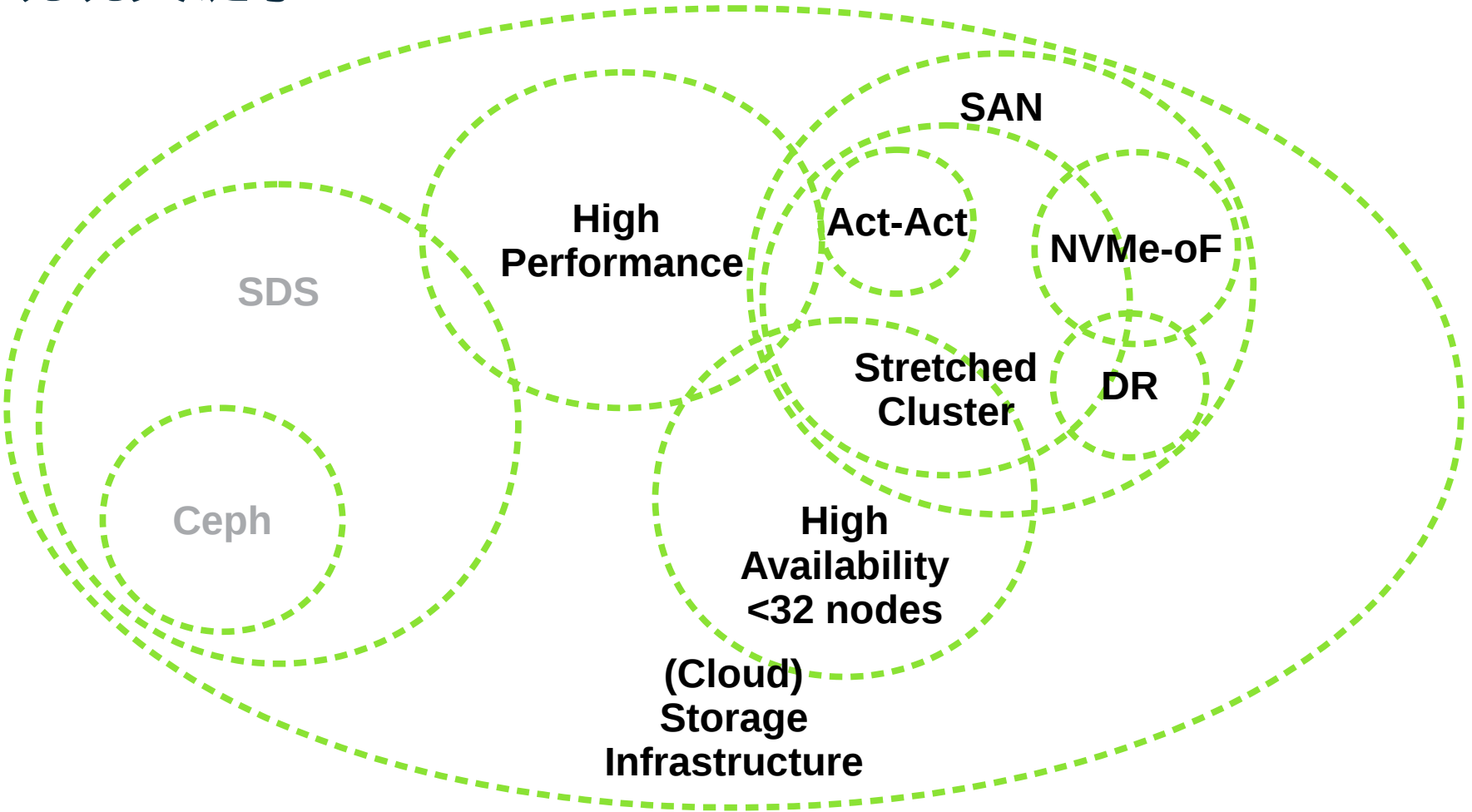
延伸 Linux 关键业务到双活 NVMe-oF 存储

- 近期 SUSE 研发人员的相关进展

周志强 (Roger)
SUSE 高级研发经理
zzhou@suse.com

2018 OpenInfra Days China

说说关键字


SAN
High
Performance
Act-Act
NVMe-oF
SDS
Stretched
Cluster
DR
Ceph
High
Availability
<32 nodes
(Cloud)
Storage
Infrastructure

Short about NVMe-oF in Linux

- **Linux storage stack catches up the hardware evolution**
  - Transport: 100M/1G → 10G/40G/100G network
  - Media: HDD → SSD Flash
  - S/W Stack: SCSI protocol → NVMe protocol
  - NVMe-oF Storage Array: Very High iops, Very Low Latency.

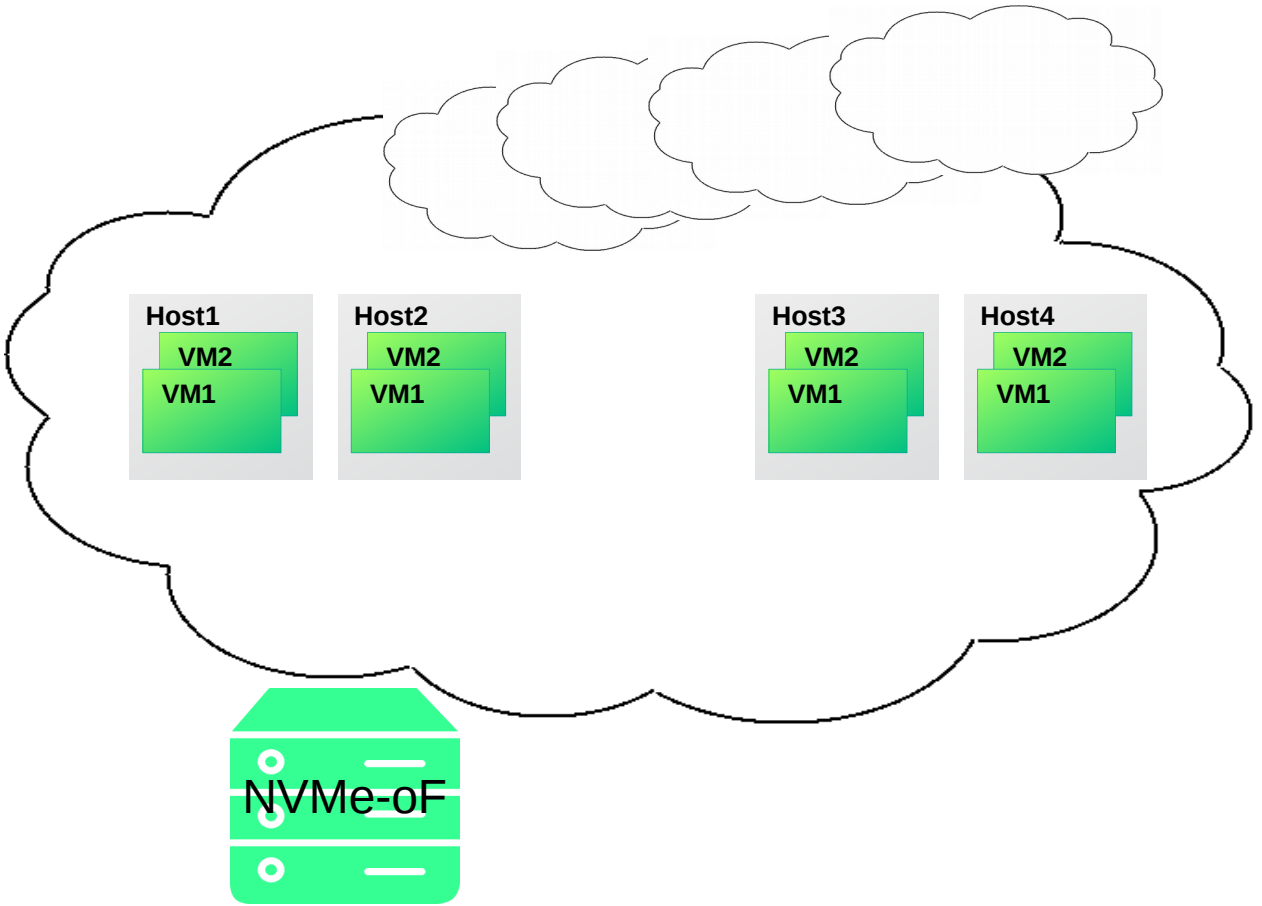
- **Linux MD RAID1 new I/O barrier, 70% NVMe speed.**
  - Contributed by Coly Li, Neil Brown, Hannes Reinecke, Guoqing Jiang, etc.
  - 2017, SLE12SP2 Maintenance Update

