Zen and
The Art of High-Availability Clustering

Lars Marowsky-Brée

Architect Storage and High-Availability, Principal Engineer

lmb@novell.com

Novell.

Agenda

Introduction

Summary of cluster architecture

Common configuration issues

Gathering cluster-wide support information

Exploring effects of cluster events

Self-written resource agents

Understanding log files

Introduction

Data Center Challenges

- Minimize unplanned downtime

- Ensure quality of service

- Contain costs

- Utilize resources

- Effectively manage multiple vendors

- Minimize risk

Pacemaker and corosync/OpenAIS-based stacks
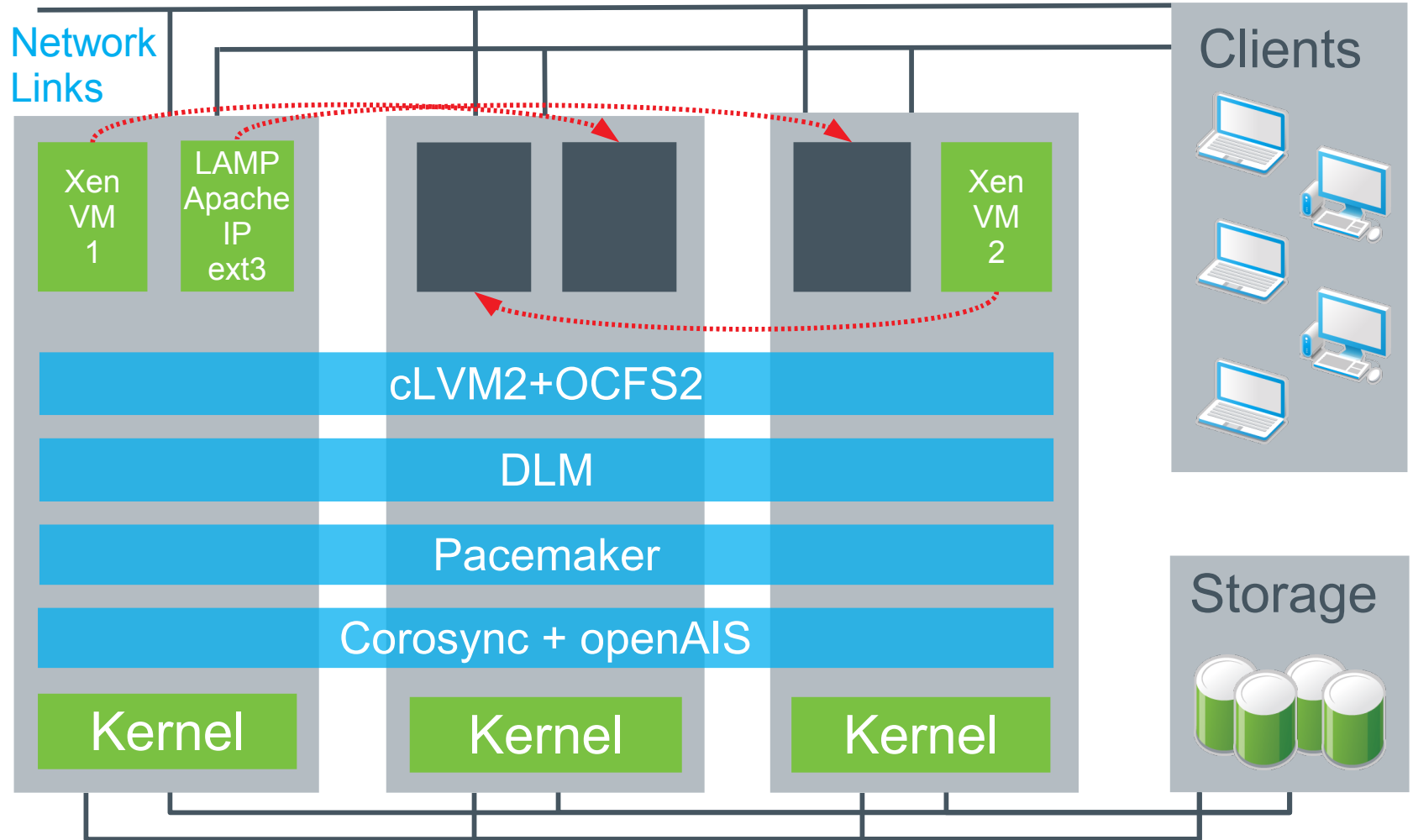# Key Features

- **Service Availability 24/7**
  - Policy driven clustering
    - OpenAIS messaging and membership layer
    - Pacemaker cluster resource manager

- **Sharing and Scaling Data-access by Multiple Nodes**
  - Cluster file system
    - OCFS2
    - Clustered logical volume manager

- **Disaster Tolerance**
  - Data replication via IP
    - Distributed replicated block device

