oVirt

Hosted Engine Deep Dive

September 9, 2013

Sandro Bonazzola
Greg Padgett

Agenda

- Introduction
- Use Cases
- Main Players
  - Hosted engine setup intro and demo
  - oVirt Engine
  - Storage architecture
  - HA services introduction and demo
- Summary
- Q&A

Introduction

- What is it all about?
  - The ability to run the oVirt engine as a VM hosted in an oVirt environment
  - Making this VM highly available
- What is it good for?
  - Providing built in HA solution for the oVirt engine
  - Better use of your hardware – no need for a dedicated oVirt host

Introduction

oVirt

- How are we gonna do that?
  - Hosted engine setup, that initializes the hosted engine environment
  - HA services – making our engine VM highly available at all times

Use Cases

oVirt

- Fresh hosted engine environment built from scratch
- Moving current oVirt environment to a hosted environment
- Future thoughts
  - Providing built-in oVirt engine VM to be configured and easily used as the hosted engine
  - Flexibility to use HA infrastructure described here to deploy other services besides the oVirt engine, and make sure they are highly available

oVirt

Main Players

Hosted Engine Setup

oVirt

It's an otopi based setup utility which takes care of:

- Configuring VDSM
- Configuring libvirt
- Creating the management bridge
- Creating the dedicated storage domain
- Creating the VM
- Validating the engine liveliness
- Transferring the control of the VM to the HA daemon

Hosted Engine Setup

Can be invoked using:

- 'ovirt-hosted-engine-setup'
- 'hosted-engine --deploy'

Outputs:

- /etc/ovirt-hosted-engine/answers.conf
  - Used for deploying additional hosts
- /etc/ovirt-hosted-engine/hosted-engine.conf
  - Used by HA daemon
- /etc/ovirt-hosted-engine/vm.conf
  - Used by HA daemon, usable by 'vdsClient create'

Hosted Engine Setup

hosted-engine also provides:

- --deploy: run ovirt-hosted-engine deployment
- --vm-start: start VM on this host
- --vm-shutdown: gracefully shutdown the VM on this host
- --vm-poweroff: forcefully poweroff the VM on this host
- --vm-status: VM status according to the monitoring daemon
- --add-console-password=<password> : create a temporary password for vnc/spice connection
- --check-liveliness: checks liveliness page of engine
- --console: open the configured console using remote-viewer on localhost (requires X)

Setup Simulation

```
# hosted-engine --deploy
[ INFO  ] Stage: Initializing
          Continuing will configure this host for serving as hypervisor and create a VM where oVirt
Engine will be installed afterwards.
          Are you sure you want to continue? (Yes, No)[Yes]:
          It has been detected that this program is executed through an SSH connection without
using screen.
          Continuing with the installation may lead to broken installation if the network
connection fails.
          It is highly recommended to abort the installation and run it inside a screen session.
          Do you want to continue anyway? (Yes, No)[No]: yes
[ INFO  ] Generating a temporary VNC password.
[ INFO  ] Stage: Environment setup
          Log file: /var/log/ovirt-hosted-engine-setup/ovirt-hosted-engine-setup-20130902143844.log
          Version: otopi-1.1.0_master
          Configuration files: []
[ INFO  ] Hardware supports virtualization
[ INFO  ] Bridge ovirtmgmt already created
[ INFO  ] Stage: Environment packages setup
[ INFO  ] Stage: Programs detection
[ INFO  ] Stage: Environment setup
[ INFO  ] Stage: Environment customization

          --== STORAGE CONFIGURATION ==--

          During customization use CTRL-D to abort.
          Please specify the storage you would like to use (glusterfs, nfs)[nfs]:
          Please specify the full shared storage connection path to use (example: host:/path):
dellserver.home:/home/images
[ INFO  ] Installing on first host
          Please provide storage domain name [hosted_storage]:
          Local storage datacenter name [hosted_datacenter]:
```

Setup Simulation

```
          --== NETWORK CONFIGURATION ==--

          Please indicate a pingable gateway IP address: 192.168.1.1
          firewalld was detected on your computer, do you wish setup to configure it? (Yes, No)
[Yes]:
          Please indicate a nic to set ovirtmgmt bridge on: (em1) [em1]:

          --== VM CONFIGURATION ==--

          Please specify the device to boot the VM from (cdrom, disk, pxe) [cdrom]:
          Please specify path to installation media you would like to use [None]: /home/Fedora-19-
x86_64-DVD.iso
          Please specify the number of virtual CPUs for the VM [Defaults to minimum requirement:
2]:
          Please specify the disk size of the VM in GB [Defaults to minimum requirement: 25]:
          Please specify the memory size of the VM in MB [Defaults to minimum requirement: 4096]:
          Please specify the console type you would like to use to connect to the VM (vnc, spice)
[vnc]:

          --== HOSTED ENGINE CONFIGURATION ==--

          Enter the name which will be used to identify this host inside the Administrator Portal
[hosted_engine_1]:
          Enter 'admin@internal' user password that will be used for accessing the Administrator
Portal:
          Confirm 'admin@internal' user password:
          Please provide the FQDN for the engine you would like to use. This needs to match the
FQDN that you will use for the engine installation within the VM: ovirt.home
[ INFO  ] Stage: Setup validation
```

