High Availability Storage

High Availability Extensions

Goldwyn Rodrigues

High Availability Storage Engineer

SUSE

High Availability Extensions

- Highly available services for mission critical systems
- Integrated suite over robust open-source technologies
- Business Continuity
- Protect Data Integrity
- Reduce Unplanned Downtime
- Commodity Hardware for high availability

Cluster

- A set of computers interacting with each other
- One goes down, another picks up its responsibility
- Should be available as much as possible
- Cluster Types:
  - Active/Active vs Active/Passive (N+1, N+M)
  - Physical vs Virtual vs Hybrid
  - Local Clusters vs Metro vs Geo Clusters

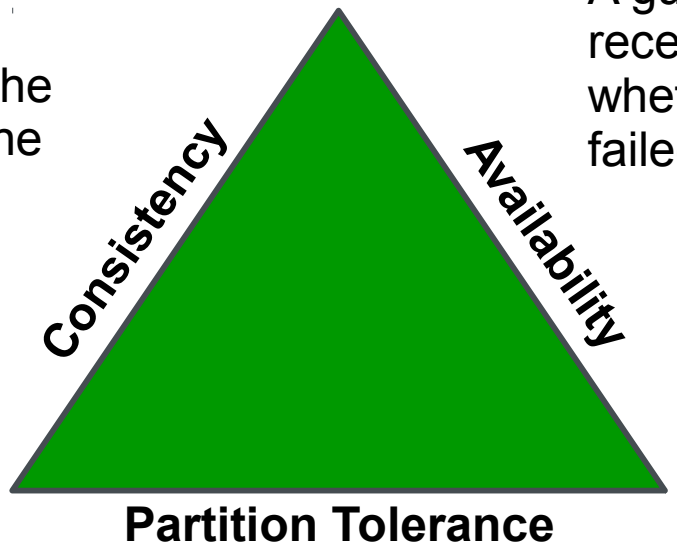
Why Cluster

- Increased Availability
- Improved Performance
- Low cost of operation
- Scalability
- Disaster Recovery
- Data Protection
- Server Consolidation
- Storage Consolidation

CAP Theorem

Brewers Theorem

All nodes see the same data at the same time



A guarantee that **every** request receives a response about whether it was successful or failed

