VDSM Overview

The node virtualization management API

Nov 2nd, 2011

Ayal Baron

Agenda

- What is VDSM?
- Responsibilities
- Why use VDSM?
- Architecture
- Packages
- Infrastructure
- Hooks
- API
- Fortune telling (Roadmap)
- How to contribute

What is VDSM?

- Node virtualization management API
- High level API
- Abstracts low level details of underlying Linux environments
- Present: RHEL-5 RHEL6 RHEV-H & Fedora
- Future: oVirt Node & Other Linux distributions (patches are welcome)

VDSM Responsibilities

- Host bootstrap and registration
- VM life cycle (via libvirt)
- Guest interaction (sso, stop, etc)
- Storage management
- Network configuration
- Monitoring host and VMs
- Policy management
  - Scheduler, KSM
  - Thin provisioning
  - Page cache
- Fencing proxy

Why VDSM?

- `$ qemu-kvm & voila`! we have a virtual machine (but read the fine print).

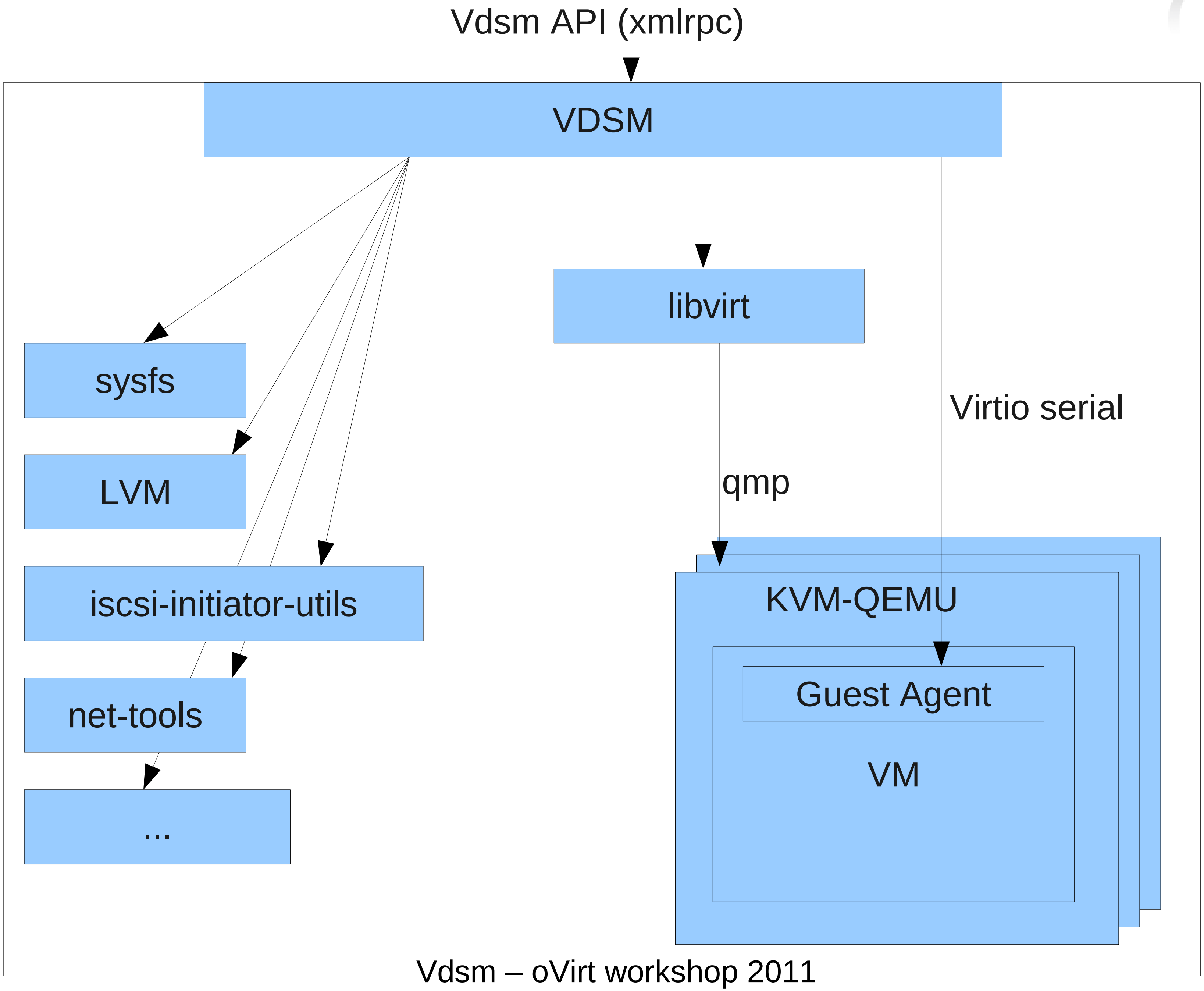
```
/usr/libexec/qemu-kvm -S -M rhel6.0.0 -cpu Conroe -enable-kvm -m 2048 -smp 1,sockets=1,cores=1,threads=1 -name z-win7x86-1 -uuid e3e19b36-f6b7-4ab9-b604-1f8b5c471bda -smbios
type=1,manufacturer=Red Hat,product=RHEL,version=6Server-6.1.0.2.el6_1,serial=50C1C6F0-B18B-11DE-ADF1-00215EC7FC0C_00:1A:64:E7:0E:E0,uuid=e3e19b36-f6b7-4ab9-b604-
1f8b5c471bda -nodefconfig -nodefaults -chardev socket,id=charmonitor,path=/var/lib/libvirt/qemu/z-win7x86-1.monitor,server,nowait -mon chardev=charmonitor,id=monitor,mode=control -rtc
base=2011-08-04T06:17:36 -boot cdn -device virtio-serial-pci,id=virtio-serial0,max_ports=16,bus=pci.0,addr=0x6 -drive file=/rhev/data-center/6927f974-c6f6-482f-aca9-
907c4acc71a9/50027e48-6cb9-4345-9c7a-c22b41ad84d2/images/5ada0ef6-5f4a-40b8-ad92-cb6758de8536/c22f4e68-439b-4a87-8e22-bc7d8e2391f1,if=none,id=drive-ide0-0-