- **Scale Network Services**
  - IP load-balancing

- **User-friendly Tools**
  - Graphical user interface
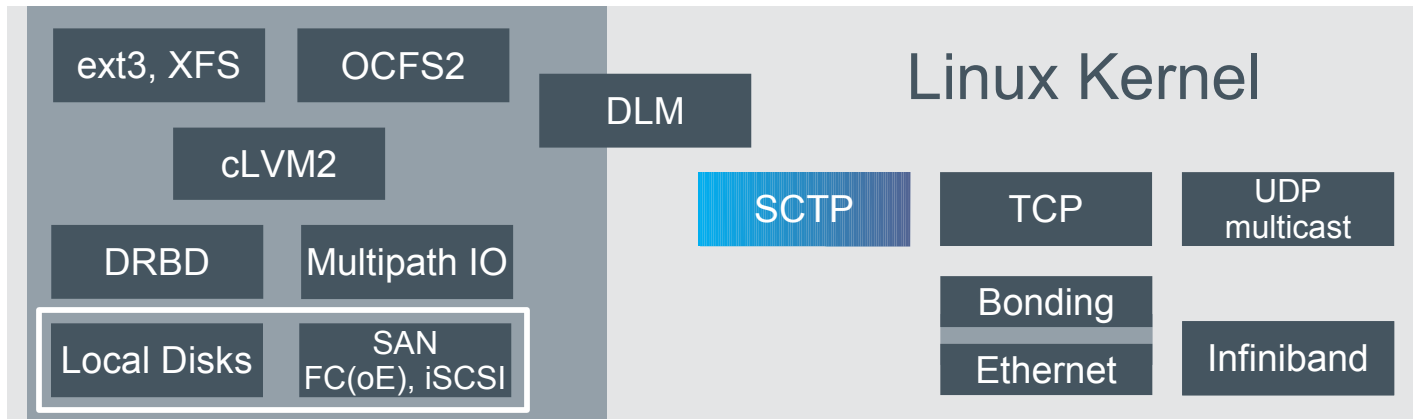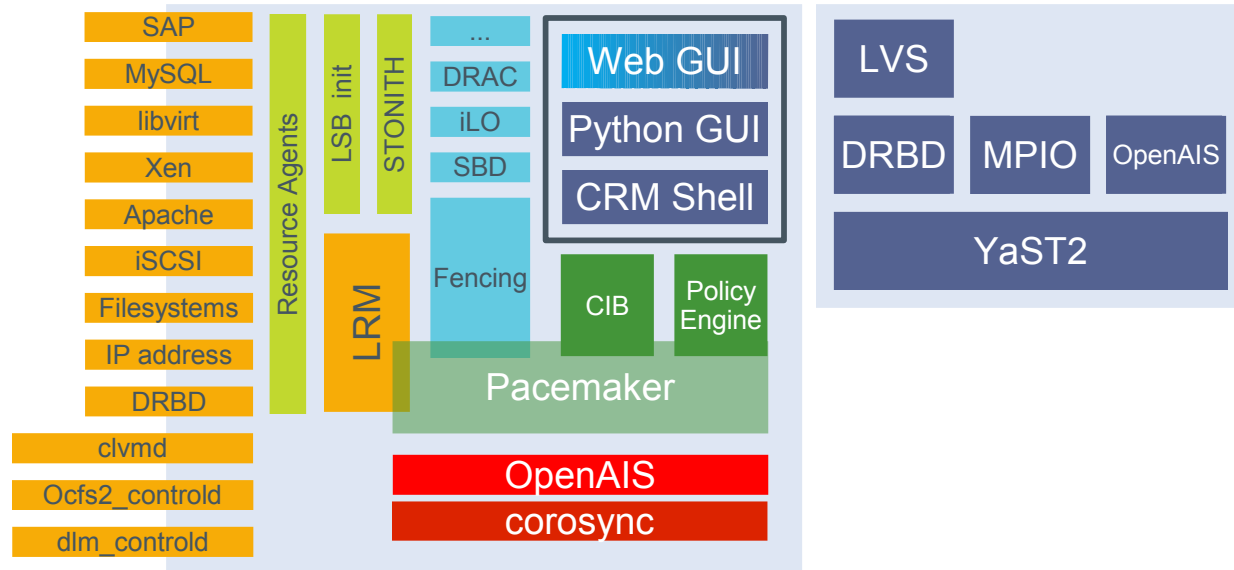 - Unified command line interface

Cluster Architecture

3 Node Cluster Overview

Network Links

Clients

| Xen VM 1 | LAMP Apache IP ext3 | | | | | Xen VM 2 |

cLVM2+OCFS2

DLM

Pacemaker

Corosync + openAIS

| Kernel | Kernel | Kernel |

Storage

Detailed View of Components Per Node:

SAP				...	**Web GUI**		LVS	
MySQL		LSB init	STONITH	DRAC	**Python GUI**			
libvirt	Resource Agents			iLO	**CRM Shell**		DRBD	MPIO	OpenAIS
Xen				SBD				
Apache				Fencing	CIB	Policy Engine	YaST2	
iSCSI		LRM						
Filesystems								
IP address				Pacemaker				
DRBD								
clvmd				OpenAIS				
Ocfs2_controld				corosync				
dlm_controld								

Linux Kernel

ext3, XFS — OCFS2 — DLM

cLVM2

DRBD — Multipath IO

Local Disks — SAN FC(oE), iSCSI

SCTP — TCP — UDP multicast

Bonding

Ethernet — Infiniband

Why Is This Talk Necessary?

- Can't you just make the software stack easy to understand?

- Why is a multi-node setup more complicated than a single node?

