# Where Does the Time Go?

## Profiling Nested KVM on x86/Intel

Jim Mattson

KVM Forum — 26 October 2017

# Test Environment

Host: Broadwell (>20 threads, >64GB RAM)
Distro: Debian 8 (except for L0: Google)
Kernel: 4.13.4 + cherry-pick 43e30258
KVM + QEMU 2.10.0
VMware Workstation 14*

*Patched to support 4.13.4.

# A Micro Benchmark

The Worst Case for Nested Performance?

# CPUID Loop

15 vCPUs / 32GB RAM (LLVM)
16 vCPUs / 42GB RAM (L1, if parent)
CPUID.0 in tight loop (4 instructions)
Terminated by SIGALRM after 30 seconds
Measure elapsed time with RDTSC
Divide elapsed time by iterations

# CPUID Performance

| Configuration | Cycles* | Relative to L1 |
|---|---|---|
| L1 KVM/QEMU | 768 ± 11 | --- |
| L2 KVM/QEMU | 18830 ± 278 | 25X |
| L1 VMware WS | 590 ± 7 | --- |
| L2 VMware WS | 6446 ± 68 | 11X |

*Average of six runs

# Ancillary L1 VM-Exits

| KVM/QEMU | |
|---|---|
| **Exit Reason** | **Frequency*** |
| VMREAD | 2X |
| VMWRITE | 1X |
| XSETBV | 1X |
| RDMSR | 1X |
| VMRESUME | 1X |

| VMware WS | |
|---|---|
| **Exit Reason** | **Frequency*** |
| VMWRITE | 1X |
| VMRESUME | 1X |
| CPUID | 1X |

*Relative to L2 CPUID VM-Exits

# Ancillary Exit Culprits (KVM/QEMU)

| VMREAD | |
|---|---|
| PML index | 49.8% |
| Interrupt status | 49.7% |
| Pin-based controls | 0.3% |
| Guest SS rights | 0.0% |

| VMWRITE | |
|---|---|
| Preemption timer | 99.0% |
| Pin-based controls | 0.7% |
| Interrupt status | 0.3% |

| RDMSR | |
|---|---|
| IA32_DEBUGCTL | 100% |

# Macro Benchmarks

# Kernel Compile Performance

| Configuration | Time (sec)* | Relative to L1 |
|---|---|---|
| L1 KVM/QEMU | 470 ± 0.3 | --- |
| L2 KVM/QEMU | 488 ± 0.4 | 1.04X |
| L1 VMware WS | 533 ± 0.5 | --- |
| L2 VMware WS | 572 ± 2.3 | 1.07X |

*Average of six runs

# iperf Network Bandwidth

| Configuration | Bandwidth (Mbits/s)* | Relative to L1 |
|---|---|---|
| L1 KVM/QEMU | 17168 ± 245 | --- |
| L2 KVM/QEMU | 19598 ± 118 | 1.1 X |
| L1 VMware WS | 1783 ± 72 | --- |
| L2 VMware WS | 494 ± 40 | 0.3X |

*Average of six runs

# schbench Scheduling Latency

| Configuration | P99 (usec)* | Relative to L1 |
|---|---|---|
| L1 KVM/QEMU | 17781 ± 824 | --- |
| L2 KVM/QEMU | 28907 ± 315 | 1.6X |
| L1 VMware WS | 28107 ± 449 | --- |
| L2 VMware WS | 30304 ± 40 | 1.1X |

*Average of six runs

# Profiling L0 with Perf

# L2 CPUID

| Frequency | Symbol |
|---|---|
| 26% | nested_vmx_disable_intercept_for_msr |
| 11% | nested_vmx_run |
| 11% | vmx_vcpu_run |
| 4% | kvm_arch_vcpu_ioctl_run |
| 3% | copy_shadow_to_vmcs12 |
| 3% | vmx_save_host_state |
| 3% | vmx_set_cr3 |

