SUSE
We adapt. You succeed.

High Availability Storage

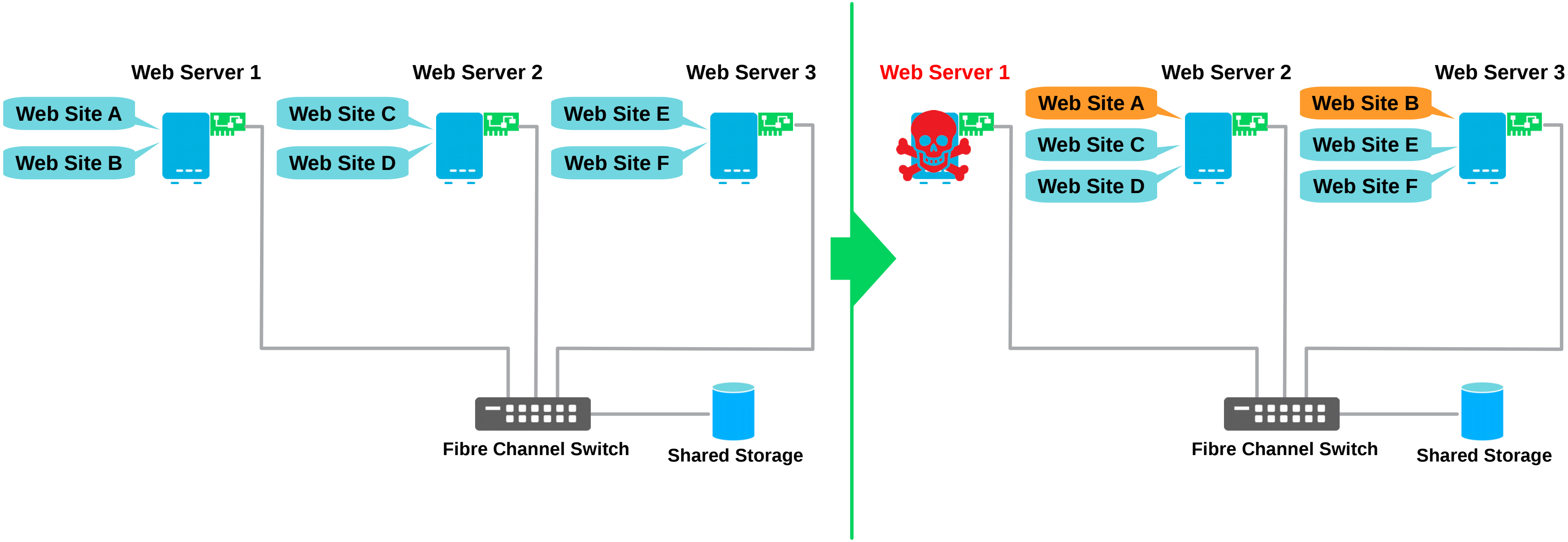
SUSE's Building Blocks

Roger Zhou
Sr. Eng Manager
zzhou@suse.com

Guoqing Jiang
HA Specialist
gqjiang@suse.com

SLE-HA Introduction

The SUSE Linux Enterprise High Availability Extension is powered by Pacemaker & Corosync to eliminate SPOF for critical data, applications, and services, to implement HA clusters.



Storage

Software Elements

- Block Device: hdX, sdX, vdX => vgX, dmX, mdX
- Filesystem: ext3/4, xfs, btrfs => ocfs2, gfs2

Enterprise Requirements (must-have)

- High Available (Active-Passive, Active-Active)
- Data Protection (Data Replication)

