



容器云在携程的落地实践

蔡峰

携程系统研发部云平台总监

Agenda



- 在线旅游与弹性需求
- 携程容器云定位
- 携程容器云现状
- 容器编排方案选型
- 容器网络方案
- 监控方案
- 案例分享
- Q&A

携程业务概况

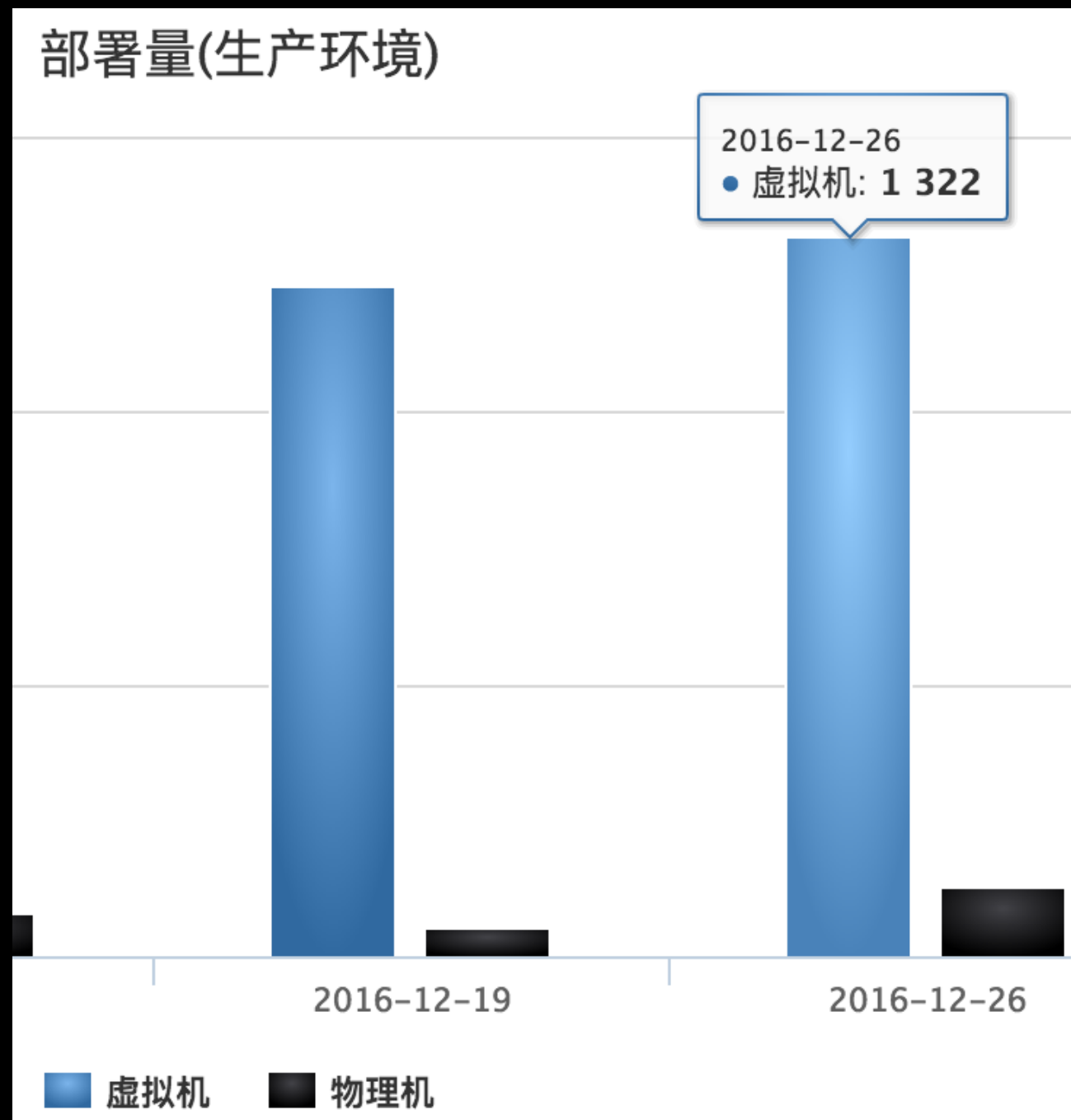


- 2016 Q4 财报
 - 交通票务营业收入同比增长98%
 - 酒店预订营业收入同比增长56%
 - 其他BU也有大幅增长
 - ...
- 2016 全年营业收入同比增长76%
- 2018 年GMV 突破10000亿
- 20+ 业务线/事业部