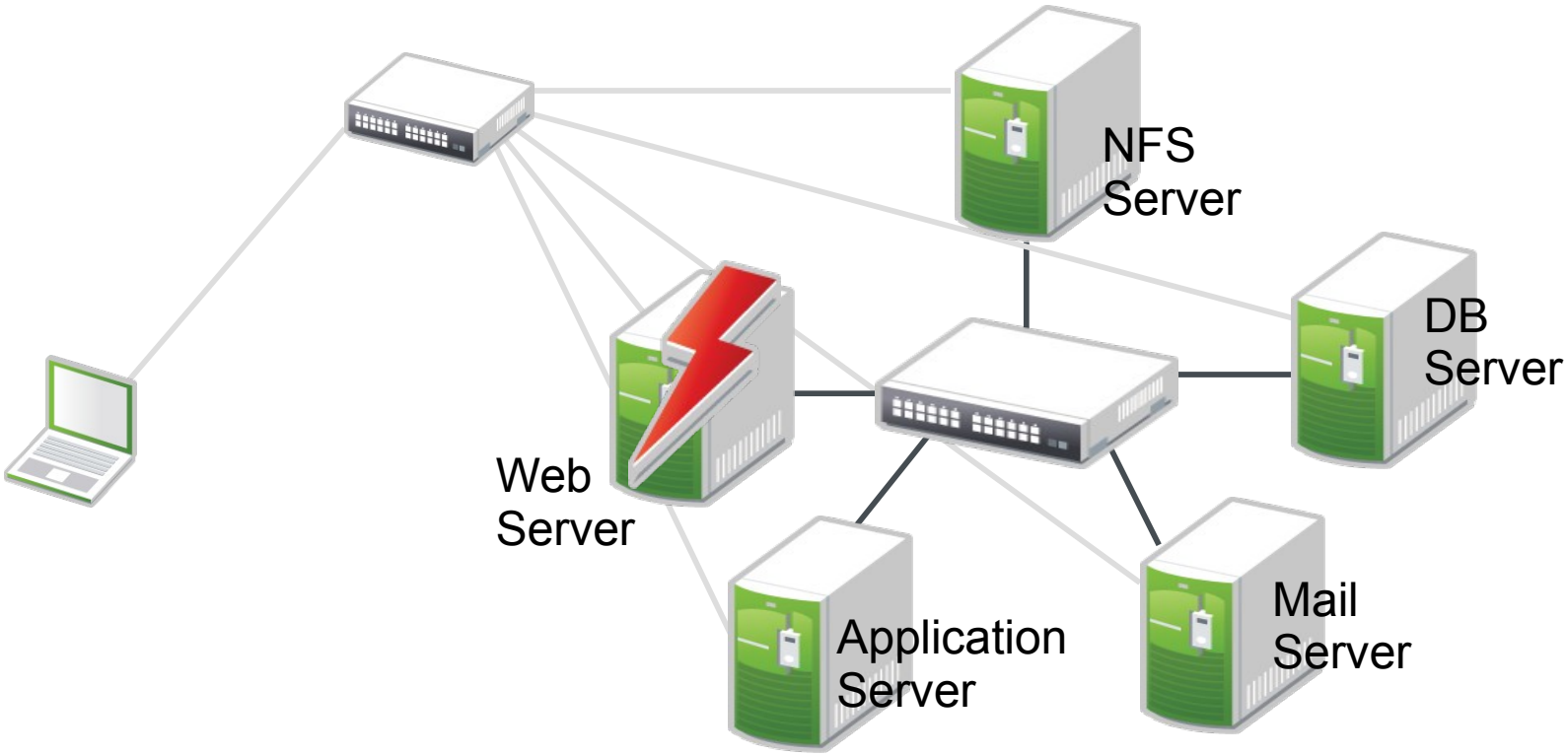
The system continues to operate despite arbitrary message loss or failure of part of the system

Fault Tolerance vs High Availability