0,format=qcow2,serial=b8-ad92-cb6758de8536,cache=none,werror=stop,rerror=stop,aio=native -device ide-drive,bus=ide.0,unit=0,drive=drive-ide0-0-0,id=ide0-0-0 -drive file=/rhev/data-
center/6927f974-c6f6-482f-aca9-907c4acc71a9/e0acfcc8-c020-413e-84cd-a93cb0ab9b2d/images/11111111-1111-1111-1111-111111111111/RHEV-
toolsSetup_3.0_12.iso,if=none,media=cdrom,id=drive-ide0-1-0,readonly=on,format=raw -device ide-drive,bus=ide.1,unit=0,drive=drive-ide0-1-0,id=ide0-1-0 -drive file=/rhev/data-
center/6927f974-c6f6-482f-aca9-907c4acc71a9/50027e48-6cb9-4345-9c7a-c22b41ad84d2/images/f52621e0-8b1e-47af-809c-45de2aa697fc/f77b5dd2-3141-4ea7-84fa-
e8cfffe9cff9,if=none,id=drive-virtio-disk0,format=qcow2,serial=af-809c-45de2aa697fc,cache=none,werror=stop,rerror=stop,aio=native -device virtio-blk-pci,bus=pci.0,addr=0x7,drive=drive-virtio-
disk0,id=virtio-disk0 -netdev tap,fd=27,id=hostnet0 -device rtl8139,netdev=hostnet0,id=net0,mac=00:1a:4a:23:11:0b,bus=pci.0,addr=0x3 -netdev tap,fd=29,id=hostnet1,vhost=on,vhostfd=30
-device virtio-net-pci,netdev=hostnet1,id=net1,mac=00:1a:4a:23:11:0c,bus=pci.0,addr=0x4 -chardev socket,id=charchannel0,path=/var/lib/libvirt/qemu/channels/z-win7x86-
1.com.redhat.rhevm.vdsm,server,nowait -device virtserialport,bus=virtio-serial0.0,nr=1,chardev=charchannel0,id=channel0,name=com.redhat.rhevm.vdsm -chardev
spicevmc,id=charchannel1,name=vdagent -device virtserialport,bus=virtio-serial0.0,nr=2,chardev=charchannel1,id=channel1,name=com.redhat.spice.0 -usb -spice port=5902,tls-
port=5903,addr=0,x509-dir=/etc/pki/vdsm/libvirt-spice,tls-channel=main,tls-channel=inputs -k en-us -vga qxl -global qxl-vga.vram_size=67108864 -device intel-
hda,id=sound0,bus=pci.0,addr=0x5 -device hda-duplex,id=sound0-codec0,bus=sound0.0,cad=0
```