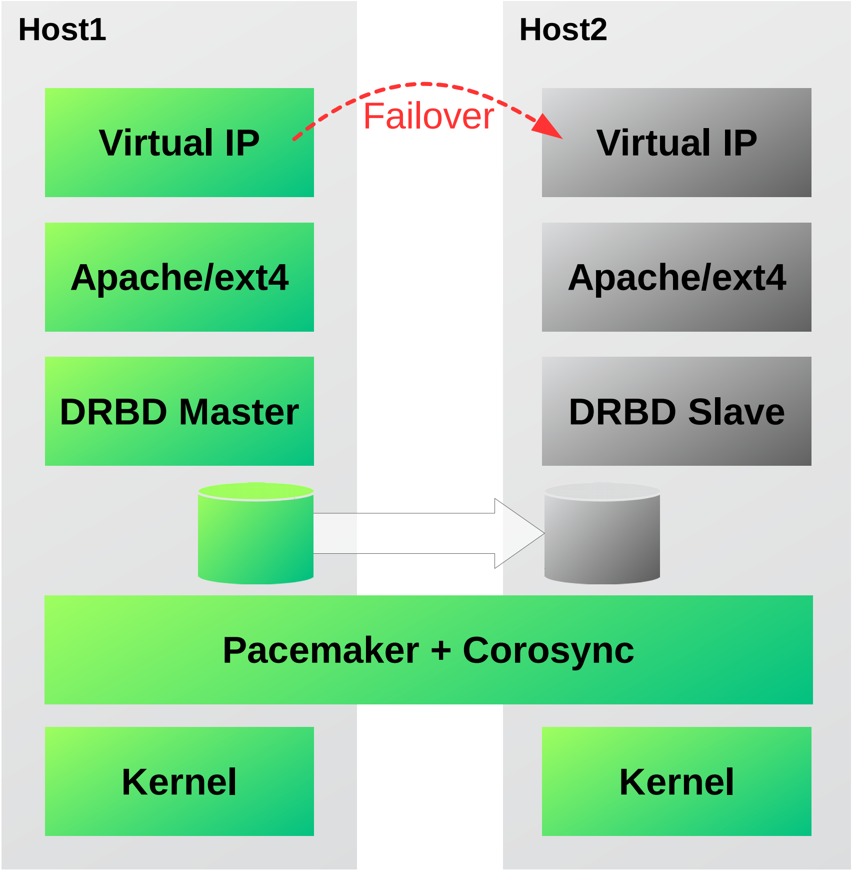
Small Business

High Availability Block Device

- active/passive

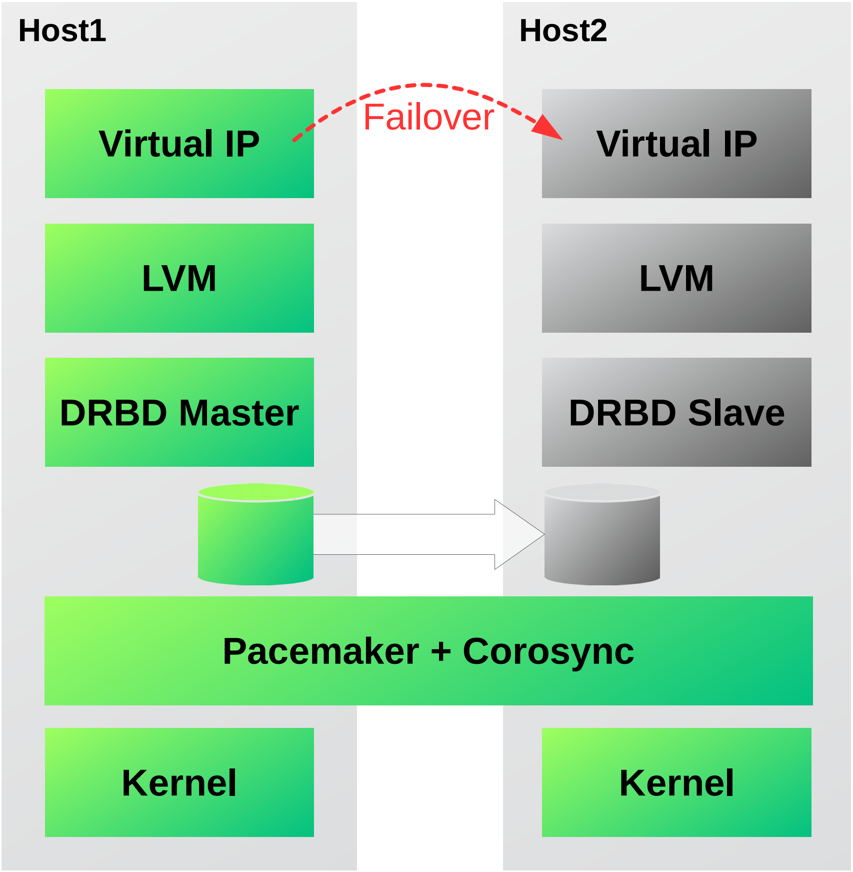
High Availability – DRBD

- DRBD – Data Replication Block Device
- A special master/slave resources is managed by pacemaker, corosync software stack
- SLE HA stack manages the service ordering, dependency, and failover
- Active-Passive approach



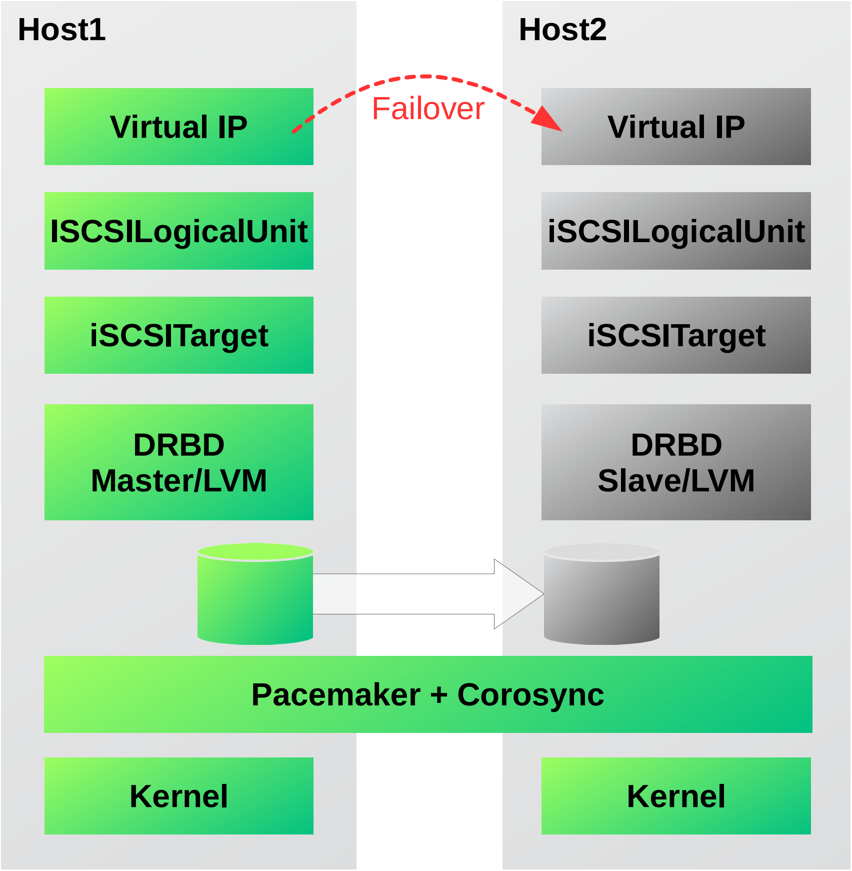
High Availability LVM

- HA-LVM resources managed by pacemaker, corosync software stack
- lvmconf --enable-halvm
- Active-Passive approach



