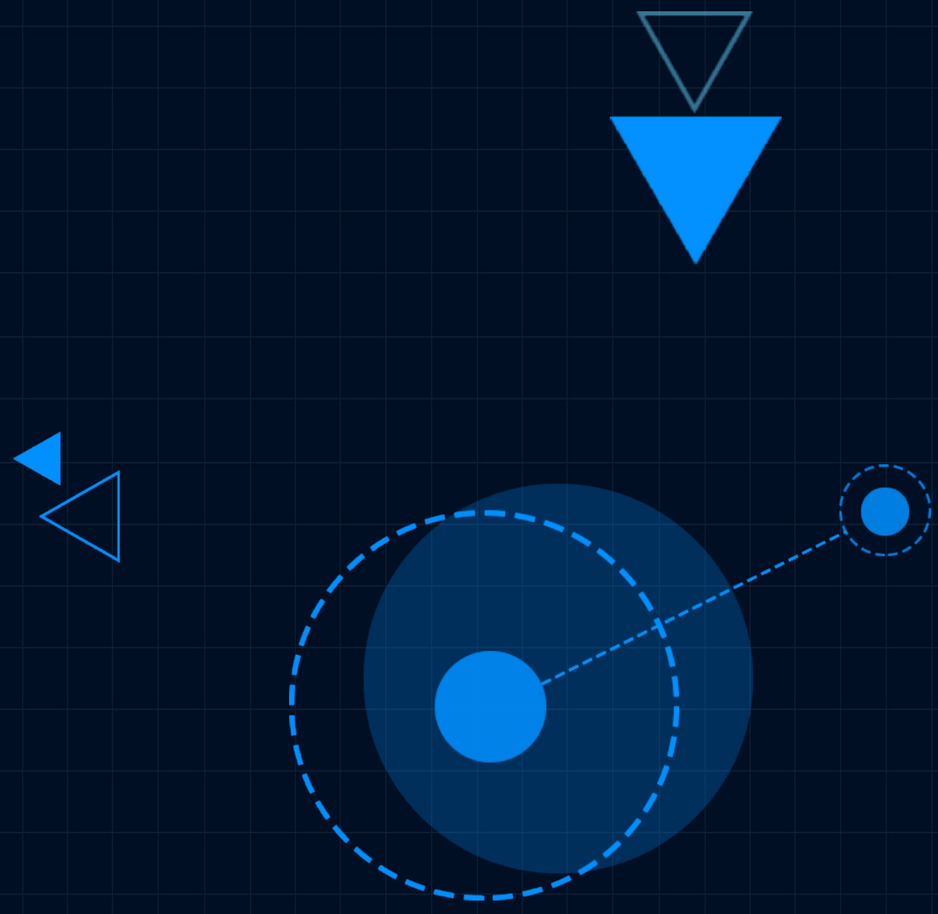
Kata BM

高密度服务器混部性能隔离方案

邓良

字节跳动STE团队工程师



高密度服务器混部面临的问题

1、传统runc容器

runc容器共用Linux内核，但随着服务器核数的增多(Genoa 384、ICELAKE 224)，Linux内核面临性能扩展性问题。Linux内核内存管理、调度、文件系统、网络协议栈、EBPF等子系统由于锁竞争、同步等因素影响，造成system time时间过长，导致业务性能下降或者抖动

runc容器共用Linux内核，内核层触发的异常状态（比如，死锁、I/O hung）会影响其他容器中的业务。

runc容器基于cgroup，难以实现完全的资源隔离，仍然可能存在资源竞争

2、传统kata容器

Kata容器基于传统虚拟化将服务器做kata VM切分，每个kata容器采用独立的内核，能够解决性能扩展性的问题。但又会引入虚拟化开销和host层的性能扰动。

高密度服务器混部面临的问题

传统Kata VM的虚拟化开销和扰动（Intel/AMD）

1. 计算虚拟化

Local APIC Timer，IPI触发的VM exit开销

HLT，MWAIT等指令触发的VM exit开销

CPUID和读写MSR指令触发的VM exit开销

Posted interrupt在硬件层仍然存在额外开销

Host层内核线程、中断的性能扰动

2. 内存虚拟化

EPT walk的开销

EPT page fault的开销

3. 设备虚拟化

Virtio设备中断注入/kick触发的VM exit开销

直通设备IOMMU page table walk的开销

高密度服务器混部面临的问题

现有硬件直通方案：Directvisor(VEE’20)