突发的扩容需求



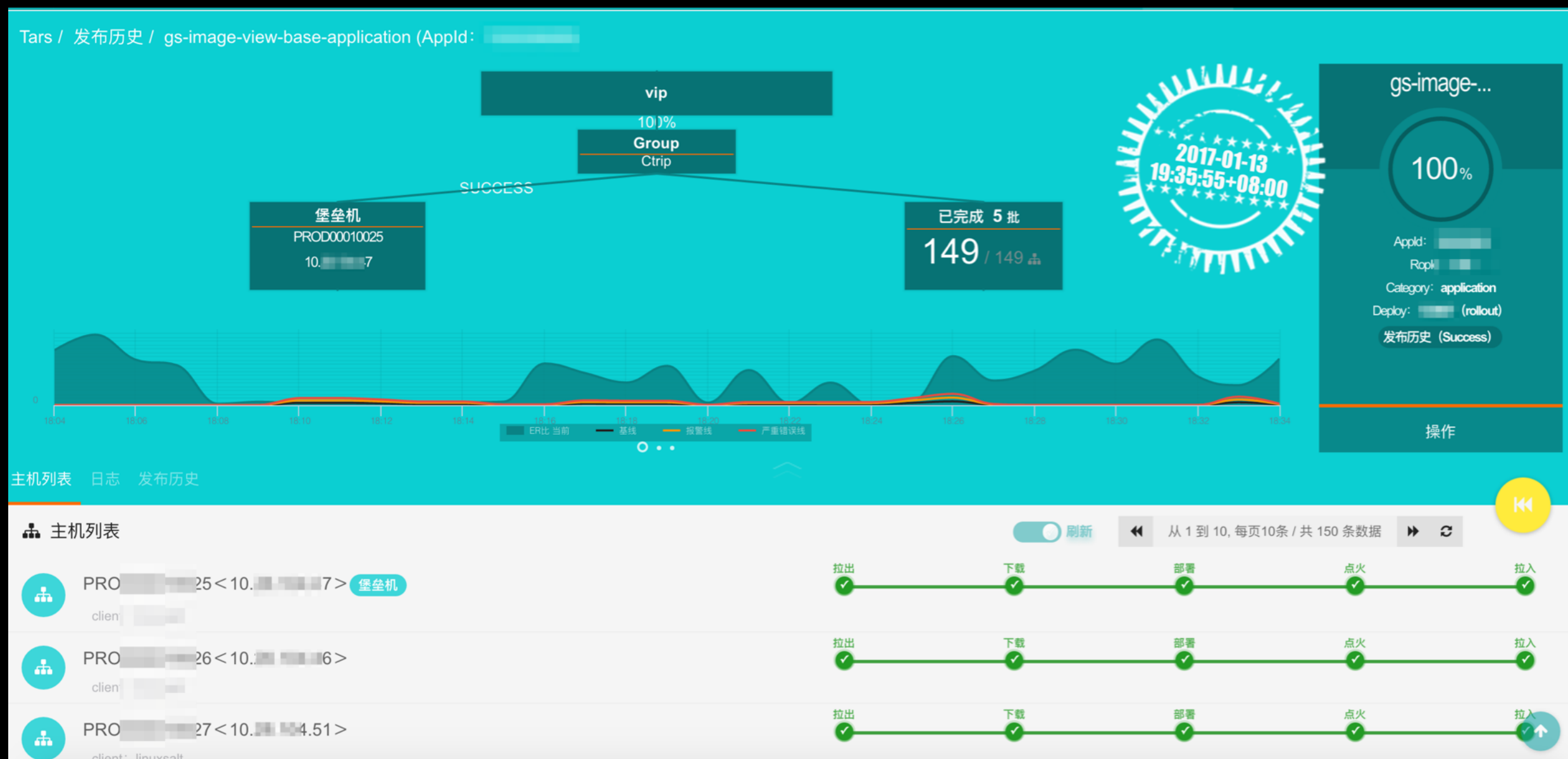
传统的虚拟机扩容，只能通过资源池的方式，提前预留一部分机器，来保证快速交付。



突发的扩容需求



基于镜像发布的容器，5分钟完成扩容 150 docker instance



携程容器云定位



- 打造极致的秒级持续交付体验 服务 20+ BU
 - 秒级 扩容、缩容、发布、回滚
 - 完善的DevOps工具链和技术服务
- 提升资源利用率
- 组件服务化(mysql/kv/mq/...) or PaaS化
- 从自动化到一定程度智能化