- Gosh, this is awfully complicated! Why is this stuff so powerful? I don't need those other features!

This session will address most of these questions.

Design and Architecture Considerations

General Considerations

- Consider the support level requirements of your mission-critical systems.

- Your staff is your key asset!
  - Invest in training, processes, knowledge sharing.
  - A good administrator will provide higher availability than a mediocre cluster setup.

- Get expert help for the initial setup, and

- Write concise operation manuals that make sense at 3am on a Saturday ;-)

- Thoroughly test the cluster at deployment & regularly!
  - Use a staging system before deploying large changes!

Manage Expectations Properly

- Clustering improves reliability, but does not achieve 100%, ever.

- Fail-over clusters reduce service outage, but do not eliminate it.

- HA protects data before the service.

- Clusters are more complex than single nodes.

- Clustering broken applications will not fix them.

- Clusters do not replace backups, RAID, or good hardware.

Complexity Versus Reliability

- Every component has a failure probability.

  - Good complexity: Redundant components.

  - Undesirable complexity: chained components.

  - Choke point → single point of failure

  - Also consider: Administrative complexity.

- Use as few components (features) as feasible.

  - Our extensive feature list is not a mandatory checklist for your deployment ;-)

- What is your fall-back in case the cluster breaks?

  - Backups, non-clustered operation

  - Architect your system so that this is feasible!

Cluster Size Considerations

- More nodes:
  - Increased absolute redundancy and capacity.
  - Decreased relative redundancy.
  - One cluster → one failure and **security** domain.
  - HA is not HPC.
- Does your work-load scale well to larger node counts?
- Choose odd node counts.
  - 4 and 3 node clusters both lose majority after 2 nodes.
- Question:
  - 5 cheaper servers, or
  - 3 higher quality servers with more capacity each?

Common Setup Issues

General Software Stack

- Please avoid chasing already solved problems!

- Please apply all available software updates, as soon as feasible.

- Consider a pre-build distribution

Networking: L2

- Switches must support multicast properly.

    - IGMP snooping some times works, some times it doesn't.

    - Consider isolating cluster traffic on separate VLANs.

- Keep firmware on switches uptodate!

- Bonding is preferable to using multiple rings:

    - Reduces complexity

    - Exposes redundancy to all cluster services and clients

- Firewall rules are not your friend.

Networking

- Make NIC names identical on all nodes

- Local hostname resolution versus DNS

- Address assignment via DHCP

  - Make sure it is static!
- Setup **NTP** for time synchronization.

From One to Many Nodes

- **Error: Configuration files not identical across nodes.**

    – /etc/drbd.conf, /etc/corosync/corosync.conf, /etc/ais/openais.conf, resource-specific configurations ...

- Symptoms: Causes weird misbehavior, works one but not on other systems, interoperability issues, and possibly others.

- Solution: Make sure they are synchronized.

    – "**csync2**" does this automatically for you.

      > You can add your own files to this list as needed.

Fencing (STONITH)

- Error: Not configuring STONITH at all

  – It defaults to enabled, resource start-up will block and the cluster simply do nothing. This is for your own protection.

- Warning: Disabling STONITH

  – DLM/OCFS2 will block forever waiting for a fence that is never going to happen.

- Error: Using "external/ssh", "ssh", "null" in production

  – These plug-ins are for testing. They will not work in production!

  – Use a "real" fencing device or external/sbd

- Error: configuring several power switches in parallel.

- Error: Trying to use external/sbd on DRBD

Virtualization and Clustering

- Hypervisor cluster managing VMs or

- Clustering VMs internally?

- Hide bonding at the hypervisor layer

- Cluster will happily live migrate a crashed kernel ...

- You can mix physical and virtual nodes in a cluster
  - But don't host the virtual nodes *on* the physical ones!

CIB Configuration Issues: Placement

- Even though I ordered them, they are starting on different nodes!

    – Use (anti-)collocation

    > You can specify who wins using "priority"

- Resources are starting up on "wrong" nodes

- My location constraints are too complex!

    – Use "utilization"

- Resources move around when something "unrelated" changes

    – # crm configure property default-resource-stickiness=100

CIB Configuration Issues: Quorum

- 2 node clusters cannot have majority with 1 node failed

    – # crm configure property no-quorum-policy=ignore

- no-quorum-policy="stop"

    – Can make OCFS2/DLM misbehave

- no-quorum-policy="freeze"

    – Safe default for 2+ nodes

CIB Configuration Issues: Misc

- Resources are starting up in "random" order, even though I collocated them!

  – Order != Placement

- Clones most frequently need to have interleave="true"

- Use simple constructs where possible!

- # crm_verify -L ; ptest -L -VVVV

  – Will point out some basic issues

- CIB is easily the most important bit of your configuration, and the most easily misunderstood ...

We'll get back to that ...

Configuring Cluster Resources

- Symptom: On start of one or more nodes, the cluster restarts resources!

- Cause: resources under cluster control are also started via the "init" sequence.

  – The cluster "probes" all resources on start-up on a node, and when it finds resources active where they should not be – possibly even more than once in the cluster –, the recovery protocol is to stop them all (including all dependencies) and start them cleanly again.

- Solution: Remove them from the "init" sequence.