- **NVMe-oF products.**
  - 2017, SLE12SP3 support NVMe-oF with NetApp, Emulex, Mellanox.
  - 2018-05, Broadcom, NetApp and SUSE Announce Production Availability

NVMe-oF in Data Centers

Data Center

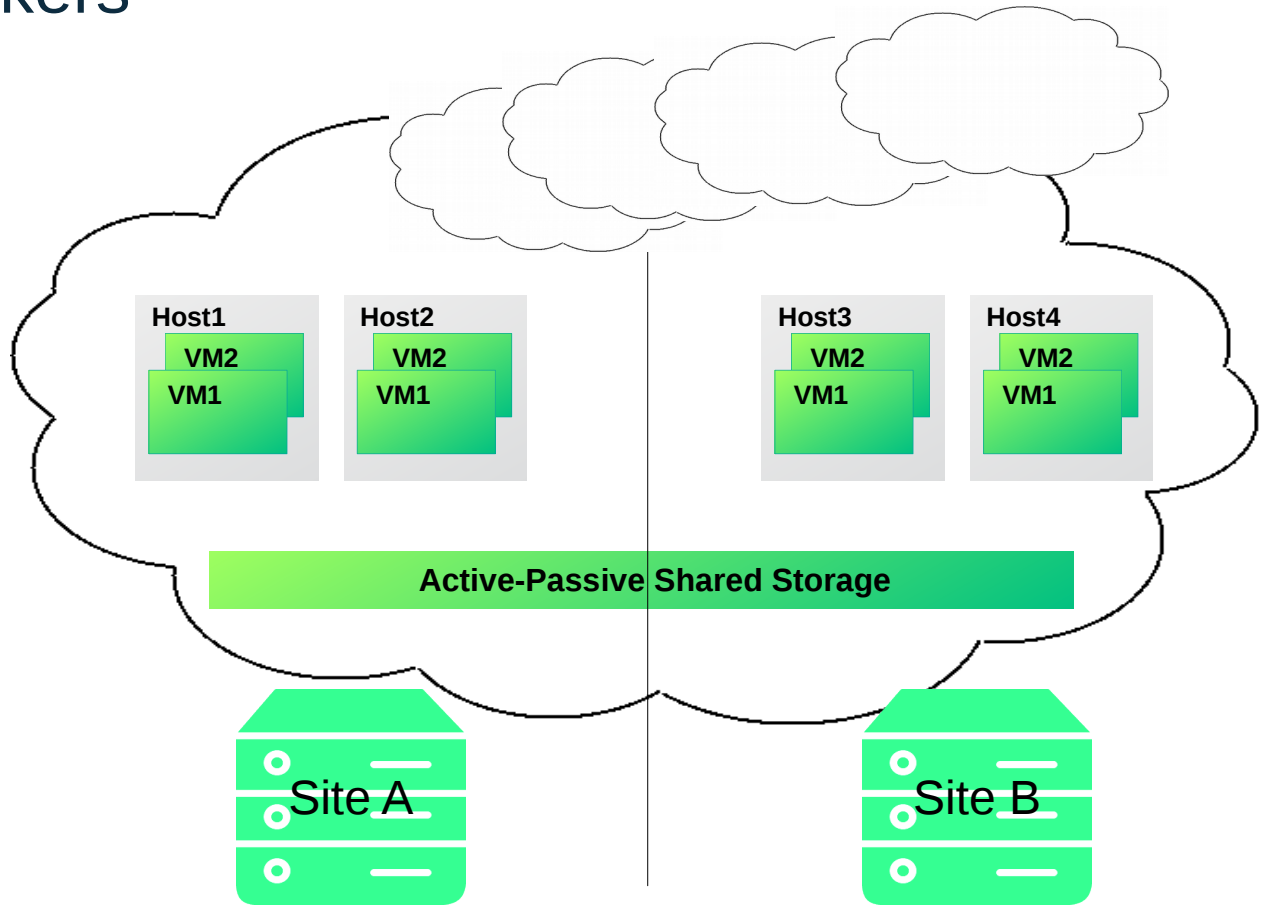
- 期望： FTT >= 2
  Failures To Tolerate
  可容忍 / 可恢复错误的数量
- 期望：接近于 0 的 RTO/RPO.
- 期望：数据保护 / 灾难恢复



Successful Stories: Stretched / Host-Based-Mirroring - Banking & Automakers

- 支持异构存储
- 解锁 供应商特定方案 , 特定存储
- 解锁 灾难恢复 (DR) 厂商专有复制工具。
- 解锁 基于存储厂商的镜像复制工具 .
- 和 Linux 无缝集成

** hundreds of clusters



Successful Stories: 里面的挑战
- Banking & Automakers

What happen during “Failover” ?

