heartbeat:Raid1	
rsc_lvm_DemoDB_oravg	running on ['suse11-4d-102'] ocf::heartbeat:LVM	
rsc_fs_DemoDB_ora	running on ['suse11-4d-102'] ocf::heartbeat:Filesystem	
rsc_fs_DemoDB_data	running on ['suse11-4d-102'] ocf::heartbeat:Filesystem	
rsc_ip_DemoDB_oracle1	running on ['suse11-4d-102'] ocf::heartbeat:IPAddr2	
rsc_oracle_DemoDB	running on ['suse11-4d-102'] ocf::heartbeat:oracle	

Validate With: pacemaker-1.2
Epoch: 65
Num Updates: 146
CRM Feature Set: 3.0.5
Have Quorum: 1
DC UUID: suse11-4d-102
CIB Last Written: Tue Oct 25 17:26:24 2011

Connected to hacluster@127.0.0.1 (Simple Mode)

- SLE11-HA Cluster add-on for SLES11
- Resource agent for Oracle is part of SLE11-HA
- Host based mirror
 - MD-RAID
 - LVM2
- SBD to handle split-brain scenarios

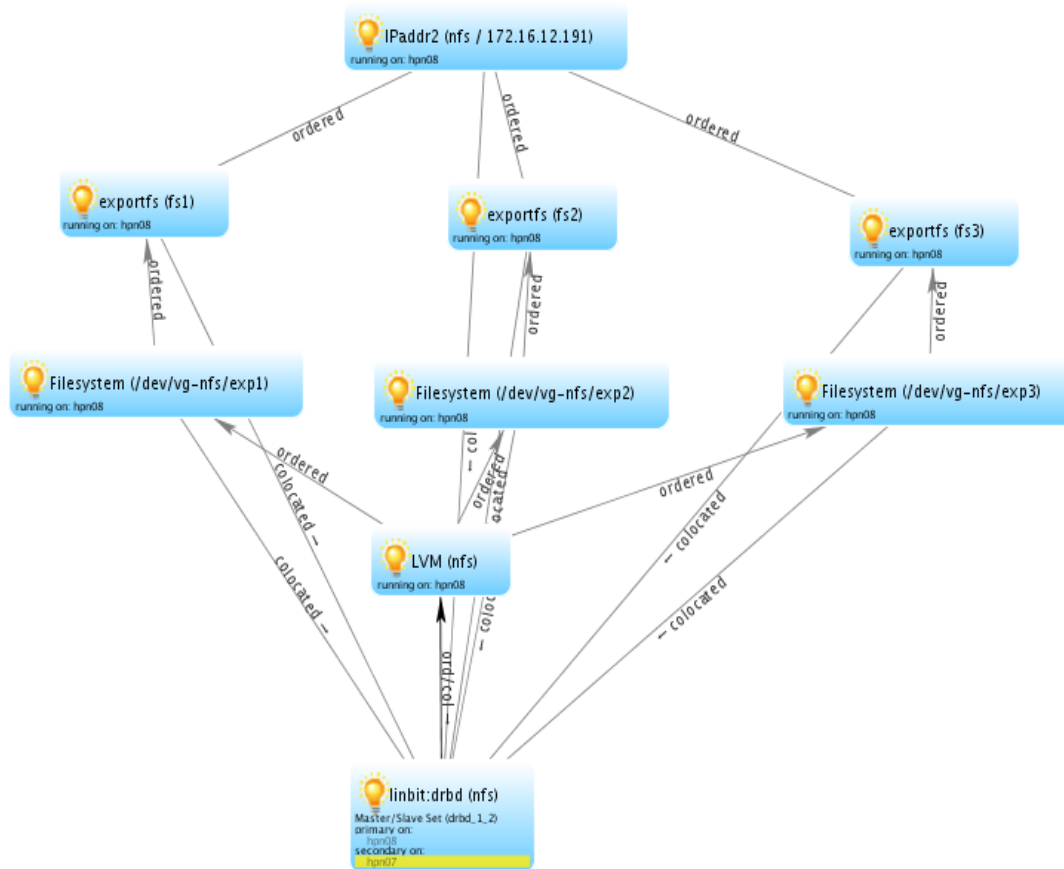
Sample: HA for Oracle

The screenshot shows the Oracle Clusterware configuration tool interface. The left sidebar contains a tree view with the following items: Configuration, CRM Config, Resource Defaults, Operation Defaults, Nodes, Resources, Constraints, and Management. The main pane displays a table of resources with columns for Name, Status, and Details. The resources are organized into a hierarchy: Cluster (have quorum), Resources (clone), grp_iti2t (group), and various rsc_* resources (running on [dezulsap94]). The status of all resources is 'running'. The details column shows the resource type and the node it is running on. At the bottom, there is a table with columns: Call ID, Operation, Interval, Return Code, Status, Last Run, Exec Time, Queue Time, and Last Return Code Change. The status bar at the bottom indicates 'Connected to 127.0.0.1 (Simple Mode)'.

Name	Status	Details
Cluster	have quorum	Openais & Pacemaker
Resources	clone	
grp_iti2t	group	
rsc_md_iti2t_md0	running on [dezulsap94]	ocf:heartbeat:Raid1
rsc_lvm_iti2t_vgora_iti2t	running on [dezulsap94]	ocf:heartbeat:LVM
rsc_fs_iti2t_oracle11	running on [dezulsap94]	ocf:heartbeat:Filesystem
rsc_fs_iti2t_oracle	running on [dezulsap94]	ocf:heartbeat:Filesystem
rsc_fs_iti2t_online	running on [dezulsap94]	ocf:heartbeat:Filesystem
rsc_fs_iti2t_mirror	running on [dezulsap94]	ocf:heartbeat:Filesystem
rsc_fs_iti2t_redo	running on [dezulsap94]	ocf:heartbeat:Filesystem
rsc_fs_iti2t_data	running on [dezulsap94]	ocf:heartbeat:Filesystem
