KVM as a Microsoft-compatible hypervisor.

Vadim Rozenfeld <vrozenfe@redhat.com>
KVM Forum, 2012
Agenda

Microsoft Enlightenment
KVM as a conformant hypervisor
Performance Improvements
Microsoft Enlightenment

An optimization to a guest operating system to make it aware of VM environments and tune its behavior for VMs. Enlightenments help to reduce the cost of certain operating system functions such as memory management. Enlightenments are accessed through the hypercall interface. Enlightened I/O can utilize the VMBus directly, bypassing any device emulation layer. An operating system that takes advantage of all possible enlightenments is said to be “fully enlightened.”

Supported Windows Operating Systems

- Windows Vista
- Windows Server 2008
- Windows 7
- Windows Server 2008 R2
- Windows 8
- Windows Server 2012
Conformant Hypervisor

Minimal HV#1 Interfaces

- CPUID leaves 0x40000000 - 0x40000005

- Hypervisor synthetic MSRs
  - HV_X64_MSR_GUEST_OS_ID
  - HV_X64_MSR_HYPERCALL
  - HV_X64_MSR_VP_INDEX
The Hypercall Environment

- Check that hypervisor is present
- Determine
  - Hypervisor version
  - Capabilities
  - Implementation recommendations
- Report the guest OS' identity
- Setup and enable the hypercall page
Hypercall page

- No hypercalls

```
RET
MOV EDX, 0
MOV EAX, 2
```

- With hypercalls

```
int kvm_emulate_hypercall(struct kvm_vcpu *vcpu) {
    unsigned long nr, a0, a1, a2, a3, ret;
    int r = 1;
    if (kvm_hv_hypercall_enabled(vcpu->kvm))
        return kvm_hv_hypercall(vcpu);
```
Partition Reference Time Enlightenment

“hv_reftime”

Guest:

- Windows 7
- Windows 7 SP1
- Windows Server 2008 R2
- Windows Server 2008 R2 SP1
(Ke)QueryPerformanceCounter

- System time sources:
  - HPET
  - PM Timer
  - iTSC
  - Reference Time
Invariant TSC

Host:
- Constant rate TSC
- HV_X64_MSRREFERENCE_TSC MSR
- allows mapping the reference TSC page

Guest:
- RDTSC as a system time source

TSC reference Page

```c
uint64_t TscOffset;
uint64_t TscScale;
uint32_t Res;
uint32_t TscSequence;
```
Reference Time Enlightenment as the fallback mechanism.

Host:

- System without invariant TSC
- HV_X64_MSR_TIME_REF_COUNT MSR
Guest Spin locks

“hv_spinlocks=xxx”

HvNotifyLongSpinWait hypercall

Guest:

- used by a guest OS to notify the hypervisor that the calling virtual processor is attempting to acquire a resource that is potentially held by another virtual processor within the same partition.

Host:

- hypervisor indicates to the guest OS the number of times a spinlock acquisition should be attempted before indicating an excessive spin situation to the hypervisor.
Guest Spin locks (KfAcquireSpinLock)

- Pause-Loop Exiting

```c
spin_lock:
    attempt lock_acquire;
    if fail {
        if(!spin_wait_count--) {
            HvNotifyLongSpinWait
        }
        PAUSE;
        jmp spin_lock;
    }
```
Local APIC Virtualization

- “hv_vapic”
- KVM provides accelerated MSR access to high usage memory mapped APIC registers.
- HV_X64_MSR_EOI Accesses the APIC EOI
- HV_X64_MSR_ICR Accesses the APIC ICR
- HV_X64_MSR_TPR Access the APIC TPR
- APIC Assist Page
Relaxed Timing

hv_relaxed

Caused by:

- Heavy loaded.
- Interrupt delivery delays.
IoMeter

<table>
<thead>
<tr>
<th></th>
<th>2K</th>
<th>4K</th>
<th>8K</th>
<th>16K</th>
<th>32K</th>
<th>64K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enlgh</td>
<td>2.31</td>
<td>1.94</td>
<td>1.50</td>
<td>1.01</td>
<td>0.80</td>
<td>0.68</td>
</tr>
<tr>
<td>kvm</td>
<td>3.34</td>
<td>2.95</td>
<td>2.45</td>
<td>1.94</td>
<td>1.20</td>
<td>0.88</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>2K</th>
<th>4K</th>
<th>8K</th>
<th>16K</th>
<th>32K</th>
<th>64K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enlgh</td>
<td>12.80</td>
<td>21.87</td>
<td>43.69</td>
<td>47.46</td>
<td>66.38</td>
<td>75.21</td>
</tr>
<tr>
<td>kvm</td>
<td>10.51</td>
<td>18.28</td>
<td>29.64</td>
<td>42.50</td>
<td>55.49</td>
<td>72.52</td>
</tr>
</tbody>
</table>
Windows Performance Toolkit

xbootmgr
Viostor (virtio-blk) ISR and DPC performance

- **DPC**

<table>
<thead>
<tr>
<th></th>
<th>Max Actual Duration (ms)</th>
<th>Avg Actual Duration (ms)</th>
<th>Actual Duration (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>kvm + Enlgh</td>
<td>0.14</td>
<td>0.001712</td>
<td>19.0116</td>
</tr>
<tr>
<td>kvm</td>
<td>0.260369</td>
<td>0.011171</td>
<td>130.587</td>
</tr>
</tbody>
</table>

- **ISR**

<table>
<thead>
<tr>
<th></th>
<th>Max Actual Duration (ms)</th>
<th>Avg Actual Duration (ms)</th>
<th>Actual Duration (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>kvm + Enlgh</td>
<td>0.1841</td>
<td>0.002236</td>
<td>25.1708</td>
</tr>
<tr>
<td>kvm</td>
<td>0.339429</td>
<td>0.015927</td>
<td>179.8165</td>
</tr>
</tbody>
</table>
Microsoft Hyper-V Virtual Machine Bus

“Third party virtualization solutions must not claim support for the Microsoft HyperV Virtual Machine Bus (VMBus) device in the virtual BIOS ACPI namespace.”

DefinitionBlock ("DSDT.aml", "DSDT", 1, "MSFTVM", "MSFTVM02", 0x00000002)
Resources:

Hypervisor Top-Level Functional Specification 2.0A: Windows Server 2008 R2

Requirements for Implementing the Microsoft Hypervisor Interface
http://msdn.microsoft.com/library/windows/hardware/hh975392