方案

基于Intel posted interrupt

Local APIC timer 中断和MSR直通

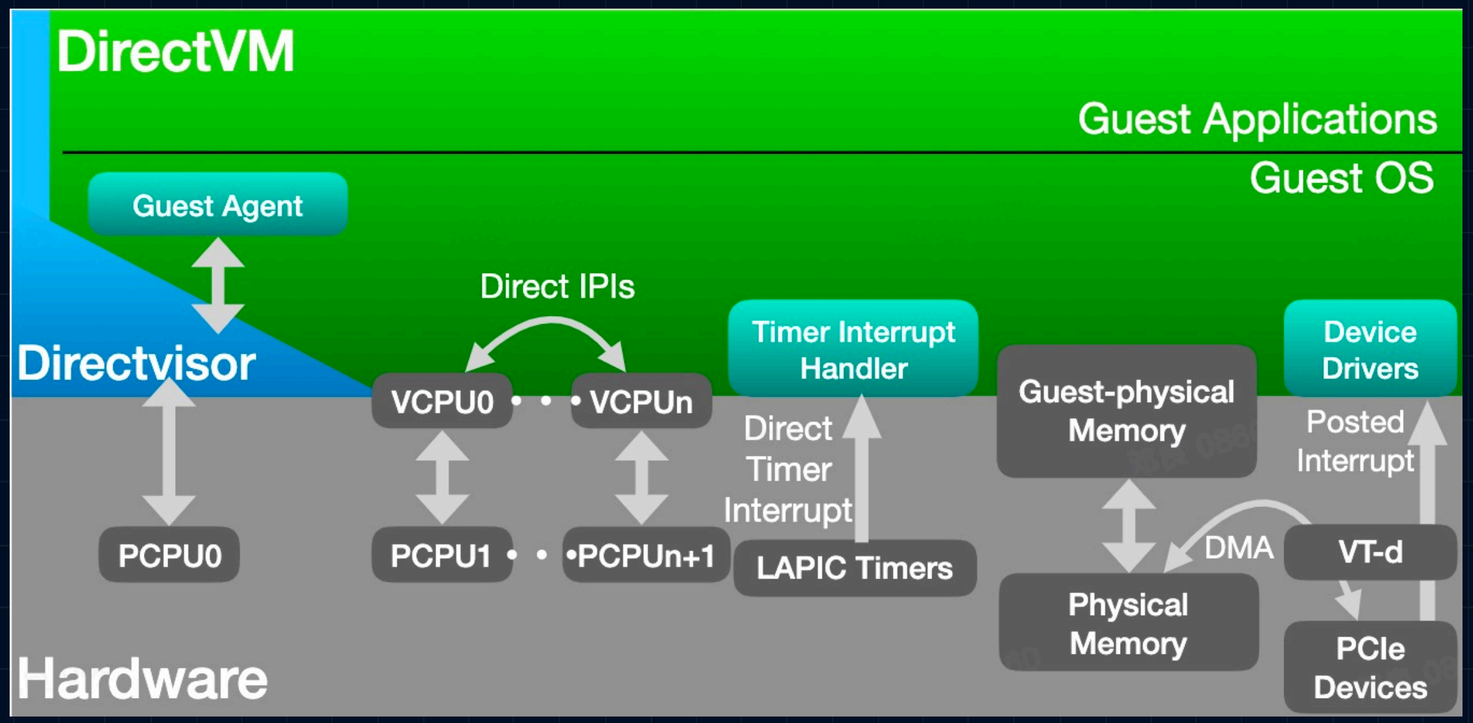
IPI中断和MSR直通

问题

Posted interrupt仍在有开销

在AMD上无法实现Local APIC timer中断的直通

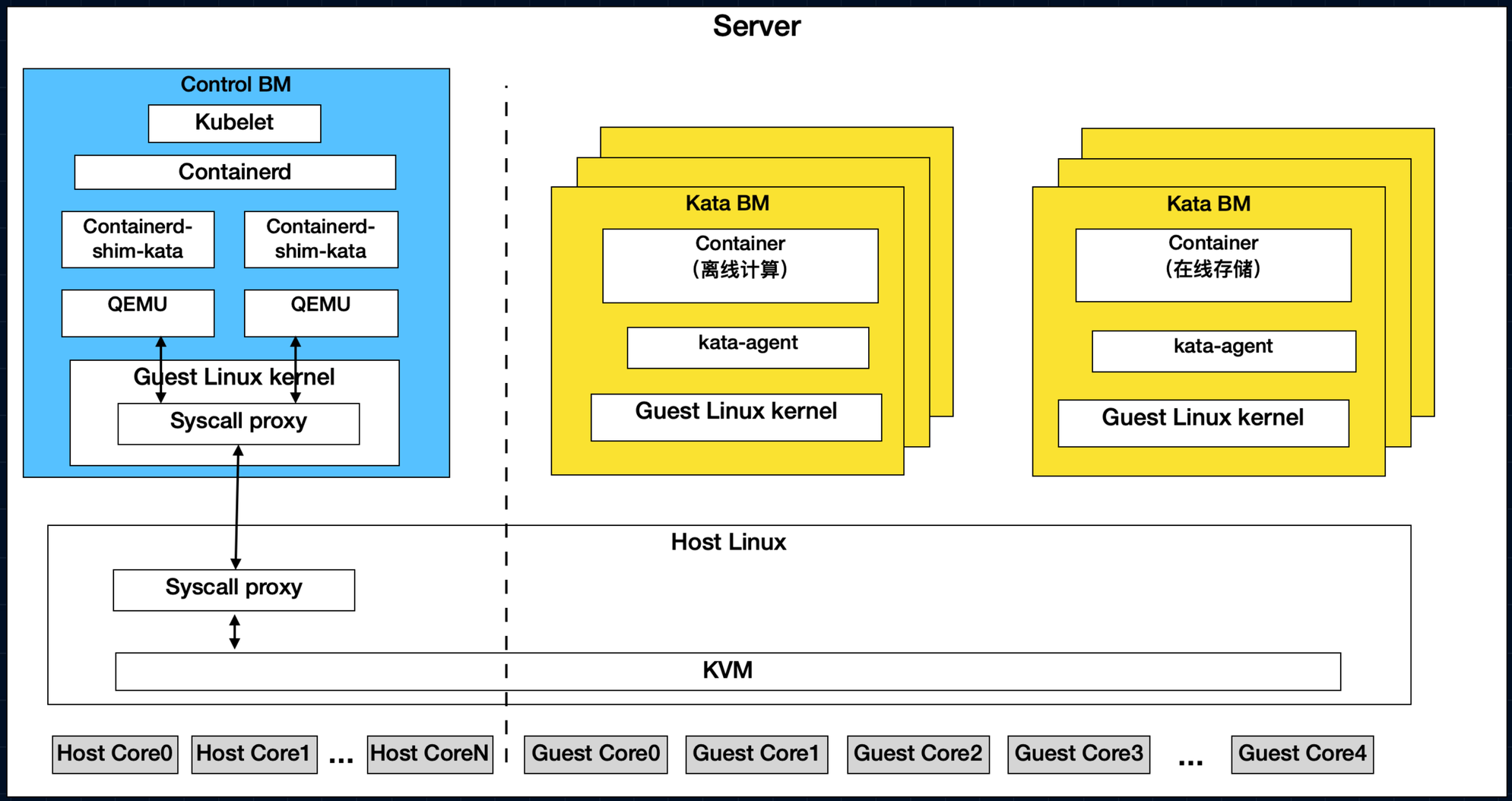
仅考虑了虚拟化的开销问题，没有考虑Host层的性能干扰问题



DirectVisor [1]

[1]https://dl.acm.org/doi/abs/10.1145/3381052.3381317

Kata BM总体架构



运行时硬件直通的Kata VM——>无虚拟化开销

Kata BM直接访问硬件资源，运行时无VM exit，无虚拟化开销

系统调用代理——>避免host层干扰

管控面和VMM（QEMU）运行在独立的Control BM中，不会干扰到Kata BM的运行

QEMU通过系统调用代理，跨OS调用Host linux的kvm模块，拉起Kata BM

Kata BM硬件直通

1. 计算虚拟化

将Local APIC timer和IPI相关的MSR直通Kata BM所在的Nonroot模式，使得Kata BM可以直接配置Local APIC timer和IPI，不产生VM exit。

禁用apicv/avic，直接将external interrupt配置为直通到Kata BM所在的Nonroot模式，使得Kata BM可以直接收到自己的Local APIC timer中断、IPI、设备中断，不产生VM exit。