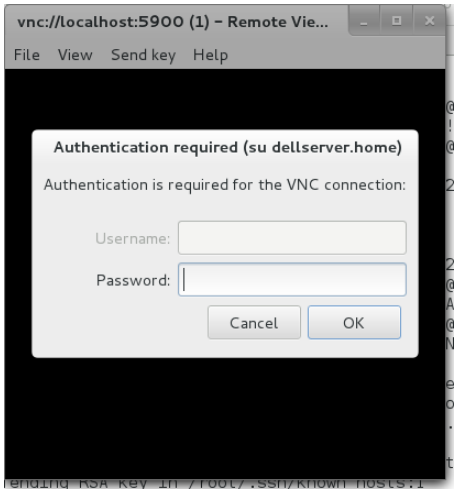
Setup Simulation

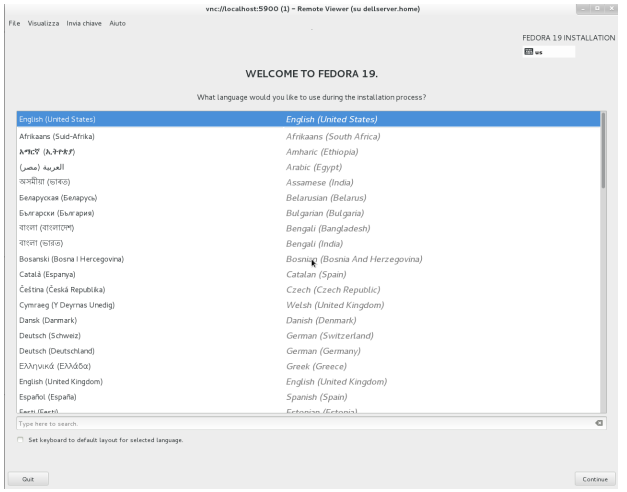
```
--== CONFIGURATION PREVIEW ==--

Bridge interface                   : em1
Engine FQDN                        : ovirt.home
Bridge name                        : ovirtmgmt
SSH daemon port                    : 22
Firewall manager                   : firewalld
Gateway address                    : 192.168.1.1
Host name for web application      : hosted_engine_1
Host ID                            : 1
Image size GB                      : 25
Storage connection                 : dellserver.home:/home/images
Number of CPUs                     : 2
Console type                       : vnc
Boot type                          : cdrom
Memory size MB                     : 4096
ISO image (for cdrom boot)         : /home/Fedora-19-x86_64-DVD.iso

Please confirm installation settings (Yes, No)[No]: yes
```

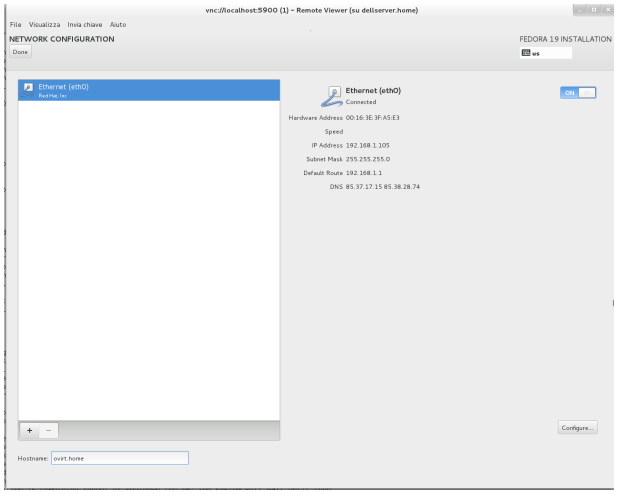
Setup Simulation

```
[ INFO  ] Generating answer file '/etc/ovirt-hosted-engine/answers.conf'
[ INFO  ] Stage: Transaction setup
[ INFO  ] Stage: Package installation
[ INFO  ] Stage: Misc configuration
[ INFO  ] Configuring the management bridge
[ INFO  ] Configuring VDSM
[ INFO  ] Generating VDSM certificates
[ INFO  ] Configuring libvirt
[ INFO  ] Starting vdsmd
[ INFO  ] Waiting for VDSM hardware info
[ INFO  ] Creating Storage Domain
[ INFO  ] Creating Storage Pool
[ INFO  ] Connecting Storage Pool
[ INFO  ] Verifying sanlock lockspace initialization
[ INFO  ] Initializing sanlock lockspace
[ INFO  ] Creating VM Image
[ INFO  ] Updating hosted-engine configuration
[ INFO  ] Disonnecting Storage Pool
[ INFO  ] Configuring VM
[ INFO  ] Stage: Transaction commit
[ INFO  ] Stage: Closing up
[ INFO  ] Creating VM
          You can now connect to the VM with the following command:
               /bin/remote-viewer vnc://localhost:5900
          Use temporary password "2615gIoD" to connect to vnc console.
          If you need to reboot the VM you can set a temporary password using the command:
          host-deploy --add-console-password=<password>
          Please install the OS on the VM.
          When the installation is completed reboot or shutdown the VM: the system will wait until
then
```