rsc_fs_iti2t_index	running on [dezulsap94]	ocf:heartbeat:Filesystem
rsc_fs_iti2t_system	running on [dezulsap94]	ocf:heartbeat:Filesystem
rsc_ip_iti2t_dezulsap92	running on [dezulsap94]	ocf:heartbeat:IPAddr2
rsc_oracle_iti2t	running on [dezulsap94]	ocf:heartbeat:oracle
rsc_oralsnr_iti2t	running on [dezulsap94]	ocf:heartbeat:oralsnr
cln_ping	clone	

- SLE11-HA Cluster add-on for SLES11
- Resource agents for Oracle Instance and Oracle Listener are part of SLE11-HA
- Storage stack with or without host based mirror
- Check for network connectivity

Sample: HA NFS Server



- SLE11-HA Cluster add-on for SLES11
- Resource agents for NFS and DRBD are part of SLE11-HA
- Network based mirror
 - DRBD sync.
 - LVM2

Sample: OCFS2 Cluster

The screenshot shows the Pacemaker GUI window titled "Pacemaker GUI <@cisbwa013>". The left sidebar has a "Live" section with a tree view containing "Configuration", "Resource Def", "Operation Def", "Nodes", "Resources", "Constraints", and "Management". The "Management" item is selected. The main pane displays a table with three columns: "Name", "Status", and "Details".

Name	Status	Details
Cluster	● have quorum	Openais & Pacemaker
cisbwa015	● online	
cisbwa013	● online (dc)	
cisbwa014	● online	
cisbwa016	● online	
Resources	●	
cln_stonith_sbd	● clone	
cln_storage	● clone	
grp_storage:0	● group	
rsc_controld:0	● running on ['cisbwa013']	ocf::pacemaker:controld
rsc_o2cb:0	● running on ['cisbwa013']	ocf::ocfs2:o2cb
rsc_clvmd:0	● running on ['cisbwa013']	ocf::lvm2:clvmd
grp_storage:1	● group	
grp_storage:2	● group	
grp_storage:3	● group	
cln_bwa	● clone	
grp_bwa:0	● group	
rsc_lvm_vgbwa:0	● running on ['cisbwa013']	ocf::heartbeat:LVM
rsc_fs_bwa1:0	● running on ['cisbwa013']	ocf::heartbeat:Filesystem
grp_bwa:1	● group	
grp_bwa:2	● group	
grp_bwa:3	● group	

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

- SLE11-HA Cluster add-on for SLES11
- 4 Nodes, further scale-out possible
- Storage Stack
 - controld + o2cb
 - cLVM
 - OCFS2

Miscellaneous

Collect logs

- Base system
 - supportconfig -A
 - sam -o ...
 - tar czf ... /var/log/messages.*
- HA cluster
 - hb_report -f ...
 - tar czf ... /usr/lib/ocf/resource.d/

<http://thm-a01.yimg.com/nimage/d6dd37b46ee7dcb8.jpeg>



Read logs 1/2

Q1: Why did node lxeval01 reboot?

