

SUSE® Linux Enterprise 11

Platform for SAP

Basic Architecture Overview

Fabian Herschel

SUSE Linux GmbH

Lars Pinne

SUSE Linux GmbH



Agenda

- Architecture for SAP
- Storage
- Network
- Cluster Manager Pacemaker/Corosync
- Cluster Test and Maintenance
- Appendix



Architecture for SAP

Architecture: Design and Goals



- Top level design and goals:

- High Availability
- Low Complexity
- Flexible Scalability
- Road Capability



- To fit these goals, we separate the SAP system into a clustered and an unclustered area. The clustered area holds all mandatory SAP components such as SAP database and needed SAP instances.



- The unclustered area holds the optional and scalable SAP components such as additional SAP instances. This allows to scale the entire SAP system without increasing the cluster complexity. The horizontal scaling is just a purpose of the unclustered area.

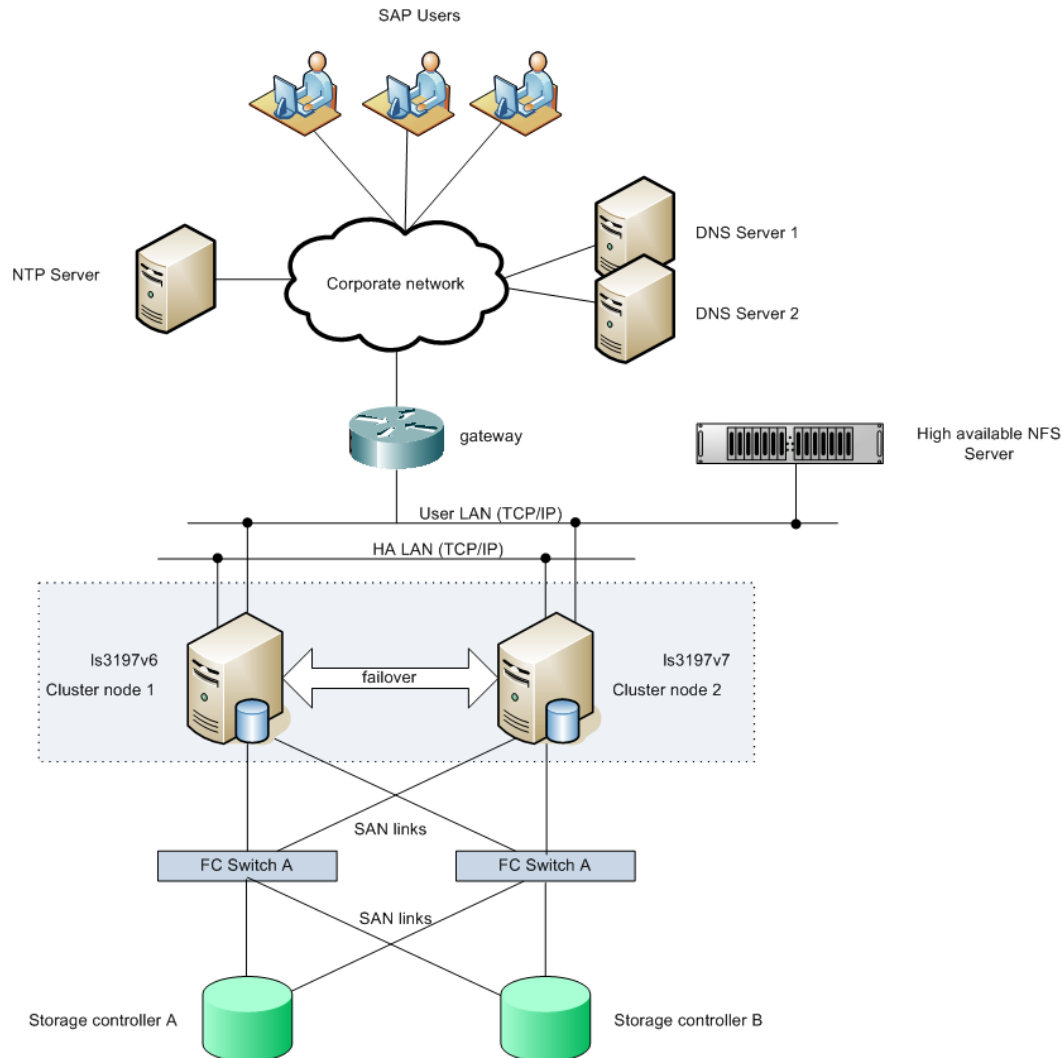


- The architecture is focused to one single SAP system, even if is possible to run more than on SAP system in the same cluster.



- Alternative scenarios “Enqueue Replication” or “Simple Stack”.

Architecture: Overview



• Concept

- Separation of un-clustered SAP components for optional scale-out
- 2node cluster with hostbased mirrored shared storage
- 2tier SAP system
- “Enqueue Replication” or “**Simple Stack**” scenario

