THE NEXT GENERATION
CERTAINTY IN SHARED STORAGE ENVIRONMENTS

Adam Litke - alitke@redhat.com
Senior Software Engineer - Red Hat
FOSDEM 2017 - 04 February 2017
AGENDA

• oVirt shared storage architecture
• 😱 MAYHEM!
• 🐮 Order.
• Examples
OVIRT SHARED STORAGE

- NFS
- iSCSI
- Gluster
- Ceph

VM: x 1000s
Engine: x 1
Storage: x 100s
OVIRT VOLUME

Metadata
- List
- Details
- Files

Lease

Data
STORAGE OPERATIONS

VM

Datapath

Metadata
STORAGE JOBS

- Encapsulate a singular storage operation on a host
- Engine selects a host and submits the job
- Host performs work asynchronously
- Engine polls host for job status
- Job status only available on host while job is active
THE WORLD IS A CHAOTIC PLACE

- Power outages
- Network outages
- Hardware failure
- Software bugs
FAILED HOST

[Diagram showing a failed host (Host A) with connections to Engine and Storage volumes (Vol 1, Vol 2, Vol 3).]
RESTORING ORDER

- Determine the status of outstanding storage jobs
- The answers to all questions must come from storage
- Wait or abort active jobs on unresponsive hosts
- Check final status of jobs that have ended
VOLUME LEASES

- Implemented using **Sanlock**
- Lockspace is on shared storage alongside the volumes
- A lease grants a host exclusive access to a volume
- Failing hosts will be fenced by Sanlock if they hold leases
VOLUME GENERATIONS

- Monotonically increasing value
- Stored in volume metadata area
- Changeable only while holding the volume lease
- Allows sequencing of storage jobs
STORAGE JOB STRUCTURE

1. Acquire volume lease
2. Validate volume generation
3. Do work
4. Increment volume generation
5. Release volume lease
EXAMPLE: NORMAL FLOW
SCHEDULE JOB
ACQUIRE VOLUME LEASE
VALIDATE VOLUME GENERATION

Engine

Host A

Host B

Storage

Vol 1

Vol 2

Vol 3
WRITE VOLUME
INCREMENT VOLUME GENERATION

Engine

Host A

Host B

Storage

Job

Vol 1

Vol 2

Vol 3

Lease

2
RELEASE VOLUME LEASE

Engine

Host A

Host B

Storage

Vol 1

Vol 2

Vol 3
DONE EVENT
SCENARIO: UNRESPONSIVE HOST
JOB SCHEDULED ON HOST A

Engine

Host A

Host B

Storage

Vol 1

Vol 2

Vol 3
HOST A BECOMES UNRESPONSIVE
VOLUME RECONNAISSANCE

- Special storage job that resolves volume status
- Checks if a job is running
- Option to use sanlock fencing to free the lease
- If volume is free, uses generation to check status
- Increments generation to preempt pending jobs
SELECT ANY AVAILABLE HOST
ACQUIRE OR REQUEST LEASE

Host B

Job
Lease
Vol 1

Job
Running

.................. or ..................

Host B

Job
Lease
Vol 1

..................................

...
COMPARE AND BUMP GENERATION

Host B

Job 1

Lease

1

2

! Job Failed

---------- or ----------

Host B

Job 1

Lease

1

2

! Job Succeeded
PREEMPTED JOB

Host A

Vol 1

Vol 2

Vol 3

Lease

Host B

Engine

Storage

2

1
FUTURE WORK

- Shared lease support
- Parallel job scheduling
- Integrate with VM leases
JOIN US!

- http://www.ovirt.org
- irc://irc.oftc.net/ovirt
- http://lists.ovirt.org/mailman/listinfo/devel
- http://lists.ovirt.org/mailman/listinfo/users
QUESTIONS?