Setting Resource Time-outs

- **Belief**: **"Shorter time-outs make the cluster respond faster."**

- **Fact:** Too short time-outs cause resource operations to "fail" erroneously, making the cluster unstable and unpredictable.

  - A somewhat too long time-out will cause a fail-over delay;

  - a slightly too short time-out will cause an **unnecessary service outage.**

- Consider that a loaded cluster node may be slower than during deployment testing.

  - Check "crm_mon -t1" output for the actual run-times of resources.

Storage in General

- Activating non-battery backed caches for performance

  - Causes **data corruption**.

- iSCSI over unreliable networks.

- Lack of multipath for storage.

- Believing that RAID replaces backups.

  - RAID and replication immediately propagate logical errors!

- Please ensure that device names are identical on all nodes.

OCFS2

- Using ocfs2-tools-o2cb (legacy mode)
  - O2CB only works with Oracle RAC; full features of OCFS2 are only available in
    combination with Pacemaker
  - *# zypper rm ocfs2-tools-o2cb*
  - Forget about /etc/ocfs2/cluster.conf, /etc/init.d/ocfs2, /etc/init.d/o2cb and /etc/sysconfig/ocfs2
- Nodes crash on shutdown
  - If you have active ocfs2 mounts, you need to umount before shutdown
  - If openais is part of the boot sequence
    - > *# insserv openais*
- Consider: Do you really need OCFS2?
  - Can your application really run concurrently?

Host-based RAID

- Active/Active:
  - Possible with cLVM2 and cmirrord
  - Disk failures are costly; consider MPIO
  - Consider centralizing your storage infrastructure!
- Active/Passive:
  - Works fine with md / Raid1 resource agent
  - Use recent versions to get reassembly on switch-over and monitor
  - Preferable if not requiring active/active

Distributed Replicated Block Device

- Myth: has no shared state, thus no STONITH needed.
  - **Fact:** DRBD still needs fencing!
- Active/Active:
  - Does not magically make ext3 or applications concurrency-safe, still can only be mounted once
  - With OCFS2, split-brain is still fatal, as data diverges!
- Active/Passive:
  - Ensures only one side can modify data, added protection.
  - Does not magically make applications crash-safe.
- Issue: Replication traffic during reads.
  - "noatime" mount option.

Exploring the Effect of Events

What Are Events?

- All state changes to the cluster are events

  – They cause an update of the CIB

  – Configuration changes by the administrator

  – Nodes going up or going down

  – Resource monitoring failures

- Response to events is configured using the CIB policies and computed by the Policy Engine

- This can be simulated using ptest

  – Available comfortably through the "crm" shell

Simulating Node Failure

```
hex-0:~ # crm

crm(live)# cib new sandbox

INFO: sandbox shadow CIB created

crm(sandbox)# cib cibstatus node hex-0
unclean

crm(sandbox)# ptest
```

Simulating Node Failure
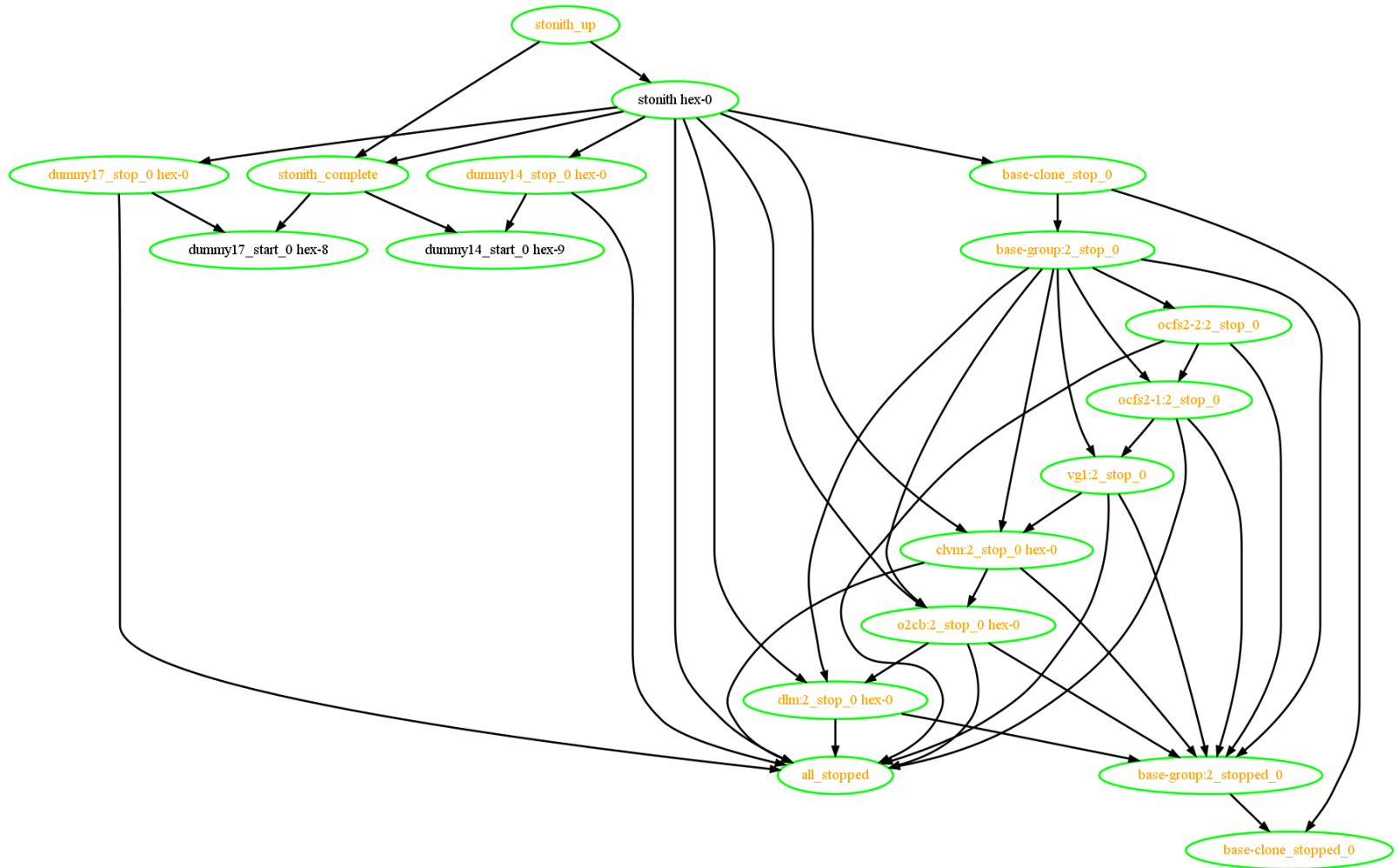
Simulating Resource Failure

