PRODUCTION READY VIRTUALIZATION WITH RED HAT

Andrew Sullivan
Principal Technical Marketing Manager

Martin Perina
Manager, Software Engineering

May 2019
RED HAT VIRTUALIZATION OVERVIEW

Directory service
- Active directory
- IPA
- Red Hat Directory Server
- IBM Tivoli Directory Server

RED HAT VIRTUALIZATION MANAGER
- Storage domain
- Console access
- Backend
- Web service
- REST API/Python SDK/Java SDK
- Web browser
- User portal

RED HAT VIRTUALIZATION HYPERVISOR
- Internal web service
- PostgreSQL
- HTTPS
- SSH/SSL
- VDSM
- libvirt
- SPICE or VNC

RED HAT ENTERPRISE LINUX
- Web app

#redhat #rhsummit
RED HAT VIRTUALIZATION MANAGER
RHV-M ARCHITECTURE
RHV-M deployed to separate host
  - Physical or virtual
  - **Not** hosted on and managing the same infrastructure
- High availability handled externally
- Deployed using standard yum practice
Deployed to same hypervisor hosts which are managed by the RHV-M instance
  ○ May use Cockpit interface, Ansible, or CLI for deployment
  ○ Hosted by either RHEL or Red Hat Virtualization Host (RHV-H) hypervisors
Intrinsic high availability using technology separate from the hypervisor clusters/datacenter
RHV-M GUIDELINES

- Standard vs self-hosted
- Necessary for VM HA operations!
- Scale
  - Tested up to 500 hosts, 5000 VMs
  - Separate the database from the RHV-M application for optimal performance
COMPUTE RESOURCES
HYPERVERSOR HOSTS

- Two different hypervisor “models”
  - Appliance: Red Hat Virtualization - Host (RHV-H)
  - Traditional OS: Red Hat Enterprise Linux (RHEL) w/RHV packages
- Both result in the same capabilities!
  - RHV-H has a smaller footprint, having only what’s needed to be a hypervisor
- Configuration and management are both handled the same by RHV-M
  - Updates/upgrades, power management, etc. all equivalent
  - Logical entities (e.g., networks and storage) are created and managed the same
- Do you want/need to customize the hypervisor OS layout and/or package set extensively?
  - Yes - RHEL
  - No - RHV-H
## RHV-H OR RHEL?

<table>
<thead>
<tr>
<th>Red Hat Enterprise Linux</th>
<th>Red Hat Virtualization Host</th>
</tr>
</thead>
<tbody>
<tr>
<td>● Manage using traditional tools and practices</td>
<td></td>
</tr>
<tr>
<td>○ Monitoring and alerting</td>
<td></td>
</tr>
<tr>
<td>○ Automation</td>
<td></td>
</tr>
<tr>
<td>● Customize to your organization’s needs</td>
<td></td>
</tr>
<tr>
<td>○ Storage layout</td>
<td></td>
</tr>
<tr>
<td>○ Packages</td>
<td></td>
</tr>
<tr>
<td>● Cockpit optional (but recommended)</td>
<td></td>
</tr>
<tr>
<td>● Fewer packages, smaller footprint</td>
<td></td>
</tr>
<tr>
<td>● Hands-off, appliance management model</td>
<td></td>
</tr>
<tr>
<td>○ Cockpit required</td>
<td></td>
</tr>
<tr>
<td>○ Automatically tuned for virtualization workloads</td>
<td></td>
</tr>
<tr>
<td>● Limited customization</td>
<td></td>
</tr>
<tr>
<td>○ Add packages only if needed</td>
<td></td>
</tr>
<tr>
<td>○ Storage layout is semi rigid</td>
<td></td>
</tr>
</tbody>
</table>

USE CASE DETERMINES WHICH ONE IS BEST FOR YOUR ENVIRONMENT
HYPERVISOR ARCHITECTURE
RESOURCES

- A datacenter has 1 or more clusters
- Clusters are composed of 1 or more hosts
- VMs are hosted by the clusters and can be migrated to any host in the cluster
- All hosts in the cluster must access the same physical networks
- All hosts in the datacenter must have access to the same shared storage