VM Installation







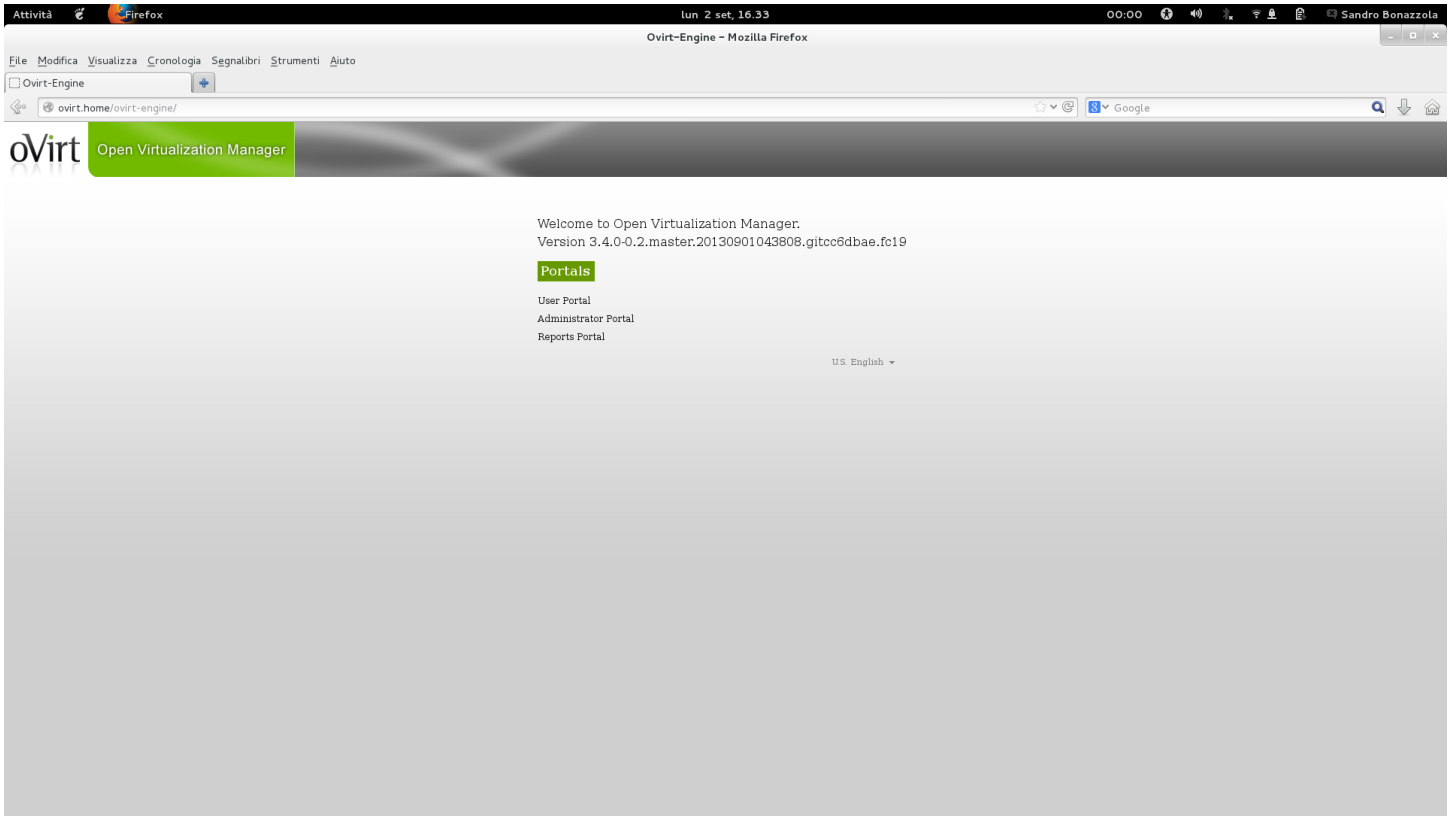
- Remember to use the same FQDN as host name!

