RHEV Hypervisor 7: Now and the Future

Karen Noel
Senior Software Engineering Manager
Platform Engineering – Virtualization Team
June 2015
Storyline

• Engineering at Red Hat – Virt Team

• KVM Performance & Scaling

• RHEV Hypervisor 7 - Feature Roadmap
Virt Team Engineering
RHEV Hypervisor – best engineered distribution of KVM!
RHEV Hypervisor & Virt Team Contribution

*Do not confuse with RHEV-H iso used with RHEV-M only
RHEV Hypervisor 7 - RHEL 7.1 example

- Rebase to upstream QEMU 2.1.2 for RHEL 7.1
  - Disable some devices (-)
  - Disable certain features (-)
  - Backport certain features: QEMU 2.2 (+)
  - Backport bugfixes: QEMU 2.2 (+)

- RHEV and RHEL OpenStack Platform
  - qemu-kvm-rhev-2.1.2*.rpm
  - Backport additional fixes (+)
<table>
<thead>
<tr>
<th>RHEV Hypervisor 7</th>
<th>RHEL 7</th>
<th>RHEL 7.1</th>
<th>RHEL 7.2</th>
<th>RHEL 7.3</th>
<th>......</th>
</tr>
</thead>
<tbody>
<tr>
<td>qemu-kvm-rhev</td>
<td>QEMU 1.5.3</td>
<td>QEMU 2.1.2</td>
<td>QEMU 2.3*</td>
<td>QEMU 3.0*</td>
<td></td>
</tr>
<tr>
<td>Included in:</td>
<td>Unsupported Features &amp; Devices Disabled</td>
<td>Unsupported Features &amp; Devices Disabled</td>
<td>Unsupported Features &amp; Devices Disabled</td>
<td>Unsupported Features &amp; Devices Disabled</td>
<td></td>
</tr>
<tr>
<td>- RHEV</td>
<td>Backport Bug fixes</td>
<td>Backport Bug fixes</td>
<td>Backport Bug fixes</td>
<td>Backport Bug fixes</td>
<td></td>
</tr>
<tr>
<td>- RHEL OpenStack Platform</td>
<td>Backport some Features up to QEMU 2.0</td>
<td>Backport few Features up to QEMU 2.2</td>
<td>Backport few Features up to QEMU 2.5*</td>
<td>Backport few Features up to QEMU 3.2*</td>
<td></td>
</tr>
<tr>
<td>Advanced Features - enabled</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>* Exact versions TBD</td>
</tr>
</tbody>
</table>
RHEV Hypervisor 7 Features

• Adopted by RHEV
  – Many OpenStack new features → hardened by RHEV

• Adopted by RHEL OpenStack Platform
  – KVM is the #1 hypervisor for OpenStack

RHEV Hypervisor – best engineered distribution of KVM!
RHEV Hypervisor Performance & Scaling
RHEV Hypervisor – Fastest and Biggest!
* 10 of 14 results are Red Hat
* All results are KVM, except 2

SPEC® and SPEC virt_sc are registered trademarks of the Standard Performance Evaluation Corporation. For more information about SPEC and its benchmarks see www.spec.org. For details about the SPECvirt_sc 2013 benchmark results in this chart see www.spec.org/virt_sc2013/results.
SPEC virt_sc® 2013 Benchmark Leadership

Highest Red Hat based result versus highest non Red Hat result by socket
(As of June 10, 2015)

<table>
<thead>
<tr>
<th></th>
<th>2-Socket</th>
<th>4-Socket</th>
<th>8-Socket</th>
</tr>
</thead>
<tbody>
<tr>
<td>FusionSphere V100R005</td>
<td>632 @ 37 VMs</td>
<td>2,086 @ 121 VMs</td>
<td></td>
</tr>
<tr>
<td>RHEL 7 (KVM)</td>
<td></td>
<td></td>
<td>3,790 @ 210 VMs</td>
</tr>
</tbody>
</table>

SPEC® and SPEC virt_sc are registered trademarks of the Standard Performance Evaluation Corporation. For more information about SPEC and its benchmarks see www.spec.org. For details about the SPEC virt_sc 2013 benchmark results in this chart see www.spec.org/virt_sc2013/results.
RHEV Hypervisor 7 – Feature Roadmap
KVM, now and in the Future!
Live Migration – RHEV 3.5 & RHEL OSP 6