- To manage multiple virtual machines you would need libvirt: virsh, virt-manager.
- To dynamically manage anything from a few VMs on a single host up to thousands of VMs on a cluster of hundreds of hosts using multiple storage targets – VDSM

Architecture and Implementation

- VDSM is the oVirt node agent, tailored for its needs
- It manages transient VMs (vm data stored centrally in db managed by oVirt)
- It is KVM centric
- moving to a more general use case, applicable to other management platforms

Architecture and Implementation



Architecture and Implementation

- Written in Python
- Multithreaded, multi-processes
- Speaks with its guest agent via virtio-serial
- Adds customized clustering support for LVM that scales to hundreds of nodes
  - Implements a distributed image repository over the supported storage types (local directory, FCP, FCoE, iSCSI, NFS, SAS)
  - Multihost system, one concurrent metadata writer
  - Scales linearly in data writers

Robustness as a Design Goal

- Evaporated NFS exports
- Faulty paths
- Node crashes
- Live-locked qemu
- Internal Python exceptions
- Self-fencing of metadata writer

Packages

- vdsm
- vdsm_cli
- vds_bootstrap
- vdsm_reg
- vdsm_hooks

Infrastructure

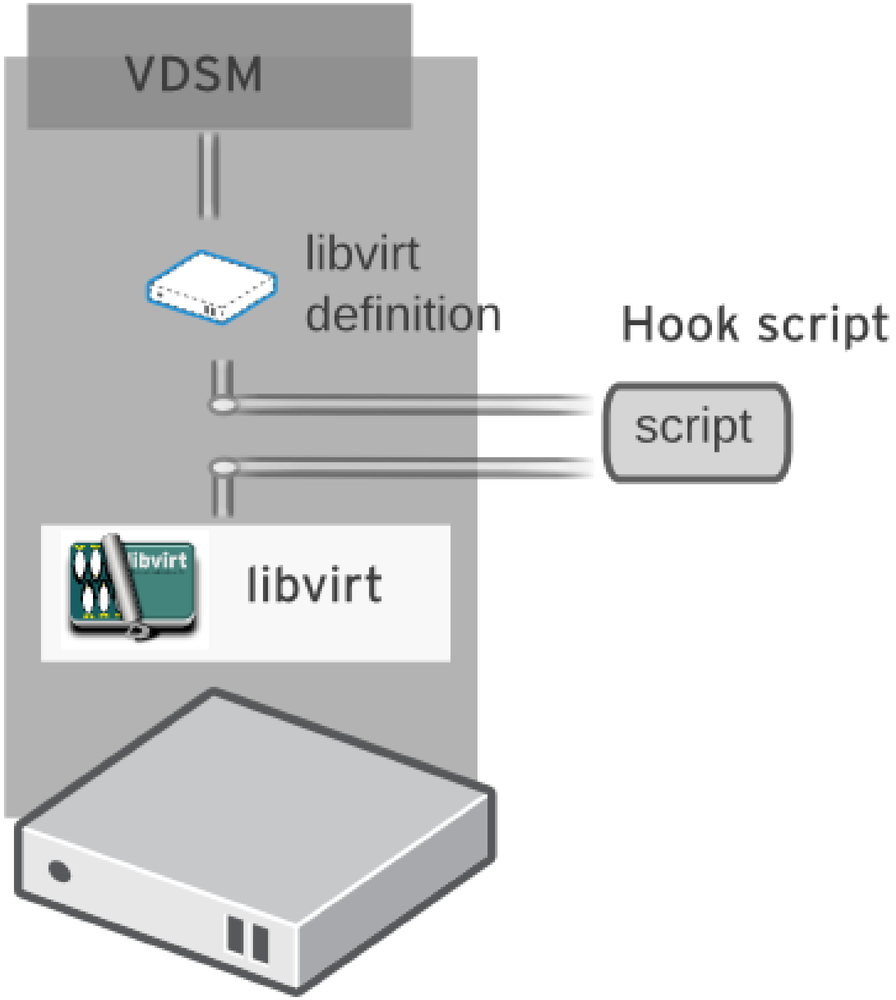
- Supervdsm
- Out of process
- Async Tasks

Infrastructure - cont

- Misc
  - execCmd
  - persistFile
  - retry
  - AsyncProc
  - [i]tmap
  - PersistantDict
  - lvm cache
- Logging
  - logskip
  - simpleLogAdapter
  - OOPLogger
- Synchronization
  - RWLock
  - DeferableContext
  - DynamicBarrier
  - SamplingMethod
  - OperationMutex
  - Safelease
  - ResourceManager
  - Securable

Hooks

- VM lifecycle hooks
  - before/after vm_start
  - before/after vm_cont
  - before/after vm_pause
  - before/after vm_hibernate
  - before/after vm_dehibernate
  - before/after vm_migrate_source
  - before/after vm_migrate_destination
  - after_vm_destroy
- Vdsm lifecycle hooks
  - before/after vdsm_start



VM Lifecycle API

- create
- destroy
- pause
- continue
- setVmTicket
- changeCD
- changeFloppy
- migrate (downtime, timeout)
- hibernate

VM Lifecycle API (agent-dependent)

- shutdown
- desktopLogin
- desktopLogoff
- desktopLock

VM Monitoring API

- list
- getAllVmStats
- getVmStats
  - Interesting applications installed
  - Logged in users
  - CPU consumption
  - Memory usage

Network Config API

- AddNetwork
- DelNetwork
- EditNetwork
- SetSafeNetworkConfig
- SetupNetworks
- ConnectivityCheck

Host Monitoring API

- getVdsCapabilities
- getVdsStats
- ping
- fenceNode

Storage API

- connectStorageServer
- getDeviceList
- createStorageDomain
- attachStorageDomain
- createImage
- prepareVolume
-
-
- (and many, many more)

- repoStats
- getSpmStatus
- spmStart
- extendVolume

Async Tasks API

- GetAllTasksStatuses
- getTaskStatus
- clearTask
- stopTask

Roadmap

- Networking
  - Vepa, VN-Link, SRIOV
  - storage network (requires bridgeless network)
  - migration network (requires bridgeless network)
  - Traffic shaping (tc, cgroups)
  - Intrusion detection
- Cgroups (CPU, Memory, I/O, Network)
- Monitoring
  - Add counters
  - Move to collectd?
- Support for self-contained single host

Roadmap - cont

- New API
  - Current API is not very clean (createVG, createStorageDomain)
  - stable
  - RESTful?
  - oVirt-api look and feel
- Support sending events
  - QMF support
- Split VDSM up into reusable autonomous parts.
  - Spin storage off as a generic image repository.
  - Policy engine (MOM?)