From node-1

1. To umount filesystem

2. To deactivate lvm

3. To remove RAID1

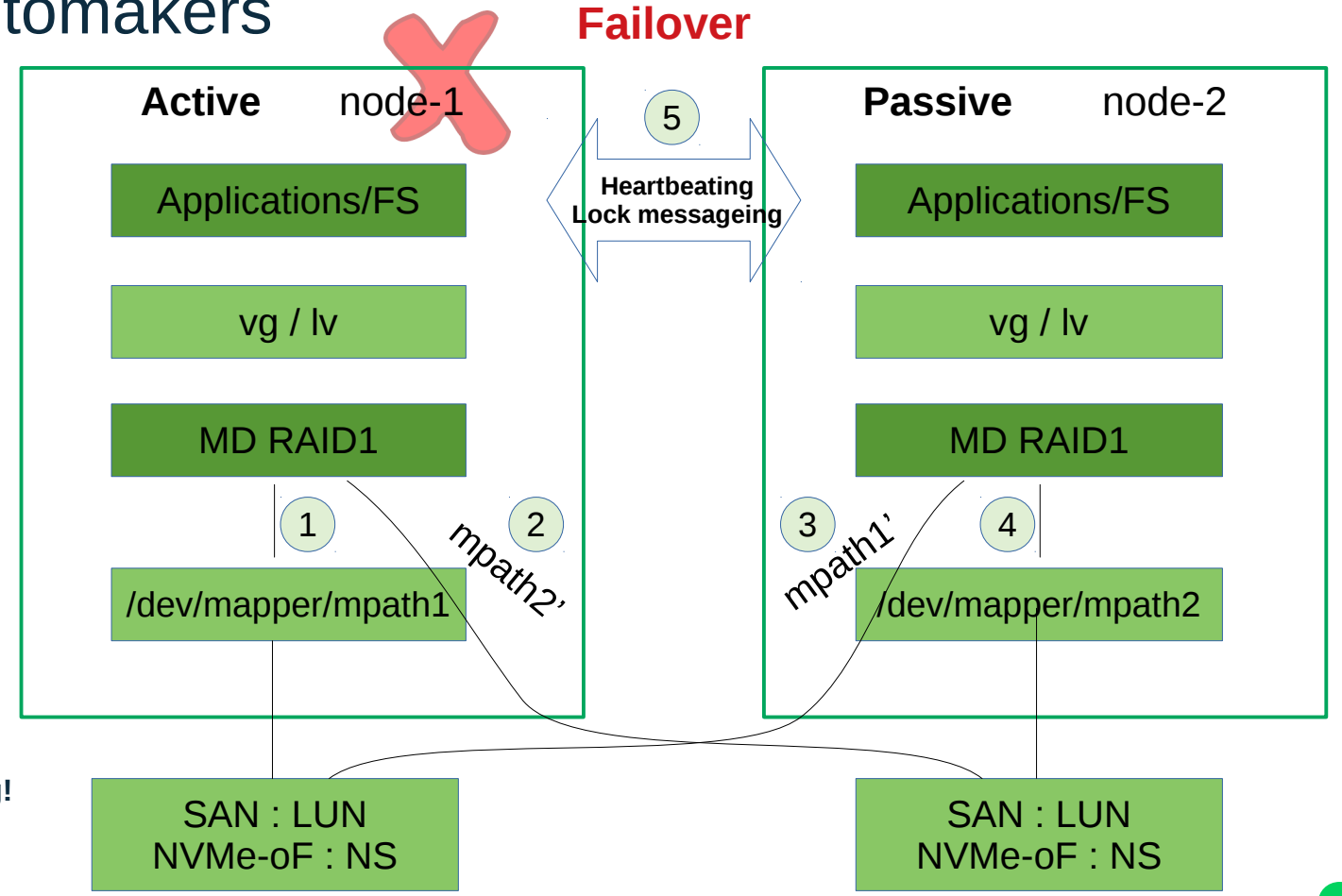
To node-2

4. To assemble RAID1

5. To activate lvm

6. To mount filesystem

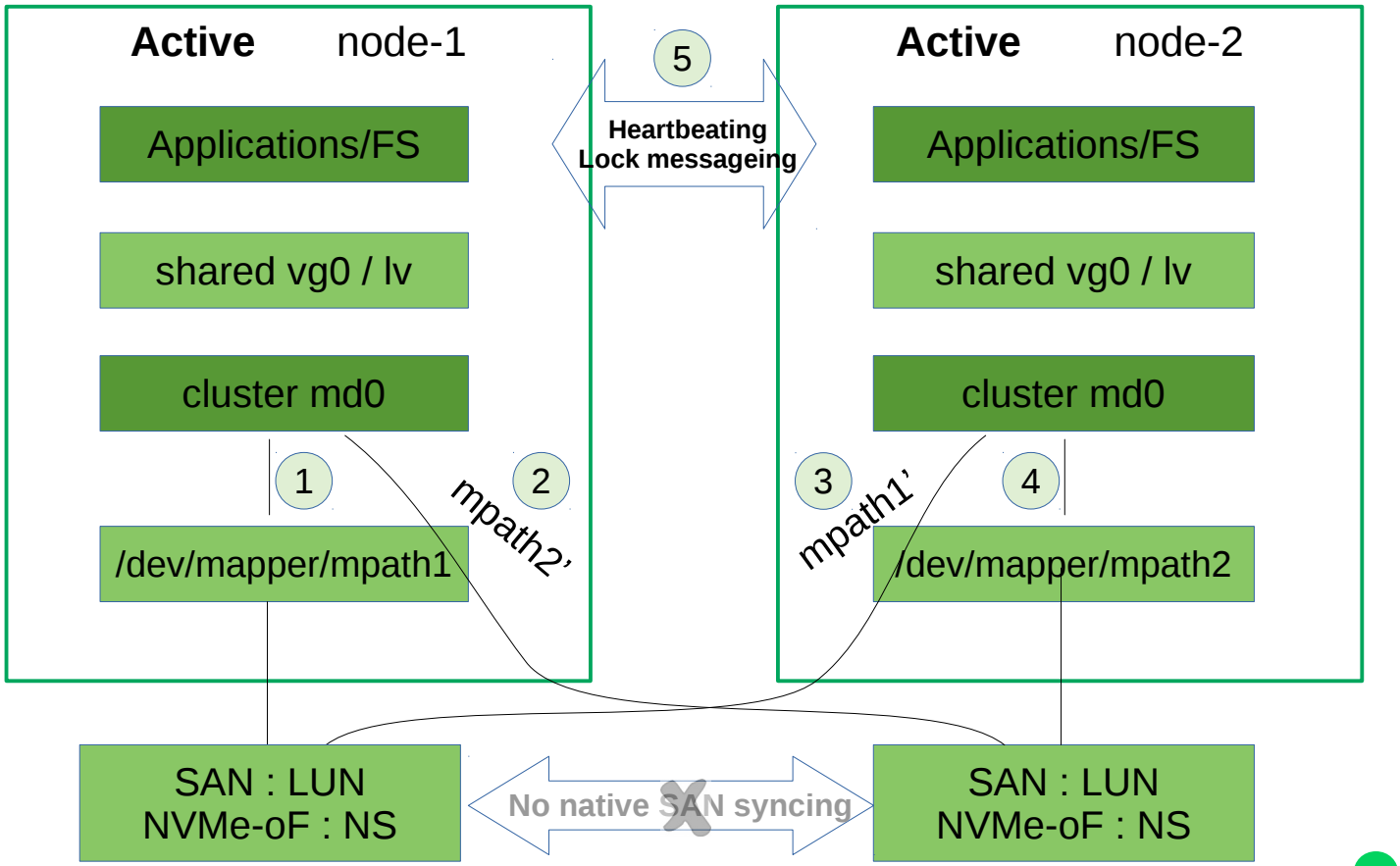
Imaging hundreds of RAID1 devices, **RTO can be very long!**



Improve the cluster to Active - Active (This Talk)