Setup Simulation

```
          Has the OS installation been completed successfully?
          Answering no will allow you to reboot from the previously selected boot media. (Yes, No)
[Yes]:
[ INFO  ] Creating VM
          You can now connect to the VM with the following command:
               /bin/remote-viewer vnc://localhost:5900
          Use temporary password "2615gIoD" to connect to vnc console.
          If you need to reboot the VM you can set a temporary password using the command:
          host-deploy --add-console-password=<password>
          Please install the engine in the VM, hit enter when finished.
```

oVirt Engine Setup

- ssh to the VM
- yum localinstall http://ovirt.org/releases/ovirt-release-fedora.noarch.rpm
- yum -y install ovirt-engine
- engine-setup



Setup Simulation

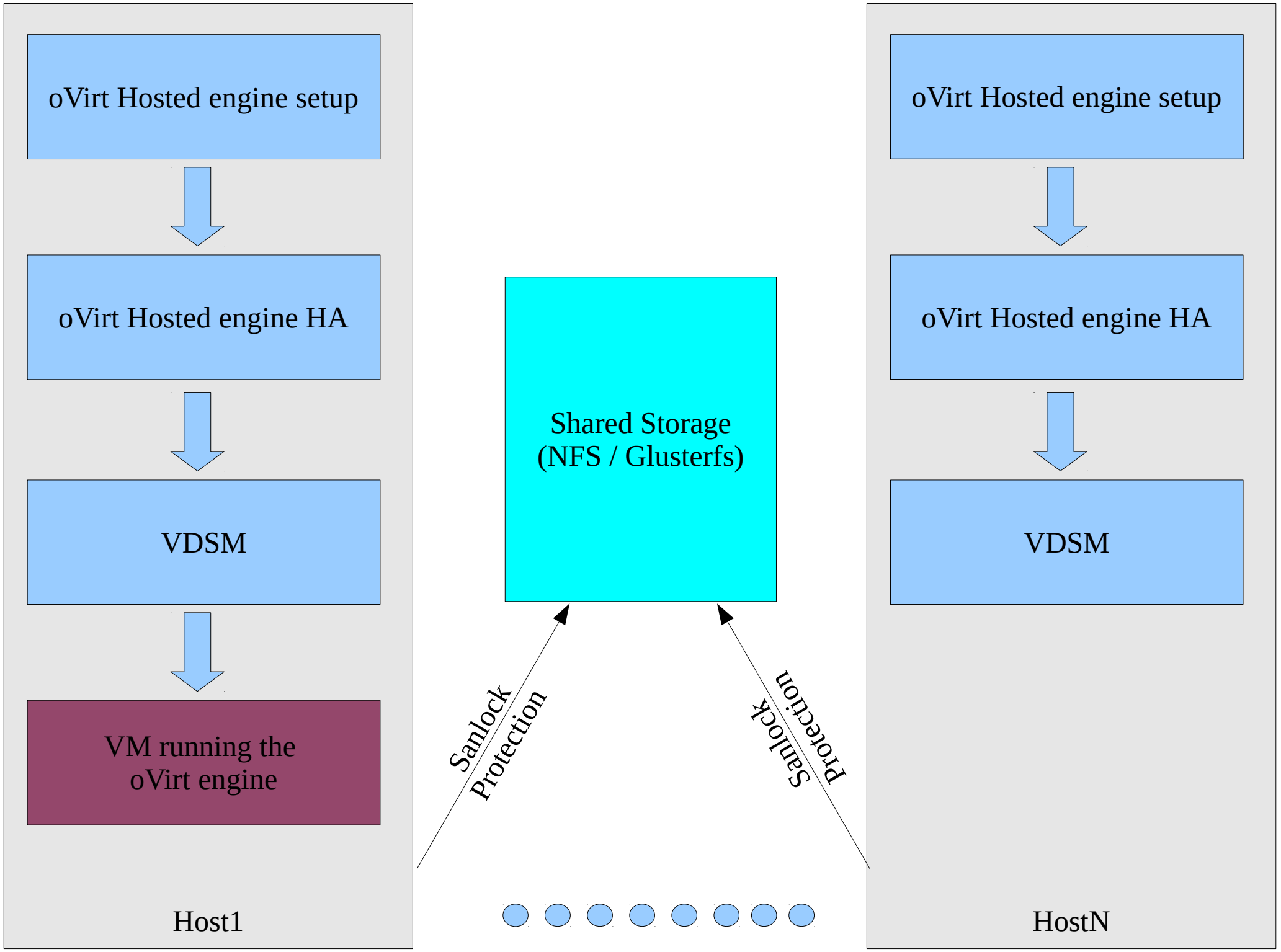
```
[ INFO  ] Creating VM
          You can now connect to the VM with the following command:
               /bin/remote-viewer vnc://localhost:5900
          Use temporary password "2615gIoD" to connect to vnc console.
          If you need to reboot the VM you can set a temporary password using the command:
          host-deploy --add-console-password=<password>
          Please install the engine in the VM, hit enter when finished.
[ INFO  ] Engine replied: DB Up!Welcome to Health Status!
[ INFO  ] Waiting for the host to become operational in the engine. This may take several
minutes...
[ INFO  ] The VDSM Host is now operational
          Please shutdown the VM allowing the system to launch it as a monitored service.
          The system will wait until the VM is down.
[ INFO  ] Enabling and starting HA services
          Hosted Engine successfully set up
[ INFO  ] Stage: Clean up
[ INFO  ] Stage: Pre-termination
[ INFO  ] Stage: Termination
```

