oVirt

THE NEXT GENERATION

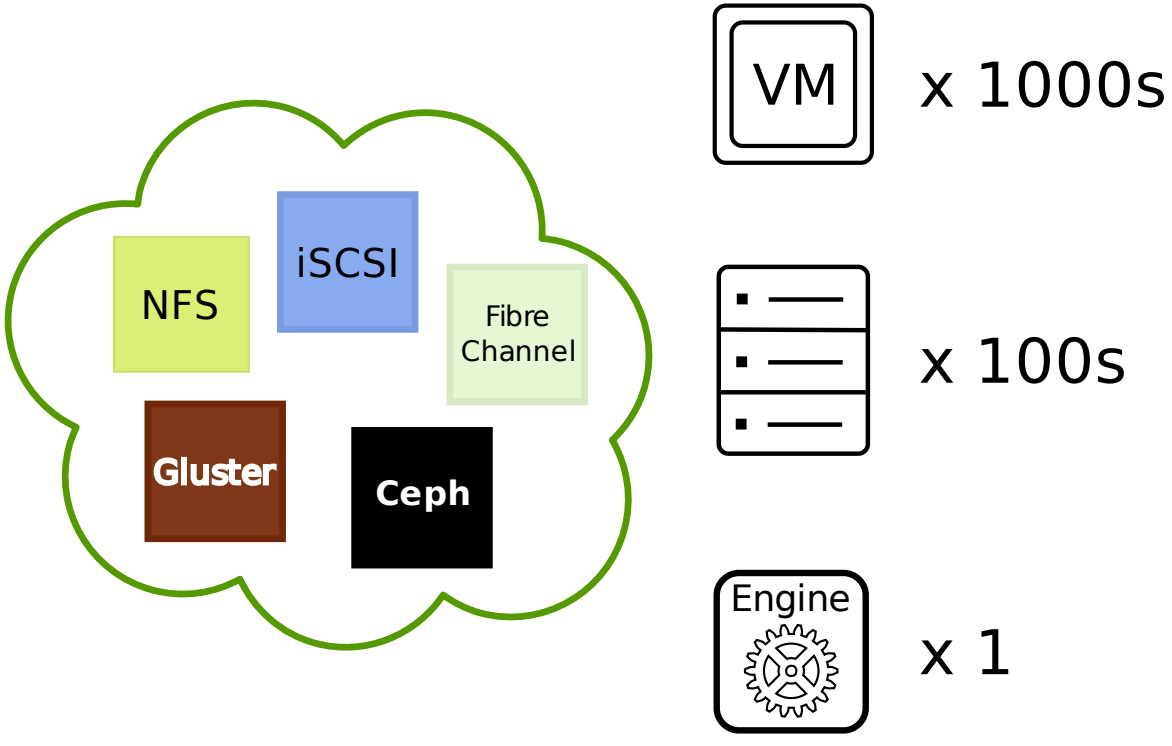
CERTAINTY IN SHARED STORAGE ENVIRONMENTS

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

AGENDA

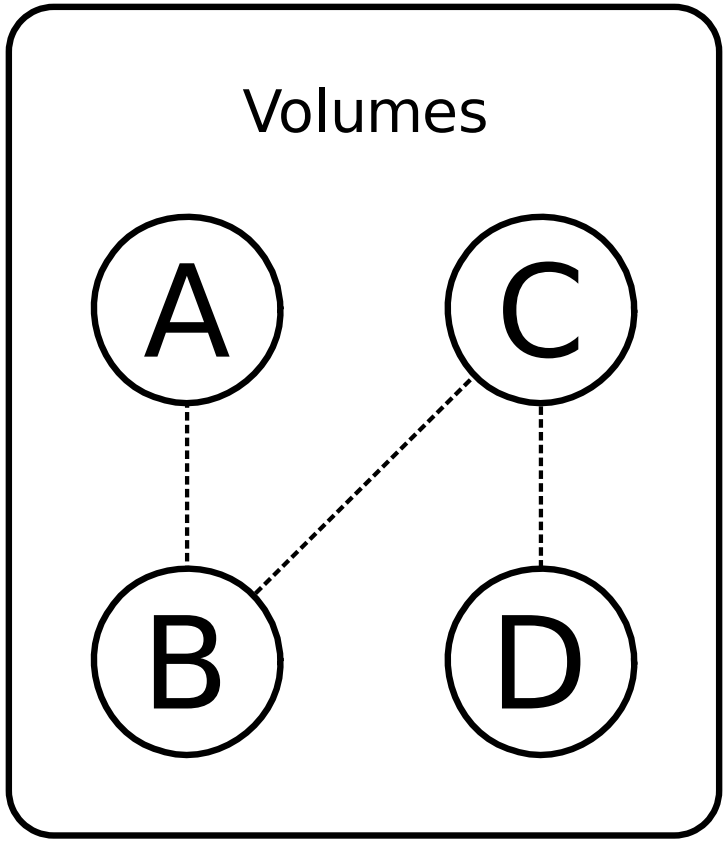
- oVirt shared storage architecture
- MAYHEM!
- Order.
- Examples