High Availability iSCSI Server

- iSCSI provides the block devices over TCP/IP
- Active-Passive approach



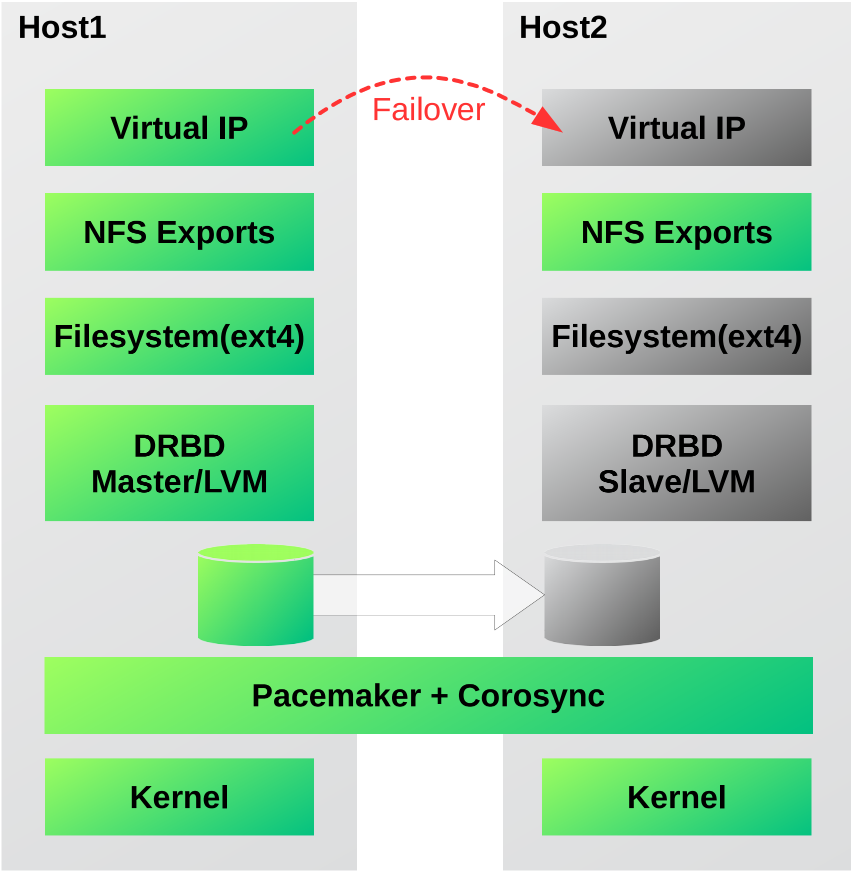
High Availability Filesystem

High Availability NFS (HA-NAS)

- NFS server on top of ext3/ext4

- Same applies to:
  - xfs
  - cifs samba
  - etc.

