Alibaba Group 阿里巴巴集团

运用新技术解决有状态应用的冷热迁移挑战
迁移策略+新容器运行时

阿里巴巴高级技术专家　叶磊（稻农）

Alibaba Group 阿里巴巴集团

Agenda

1. 容器迁移背景及现状
2. 管理面支撑Pod迁移
3. runC引擎的可迁移性
4. 新运行时带来的机会

Alibaba Group 阿里巴巴集团

阿里集团100%容器化的秘诀

- 架构演变
  运行：从集中式到分布式
  运维：从分散到集中

- 资源使用的演变
  从物理机到VM--共用到隔离
  从VM到容器--提升资源利用率

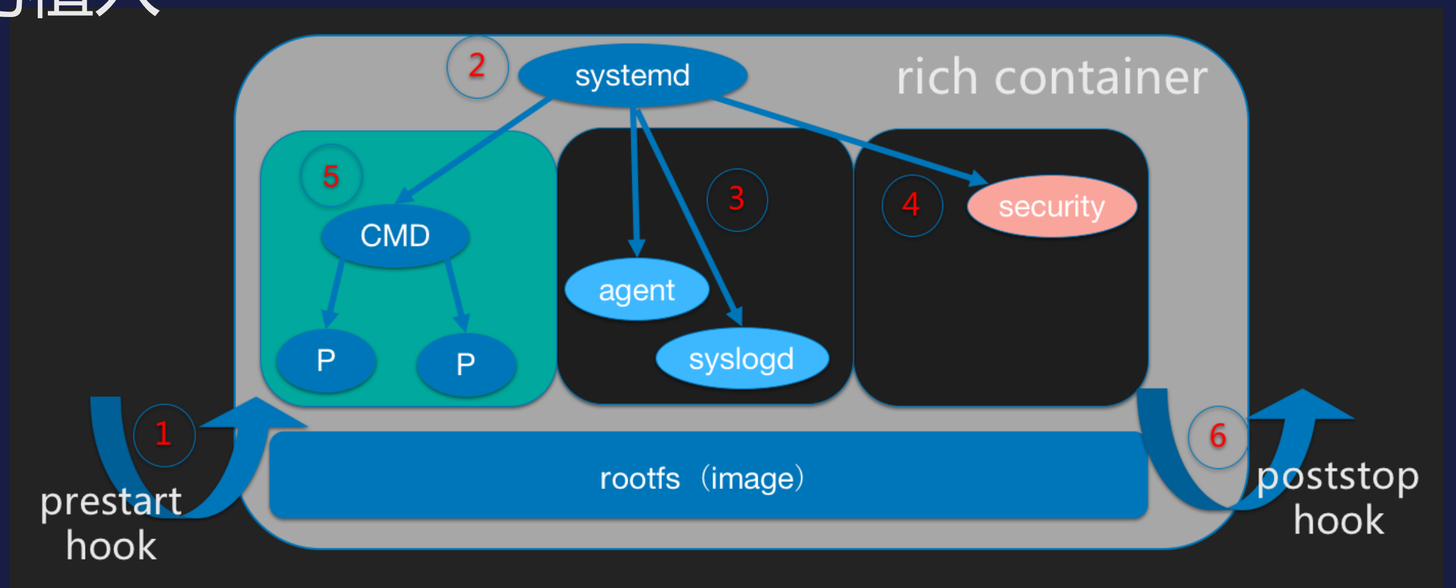
容器的要素 阿里内部运维和应用视角
有独立IP--能够被独立访问
能够ssh登陆--必要时能上得去
登陆后能够看到一个独立的环境
资源隔隔离--使用量和可见性

和物理机的
使用体验一致

Alibaba Group 阿里巴巴集团

富容器

- 存量业务最容易的接入方式
- 完善的进程管理模型：信号传递，退出清理和回收
- 友好的用户视角：用户可以直接像vm一样使用
- 运维和安全插件用户无感知的植入