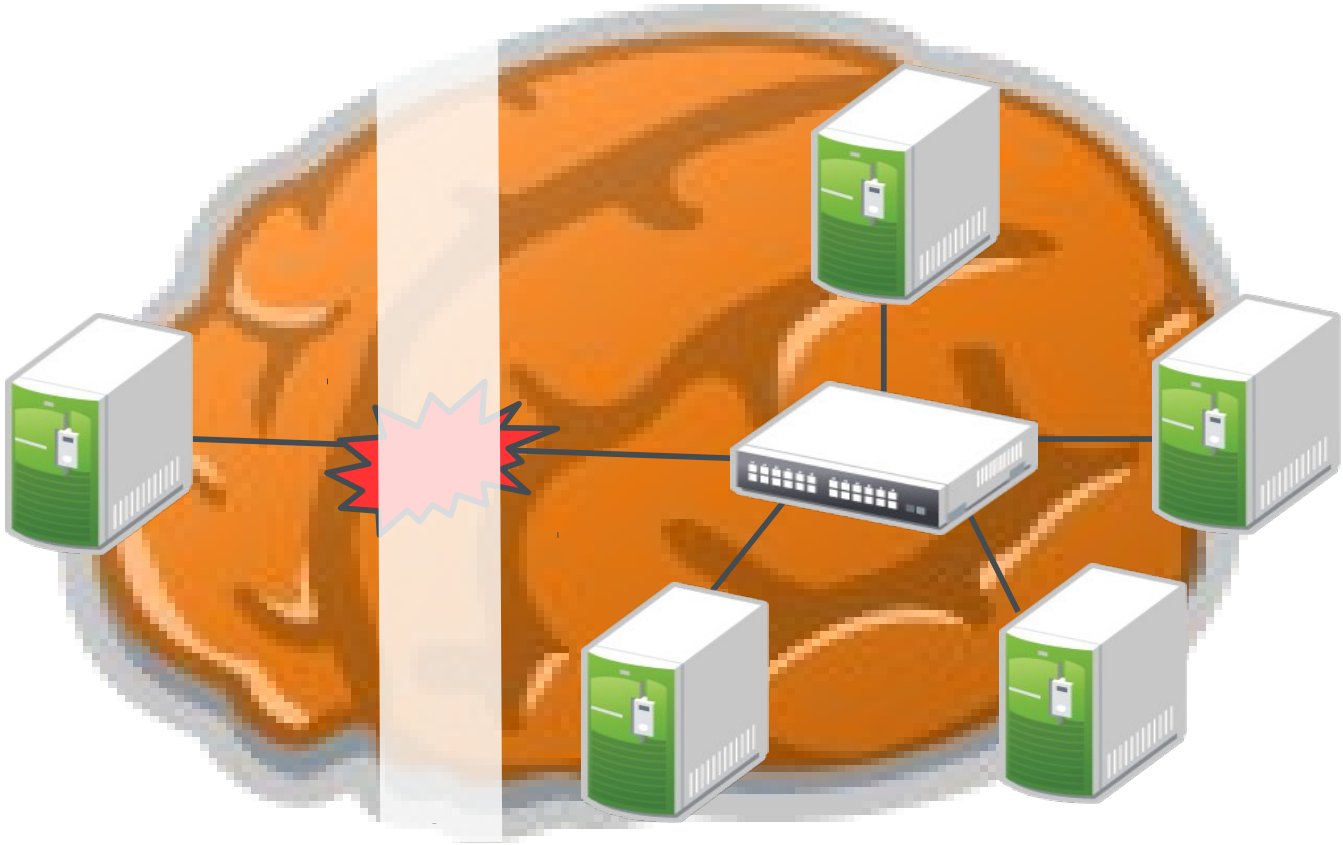
- Specialized hardware to detect a hardware fault and switch to redundant hardware
- Expensive redundant and replicated components