Linux
MD RAID
cluster aware
2016
Guoqing Jiang, Neil Brown

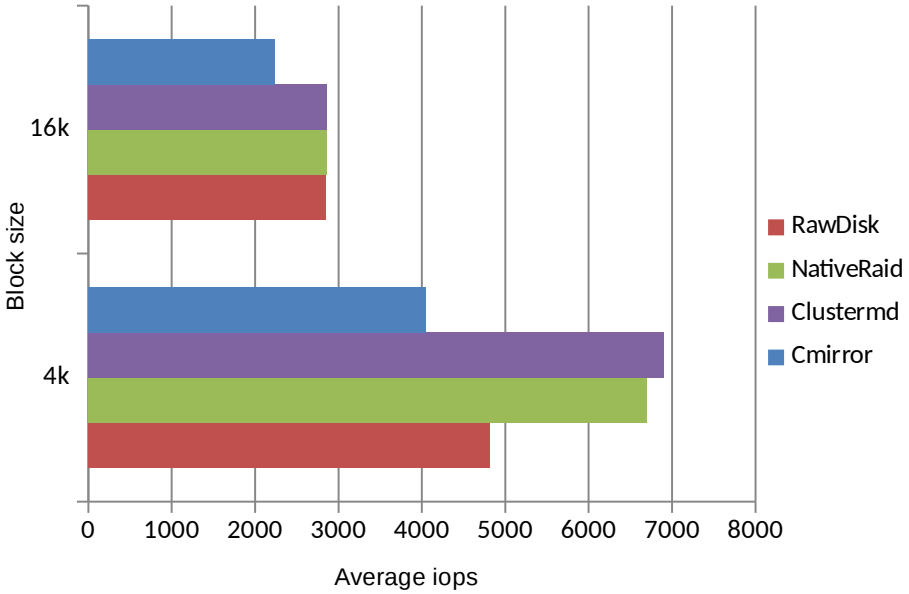
- **Assemble MD RAID1 on both datacenters**
- **Activate shared LV on both datacenters**
- **Mount OCFS2 on both datacenters**



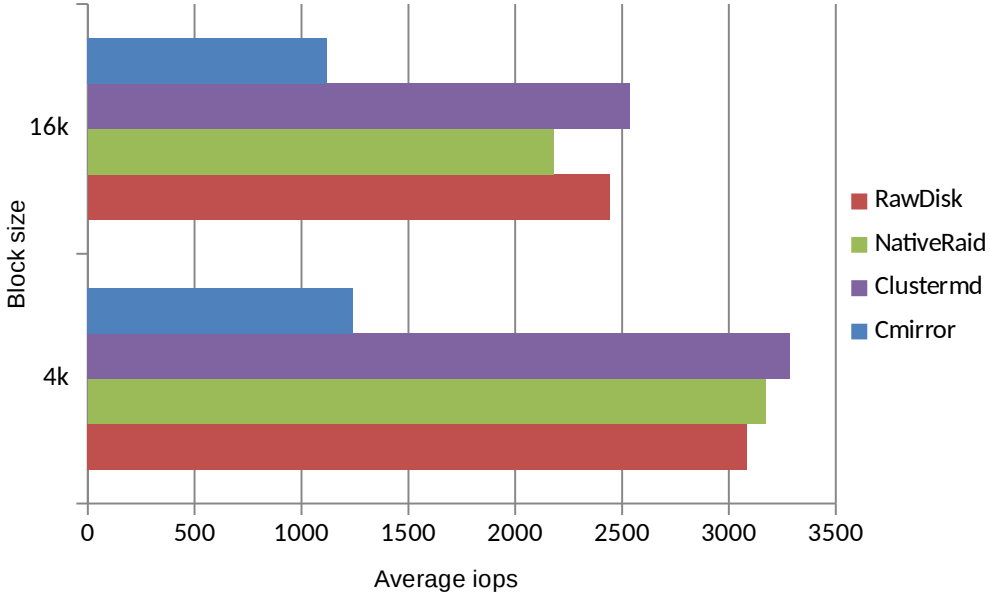
Cluster RAID1 performance is nearly same as native

FIO test with sync engine

Read



Write

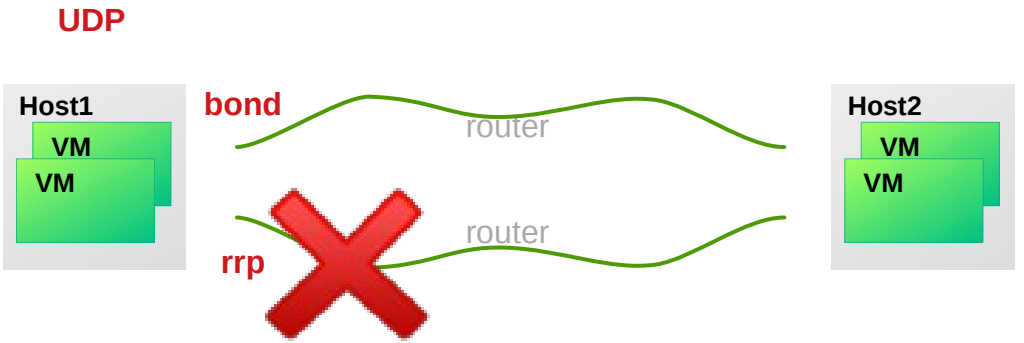


Failures in Stretched Cluster
- Ethernet / Cluster Communication

Keep stretching – Ethernet perspective

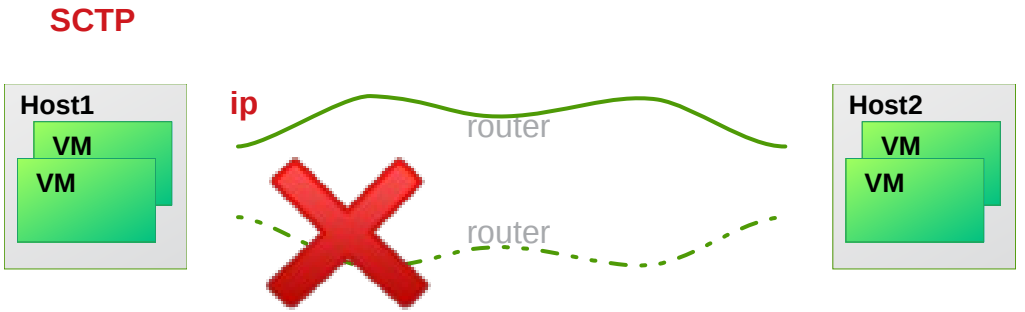
FTT = 1

- **Heartbeating**
  - Network Bonding (L2)
  - Redundant Rings (L3)



