Alibaba Cloud | Worldwide Cloud Services Partner

The traps of using Hyper V features in KVM environment

Liang Li Aug 2021

Content

Alibaba Cloud | Worldwide Cloud Services Partner

Background

Alibaba Cloud | Worldwide Cloud Services Partner

- Application scenarios of Window guest
  - Cloud Desktop
  - Cloud Game
- Hyper V
  - Windows guest support is good
- KVM
  - Try to support Windows guest better by simulating Hyper V functions
- Hyper V related features
  - hv-relaxed, hv-time, hv-stimer, hv-stimer-direct, hv-vapic, hv-synic, hv-tlbflush, hv-ipi, hv-spinlocks …
  - Common usage: turn on all features

Workload characteristics of cloud gaming

- 3D rendering
  - High CPU & GPU usage
  - IPI intensive
    - 35000+ IPIs per second
    - 1200+ extra IPIs per second for when Microsoft Remote Desktop for Mac is used
      - IPI send to some of the VCPUs
      - Which can be accelerate by hv-ipi
- Performance drops significantly when running on a VM
  - Compared with running on a bare metal server
  - With Hyper V features enabled
  - Average FPS drops by 1
  - The proportion of [>55] FPS decreased by 10%

Performance comparison with different config

Alibaba Cloud | Worldwide Cloud Services Partner

None of Hyper V feature is set

```
<features>
  <acpi/>
  <apic/>
  <pae/>
  <pmu/>
  <kvm><hidden state='on'/></kvm>
</features>
```

```
Analyze events for all VMs, all VCPUs:

             VM-EXIT    Samples  Samples%     Time%    Min Time    Max Time          Avg time

      EPT_MISCONFIG     893519    72.16%    88.46%      4.38us    662.31us      6.50us ( +-   0.05% )
          APIC_WRITE     215110    17.37%     4.83%      0.67us     49.55us      1.47us ( +-   0.17% )
  EXTERNAL_INTERRUPT     100693     8.13%     6.20%      0.82us     61.70us      4.04us ( +-   0.14% )
           DR_ACCESS      22842     1.84%     0.29%      0.57us     42.28us      0.83us ( +-   0.37% )
        EOI_INDUCED       5060     0.41%     0.12%      0.98us      3.44us      1.59us ( +-   0.23% )
     IO_INSTRUCTION        734     0.06%     0.08%      3.31us    637.88us      7.53us ( +-  11.95% )
      EPT_VIOLATION        224     0.02%     0.02%      1.15us     15.87us      6.05us ( +-   3.80% )
              CPUID         45     0.00%     0.00%      0.59us      2.68us      1.70us ( +-   6.06% )
      EXCEPTION_NMI         10     0.00%     0.00%      2.61us     20.12us     11.91us ( +-  20.50% )

Total Samples:1238237, Total events handled time:6563955.39us.
```