2. 内存虚拟化

Kata BM的GPA直接等于HPA，不使用EPT

3. 设备虚拟化

Virtio设备的中断注入直接将IPI直通到Kata BM，不产生VM exit；kick操作直接在Kata BM中向host发送IPI，不产生VM exit

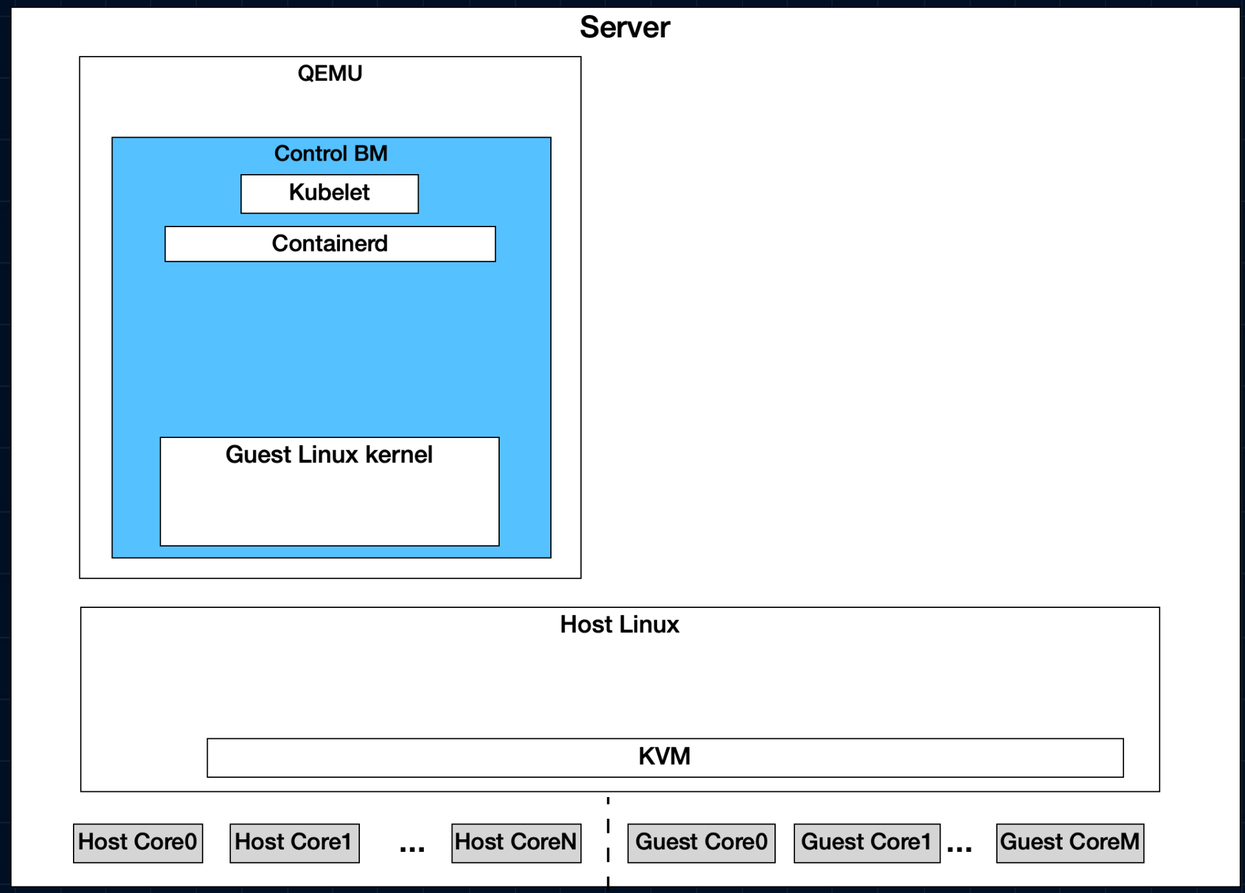
对于直通设备， Kata BM的GPA直接等于HPA ，不使用IOMMU页表

Kata BM系统调用代理

1. 启动管控面

Host Linux通过QEMU+KVM启动Control BM