- **Distributed Lock Messaging**
  - SCTP

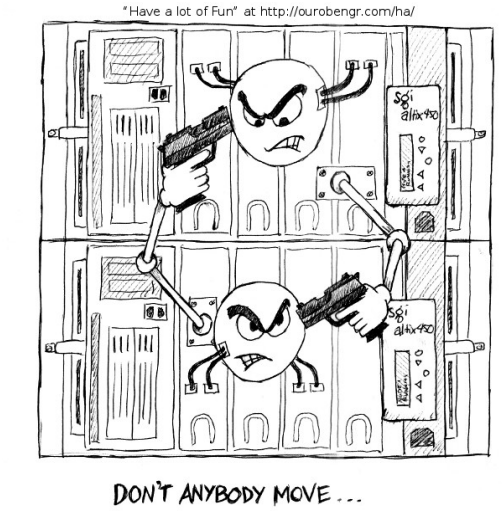
2018
Gang He, Michal Kubecek

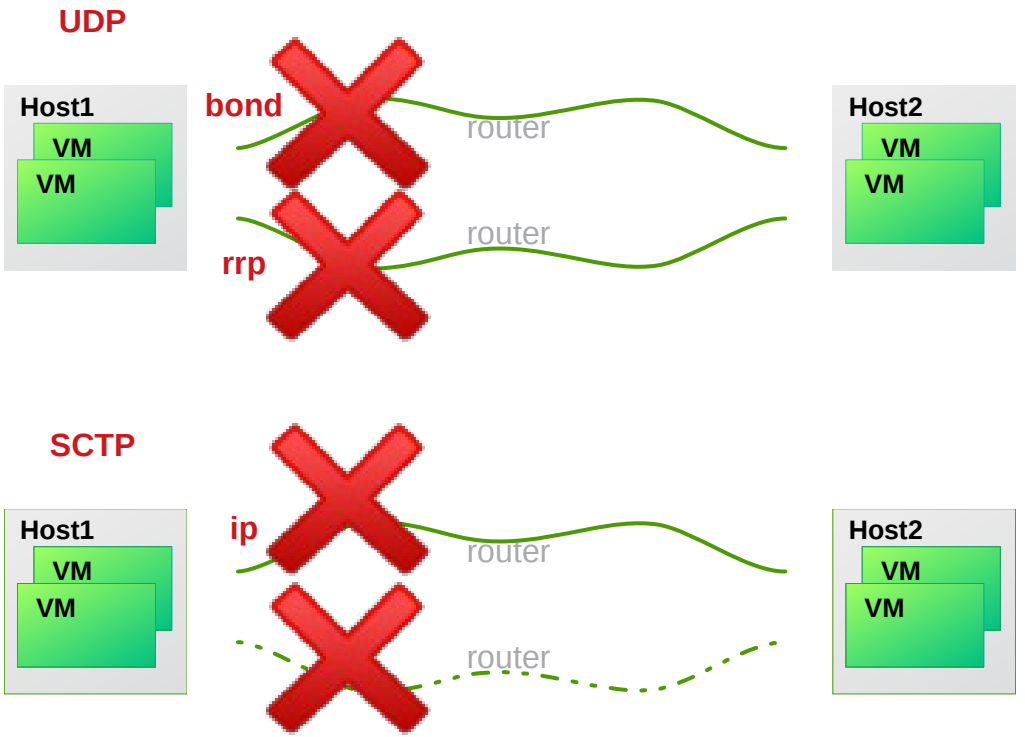


Mature Linux HA stack to deal with SPLIT BRAIN

FTT = 2

- Pacemaker
- Corosync
- STONITH

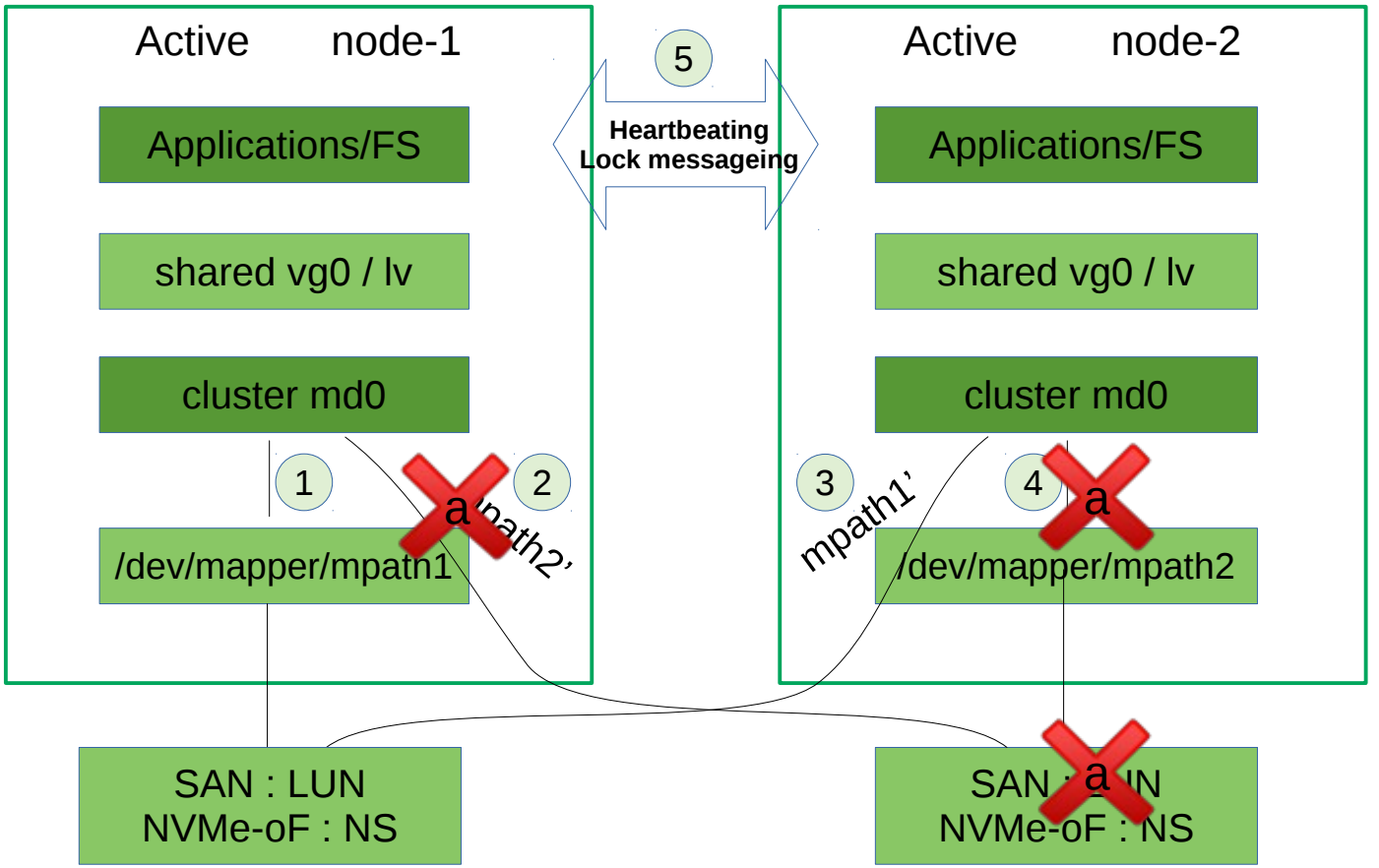




Failures in Stretched Cluster
- SAN Storage (eg. NVMe-oF)

Failure 1: SAN Storage(NVMe-oF) lose power

- Node-2 RAID1 marks mpath2 as FAULTY device.
- Node-1 RAID1 marks mpath2’ as FAULTY device.
- Both sites working well via node-1’s SAN storage.