Q&A

After the Installation...

oVirt

- Hosts are set up, what's next?
  - HA services manage high availability of engine
  - Sanlock used to synchronize activity between hosts
  - Shared storage used to pass state between hosts


oVirt Hosted engine setup
oVirt Hosted engine HA
VDSM
VM running the oVirt engine
Host1
Shared Storage (NFS / Glusterfs)
Sanlock Protection
Sanlock Protection
oVirt Hosted engine setup
oVirt Hosted engine HA
VDSM
HostN

What is Sanlock?

oVirt

- Sanlock daemon manages leases for applications running on a cluster of hosts with shared storage
- Lease management and coordination is done through reading and writing blocks on the shared storage

Sanlock Entities

oVirt

- Lockspaces – are slow to acquire and require regular I/O to shared storage
  - Sanlock uses Lockspaces internally to hold a lease on a Host ID
  - Host ID leases prevent two hosts from using the same Host ID and provide basic host liveness information based on the renewals
- Resources – are fast to acquire
  - Sanlock makes them available to applications as general purpose resource leases
  - Resources use Host ID's internally to indicate the owner of the lease

VDSM and Sanlock

oVirt

- VDSM is using Sanlock on Storage Domains since V3
  - Acquire and release the SPM Role (Resource)
- Storage Domains V3 are also providing Volume Leases (disabled by default)
  - Acquire and release exclusive/shared Sanlock Resources on Volumes
  - Prevent two different VMs from using the same Disk at the same time
  - Prevent the SPM from managing Volumes that are in use by a VM on another host
  - Requires Libvirt cooperation

VDSM and Sanlock

oVirt

- On connectStoragePool
  - VDSM acquires the Lockspaces using the Host ID on all the Storage Domains (V3) in order to acquire the SPM resource (when needed) and to run VMs
- On disconnectStoragePool
  - VDSM releases the Lockspaces on all the Storage Domains (V3)
- VDSM also monitors the Lockspaces re-acquiring them when needed (e.g. after a Storage Domain connectivity issue)

Libvirt and Sanlock

- VDSM according to some basic rules determine when a Volume Lease is requested (Exclusive/Shared) and uses the relevant XML to start the VM:

```
<disk device="disk" snapshot="no" type="block">
   <address bus="0" controller="0" target="0".../>
   <source dev=".../3812141a-5e41-4f98-ab07-4143160fdb7e"/>
   <driver cache="none" error_policy="stop" io="native".../>
</disk>
<lease>
   <key>3812141a-5e41-4f98-ab07-4143160fdb7e</key>
   <lockspace>d26915e8-9049-43a3-ba74-e403730875dc</lockspace>
   <target offset="115343360"
      path="/dev/d26915e8-9049-43a3-ba74-
e403730875dc/leases"/>
</lease>
```