All Hyper V features are set

```
<features>
  <acpi/>
  <apic/>
  <pae/>
  <pmu/>
  <kvm><hidden state='on'/></kvm>
  <hyperv>
    <relaxed state='on'/>
    <vapic state='on'/>
    <spinlocks state='on' retries='4096'/>
    <vpindex state='on'/>
    <runtime state='on'/>
    <synic state='on'/>
    <stimer state='on'>
      <direct state='on'/>
    </stimer>
    <reset state='on'/>
    <vendor_id state='on' value='KVM Hv'/>
    <frequencies state='on'/>
    <reenlightenment state='on'/>
    <tlbflush state='on'/>
    <ipi state='on'/>
  </hyperv>
</features>
```

```
Analyze events for all VMs, all VCPUs:

             VM-EXIT    Samples  Samples%     Time%    Min Time    Max Time          Avg time

  EXTERNAL_INTERRUPT     429372    65.13%    69.99%      0.76us     59.48us      2.12us ( +-   0.10% )
          MSR_WRITE     176410    26.76%    22.59%      0.69us     49.55us      1.67us ( +-   0.19% )
           DR_ACCESS      21562     3.27%     1.47%      0.51us     46.78us      0.89us ( +-   0.34% )
 TPR_BELOW_THRESHOLD      12999     1.97%     1.00%      0.62us      7.23us      1.01us ( +-   0.21% )
    INTERRUPT_WINDOW       8971     1.36%     0.68%      0.71us      6.30us      0.98us ( +-   0.31% )
              VMCALL       6612     1.00%     2.95%      1.18us     45.14us      5.81us ( +-   0.53% )
      EPT_VIOLATION       1574     0.24%     0.51%      0.96us     70.32us      4.25us ( +-   2.07% )
     IO_INSTRUCTION       1567     0.24%     0.79%      3.20us     53.15us      6.58us ( +-   2.76% )
              CPUID        158     0.02%     0.02%      0.54us      3.31us      1.40us ( +-   3.72% )
     PREEMPTION_TIMER         46     0.01%     0.00%      0.76us      1.34us      0.91us ( +-   2.24% )
      EXCEPTION_NMI          5     0.00%     0.00%      2.05us     16.75us      8.08us ( +-  43.34% )

Total Samples:659276, Total events handled time:1303363.15us.
```

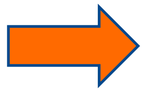
Hyper V features can help to reduce the virtualization overhead a lot

Performance comparison with different config

Alibaba Cloud | Worldwide Cloud Services Partner

Disable hypervisor CPUID

```
<cpu mode='host-passthrough'>
    <topology sockets='1' cores='10' threads='2'/>
    <feature policy='disable' name='hypervisor'/>
</cpu>
```

