High Availability on Linux

SUSE way

Roger Zhou
zzhou@suse.com

openSUSE.Asia Summit 2015

CURIOSITY in the land of Linux High Availability

Agenda

- HA architectural components
- Use case examples
- Future outlook & Demo

What is Cluster?

- HPC (super computing)
- Load Balancer (Very high capacity)
- High Availability
  - 99.999% = 5 m/year MTTR
  - SPOF(single point of failure)
  - Murphy's Law

    "Everything that can go wrong will go wrong"

"HA", widely used, often confusing

- VMWare vSphere HA
  - hypervisor and hardware level. Close-source.
  - Agnostic on Guest OS inside the VM.
- SUSE HA
  - Inside Linux OS.
  - That said, Windows need Windows HA solution.
- Different Industries
  - We are Enterprise.
  - HADOOP HA (dot com)
  - OpenStack HA (paas)
  - OpenSAF (telecom)

History of HA in Linux OS

- 1990s, Heartbeat project. Simply two nodes.
- Early 2000s, Heartbeat 2.0 too complex.
  - Industry demands to split.

    1) one for cluster membership

    2) one for resource management
- Today, ClusterLabs.org
  - A completely different solution in early days, pacemaker + corosnc
  - While merged Heartbeat project.
  - 2015 HA Summit

Typical HA Problem - Split Brain

- Clustering
  - multiple nodes share the same resources.
- Split partitions run the same service
  - It just breaks data integrity !!!
- Two key concepts as the solution:
  - Fencing

    Cluster doesn't accept any confusing state.

    STONITH - "shoot the other node in the head".
  - Quorum

    It stands for "majority". No quorum, then no actions, no resource management, no fencing.

HA Hardware Components

- Multiple networks
  - A user network for end user access.
  - A dedicated network for cluster communication/heartbeat.
  - A dedicated storage network infrastructure.
- Network Bonding
  - aka. Link Aggregation
- Fencing/STONITH devices
  - remote “powerswitch”
- Shared storage
  - NAS(nfs/cifs), SAN(fc/iscsi)

Architectural Software Components

"clusterlabs.org"

- Corosync
- Pacemaker
- Resource Agents
- Fencing/STONITH Devices
- UI(crmsh and Hawk2)
- Booth for GEO Cluster

Outside of "clusterlabs.org"

- LVS: Layer 4, ip+port, kernel space.
- HAproxy: Layer 7/ HTTP, user space.
- Shared filesystem: OCFS2 / GFS2
- Block device replication:
  DRBD, cLVM mirroring, cluster-md
- Shared storage:
  SAN (FC / FCoE / iSCSI)
- Multipathing