OVIRT SHARED STORAGE



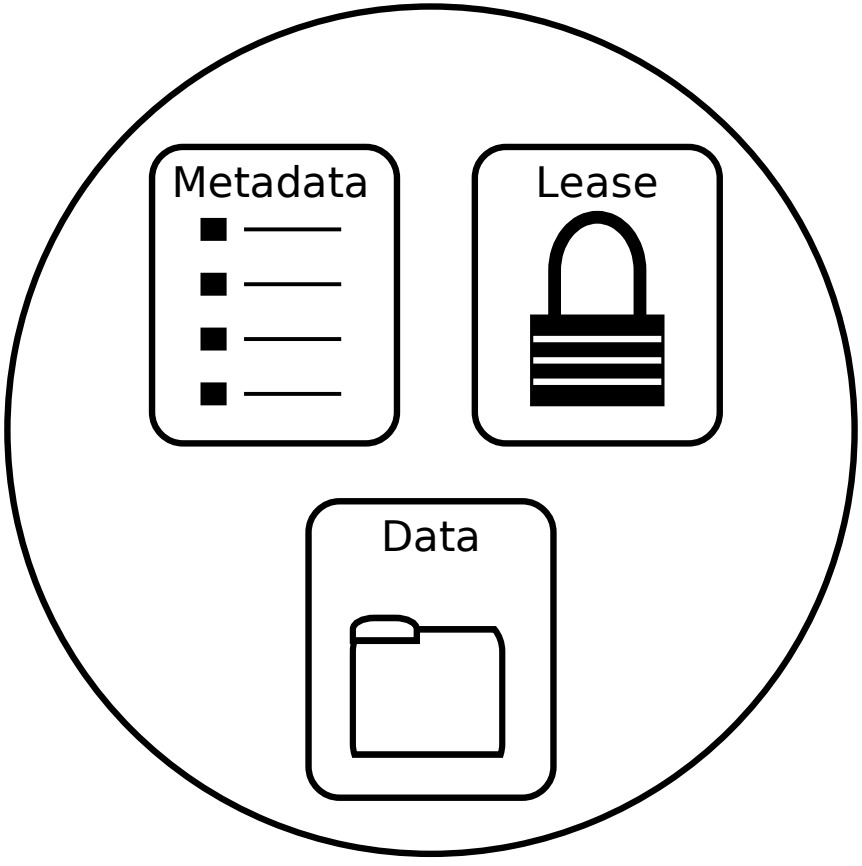
oVirt

OVIRT IMAGE



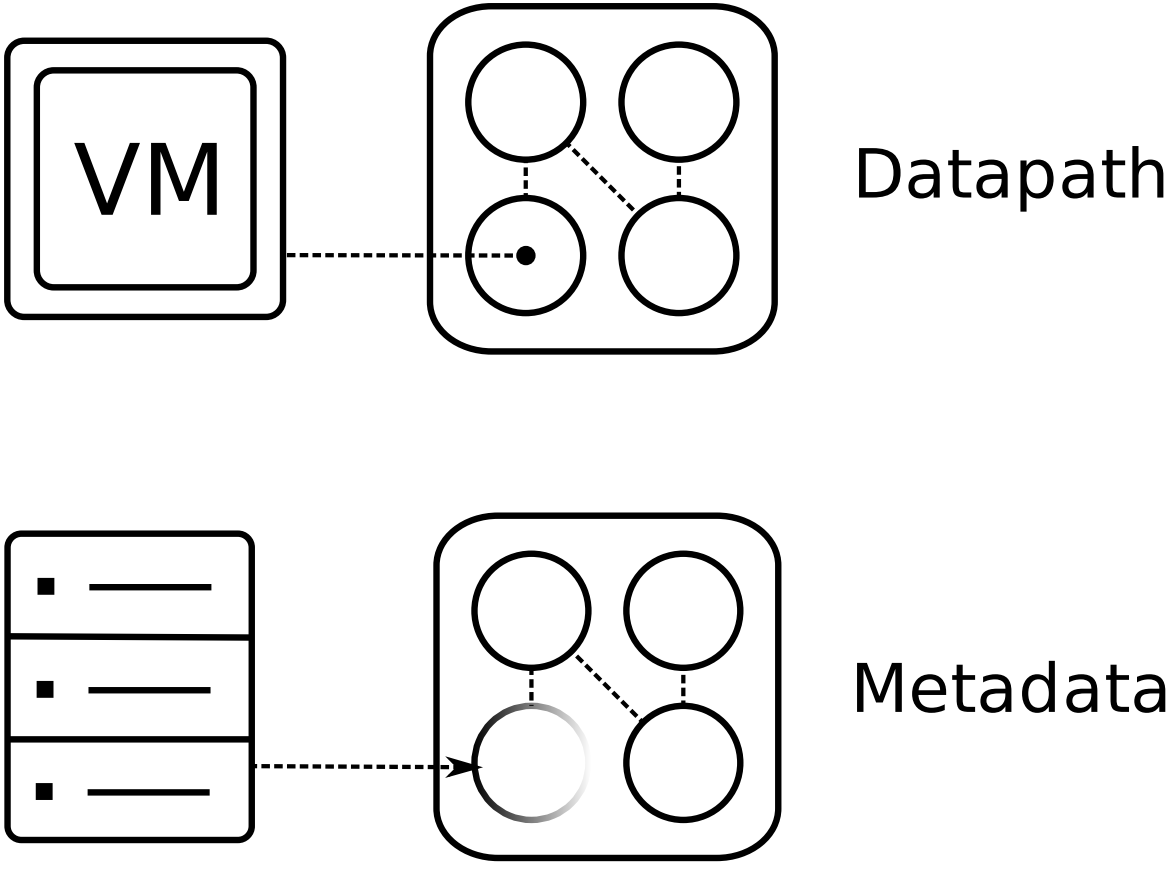
oVirt

OVIRT VOLUME



oVirt

STORAGE OPERATIONS



STORAGE JOBS

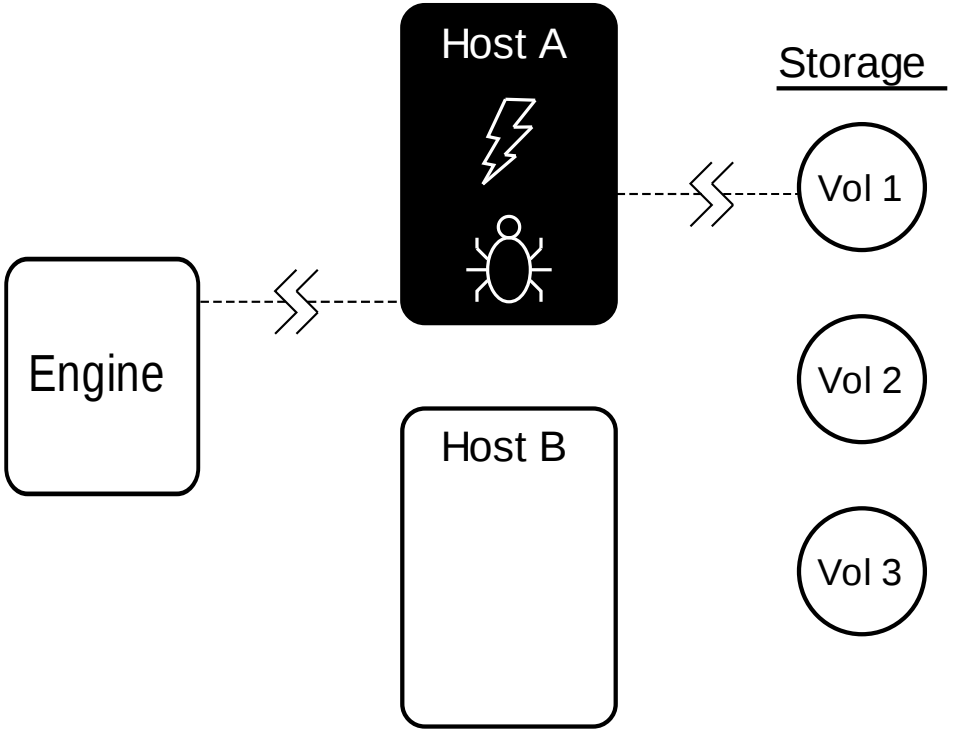
- Encapsulate a singular storage operation on a host
- Engine selects a host and submits the job
- Host performs work asynchronously
- Engine polls host for job status
- Job status only available on host while job is active

THE WORLD IS A CHAOTIC PLACE

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

oVirt

FAILED HOST

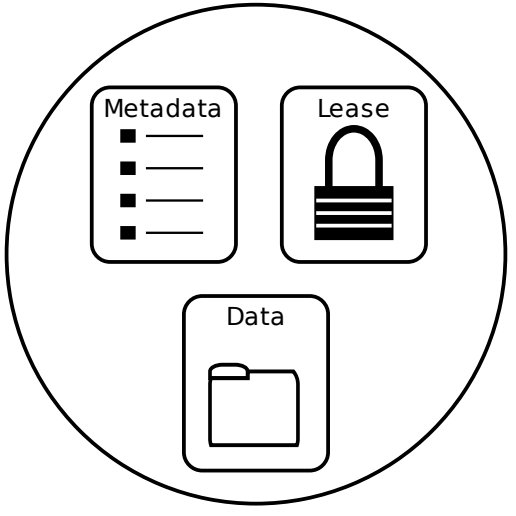