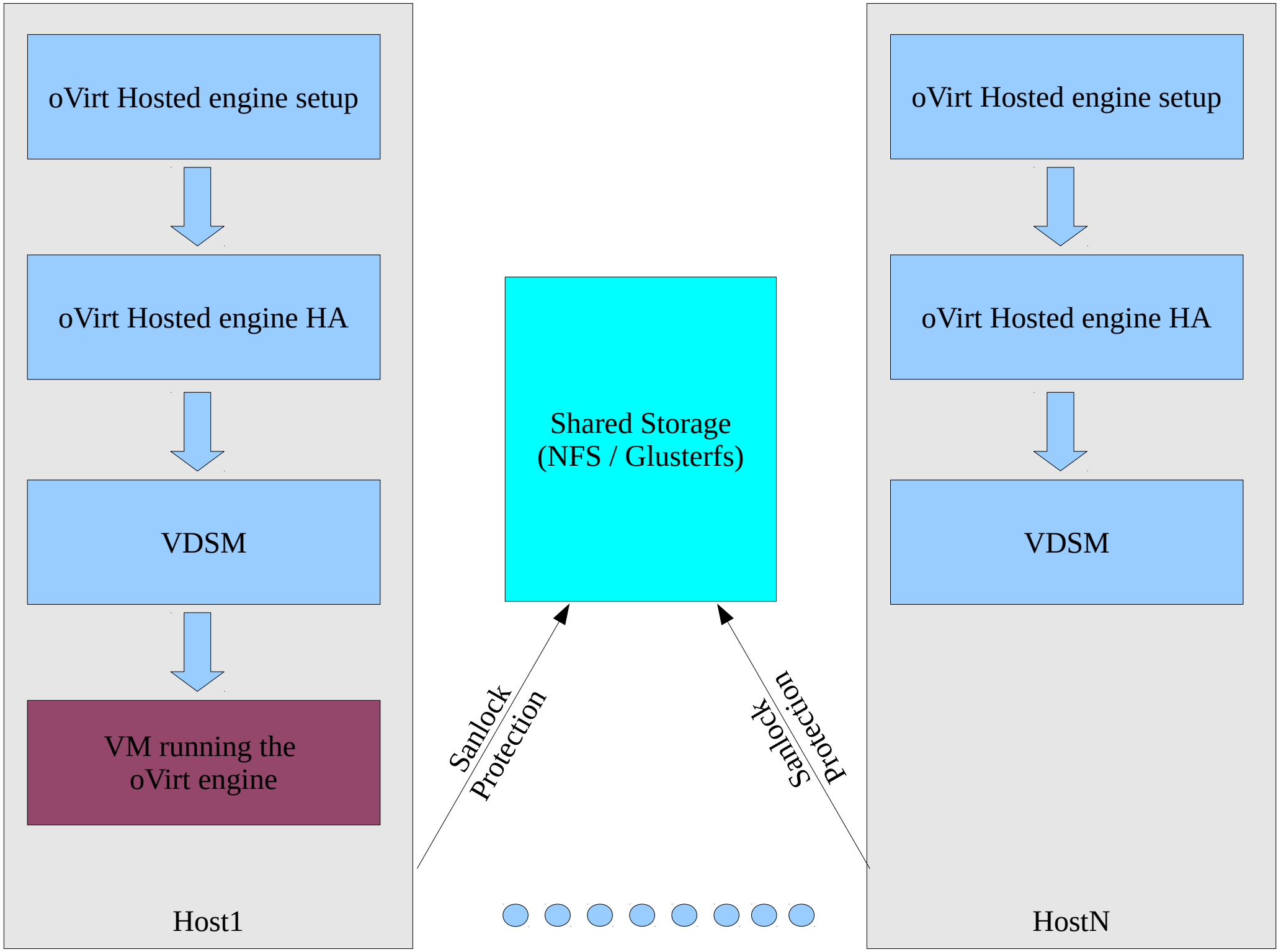
Libvirt and Sanlock

- At the moment VDSM blocks some advanced operations on Disks with enabled Volume Leases (e.g. Live Snapshots, Hotplug, Hotunplug, etc.)
  - These operations will be supported in the future with the relevant Hotplug/Hotunplug of Volume Leases
- Live Migration is supported and Libvirt is managing the Volume Leases handover between the source host and the destination host

VDSM Storage and Hosted Engine

oVirt

- New VDSM capabilities:
  - Run VMs without a Storage Pool
  - Run few selected VMs (Hosted Engine) with Volume Leases enabled (Sanlock protection)
  - Monitor Sanlock Lockspaces for Domains that are not in a Pool


oVirt Hosted engine setup
oVirt Hosted engine HA
VDSM
VM running the oVirt engine
Host1
Shared Storage (NFS / Glusterfs)
Sanlock Protection
Sanlock Protection
oVirt Hosted engine setup
oVirt Hosted engine HA
VDSM
HostN

Hosted Engine Services

oVirt

- 2 daemons
- ovirt-ha-agent
  - Monitors local host state, engine VM status
  - Takes action if needed to ensure high availability
- ovirt-ha-broker
  - Liason between ovirt-ha-agent and:
    - Shared storage (metadata)
    - Local host status (monitoring)
  - Serializes requests
  - Separate, testable entity distinct from ovirt-ha-agent

ovirt-ha-broker

- Used by ovirt-ha-agent to read to/write from storage
- Has set of monitors for host status:
  - Ping
  - Cpu load
  - Memory use
  - Management network bridge status
  - Engine VM status
- Listening socket:

  /var/run/ovirt-hosted-engine-ha/broker.socket

ovirt-ha-agent

oVirt

- Job is to ensure ovirt-engine VM high availability
- Uses configuration file written by setup
  - Host id, storage config, gateway address to monitor, ...

- If Engine VM is not running, it's started
- If Engine is non-responsive, VM is restarted

- VM status read from vdsm getVmStats verb
- Engine status via engine liveness page:

  http://<engine-fqdn>/OvirtEngineWeb/HealthStatus

ovirt-ha-agent

oVirt

- Main loop (every ~10 seconds):
  - Ensure connection to ovirt-ha-broker is good
  - Check vdsmd status, ensure storage is connected
  - Check sanlock status, ensure host_id lock is acquired
  - Read localhost monitors, write status to storage
  - Read all host statuses from storage, perform actions on engine vm if necessary
    - Decisions based only on what is in shared storage... helps us ensure that all hosts independently arrive at the same conclusion

ovirt-ha-agent

- Host Score
  - Single number representing a host's suitability for running the engine VM
  - Range is 0 (unsuitable) to 2400 (all is well)
    - May change
  - Calculated based on host status: each monitor (ping, cpu load, gateway status, ...) has a weight and contributes to the score

Score weights:
1000 - gateway address is pingable
800 - host's management network bridge is up
400 - host has 4GB of memory free to run the engine VM
100 - host's cpu load is less than 80% of capacity
100 - host's memory usage is less than 80% of capacity