容器部署基本原则



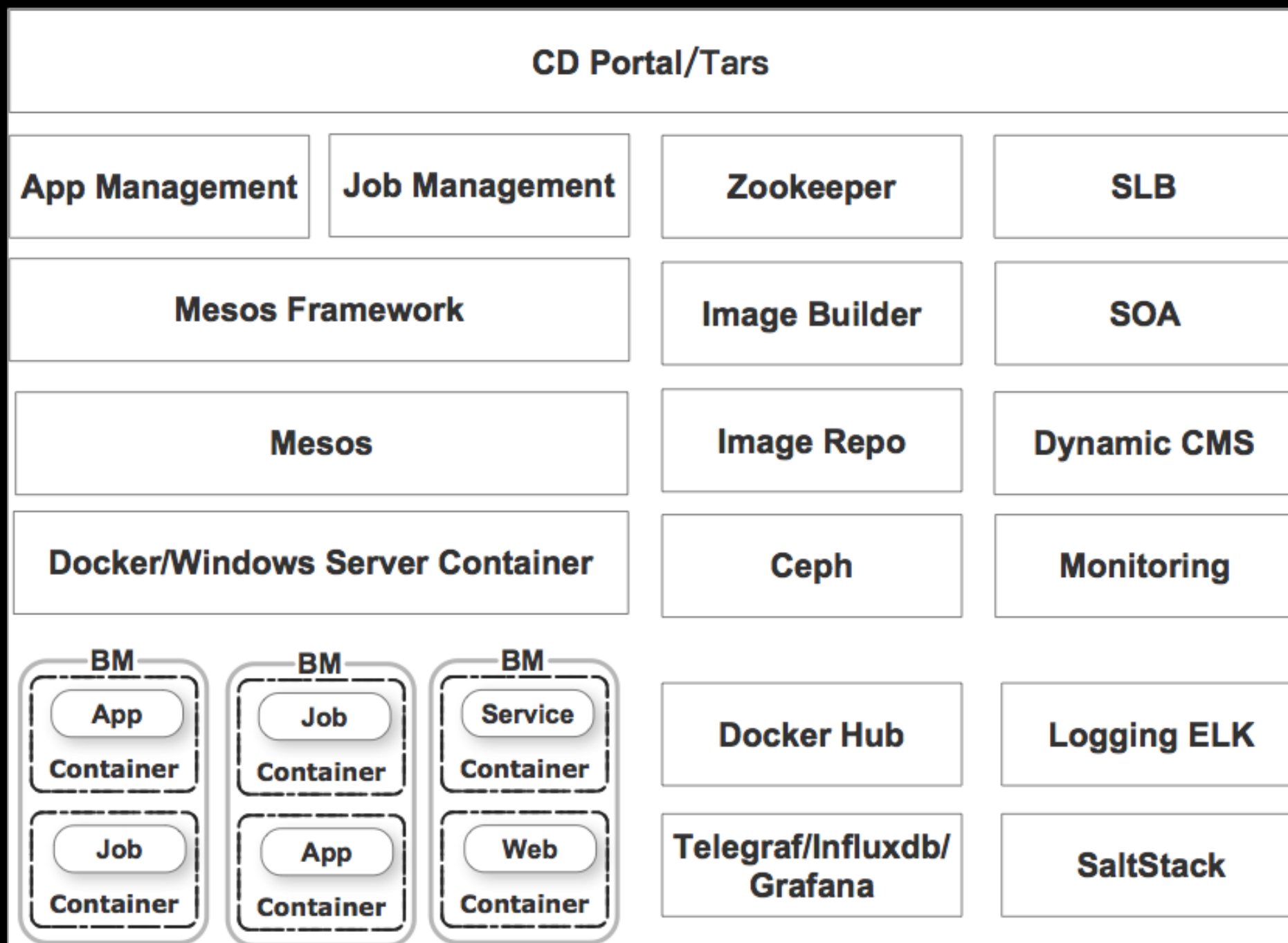
- 单容器单应用
- 单容器单IP，可路由的IP
- 容器镜像发布
- immutable infrastructure
- 容器内部只运行App，所有agent部署在host层面
 - monitoring/ES/salt/...

容器云现状



- CentOS 7.1 + Docker 1.11.1 + DeviceMapper
- 胖容器，瘦容器，windows容器共存
- 支持web service，cronjob，Mysql，Redis等服务
- FAT测试环境部署容器数量>3000
- 生产环境开始逐步开始接入

容器云的架构图



携程容器云主要建设项目



- Docker Hub — Harbor
- 服务编排，调度系统 - Mesos
- 网络管理 - OpenVSwitch Vlan
- 容器监控 - 内部系统hickwall
- 日志采集 - syslog/filebeat + ELK

Why Mesos?