Software Components in details

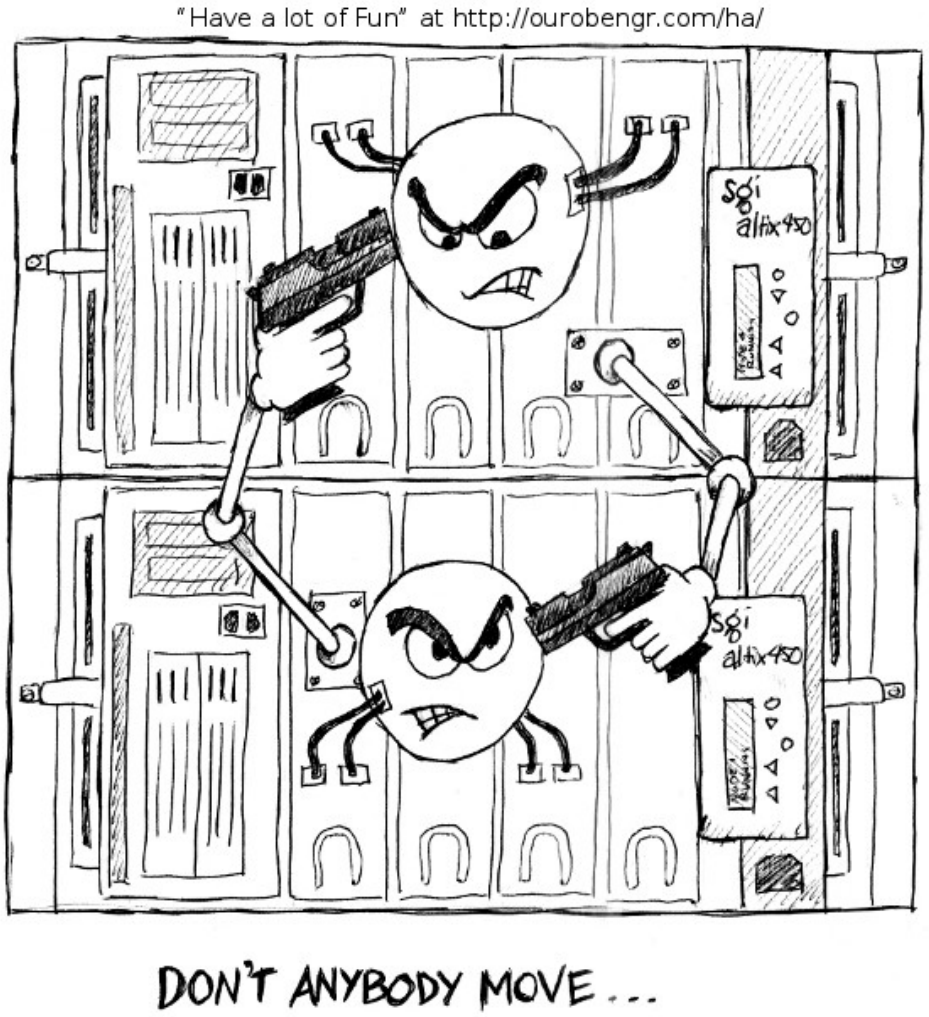
Corosync: messaging and membership

- Consensus algorithm
  - "Totem Single Ring Ordering and Membership protocol"
- Closed Process Group
  - Analogue “TCP/IP 3-way hand shaking”
    - Membership handling.
    - Message ordering.
- A quorum system
  - notifies apps when quorum is achieved or lost.
- In-memory object database
  - for Configuration engines and Service engines.
  - Shared-nothing cluster.

Pacemaker: the resources manager

- The brain of the cluster.
- Policy engine for decision making.
  - To start/stop resources on a node according to the score.
  - To monitor resources according to interval.
  - To restart resources if monitor fails.
  - To fence/STONITH a node if stop operation fails.

Shoot The Other Node In The Head



- Data integrity does not tolerate any confusing state. Before migrating resources to another node in the cluster, the cluster must confirm the suspicious node really is down.
- STONITH is mandatory for *enterprise* Linux HA clusters.

Popular STONITH devices

- APC PDU
  - network based powerswitch
- Standard Protocols Integrated with Servers
  - Intel AMT, HP iLO, Dell DRAC, IBM IMM, IPMI Alliance
- Software libraries
  - to deal with KVM, Xen and VMware Vms.
- Software based
  - SBD (STONITH Block Device) to do self termination.

    The last implicit option in the fencing topology.
- NOTE: Fencing devices can be chained.

