阿里云 × CLOUD NATIVE COMPUTING FOUNDATION

云原生技术公开课

第 29 讲

安全容器技术

王旭 蚂蚁金服资深技术专家

关注“阿里巴巴云原生”公众号
获取第一手技术资料

1 缘起
安全容器的命名

2 间接层
安全容器的精髓

3 Kata Containers
云原生化的虚拟化

4 gVisor
进程级虚拟化

5 安全容器
不止于安全

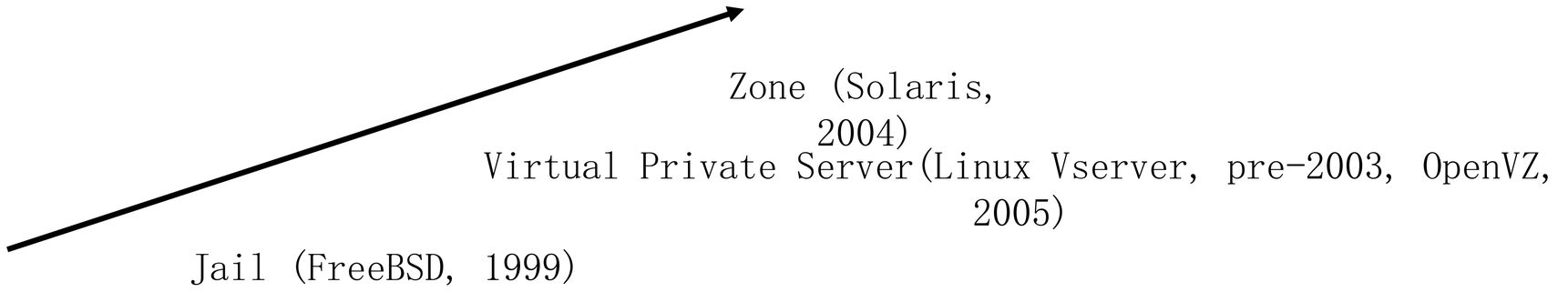
计算机科学界只有两个真正的难题——缓存失效和命名。

——Phil Karlton

“OS-Level” Virtualization

OS-level virtualization refers to an OS paradigm in which the kernel allows the existence of multiple isolated user-space instances. Such instance, called containers, Zones, virtual private servers, partitions, virtual environments, virtual kernel or jails…

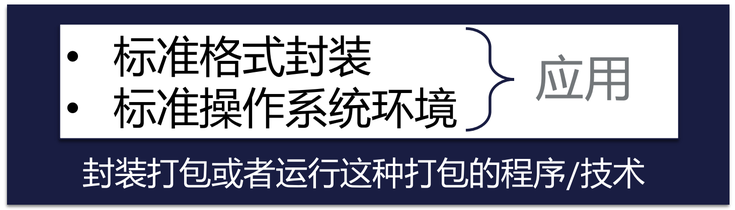
--- Wikipedia



云原生语境下的容器与安全容器

云原生语境下的容器，实质是“应用容器”——以标准格式封装的，运行于标准操作系统环境（常常是Linux ABI）上的应用打包——或运行这一应用打包的程序/技术。

——王旭



☜应用容器

安全容器是一种运行时技术，为容器应用提供一个完整的操作系统执行环境（常常是Linux ABI），但将应用的执行与宿主机操作系统隔离开，避免应用直接访问主机资源，从而可以在容器主机之间或容器之间提供额外的保护。