Adjustments:
-50 - subtraction for each failed vm startup attempt
0 - score reset to 0 after 3 attempts, for 10 minutes

ovirt-ha-agent

oVirt

- If VM is starting, starts on host with highest score
  - If startup fails, host score is temporarily reduced, allows other hosts to try starting the VM
- If VM is running, and a different host has much higher score, VM is migrated to the better host
  - Current migration threshold is 800 points
  - E.g. gateway failure will trigger migration, cpu load won't

ovirt-ha-agent

oVirt

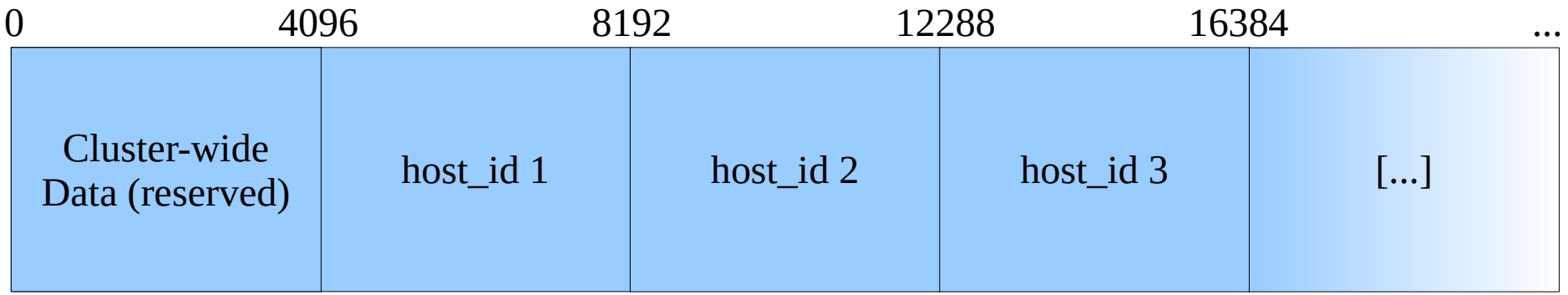
- Only live hosts are considered for startup/migration... if a host hasn't updated its metadata in a short while, it is considered dead
- VM startup algorithm is eager/optimistic: if two+ hosts have same score, both will try to start VM. Sanlock will allow only one to succeed
  - Race can also happen due to hosts not seeing metadata updates at the same time
- VM migration initiated in agent vs engine under discussion

Hosted Engine Storage

- Storage domain created during setup
  - First host only
  - Holds engine vm, sanlock metadata, agent metadata
  - NFS/GlusterFS only (support for iSCSI/FC coming later)
- Special files:
  - /rhev/data-center/mnt/<host:domain>/<uuid>/ha_agent/
  - [...] hosted-engine.lockspace – for sanlock
  - [...] hosted-engine.metadata – for agent
  - (both files created during setup)