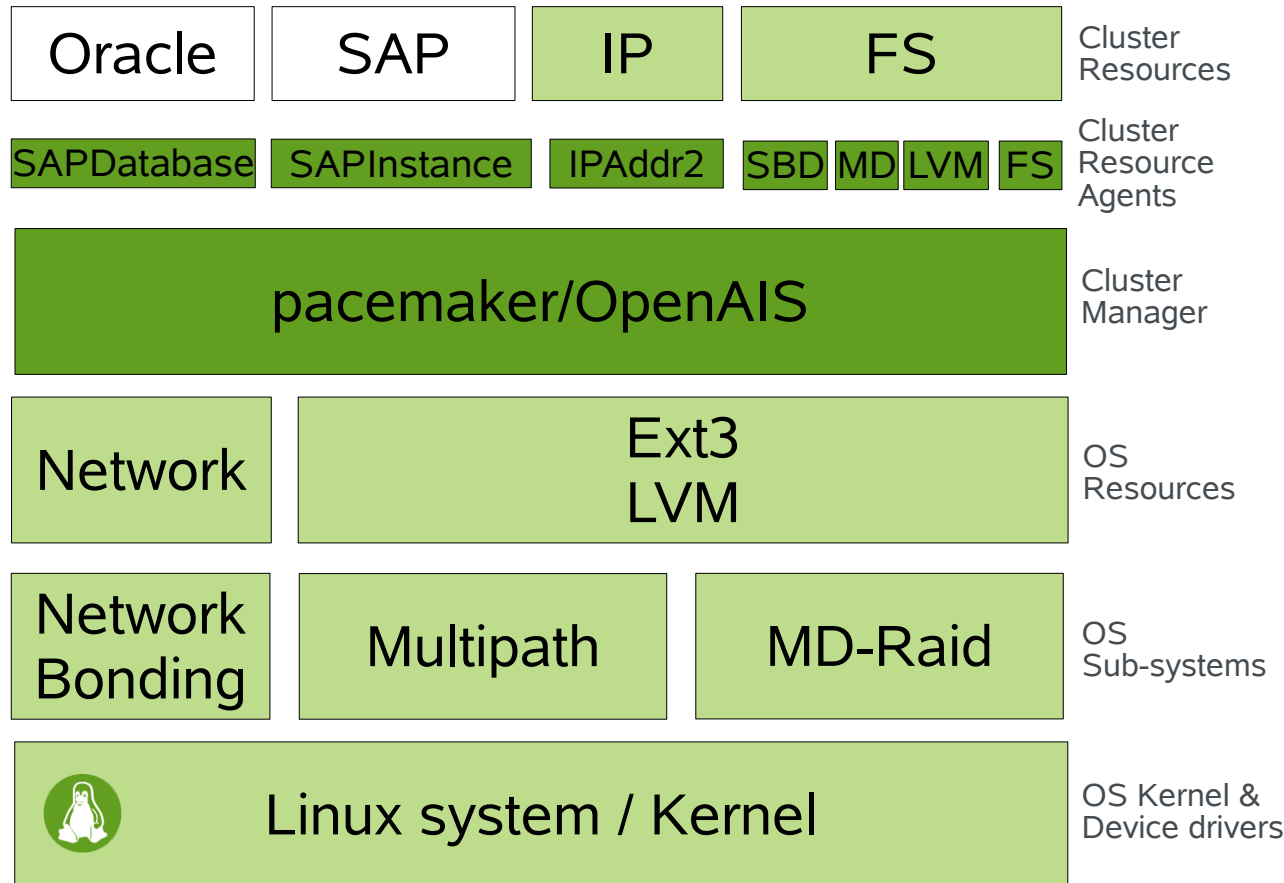
• Software

- SAP Netweaver 7.30
- MaxDB 7.7
- SUSE Linux Enterprise Server for SAP Applications 11sp2 x86_64