RESTORING ORDER

- 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

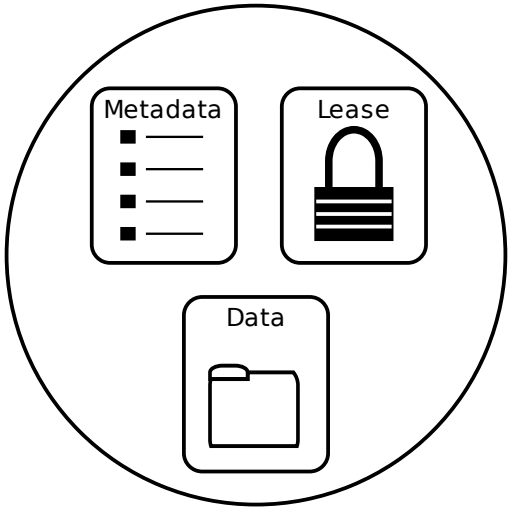
- Implemented using Sanlock
- Lockspace is on shared storage alongside the volumes
- A lease grants a host exclusive access to a volume
- Failing hosts will be fenced by Sanlock if they hold leases



VOLUME GENERATIONS

oVirt

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



STORAGE JOB STRUCTURE

oVirt

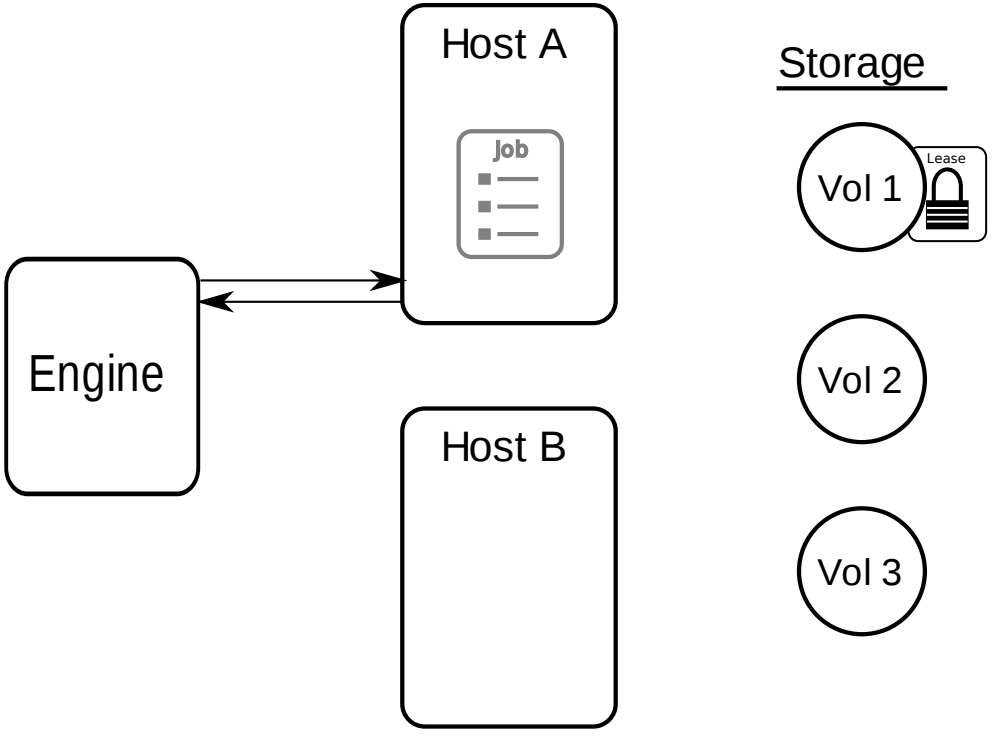
1. Acquire volume lease
2. Validate volume generation
3. Do work
4. Increment volume generation
5. Release volume lease