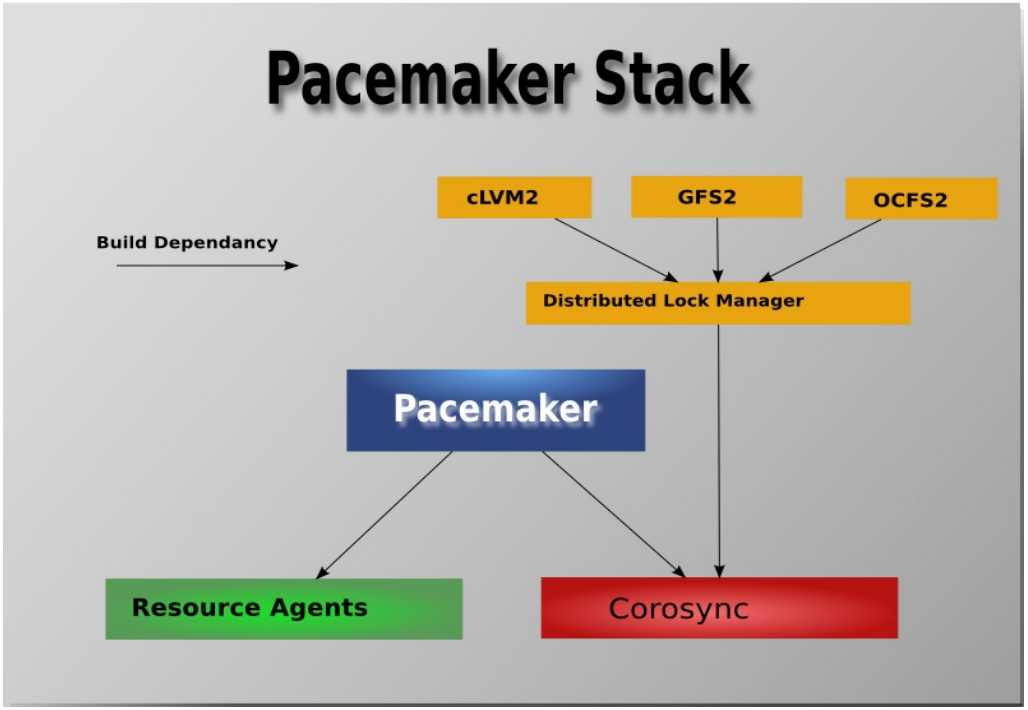
Resources Agents (RAs)

- Write RA for your applications
- LSB shell scripts:

  start / stop / monitor
- More than hundred contributors in upstream github.

Cluster Filesystem

- OCFS2 / GFS2
  - On the shared storage.
  - Multiple nodes concurrently access the same filesystem.



http://clusterlabs.org/doc/fr/Pacemaker/1.1/html/Clusters_from_Scratch/_pacemaker_architecture.html

