High Availability with No Split Brains!

Arik Hadas
Principal Software Engineer
Red Hat
27/01/2018
Virtual Data Center – Physical Servers
Virtual Data Center – Virtual Machines
Virtual Data Center - Applications
Some Applications are More Critical
High Availability - Application-Level
• Higher resource consumption
• More responsibility on the application
• Backup starts in a different environment
  – Different IP address(es)
  – Different disk(s)
High Availability - VM-Level
High Availability - VM-Level

- More efficient resource consumption
- Implemented at the infrastructure level
- VM always start in the same environment
  - Same IP address(es)
  - Same disk(s)
Fault Detection

HA VM went down!
Automatic Restart

- Restart the VM

DevConf.cz, January 2018
What if:
- Inaccessible resources
- VM is locked
- VM is being intentionally shut down

Restart the VM
Automatic Restart – Not That Simple

What if:
- Inaccessible resources
- VM is locked
- VM is being intentionally shut down

AutoStartVmsRunner

AutoStartVmsRunner

- Lock VM
  - Should Restart?
    - Run
      - No More Tries

DevConf.cz, January 2018
Fault Detection – Even More Complex
Fault Detection – Even More Complex
Fault Detection – Even More Complex

Is the left server alive?
Fault Detection – Even More Complex

Is the HA VM running?
The server has been rebooted
Fault Detection – Manual Confirmation

Restart the VM

DevConf.cz, January 2018
Fault Detection – Manual Confirmation

- Slow
- Error-prone
  - Mistakes may lead to a split-brain
Split Brain of Virtual Machines

A scenario in which several instances of the same VM run simultaneously
Split Brain Due to a False Confirmation

May lead to data corruption!
Split Brains May Happen Due to Bugs

Only the right VM is reported
Split Brains May Happen Due to Bugs

- Restart the left VM
VM Leases: Our Solution to Split Brains
VM Leases: Our Solution to Split Brains

VM will not start while its lease exists.

DevConf.cz, January 2018
## VM Lease Creation

### General
- **Cluster**: Default
- **Data Center**: Default

### System
- **Template**: Blank (0)
- **Operating System**: Debian 7

### Initial Run
- **Instance Type**: Custom
- **Optimized for**: Server

### Console

### Host

### High Availability
- **Highly Available**: On
- **Target Storage Domain for VM Lease**: Default
- **Resume Behavior**: KILL

### Resource Allocation

### Boot Options

### Random Generator

### Custom Properties

### Icon

### Foreman/Satellite

### Affinity Labels

---

**DevConf.cz, January 2018**
VM Lease Creation

- Highly Available
- Target Storage Domain for VM Lease: Default
- Resume Behavior: KILL
"Create a VM Lease for VM X in storage domain Y"
VM Lease Creation

“Create a Lease X in lockspace Y”

“Create a VM Lease for VM X in storage domain Y”
VM Lease Creation

“Create a Lease X in lockspace Y”

“Create a VM Lease for VM X in storage domain Y”

“Path P to xleases volume and Lease offset O”
• Sanlock does not manage leases allocation
• Volume layout:

<table>
<thead>
<tr>
<th>lockspace</th>
<th>index</th>
<th>master lease</th>
<th>user lease 1</th>
<th>user lease 2</th>
<th>....</th>
</tr>
</thead>
</table>

• Same format in block and file storage
• **Deep Dive - VM leases** (youtube)
<domain type='kvm' id='6'>
  <name>fedora8</name>
  ... skipped ...
  <devices>
  ... skipped ...
  <lease>
    <lockspace>571184ae-79da-41fb-a3fb-c3117991abae</lockspace>
    <key>cbd783e4-45f8-4b51-93ca-4460d4dad772</key>
    <target path='rhev/data-center/mnt/10.35.1.90:_srv_Default/571184ae-79da-41fb-a3fb-c3117991abae/dom_md/xleases' offset='3145728'/>
  </lease>
  ... skipped ...
</domain>
Running a VM with a Lease

Acquires the Lease using Sanlock

Domain XML with Lease

Lease
Non-Responsive Host Treatment
Non-Responsive Host Treatment
Non-Responsive Host Treatment

60+ sec of grace period
Non-Responsive Host Treatment

Fence (power management)
Non-Responsive Host Treatment

Restart VMs with a Lease
(1) Non-Responsive Host + VM is Down

restart VMs with a Lease
(1) Non-Responsive Host + VM is Down

VM starts on another host
(2) Non-Responsive Host + VM is UP

Restart VMs with a Lease
(2) Non-Responsive Host + VM is UP

Restart VMs with a Lease
Disconnection From Storage Device
Disconnection From Storage Device (1)

(1) Lease expires
Disconnection From Storage Device (1)

(1) Lease expires

(2) VM is terminated
Disconnection From Storage Device (2)

(1) VM is paused

(2) Lease is released
Summary

- VM Lease – an important new element
  - Prevents split-brains
  - Enables automatic restart of unreported VMs
- Available since oVirt 4.1
  - Polished in oVirt 4.2
- Possible future enhancements:
  - May be used to restart paused VMs
  - Move together with the bootable disk
THANK YOU!

http://www.ovirt.org
ahadas@redhat.com
ahadas@irc.oftc.net#ovirt