

Apic Virtualization: Fastpaths for most common exit reasons

1. 目前下两个 feature

- Fast path for EXIT_REASON_PREEMPTION_TIMER
- Fast path for EXIT_REASON_MSR_WRITE { APIC_ICR IPI related and TSC_DEADLINE }

另一个相关 feature:

Exitless Timer: 目前看只对于企业云场景, 须在 qemu 侧使能 cpu-pm feature, 使能后, c82、c84 默认开启;

2.原理

- VM 虚拟机系统更新 TSC_DEADLINE 定时器时, 会写 TSC_DEADLINE 相关的 MSR 寄存器, 这个操作会触发 guest exit, 由 host 模拟定时器的设定以及触发, 正常的模拟路径需要在 host 进行诸多其他的处理后会执行到相关的定时器模拟函数, 这期间大概由几百个 cycles 的时长, 对云场景来说, 这种退出能够占总 exit 的 40%或者更多, 所以将这个功能处理提前至虚拟机退出后更加紧邻的位置, 在云场景下, 这种提前能节省数百 cycle/per exit, 对性能提升有明显的收益;
- 对于公有云场景, 我们会用 VMX Preemption Timer 代替 host 端模拟的定时器, 节省 host 端软件模拟定时器的开销, 当 guest 是由于 VMX Preemption Timer 到期而导致的退出时, 这个 fastpath 的操作也能节省数百个 cycle 的时间; 这个功能以及 1 仅在公有云场景生效;
- Fastpath for APIC_ICR IPI, 这个优化的原理同前, VM 虚拟机发送核间消息更新 APIC_ICR 也是云场景占比比较大的 guestexit 场景, 通过将处理提前, 可以在云场景节省数百 cycle 的时间; 这个 feature 对公有云场景和企业云场景均有效;

3.Test Result:

功能性测试

FastPath		Fastpath Enabled (cycles)	Fastpath Disabled (cycles)	提升
PREEMPTION_TIME MSR_WRITE TSC_DEADLINE	tscdeadline	----	----	----
APIC_ICR IPI related	ipi	----	----	----

上述结果是专门的针对性测试, 所以性能提升相对明显, 但是到实际场景中, 收益会降低; benchmark 性能数据:

Guest Benchmark	Fastpath Enabled	Fastpath Disabled	提升
single-core	-----	----	----
multi-core	-----	-----	-----

稳定性以及迁移兼容性测试:

migration	结果
Fastpath to nonfastpath	正常迁移, 结果正常(fastpath 路径 fallback 到慢速路径)

non-Fastpath to Fastpath	正常迁移，结果正常（慢速路径切到快速路径）
--------------------------	-----------------------

4.遇到的问题

1. 对于使能 DPDK 的云场景，Fastpath for TSC_DEADLINE 判断逻辑有问题，会导致 VMX Preemption Timer 被 disable，相对应的 fastpath 也被 disable，导致了负向的收益，社区的解决方案尚在沟通中，目前测试结果以金山的 patch 进行测试；
 - [0001-KVM-LAPIC-Per-vCPU-control-over-kvm_can_post_timer_i.patch](#)
 - <https://lore.kernel.org/kvm/20211124125409.6eec3938@gmail.com/>
 - <https://lore.kernel.org/kvm/1637733585-47113-1-git-send-email-wanpengli@tencent.com/>
2. Fastpath for ipi 相关的 patch 性能低于社区数据，经过分析，原因有三：
 - c82 的内核缺少“kvm: Replace vcpu->swait with rcuwait”相关的优化 patch，当然 c82 比起社区还缺少很多小的优化 patch
 - 以及改进，但是这个 patch 影响较大，目前已经将此 patch 合入 c82；
TAA: Vulnerable: Clear CPU buffers attempted, no microcode intel CPU 漏洞抑制策略不同，会导致 IPI 相关的优化性能有较大的差异，社区的数字应该只是 优场景的表现；目前看 skylakeCPU 会包含这个漏洞，而不同的服务器厂商对此支持的方式不同，所以会有不同的性能表现；icelake 已经修复此漏洞；
 - Fastpath 相关的 patch 合入社区的时间比较近，在此基础上测试依赖于所在内核版本的表现，c82 相对于社区版本有比较大的 gap；

5.目前的分支

<http://newgit.op.ksyun.com/kvm/kernel-c82/tree/ApicV>

6.AMD 支持

这个 feature 似乎 AMD 也支持，需要找时间调研测试一下；待定