How to contribute

- **Repository:**
  - http://git.fedorahosted.org/git/?p=vdsm.git
- **Mailing lists:**
  - vdsm-devel@lists.fedorahosted.org
  - vdsm-patches@lists.fedorahosted.org
- **IRC:**
  - #vdsm on Freenode
- **Core Team:**
  Dan Kenigsberg, Saggi Mizrahi, Igor Lvovsky, Eduardo Warszawasky, Yotam Oron, Ayal Baron

Q&A

oVirt

THANK YOU !

http://www.ovirt.org

Bootstrap

- Verifies node compatibility with oVirt
  - Check os/cpu/vdsm compatibility
  - Check RPMs (Install if needed)
  - Configure node (certificate, networking, services, etc.)
- Currently supports only RHEL 5.X and RHEL 6.X
- Working on support for Fedora

VDSM Overview

The node virtualization management API

Nov 2nd, 2011

Ayal Baron

Agenda

- What is VDSM?
- Responsibilities
- Why use VDSM?
- Architecture
- Packages
- Infrastructure
- Hooks
- API
- Fortune telling (Roadmap)
- How to contribute

What is VDSM?

- Node virtualization management API
- High level API
- Abstracts low level details of underlying Linux environments
- Present: RHEL-5 RHEL6 RHEV-H & Fedora
- Future: oVirt Node & Other Linux distributions (patches are welcome)

VDSM Responsibilities

- Host bootstrap and registration
- VM life cycle (via libvirt)
- Guest interaction (sso, stop, etc)
- Storage management
- Network configuration
- Monitoring host and VMs
- Policy management
  - Scheduler, KSM
  - Thin provisioning
  - Page cache
- Fencing proxy

Why VDSM?

- `$ qemu-kvm & voila`! we have a virtual machine (but read the fine print).

