oVirt Hosted Engine

The Egg That Hosts its Parent Chicken

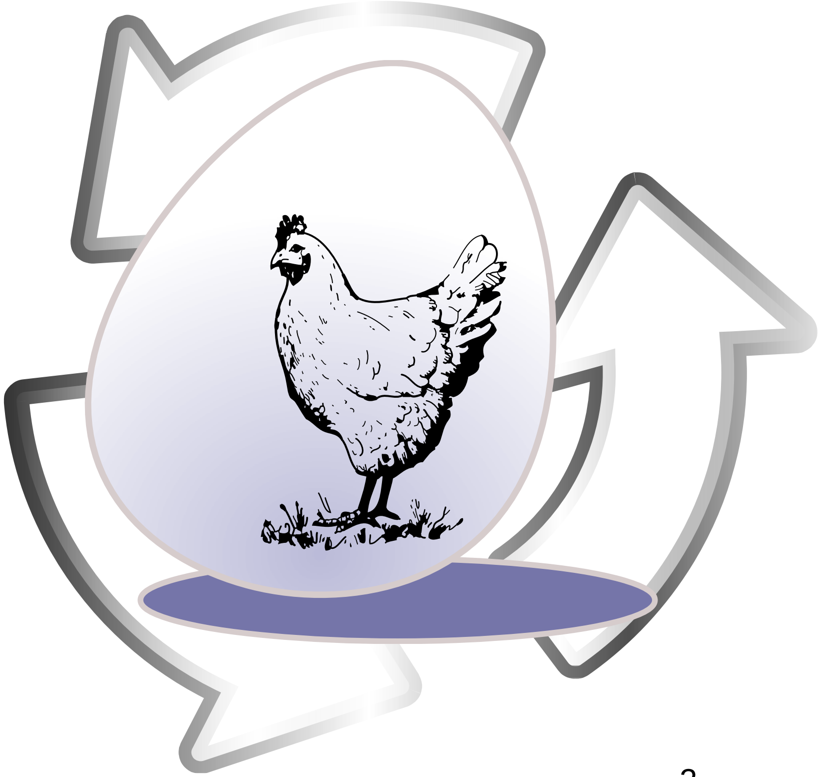
Doron Fediuck
Red Hat

KVM Forum
October 2013

Agenda

oVirt

- Fundamental question
- What is it?
- Why do we need it?
- Challenges
- Solutions
- Hosted engine architecture
- Hosted engine storage
- Simulations
- Summary



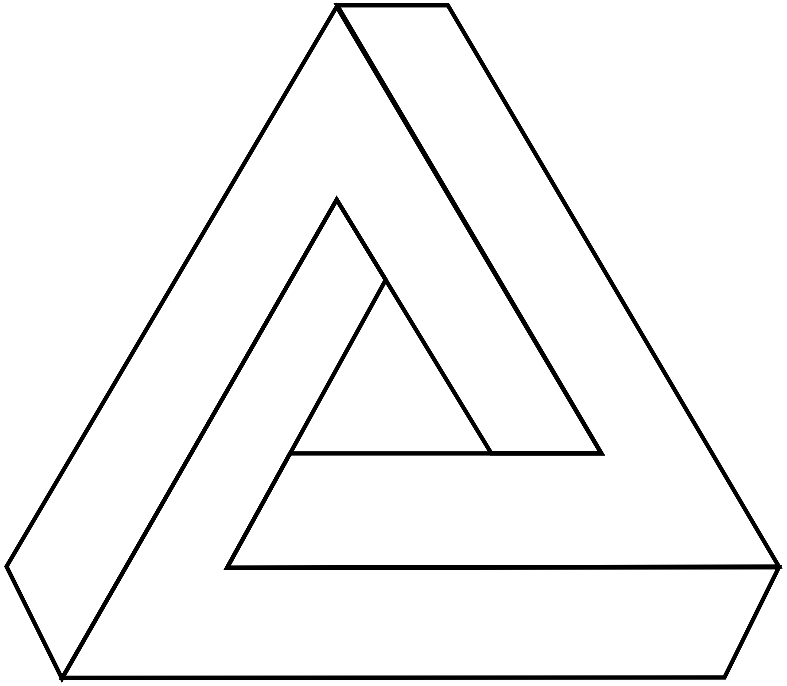
Fundamental question

Why did the chicken cross the road?

What is it?

- Standard oVirt installation
- Running in a highly available VM
- The VM is managed... by the engine it's hosting

- Sound challenging?...



Why do we need it?

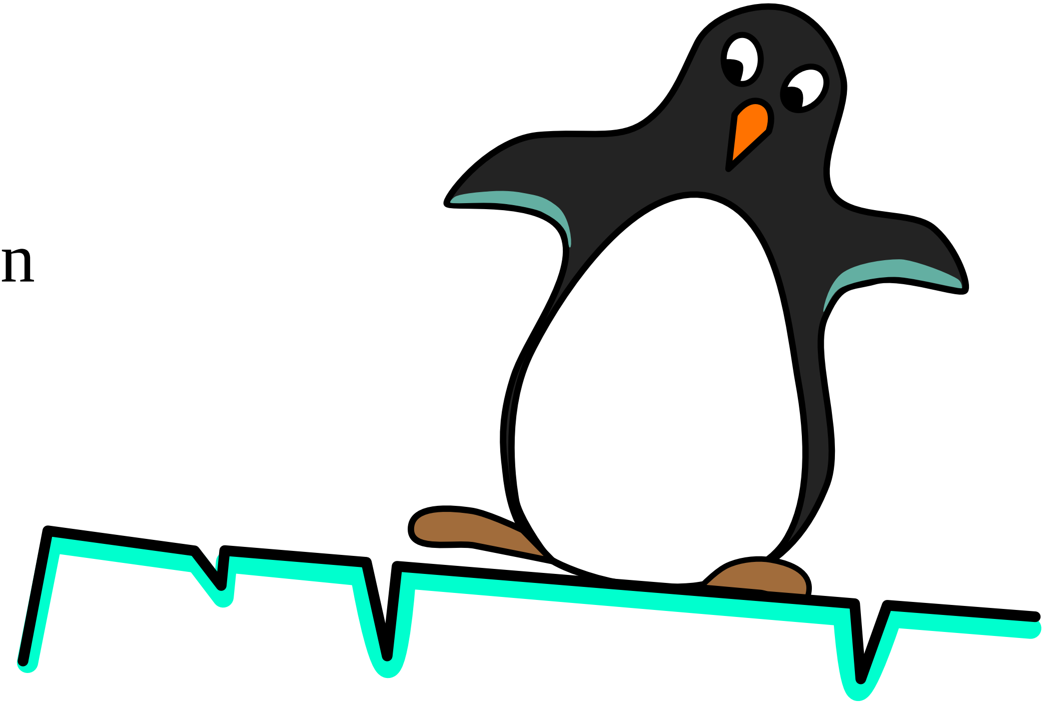
- Saves $ / £ / € / ₪ /...
  - No need for dedicated box

- Actually, saves $$$ / £££ / €€€ / ₪₪₪ /...
  - If you have a failover solution

Challenges

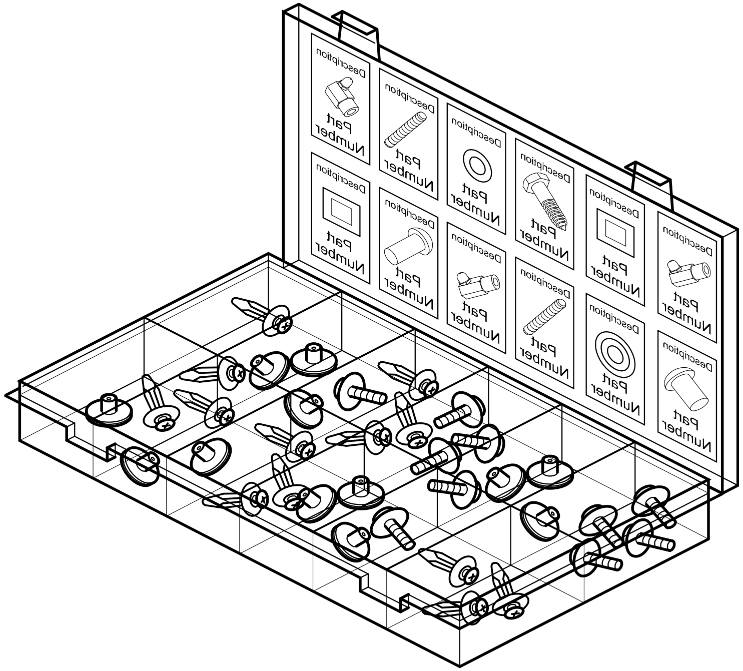
- Setup...
  - **How do we set up an egg (VM) that hosts its parent chicken (oVirt engine)**?