Alibaba Group 阿里巴巴集团

容器可携带状态迁移成为规模化运维的痛点

典型场景

- 容器故障，管理平台针对故障容器发起迁移
- 管理员发起的大规模容器迁移；例如机柜下线、机房搬迁；一般走先扩容、再缩容流程
- 大量应用携带有状态标签，迁移需要开发同学参与才能实现

难点

- 管理面：Kubernetes作为容器编排系统的事实标准，只有Pod扩缩容的机制，迁移需求讨论15年已经有Issue，一直无下文
- 执行层面：runC作为容器运行时主流，虽有CRIU的项目辅助，仍然无法提供完善可靠的迁移机制

Alibaba Group 阿里巴巴集团

Agenda

1. 容器迁移背景及现状
2. 管理面支撑Pod迁移
3. runC引擎的可迁移性
4. 新运行时带来的机会

Alibaba Group 阿里巴巴集团

当前K8s系统的Pod迁移为空白

- Pod作为基础部署单位，拥有独特标识及IP
- K8s主推无状态容器的扩缩容
- K8s骨干系统不支持Pod标识及IP冲突
- 阿里周边还有大量管理系统，依赖标识来管理容器应用
- K8s生态内伸缩过程与迁移过程直接冲突
- Volume如为远程盘，很多不支持同时异地Mount
- IP管理系统不允许出现地址冲突
- ……