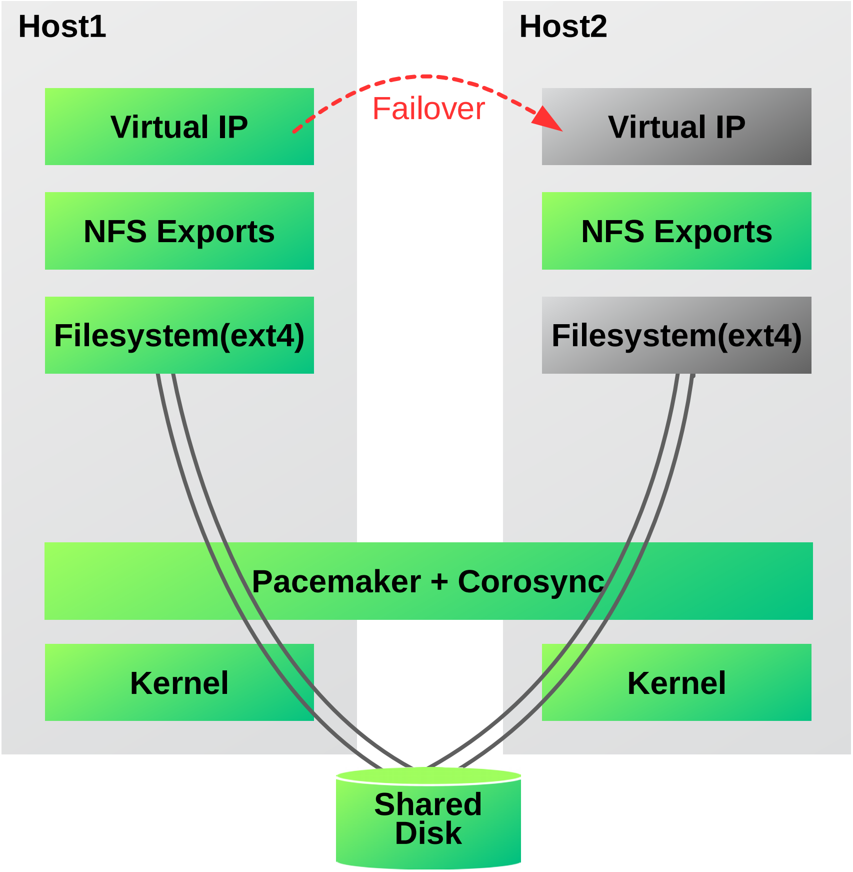
- Active-Passive approach



High Availability NFS (HA-NAS) with Shared Disk

- Active-Passive approach

NOTE: multipathing is “must-have” for Enterprise



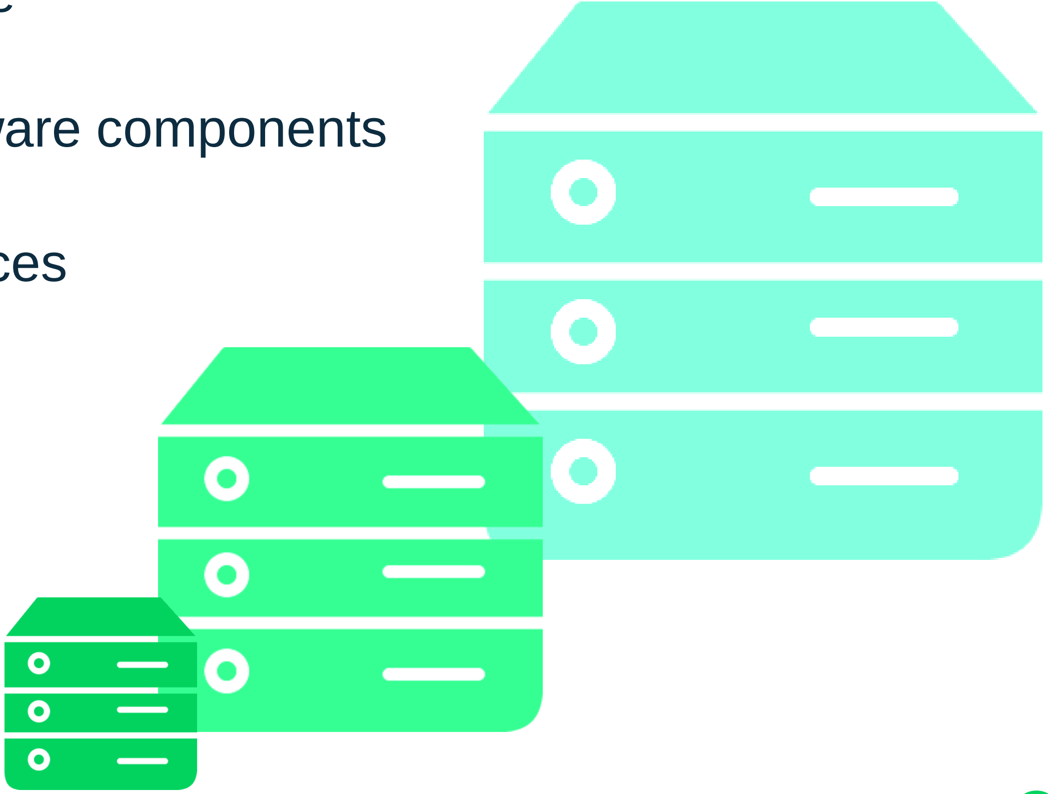
Motivation for Your Storage Growth

Scalability —► extendable storage

Easy Management —► cluster aware components

Cost Effective —► shared resources

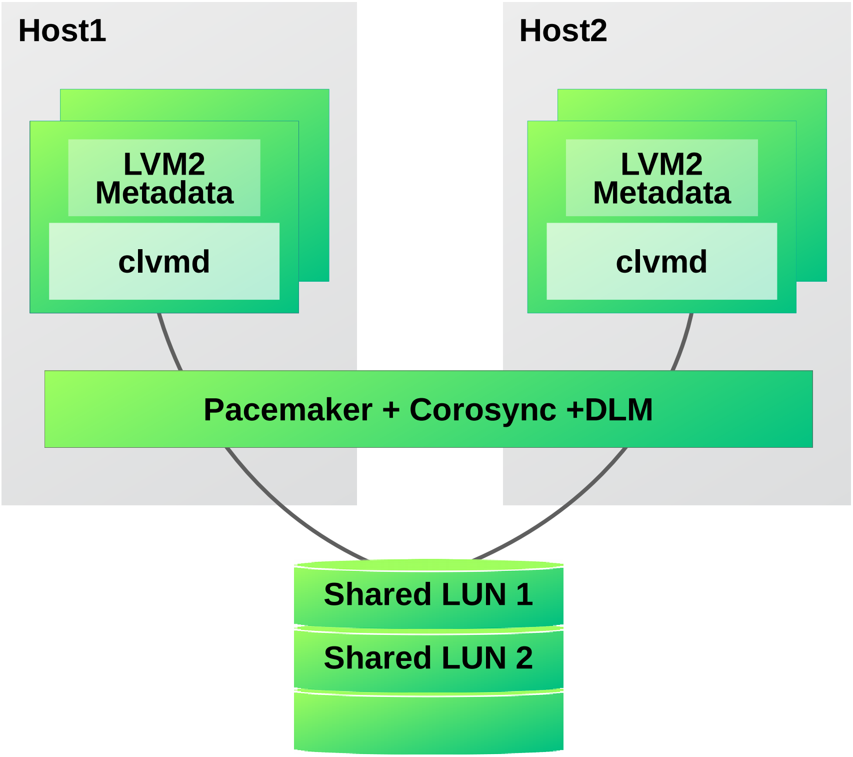
Performance —► active – active



Concurrent Sharing – you need a lock

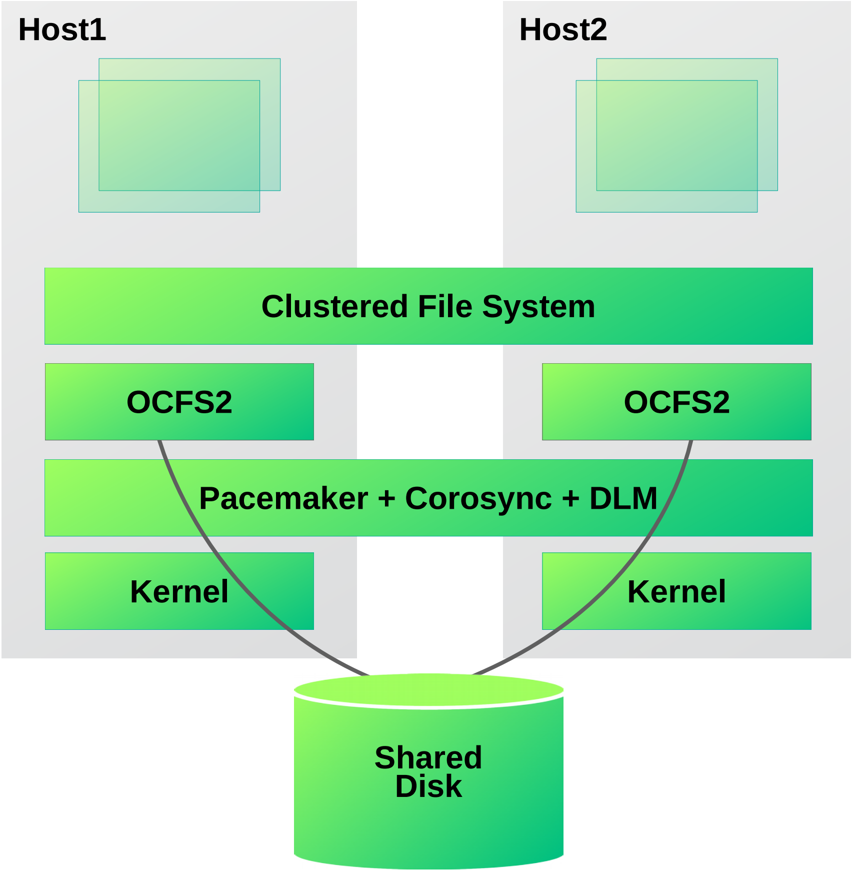
Clustered LVM2 (cLVM2)