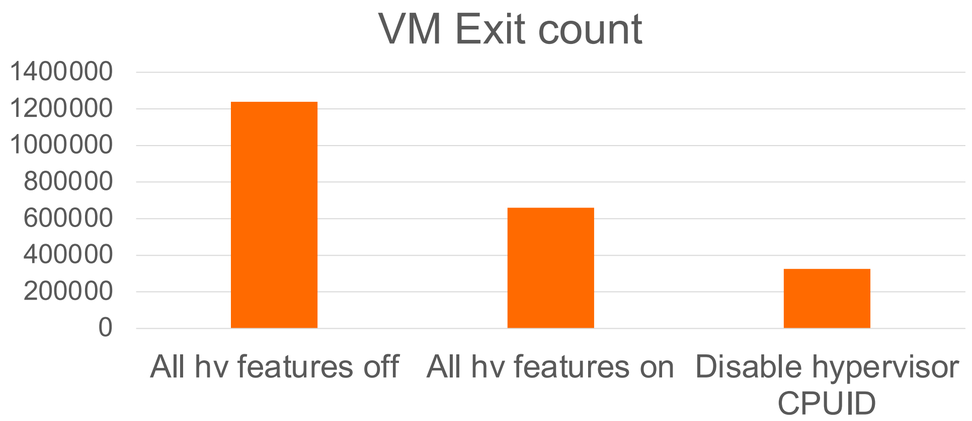


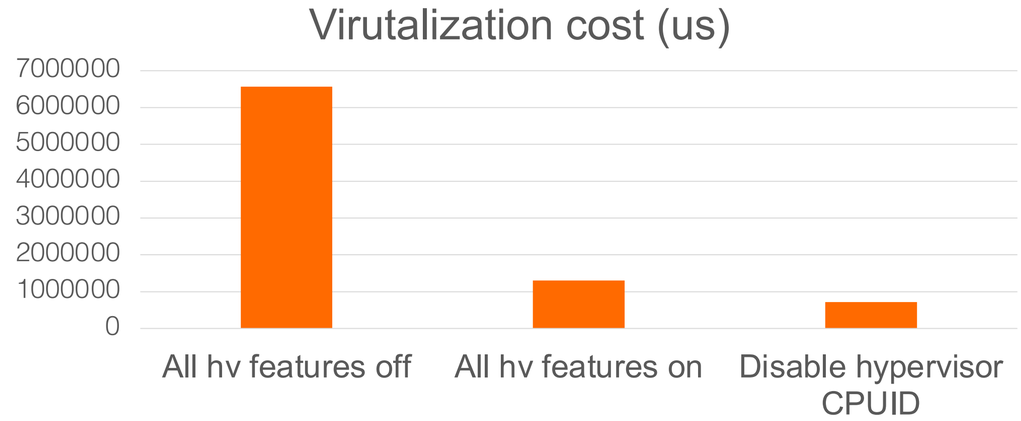
```
Analyze events for all VMs, all VCPUs:

             VM-EXIT    Samples  Samples%     Time%   Min Time   Max Time          Avg time

          APIC_WRITE     185817     56.97%     37.07%      0.57us    78.45us     1.43us ( +-   0.18% )
  EXTERNAL_INTERRUPT     114446     35.09%     58.73%      0.80us    53.05us     3.69us ( +-   0.14% )
           DR_ACCESS      24425      7.49%      2.98%      0.55us    45.56us     0.88us ( +-   0.33% )
      IO_INSTRUCTION        804      0.25%      0.88%      3.30us    59.74us     7.87us ( +-   3.97% )
       EPT_VIOLATION        433      0.13%      0.27%      1.03us    23.21us     4.40us ( +-   4.51% )
               CPUID        213      0.07%      0.07%      1.00us     8.92us     2.43us ( +-   2.14% )
       EXCEPTION_NMI          6      0.00%      0.01%      2.66us    17.36us     9.71us ( +-  32.24% )

Total Samples:326144, Total events handled time:718992.36us.
```





Disable hypervisor CPUID has the lowest virtualization overhead

How Windows guest choose system timer

- Expose hypervisor CPUID
  - Priority: Stimer > HPET > RTC
  - Stimer is used when all the Hyper V features are turned on
  - HPET or RTC is used when the hyper v features are turned off
- Hide hypervisor CPUID
  - Priority: LAPIC timer > HPET > RTC
  - LAPIC timer is the default system timer
  - Hyper V related features are invalid

Virtualization efficiency of different system timers

- RTC & HPET
  - RTC trapped by PIO access
  - HPET trapped by MMIO access
  - Emulated in user space
- Stimer
  - Trapped by MSR access
  - Emulated in Kernel
- LAPIC timer
  - Trapped by APIC access
  - Emulated in Kernel
- Virtualization overhead
  - LAPIC timer == Stimer < HPET < RTC

Cause Analysis

- Why virtualization overhead is lower when Hyper V features are enabled
  - Stimer has lower virtualization overhead than HPET & RTC
- Why virtualization overhead is the lowest after hiding hypervisor CPUID
  - Stimer has side effects

Cause Analysis

- Some facts about Stimer
  - hv-stimer depends on hv-synic
  - The Auto EOI feature of hv-synic conflicts with APICv
    - APICv can reduce interrupt injection overhead
  - The hardware APICv feature is invalid when Stimer is on
- IPI virtualization for Intel CPU
  - Trapped by ICR access
  - Inject interrupts into the VCPU will cause vm exit if APICv is off
- Stimer will increase the IPI virtualization overhead
  - LAPIC timer does not have this problem
  - Turn off Stimer will increase overall virtualization overhead

Solutions

- Hide hypervisor CPUID for scenarios with intensive IPIs
  - Disable all Hyper V features at the same time
  - Can't enjoy the benefits of hv-tlbflush, hv-ipi, hv-spinlocks and hv-xxx
- Adjust the logic of Windows' selection of system timers
  - Decoupling hypervisor CPUID and LAPIC timer
  - Give priority to using LAPIC timer when Hypervisor CPUID is exposed