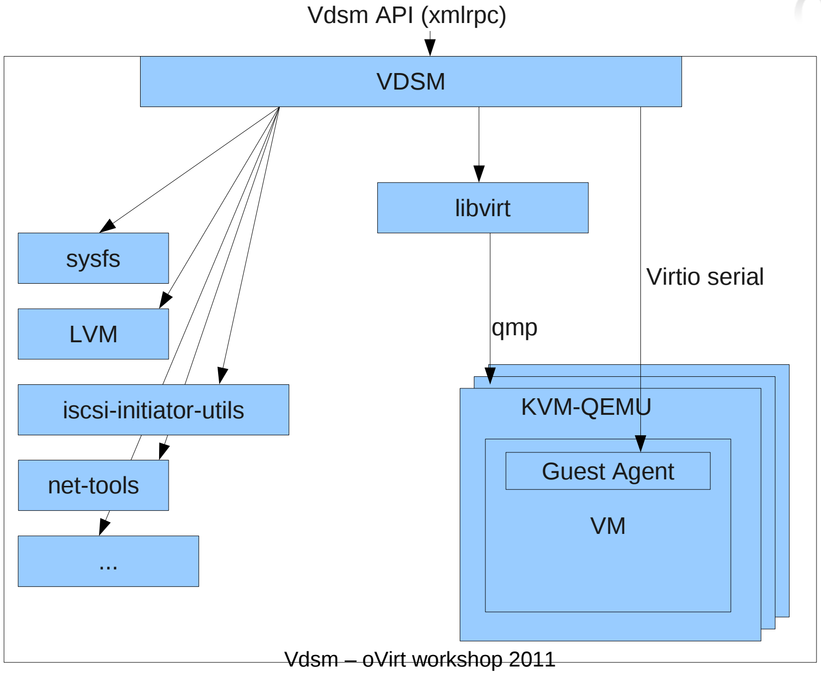
```
/usr/libexec/qemu-kvm -S -M rhel6.0.0 -cpu Conroe -enable-kvm -m 2048 -smp 1,sockets=1,cores=1,threads=1 -name z-win7x86-1 -uuid e3e19b36-f6b7-4ab9-b604-1f8b5c471bda -smbios
type=1,manufacturer=Red Hat,product=RHEL,version=6Server-6.1.0.2.el6_1,serial=50C1C6F0-B18B-11DE-ADF1-00215EC7FC0C_00:1A:64:E7:0E:E0,uuid=e3e19b36-f6b7-4ab9-b604-
1f8b5c471bda -nodefconfig -nodefaults -chardev socket,id=charmonitor,path=/var/lib/libvirt/qemu/z-win7x86-1.monitor,server,nowait -mon chardev=charmonitor,id=monitor,mode=control -rtc
base=2011-08-04T06:17:36 -boot cdn -device virtio-serial-pci,id=virtio-serial0,max_ports=16,bus=pci.0,addr=0x6 -drive file=/rhev/data-center/6927f974-c6f6-482f-aca9-
907c4acc71a9/50027e48-6cb9-4345-9c7a-c22b41ad84d2/images/5ada0ef6-5f4a-40b8-ad92-cb6758de8536/c22f4e68-439b-4a87-8e22-bc7d8e2391f1,if=none,id=drive-ide0-0-
0,format=qcow2,serial=b8-ad92-cb6758de8536,cache=none,werror=stop,rerror=stop,aio=native -device ide-drive,bus=ide.0,unit=0,drive=drive-ide0-0-0,id=ide0-0-0 -drive file=/rhev/data-
center/6927f974-c6f6-482f-aca9-907c4acc71a9/e0acfcc8-c020-413e-84cd-a93cb0ab9b2d/images/11111111-1111-1111-1111-111111111111/RHEV-
toolsSetup_3.0_12.iso,if=none,media=cdrom,id=drive-ide0-1-0,readonly=on,format=raw -device ide-drive,bus=ide.1,unit=0,drive=drive-ide0-1-0,id=ide0-1-0 -drive file=/rhev/data-
center/6927f974-c6f6-482f-aca9-907c4acc71a9/50027e48-6cb9-4345-9c7a-c22b41ad84d2/images/f52621e0-8b1e-47af-809c-45de2aa697fc/f77b5dd2-3141-4ea7-84fa-
e8cfffe9cff9,if=none,id=drive-virtio-disk0,format=qcow2,serial=af-809c-45de2aa697fc,cache=none,werror=stop,rerror=stop,aio=native -device virtio-blk-pci,bus=pci.0,addr=0x7,drive=drive-virtio-
disk0,id=virtio-disk0 -netdev tap,fd=27,id=hostnet0 -device rtl8139,netdev=hostnet0,id=net0,mac=00:1a:4a:23:11:0b,bus=pci.0,addr=0x3 -netdev tap,fd=29,id=hostnet1,vhost=on,vhostfd=30
-device virtio-net-pci,netdev=hostnet1,id=net1,mac=00:1a:4a:23:11:0c,bus=pci.0,addr=0x4 -chardev socket,id=charchannel0,path=/var/lib/libvirt/qemu/channels/z-win7x86-
1.com.redhat.rhevm.vdsm,server,nowait -device virtserialport,bus=virtio-serial0.0,nr=1,chardev=charchannel0,id=channel0,name=com.redhat.rhevm.vdsm -chardev
spicevmc,id=charchannel1,name=vdagent -device virtserialport,bus=virtio-serial0.0,nr=2,chardev=charchannel1,id=channel1,name=com.redhat.spice.0 -usb -spice port=5902,tls-
port=5903,addr=0,x509-dir=/etc/pki/vdsm/libvirt-spice,tls-channel=main,tls-channel=inputs -k en-us -vga qxl -global qxl-vga.vram_size=67108864 -device intel-
hda,id=sound0,bus=pci.0,addr=0x5 -device hda-duplex,id=sound0-codec0,bus=sound0.0,cad=0
```

- To manage multiple virtual machines you would need libvirt: virsh, virt-manager.
- To dynamically manage anything from a few VMs on a single host up to thousands of VMs on a cluster of hundreds of hosts using multiple storage targets − VDSM

Architecture and Implementation

- VDSM is the oVirt node agent, tailored for its needs
- It manages transient VMs (vm data stored centrally in db managed by oVirt)
- It is KVM centric
- moving to a more general use case, applicable to other management platforms

Architecture and Implementation



Architecture and Implementation

- Written in Python
- Multithreaded, multi-processes
- Speaks with its guest agent via virtio-serial
- Adds customized clustering support for LVM that scales to hundreds of nodes
  - Implements a distributed image repository over the supported storage types (local directory, FCP, FCoE, iSCSI, NFS, SAS)
  - Multihost system, one concurrent metadata writer
  - Scales linearly in data writers