- cLVM2 enables multiple nodes to use LVM2 on the shared disk
- cLVM2 coordinates LVM2 metadata, does not coordinate access to the shared data
- That said, multiple nodes access data of different dedicated VG are safe
- lvmconf --enable-cluster
- Active-Active approach



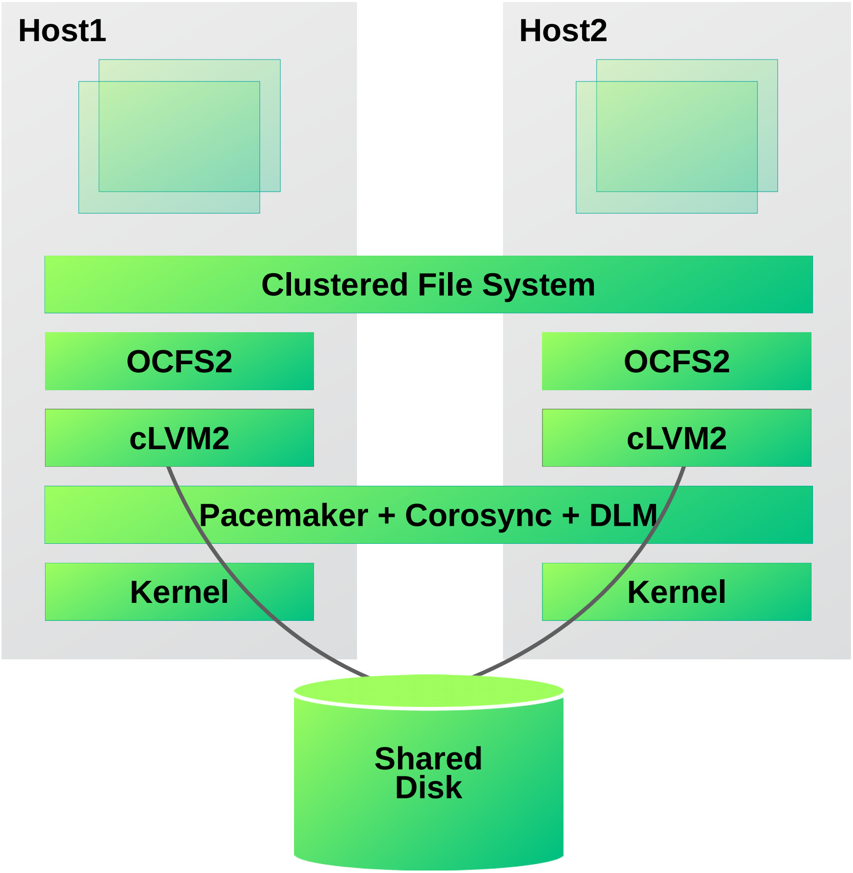
Clustered FS – OCFS2 on shared disk

- OCFS2 resources is managed by using pacemaker, corosync, dlm software stack
- Multiple host can modify the same file with performance penalty
- DLM lock protects the data access
- Active-Active approach