- VM availability
  - Network connectivity lost
  - Engine unexpectedly down
  - Load balancing
  - Maintenance
  - ...



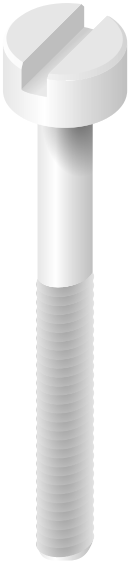
Solutions

- Existing solutions
  - Clustering File system + file locking
    - Proprietary
  - RHCS / Pacemaker
    - Standard file system
    - Uses Corosync
    - Limits number of nodes
    - No oVirt node support



Solutions

- Here's a thought
  - Standard file system
  - Sanlock leases
- Simpler
- Focused on VMs
- Less logic



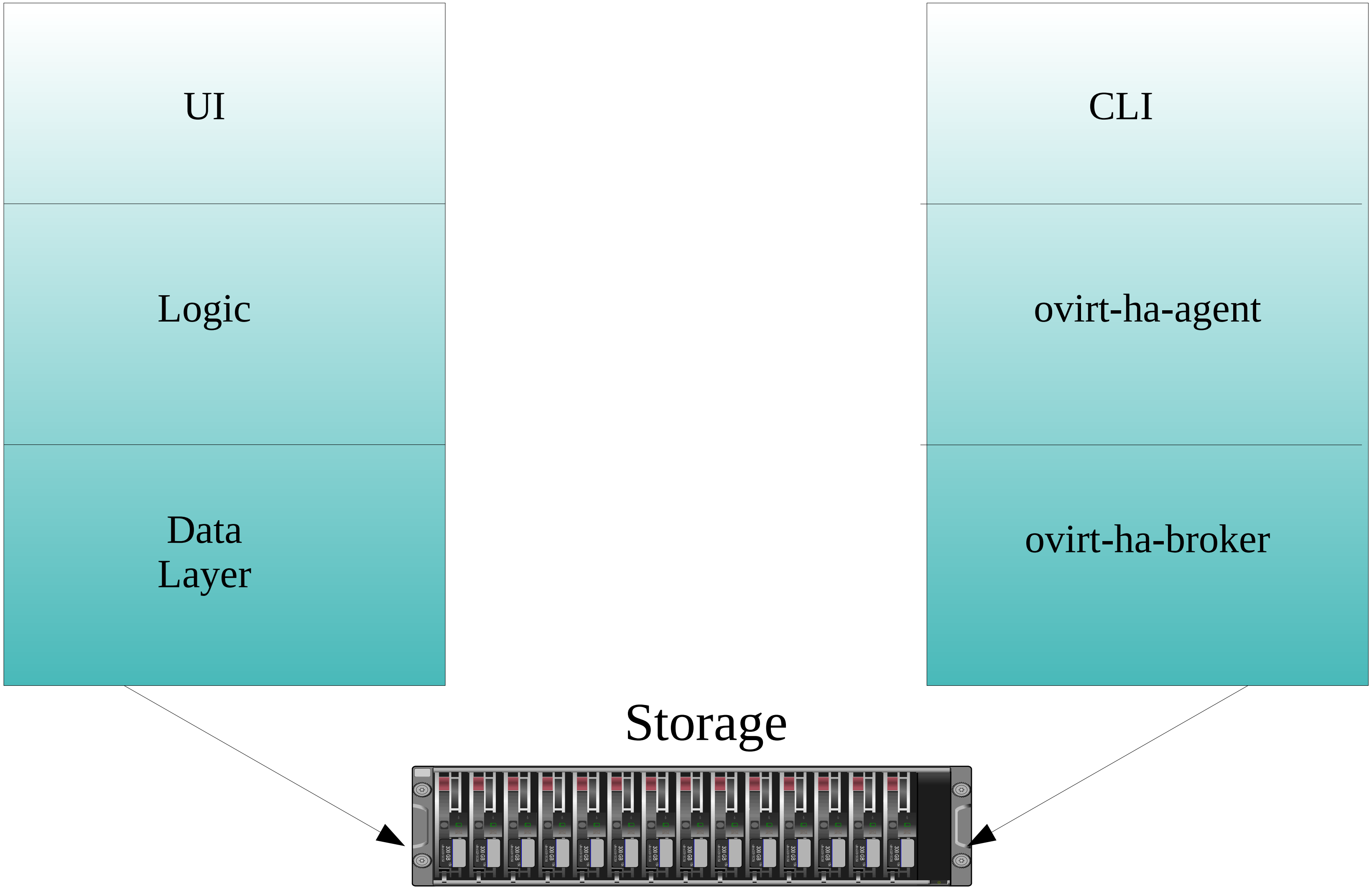
Architecture

CAUTION!

THIS PRODUCT MAY CONTAIN COMICS

Architecture

Classic 3-layers architecture



Architecture

- CLI: /usr/sbin/hosted-engine
  - --help
    - show this help.
  - --deploy
    - run ovirt-hosted-engine deployment
  - --vm-start
    - start VM on this host
  - --vm-shutdown
    - gracefully shut down the VM on this host
  - --vm-poweroff
    - forcefully power off the VM on this host
  - --vm-status
    - VM status according to HA agent

Architecture

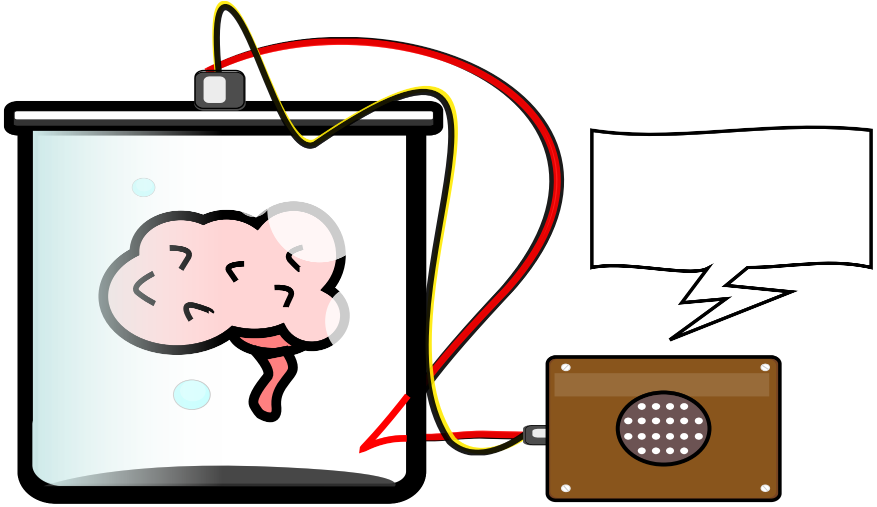
- CLI: /usr/sbin/hosted-engine
  - --add-console-password=<password>
    - Create a temporary password for vnc/spice connection
  - --check-liveliness
    - Checks liveliness page of engine
  - --connect-storage
    - Connect the storage domain
  - --start-pool
    - Start the storage pool manually
  - --console
    - Open the configured console using remote-viewer on localhost

Coming soon:

  - --set-maintenance=<local|global|none>

Architecture

- ovirt-ha-agent
  - AKA 'The Brain'
  - Standalone system service
  - Contains the HA logic, state machine, etc
  - Takes action if needed to ensure high availability
  - Communicates locally with the broker to get data



Architecture

- 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

Architecture

- ovirt-ha-broker
  - Standalone system service
  - Liason between ovirt-ha-agent and:
    - Shared storage
    - Monitoring
  - Serializes requests
  - Separate, testable entity distinct from ovirt-ha-agent