——王旭

1 缘起
安全容器的命名

2 间接层
安全容器的精髓

3 Kata Containers
云原生化的虚拟化

4 gVisor
进程级虚拟化

5 安全容器
不止于安全

安全问题的唯一正解在于允许那些（导致安全问题的）**Bug** 发生，但通过额外的隔离层来阻挡住它们。

—— LinuxCon NA 2015, Linus Torvalds

1 缘起
安全容器的命名

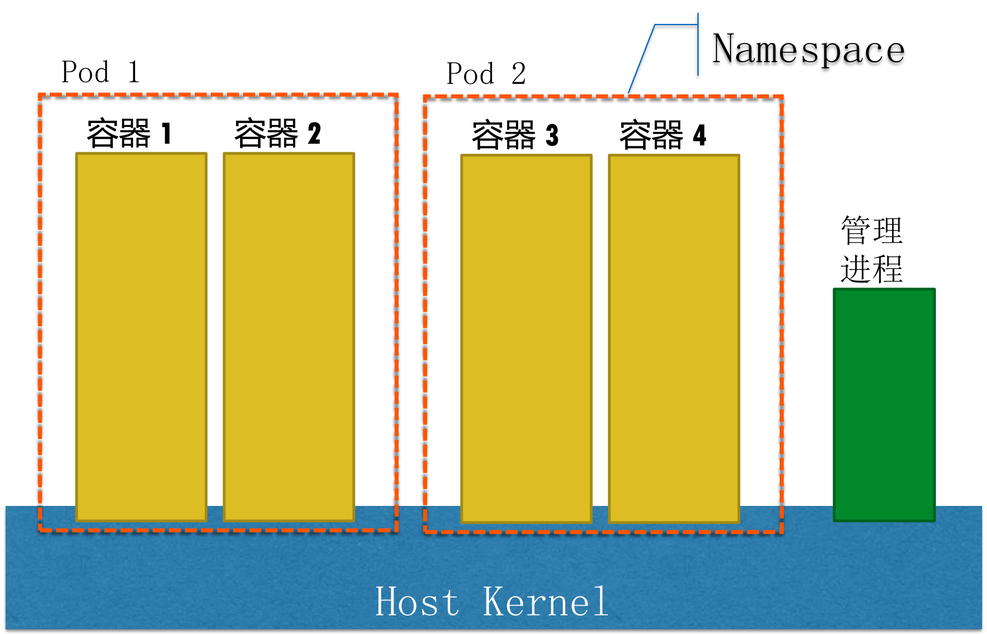
2 间接层
安全容器的精髓