Keep stretching

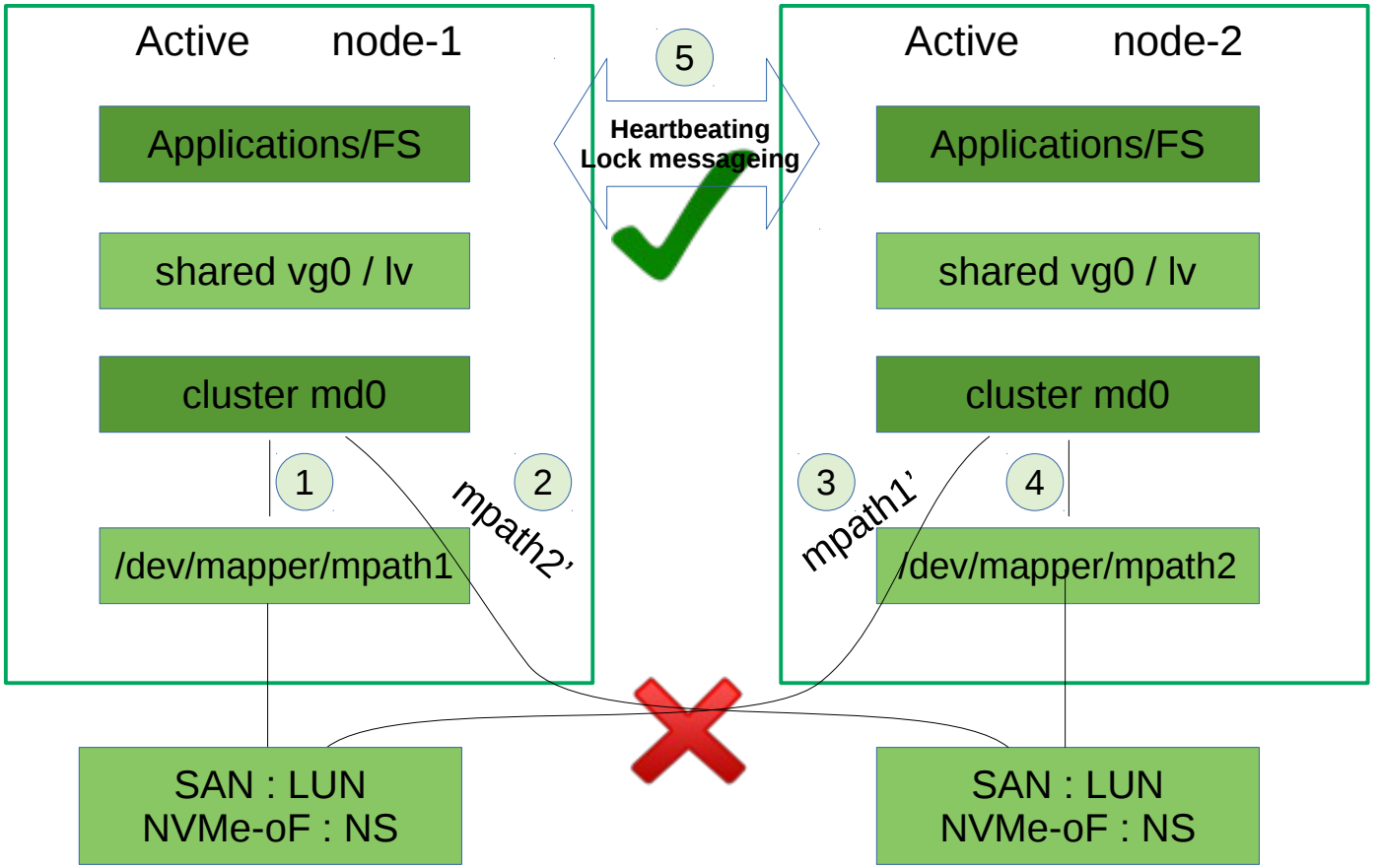
- **Storage links failures in between**
  (蓝翔挖掘机和光缆的恩怨)

Original image: baike.baidu

Failure: SAN Partitioned

Byzantine Failures
(Wikipedia)

组件在故障检测系统中的呈现可能不一致，不同的观察者有不同的症状：一个角度看正常工作，另一个看已经失败。



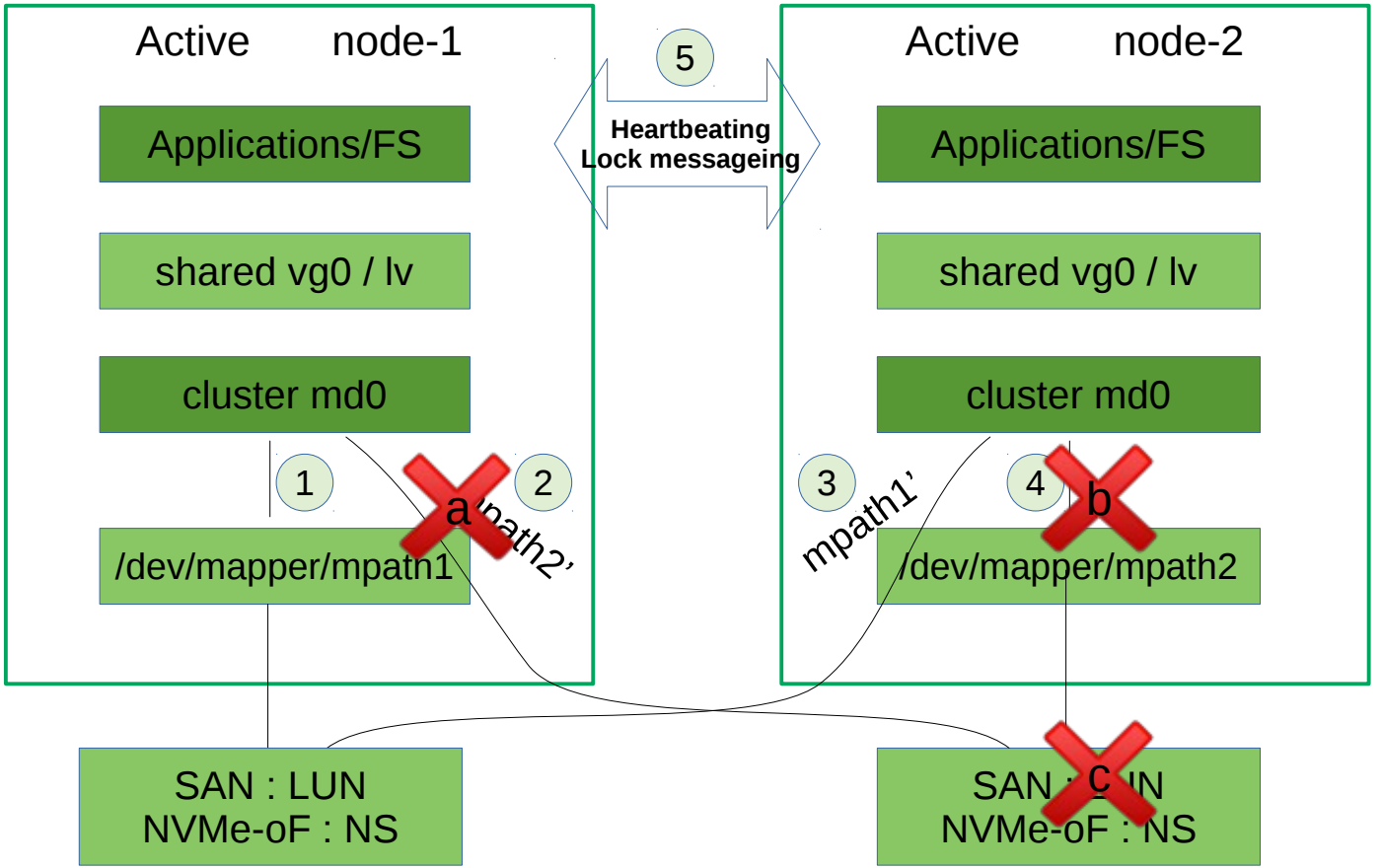
Failure 2: one storage link failed

a) **Assuming, Link② failed**. Node1 RAID1 marks mpath2' as FAULTY

b) **Cluster RAID1 will populate FAULTY device role** of mpath2' in superblock (*), and Node2 mpath2 becomes as FAULTY too.

c) That says, Cluster RAID1 will populate FAULTY disk. In the end. **Just like a whole SAN failure** .