A1: Node lxeval01 was fenced by lxeval02, as shown in sbd slots.

Q2: Why was lxeval01 fenced?

A2: MD and LVM resources could not be stopped on lxeval01, as log on lxeval02 shows:

```
Feb 24 12:08:07 lxeval02 crmd: [9633]: WARN: update_failcount: Updating failcount for rsc_lvm_ZM2_ZM2db01 on lxeval01 after failed
start: rc=1 (update=INFINITY, time=1298545687)
Feb 24 12:08:08 lxeval02 crmd: [9633]: WARN: update_failcount: Updating failcount for rsc_md_ZM2_md10 on lxeval01 after failed stop:
rc=1 (update=INFINITY, time=1298545688)
...
Feb 24 12:08:08 lxeval02 pengine: [9632]: notice: LogActions: Move resource rsc_md_ZM2_md10 (Started lxeval01 -> lxeval02)
Feb 24 12:08:08 lxeval02 pengine: [9632]: notice: LogActions: Move resource rsc_lvm_ZM2_ZM2db01 (Started lxeval01 -> lxeval02)
...
Feb 24 12:08:08 lxeval02 crmd: [9633]: WARN: status_from_rc: Action 50 (rsc_md_ZM2_md10_stop_0) on lxeval01 failed (target:
0 vs. rc: 1): Error
Feb 24 12:08:08 lxeval02 crmd: [9633]: WARN: status_from_rc: Action 85 (rsc_lvm_ZM2_ZM2ci01_start_0) on lxeval01 failed (target:
0 vs. rc: 1): Error
...
Feb 24 12:08:08 lxeval02 pengine: [9632]: info: determine_online_status: Node lxeval01 is online
Feb 24 12:08:08 lxeval02 pengine: [9632]: WARN: unpack_rsc_op: Processing failed op rsc_md_ZM2_md10_stop_0 on lxeval01:
unknown error (1)
Feb 24 12:08:08 lxeval02 pengine: [9632]: WARN: pe_fence_node: Node lxeval01 will be fenced to recover from resource failure(s)
```


Read logs 2/2

Q3: Why could resources not be stopped?

A3: I/O to the SAN storage did not work, as log on lxeval01 shows:

```
Feb 24 12:08:04 lxeval01 kernel: [65951.343186] end_request: I/O error, dev dm-2, sector 144
Feb 24 12:08:04 lxeval01 kernel: [65951.344457] end_request: I/O error, dev dm-16, sector 144
Feb 24 12:08:04 lxeval01 kernel: [65951.344461] raid1: dm-16: rescheduling sector 0
Feb 24 12:08:04 lxeval01 kernel: [65951.344547] end_request: I/O error, dev dm-16, sector 144
...
Feb 24 12:08:05 lxeval01 kernel: [65951.348395] raid1: Disk failure on dm-19, disabling device.
Feb 24 12:08:05 lxeval01 kernel: [65951.348397] raid1: Operation continuing on 1 devices.
Feb 24 12:08:05 lxeval01 kernel: [65951.349345] raid1: Disk failure on dm-13, disabling device.
Feb 24 12:08:05 lxeval01 kernel: [65951.349346] raid1: Operation continuing on 1 devices.
Feb 24 12:08:05 lxeval01 kernel: [65951.349345] raid1: Disk failure on dm-13, disabling device.
Feb 24 12:08:05 lxeval01 kernel: [65951.349346] raid1: Operation continuing on 1 devices.
Feb 24 12:08:08 lxeval01 lrmd: [10119]: info: cancel_op: operation monitor[54] on ocf::Raid1::rsc_md_ZM2_md10 for client 10122,
its parameters: raidconf=[/etc/mdadm.conf.cluster] on_fail=[fence] crm_feature_set=[3.0.2] raiddev=[/dev/md10] CRM_meta_on_fail=[fence]
CRM_meta_name=[monitor] CRM_meta_interval=[120000] CRM_meta_timeout=[60000] cancelled
...
Feb 24 12:08:08 lxeval01 lrmd: [10119]: info: rsc:rsc_md_ZM2_md10:111: stop
Feb 24 12:08:08 lxeval01 crmd: [10122]: info: process_lrm_event: LRM operation rsc_md_ZM2_md10_monitor_120000 (call=54, status=1,
cib-update=0, confirmed=true) Cancelled
Feb 24 12:08:08 lxeval01 lrmd: [10119]: info: RA output: (rsc_md_ZM2_md10:stop:stderr) mdadm: failed to stop array /dev/md10:
Device or resource busy Perhaps a running process, mounted filesystem or active volume group?
Feb 24 12:08:08 lxeval01 lrmd: [10119]: info: RA output: (rsc_md_ZM2_md10:stop:stderr) mdadm: failed to set readonly for /dev/md10:
Device or resource busy
Feb 24 12:08:08 lxeval01 lrmd: [10119]: WARN: Managed rsc_md_ZM2_md10:stop process 12921 exited with return code 1.
Feb 24 12:08:08 lxeval01 crmd: [10122]: info: process_lrm_event: LRM operation rsc_md_ZM2_md10_stop_0 (call=111, rc=1,
cib-update=119, confirmed=true) unknown error
```