3 Kata Containers
云原生化的虚拟化

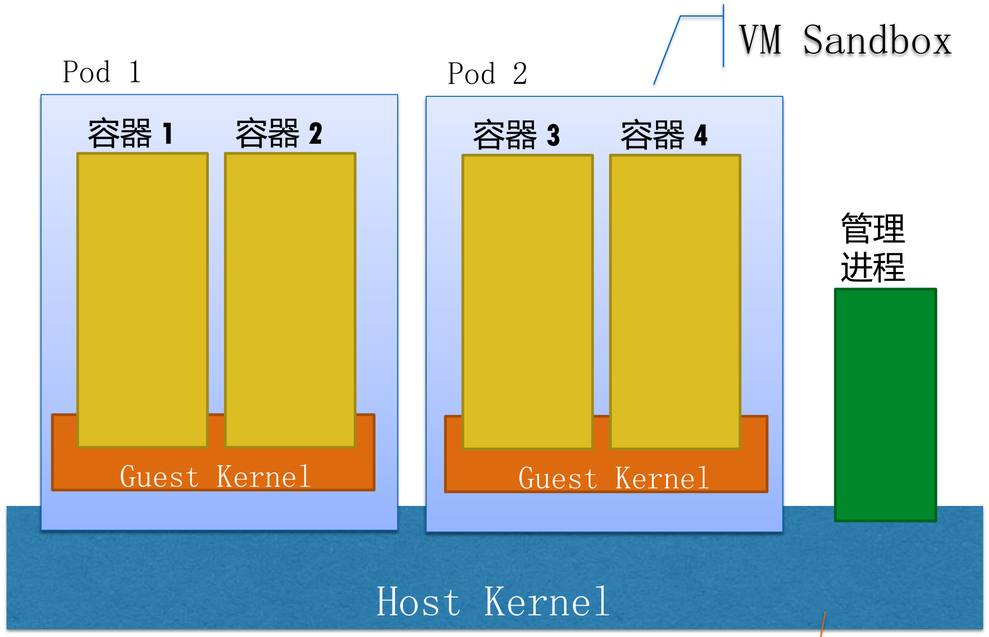
4 gVisor
进程级虚拟化

5 安全容器
不止于安全

用虚拟机做PodSandbox



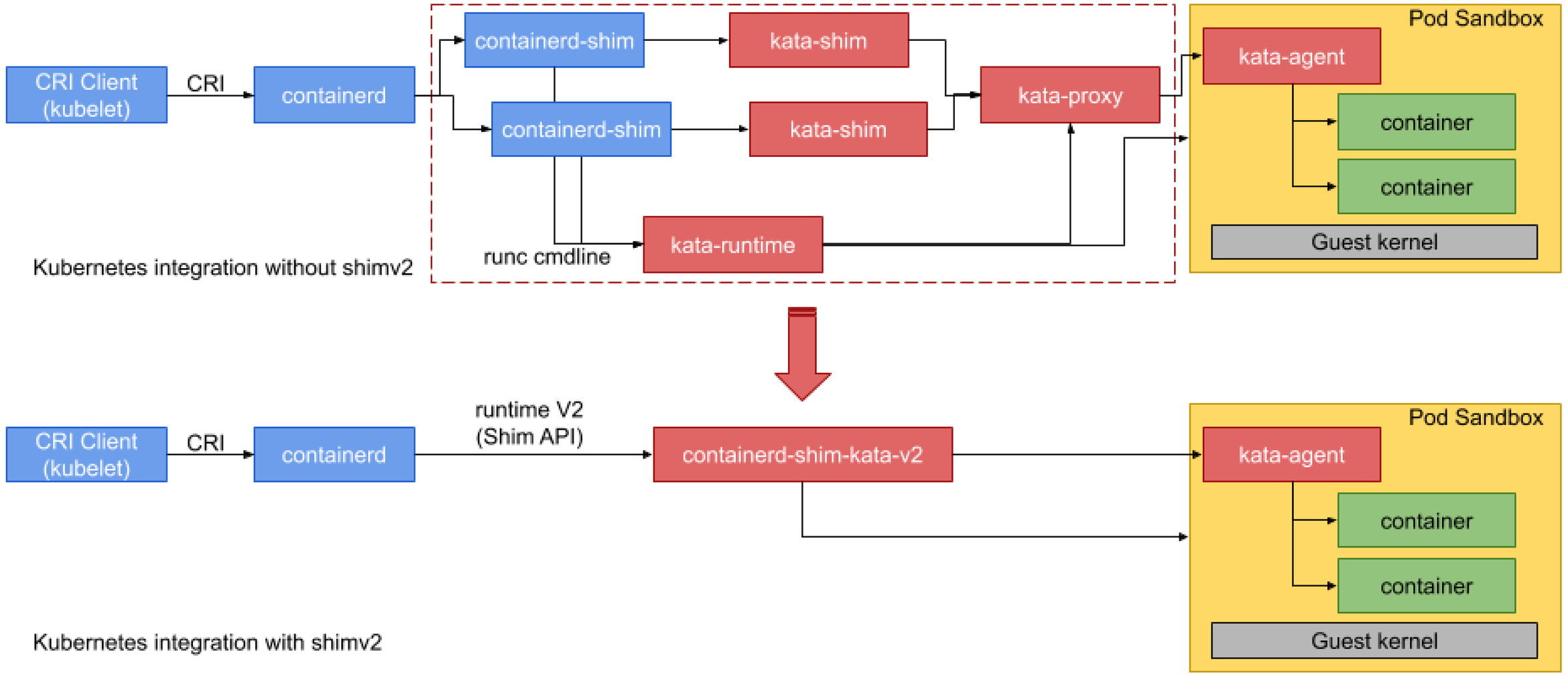
传统Container



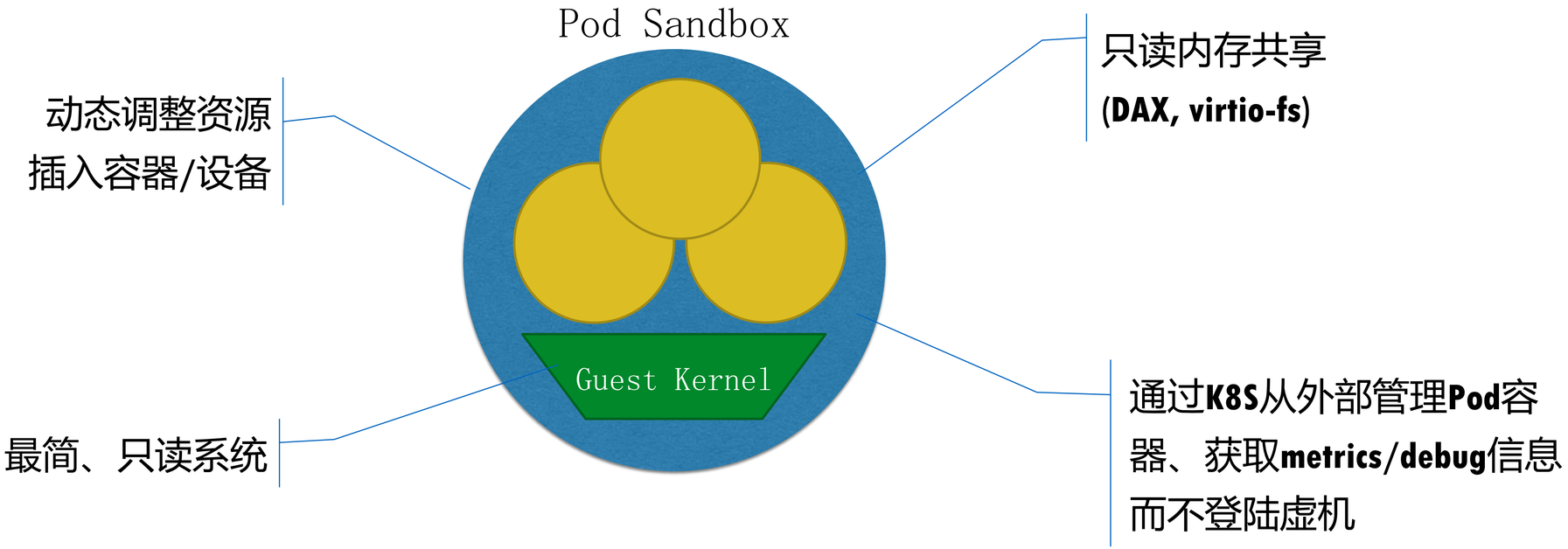
Kata 安全容器

1. 虚拟机实现 PodSandbox (qemu, firecracker, ACRN, cloud-hypervisor)
2. Guest 内有 kernel, 但没有完整操作系统环境，只负责运行容器
3. 符合OCI规范

通过 shim-v2 联接 Containerd/CRI-O



云原生化的虚拟化


Pod Sandbox
Guest Kernel
动态调整资源
插入容器/设备
只读内存共享
(DAX, virtio-fs)
最简、只读系统
通过K8S从外部管理Pod容器、获取metrics/debug信息而不登陆虚机