• Hardware

- 2 servers x86_64, 8GB RAM, 4 NIC, 2 HBA, remote mgmt. board
- 2 SAN LUNs 100GB

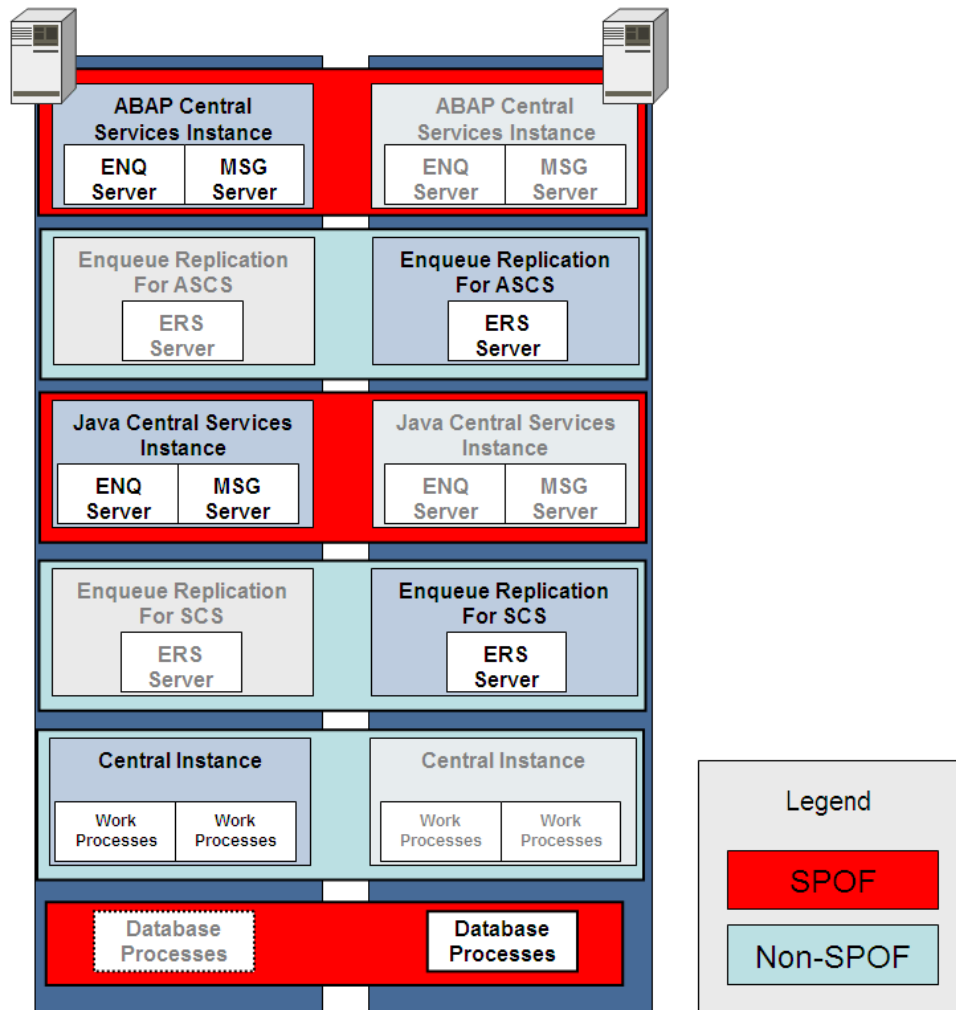
HA Stack for SAP



- Cluster manager pacemaker/openAIS from SLE11-HAE
- Resource agents for SAP Instance and SAP Database are part of SLE11-HAE
- Supported databases
 - DB/2
 - MaxDB
 - Sybase ASE
 - Oracle
- SBD to handle split-brain scenarios
 - SFEX also possible

SAP Architecture #1

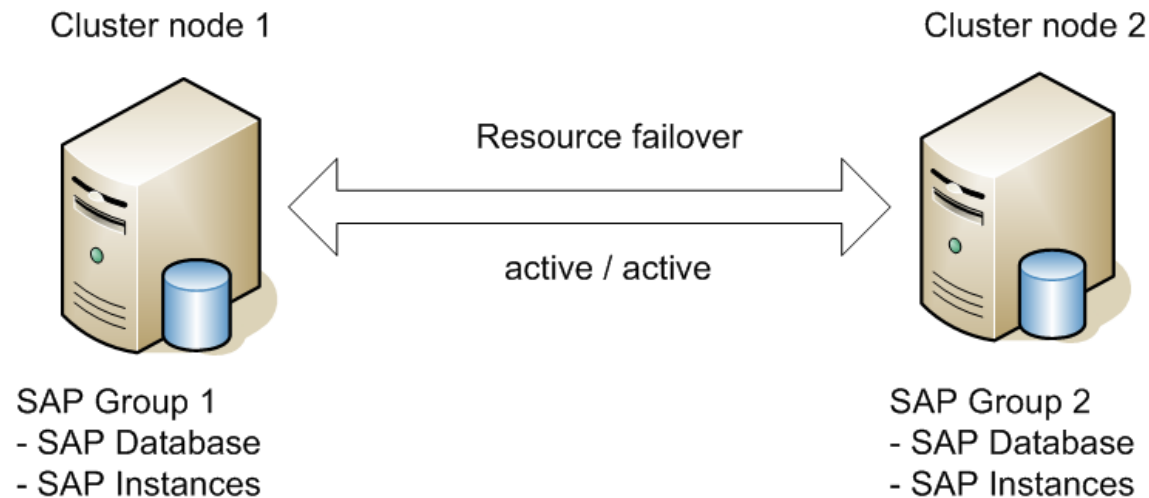
Enqueue Replication



- SAP system balanced (database/instances) on both cluster nodes.
- SAP certified HA for enqueue replication server.
- Multiple SAP systems are possible from the technical point of view, but heavily increase the complexity.
- Fits well for running one large SAP system on a dedicated cluster.

SAP Architecture #2

Simple Stack



- Entire SAP system runs on one node.
- No enqueue replication server.
- Multiple SAP systems are possible. For two SAP system slightly increased complexity .
- Fits well for running multiple SAP systems of small to medium size on one cluster.

Storage

Filesystems for SAP #1

Enqueue Replication

- Local filesystems

- /usr/sap

- /usr/sap/HB2

- /usr/sap/HB2/ERS10

- /usr/sap/HB2/ERS11

- Shared storage filesystems

- /usr/sap/HB2/DVEBMSG02

- /sapdb

- /sapdb/HB2/sapdata

- /sapdb/HB2/saplog

- Network-mounted filesystems

- /sapmnt/HB2

- /usr/sap/HB2/ASC00

- /usr/sap/HB2/SCS01

Note: HB2 is the name of this particular SAP system instance

- Filesystems for ERS instances are local on every node because of the Master-Slave mechanism of the SAPInstance resource agent.
- Filesystems for database tables and transaction logs are on a shared storage for active/passive clustered DB operation.
- Filesystems for ASCS and SCS are on a network filesystem for optional scale-out.