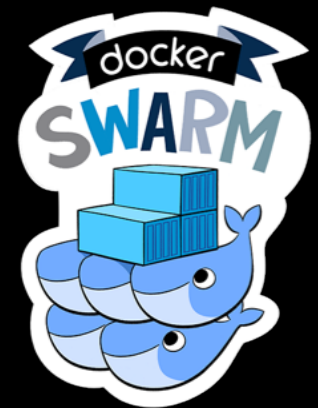


大规模部署、有
技术储备
切入CAAS领域
太重、性能问题

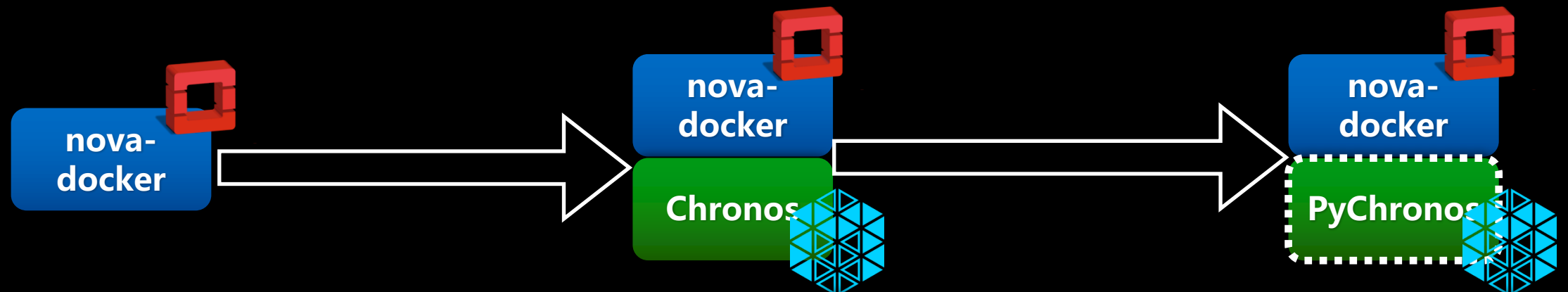
Mesos自身比
较稳定
架构简单，适合
集成

大而全，变化太
大，控制不住
Pod模型落地困
难
服务发现\域名
服务有冲突

版本太新，没有
大规模使用用例



容器编排 - 我们的探索之路

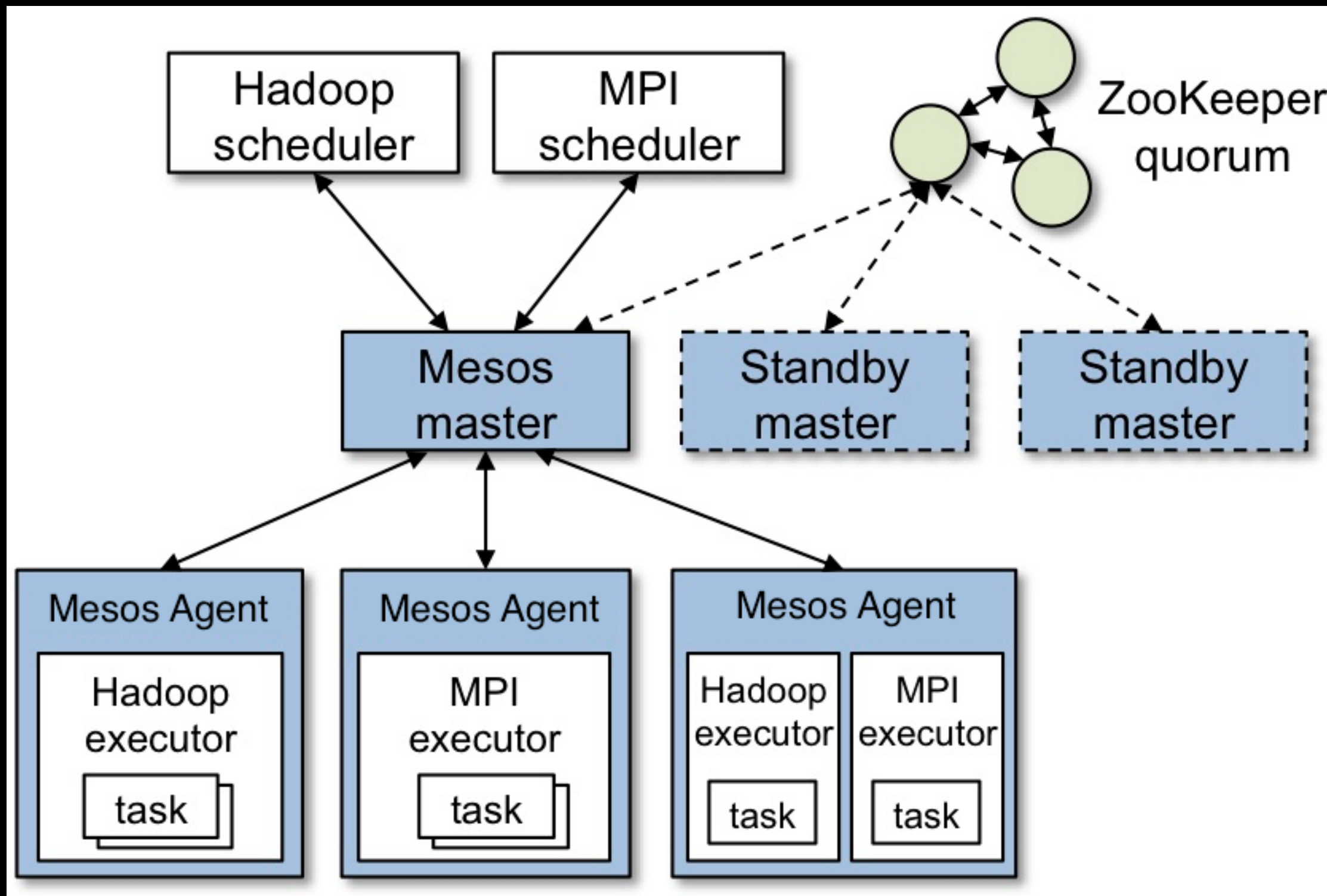


- 通过openstack提供container-as-vm
- 基于docker-compose和dwait
- 目前已在测试环境提供数千实例

- 引入Mesos和Chronos, 提供分布式Cronjob调度
- 踩了很多坑

- 使用python重新实现Chronos
- 加入long running service支持
- 接入生产 500 + Job
- 上层管理平台实现扩容、缩容等应用逻辑