Solutions

- Resolve the conflict between hv-stimer and APICv
  - Disable the Auto EOI feature of hy-synic
    - Solved by expose HV_DEPRECATING_AEOI_RECOMMENDED
  - Recommend guests use hardware APICv MSR
    - Can be solved by clear HV_X64_APIC_ACCESS_RECOMMENDED
  - Optimize the cost of EOI induced vm-exit
    - Avoid EOI induced vm-exit for Stimer

Effect of optimization

Alibaba Cloud | Worldwide Cloud Services Partner

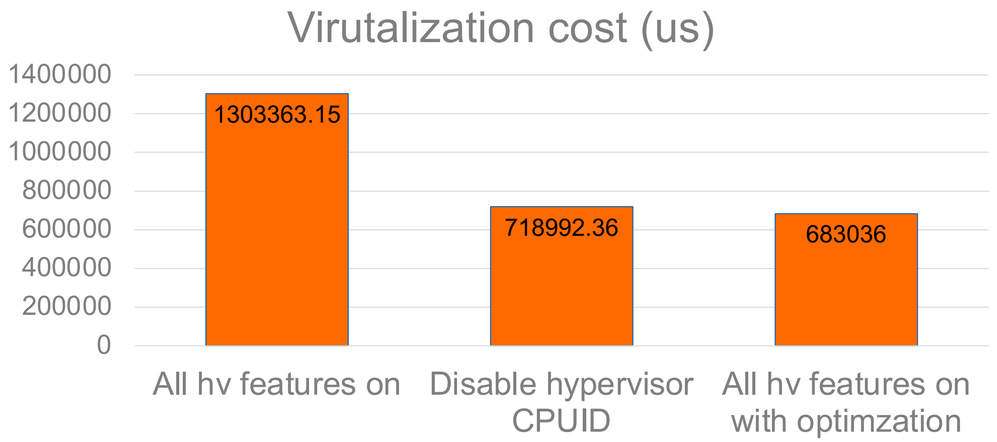
Set all HV features (with hv-stimer optimization)

```
Analyze events for all VMs, all VCPUs:

             VM-EXIT    Samples  Samples%     Time%    Min Time    Max Time         Avg time

          APIC_WRITE     173738     56.60%     35.57%      0.67us     47.68us      1.39us ( +-   0.18% )
  EXTERNAL_INTERRUPT     105217     34.28%     56.78%      0.89us     65.33us      3.70us ( +-   0.14% )
         DR_ACCESS       21740      7.08%      2.61%      0.57us      7.79us      0.81us ( +-   0.22% )
             VMCALL       5851      1.91%      4.52%      1.19us     16.51us      5.30us ( +-   0.50% )
      IO_INSTRUCTION       198      0.06%      0.45%      3.52us    643.97us     15.59us ( +-  30.29% )
          MSR_WRITE       120      0.04%      0.03%      0.90us      3.55us      1.48us ( +-   3.01% )
              CPUID        46      0.02%      0.01%      1.08us      2.95us      1.99us ( +-   2.83% )
       EPT_VIOLATION        20      0.01%      0.03%      2.67us     16.25us      9.78us ( +-   9.25% )
       EXCEPTION_NMI         7      0.00%      0.01%      2.12us     17.97us      9.51us ( +-  28.03% )

Total Samples:306939, Total events handled time:683036.55us.
```



- After optimization, turn on all Hyper V features perform best

Conclusion

Alibaba Cloud | Worldwide Cloud Services Partner

- Hyper V features in KVM have room for improvement
  - Any feature should not cause performance degradation
  - Avoid the need for users to decide which feature to use according to workload
- Pay attention to the pitfalls when using the Hyper V features
  - Turning on all Hyper V features is not necessarily the best way
  - Pay attention to business scenarios with intensive IPIs
  - Before related problems are solved, performance tests are required

Q&A

Alibaba Cloud | Worldwide Cloud Services Partner

foxi.ll@alibaba-inc.com

Worldwide Cloud Services Partner