<table>
<thead>
<tr>
<th>Datacenter</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>VMs</td>
<td>![VMs image]</td>
<td>![VMs image]</td>
</tr>
<tr>
<td>Logical networks</td>
<td>![Logical networks image]</td>
<td>![Logical networks image]</td>
</tr>
<tr>
<td>Hosts</td>
<td>![Hosts image]</td>
<td>![Hosts image]</td>
</tr>
<tr>
<td>Physical network</td>
<td>![Physical network image]</td>
<td>![Physical network image]</td>
</tr>
<tr>
<td>Shared storage resources</td>
<td>![Shared storage resources image]</td>
<td>![Shared storage resources image]</td>
</tr>
</tbody>
</table>
## LIMITS

<table>
<thead>
<tr>
<th>COMPONENT</th>
<th>LIMIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logical CPUs</td>
<td>768</td>
</tr>
<tr>
<td>RAM</td>
<td>12TB</td>
</tr>
<tr>
<td>VMs</td>
<td>Unlimited</td>
</tr>
<tr>
<td>Hosts</td>
<td>400</td>
</tr>
<tr>
<td>VMs</td>
<td>Unlimited</td>
</tr>
<tr>
<td>CPUs</td>
<td>240</td>
</tr>
<tr>
<td>RAM</td>
<td>4TB</td>
</tr>
</tbody>
</table>

Maximum per hypervisor

Maximum per cluster

Maximum per VM
VIRTUAL MACHINES

- RHEL and Microsoft Windows (server and desktop) operating systems supported
- Templates and instance sizing helpers to simplify provisioning
- Optimization profiles automatically pre-tune the virtual machine for the workload
  - Desktop, server, high performance
- VMs support Non-uniform memory access (NUMA) pinning, CPU pinning, CPU masking and passthrough, large (2MiB) and huge pages (1GiB)
- VM resources may be reserved or allowed to overcommit, as desired
- High availability is enabled on a per-VM basis
  - Specify startup priority
- May have as many disks, NICs, etc. as the underlying virtual hardware (i440fx, Q35, etc.) will support
- PCI passthrough and SR-IOV for direct access to host resources
GUIDELINES

● Create a datacenter when
  ○ Isolating compute resources which do not share network or storage

● Create a cluster when
  ○ Different configuration (e.g. memory overcommit) and/or hardware
  ○ Isolating virtual machines for resources or permissions
    ■ Use QoS policies to limit resource consumption by a group of VMs
    ■ Use quotas to limit resource consumption by a group of users
  ○ No cross-cluster live migration

● Cluster configuration guidelines
  ○ Use the highest common CPU version available
  ○ Use IBRS SSBD Spectre/Meltdown mitigations

● Virtual machine configuration guidelines
  ○ Use Q35 when ...
NETWORK
- Single/standalone or bonded interfaces
  - Bond modes 1 (active-passive), 2 (XOR), 3 (broadcast), and 4 (802.3ad / LACP) are supported for virtual machine networks
  - Modes 0 (round robin), 5 (xmit LB), and 6 (xmit + rcv LB) are not supported for virtual machine networks
- Bonds of bonds are supported (e.g., 2 mode-1 bonds aggregated into a mode-2 bond)
Logical networks are created “on top” of physical interfaces
  ○ Multiple logical networks (e.g., VLAN) may be associated with each physical interface

Linux bridge devices are created for each logical network on the interface

Pre-defined traffic type isolation:
  ○ Management (communication with RHV-M and administrator access)
  ○ Virtual machine
  ○ Display (VM console)
  ○ Migration (live migration)
  ○ Gluster

Additional logical networks may be created for any reason (e.g., iSCSI or NFS)
Virtual network interface card (VNIC) profiles are created to modify characteristics of how the host or virtual machine connects to the logical network:
- Network filter - apply rules to how traffic is handled to/from the VNIC
- Passthrough - for VNF / SR-IOV functionality
- Port mirroring - mirrors traffic to another port for analysis