oVirt

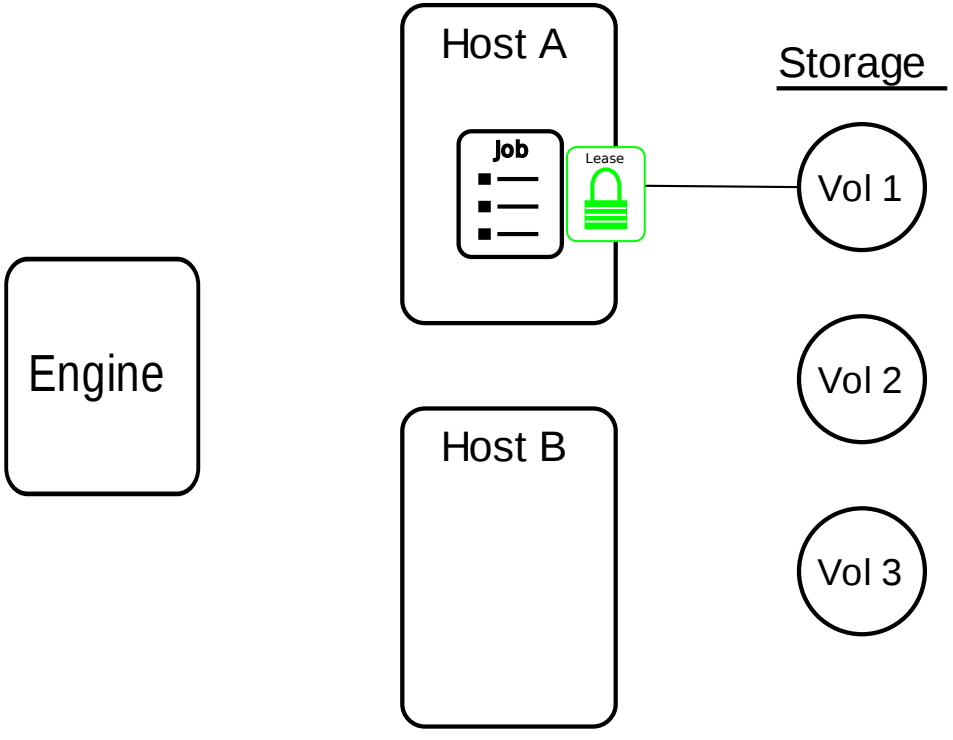
EXAMPLE: NORMAL FLOW

oVirt

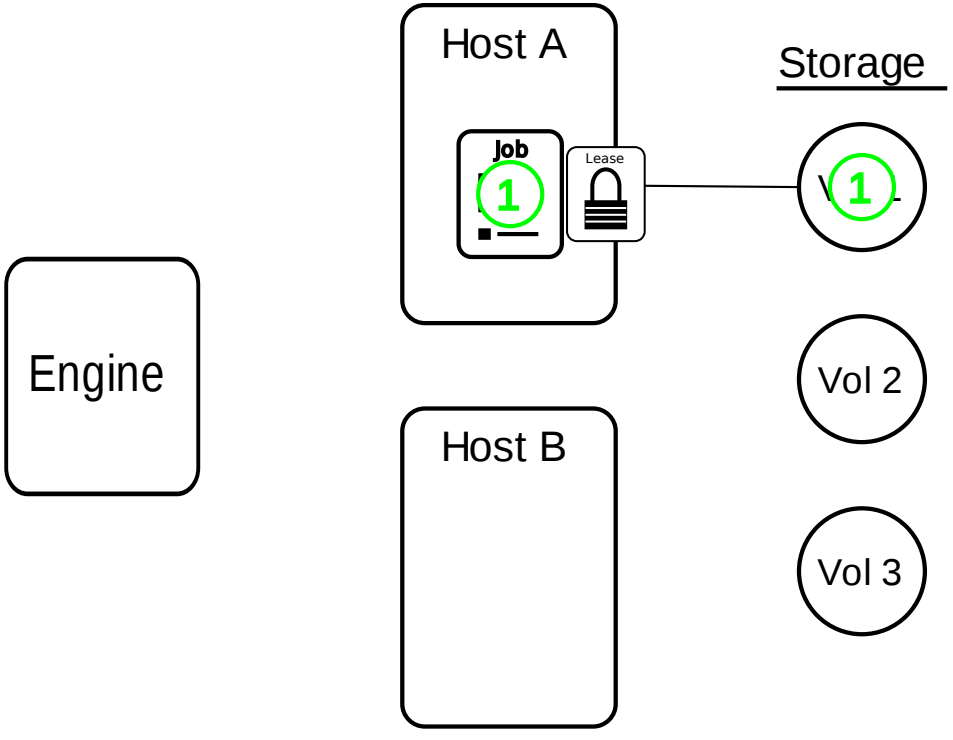
SCHEDULE JOB



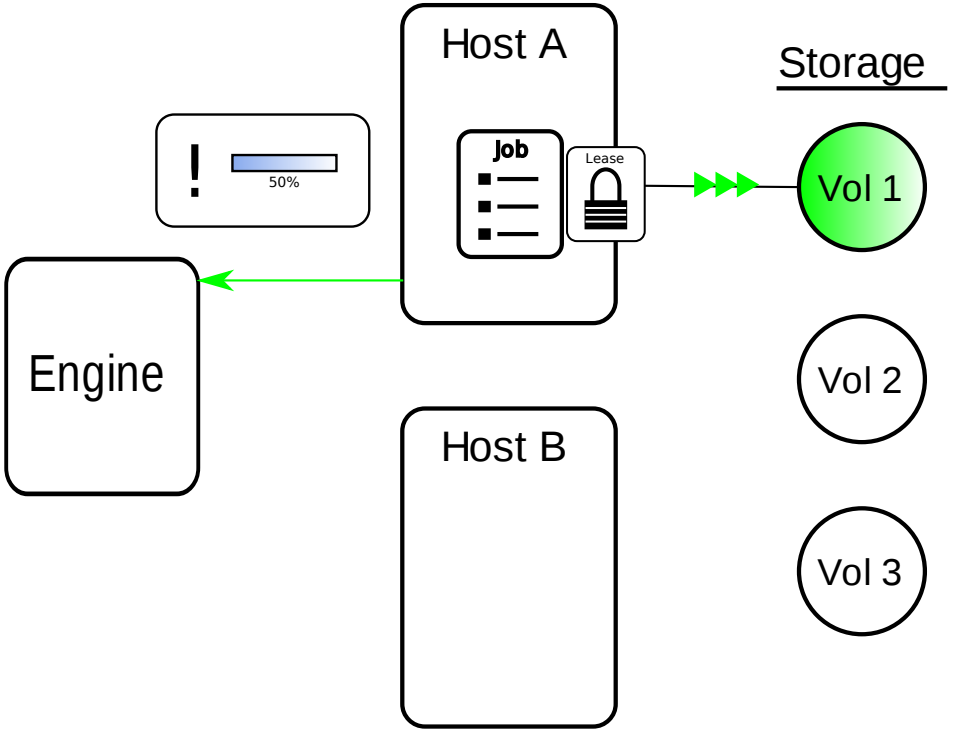
ACQUIRE VOLUME LEASE



VALIDATE VOLUME GENERATION

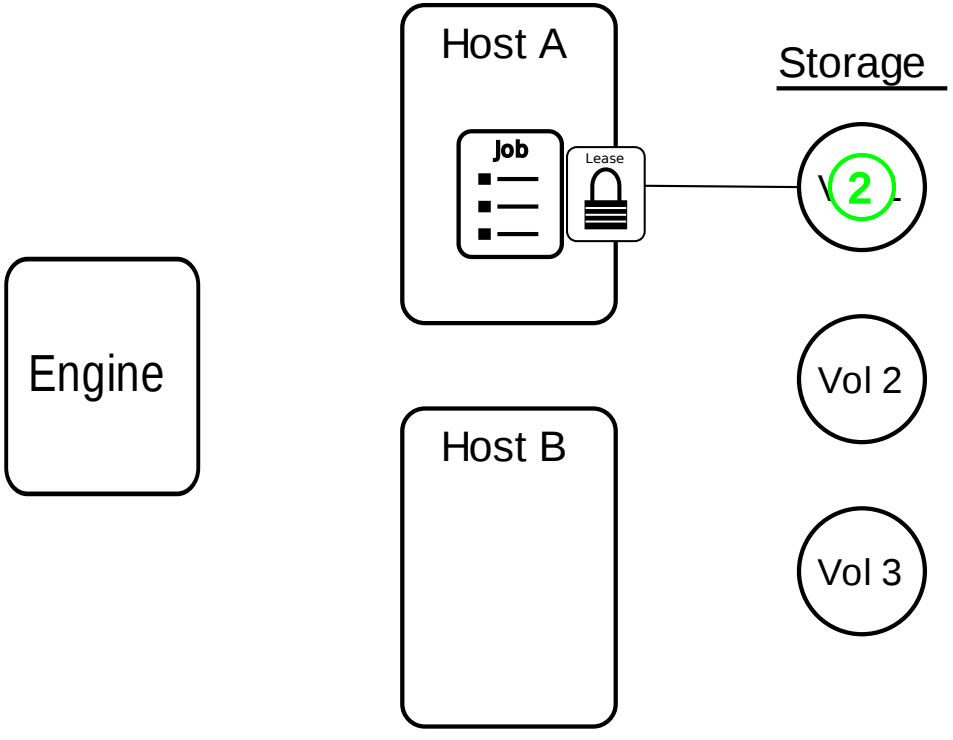