Architecture

- ovirt-ha-broker (continued)
  - Used by ovirt-ha-agent to read to/write from storage
  - Pluggable monitoring (…/submonitors/)
  - 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

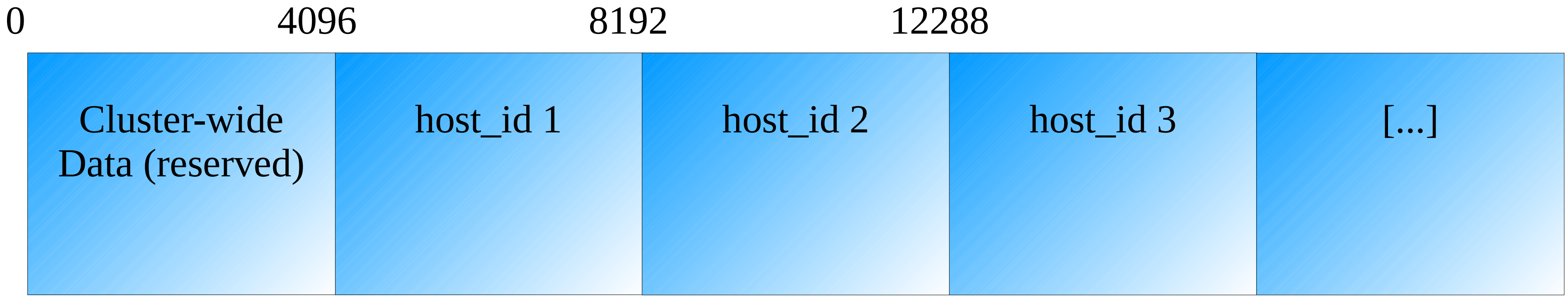
Hosted engine storage

- 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

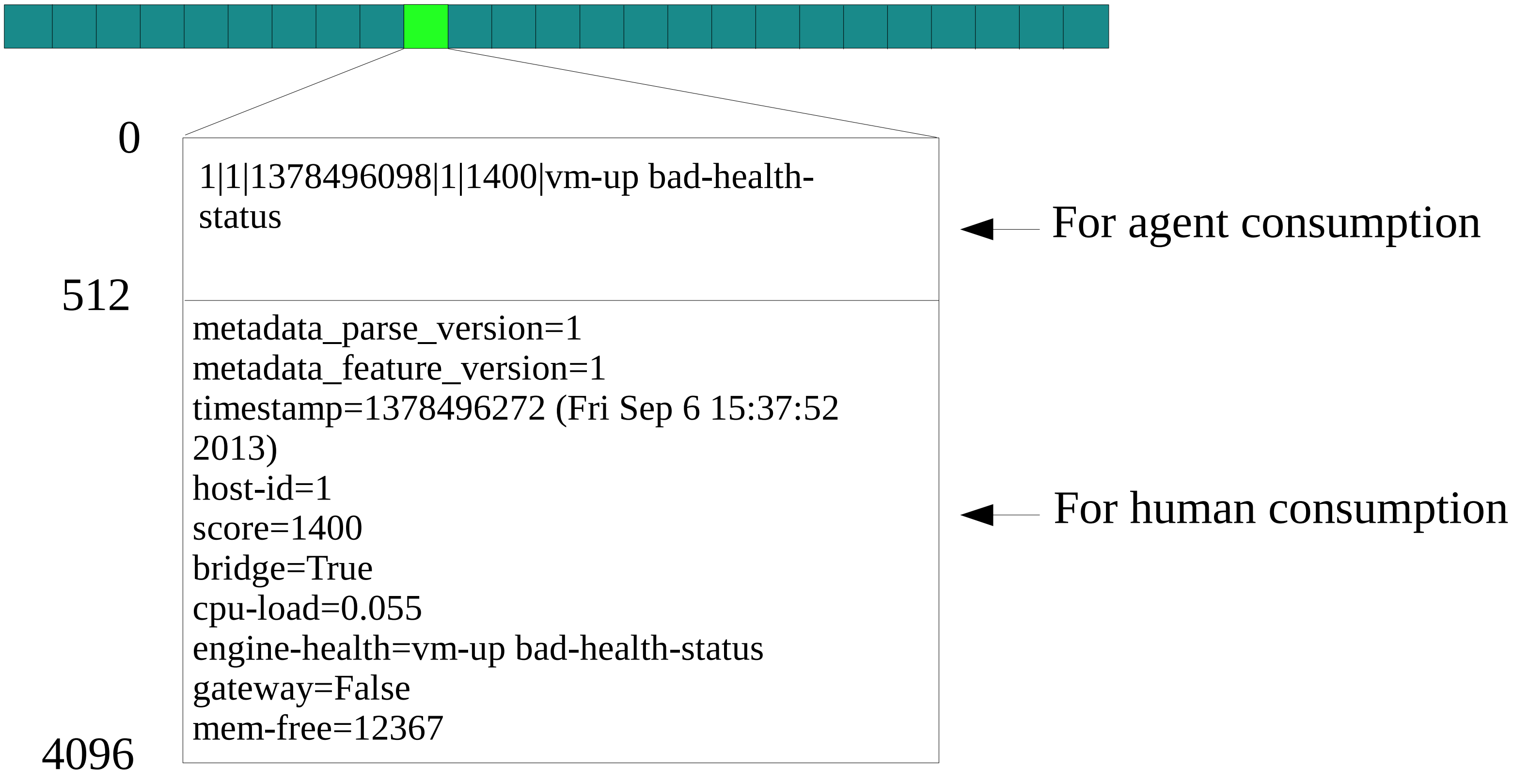


0 4096 8192 12288

Cluster-wide Data (reserved) | host_id 1 | host_id 2 | host_id 3 | [...]

Hosted engine storage

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



0

1|1|1378496098|1|1400|vm-up bad-health-status

For agent consumption

512

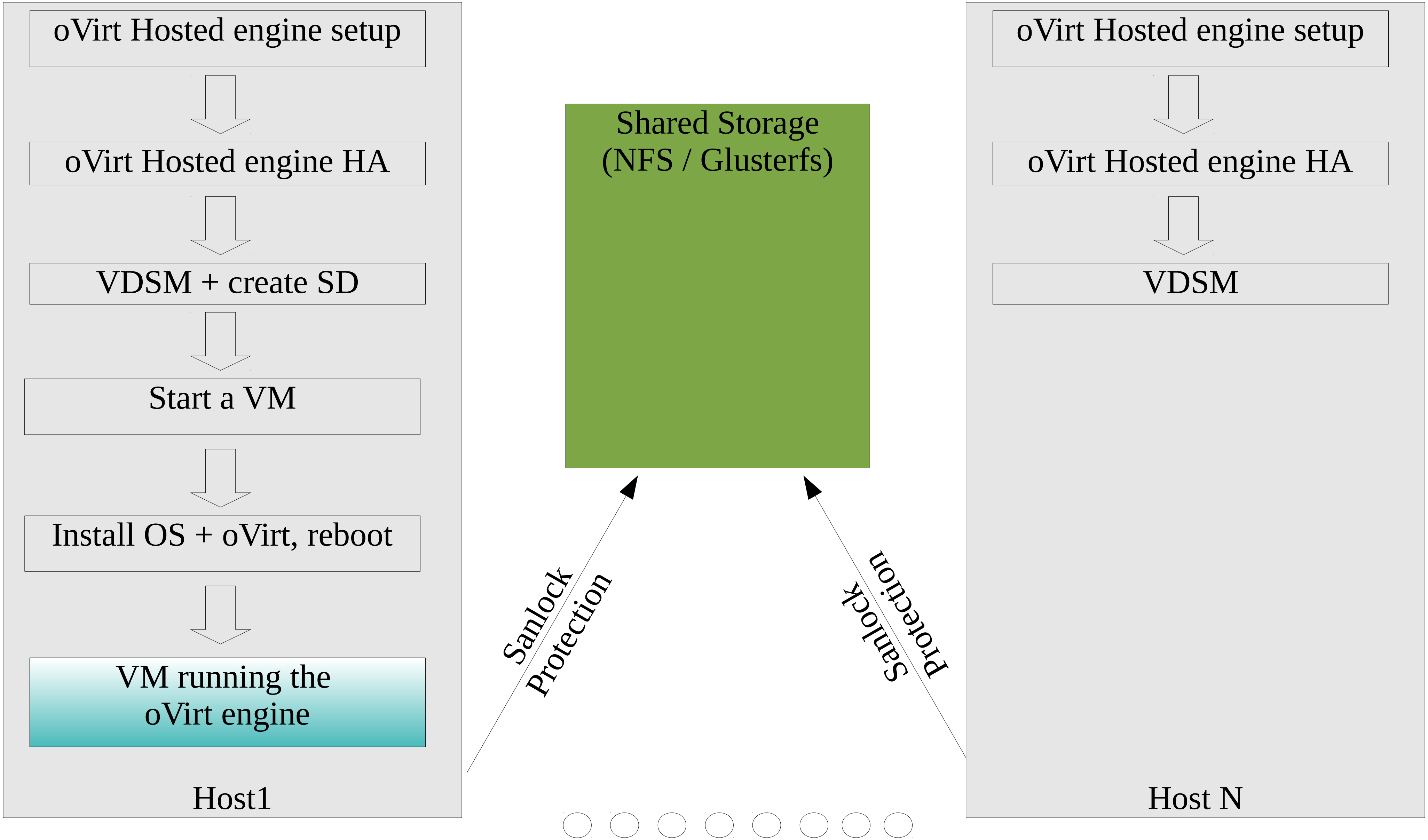
```
metadata_parse_version=1
metadata_feature_version=1
timestamp=1378496272 (Fri Sep 6 15:37:52 2013)
host-id=1
score=1400
bridge=True
cpu-load=0.055
engine-health=vm-up bad-health-status
gateway=False
mem-free=12367
```

