Sanlock for your shared SAN storage

- use case highlights
- and its consensus algorithms

Roger Zhou
2017-07
zzhou@suse.com

What is Sanlock
是什么东西？

是一个基于 SAN 的分布式锁管理器
A distributed lock manager based on SAN.

每个节点都各自运行
sanlock daemon runs on every node.

哪些前提假设
Assumptions ？

(1)

分布式锁不会频繁的获取和释放

Distributed Lock is not acquired and released very often.

(2)

可靠性要比集群中的主机高

是选择以共享存储作为 sanlock 服务的基础

It’s a truth. SAN is more reliable than Host Severs.

Use Case 1

Virtual Machine Disk mutual exclusion

- by default preferred by oVirt.
- a optional plugin of libvirt.

Use Case 2

to protect the metadata on the shared storage

an option of lvmlockd

Two Consensus Algorithms in Sanlock

锁的状态都被写到共享存储上

Lock state is written in the shared storage.

Terminology: Lease aka. Lock

Delta Lease

- aka. Lockspace.
- Purpose is to ensure host_id uniqueness, the prerequisite of paxos lease.
- The size is small.
- Can be very slow

Paxos Lease

- aka. resource lease.
- Purpose is to protect a resource
  … at the application level.
- Disk Paxos is the second Paxos variant out of 9.
- The size is costly.
- Fast speed for max of 2000 nodes.

To use sanlock from an application

- Allocate shared storage(lun or file) for an application.
- The application uses this storage with libsanlock to create a lockspace and resources for itself.
- The application joins the lockspace when it starts.
- The application acquires and releases leases on resources.

Why implement sanlock fencing ?

WDMD
watchdog multiplexing daemon
(多路复用 watchdog)

how live nodes being fenced?

每个节点有一个 Paxos Lease ，
被别的节点 acquired 即完成 fence 动作 .
A special Paxos Lease for every node, as long as it is acquired by other node, it proves the node get fenced.

Like SBD poison pill?

sanlock fencing can be chained

Usually, as a last resort

Like pacemaker SBD ?

Failure Scenarios

What sanlockd will response
if APP get killed ?

===> all resource leases will be force released.

What sanlockd will response
if share storage not accessible ?

===> recovery mode in next slide

Recovery Mode

- Graceful shutdown app – SIGTERM
- Forced shutdown app – SIGKILL
- Host reset – watchdog

What wdmd will react
when sanlock daemon is died?
eg. because of sanlock rpm update.

===> reset host get reset, vm vanished and
impact service inside vm. BAD!

Find the Difference

- comparing to dlm/corosync?
- comparing to SBD?
- sanlock vs. Persistent Reservation?

REFERENCE

- 2014, https://www.ibm.com/developerworks/cn/linux/1404_zhouzs_sanlock/
- man sanlock

Backup

man sanlock

… lockspace

… resources