Filesystems for SAP #2

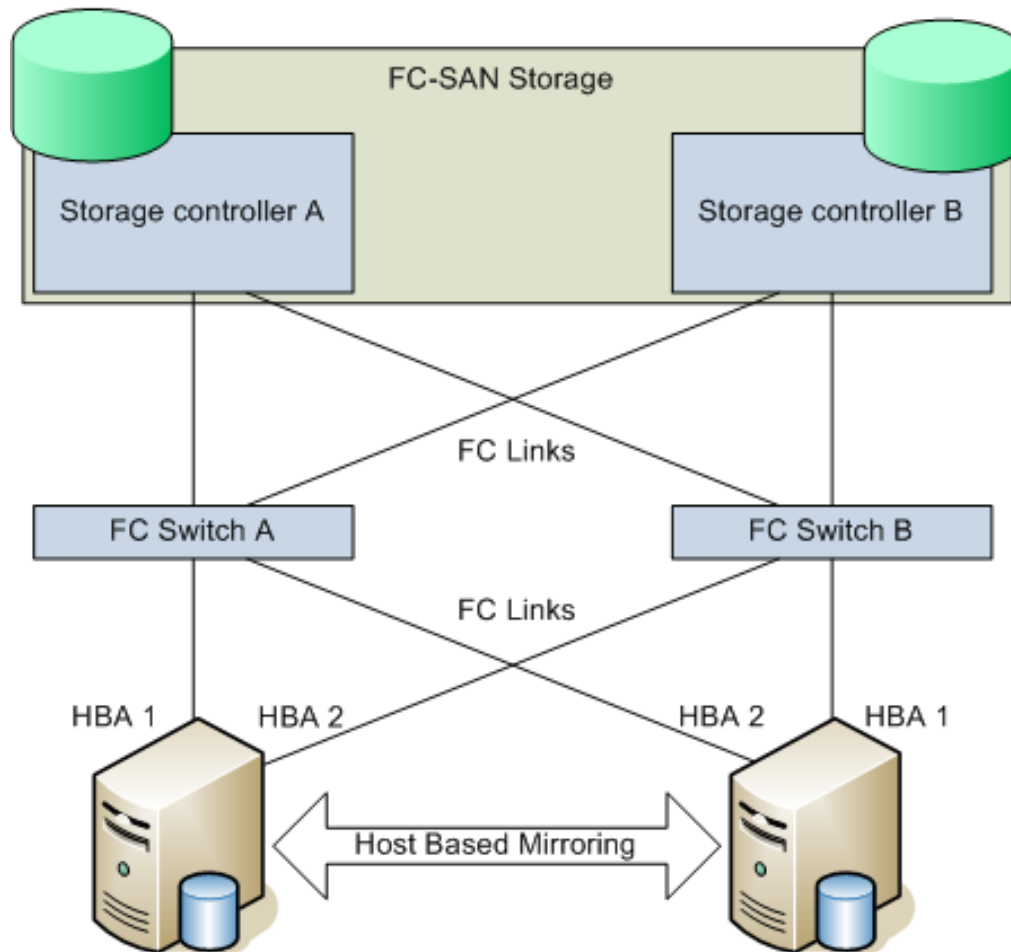
Simple Stack

- Local filesystems
- Shared storage filesystems
 - `/usr/sap/HB2`
 - `/sapdb`
 - `/sapdb/HB2/sapdata`
 - `/sapdb/HB2/saplog`
- Network-mounted filesystems
 - `/sapmnt/HB2`

Note: HB2 is the name of this particular SAP system instance

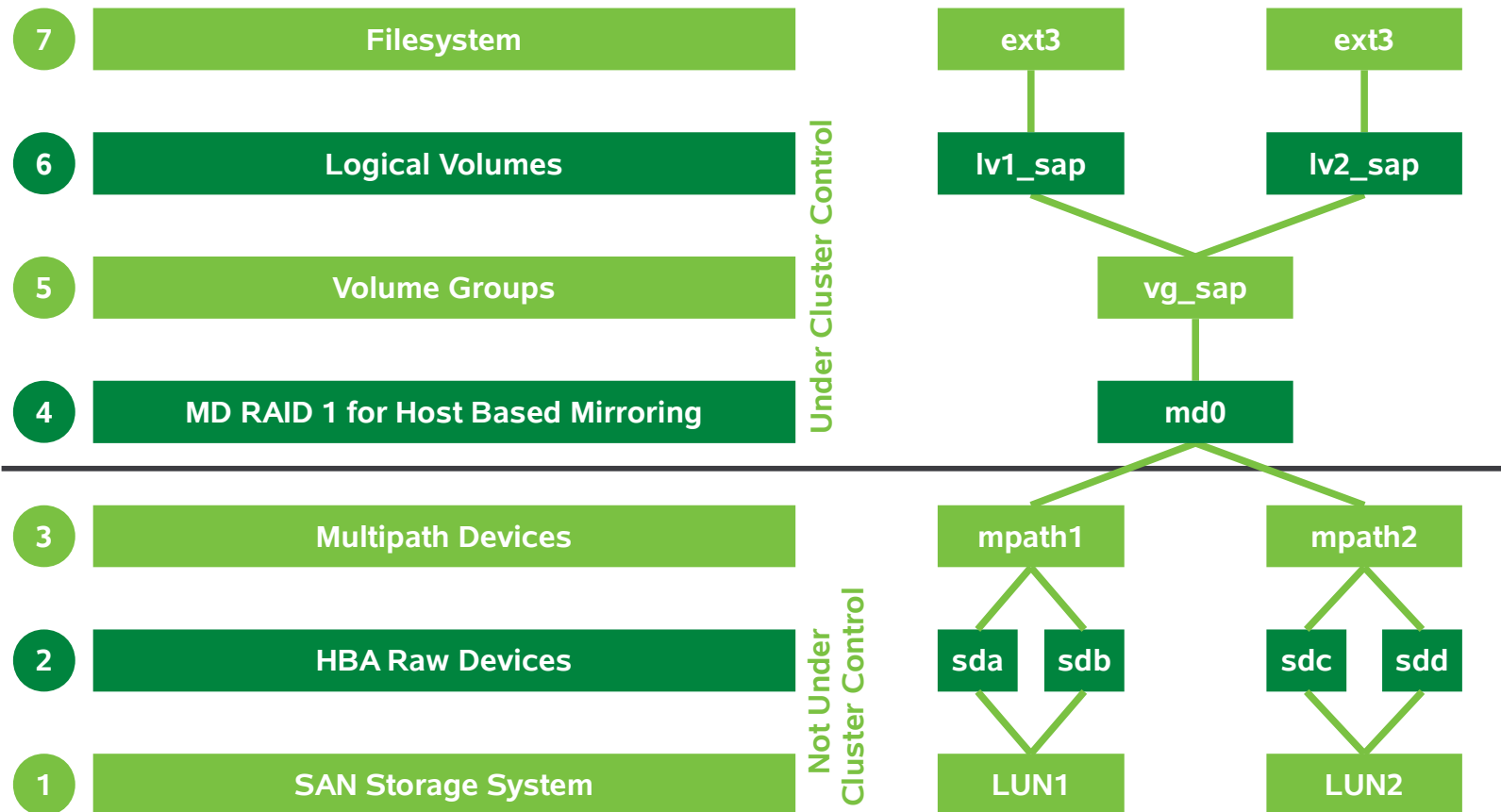
- No local filesystems for SAP.
- Filesystems for SAP instances are on a shared storage for active/passive cluster operation.
- Filesystems for database tables and transaction logs are on a shared storage for active/passive clustered DB operation.