WRITE VOLUME

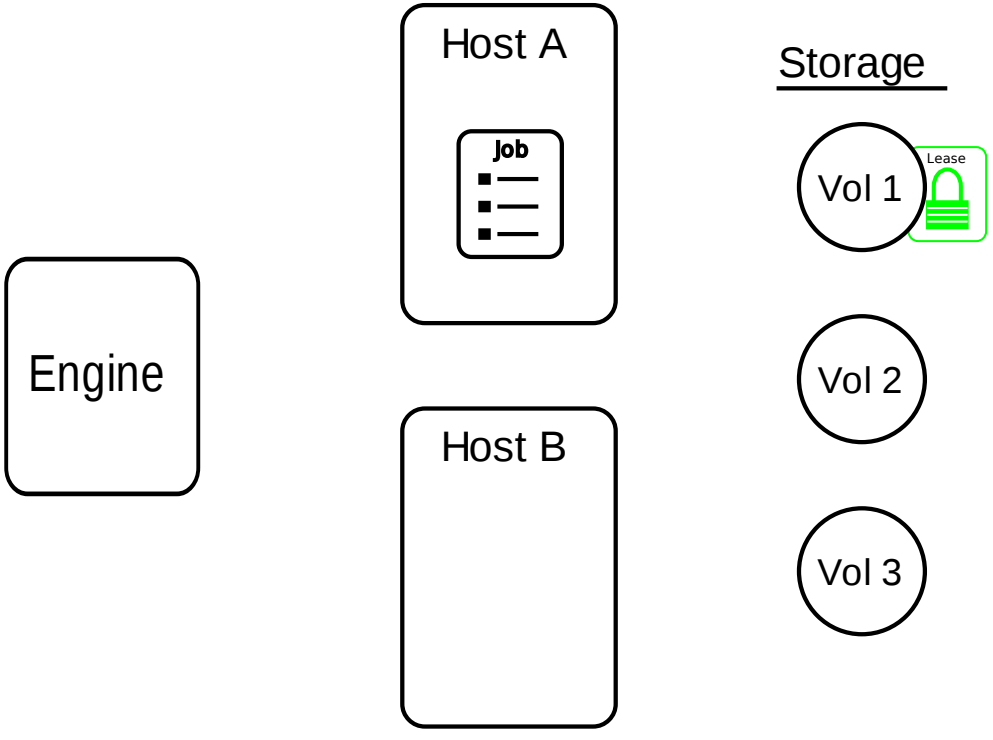


oVirt

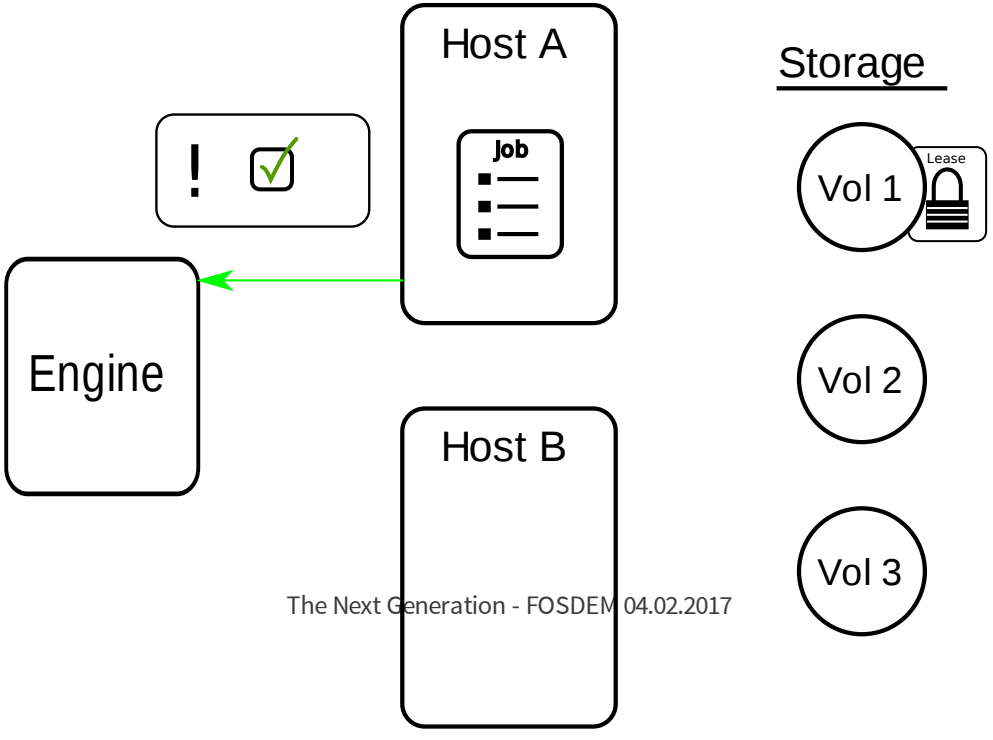
INCREMENT VOLUME GENERATION



RELEASE VOLUME LEASE



DONE EVENT



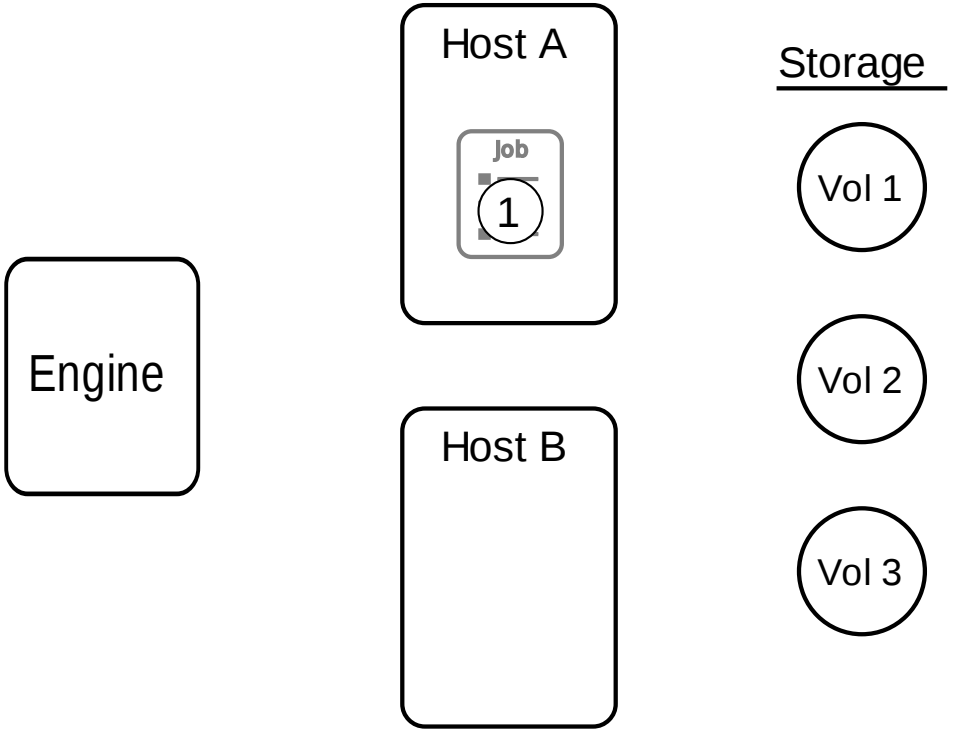