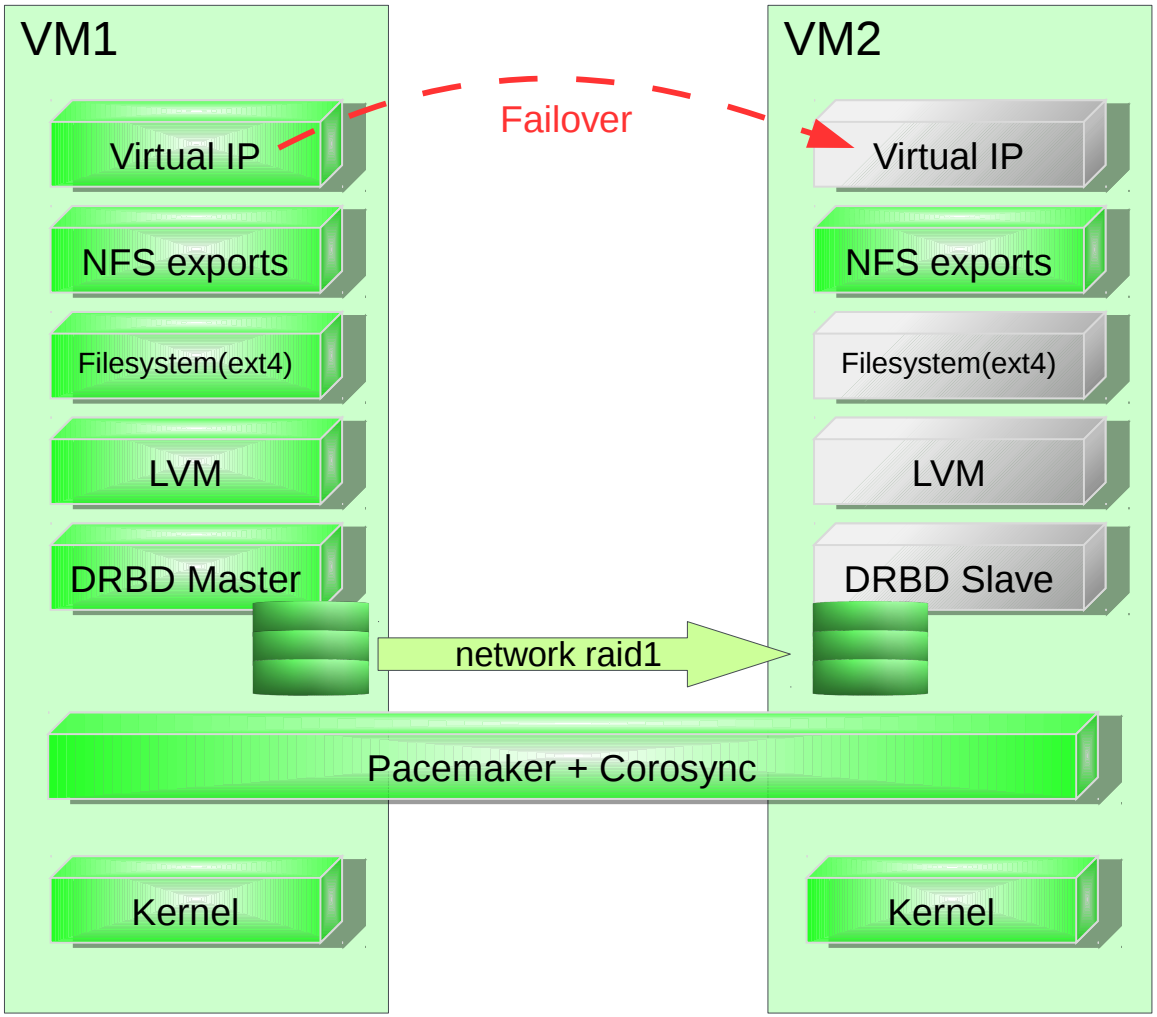
Cluster Block Device

- DRBD
  - network based raid1.
  - high performance data replication over network.
- cLVM2 + cmirrord
  - Clustered lvm2 mirroring.
  - Multiple nodes can manipulate volumes on the shared disk.
  - clvmd distributes LVM metadata updates in the cluster.
  - Data replication speed is way too slow.
- Cluster md raid1
  - multiple nodes use the shared disks as md-raid1.
  - High performance raid1 solution in cluster.

Cluster Examples

NFS Server (High Available NAS)