- RHEL 6 → RHEL 7 hosts
- RHEL 7.x ↔ RHEL 7.y hosts
- RHEL 6 machine types supported on RHEV Hypervisor 7
- Massive improvements in live migration technology!
- Joint engineering collaboration
  - Red Hat: RHEV, Nova, libvirt, QEMU, Virt QE
  - Key partners
Live Snapshot Active Merge – RHEV 3.5.1

- Live snapshots: capture disks and memory at a point in time - qcow2 volume chains
- RHEV Hypervisor 7.1
  - Block mirroring
  - Live snapshot merge, forward <=> backward, internal or **active layer**
  - Live snapshot delete
- Joint engineering collaboration
  - Red Hat: RHEV-M, vdsm, libvirt, QEMU, field account managers
  - Customer requirements
RHEV Hypervisor 7 - NFV Roadmap (now + future)

- NFV Configurations:
  - Config I – NUMA Pinning, Topology Awareness, SR-IOV, DPDK
  - Config II – DPDK with Open vSwitch, vhost-user
  - Future – In planning
    - OPNFV: [https://wiki.opnfv.org/nfv_hypervisors-kvm](https://wiki.opnfv.org/nfv_hypervisors-kvm)
- Real-time KVM – low avg/max latency requirements
NUMA Pinning/Topology Awareness, SR-IOV

- RHEL OpenStack Platform 7 (plan)
  - Nova scheduler NUMA support – based on PCI bus location
- RHEV Hypervisor 7.1
  - VFIO device assignment (PCI passthrough)
  - Libvirt API – host NUMA topology
  - Huge pages allocated to VM – 2MB and 1GB – NUMA awareness
  - Guest NUMA node <=> host NUMA node alignment
- Joint engineering collaboration
  - Red Hat teams: OpenStack Nova, libvirt, QEMU, kernel, perf
  - NFV partners
NUMA Pinning and Topology Awareness

Virtual Machine

NUMA Node 0

Socket

CPU core 0

CPU core 1

CPU core 2

CPU core 3

Per NUMA Node huge pages

vCPU

vCPU

vCPU

vCPU
NUMA Pinning and Topology Awareness – two sockets

- NUMA Node 0
  - CPU core 0
  - CPU core 1
  - CPU core 2
  - CPU core 3
  - vCPU
  - vCPU

- NUMA Node 1
  - CPU
  - CPU
  - CPU
  - vCPU
  - vCPU

Per NUMA Node huge pages

Virtual Machine

Socket
Config I - DPDK with VFIO device assignment (future)
Network Function Virtualization (NFV)
Throughput and Packets/sec (RHEL7.x+DPDK)

NFV: Millions of Packets Per Second
RHEL7.x, L2 Forwarding, 12 x 40Gb NiCs

- KVM: 208Mpps+
- Docker: 215
- Bare-metal: 218
- HW Maximum: 225

#redhat #rhsummit
#### Config II - DPDK w/ Open vSwitch, vhost-user (future)

- **Server Hardware**: Intel XL710
- **RHEL 7 host kernel**: KVM, VFIO
- **Open vSwitch**: OVS-dpdk
- **QEMU**: vhost-user
- **Libvirt**:
  - **Linux Guest**: dpdk-lib, uio_pci_generic
  - **QEMU**: vhost-user
- **OpenStack Nova**:
  - **Linux Guest**: dpdk-lib, uio_pci_generic
  - **QEMU**: vhost-user
- **DPDK apps**
- **Linux Guest**
- **Linux Guest**
- **Linux Guest**
- **Linux Guest**
- **Linux Guest**

#redhat #rhsummit
OPNFV KVM Hypervisor Project (future)

https://wiki.opnfv.org/nfv_hypervisors-kvm
Real-Time KVM for NFV (future)

- Low average and max latency requirements – **no latency spikes**!
- RHEL-RT in host and guests
- New tuned profiles in host and guest
- RHEV Hypervisor 7.2 (plan)
  - KVM and kernel patches
  - Libvirt support
- Joint engineering collaboration
  - Red Hat: Nova, libvirt, KVM, kernel
    - DPDK, Open vSwitch, Real-time
    - performance, tuned
  - NFV Partners
RHEL7.x Real-time Scheduler Latency Jitter Plot

10 Million Samples
RHEL7.x Scheduler Latency (cyclictest)
Intel Ivy Bridge 2.4 Ghz, 128 GB mem

Cyclictest Latency

Remove maxes to zoom in
KVM Real-Time Patches (as of 18-Jun-2015)

- Kernel patches
  - vmstat: 6
  - timer: 4
  - sched: 45
  - rcu, nohz, kvm: 12
  - cpusets, isolcpus: 2
  - workqueue: 4
  - kvm: 4
  - kernel-rt: 4