crm(sandbox)# *cib cibstatus load live*

crm(sandbox)# *cib cibstatus op*

usage: op <operation> <resource> <exit_code> [<op_status>] [<node>]

crm(sandbox)# *cib cibstatus op start dummy1 not_running done hex-0*

crm(sandbox)# *cib cibstatus op start dummy1 unknown timeout hex-0*

crm(sandbox)# *configure ptest*

ptest[4918]: 2010/02/17_12:44:17 WARN: unpack_rsc_op: Processing failed op dummy1_start_0 on hex-0: unknown error (1)

Simulating Resource Failure
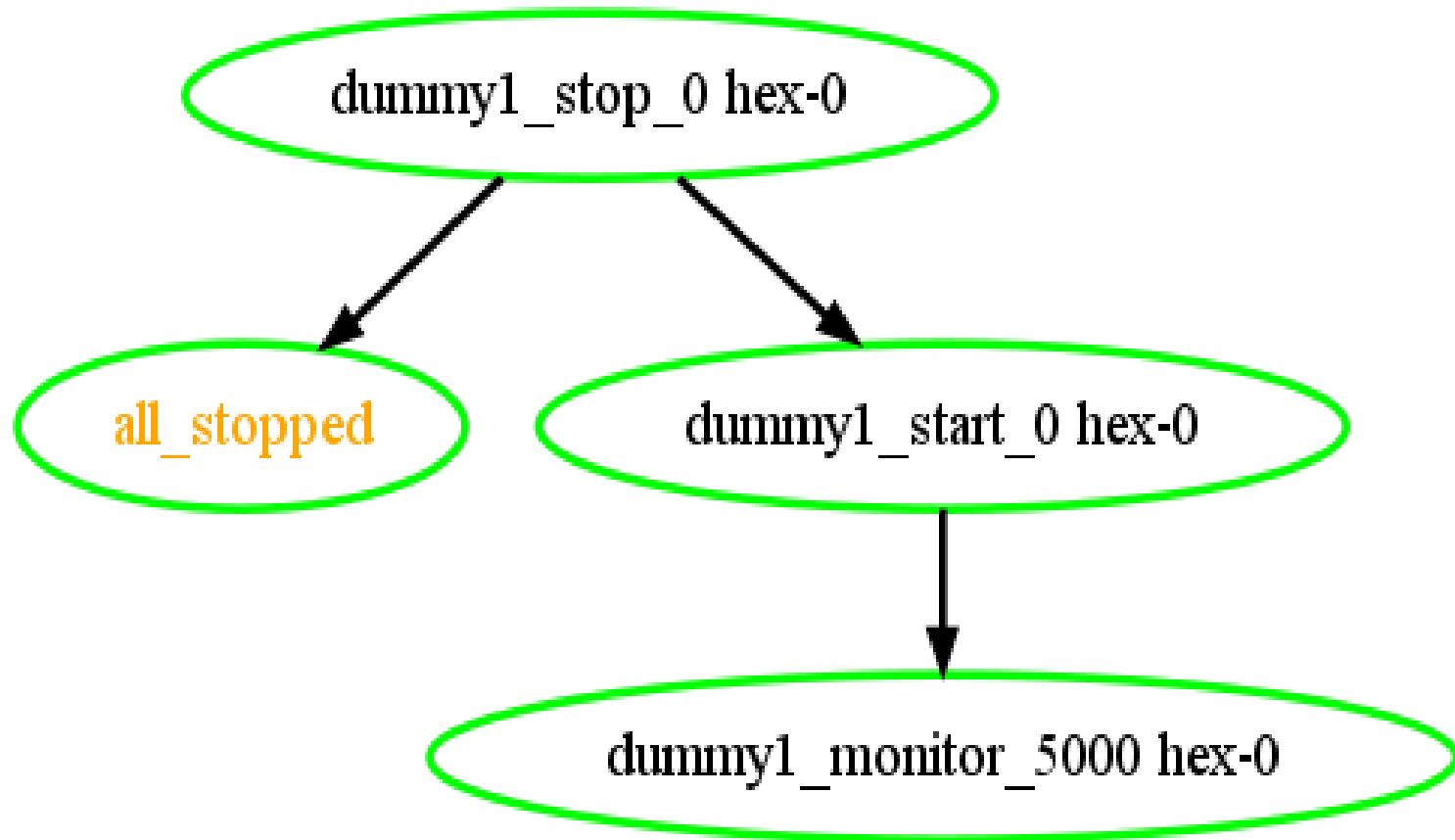
Exploring Configuration Changes