# L2 Kernel Compile

| Frequency | Symbol |
|---|---|
| 16% | queued_spin_lock_slowpath |
| 10% | nested_vmx_disable_intercept_for_msr |
| 8% | vmx_vcpu_run |
| 7% | nested_vmx_run |
| 4% | kvm_arch_vcpu_ioctl_run |
| 3% | _raw_spin_lock |
| 2% | copy_shadow_to_vmcs12 |

# L2 iperf

| Frequency | Symbol |
|---|---|
| 13% | nested_vmx_disable_intercept_for_msr |
| 13% | vmx_vcpu_run |
| 6% | nested_vmx_run |
| 5% | kvm_arch_vcpu_ioctl_run |
| 4% | read_tsc |
| 3% | _raw_spin_lock |
| 3% | __srcu_read_lock |

# L2 schbench

| Frequency | Symbol |
|---|---|
| 15% | queued_spin_lock_slowpath |
| 14% | __lock_acquire |
| 8% | nested_vmx_disable_intercept_for_msr |
| 7% | lock_is_held_type |
| 4% | vmx_vcpu_run |
| 4% | nested_vmx_run |
| 3% | do_raw_spin_lock |

# Hot Spots: nested_vmx_run

| Frequency | Code |
|---|---|
| 32% | for (msr = 0x800; msr <= 0x8ff; msr++)<br>  nested_vmx_disable_intercept_for_msr(<br>  msr_bitmap_l1, msr_bitmap_l0,<br>  msr, MSR_TYPE_R); |
| 15% | memset(msr_bitmap_l0, 0xff, PAGE_SIZE); |

# Hot Spots: vmx_vcpu_run

| Frequency | Code |
|---|---|
| 25% | debugctlmsr = get_debugctlmsr(); |
| 23% | vmresume<br>mov %rcx,0x8(%rsp)<br>pop %rcx |
| 6% | tscl = rdtsc(); |

# Lessons

Continuously monitor performance
Avoid reconstructing the VMCS02 MSR permission bitmap
Use VMCS shadowing effectively
Cache IA32_DEBUGCTL in memory

# Backup Slides

# L1 QEMU Command Line

```
qemu --enable-kvm -cpu host -smp 16 -m 43008 \
 -nographic -serial mon:stdio \
 -drive file=Debian8.img,format=raw,id=hd0,if=none,readonly=off \
 -drive file=Debian8-2.1.img,format=raw,id=hd1,if=none,readonly=off \
 -drive file=Debian8-2.2.img,format=raw,id=hd2,if=none,readonly=off \
 -device driver=virtio-scsi-pci,id=scsi \
 -device drive=hd0,driver=scsi-hd \
 -device drive=hd1,driver=scsi-hd \
 -device drive=hd2,driver=scsi-hd \
 -netdev tap,id=eth,ifname=tap16,script=no,downscript=no \
 -device virtio-net-pci,netdev=eth,mac=52:54:00:00:00:01
```

Blue fields modified when L1 is the LLVM (-smp and -m to match L2 test; unique ifname and mac)

# L2 QEMU Command Line(s)

```
qemu --enable-kvm -cpu host -smp 15 -m 32768 \
 -nographic -serial mon:stdio \
 -drive file=/dev/sdb,format=raw,id=hd0,if=none,readonly=off \
 -device driver=virtio-scsi-pci,id=scsi \
 -device drive=hd0,driver=scsi-hd \
 -netdev tap,id=eth,ifname=tap1,script=no,downscript=no \
 -device virtio-net-pci,netdev=eth,mac=52:54:00:00:00:02
```

iperf only: -smp 7 -m 16384 & a second L2 with file=/dev/sdc and mac=52:54:00:00:00:03

# CPUID Loop Code

```
        movl  $0, %esi
.L9:
        movl  %esi, %eax
        cpuid
        addq $1, count(%rip)
        jmp    .L9
```

# CPUID Raw Data

| Configuration | Timings (cycles) | | | | | |
|---|---|---|---|---|---|---|
| L1 KVM/QEMU | 786 | 734 | 792 | 768 | 789 | 736 |
| L2 KVM/QEMU | 19294 | 18409 | 19428 | 18181 | 19592 | 18075 |
| L1 VMware WS | 604 | 570 | 615 | 586 | 587 | 575 |
| L2 VMware WS | 6672 | 6251 | 6607 | 6339 | 6466 | 6341 |