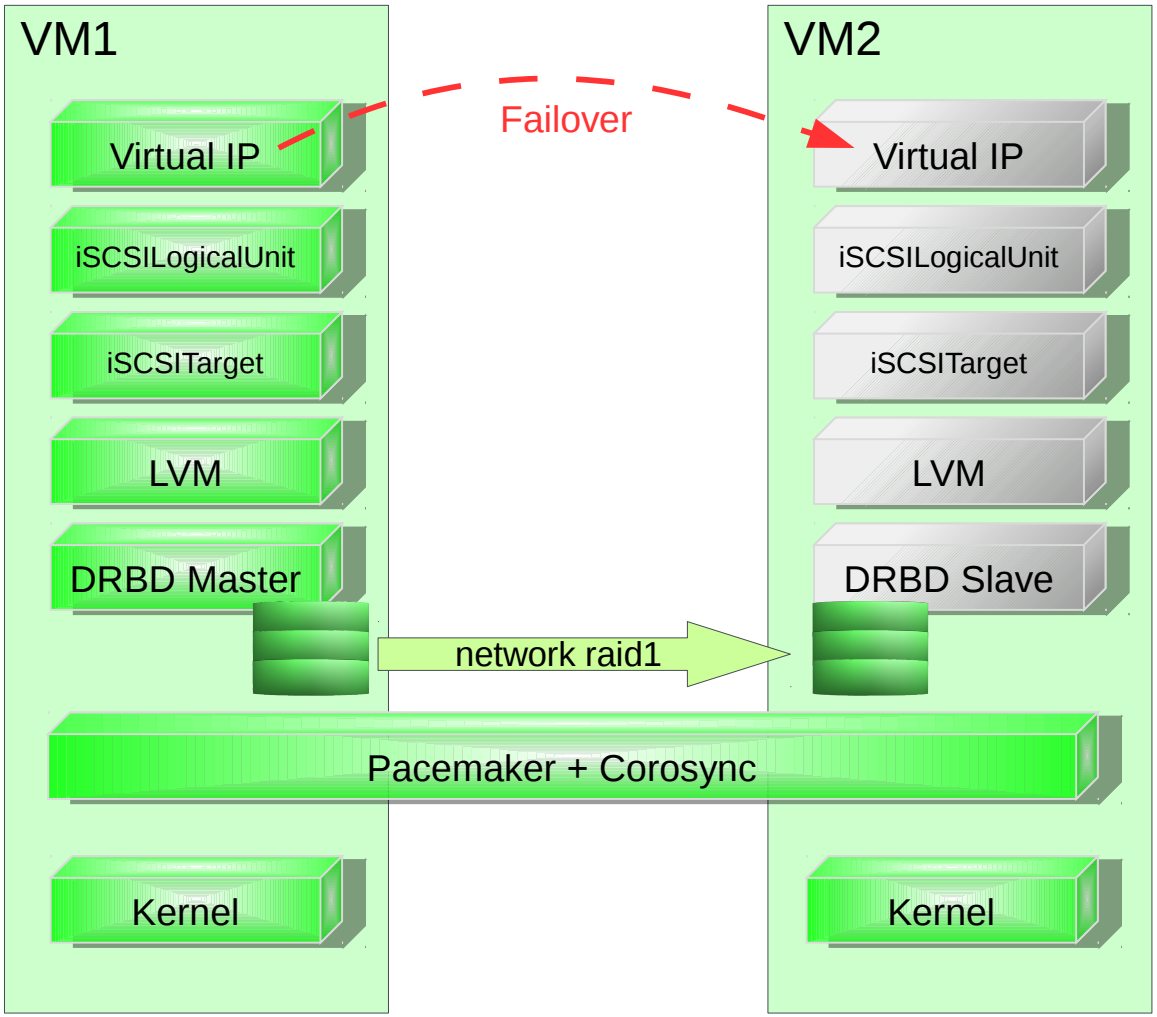
Cluster Example in Diagram



Leap

HA iSCSI Server (Active/Passive)