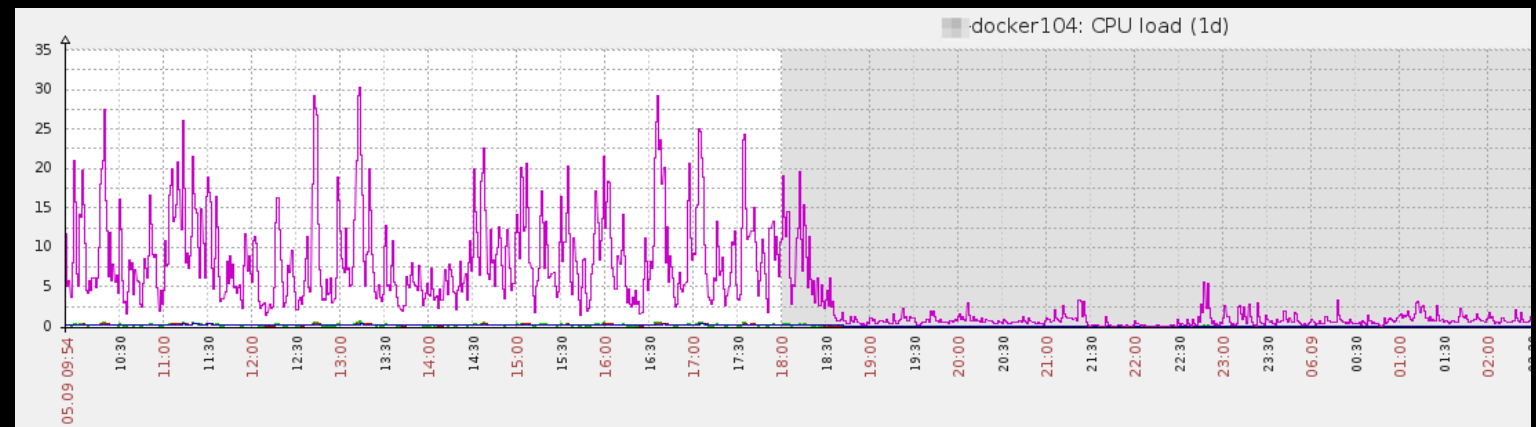
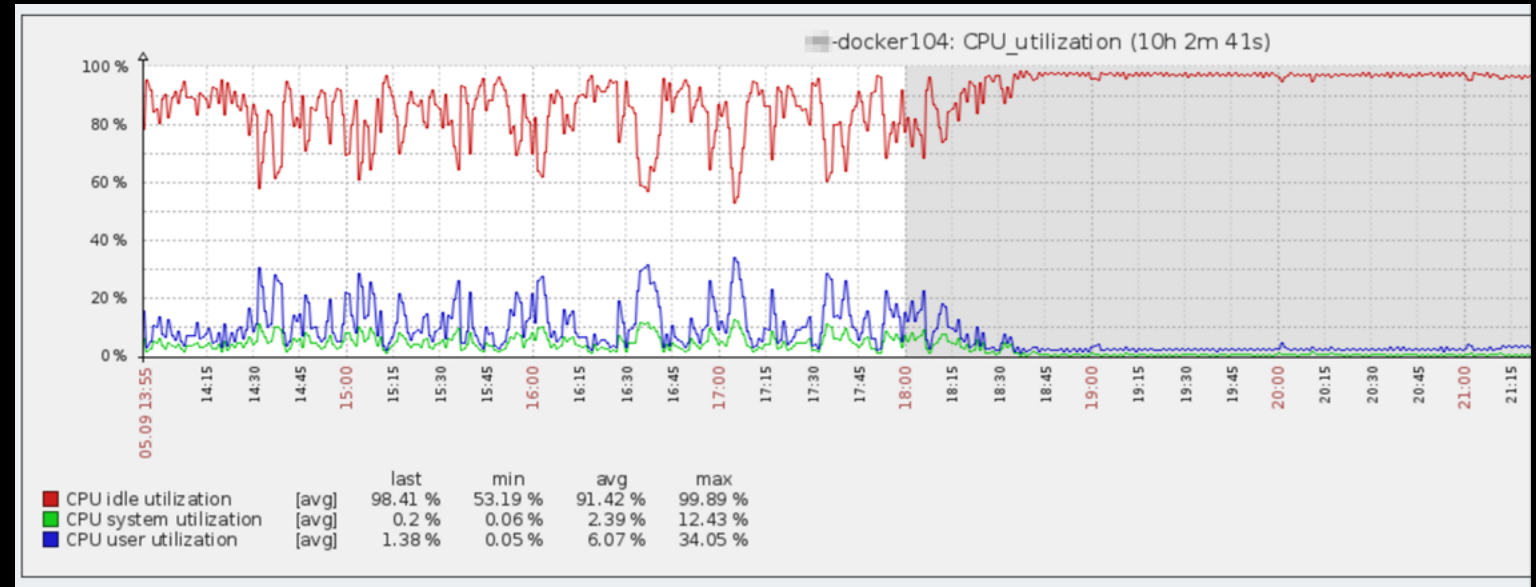
Mesos 调度模型