# Kernel Compile

15 vCPUs / 32GB RAM (LLVM)
16 vCPUs / 42GB RAM (L1, if parent)
mount none -t tmpfs -o size=80% /kernel
4.13.4+ source copied to /kernel
make oldconfig; make clean
/usr/bin/time make -j 15 vmlinux modules

# Kernel Compile Raw Data

| Configuration | Timings (seconds) | | | | | |
|---|---|---|---|---|---|---|
| L1 KVM/QEMU | 469.42 | 468.94 | 468.67 | 470.23 | 469.8 | 470.5 |
| L2 KVM/QEMU | 487.4 | 488.08 | 489.12 | 489.98 | 488.46 | 487.55 |
| L1 VMware WS | 534.79 | 532.25 | 534.46 | 533.13 | 532.31 | 532.21 |
| L2 VMware WS | 572.61 | 582.84 | 568.45 | 569.99 | 567.42 | 568.94 |

# Kernel Compile Top VM-exits

| KVM/QEMU VM-exits per second | | |
|---|---|---|
| L1 | VMREAD | 50451 |
| L1 | VMWRITE | 25641 |
| L2 | INTR | 17311 |
| L1 | RDMSR | 16822 |
| L1 | XSETBV | 16809 |
| L1 | VMRESUME | 16809 |
| L2 | WRMSR | 12786 |
| L1 | WRMSR | 6939 |
| L2 | PREEMPT | 3374 |
| L2 | RDTSC | 3358 |
| L2 | HLT | 592 |
| L1 | HLT | 552 |
| L1 | TIMER | 274 |
| L1 | INTR | 149 |
| L2 | RDMSR | 12 |
| L2 | EXC/NMI | 11 |

| VMware WS VM-exits per second | | |
|---|---|---|
| L1 | WRMSR | 31919 |
| L1 | VMWRITE | 29505 |
| L1 | RDMSR | 21554 |
| L1 | MOV-DR | 21553 |
| L1 | MOV-CR | 21551 |
| L2 | INTR | 18479 |
| L1 | VMRESUME | 16361 |
| L2 | WRMSR | 12374 |
| L1 | VMCLEAR | 10778 |
| L1 | INVEPT | 10771 |
| L1 | VMXOFF | 10771 |
| L1 | VMXON | 10771 |
| L1 | VMPTRLD | 10771 |
| L1 | INVVPID | 10771 |
| L1 | VMLAUNCH | 10746 |
| L2 | RDTSC | 5075 |

(Sampled over 30 seconds)

# iperf

7 vCPUs / 16GB RAM (LLVM)
16 vCPUs / 42GB RAM (L1, if parent)
L2 VMs on private (host-only) network
Paravirtual NICs
    KVM/QEMU: virtio-net
    VMware WS: vmxnet3
iperf -c <sibling> -P6 -t 30 -f m

# iperf Raw Data

| Configuration | Bandwidth (Mbits/sec) | | | | | |
|---|---|---|---|---|---|---|
| L1 KVM/QEMU | 16564 | 17305 | 17252 | 18142 | 17251 | 16496 |
| L2 KVM/QEMU | 19433 | 19378 | 20162 | 19527 | 19631 | 19455 |
| L1 VMware WS | 1618 | 1908 | 1556 | 1724 | 1898 | 1995 |
| L2 VMware WS | 416 | 467 | 383 | 485 | 642 | 570 |