For human consumption

4096

Setup

Setup flow



oVirt Hosted engine setup

oVirt Hosted engine HA

VDSM + create SD

Start a VM

Install OS + oVirt, reboot

VM running the oVirt engine

Host1

Shared Storage (NFS / Glusterfs)

Sanlock Protection

Sanlock Protection

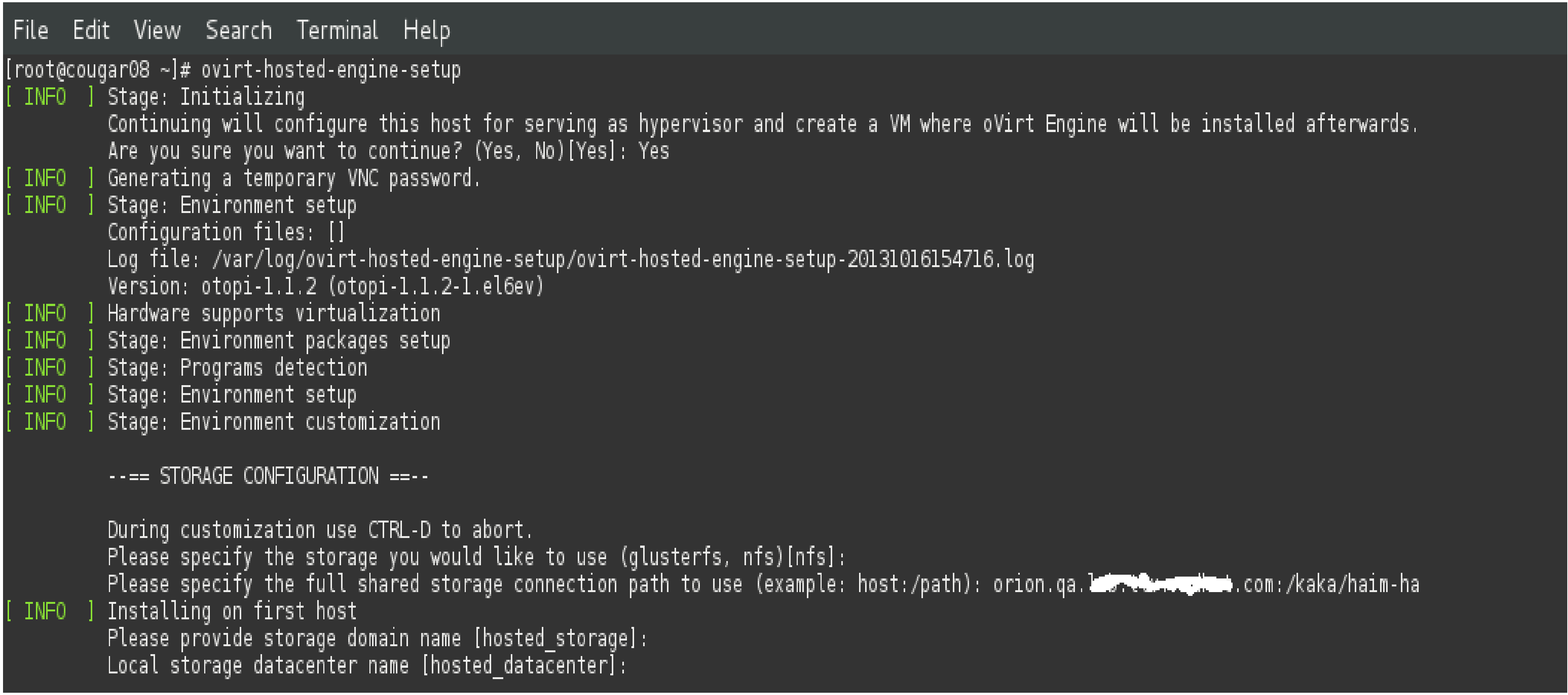
oVirt Hosted engine setup

oVirt Hosted engine HA

VDSM

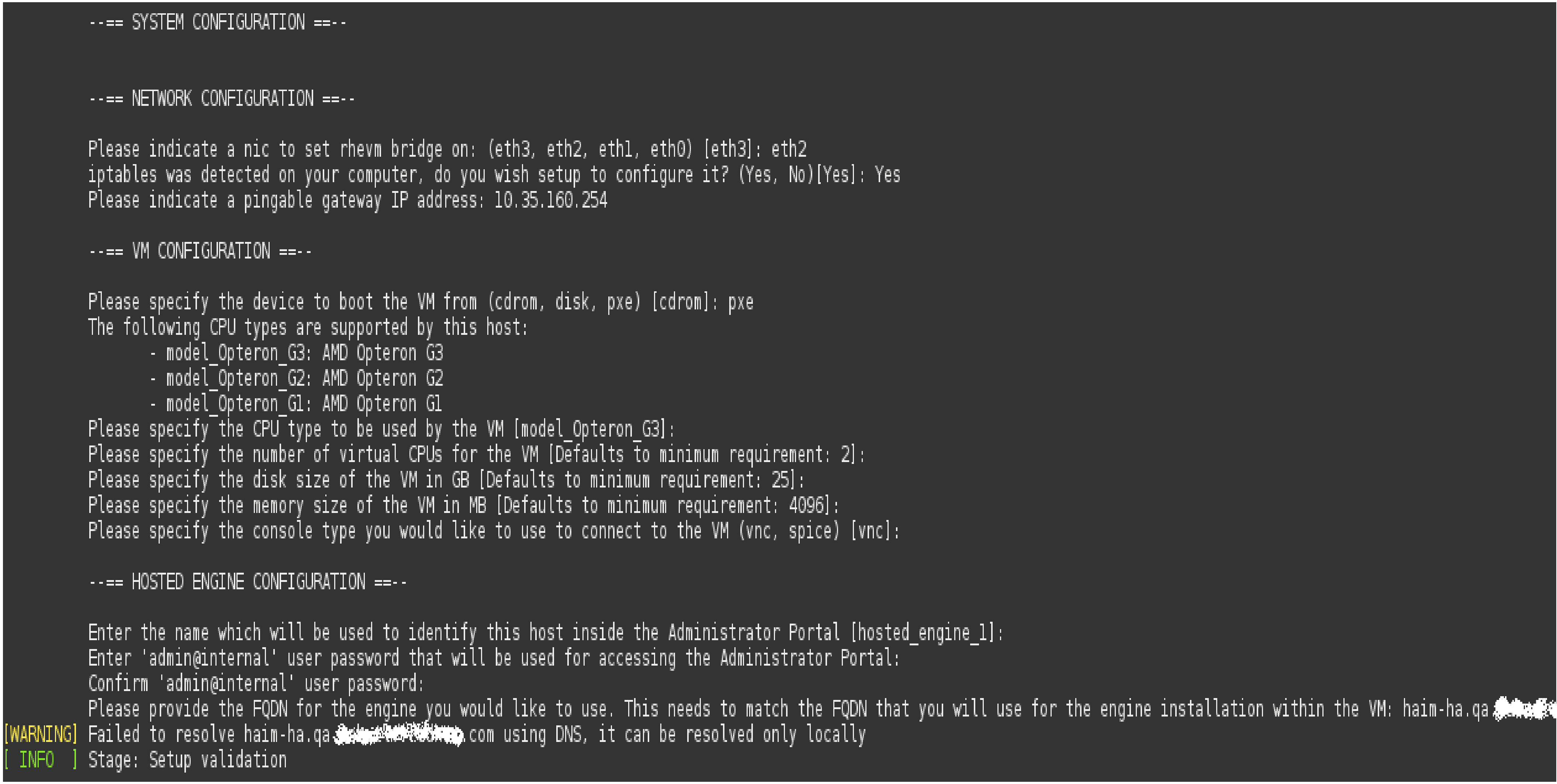
Host N

Setting up the first node



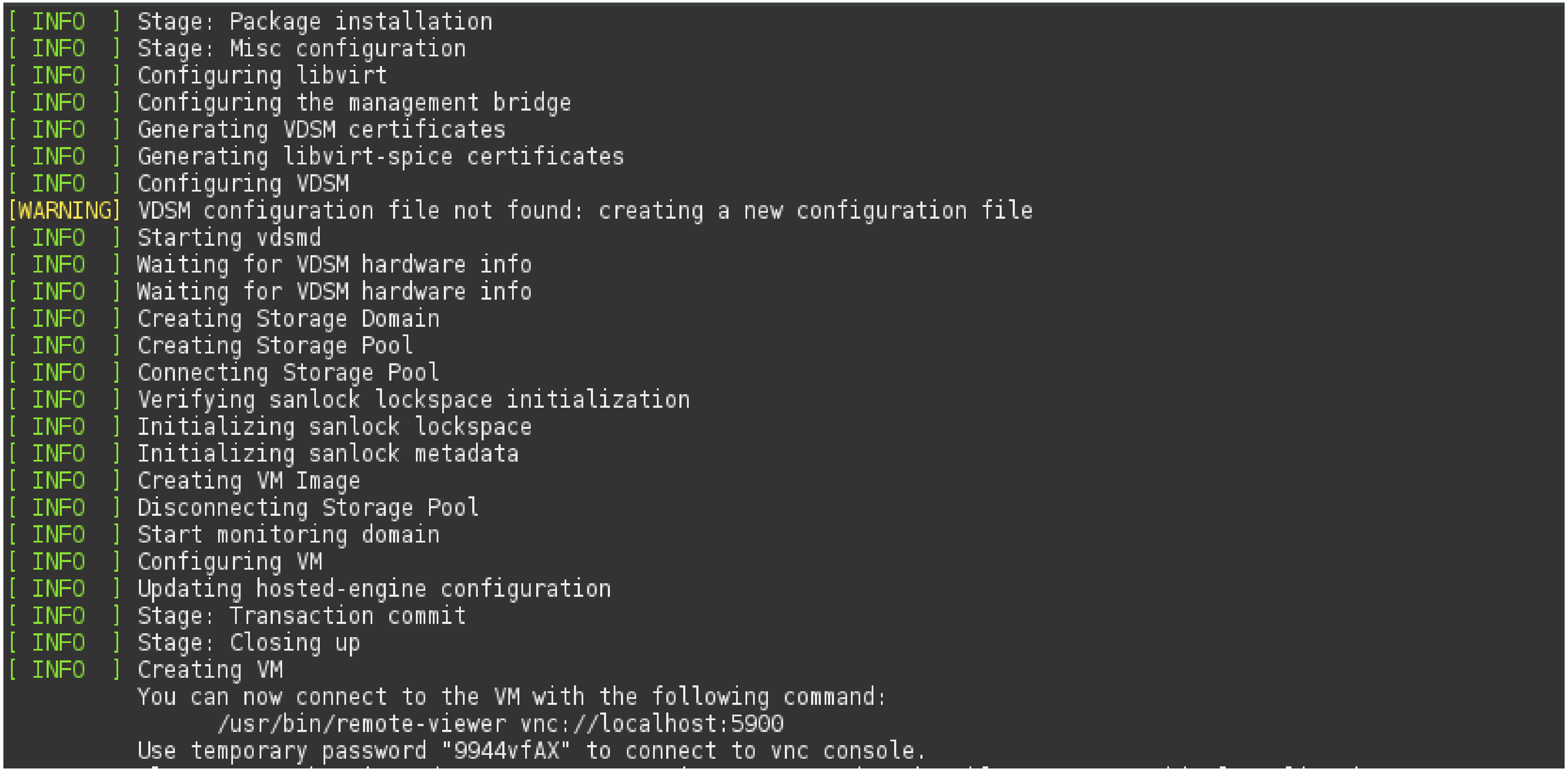
```
[root@cougar08 ~]# ovirt-hosted-engine-setup
[ INFO  ] Stage: Initializing
          Continuing will configure this host for serving as hypervisor and create a VM where oVirt Engine will be installed afterwards.
          Are you sure you want to continue? (Yes, No)[Yes]: Yes
[ INFO  ] Generating a temporary VNC password.
[ INFO  ] Stage: Environment setup
          Configuration files: []
          Log file: /var/log/ovirt-hosted-engine-setup/ovirt-hosted-engine-setup-20131016154716.log
          Version: otopi-1.1.2 (otopi-1.1.2-1.el6ev)
[ INFO  ] Hardware supports virtualization
[ 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): orion.qa.[redacted].com:/kaka/haim-ha
[ INFO  ] Installing on first host
          Please provide storage domain name [hosted_storage]:
          Local storage datacenter name [hosted_datacenter]:
```