Alibaba Group 阿里巴巴集团

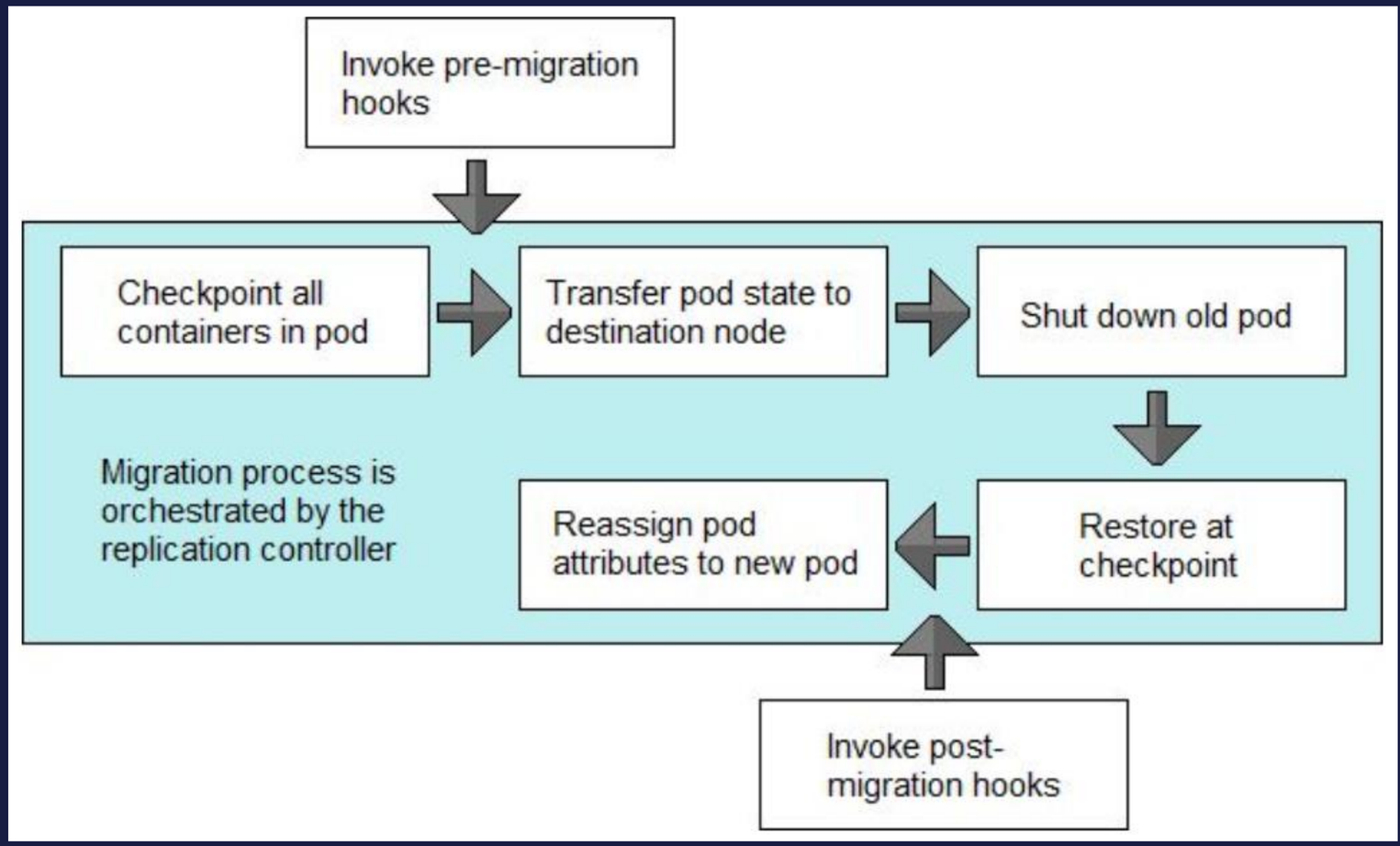
全身整容式的Pod异地重生过程

- **无差别新生**： 异地创建全新的资源等同的空Pod，请注意业务容器也选择Pause，这个过程和普通Pod创建一致；
- **唤醒记忆**：通过新旧两个宿主机上驻留的迁移agent，把旧Pod的镜像/状态灌入这个空Pod，本阶段末切流+停掉源Pod；
- **重领身份**：设置真正的IP及全局标识到目标Pod，完成迁移.

Alibaba Group 阿里巴巴集团

全流程图示

- 宿主机上的迁移Agent + Hook为视角组织流程
- 预迁移Hook，进行新Pod资源申请及占位新Pod创建
- 后迁移Hook，完成身份信息的惊险一跃



图片来自https://www.cs.ubc.ca/~bestchai/teaching/cs416_2018w1/project2/project_m6r8_s8u8_v5v8_y6x8_proposal.pdf