# iperf Top VM-exits

| KVM/QEMU VM-exits per second | | |
|---|---|---|
| L1 | VMREAD | 422462 |
| L1 | VMWRITE | 166269 |
| L1 | RDMSR | 70169 |
| L1 | XSETBV | 70157 |
| L1 | VMRESUME | 70153 |
| L2 | WRMSR | 54385 |
| L1 | WRMSR | 29951 |
| L2 | HLT | 26354 |
| L1 | HLT | 4683 |
| L1 | INTR | 2132 |
| L2 | INTR | 1829 |
| L1 | VMPTRLD | 1788 |
| L2 | EPTVIOL | 1001 |
| L1 | PREEMPT | 754 |
| L2 | PREEMPT | 450 |
| L1 | PAUSE | 22 |

| VMware WS VM-exits per second | | |
|---|---|---|
| L1 | WRMSR | 253021 |
| L1 | MOV-DR | 138996 |
| L1 | MOV-CR | 138957 |
| L1 | RDMSR | 138887 |
| L1 | VMCLEAR | 69808 |
| L1 | INVVPID | 69421 |
| L1 | INVEPT | 69421 |
| L1 | VMPTRLD | 69421 |
| L1 | VMXON | 69421 |
| L1 | VMXOFF | 69421 |
| L1 | VMWRITE | 48960 |
| L1 | HLT | 41368 |
| L1 | VMRESUME | 23971 |
| L1 | VMLAUNCH | 23457 |
| L2 | EPTVIOL | 23048 |
| L2 | WRMSR | 17787 |

(Sampled over 30 seconds)

# schbench

15 vCPUs / 32GB RAM (LLVM)
16 vCPUs / 42GB RAM (L1, if parent)
git://git.kernel.org/pub/scm/linux/kernel/git/mason/schbench.git
(commit 3d16b2322fbc)
Default options: -m 2 -t 16 -r 30 -s 10000 -c 10000

# schbench Raw Data

| Configuration | P99 Latency (usec) | | | | | |
|---|---|---|---|---|---|---|
| L1 KVM/QEMU | 15568 | 19360 | 20512 | 15536 | 18336 | 17376 |
| L2 KVM/QEMU | 28832 | 29856 | 27936 | 29600 | 29088 | 28128 |
| L1 VMware WS | 27104 | 27040 | 28448 | 28000 | 28000 | 30048 |
| L2 VMware WS | 30368 | 30368 | 30432 | 30240 | 30240 | 30176 |

# schbench Top VM-exits

| KVM/QEMU VM-exits per second | | |
|---|---|---|
| L1 | VMREAD | 27532 |
| L1 | VMWRITE | 17008 |
| L2 | INTR | 14405 |
| L1 | RDMSR | 10013 |
| L1 | XSETBV | 9998 |
| L1 | VMRESUME | 9998 |
| L2 | WRMSR | 5022 |
| L1 | WRMSR | 3964 |
| L2 | PREEMPT | 2927 |
| L1 | HLT | 239 |
| L2 | HLT | 233 |
| L1 | INTR | 38 |
| L2 | RDMSR | 15 |
| L2 | EXC/NMI | 9 |
| L1 | PREEMPT | 5 |
| L1 | PAUSE | 0 |

| VMware WS VM-exits per second | | |
|---|---|---|
| L1 | WRMSR | 14641 |
| L1 | VMWRITE | 10257 |
| L2 | INTR | 8643 |
| L1 | RDMSR | 8547 |
| L1 | MOV-CR | 8544 |
| L1 | MOV-DR | 8537 |
| L1 | VMRESUME | 4681 |
| L2 | WRMSR | 4491 |
| L1 | VMCLEAR | 4267 |
| L1 | INVVPID | 4267 |
| L1 | VMXON | 4267 |
| L1 | VMPTRLD | 4267 |
| L1 | VMXOFF | 4267 |
| L1 | INVVPID | 4267 |
| L1 | VMLAUNCH | 4256 |
| L1 | HLT | 1304 |

(Sampled over 30 seconds)

# Hot Spots: kvm_arch_vcpu_ioctl_run

| Frequency | Code |
|---|---|
| 16% | if (kvm_cpu_has_pending_timer(vcpu)) |
| 11% | xsetbv(XCR_XFEATURE_ENABLED_MASK,<br>vcpu->arch.xcr0); |
| 9% | vcpu->srcu_idx =<br>srcu_read_lock(&vcpu->kvm->srcu); |
| 8% | vcpu->arch.last_guest_tsc =<br>kvm_read_l1_tsc(vcpu, rdtsc()); |
| 7% | rcu_note_context_switch(false); |