The Next Generation - FOSDEM 04.02.2017

oVirt

SCENARIO: UNRESPONSIVE HOST

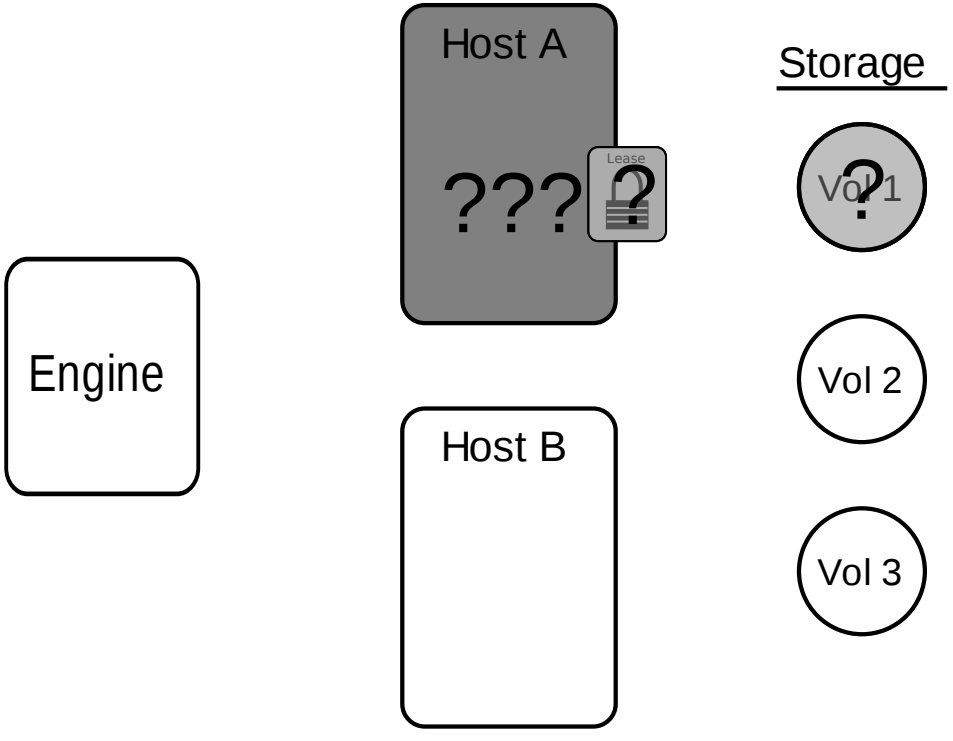
oVirt

JOB SCHEDULED ON HOST A



oVirt

HOST A BECOMES UNRESPONSIVE

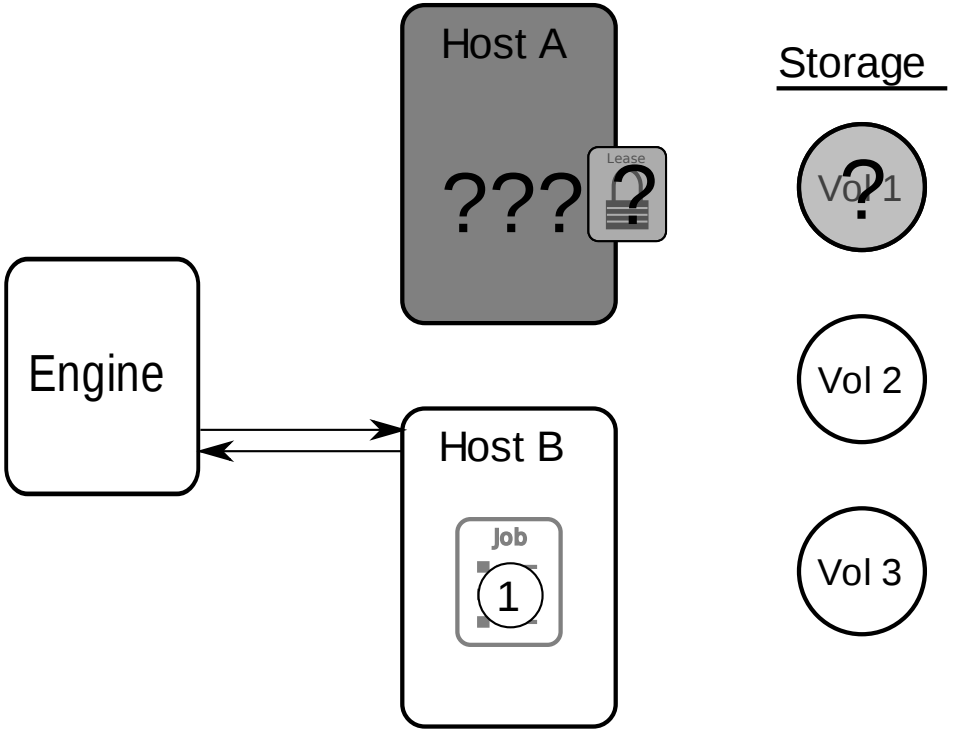


VOLUME RECONNAISSANCE