- A set of computers system-wide, shared resources that cooperate to guarantee essential services
- Software and hardware to quickly restore services

Cluster

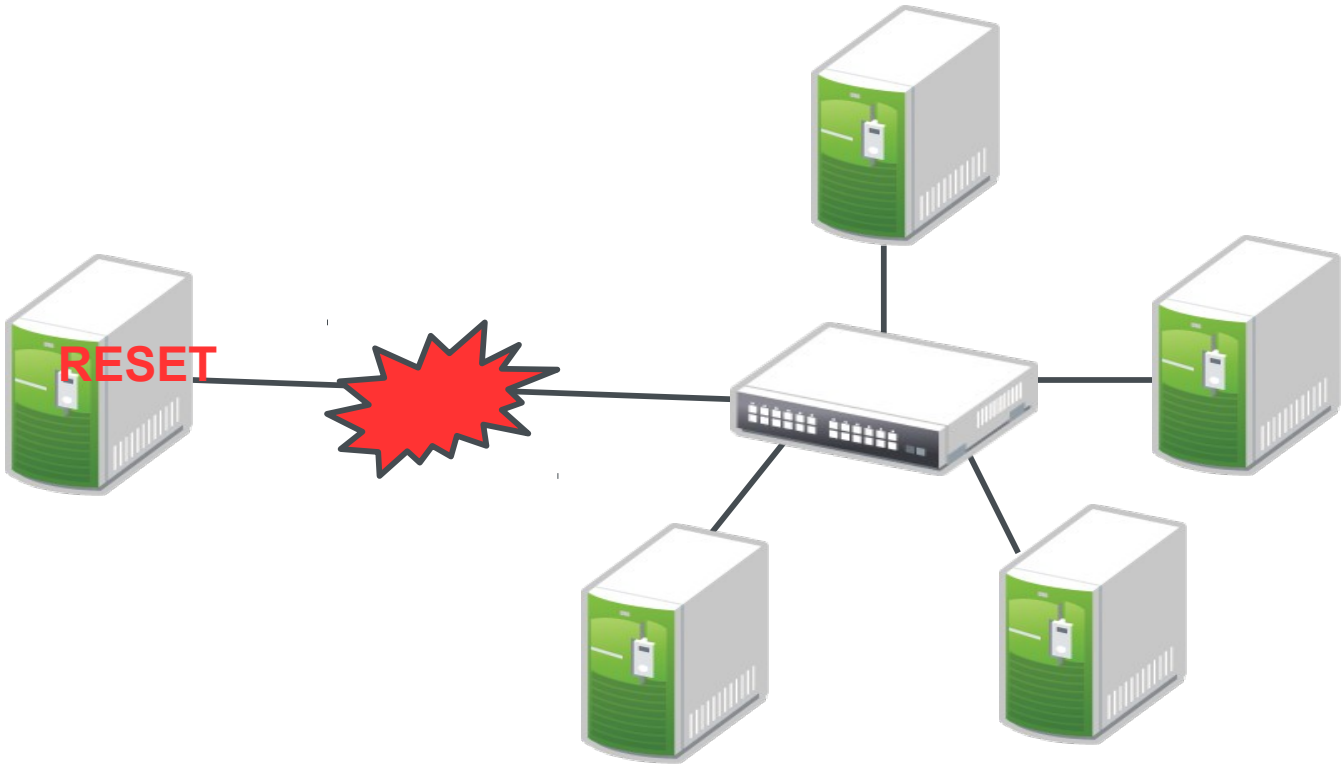

NFS
Server
DB
Server
Web
Server
Application
Server
Mail
Server

A Dysfunctional Cluster



STONITH

Shoot the Other Node In the Head



Quorum

Quorum Policies

- Ignore
  - Continue cluster operations as usual
- Freeze
  - Resource management continues
  - New resources are not started
- Stop
  - All resources affected partition are stopped
- Suicide
  - Fence all nodes in affected partition