Control BM是一个特殊的硬件直通的VM，里面运行管控面和VMM

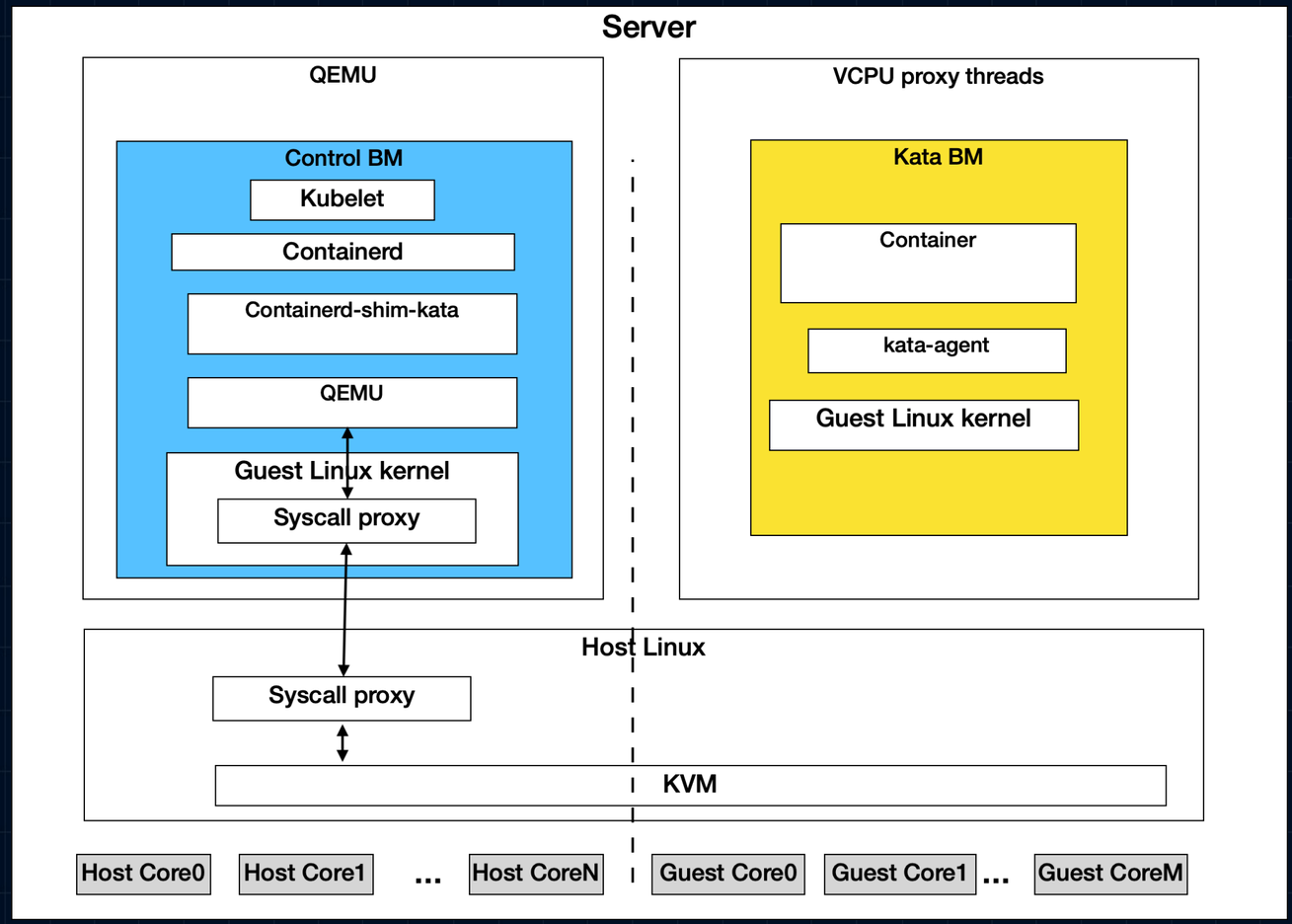


Kata BM系统调用代理

2. 启动Kata BM

运行在Control BM中的管控面通过Containerd-shim-kata启动QEMU

QEMU启动Kata BM时，并不使用本地内核的KVM，而是将对KVM的系统调用请求通过syscall proxy发送到Host linux kernel的KVM进行处理，在Host linux上启动Kata BM