OCFS2 + cLVM2 (both volumes and data are safe)

- cLVM provides clustered VG. Metadata is protected
- OCFS2 protect the data access inside the volume
- Safe to enlarge the filesystem size
- Active-Active approach

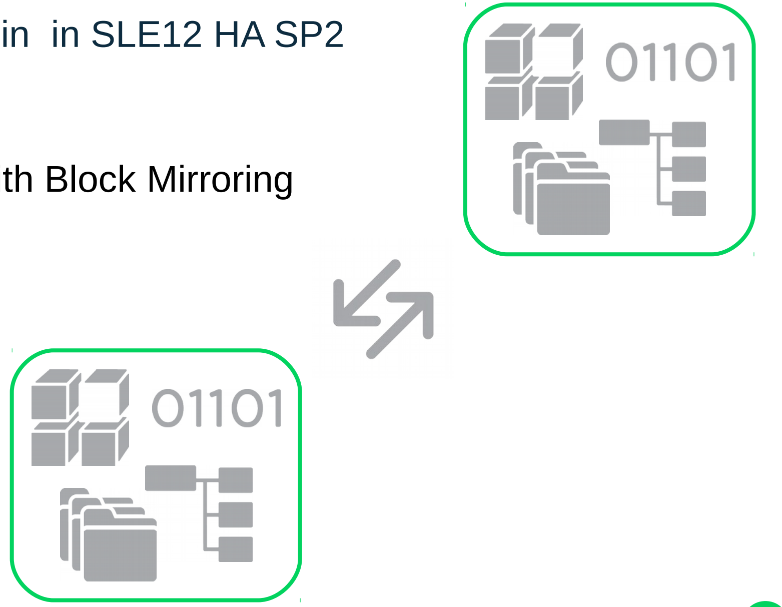


Is Data Safe Enough?

Data Replication

Data Replication in general

- **File Replication** (Not the focus of this talk.)
- **Block Level Replication**
  - A great new feature “Clustered MD RAID 1” now is available in in SLE12 HA SP2
- **Object Storage Replication**
  - TUT89016 - SUSE Enterprise Storage: Disaster Recovery with Block Mirroring



Requirements in HA context

- Multiple nodes
- Active-active

Data Replication – DRBD

DRBD can be thought as a networked RAID1.

DRBD allows you to create a mirror of two block devices that are located at two different sites across the network.

It mirrors data in real-time, so its replication occurs continuously, and works well for long distance replication.

SLE 12 HA SP2 now supports DRBD 9.

Data Replication – clvm/cmirrord

- There are different types of LV in CLVM: Snapshot, Striped and Mirrored etc.

- CLVM has been extended from LVM to support transparent management of volume groups across the whole cluster.

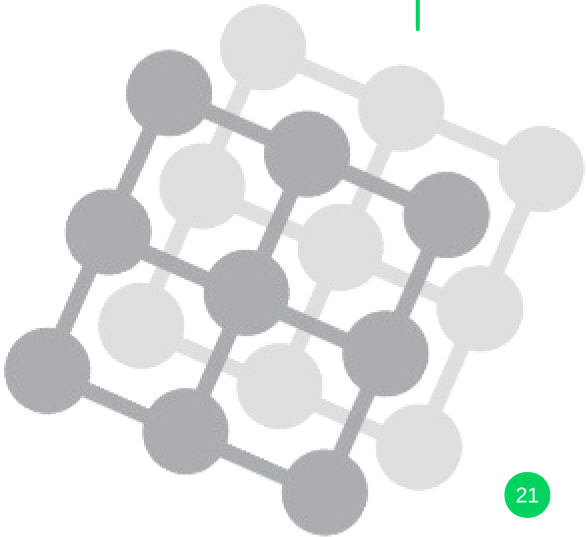
- For CLVM, we can also created mirrored lv to achieve data replication, and cmirrord is used to tracks mirror log info in a cluster.

Data Replication – clustered md/raid1

The cluster multi-device (Cluster MD) is a software based RAID storage solution for a cluster.

The biggest motivation for Cluster MD is that CLVM/cmirrord has severe performance issue and people are not happy about it.

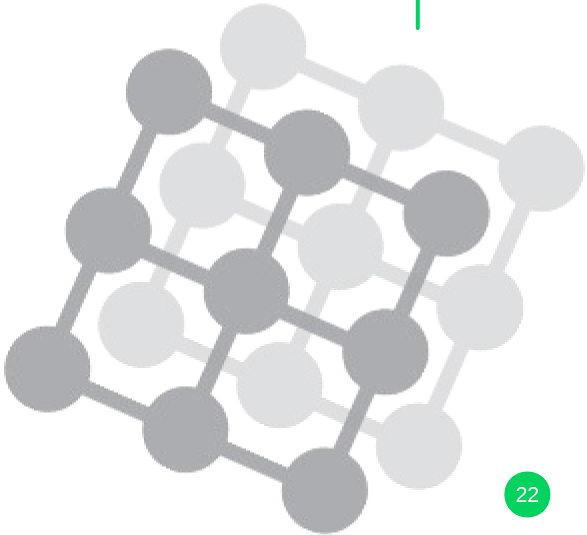
Cluster MD provides the redundancy of RAID1 mirroring to the cluster. SUSE considers to support other RAID levels with further evaluation.