CExecutor



- 自定义Executor, Go 语言实现
- 避免过于频繁创建删除容器，带来的副作用
 - load/cpu
 - kernel/docker bug
- task: container 1:1 -> N:1
- 容器持久化



容器网络选型

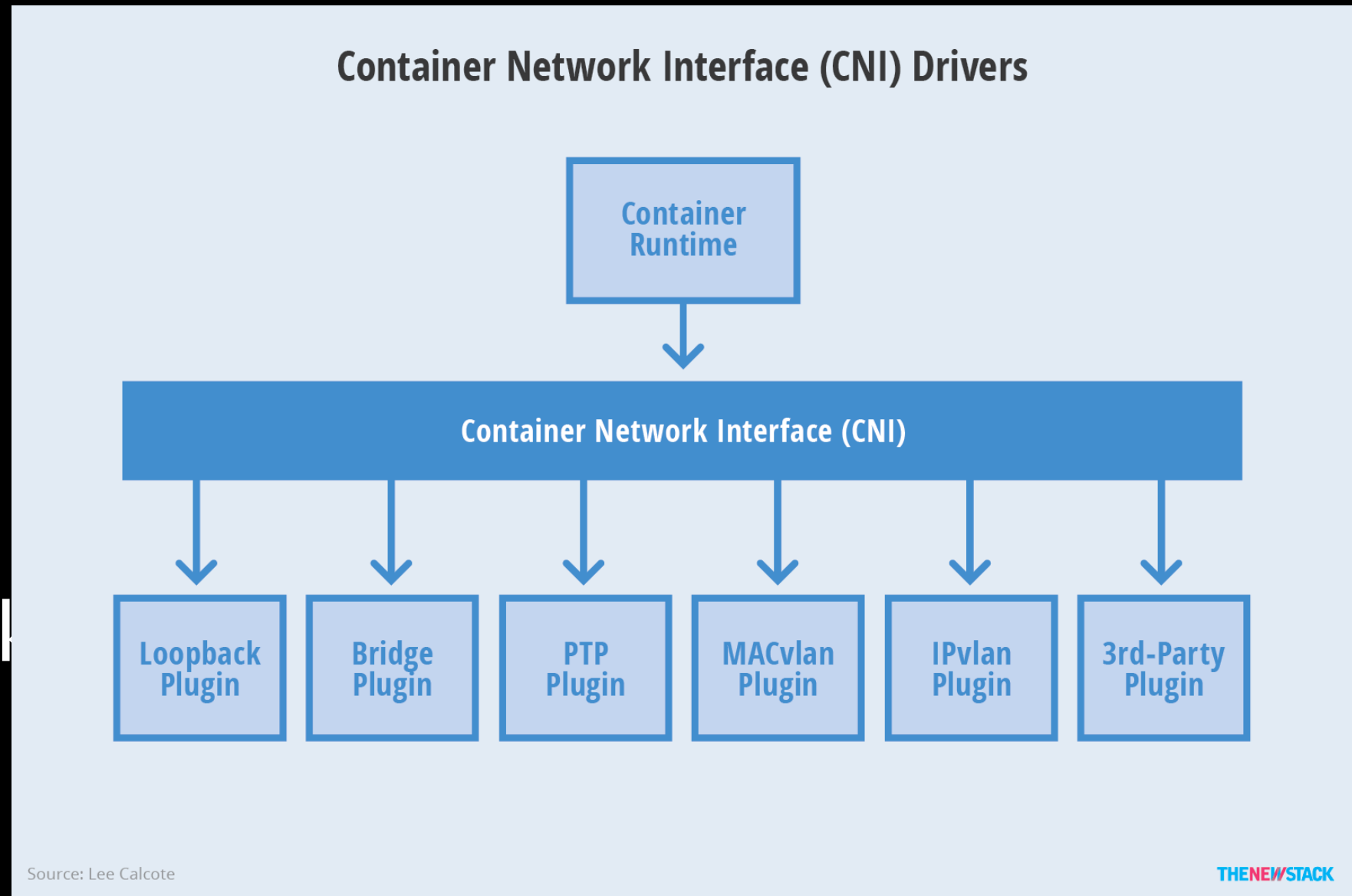


- 核心问题
 - 容器跨宿主机间的互访
- 需要做什么
 - IP管理和分配
 - 物理上的网络配置
 - Overlay方案, Vxlan
 - OpenVswitch
- CNI & CNM