Kata BM性能测试结果

测试环境

1、Kata BM容器

服务器总共128个核，管控面（Control BM）占16个核

启动7个Kata BM容器，每个占16个核，运行redis server/kernel build

2、Runc容器

服务器总共128个核，管控面占16个核

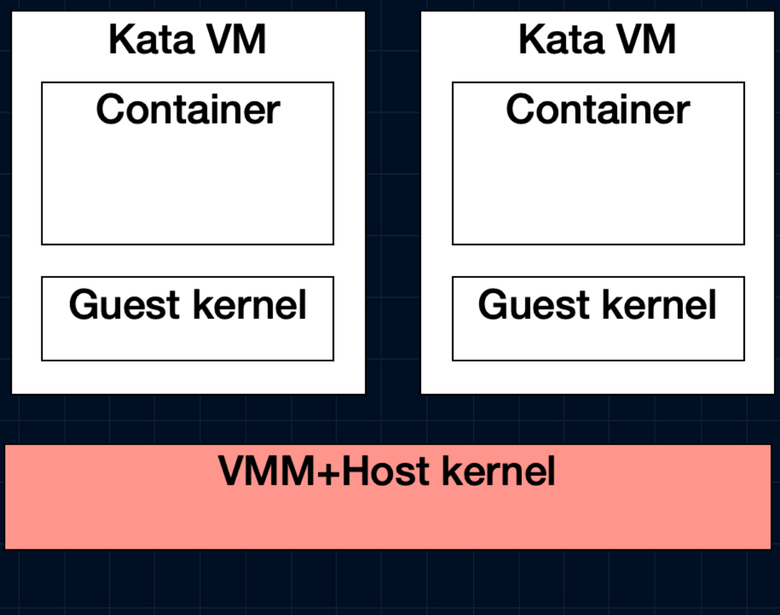
启动7个runc容器，每个占16个核，运行redis server/kernel build

测试结果

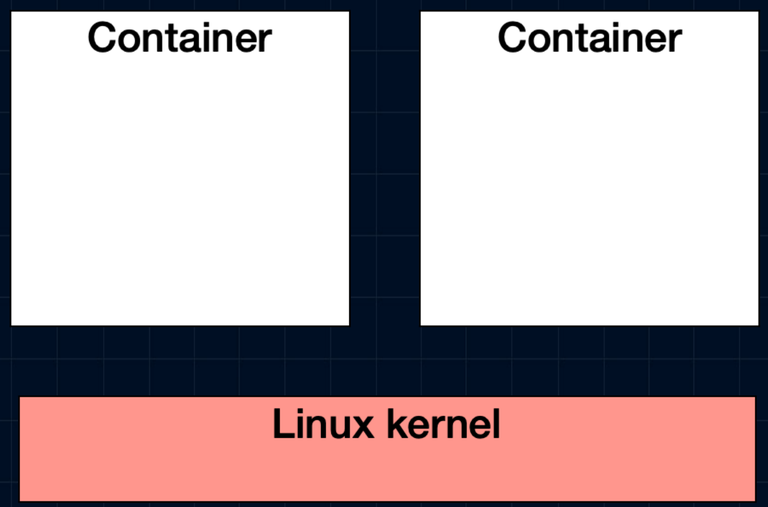
Redis测试	QPS	平均时延(ms)	P99时延(ms)
Runc容器	53200	1.21	3.12
Kata BM容器	54789	1.17	2.13
提升比	+3.4%	+3.4%	+46%

Kernel Build测试	Total Time(s)
Runc容器	140.75
Kata BM容器	135.75
提升比	+3.6%

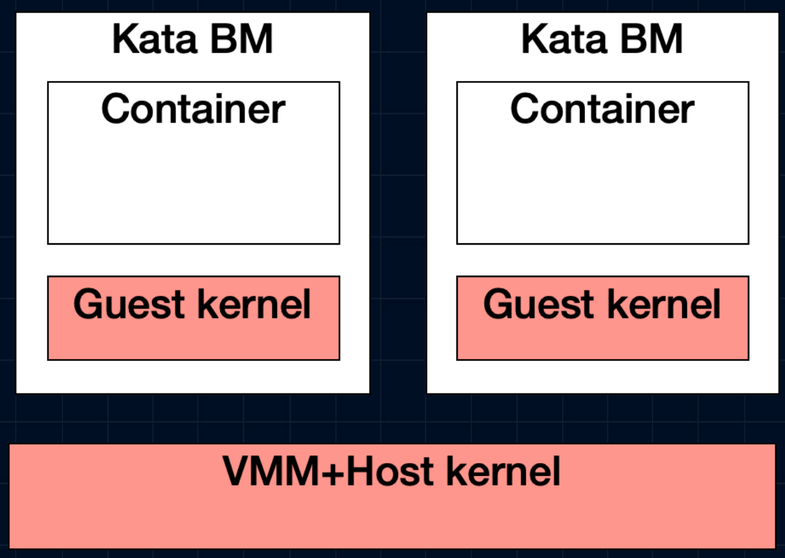
Kata BM安全性分析



传统kata容器



传统runc容器



Kata BM容器

Q & A

字节跳动 STE 团队技术交流

该二维码 7 天内 (10 月 27 日前) 有效，重新进入将更新