Data Replication – clustered md/raid1 (cont.)

Internals:

- Cluster MD keeps write-intent-bitmap for each cluster node.
- During "normal" I/O access, we assume the clustered filesystem ensures that only one node writes to any given block at a time.
- With each node have it's own bitmap, there would be no locking and no need to keep sync array during normal operation.
- Cluster MD would only handle the bitmaps when resync/recovery etc happened.



Data Replication – clustered md/raid1 (cont.)

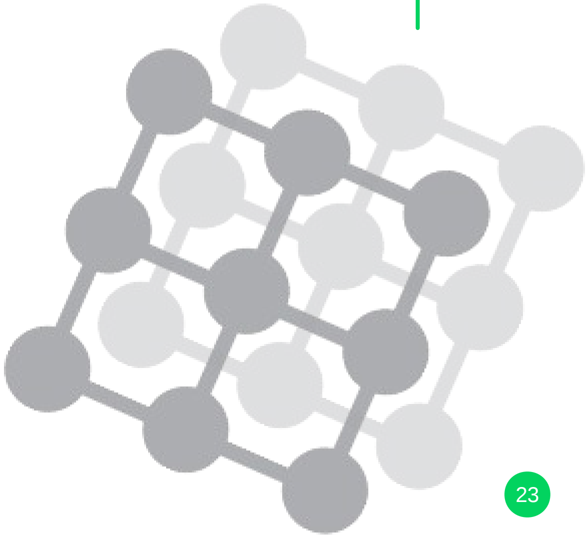
It coordinates RAID1 metadata, not coordinate access to the shared data, and it’s performance is close to native RAID1.

CLVM must send a message to user space, that must be passed to all nodes. Acknowledgments must return to the originating node, then to the kernel module.

Some related links:

https://lwn.net/Articles/674085/

http://www.spinics.net/lists/raid/msg47863.html



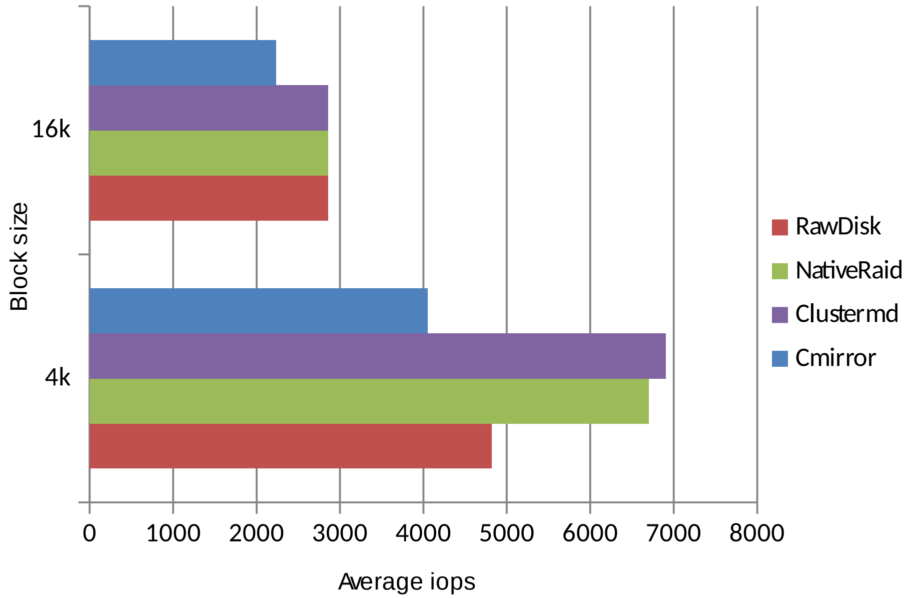
Data Replication – Comparison

	Active/Active mode	Suitable for Geo	Shared Storage
DRBD	Supported (limited to two nodes)	Yes	No, storage is dedicated to each node.
CLVM (cmirrord)	Supported, the node number is limited by pacemaker and corosync	No	Yes
Clustered md/raid1	Supported, the node number is limited by pacemaker and corosync	No	Yes

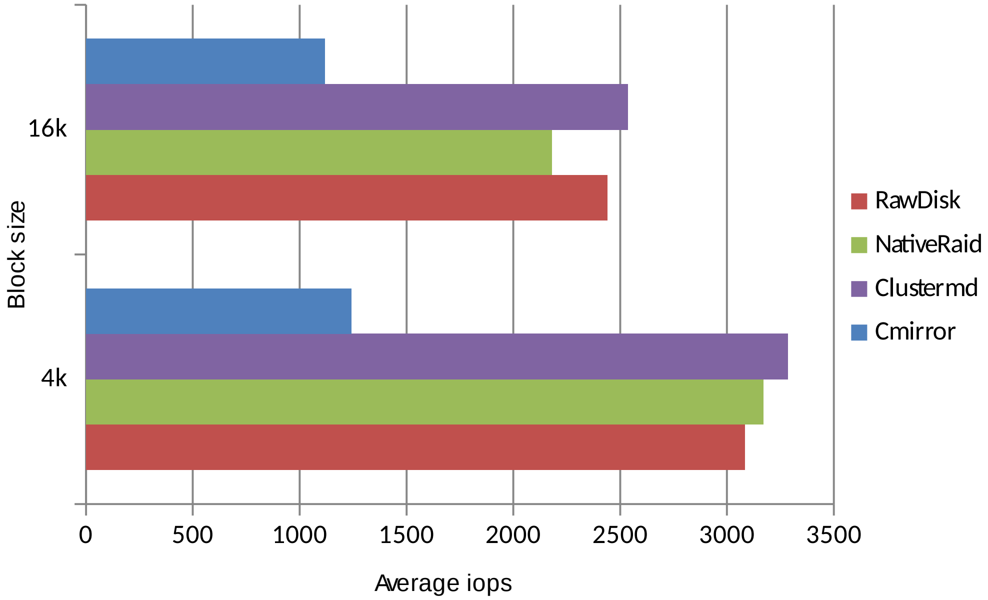
Data Replication – Performance Comparison