QoS policies place controls on network min/max throughput and share prioritization:
- Rate limit
- Committed rate
- Weighted share
- QoS policies may be defined on a per-(logical) network basis or on a per-VNIC profile basis

VNIC profiles may apply to both VM and host logical network interfaces.
NETWORK FILTERS

- Enforce rules about how VNIC traffic is handled, what is and is not allowed
  - Example: prevent packets not matching the VM's MAC from passing outbound (MAC spoofing)
- Applied at the hypervisor
- Commonly used filters:
  - vdsrn-no-mac-spoofing (default) - prevent MAC spoofing
  - clean-traffic - prevent MAC, IP, and ARP spoofing
  - clean-traffic-gateway - As with clean-traffic, but only allow communication to provided MAC(s) as the gateway. Used for private VLAN functionality.
  - No filter - allows all traffic, unaffected, to/from the VNIC
GUIDELINES

● Most important: work with your network administrator!
● At a minimum, logically isolate network traffic
  ○ Management, migration, IP storage, virtual machine
● Physically isolate traffic types if possible
  ○ Multiple 1GbE links possible, but depends heavily on other factors
  ○ Multiple 10GbE links recommended when possible
● Physical redundancy at a minimum (mode 1), redundancy and increased throughput ideally (mode 4)
  ○ Use multiple upstream switches
  ○ Link aggregation increases total available bandwidth
  ○ Individual sessions are still limited to the maximum speed of a single link
● Be conscious of East-West vs North-South traffic flows
● SDN provides logical isolation for virtual machines without additional VLANs, but adds complexity to troubleshooting. Work with your network admin!
STORAGE DOMAINS

- Storage domains exist at the datacenter level
- One Storage Pool Manager (SPM) per datacenter manages metadata
- Data storage domain - used for virtual machine data disks, leases, ISOs, cluster/datacenter metadata
- Export and ISO domains are deprecated

STORAGE TYPE

- Block - uses SCSI commands to communicate with the underlying disks. Objects are stored as logical volumes.
- File - uses file protocols (e.g., NFS) to communicate with a remote storage host. Objects are stored as files.

Both are valid. Choose based on application requirements
STORAGE DOMAINS

- Up to 40 storage domains per datacenter
  - No limit to storage domain size
- VM disks limited to 500TiB
- Block (iSCSI / FCP) storage
  - Thick / preallocated VM disks by default
  - Up to 1,300 VM disks + snapshots + other objects per domain
  - LUN size, count, and path count limits are same as for RHEL
- File (NFS) storage
  - Thin / sparse VM disks by default
- Thick / preallocated disks are recommended for high-performance VMs which write more than 1GB every 4 seconds
- Snapshots are always sparse
STORAGE POOL MANAGER (SPM)

- SPM is a role assigned to a host in the datacenter
  - One host per datacenter
  - Can run on any host
  - All hosts in the datacenter must be able to access all storage

- Coordinates metadata operations across the storage domains
  - Create / delete / modify virtual disks
  - Create / delete / modify snapshots
  - Allocate (extend) capacity for sparse disks (expanded in 1GB increments)
  - Copy / move images between storage domains

- RHV-M will move SPM if it becomes unresponsive (e.g., node failure) or it cannot reach all domains
VIRTUAL MACHINE DISKS

- Multiple disk options
  - Disk image - hosted in a storage domain
  - Direct LUN - a SAN LUN passed to the VM
  - Cinder volume - a Cinder provisioned volume connected to the VM
- Image based disks have several disk interface options:
  - IDE, VirtIO-SCSI, VirtIO
- Disk allocation
  - Thin / sparse disks = QCOW2
  - Thick / preallocated disks = raw
- Discard is supported to keep thin disks thin
- Shared disks are supported
- SCSI reservations are supported (Windows Server Failover Clustering)
- Per-disk profiles can be applied for QoS
GUIDELINES