CNI 模型



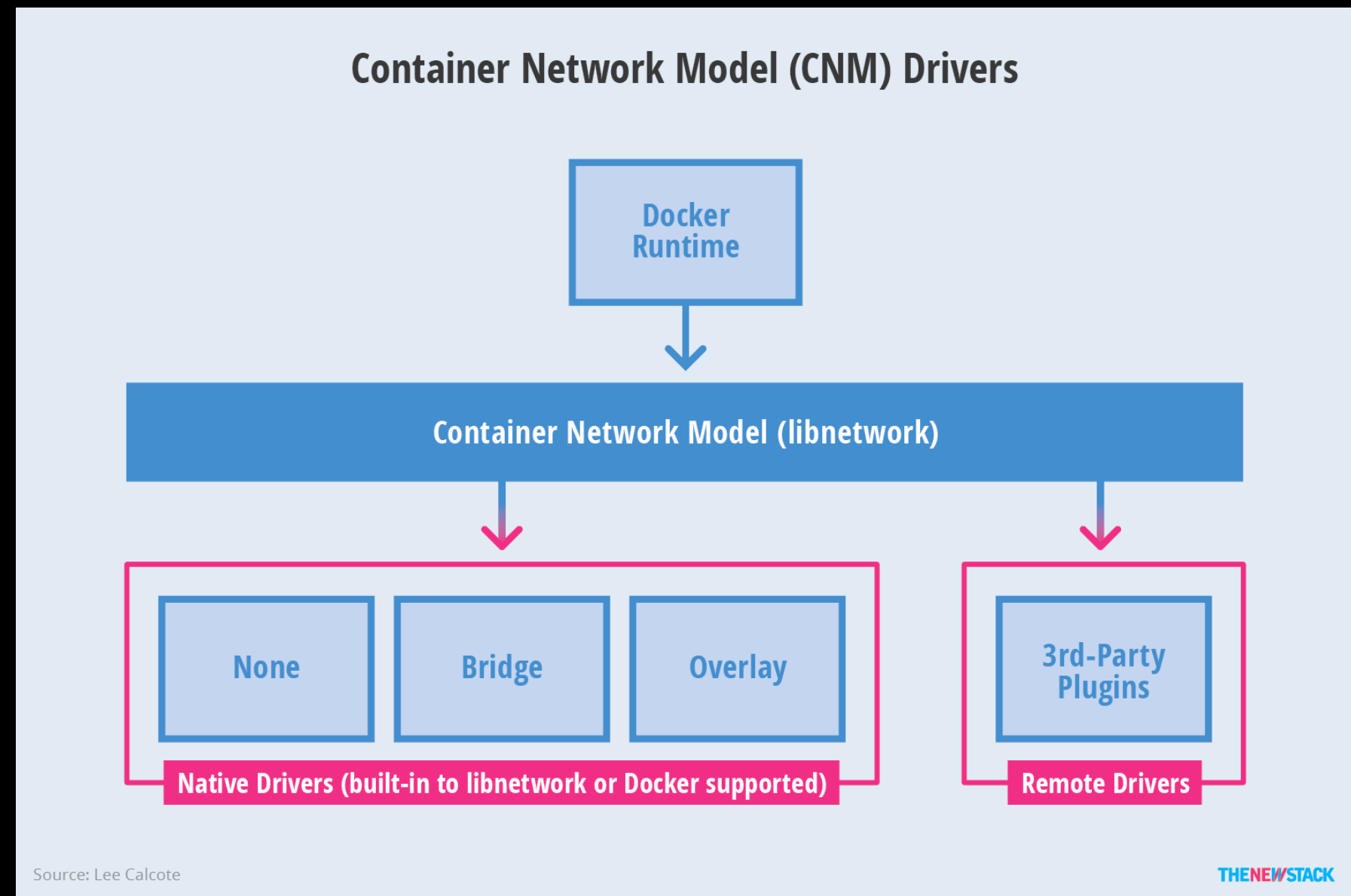
- 接口简单
- Add Network
- Delete Network



CNM 模型



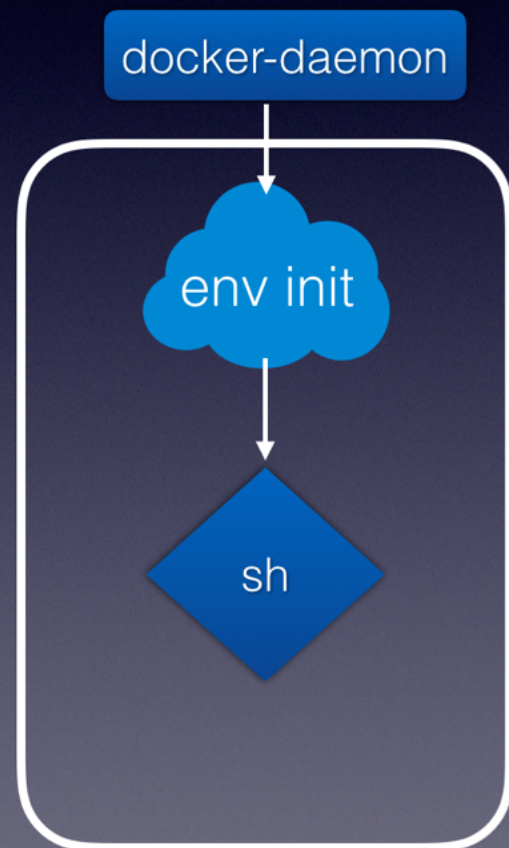