crm(sandbox)# *cib cibstatus load live*

crm(sandbox)# *configure primitive dummy42 ocf:heartbeat:Dummy*

crm(sandbox)# *ptest*

Configuration changes - Woah!
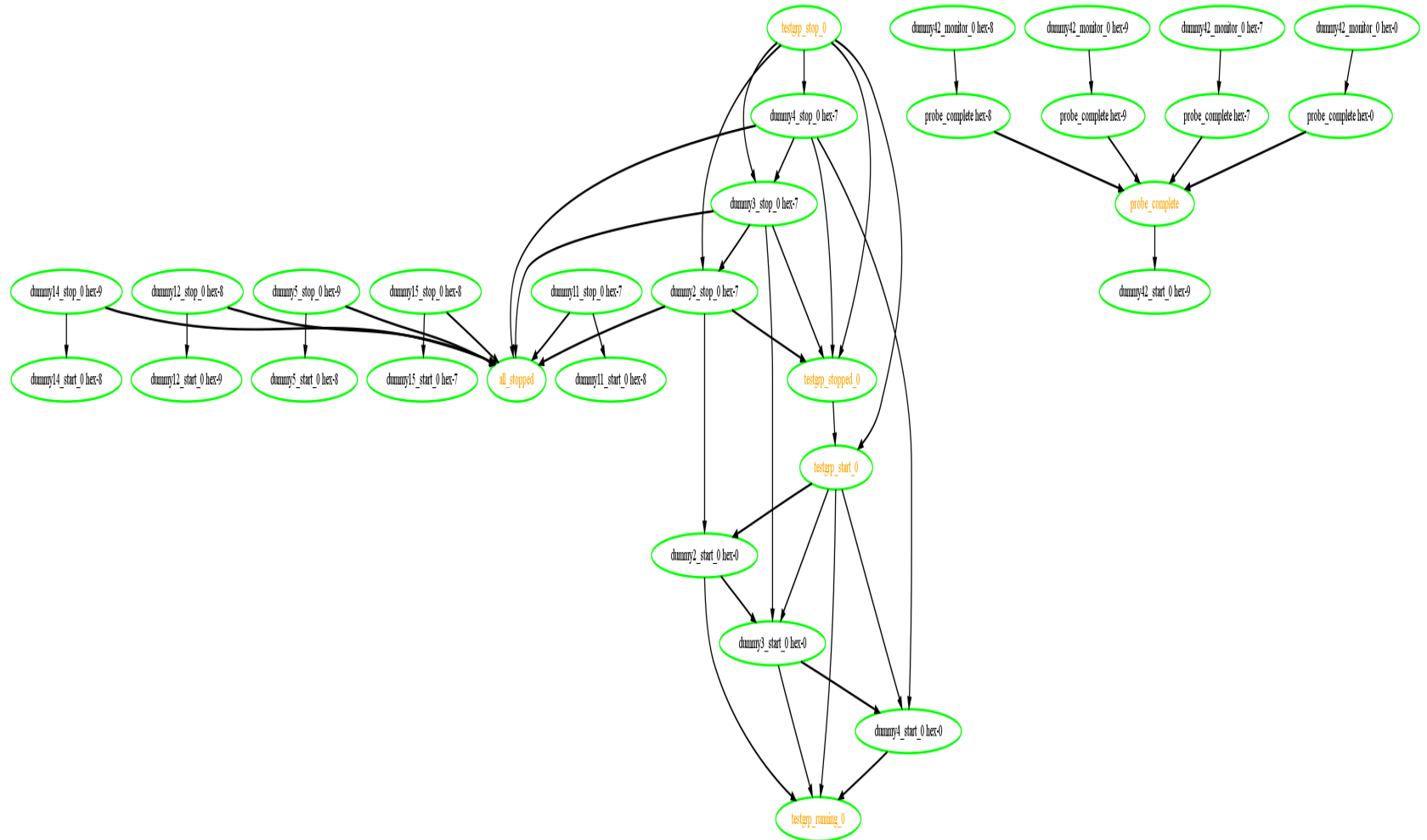
Exploring Configuration Changes