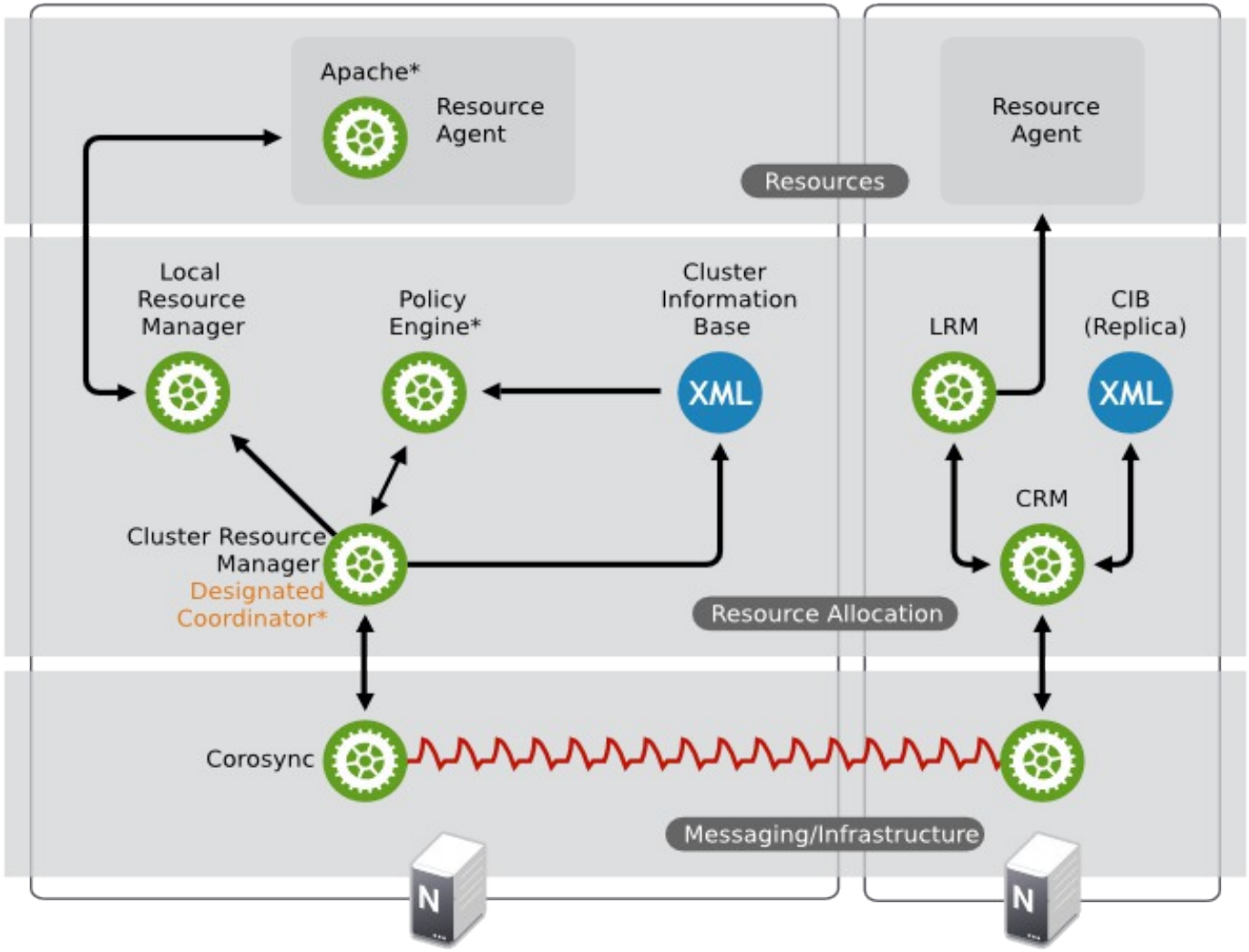
Resource Agents

- Open Cluster Framework (OCF)
- Manage resources
  - Web Server
  - IP Address
  - Shared Filesystem
- Resource Operations
  - Id
  - Name
  - Interval
  - timeout

Configuration Tools

- Cluster Resource Manager (CRM)
  - Powerful Command line tool
- YasT
  - Basic Cluster setup
  - DRBD
  - IP Load balancing
- High Availability Web Konsole (Hawk)
  - Web based