Hostbased Mirroring



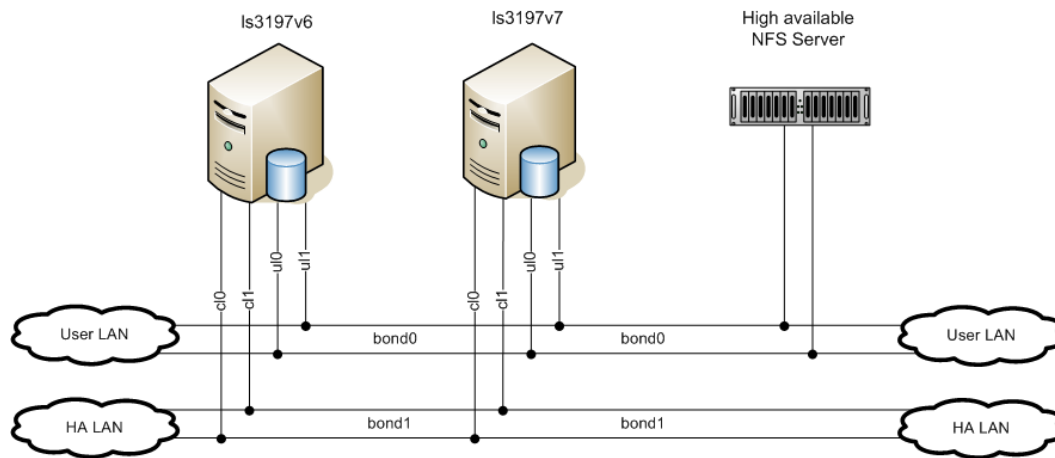
- No Single Point of Failure
- Supports two separate fire sections
- Limited distance
- Integrated with cluster manager
- Components are part of SLES
- Mostly hardware-independent
- Split-brain handling needs special care
- In responsibility of OS or SAP operating team

IO Layers of HA Stack for SAP



Network





Network



- Separate networks for internal and external communication
- Network bonding to increase network stableness.
- Network infrastructure should be redundant, too.

Cluster Manager Pacemaker/Corosync

Resource Agents for SAP 1/3

- . Operations of the SAPInstance resource agent are done by using the startup framework called SAP Management Console or sapstartsrv (see Note #1014480). Defaults paths fit to SAP kernel directory location after the default SAP installation..
- . sapstartsrv knows 4 status colours:
 -  GREEN = everything is fine
 -  YELLOW = something is wrong, but the service is still working
 -  RED = the service does not work
 -  GRAY = the service has not been started
- . SAPInstance resource agent will interpret GREEN and YELLOW as OK, statuses RED and GRAY are reported as NOT_RUNNING to the cluster.
- . Depending on the status the cluster expects from the resource, it will do a restart, failover or just nothing.

Resource Agents for SAP 2/3

- SAPInstance resource agent
 - disp+work
 - msg_server
 - enservice
 - enrepserver
 - jcontrol
 - jstart
 - wdisp
- SAPDatabase resource agent
 - DB/2
 - MaxDB
 - Oracle
 - Sybase

Supported Releases:

- SAP WebAS ABAP Release 4.6C - 7.10
- SAP WebAS Java Release 6.40 - 7.10 (min. 6.40 SP22, 7.00 SP15, 7.10 SP00)
- SAP WebAS ABAP + Java Add-In Release 6.20 - 7.30
- DB/2 UDB 9.x, 10.x
- Oracle 10gR2, 11gR2
- SAP-DB / MaxDB 7.6, 7.7
- Sybase ASE 15.7