● Most important: work with your storage admin!
● Follow vendor recommendations for RHV and RHEL
● Domain size should be limited by the ability to protect data and use vendor features
● Choose the protocol which fits your needs
  ○ Performance, reliability, knowledge
● Choose storage-level disk failure protection (e.g. RAID) according to needs
  ○ RAID 1/5/6/10/DP/TEC/Z/Z2, Erasure Coding, etc
  ○ Lots of vendors with lots of features and lots of implications
● Storage latency can come from multiple sources
  ○ Virtual machine
  ○ Physical host
  ○ Storage network
  ○ Storage array
● Snapshots are copy-on-write, this has performance and capacity implications!
OTHER CONSIDERATIONS
RHV-M HIGH AVAILABILITY

- Standard or traditional deployment
  - Administrator provides HA, no intrinsic capability
- Self-hosted deployment
  - Hosts are added to a RHV-M resource cluster
    - Cluster is managed via Cockpit
  - Resource cluster exists outside of standard RHV-M managed datacenter and cluster paradigm
  - Intrinsic HA mechanism will ensure RHV-M is running
  - Host is chosen by score, score determined by multiple factors
    - Gateway response, up-to-date VM config, management network bridge status, available memory, CPU load, previous failed attempts

--- Host 1 status ---

<table>
<thead>
<tr>
<th>Status</th>
<th>True</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hostname</td>
<td>hypervisor.example.com</td>
</tr>
<tr>
<td>Host ID</td>
<td>1</td>
</tr>
<tr>
<td>Engine status &quot;up&quot;</td>
<td>{&quot;health&quot;: &quot;good&quot;, &quot;vm&quot;: &quot;up&quot;, &quot;detail&quot;:}</td>
</tr>
<tr>
<td>Score</td>
<td>3400</td>
</tr>
<tr>
<td>stopped</td>
<td>False</td>
</tr>
<tr>
<td>Local maintenance</td>
<td>False</td>
</tr>
<tr>
<td>crc32</td>
<td>99e57eba</td>
</tr>
<tr>
<td>Host timestamp</td>
<td>248542</td>
</tr>
</tbody>
</table>
- HA is configured on a per-VM basis
- Configure a watchdog to monitor guest operating system
- Host failure detected using multiple methods:
  - Fencing - relies on out-of-band connectivity via a fence agent; host failure is validated using agent
  - Leases - host creates a per-VM lease on shared storage; when lease is lost, VM is assumed dead
- RHV-M must be running for VM HA
ACTIVE-ACTIVE DISASTER RECOVERY

- Cluster is “stretched” across multiple sites
- Automatic failover, no manual process
- VM soft-affinity rules keep VMs at primary site until failure
- Storage must be synchronously replicated and available at both sites
- Layer 2 network (VLAN) must be available at both (no network re-map)
- RHV-M must be recovered first!
  - Use self-hosted RHV-M with nodes at both sites
ACTIVE-PASSIVE DISASTER RECOVERY

- Ansible-based, manual failover between primary and backup sites
- Asynchronous storage replication: Replicate based on your RTO
- Networks do not have to be the same
- Failover process will:
  - Mount storage domains
  - Discover VMs
  - Remap networks, affinity and anti-affinity, and direct mapped LUNs
  - Restart VMs according to HA priority
- Use Ansible to reverse DR if primary site returns to service
ON THE HORIZON
TECH PREVIEW

● Network
  ○ Cluster support for Open vSwitch

● Storage
  ○ Storage offload and integration via CinderLib
  ○ 4k native disk support for local datastores
  ○ Shared and local domains in the same datacenter

● Backup
  ○ Changed block tracking for virtual machines

● RHV-M
  ○ HTML5 console
  ○ Nested virtualization
  ○ Import Debian and Ubuntu VMs from VMware an Xen
  ○ Mobile app: moVirt
THANK YOU

linkedin.com/company/Red-Hat
youtube.com/user/RedHatVideos
facebook.com/RedHatinc
twitter.com/RedHat