(*) That says, MD RAID superblock plays the role to **populate FAULTY device role** in the cluster



Failure 3: SAN Partitioned : both links failed

a) Assume Link② is the first failure detected by the cluster.

- FAULTY is populated, and
- just like a whole SAN failure.

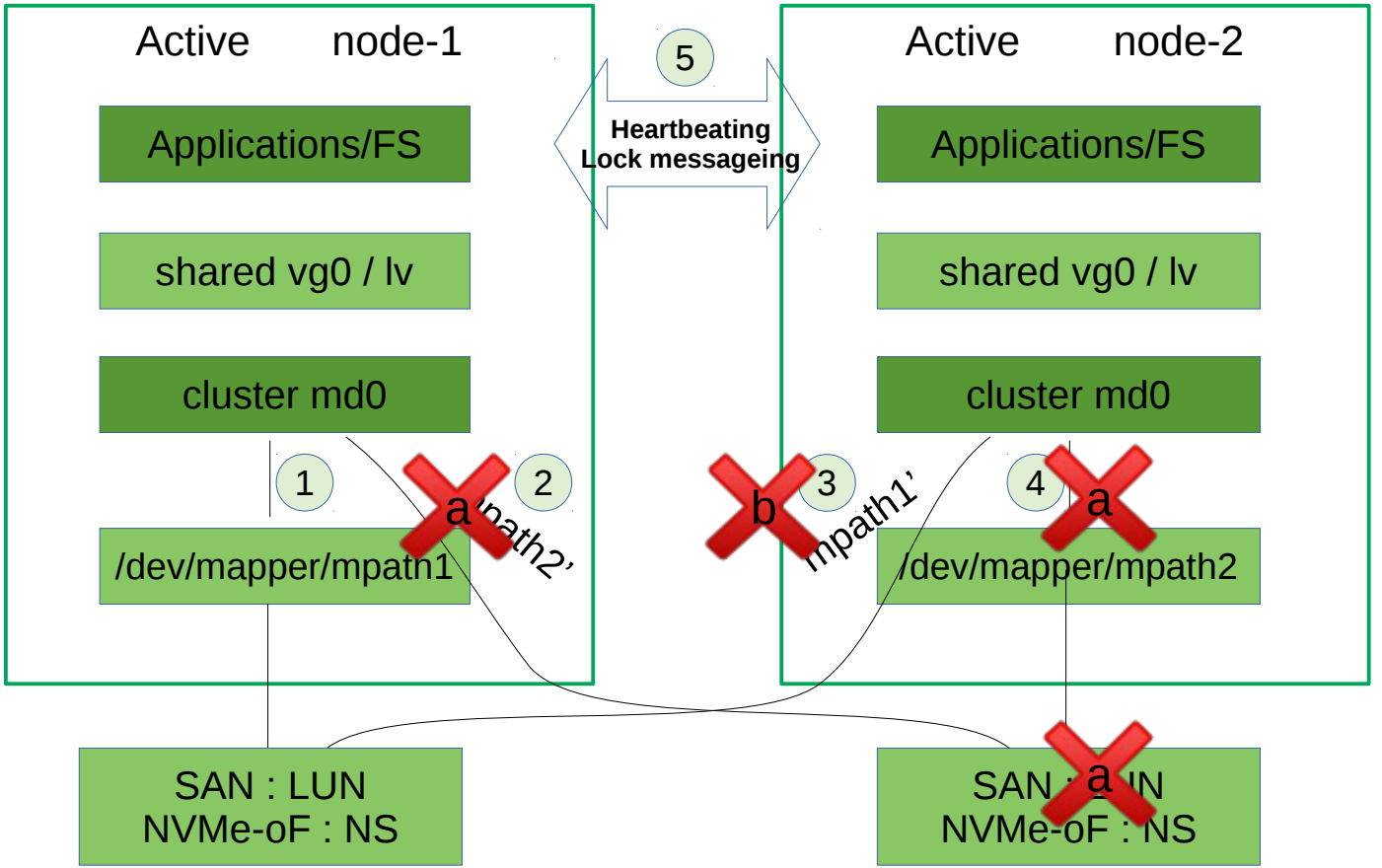
b) Sequentially(*), the cluster deals with Link③ failure.

- MD RAID1 on node-2 lose all devices.
- Cluster MD on node-2 is disabled. dmesg report: “ [79.942305] md: md0 stopped”.
- RA RAID will fail.
-

c) Services failover to node-1.

- Only one site keeps running.