FIO test with sync engine

Read

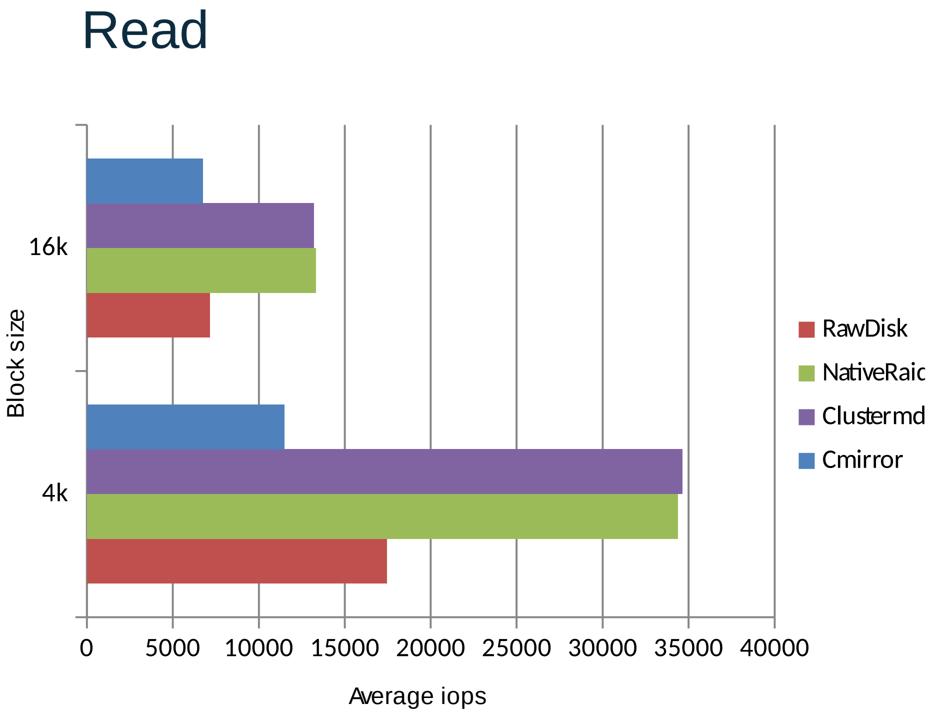


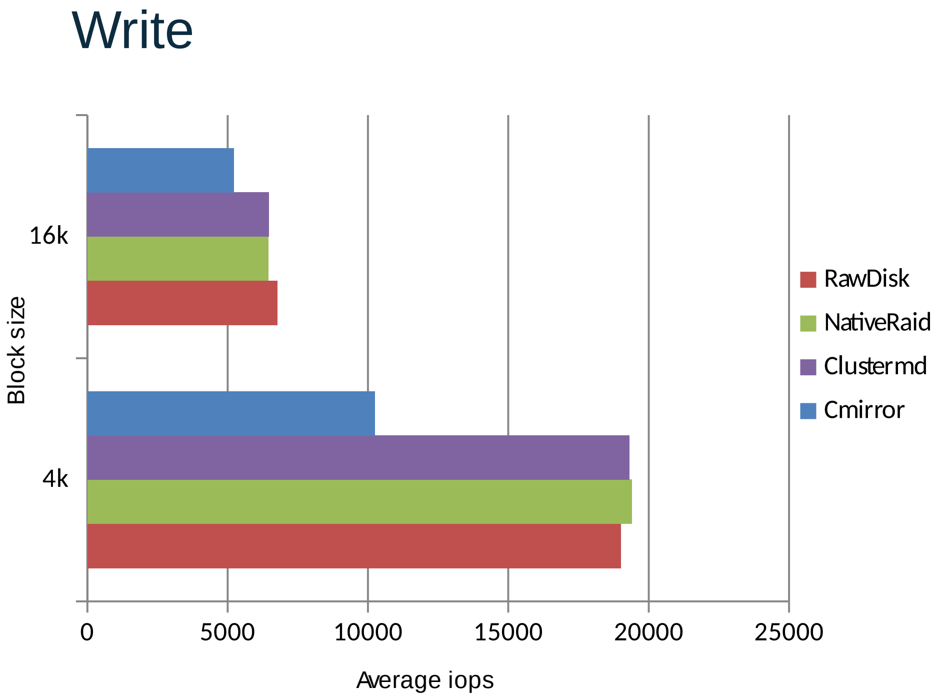
Write



Data Replication – Performance Comparison

FIO test with libaio engine





Recap

Recap: HA Storage Building Blocks

Block-level

- HA DRBD
- HA LVM2
- HA iSCSI
- cLVM2
- Clustered MD RAID1

Filesystem-level

- HA NFS / CIFS
- HA EXT3/EXT4/XFS
- OCFS2 (GFS2)

Question & Answer

SUSE®

We adapt. You succeed.