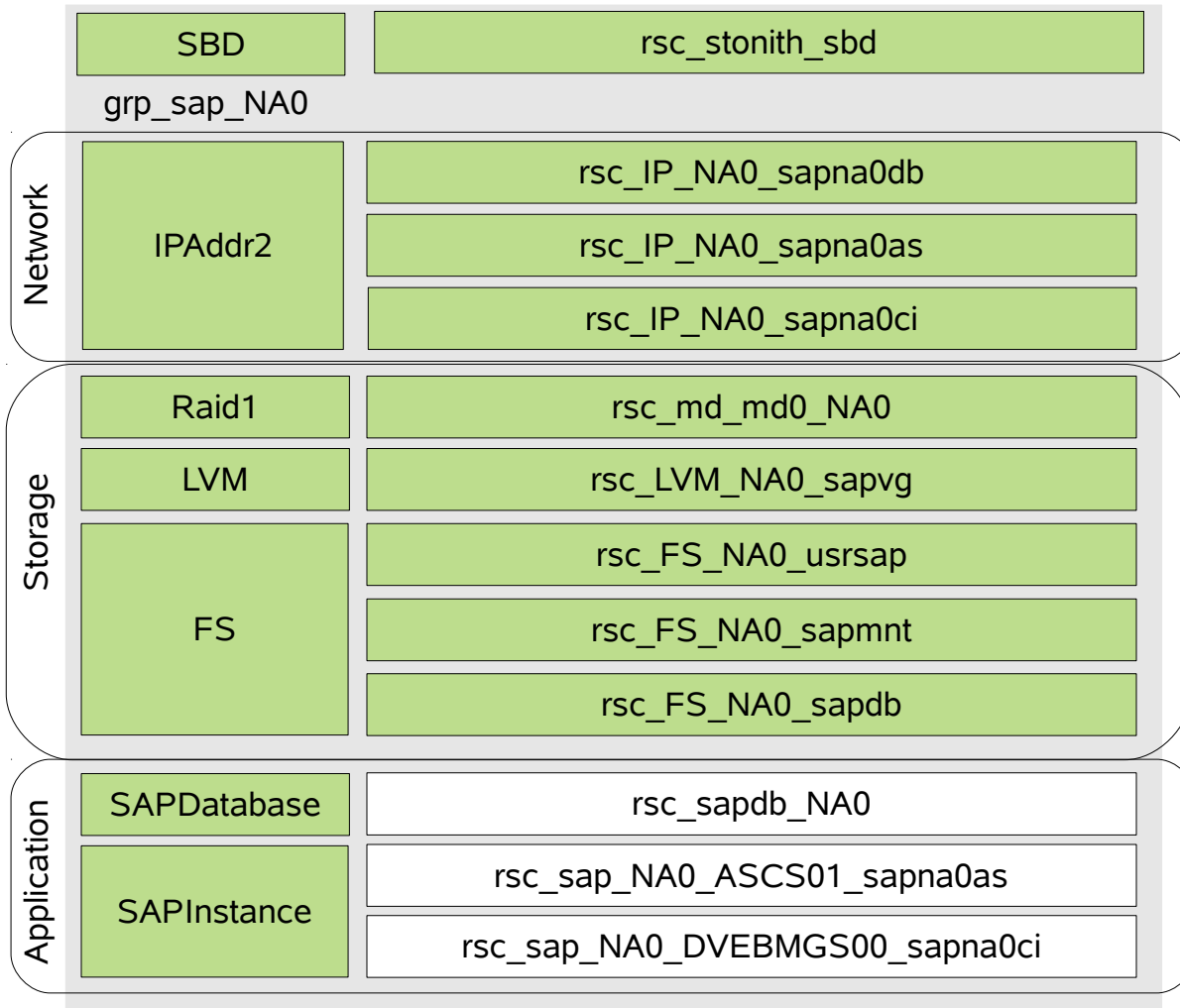
<http://www.suse.com/products/sles-for-sap/resource-library>

Resource Agents for SAP 3/3

- SAPInstance resource agent (old style)
 - **Start:** SAP Instances are started by 'sapcontrol -nr \$InstanceNr -function Start'. Optionally the RA tries to recover a failed start attempt one time.
 - **Stop:** SAP Instances are stopped by 'sapcontrol -nr \$InstanceNr -function Stop'. Optionally a faster kill plus cleanipc could be used.
 - **Monitor:** sapstartsrv uses SOAP messages to request the status of running SAP processes. RA is calling 'sapcontrol -nr \$InstanceNr -function GetProcessList -format script'.
 - Start and Stop allow pre- and post-userexit scripts.
 - See ocf_heartbeat_SAPInstance(7),
/usr/lib/ocf/resource.d/heartbeat/SAPInstance

