Painless switch from proprietary hypervisor to QEMU/KVM

Denis V. Lunev
den@openvz.org
Contents

● Some words about Parallels Cloud Server 6
● Choices to be made
● Performance results
● Status and future plans
Parallels Cloud Server

Proprietary details
Parallels Cloud Server

- Cross-platform hypervisor (Mac, Linux, Windows)
- HAL architecture
- Loose integration with host kernel
- Tight integration with guest OS
- Desktop origin
PCS: Some good stuff

- Very good Windows support
- 3D support in guest (Direct3D to OpenGL conversion)
- Decent guest tools
- Good support for block job operations
Architectures

**PSBM**

- **host user-space**
  - prl_vm_app
  - devices: host API access

- **host kernel-space**
  - hypervisor
  - resource

- **separate space**
  - monitor
  - vmstart
  - devices: vmexit processing

**KVM**

- **host user-space**
  - qemu
  - app
  - devices: vmexit processing

- **host kernel-space**
  - kvm
  - resource
  - vmstart
  - devices: vmexit processing

- Fast processing of devices commands that doesn’t require host device access
- No transparent access to resources
- Long switch between getting device command and passing it to a host device

- Most of devices operations require inter-component switch
- Native access to resources
PCS: Missed server features

- VM memory locked
- Memory deduplication (KSM)
- NUMA support (host/guest)
- SPICE support including remote USB
- Paravirtualized storage for the Linux guest
- PCI passthrough (implemented, not production)
- Nested virtualization (implemented, not production)
## VM limits

<table>
<thead>
<tr>
<th>PCS</th>
<th>QEMU/KVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>32 VCPU</td>
<td>255 VCPU</td>
</tr>
<tr>
<td>128 GB RAM</td>
<td>2 TB RAM</td>
</tr>
<tr>
<td>25 storage devices</td>
<td>Flexible configuration with</td>
</tr>
<tr>
<td>total</td>
<td>ability to setup necessary</td>
</tr>
<tr>
<td>16 network cards</td>
<td>amount of any devices</td>
</tr>
<tr>
<td></td>
<td>controllers</td>
</tr>
</tbody>
</table>
The switch: options

Hard decision to made
Switch options (low level)

<table>
<thead>
<tr>
<th>KVM + QEMU</th>
<th>KVM + Parallels App</th>
</tr>
</thead>
<tbody>
<tr>
<td>● More devices supported</td>
<td>● Backups</td>
</tr>
<tr>
<td>● QEMU is dumb and simple, does not make decisions</td>
<td>● Space effective block jobs</td>
</tr>
<tr>
<td>● Community</td>
<td>● Better Windows performance/support</td>
</tr>
<tr>
<td>● Stability</td>
<td>● Online shrink without PUNCH_HOLE on the host</td>
</tr>
<tr>
<td>● Libvirt/OpenStack</td>
<td>● <strong>PCS6 compatibility: live migration</strong></td>
</tr>
<tr>
<td>● Better Linux performance/support</td>
<td></td>
</tr>
<tr>
<td>● Better VM limits</td>
<td></td>
</tr>
</tbody>
</table>

Virtuozzo
Decision

- KVM + Parallels emulation
- 6 man/months to implement
- Lot of integration problems
- Pour performance
- Good progress?
Final decision

- Company split!
- Server team: no rights for latest Parallels code base
- Virtualization team reduction: 25 -> 5
- Live migration requirement from PCS6 to VZ7 was dropped!
- Natural choice: KVM + QEMU
Consequences

- Emulation level working
- Completely new middleware
- Virtualization team could work on features
Final Virtuozzo 7 architecture

Vzkernel 3.10 (cgroups, namespaces, etc.)
Management level constraints

- Command line kept compatible with PCS6
- PCS6 configuration file is kept
- PCS6 configuration <-> QEMU domain XML
- Domain XML keeps priority
- Dispatcher is driven by Libvirt events
Switch options (upstream)

- The kernel - RedHat 7, 3.10
  - Security
  - Hardware support (!)
- QEMU - RHEV, 2.1 (later 2.3)
  - Security
  - Actually does not matter, follow kernel
# Switch options (disc format)

<table>
<thead>
<tr>
<th>QCOW2</th>
<th>Parallels</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Public</td>
<td>• Not fully implemented</td>
</tr>
<tr>
<td>• In-place conversion from Parallels available</td>
<td>• Simpler =&gt; faster</td>
</tr>
<tr>
<td>• Compressed images</td>
<td>• Space effective block jobs</td>
</tr>
<tr>
<td>• Internal snapshots</td>
<td>• LBA padding (Win2k3)</td>
</tr>
<tr>
<td></td>
<td>• Dirty bitmaps support</td>
</tr>
</tbody>
</table>
Upgrade path

- “Hot” migration of the disk from PCS6 to Vz7
- Register new VM
  - Guest re-configuration via v2v (in-place mode)
  - Guest tools replacement
- Guest boots -> all fine
- QCOW2 overlay for in-place upgrade for safe rollback
PCS guest tools

- Enter into VM from host
- Execute command in the VM
- Guest network settings
- Guest user/password management
- Change guest SID (Windows)
Backups

● QEMU (mainstream) support is limited
  ○ No migration, no persistence, very slow
● Specification is negotiated in 02.2016
● Patches in are in the list around the year
  ○ Working without a specification was a BIG mistake
● Interface for external tools is in specification stage
● Backup compression
Performance tweaks

Welcome to the real world!
Noticeable tweaks

- QEMU: QCOW2 metadata cache increased
- QEMU: not needed flushes removed (Win 2k8)
- KVM: adaptive ‘halt_poll_ns’
- CPU governor: performance
- Non-parallel backup/mirror
- Migration auto-converge, compression
- Active sync mirror
- QEMU: fragmented heap (khugepaged)
Virtuozzo performance

Virtuozzo 7 VMs are 25% faster than Virtuozzo 6 VMs
Virtuozzo performance

Virtuozzo 7 VMs are 53% faster than Virtuozzo 6 VMs
Virtuozzo performance

Virtuozzo 7 VMs are 57% faster than Virtuozzo 6 VMs
Virtuozzo performance

Windows VM
(lower is better)
- ✰ Virtuozzo 6
- ✰ Virtuozzo 7

Linux VM
(lower is better)
- ✰ Virtuozzo 6
- ✰ Virtuozzo 7

Acceptable migration time ~ 3 min
Acceptable migration time ~ 45 sec
Virtuozzo performance

Migration downtime

Windows VM
(lower is better)

✪ Virtuozzo 6
✪ Virtuozzo 7

Linux VM
(lower is better)

✪ Virtuozzo 6
✪ Virtuozzo 7

Acceptable downtime ~ 5-15 sec (depending on load)
Some tweaks for the future

- Hyper-V TSC page
- Statistics collection
- Hyper-V remote TLB flush
Future plans

Something shiny
Status

- We are able to survive with the VERY small team
- The product is really stable
- The performance looks great
- The life is better!
Really important stuff

- Mainstream backups
- Block job improvements
- Hyper-V emulation
- VMs en-mass on one host
Questions?

V www.virtuozzo.com

@VirtuozzoInc

www.linkedin.com/company/virtuozzo