Setting up the first node



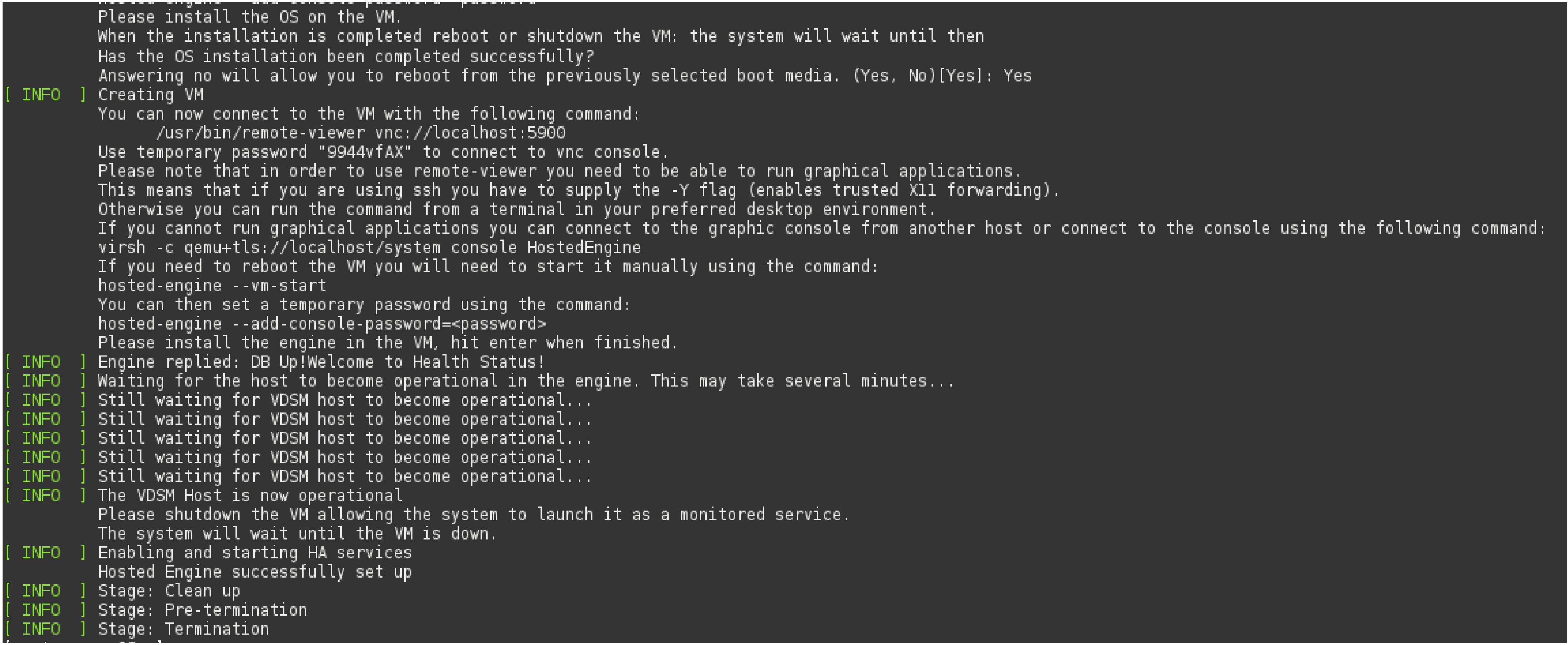
```
          --== SYSTEM CONFIGURATION ==--



          --== NETWORK CONFIGURATION ==--

          Please indicate a nic to set rhevm bridge on: (eth3, eth2, eth1, eth0) [eth3]: eth2
          iptables was detected on your computer, do you wish setup to configure it? (Yes, No)[Yes]: Yes
          Please indicate a pingable gateway IP address: 10.35.160.254

          --== VM CONFIGURATION ==--

          Please specify the device to boot the VM from (cdrom, disk, pxe) [cdrom]: pxe
          The following CPU types are supported by this host:
                 - model_Opteron_G3: AMD Opteron G3
                 - model_Opteron_G2: AMD Opteron G2
                 - model_Opteron_G1: AMD Opteron G1
          Please specify the CPU type to be used by the VM [model_Opteron_G3]:
          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: haim-ha.qa[redacted]
[WARNING] Failed to resolve haim-ha.qa[redacted].com using DNS, it can be resolved only locally
[ INFO  ] Stage: Setup validation
```

Setting up the first node



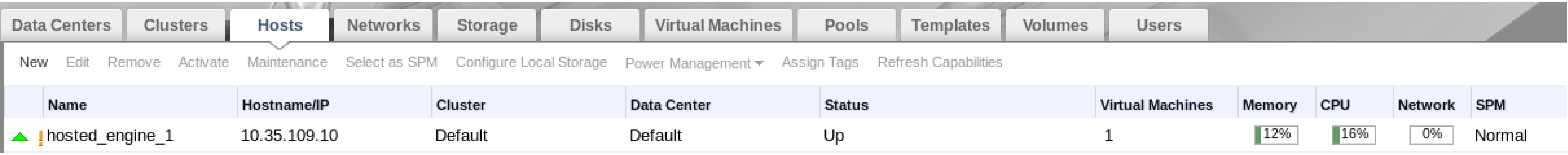
```
[ INFO  ] Stage: Package installation
[ INFO  ] Stage: Misc configuration
[ INFO  ] Configuring libvirt
[ INFO  ] Configuring the management bridge
[ INFO  ] Generating VDSM certificates
[ INFO  ] Generating libvirt-spice certificates
[ INFO  ] Configuring VDSM
[WARNING] VDSM configuration file not found: creating a new configuration file
[ INFO  ] Starting vdsmd
[ INFO  ] Waiting for VDSM hardware info
[ 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  ] Initializing sanlock metadata
[ INFO  ] Creating VM Image
[ INFO  ] Disconnecting Storage Pool
[ INFO  ] Start monitoring domain
[ INFO  ] Configuring VM
[ INFO  ] Updating hosted-engine configuration
[ INFO  ] Stage: Transaction commit
[ INFO  ] Stage: Closing up
[ INFO  ] Creating VM
          You can now connect to the VM with the following command:
                /usr/bin/remote-viewer vnc://localhost:5900
          Use temporary password "9944vfAX" to connect to vnc console.
```

Setting up the first node

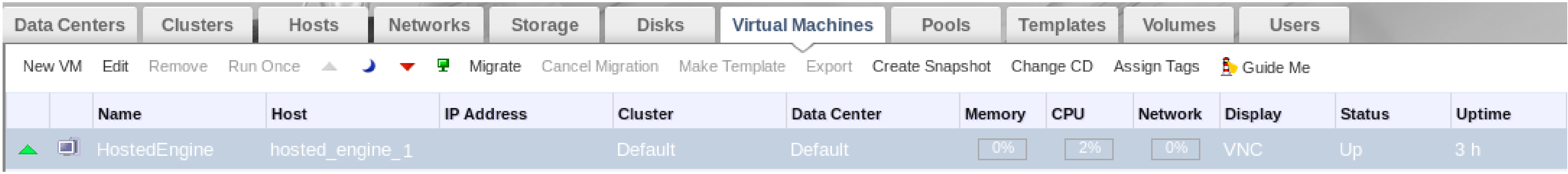


```
         Please install the OS on the VM.
         When the installation is completed reboot or shutdown the VM: the system will wait until then
         Has the OS installation been completed successfully?
         Answering no will allow you to reboot from the previously selected boot media. (Yes, No)[Yes]: Yes
[ INFO  ] Creating VM
         You can now connect to the VM with the following command:
               /usr/bin/remote-viewer vnc://localhost:5900
         Use temporary password "9944vfAX" to connect to vnc console.
         Please note that in order to use remote-viewer you need to be able to run graphical applications.
         This means that if you are using ssh you have to supply the -Y flag (enables trusted X11 forwarding).
         Otherwise you can run the command from a terminal in your preferred desktop environment.
         If you cannot run graphical applications you can connect to the graphic console from another host or connect to the console using the following command:
         virsh -c qemu+tls://localhost/system console HostedEngine
         If you need to reboot the VM you will need to start it manually using the command:
         hosted-engine --vm-start
         You can then set a temporary password using the command:
         hosted-engine --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  ] Still waiting for VDSM host to become operational...
[ INFO  ] Still waiting for VDSM host to become operational...
[ INFO  ] Still waiting for VDSM host to become operational...
[ INFO  ] Still waiting for VDSM host to become operational...
[ INFO  ] Still waiting for VDSM host to become operational...
[ 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
```

Hosted engine is alive!



	Name	Hostname/IP	Cluster	Data Center	Status	Virtual Machines	Memory	CPU	Network	SPM
▲ !	hosted_engine_1	10.35.109.10	Default	Default	Up	1	12%	16%	0%	Normal



		Name	Host	IP Address	Cluster	Data Center	Memory	CPU	Network	Display	Status	Uptime
▲		HostedEngine	hosted_engine_1		Default	Default	0%	2%	0%	VNC	Up	3 h

Setting up the 2nd+ node

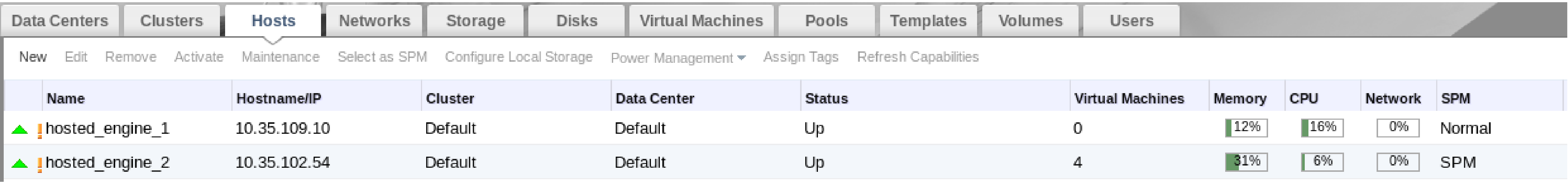
```
[root@thinkerbell ~]# hosted-engine --deploy --config-append=answers.conf
[ INFO  ] Stage: Initializing
          Continuing will configure this host for serving as hypervisor and create a VM where oVirt Engine
will be installed afterwards.
          Are you sure you want to continue? (Yes, No)[Yes]:
[ INFO  ] Generating a temporary VNC password.
[ INFO  ] Stage: Environment setup
          Configuration files: ['/root/answers.conf']
          Log file: /var/log/ovirt-hosted-engine-setup/ovirt-hosted-engine-setup-20131018091350.log
          Version: otopi-1.2.0_master (otopi-1.2.0-0.0.master.20131007.git6f8ac6d.fc19)
[ 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.
          The specified storage location already contains a data domain. Is this an additional
host setup (Yes, No)[Yes]?
[ INFO  ] Installing on additional host
           Please specify the Host ID [Must be integer, default: 2]:
```