Q4: Why did I/O not work?

A4: Further investigation of I/O stack needed to find the answer.



Grep the logs

Q5: Are there tools to grep the logs?

A5: grep_error_patterns, grep_cluster_patterns, grep_cluster_transition
are part of ClusterTools2.

<TODO>

Q6: Are there tools to grep supportconfig?

A6: grep_supportconfig is part of ClusterTools2

<TODO>

Observations

- Pre-defined, well tested stacks are crucial for quality, supportability, and acceptance.
- Storage is most important part for production. Storage stack has to be built out of a modular set of options. Many ways to fail, few to succeed.
- Customer fall into one out of four categories regarding storage:
 - a) host based mirror
 - b) storage based mirror
 - c) no mirror at all
 - d) NFS, what so everCombination 1)+b) not possible yet, except with 3rd party- or selfmade RA.
- Customers fall into one of these two categories:
 - 1) Your cluster does not do everything magically, why that?
 - 2) Any cluster should not do anything, except the server is de-materialised.

Timing

- How long a cluster survives a storage outage depends on the watchdog timeout and the sbd retry cycle. All other timeouts should be aligned with that settings. That means they have to be longer.
- If the watchdog should avoid MD mirror splitting in case of IO errors, the watchdog timeout has to be shorter than the total MPIO failure timeout. Thus, a node is fenced before the MD mirror is split.
- Storage resources - as Raid1, LVM, Filesystem - have operation timeouts. These should be longer than the MPIO failure timeout. This avoids non-needed
- failure actions. It does not define how long the cluster survives a storage outage.
- The watchdog timeout must be shorter than sbd message wait timeout. The sbd message wait timeout must be shorter than the cluster stonith-timeout.
- The cluster sbd resource should have a start interval longer than the sbd daemon loop cycle.
- <TODO> by factor 1.2
- <TODO> sbd recover

See man pages for sbd, multipath.conf.

Appendix

Info Resources 1/2

<http://www.novell.com/products/server/>
http://www.novell.com/linux/releasenotes/x86_64/SUSE-SLES/11-SP3/
<http://www.novell.com/linux/techspecs.htm>
http://wiki.novell.com/index.php/Linux_Data_Management
<http://www.novell.com/products/highavailability>
http://www.novell.com/linux/releasenotes/x86_64/SLE-HA/11-SP3/
<http://www.novell.com/partners/sap/>
http://www.novell.com/partners/sap/novell_sap_partnership.html
<http://www.novell.com/products/server/sap/matrix.html>
http://www.novell.com/documentation/sles11/stor_admin/?page=/documentation/sles11/stor_admin/data/bookinfo.html
http://www.novell.com/documentation/sle_ha/book_sleha/?page=/documentation/sle_ha/book_sleha/data/book_sleha.html
http://doc.opensuse.org/products/draft/SLE-HA/SLE-ha-guide_sd_draft/book.sleha.html
http://software.opensuse.org/search?q=ClusterTools2&baseproject=SUSE%3ASLE-11%3ASP1&lang=en&include_home=true&exclude_debug=true
http://www.clusterlabs.org/doc/en-US/Pacemaker/1.1/html/Pacemaker_Explained/s-operation-defaults.html#s-operation-timeouts
https://bugzilla.novell.com/show_bug.cgi?id=776386
<https://www.suse.com/support/kb/doc.php?id=7010781>