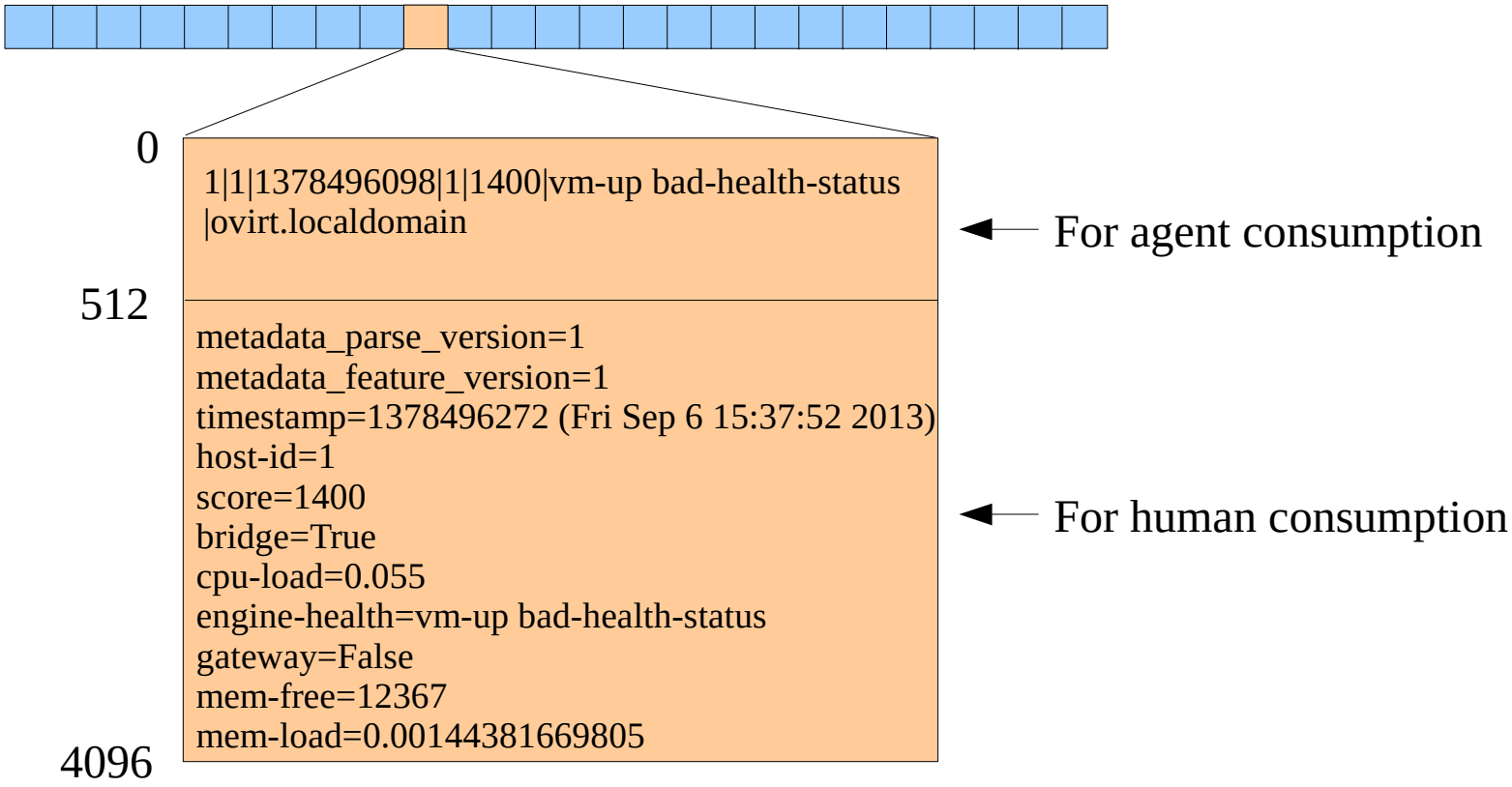
Hosted Engine Storage

- hosted-engine.metadata
  - 4KiB chunks, one per host
  - Chunk ownership defined by host_id (sanlock)
  - host_id starts at 1... offset 0 reserved for cluster-wide settings such as maintenance bit



Hosted Engine Storage

- hosted-engine.metadata: each 4KiB
  - First 512 bytes of chunks store critical data, atomic
  - Remaining space to assist in debugging



The oVirt Engine

- The oVirt engine will show the hosted engine VM as a special VM
- Valid operations on such a VM are
  - Migration (to one of the HA hosts)
  - Connect to console
- When moving a host to maintenance, the engine will make sure to migrate the hosted engine VM to another HA host

To-Do

oVirt

- Migration: initiated by engine or ha-agent?
- Global HA maintenance bit, utility to set/unset it
- Email alerts:
  - VM migration between hosts
  - VM startup failure
  - ... others?
- Block storage support (future release)
- CPU model and emulation level
- Resource reservations on participating hosts
- Moving vm configuration to shared storage
- Allowing to configure vm / hosts via engine
- Pre-installed engine image
- Adding ovirt-node support

oVirt

Hosted Engine Simulation

Hosted Engine Simulation

- Initial state: VM up on host 1, both hosts healthy

```
--== Host 1 status ==--

Hostname                           : ovirt.localdomain
Host ID                            : 1
Engine status                      : vm-up good-health-status
Score                              : 2400
Host timestamp                     : 1378510362
Extra metadata                     :
    timestamp=1378510362 (Fri Sep  6 19:32:42 2013)
    host-id=1
    score=2400
    engine-health=vm-up good-health-status
    gateway=True


--== Host 2 status ==--

Hostname                           : altovirt.localdomain
Host ID                            : 2
Engine status                      : vm-down
Score                              : 2400
Host timestamp                     : 1378510365
Extra metadata                     :
    timestamp=1378510365 (Fri Sep  6 19:32:45 2013)
    host-id=2
    score=2400
    engine-health=vm-down
    gateway=True
```

Hosted Engine Simulation

- Host 1's gateway down; VM migrated to host 2

```
--== Host 1 status ==--

Hostname                           : ovirt.localdomain
Host ID                            : 1
Engine status                      : vm-down
Score                              : 1400
Host timestamp                     : 1378510422
Extra metadata                     :
    timestamp=1378510422 (Fri Sep  6 19:33:42 2013)
    host-id=1
    score=1400
    engine-health=vm-down
    gateway=False


--== Host 2 status ==--

Hostname                           : altovirt.localdomain
Host ID                            : 2
Engine status                      : vm-up good-health-status
Score                              : 2400
Host timestamp                     : 1378510425
Extra metadata                     :
    timestamp=1378510425 (Fri Sep  6 19:33:45 2013)
    host-id=2
    score=2400
    engine-health=vm-up good-health-status
    gateway=True
```

Q&A

oVirt

THANK YOU !

http://www.ovirt.org

#ovirt (irc.oftc.net)