Architecture

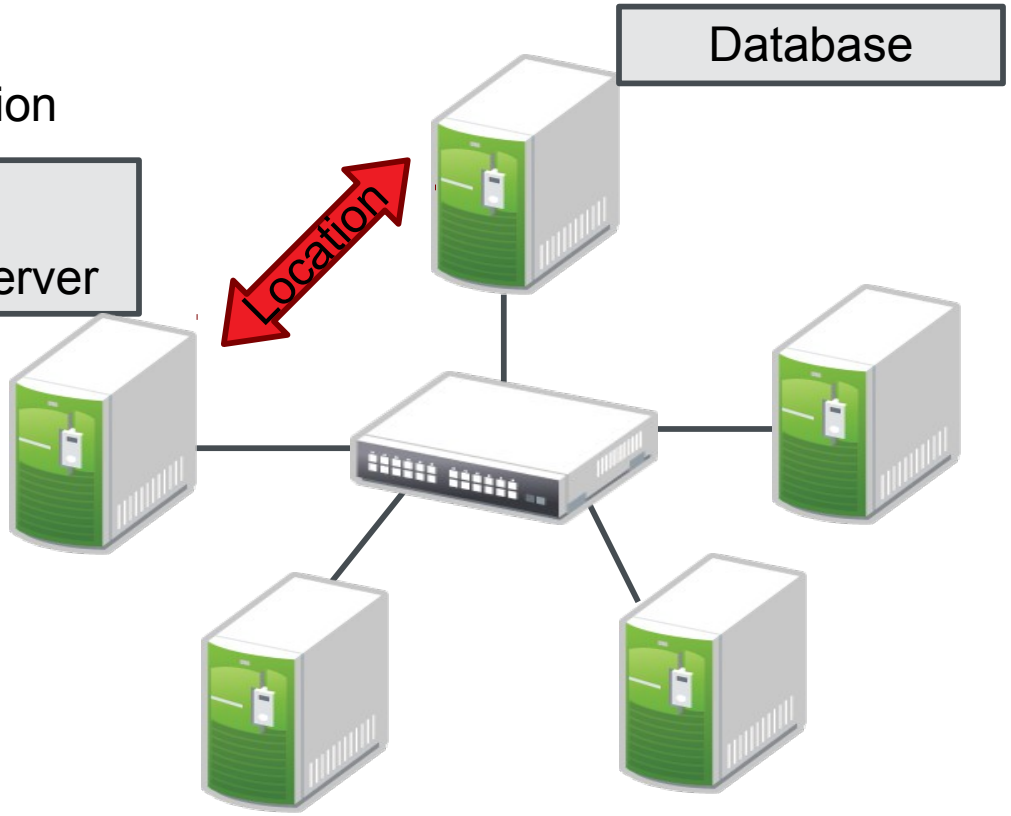


Resource Agent Constraints

Colocation

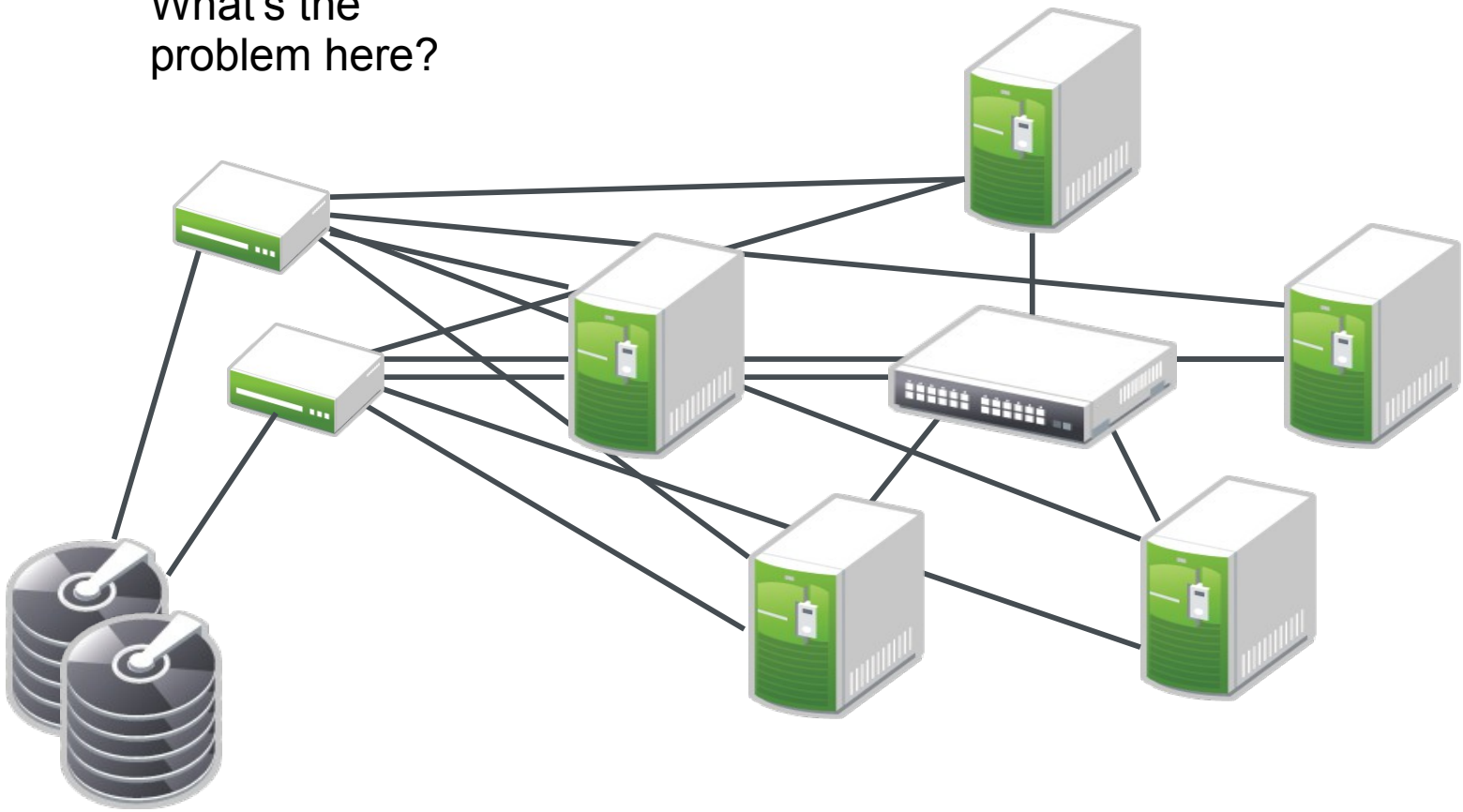
1. IPAddr
2. Web Server

Order



Shared Storage

What's the problem here?