<http://www.sap.com/linux>

<http://scn.sap.com/community/linux/blog/2014/02/17/installation-wizard-from-suse-linux-enterprise-server-ready-to-run-sap-solutions-in-just-a-few-steps>
<http://scn.sap.com/community/hana-in-memory/blog/2014/04/04/fail-safe-operation-of-sap-hana-suse-extends-its-high-availability-solution>

http://www.saphana.com/servlet/JiveServlet/previewBody/2775-102-4-9467/HANA_HA_2.1.pdf
<http://www.sdn.sap.com/irj/scn/index?rid=/library/uuid/d079c1f2-5c6b-2f10-dcbb-a29c20865e8>
<http://scn.sap.com/docs/DOC-52522>

<https://www.suse.com/promo/saphana-replication.html>

<click>



Info Resources 2/2

<http://www.linux-ha.org/>
<http://www.openais.org/>
<http://www.clusterlabs.org/>
http://www.clusterlabs.org/mediawiki/images/f/fb/Configuration_Explained.pdf
<http://www.anchor.com.au/blog/2012/05/hunting-down-unexpected-behaviour-in-corosyncs-ip-address-selection/>
http://www.linux-ha.org/wiki/OCF_Resource_Agent
<http://www.infokey.com/find/index.php?page=search/special&search=open+source&type=PDF&startpage=3>
<http://www.querzone.de/wiki/Wiki.jsp?page=Multipathing>
https://raid.wiki.kernel.org/index.php/Linux_Raid
http://www.corosync.org/doku.php?id=faq:cisco_switches
<http://de.wikipedia.org/wiki/Multicast>
http://en.wikipedia.org/wiki/Internet_Group_Management_Protocol
<http://www.sebastien-han.fr/blog/2012/08/01/corosync-rrp-configuration/>
<http://www.hastexo.com/resources/hints-and-kinks/fencing-vmware-virtualized-pacemaker-nodes>

<click>

Trainings

<https://training.sap.com/v2/course/wdehsu-high-availability-sap-with-suse-linux-enterprise-server-11-for-sap-application-classroom-001-de-de/>

<https://www.suse.com/training/high-availability/course-9211.html>

<https://www.suse.com/training/sles-for-sap/course-9073.html>

<https://www.suse.com/training/server/course-9064.html>

<https://www.suse.com/training/sles-for-sap/course-9074.html>

<click>

Quiz time



Corporate Headquarters
Maxfeldstrasse 5
90409 Nuremberg
Germany

+49 911 740 53 0 (Worldwide)
www.suse.com

Join us on:
www.opensuse.org

Unpublished Work of Novell, Inc. All Rights Reserved.

This work is an unpublished work and contains confidential, proprietary, and trade secret information of Novell, Inc. Access to this work is restricted to Novell employees who have a need to know to perform tasks within the scope of their assignments. No part of this work may be practiced, performed, copied, distributed, revised, modified, translated, abridged, condensed, expanded, collected, or adapted without the prior written consent of Novell, Inc. Any use or exploitation of this work without authorization could subject the perpetrator to criminal and civil liability.

General Disclaimer

This document is not to be construed as a promise by any participating company to develop, deliver, or market a product. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. Novell, Inc. makes no representations or warranties with respect to the contents of this document, and specifically disclaims any express or implied warranties of merchantability or fitness for any particular purpose. The development, release, and timing of features or functionality described for Novell products remains at the sole discretion of Novell. Further, Novell, Inc. reserves the right to revise this document and to make changes to its content, at any time, without obligation to notify any person or entity of such revisions or changes. All Novell marks referenced in this presentation are trademarks or registered trademarks of Novell, Inc. in the United States and other countries. All third-party trademarks are the property of their respective owners.