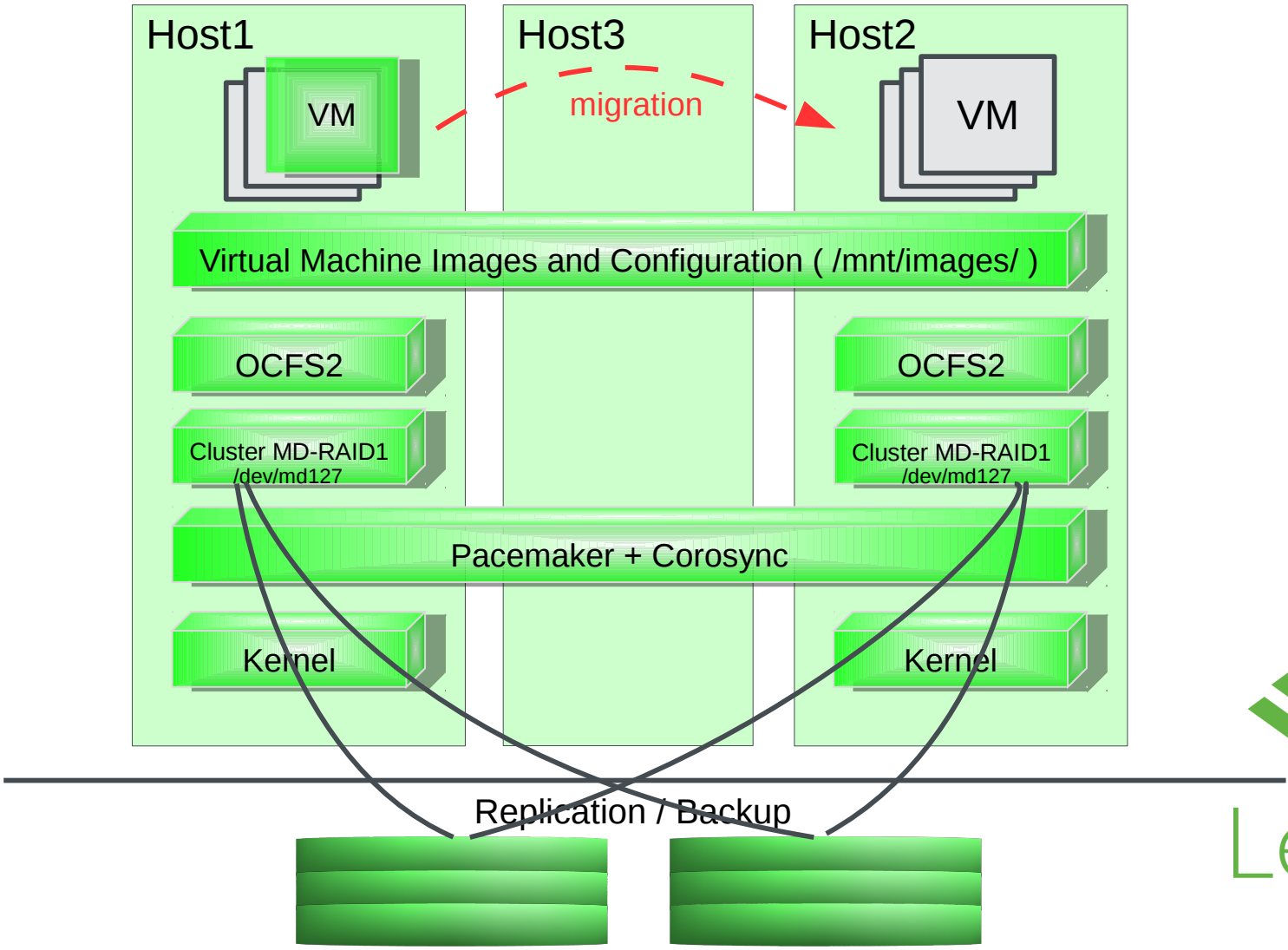
Cluster Example in Diagram



Leap

Cluster FS - OCFS2 on shared disk

Cluster Example in Diagram



Leap

Future outlook

- Upstream activities
  - OpenStack: from the control plane into the compute domain.
  - Scalability of corosync/pacemaker
  - Docker adoption
  - “Zero” Downtime HA VM
  - ...

Leap

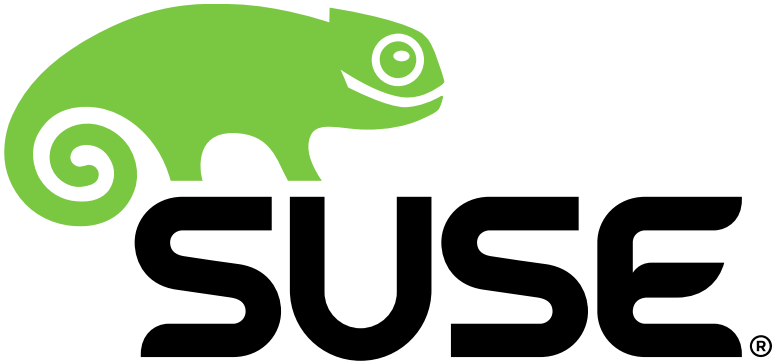
Join us (all *open-source*)

- Play with Leap 42.1: http://www.opensuse.org

  Doc: https://www.suse.com/documentation/sle-ha-12/
- Report and Fix Bugs: http://bugzilla.opensuse.org
- Discussion: opensuse-ha@opensuse.org

- HA ClusterLabs: http://clusterlabs.org/
- General HA Users: users@clusterlabs.org

Demo + Q&A + Have fun


SUSE®
We adapt. You succeed.