crm(sandbox)# *configure rsc_defaults resource-stickiness=1000*

crm(sandbox)# *ptest*

crm(sandbox)# *configure colocation order-42 inf: dummy42 dummy1*

crm(sandbox)# *ptest*

Configuration changes – Almost ...
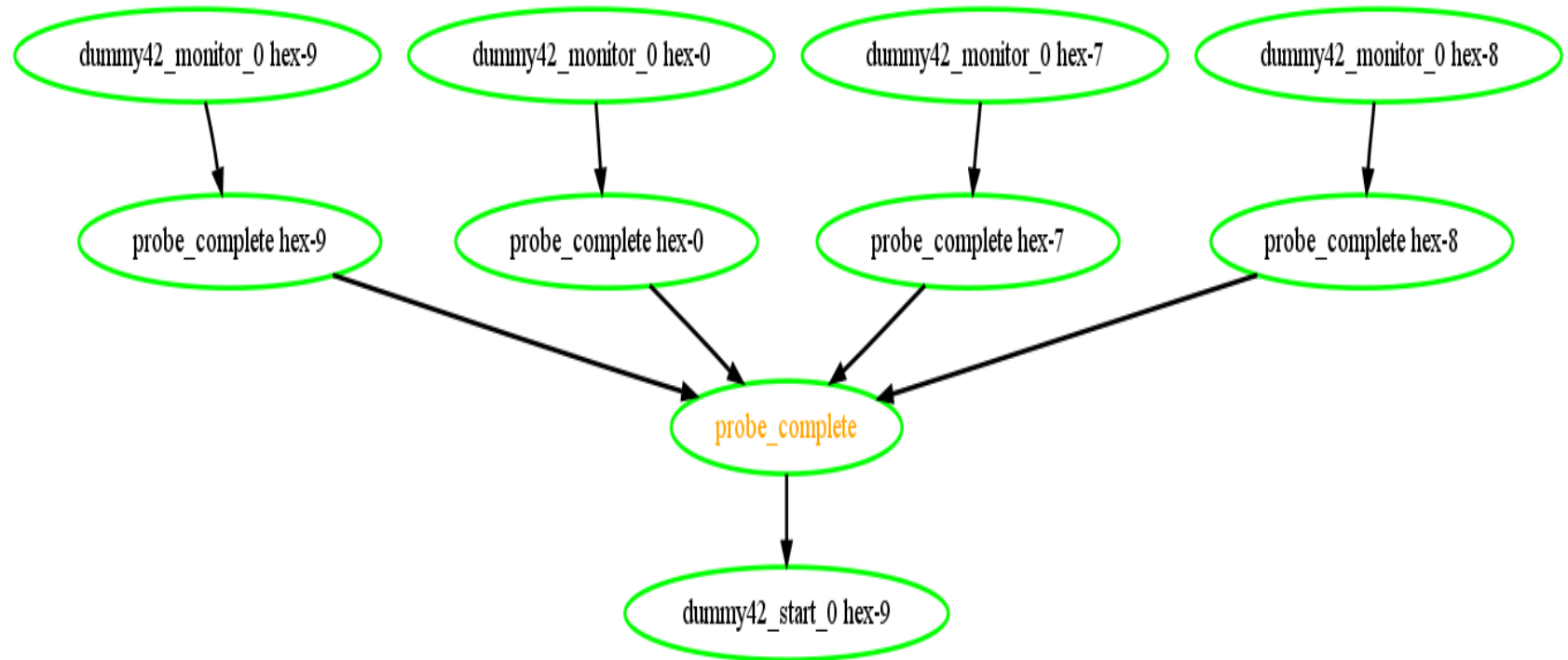
Configuration changes - done
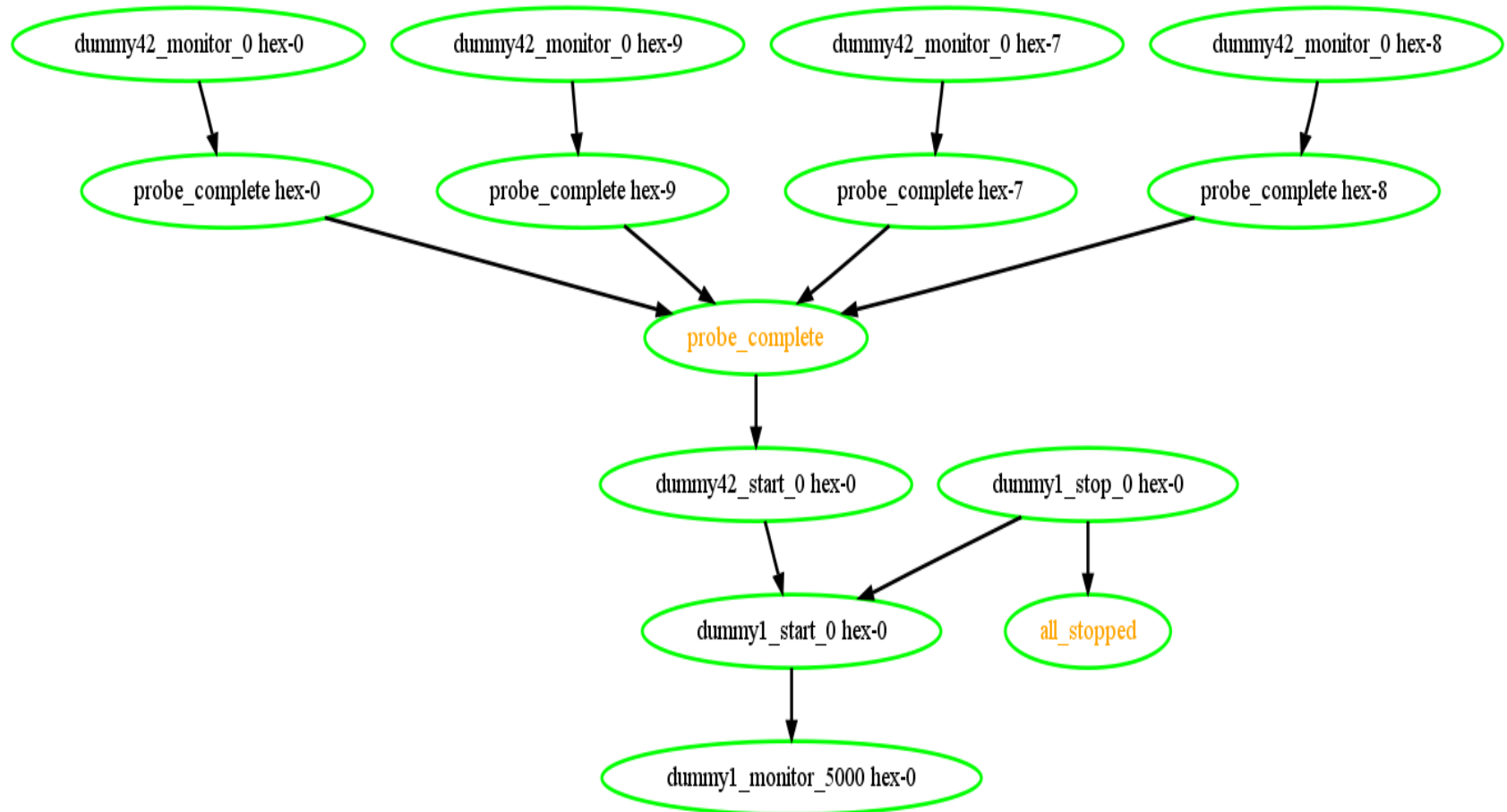
Log Files and Their Meaning