Cluster Resources #1

Simple Stack with SBD



Cluster Resources #1

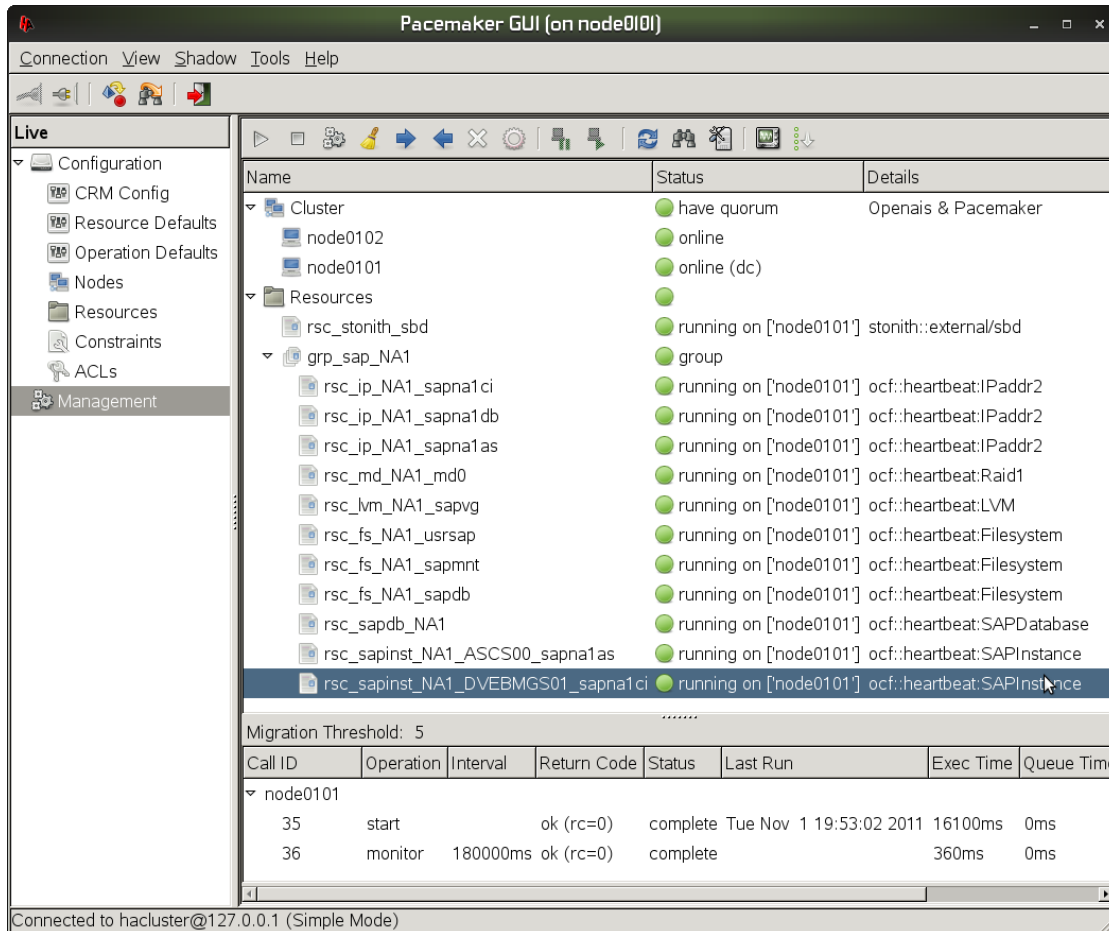
Simple Stack with SBD

- SBD as primitive resource.
- One common resource group for all SAP resources.
- Location constraints are needed to ensure that the SAP system is only started if the node fulfills all operational needs.
- Co-location constraints are implicitly defined by adding all resources to one common group.
- Order constraints are implicitly defined by adding all resources to one common group.

- Cluster Resources:
 - SBD: Server fencing
 - IPAddr2: Virtual IP addresses
 - Raid1: MD Raid1 arrays
 - LVM: LVM Volume Groups
 - Filesystem: Ext3 filesystems
 - SAPDatabase: MaxDB Database
 - SAPInstance: Java Central services instance
 - SAPInstance: Central instance
 - SAPInstance: Dialog instance

Cluster Resources #1

Simple Stack with SBD



The screenshot shows the Pacemaker GUI (on node0101) with the following components:

- Left Panel (Live):** Configuration, CRM Config, Resource Defaults, Operation Defaults, Nodes, Resources, Constraints, ACLs, Management.
- Main Panel:** A table listing cluster resources and their status.
- Bottom Panel:** Migration Threshold: 5, and a table showing recent operations.

Name	Status	Details
Cluster	● have quorum	Openais & Pacemaker
node0102	● online	
node0101	● online (dc)	
Resources	●	
rsc_stonith_sbd	● running on [node0101]	stonith::external/sbd
grp_sap_NA1	● group	
rsc_ip_NA1_sapna1ci	● running on [node0101]	ocf::heartbeat:IPAddr2
rsc_ip_NA1_sapna1db	● running on [node0101]	ocf::heartbeat:IPAddr2
rsc_ip_NA1_sapna1as	● running on [node0101]	ocf::heartbeat:IPAddr2
rsc_md_NA1_md0	● running on [node0101]	ocf::heartbeat:Raid1
rsc_lvm_NA1_sapvg	● running on [node0101]	ocf::heartbeat:LVM
rsc_fs_NA1_usrsap	● running on [node0101]	ocf::heartbeat:Filesystem
rsc_fs_NA1_sapmnt	● running on [node0101]	ocf::heartbeat:Filesystem
rsc_fs_NA1_sapdb	● running on [node0101]	ocf::heartbeat:Filesystem
rsc_sapdb_NA1	● running on [node0101]	ocf::heartbeat:SAPDatabase
rsc_sapinst_NA1_ASCS00_sapna1as	● running on [node0101]	ocf::heartbeat:SAPInstance
rsc_sapinst_NA1_DVEBMS01_sapna1ci	● running on [node0101]	ocf::heartbeat:SAPInstance

Call ID	Operation	Interval	Return Code	Status	Last Run	Exec Time	Queue Time
node0101							
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)

- Symmetrical two-node cluster
- SAP Simple Stack
- Database MaxDB
- Three filesystems
- Logical Volume Manager
- MD Host based mirror
- SBD Server fencing

Cluster Resources #2

Enqueue Replication with external Database

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".

The interface is divided into three main sections:

- Online Resources:** A list of resources that are currently online or started, each with a magnifying glass icon for details.
 - cl2n01: Online
 - cl2n02: Online
 - Master/Slave Set: msl_sap_enqrepl_HA0
 - rsc_sap_HA0_ASCS00:0: Master
 - rsc_sap_HA0_ASCS00:1: Slave
 - 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
 - rsc_sap_HA0_D02: Started
- Inactive Resources:** A section on the right side of the interface, currently empty.

SAP[®] Certified
Integration with SAP NetWeaver[®]

Cluster Test and Maintenance

Failures and Solutions

DRAFT - Example Two SBDs plus corosync

	Multipath	Bonding	MD Raid	Watchdog – SBD	Stonith – SBD	Resource Failover	Corosync
One node crashes (no response)					1	2 (for resources on fenced node)	
Complete loss of SAN connectivity on one node				1	2 (try)	3 (for resources on fenced node)	
Network outage, cluster intercommunication fails					1	2 (for resources on fenced node)	
Partial network outage on any node (one ring fails)							1
Partial network outage on any node (one link fails)		1					
SAN outage of one disk array (with one SBD device)			1 (Raid1 degraded)				
Partial SAN outage on any node (one link)	1						
SAN interconnect between sites fails			1 (Raid1 degraded)				
SAN interconnect between sites fails, Partial network outage (one ring fails), One node crashes			1 (Raid1 degraded*)	2 (node suicide)		3 (manual restart of resources)	
Split-Site (local LAN and SAN working)			1 (Raid1 degraded*)	2 (node suicide)		3 (manual restart of resources)	
Complete outage of one data center			1 (Raid1 degraded*)	2 (node suicide)		3 (manual restart of resources)	