Setting up the 2nd+ node

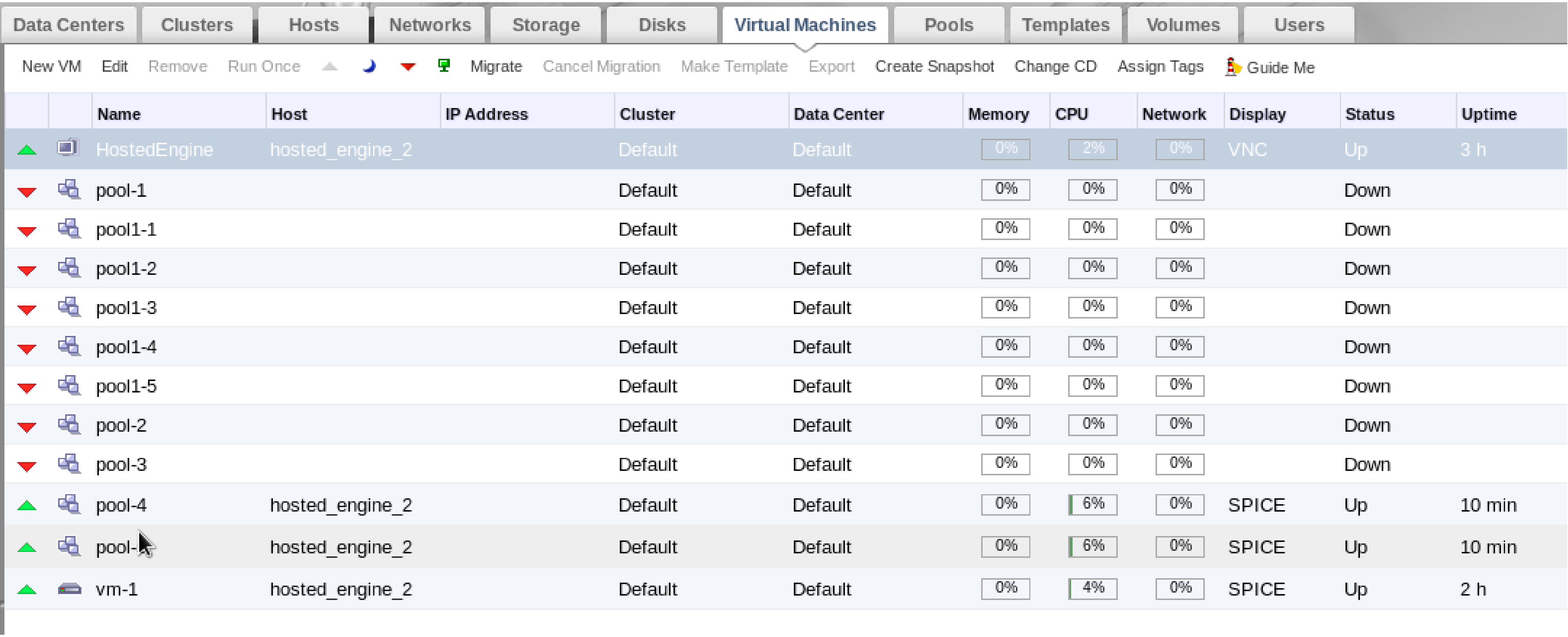
```
        --== HOSTED ENGINE CONFIGURATION ==--

        Enter the name which will be used to identify this host inside the Administrator Portal
[hosted_engine_2]:
        Enter 'admin@internal' user password that will be used for accessing the Administrator Portal:
        Confirm 'admin@internal' user password:
[ INFO  ] Stage: Setup validation

….

[ INFO  ] The VDSM Host is now operational
[ INFO  ] Enabling and starting HA services
        Hosted Engine successfully set up
[ INFO  ] Stage: Clean up
[ INFO  ] Stage: Pre-termination
[ INFO  ] Stage: Termination
```

Hosted engine is alive, 2 nodes running

oVirt



	Name	Hostname/IP	Cluster	Data Center	Status	Virtual Machines	Memory	CPU	Network	SPM
▲ !	hosted_engine_1	10.35.109.10	Default	Default	Up	0	12%	16%	0%	Normal
▲ !	hosted_engine_2	10.35.102.54	Default	Default	Up	4	31%	6%	0%	SPM



		Name	Host	IP Address	Cluster	Data Center	Memory	CPU	Network	Display	Status	Uptime
▲		HostedEngine	hosted_engine_2		Default	Default	0%	2%	0%	VNC	Up	3 h
▼		pool-1			Default	Default	0%	0%	0%		Down	
▼		pool1-1			Default	Default	0%	0%	0%		Down	
▼		pool1-2			Default	Default	0%	0%	0%		Down	
▼		pool1-3			Default	Default	0%	0%	0%		Down	
▼		pool1-4			Default	Default	0%	0%	0%		Down	
▼		pool1-5			Default	Default	0%	0%	0%		Down	
▼		pool-2			Default	Default	0%	0%	0%		Down	
▼		pool-3			Default	Default	0%	0%	0%		Down	
▲		pool-4	hosted_engine_2		Default	Default	0%	6%	0%	SPICE	Up	10 min
▲		pool-	hosted_engine_2		Default	Default	0%	6%	0%	SPICE	Up	10 min
▲		vm-1	hosted_engine_2		Default	Default	0%	4%	0%	SPICE	Up	2 h

HA simulation

Hosted engine simulation

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

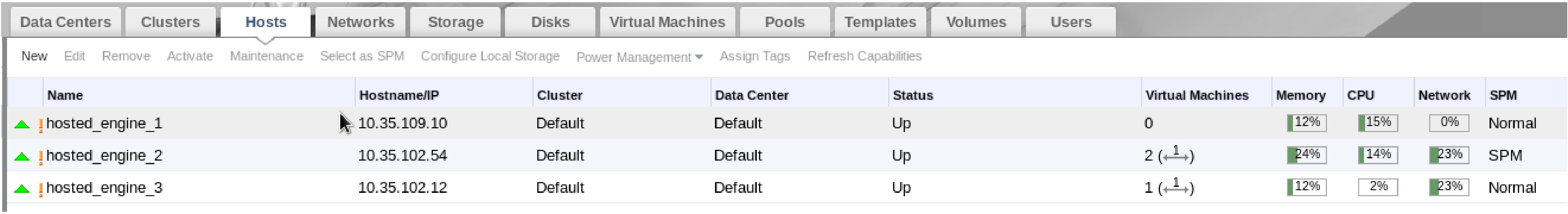
```
--== Host 1 status ==--

Hostname                           : hosted_engine_2
Host ID                            : 1
Engine status                      : vm-up good-health-status
Score                              : 2400
Host timestamp                     : 1378510362
Extra metadata                     :
    timestamp=1378510362 (Sun Oct 20 19:32:42 2013)
    host-id=1
    score=2400
    engine-health=vm-up good-health-status
    gateway=True


--== Host 2 status ==--

Hostname                           : hosted_engine_3
Host ID                            : 2
Engine status                      : vm-down
Score                              : 2400
Host timestamp                     : 1378510365
Extra metadata                     :
    timestamp=1378510365 (Sun Oct 20 19:32:45 2013)
    host-id=2
    score=2400
    engine-health=vm-down
    gateway=True
```