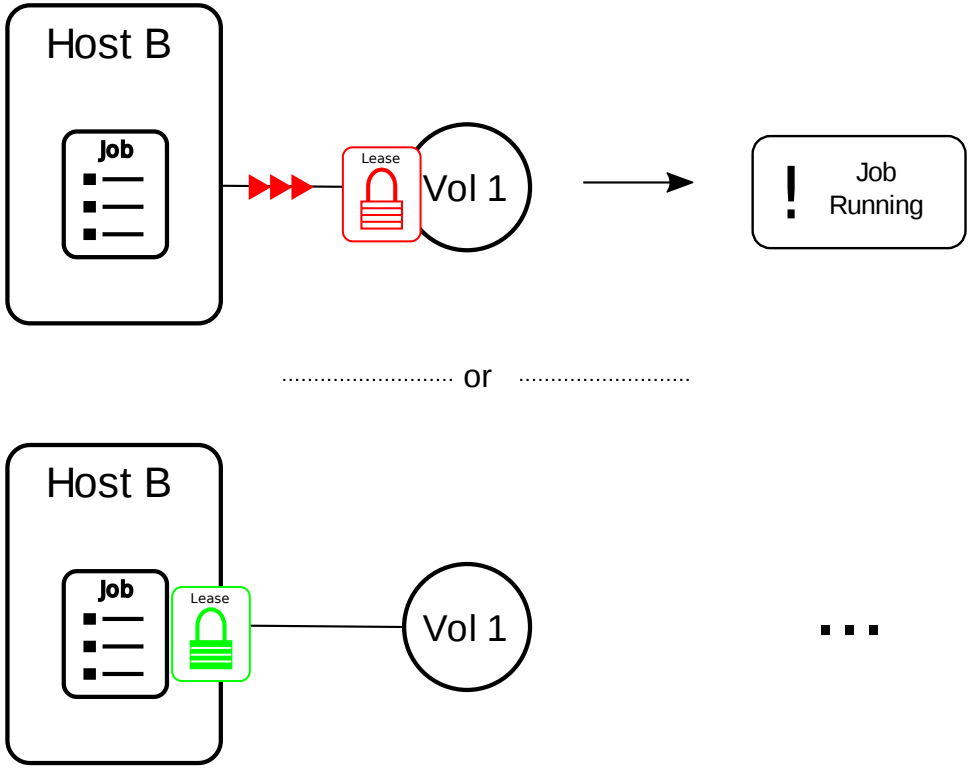
- Special storage job that resolves volume status
- Checks if a job is running
- Option to use sanlock fencing to free the lease
- If volume is free, uses generation to check status
- Increments generation to preempt pending jobs