Alibaba Group 阿里巴巴集团

Agenda

1. 容器迁移背景及现状
2. 管理面支撑Pod迁移
3. runC引擎的可迁移性
4. 新运行时带来的机会

Alibaba Group 阿里巴巴集团

runC的迁移靠CRIU

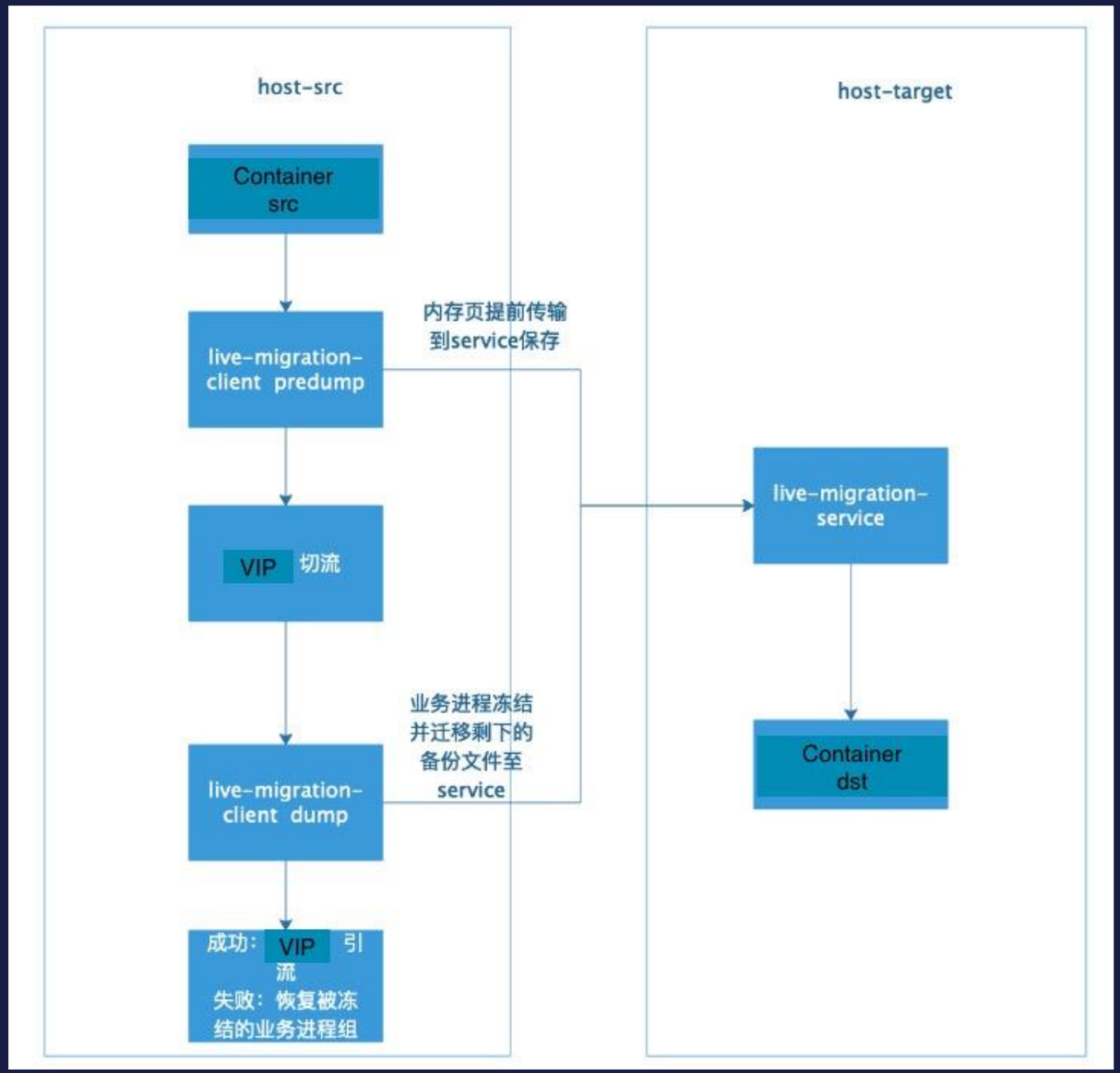
- **CRIU全称是Checkpoint/Restart In Userspace，这项技术仅仅通过用户态的代码实现，无需对Linux内核做任何修改**；
- 内存、进程执行上下文、Socket、文件句柄等迁移已经逐一攻克；
- 实测最耗时的是内存迁移，带宽不是瓶颈情况下，大约1G/s;

- **目前支持的还不好的部分**：

  外部资源、文件锁、依赖的外部设备、不同用户的进程等

容器引擎视角的迁移流程

- OCIagent-src发起热迁移请求，调用live-migration-src对容器进行predump；
- live-migration-src完成predump并将文件传输到migration-target侧
- VIP进行切流后，同时OCIagent在target端创建相同IP的pause容（runc）和未运行的业务容器（仅ccontainer Image，容器namespace、create参数等与src端相同）。
- OCIagent-src调用live-migration-src进行dump，live-migration-target接收最终的备份文件；同时冻结src容器。
- live-migration-target完成容器进程组restore，完成容器热迁移；
- 如果失败，则live-migration-target通知live-migration-src，恢复src容器继续运行。
- live-migration-src把结果（成功或者失败）返回给OCIagent-src
- VI 引流到target端（成功）或者恢复src容器的流（失败）



可基于任意OCI引擎修改

Alibaba Group 阿里巴巴集团

Agenda

1. 容器迁移背景及现状
2. 管理面支撑Pod迁移
3. runC引擎的可迁移性
4. 新运行时带来的机会

Alibaba Group 阿里巴巴集团

容器运行时从非黑即白到百花齐放

- **容器运行时，长期表现为不是纯容器就是纯虚拟机（运行容器镜像）；**
- 云厂商从长期实践中迸发出新的形态，可以称之为进程级虚拟化，兼具两者之长，努力修补短板；
- 以gVisor/FireCraker为代表，表现为采用现代语言重写进程级内核服务，摒弃与应用无关的设备模拟及其他冗余功能;