Shared Storage

- A common view for all nodes
- Node Failure: One node can pick up from where the other left
- Local filesystems don't work
  - Data corruptions because of writing files other nodes access
  - Node Cache Inconsistencies

DLM

Distributed Lock Manager

- Provides a Cluster-wide locking for data access
- Different Level for wide variety of uses
- No centralized-control
  - Easy for take over
  - The node accessing the object first gets to create the distributed lock object
- Lock Value Block (LVB) for data synchronization

DLM

Distributed Lock Manager

Mode	NL	CR	CW	PR	PW	EX
NL	Yes	Yes	Yes	Yes	Yes	Yes
CR	Yes	Yes	Yes	Yes	Yes	No
CW	Yes	Yes	Yes	No	No	No
PR	Yes	Yes	No	Yes	No	No
PW	Yes	Yes	No	No	No	No
EX	Yes	No	No	No	No	No

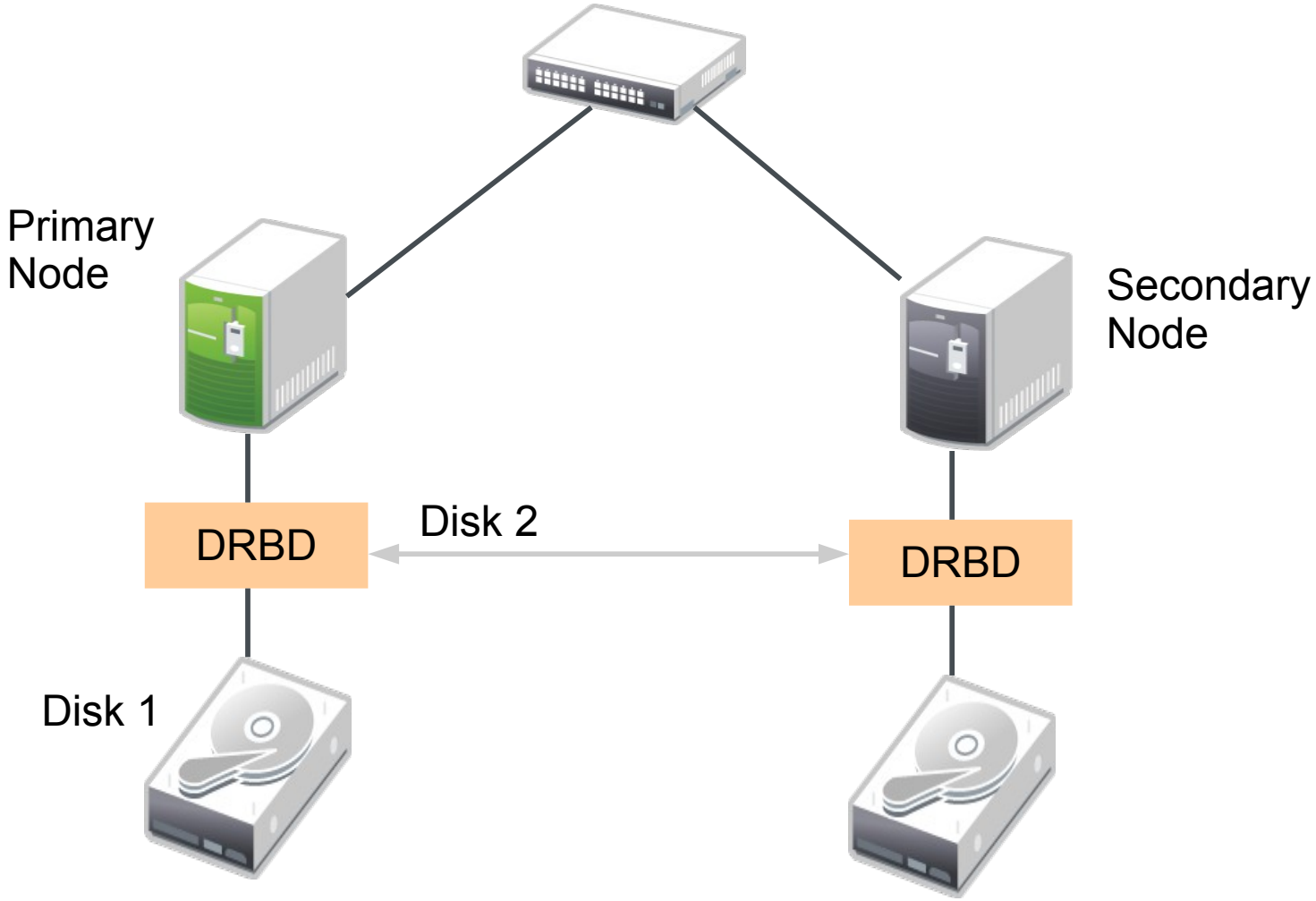
cLVM

Clustered Logical Volume Manager

- Logical Volume Manager for the cluster
- Add or remove devices as storage needs change
- Linear
  - Add storage as required
  - Simple addition of devices in a linear form
- Mirrored
  - Redundancy of devices over the cluster

DRBD

Distributed Replicated Block Device



DRBD

Distributed Replicated Block Device

- Shared Storage without a SAN
  - Governed by network speeds
- RAID1 over network block device and local device
- Fully synchronous, memory synchronous or asynchronous modes of operation
- Dual Primary for clustered filesytems

Shared Filesystem

ocfs2

- Simultaneously mounted on all nodes