Robustness as a Design Goal

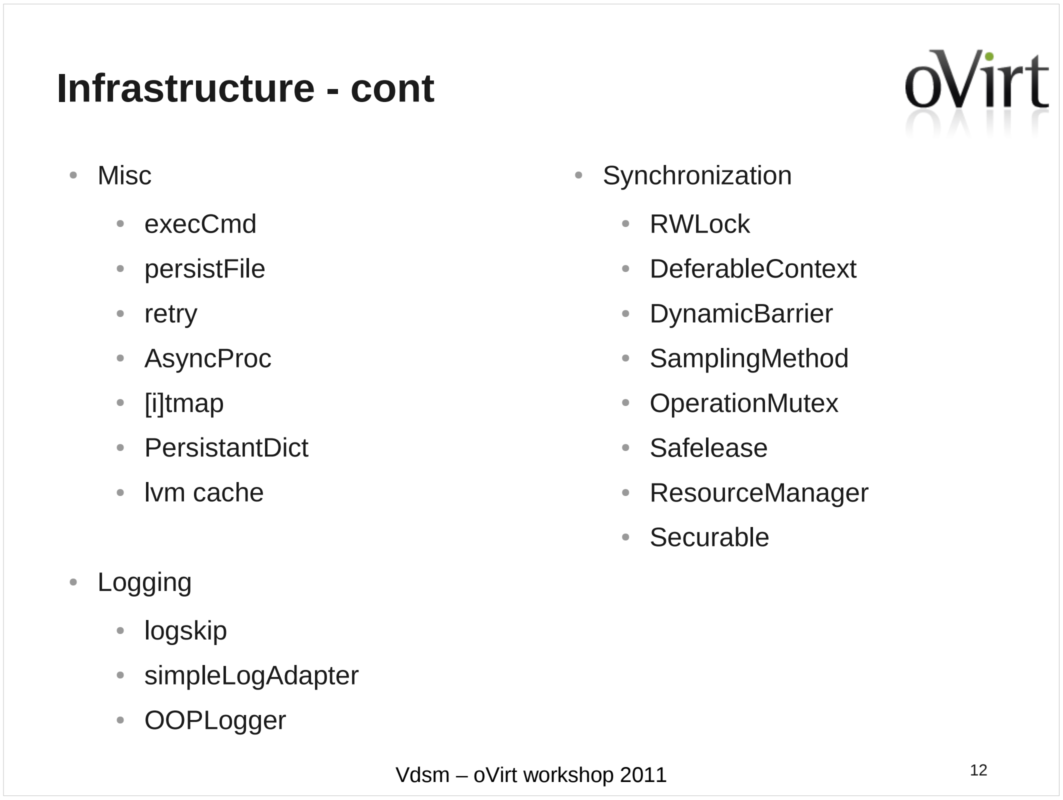
- Evaporated NFS exports
- Faulty paths
- Node crashes
- Live-locked qemu
- Internal Python exceptions
- Self-fencing of metadata writer