Hosted engine simulation

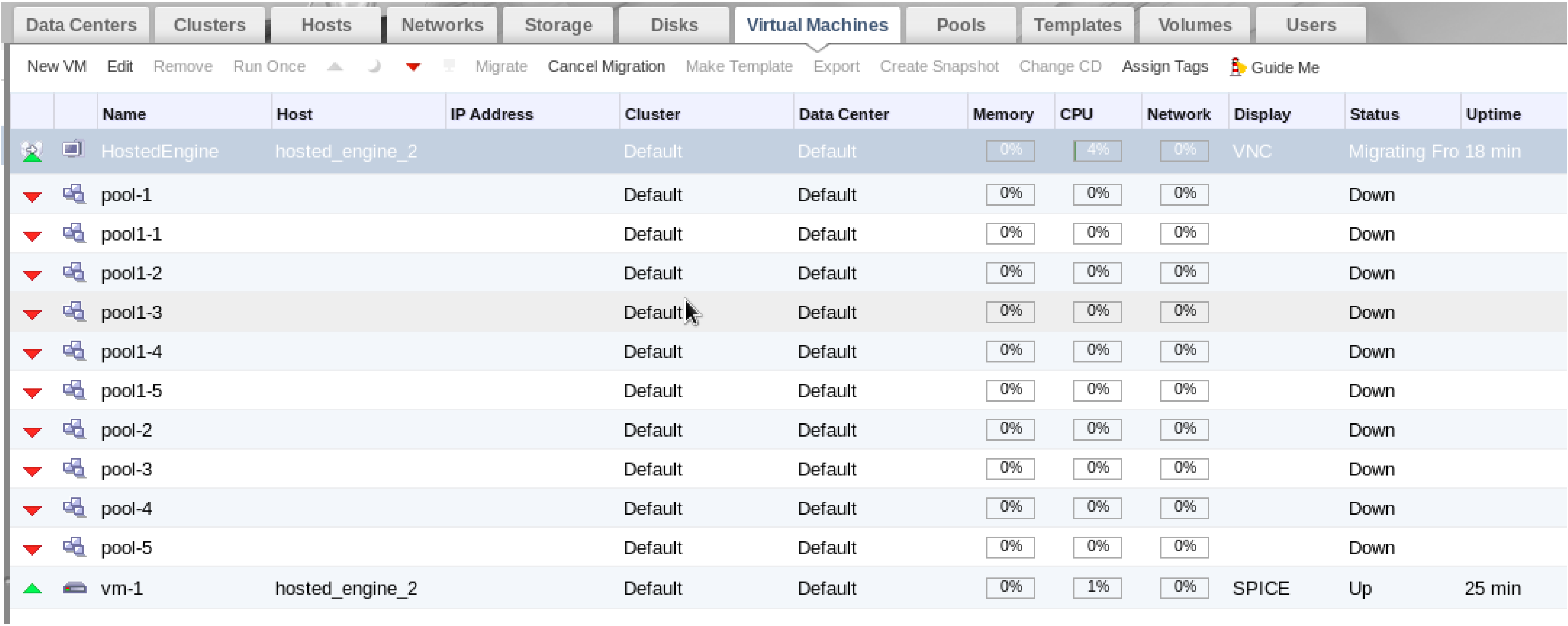
Now, let's block GW in hosted_engine_2....

Hosted engine simulation

oVirt



	Name	Hostname/IP	Cluster	Data Center	Status	Virtual Machines	Memory	CPU	Network	SPM
▲ !	hosted_engine_1	10.35.109.10	Default	Default	Up	0	12%	15%	0%	Normal
▲ !	hosted_engine_2	10.35.102.54	Default	Default	Up	2 (↔ 1)	24%	14%	23%	SPM
▲ !	hosted_engine_3	10.35.102.12	Default	Default	Up	1 (↔ 1)	12%	2%	23%	Normal



		Name	Host	IP Address	Cluster	Data Center	Memory	CPU	Network	Display	Status	Uptime
▲		HostedEngine	hosted_engine_2		Default	Default	0%	4%	0%	VNC	Migrating Fro	18 min
▼		pool-1			Default	Default	0%	0%	0%		Down	
▼		pool1-1			Default	Default	0%	0%	0%		Down	
▼		pool1-2			Default	Default	0%	0%	0%		Down	
▼		pool1-3			Default	Default	0%	0%	0%		Down	
▼		pool1-4			Default	Default	0%	0%	0%		Down	
▼		pool1-5			Default	Default	0%	0%	0%		Down	
▼		pool-2			Default	Default	0%	0%	0%		Down	
▼		pool-3			Default	Default	0%	0%	0%		Down	
▼		pool-4			Default	Default	0%	0%	0%		Down	
▼		pool-5			Default	Default	0%	0%	0%		Down	
▲		vm-1	hosted_engine_2		Default	Default	0%	1%	0%	SPICE	Up	25 min

Hosted engine simulation

- Node 1's gateway down; VM migrated to node 2

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

Hostname                           : hosted_engine_2
Host ID                            : 1
Engine status                      : vm-down
Score                              : 1400
Host timestamp                     : 1378510422
Extra metadata                     :
    timestamp=1378510422 (Sun Oct 20 19:33:42 2013)
    host-id=1
    score=1400
    engine-health=vm-down
    gateway=False


--== Host 2 status ==--

Hostname                           : hosted_engine_3
Host ID                            : 2
Engine status                      : vm-up good-health-status
Score                              : 2400
Host timestamp                     : 1378510425
Extra metadata                     :
    timestamp=1378510425 (Sun Oct 20 19:33:45 2013)
    host-id=2
    score=2400
    engine-health=vm-up good-health-status
    gateway=True
```

Summary

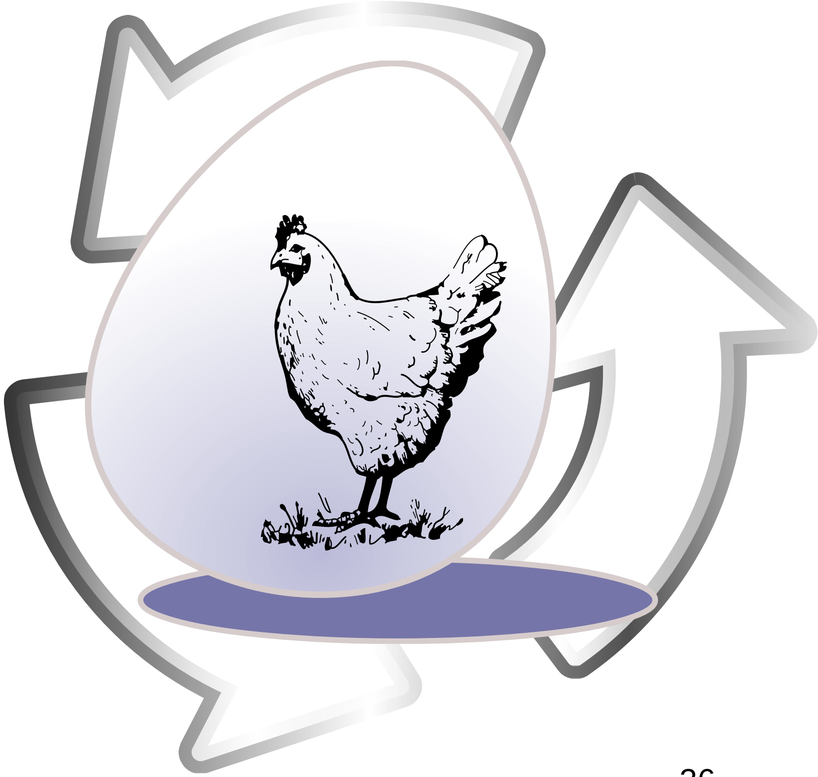
Back to the fundamental question...

Why did the chicken cross the road?

Summary

It did not,

It was migrated by the HA services.



Questions?

THANK YOU !

http://www.ovirt.org
http://www.ovirt.org/Category:SLA

http://lists.ovirt.org/mailman/listinfo
vdsm-devel@lists.fedorahosted.org

#ovirt irc.oftc.net

doron@redhat.com