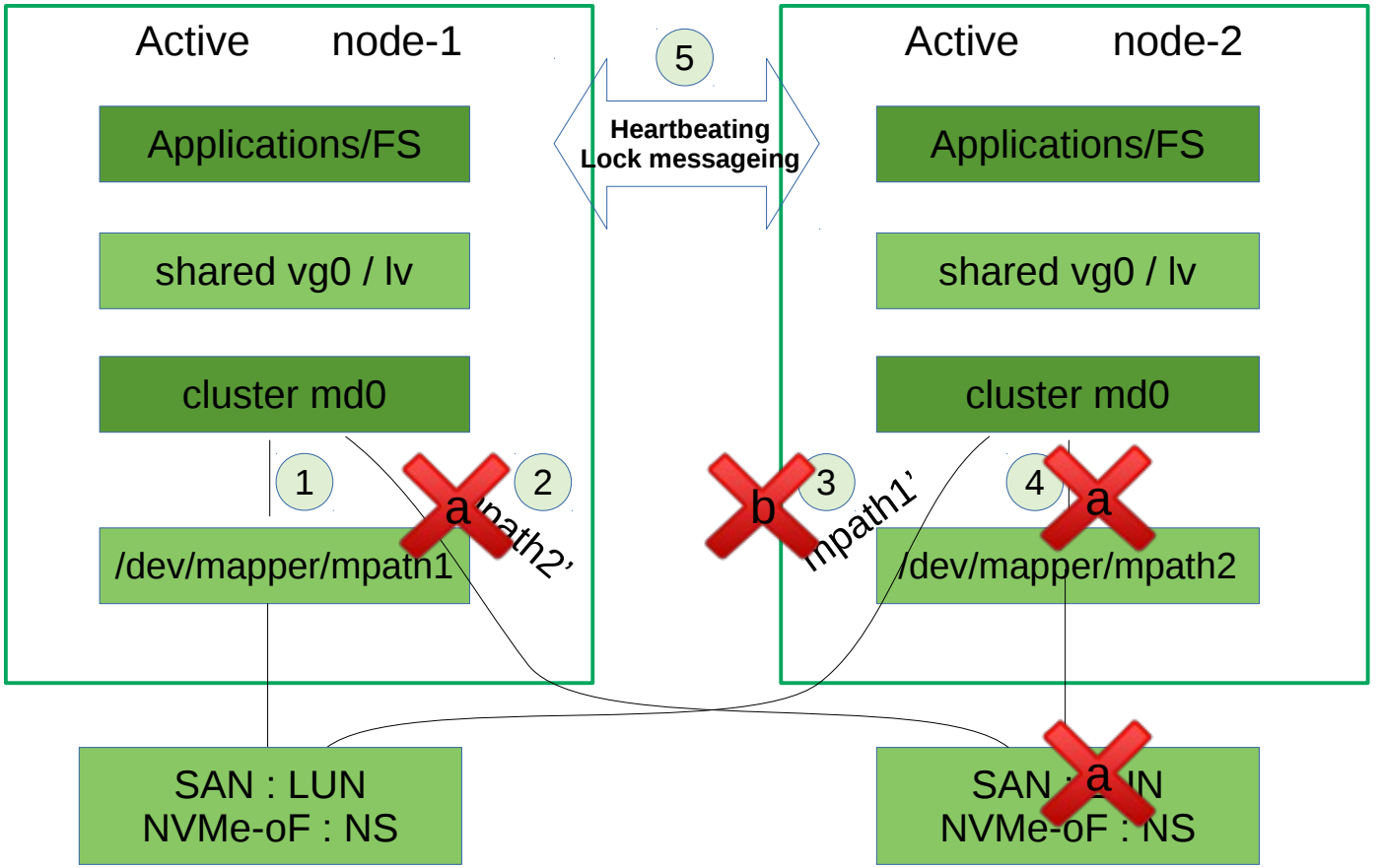
(*) the distributed lock play the game here.



Failure 4: SAN switch broken

same as

Failure 3:
SAN Partitioned



Now, you have Act-Act NVMe-oF in stretched cluster!

NVMe-oF in OpenStack

openstack™

- **Aug 2018, Rocky release**

  - Nova:

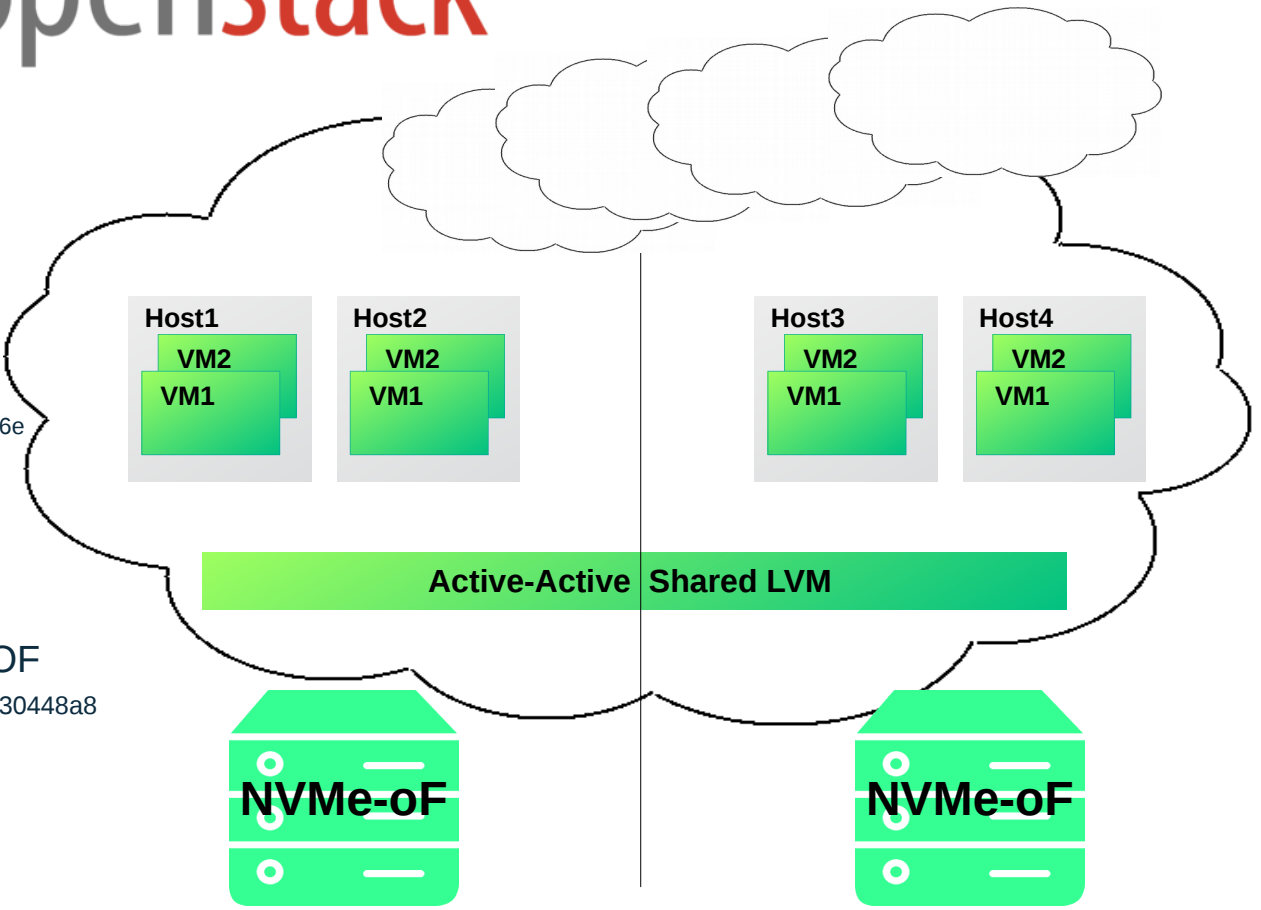
    Adding NVMEoF libvirt driver for supporting NVMEoF initiator CLI

    commit a833bcd05f811325f40cb3c8cce7f94c93cd6b6e
    Author: Rawan Herzallah <rawanh@mellanox.com>
    Date: Tue Jul 11 20:18:07 2017 +0300

  - Cinder:

    Adding NVMET target for NVMeOF

    commit d2b3e1011e238ce1c29157e0614a0416a30448a8
    Merge: f6cad8178 8d7e131c5
    Author: Zuul <zuul@review.openstack.org>
    Date: Wed May 9 22:01:16 2018 +0000



Let’s play with it !

Challenges ahead

- **Cluster RAID10**
- **Cluster RAID5**
- **Preferred site in case stretched SAN partitioned**

Welcome to join in Open Source!

SUSE 抽奖活动及规则介绍

参与方式：

① 扫描左侧二维码，关注 SUSE 官方微信；

② 发送“抽奖”至 SUSE 官方微信；

③ 简单填写信息后，进入幸运大转盘抽取礼品；

④ 凭中奖页面，前往 SUSE 展台领取礼品。