oVirt

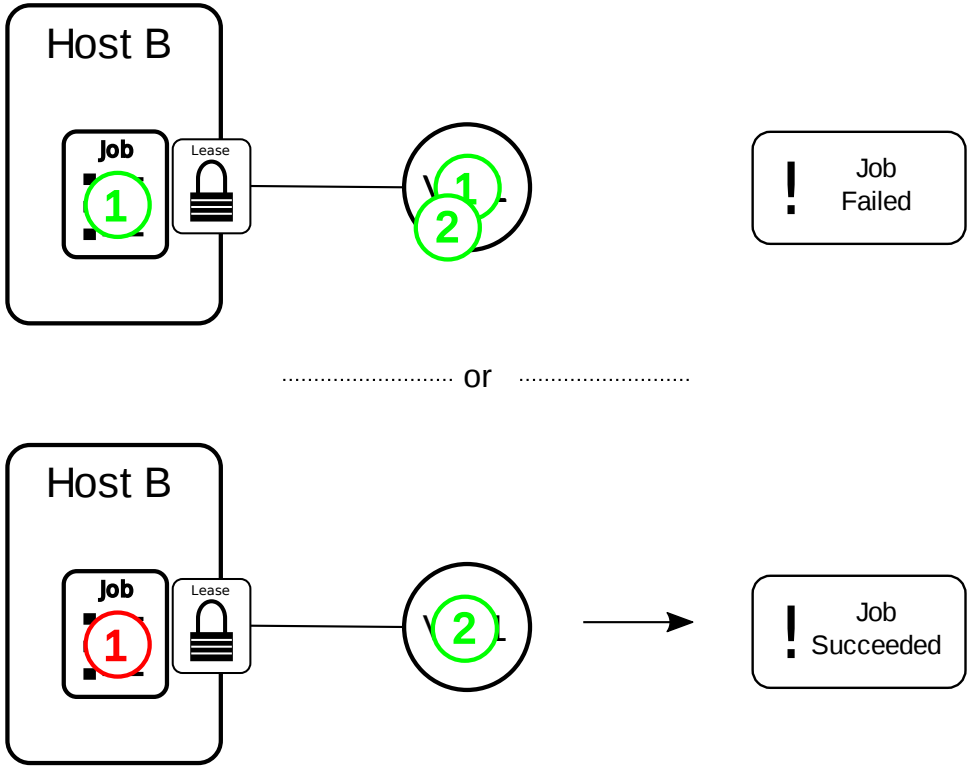
SELECT ANY AVAILABLE HOST



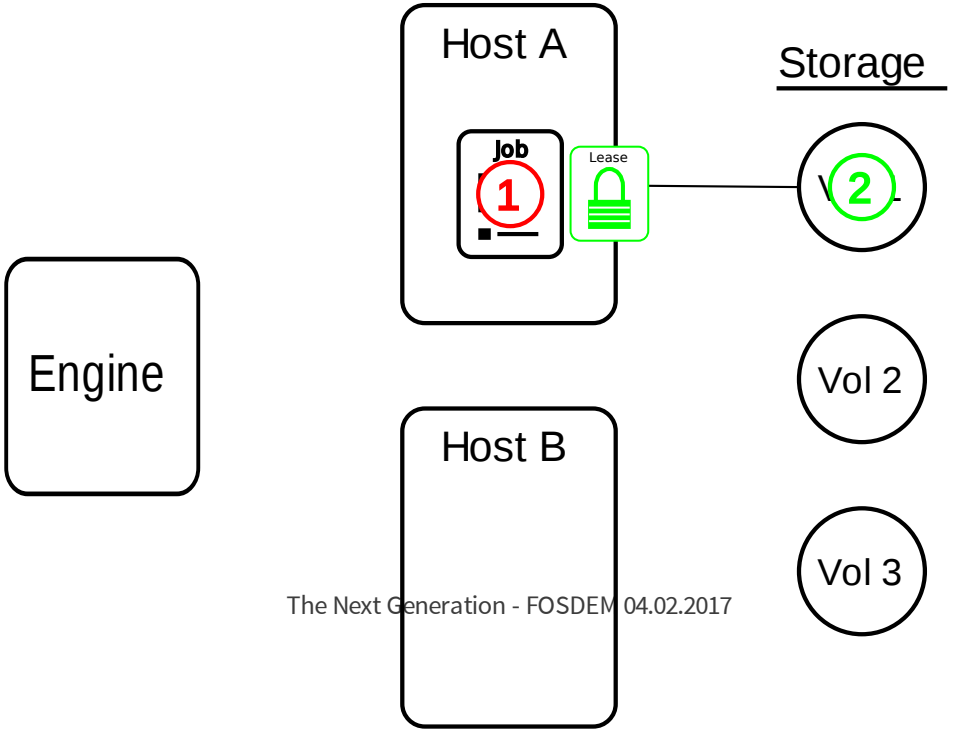
ACQUIRE OR REQUEST LEASE



COMPARE AND BUMP GENERATION



PREEMPTED JOB



The Next Generation - FOSDEM 04.02.2017

FUTURE WORK

- Shared lease support
- Parallel job scheduling
- Integrate with VM leases

oVirt

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

oVirt

QUESTIONS?