1 缘起
安全容器的命名

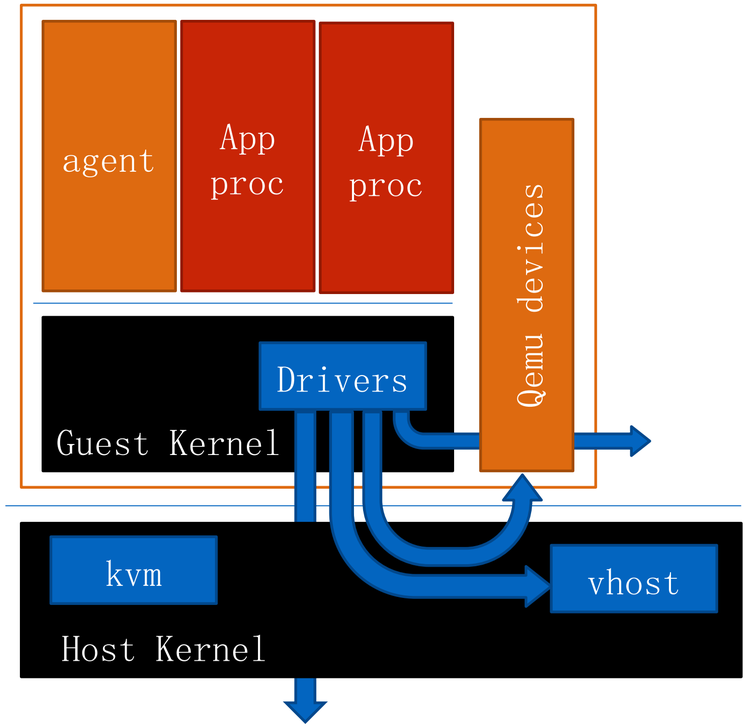
2 间接层
安全容器的精髓

3 Kata Containers
云原生化的虚拟化

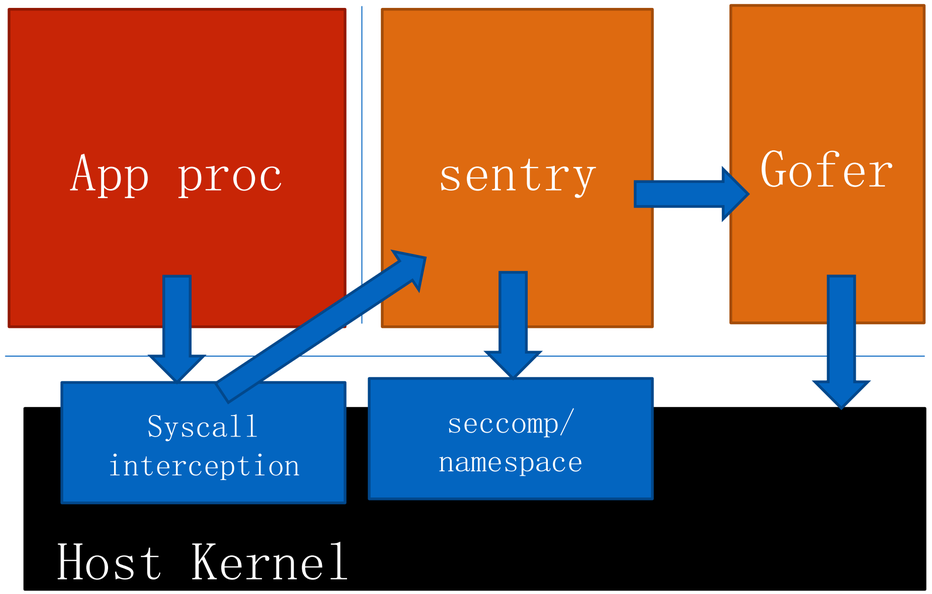
4 gVisor
进程级虚拟化

5 安全容器
不止于安全

Kata Containers vs. gVisor



Kata 架构示意



gVisor 架构示意

- 宿主操作系统将只为沙箱里的应用执行大约**20%**的 **Linux** 系统调用
- 将真有必要的 **open()** 调用交给了一个专门的称为 **Gopher** 的进程来执行

1 缘起
安全容器的命名

2 间接层
安全容器的精髓

3 Kata Containers
云原生化的虚拟化

4 gVisor
进程级虚拟化

5 安全容器
不止于安全

隔离，让云原生基础设施更完美。

用户数据保护

服务质量

调度效率

安全容器不仅仅是在做安全隔离，因为安全容器隔离层的内核，相对于宿主机的内核是独立的，专门对应用服务，从这个角度说，主机/应用的功能之间做合理的功能分配和优化，展现出让人期待的潜力，将来的安全容器，可能不仅是隔离性开销的降低，甚至是提升应用的效能

思考

- 当安全容器的开销下降到什么程度的时候，安全容器可能成为主流运行时？
- 未来的安全容器技术，**Kata** 和 **gVisor** 是否有可能统一成一个，这是否需要硬件和指令集有什么改进？

谢谢观看

THANK YOU

关注“阿里巴巴云原生”公众号
获取第一手技术资料