Remarks:

* if watchdog timeout is shorter than multipath

Failures and Solutions

Example SFEX

	Multipath	Bonding	Resource restart	Resource failover	Stonith	SFEX	Ping Nodes
Application crash on active node			1 st action 1)	2 nd action			
Active node crashes				2 nd action	1 st action 2)		
Network outage, cluster inter-communication fails (Split-Brain)					1 st action 3)	1 st action 4)	
Partial network outage on any node (one link fails)		Switch to 2 nd 5)					
Partial network outage on active node (gateway not reachable)				2 nd action			1 st action 6)
SAN outage on active node				2 nd action	1 st action 7)	SFEX 8)	
Partial SAN outage on any node (one link)	1 st action 9)						
Power outage of the active node				2 nd action	1 st action 10)		
Split-Site (not described in this Use Case)						1 st action 11)	

Remarks:

1. 3 times
2. Ensures that the node is really dead
3. if Stonith is enabled
4. if Stonith is disabled
5. Ensures that the node is really dead
6. Shutdown of active resources
7. triggered by filesystem monitoring
8. detects missing locking disk
9. Failover to 2nd path
10. Requires operator interaction if **no** external APC device for Stonith is used
11. Disk locking secures SAN devices

Cluster Test

Enqueue Replication with SFEX



For productive clusters a test plan for all possible failure scenarios is necessary. In the workshop, some important basic tests are used:

Test Case	Expected Result
Set one node to standby	Database and CI instance group is running on remaining node. Virtual IPs are running on remaining node.
Shutdown one node gracefully	The node leaves the cluster. The enqueue replication table is moved. Database and CI instance group start on remaining host.
Turn off one node (power off)	The remaining node will try STONITH. If the STONITH was successful, the remaining node takes over all resources.
Turn on node again, start openais	The node rejoins the cluster and starts resources. Also the CI host moves to the rejoined node.
Plug out User LAN	The second node takes over all resources. Depending on the configuration, the affected node gets STONITHed.
Plug out both SAN links	The monitoring of the SAN resources fail. The affected node gets STONITHed.
Plug out all network (split-brain)	The cluster stops doing anything (resource keep running), since both nodes can't determine the state of their counterpart.
Kill SAP CI instance	The CI instance restarts 3 times. After the 3rd try, it fails over to the second node
Kill database	The database instance restarts 3 times. After the 3rd try, it fails over to the second node
Shut down one SAN storage	The MD-Mirrors get degraded but continue to work.

Cluster Tools

```
+-----+
| NSO      : set node online          |
| NSS      : set node standby         |
| CSAN     : show status all nodes    |
| NSSTAT   : show status specific node |
| RDN      : set target-role stopped  |
| RUP      : set target-role started  |
| CSS      : show status of all resources |
| RSU      : set unmanaged for resource |
| RSM      : set managed for resource  |
| RMI      : migrate resource         |
| RUMH     : migrate resource to specific node |
| RUM      : unmigrate resource       |
| CSA      : show (open) cluster actions using ptest |
| RCL      : cleanup resource         |
| RCLN     : cleanup resource node    |
| CSFN     : show failcount of resource on node |
| CDFN     : delete failcount of resource on node |
| NFENCE   : FENCE node!!            |
| MP       : multipath status         |
| LS       : show mapper devices      |
| LOG      : show CLX log entries     |
| LF       : show failcounts          |
| RF       : reset failcounts         |
+-----+
command: RMI grp_sap_NA2
```

ClusterTools2 RPM is a collection of tools, used in projects by SUSE consulting. Examples:

- `precheck_for_sap` is a script to check the OS base-installation for pre-requisites needed by SAP and OpenAIS.
- `ClusterService` is a frontend for most used cluster operations. It is designed along the use cases that are common for operating a clustered SAP system.
- `clusterstate` is a script to analyse cluster status.
- `Wow` is a simple program, which helps to create cluster configurations. It is based on `crm`-snippets and scripts. The snippets may contain variable references which are resolved by the scripts and the `wow` engine.
- `grep_cluster_patterns` is a script to filter syslog entries.

Appendix

Links

SAP Notes

0171356	SAP on Linux
0816097	Availability of SAP on Linux for x86_64
0958253	SLES10
1310037	SLES11
1275775	sapconf
1172419	Java VM on x86_64
1008828	SAP Kernel 4.6D
0877795	SAP sapstartsrv before 7.0
0995116	SAP backporting sapstartsrv
1014480	SAP Mgmt. Console
1398634	Oracle 11
0618104	sapsysinfo

Links

<http://www.sap.com/linux>
<http://wiki.sdn.sap.com/wiki/display/HOME/SAPonLinuxNotes>
<http://www.novell.com/de-de/partners/sap>
<http://www.novell.com/products/server/sap/matrix.html>
<http://www.suse.com/products/sles-for-sap/resource-library>
http://developer.novell.com/wiki/index.php/SAP_on_hasi_v2
<http://www.novell.com/saptechdocs>
<http://www.novell.com/rc/docrepository/public/7/basedocument.2009-04-30.4730280192>
http://download.opensuse.org/repositories/home:/fmherschel/SLE_11_SP1/noarch/

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 SUSE. All Rights Reserved.

This work is an unpublished work and contains confidential, proprietary and trade secret information of SUSE. Access to this work is restricted to SUSE 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 SUSE. 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. SUSE 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 SUSE products remains at the sole discretion of SUSE. Further, SUSE 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 SUSE 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.