- Non-kernel patches
  - irqbalance: 3
  - tuned profile: 9
  - tuned adm: 1
  - python-linux-procfs: 1
  - libvirt: 6
Virtual Storage Improvements (future)

- Single VM scaling w/IOthreads:
  - fine-grained locking in QEMU
  - virtio data-plane
  - IOthread pinning
  - thread per disk/initiator or host CPU
- Multi-queue QMEU block layer
- Fine grained block layer config
RHEV for Power Little Endian (future)

POWER Server Hardware
Linux kernel LE
OPAL/Sapphire

libvirt LE
RHEL 6.5+ guest BE
RHEL 7.1+ guest BE
RHEL 7.1+ guest LE

KVM vfi vhost
QEMU LE
vdsm LE

RHEL 7.2 LE
Linux kernel LE

RHEL 7.2 LE

POWER Server Hardware

RHEV-M x86_64

IBM

redhat rhsummit
## RHEV Hypervisor 7 - highlights

<table>
<thead>
<tr>
<th>RHEV host</th>
<th>Integration</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>RHEL 7+</td>
<td>RHEV and OSP</td>
<td>Live migration from RHEL 6 to RHEL 7 host</td>
</tr>
<tr>
<td>RHEL 7+</td>
<td>RHEV and OSP</td>
<td>Live migration improvements</td>
</tr>
<tr>
<td>RHEL 7.1</td>
<td>RHEV 3.5.1</td>
<td>Live snapshot active merge</td>
</tr>
<tr>
<td>RHEL 7.1</td>
<td>RHEL OSP 7*</td>
<td>NFV: huge pages, NUMA, pinning, SR-IOV</td>
</tr>
<tr>
<td>RHEL 7.2*</td>
<td>RHEL OSP 8*</td>
<td>NFV: vhost-user – for OVS and DPDK</td>
</tr>
<tr>
<td>RHEL 7.2*</td>
<td>RHEL OSP 8*</td>
<td>NFV: real-time KVM</td>
</tr>
<tr>
<td>RHEL 7.2*</td>
<td>RHEV 3.6*</td>
<td>IOthreads – virtio-blk data-plane</td>
</tr>
<tr>
<td>RHEL 7.2*</td>
<td>RHEV 3.5/6*</td>
<td>RHEV for Power, RHEV Hypervisor LE Host</td>
</tr>
</tbody>
</table>

*RHEV Hypervisor, RHEV and OSP versions are specified for planning purposes only*
# RHEV Hypervisor – future plans

<table>
<thead>
<tr>
<th>RHEV host</th>
<th>Integration</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>RHEL 7.2*</td>
<td>RHEV 3.6*</td>
<td>Flexibility: Memory hot-plug</td>
</tr>
<tr>
<td>RHEL 7.2*</td>
<td>RHEV and OSP</td>
<td>Windows 10, Server 2016 guest support</td>
</tr>
<tr>
<td>RHEL 7.3*</td>
<td>RHEV</td>
<td>IOthreads – virtio-scsi</td>
</tr>
<tr>
<td>RHEL 7.3*</td>
<td>RHEV</td>
<td>Error reporting/recovery: pci-express</td>
</tr>
<tr>
<td>RHEL 7.3*</td>
<td>RHEV</td>
<td>OVMF with UEFI secure boot</td>
</tr>
<tr>
<td>RHEL 7.3*</td>
<td>RHEV</td>
<td>Post-copy live migration</td>
</tr>
<tr>
<td>RHEL.next</td>
<td>RHEL OSP</td>
<td>OpenStack on ARM</td>
</tr>
<tr>
<td>RHEL.next</td>
<td>RHEV and OSP</td>
<td>Nested KVM – tech preview</td>
</tr>
<tr>
<td>RHEL.next</td>
<td>RHEV and OSP</td>
<td>Virtual GPU (compute and graphics)</td>
</tr>
</tbody>
</table>

*RHEV Hypervisor, RHEV and OSP versions are specified for planning purposes only.
Speakers:
- Red Hat
- Linaro
- Google
- INRIA
- ARM
- Siemens
- GE
- IBM
- SUSE
- Fujitsu
- Intel
- Huawei
- Xilinx
- Samsung
- Montavista
- op5

Communities:
- KVM
- QEMU
- libvirt
- Kernel
- OVMF
- OpenStack
- Nova
- oVirt
LEARN. NETWORK. EXPERIENCE OPEN SOURCE.