hb_report Is the Silver Support Bullet

- Compiles

  - Cluster-wide log files,

  - Package state,

  - DLM/OCFS2 state,

  - System information,

  - CIB history,

  - parses core dump reports (install debuginfo packages!)

  - into a single tarball for all support needs.

- *# hb_report -n "node1 node2 node3" -f 12:00 /tmp/hb_report_example1*

Logging

- "The cluster generates too many log messages!"

  - Alas, customers are even more upset when asked to reproduce a problem on their production system ;-)

  - Incidentially, all command line invocations are logged.

- System-wide logs: /var/log/messages

- CIB history: /var/lib/pengine/*

  - All cluster events are logged here and can be analyzed with hindsight (python GUI, ptest, and the crm shell).

Where Is the Real Cause?

The answer is always in the logs.

Even though the logs on the DC may print a reference to the error, the real cause may be on another node.

Most errors are caused by resource agent misconfiguration.

Correlating Messages to Their Cause

- **Feb 17 13:06:57 hex-8 pengine: [7717]: WARN: unpack_rsc_op: Processing failed op ocfs2-1:2_monitor_20000 on hex-0: not running (7)**

  - This is **not** the failure, just the Policy Engine reporting on the CIB state! The real messages are on hex-0, grep for the operation key:

- **Feb 17 13:06:57 hex-0 Filesystem[24825]: [24861]: INFO: /filer is unmounted (stopped)**

- **Feb 17 13:06:57 hex-0 crmd: [7334]: info: process_lrm_event: LRM operation ocfs 2-1:2_monitor_20000 (call=37, rc=7, cib-update=55, confirmed=false) not running**

  - Look for the error messages from the resource agent before the lrmd/pengine lines!

Debugging Resource Agents

Common Resource Agent Issues

- Operations must succeed if the resource is already in the requested state.

- "monitor" must distinguish between at least "running/OK", "running/failed", and "stopped"

    - Probes deserve special attention

- Meta-data must conform to DTD.

- 3rd party resource agents do *not* belong under /usr/lib/ocf/resource.d/heartbeat – chose your own provider name!

- Use ocf-tester to validate your resource agent.

ocf-tester Example Output

```
hex-0:~ # ocf-tester -n Example
/usr/lib/ocf/resource.d/bs2010/Dummy

Beginning tests for /usr/lib/ocf/resource.d/bs2010/Dummy...

* Your agent does not support the notify action (optional)

* Your agent does not support the demote action (optional)

* Your agent does not support the promote action (optional)

* Your agent does not support master/slave (optional)

* rc=7: Stopping a stopped resource is required to succeed

Tests failed: /usr/lib/ocf/resource.d/bs2010/Dummy failed 1
tests
```

01101

Questions and Answers

Community resources

- Documentation is excellent:

  – http://www.novell.com/documentation/sle_ha/

  – http://www.clusterlabs.org/wiki/Documentation

    > Full reference documentation, step-by-step guides, pointers to more documentation

- Very active community & discussions:

  – http://www.clusterlabs.org/wiki/Mailing_lists

- Bug reports:

  – http://developerbugs.linux-foundation.org/

    > Remember to include hb_report!