Packages

- vdsm
- vdsm_cli
- vds_bootstrap
- vdsm_reg
- vdsm_hooks

Infrastructure

- Supervdsm
- Out of process
- Async Tasks

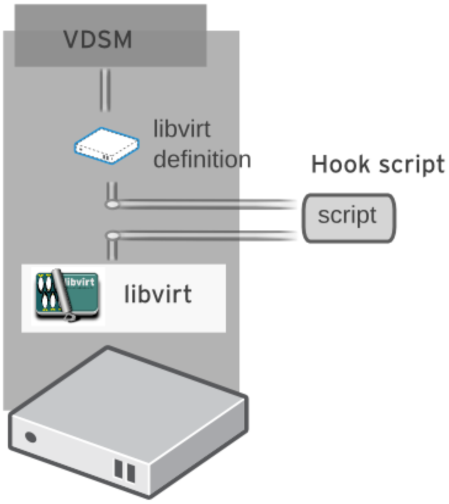


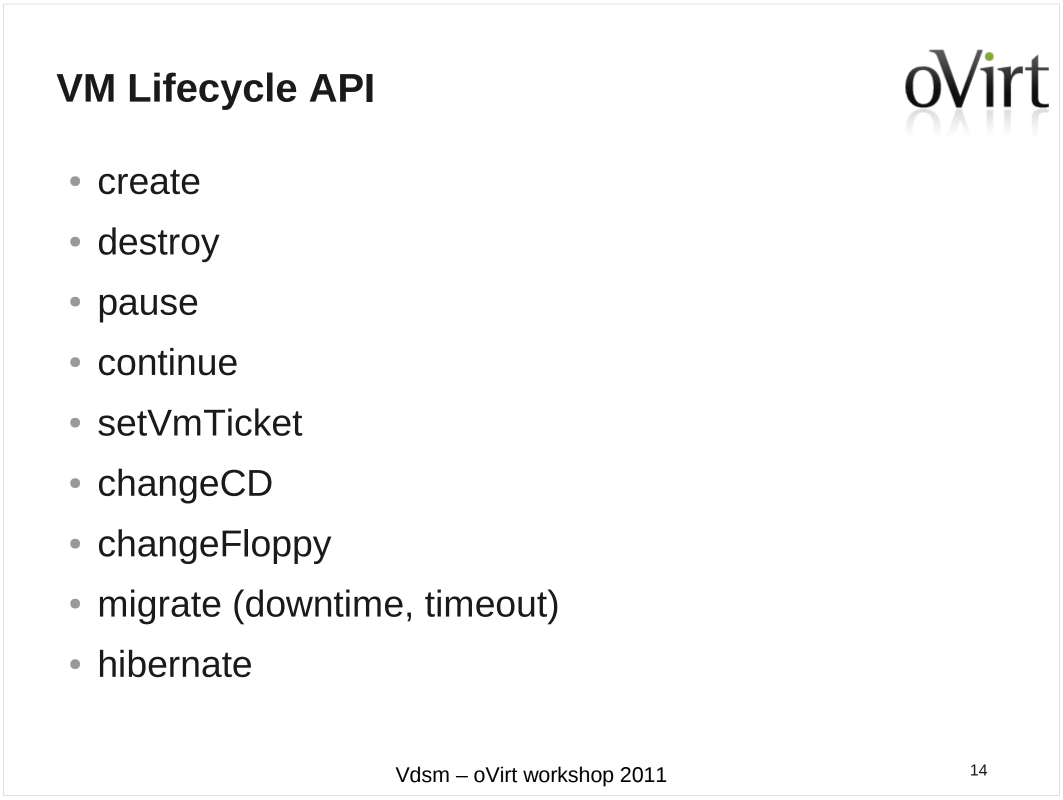
Infrastructure - cont

- Misc
  - execCmd
  - persistFile
  - retry
  - AsyncProc
  - [i]tmap
  - PersistantDict
  - lvm cache
- Logging
  - logskip
  - simpleLogAdapter
  - OOPLogger
- Synchronization
  - RWLock
  - DeferableContext
  - DynamicBarrier
  - SamplingMethod
  - OperationMutex
  - Safelease
  - ResourceManager
  - Securable

Hooks

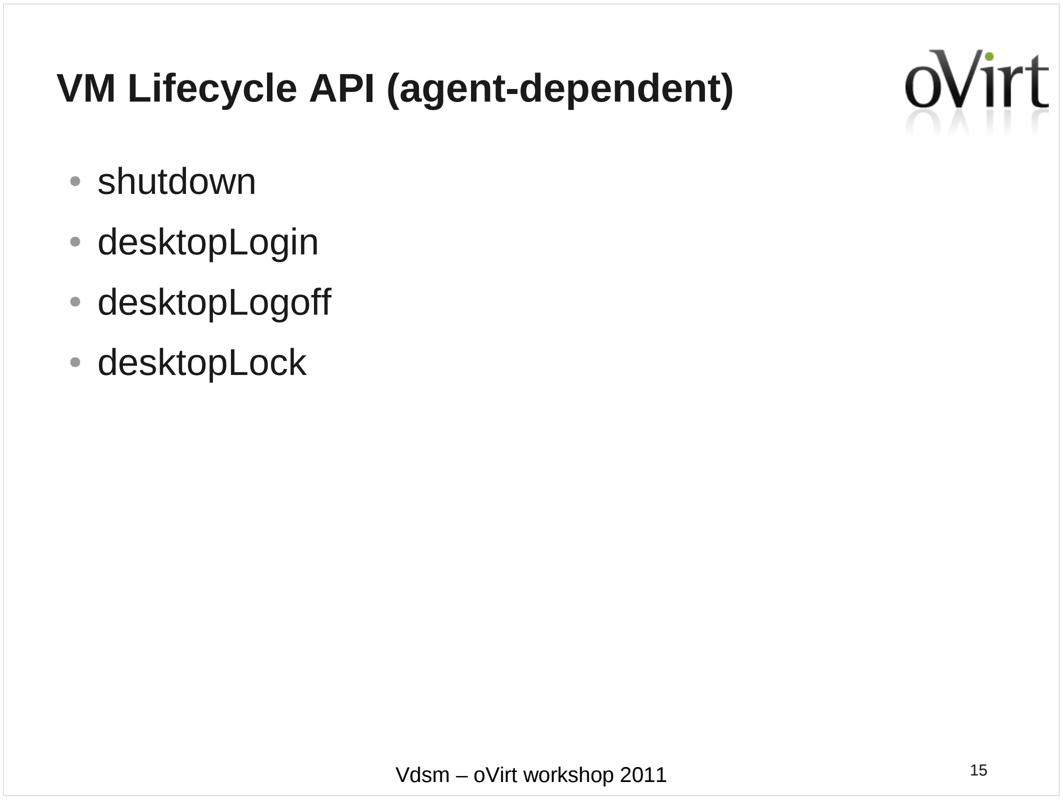
- VM lifecycle hooks
  - before/after vm_start
  - before/after vm_cont
  - before/after vm_pause
  - before/after vm_hibernate
  - before/after vm_dehibernate
  - before/after vm_migrate_source
  - before/after vm_migrate_destination
  - after_vm_destroy
- Vdsm lifecycle hooks
  - before/after vdsm_start





VM Lifecycle API

- create
- destroy
- pause
- continue
- setVmTicket
- changeCD
- changeFloppy
- migrate (downtime, timeout)
- hibernate



VM Lifecycle API (agent-dependent)

- shutdown
- desktopLogin
- desktopLogoff
- desktopLock

VM Monitoring API

- list
- getAllVmStats
- getVmStats
  - Interesting applications installed
  - Logged in users
  - CPU consumption
  - Memory usage