  - All nodes should have access to the data
- Cluster Filesystem with a good throughput
  - Should not be bogged down with multiple access
  - Force cache flush on other nodes using filesystem when current node access the file

OCFS2 Features

- B-tree Extent based
- Inline data
- Indexed Directories
- Metadata ECC
- Refcount
- Extended Attributes
  - ACL
- Quota

Read man mkfs.ocfs2 for more information..

Usage Scenarios

- Database Server Applications
  - Common database repository
  - CTDB (Samba)
- Web Servers
  - WWW root
- Highly Available Virtual Machines
  - VM Disks

Generic Clustering Tips

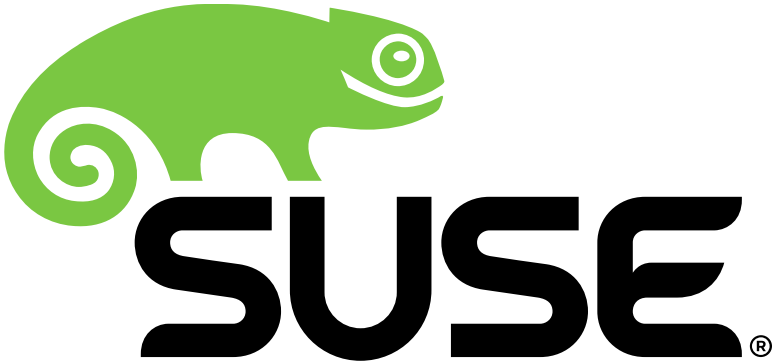
- Always use STONITH
  - Recommend multiple STONITH devices
  - SBD is an alternative
- Time synchronization
- Read the logs when things are not working
  - Record Time of event
- Redundant Communications
  - Network Device bonding
  - Redundant Ring Protocol

OCFS2 Tips

- Prefer hardware based RAID (mirroring)
- If you don't want a feature don't enable it
  - Quota, ACL, Inline directories use additional data on disk and additional lookups
  - You can always enable it later using tunefs.ocfs2
- Inline directories for large number of files
- MetaECC for better data protection
  - Protects filesystem from getting corrupted further
  - Immediately run fsck.ocfs2

SUSE® Linux Enterprise High Availability Extension
https://www.suse.com/products/highavailability/

Thank you.


SUSE®
We adapt. You succeed.