- 这些机制拥有进程级私有内核，文件句柄、虚拟设备、协议栈都无需借重宿主机，表现出了与虚拟机等同的自包含特性，为容器迁移提供新的动力

Alibaba Group 阿里巴巴集团

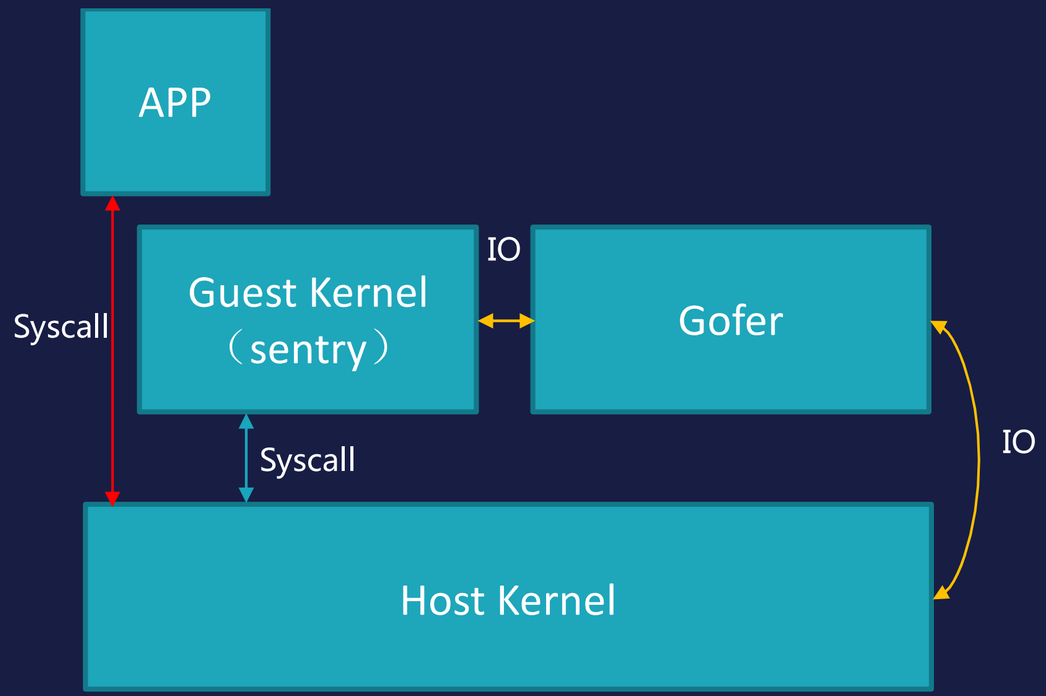
新运行时

云厂商	类别	云容器产品	安全容器技术	技术来源	状态自包含指数
AWS	国外	AWS Fargate/Lambda	Firecracker	用Google的Crosvm替换了Kata中的QEMU	90%以上，有独立内核及协议栈
Azure	国外	Azure Container Instance	Nested VM	自研	未知
Google	国外	Google Kubernetes Engine / App Engine / Cloud Functions	gVisor	自研，重写了Linux内核大部分API，去掉了VMM	90%以上，有独立内核及协议栈
华为	国内	Huawei CCE/CCI	Kata Container	基于社区的QEMU	90%以上，有独立内核及协议栈

Alibaba Group 阿里巴巴集团

gVisor运行架构

- **核心机制需要截获进程的系统调用，可采用pTrace（软）或VT-x（硬）；**
- Guest Kernel采用golang重写，自动获得了垃圾回收，runtime用户态调度等特性，相比C实现更加高效安全；
- 采用隔离实体Gofer操作Host文件，摒弃runC的文件IO直接操作;
- 下一步的工作，主要是补足广泛的场景支持及进一步提升运行效率



Alibaba Group 阿里巴巴集团

Thanks

&欢迎大家更多参与阿里开源项目