Network Config API

- AddNetwork
- DelNetwork
- EditNetwork
- SetSafeNetworkConfig
- SetupNetworks
- ConnectivityCheck

Host Monitoring API

- getVdsCapabilities
- getVdsStats
- ping
- fenceNode

Storage API

- connectStorageServer
- getDeviceList
- createStorageDomain
- attachStorageDomain
- createImage
- prepareVolume
-
-
- (and many, many more)

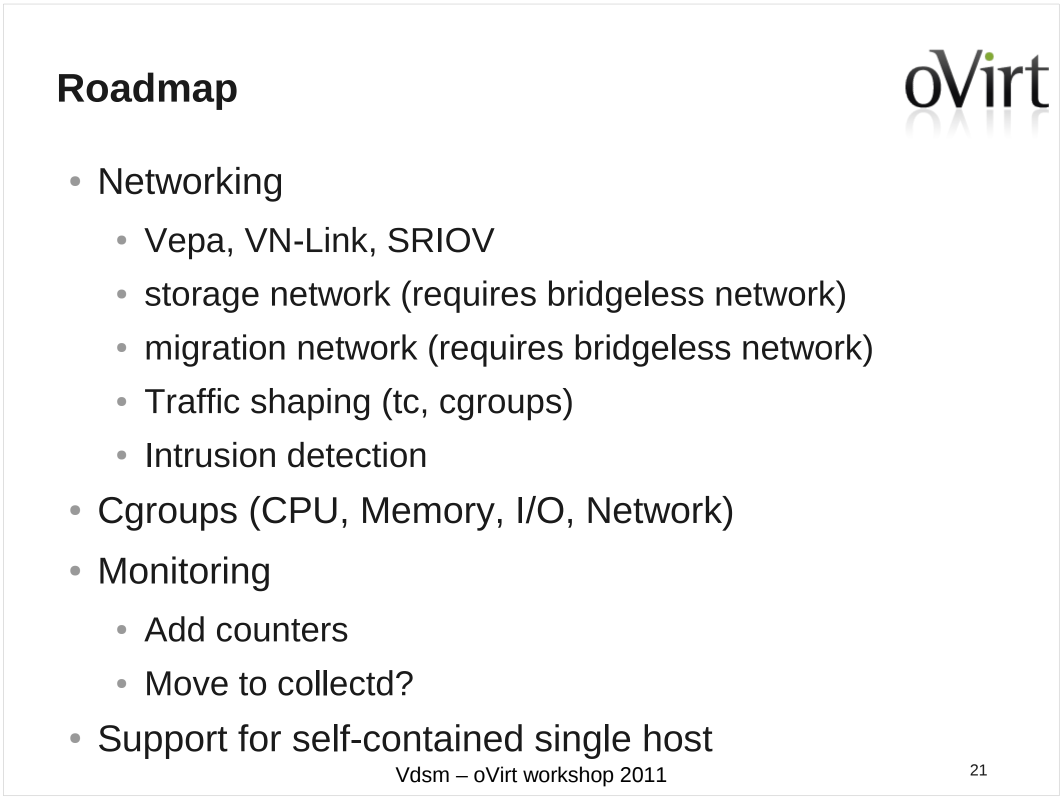
- repoStats

- getSpmStatus
- spmStart

- extendVolume

Async Tasks API

- GetAllTasksStatuses
- getTaskStatus
- clearTask
- stopTask



Roadmap

- Networking
  - Vepa, VN-Link, SRIOV
  - storage network (requires bridgeless network)
  - migration network (requires bridgeless network)
  - Traffic shaping (tc, cgroups)
  - Intrusion detection
- Cgroups (CPU, Memory, I/O, Network)
- Monitoring
  - Add counters
  - Move to collectd?
- Support for self-contained single host

Roadmap - cont

- New API
  - Current API is not very clean (createVG, createStorageDomain)
  - stable
  - RESTful?
  - oVirt-api look and feel
- Support sending events
  - QMF support
- Split VDSM up into reusable autonomous parts.
  - Spin storage off as a generic image repository.
  - Policy engine (MOM?)

How to contribute

- **Repository:**
  - http://git.fedorahosted.org/git/?p=vdsm.git
- **Mailing lists:**
  - vdsm-devel@lists.fedorahosted.org
  - vdsm-patches@lists.fedorahosted.org
- **IRC:**
  - #vdsm on Freenode
- **Core Team:**
  Dan Kenigsberg, Saggi Mizrahi, Igor Lvovsky, Eduardo Warszawasky, Yotam Oron, Ayal Baron

Q&A

oVirt

THANK YOU !

http://www.ovirt.org

Bootstrap

- Verifies node compatibility with oVirt
  - Check os/cpu/vdsm compatibility
  - Check RPMs (Install if needed)
  - Configure node (certificate, networking, services, etc.)
- Currently supports only RHEL 5.X and RHEL 6.X
- Working on support for Fedora