- 接口复杂
- Remote Plugin



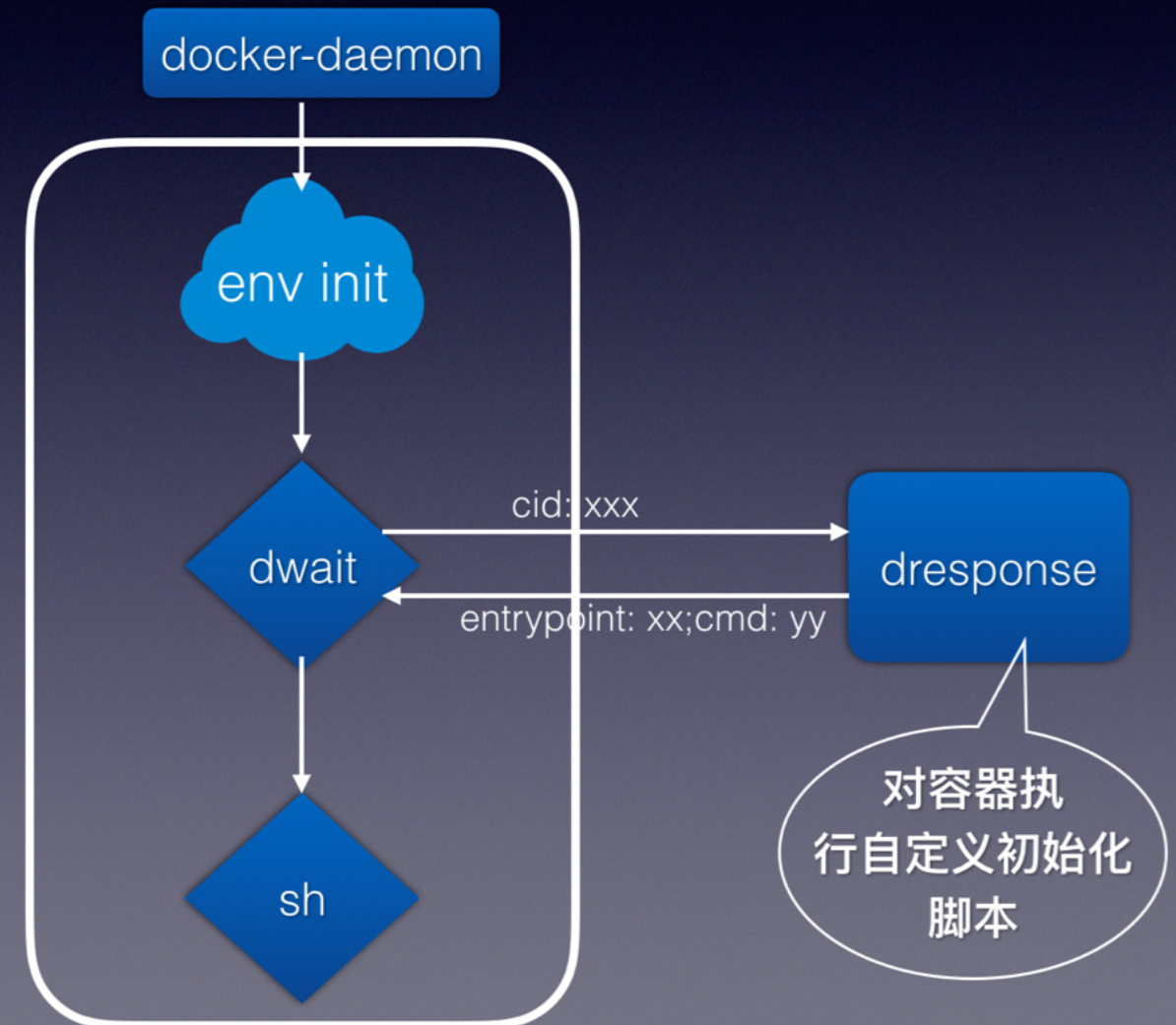
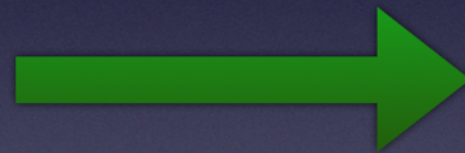
如何配置容器网络

dwait/dresponse 用Go开发

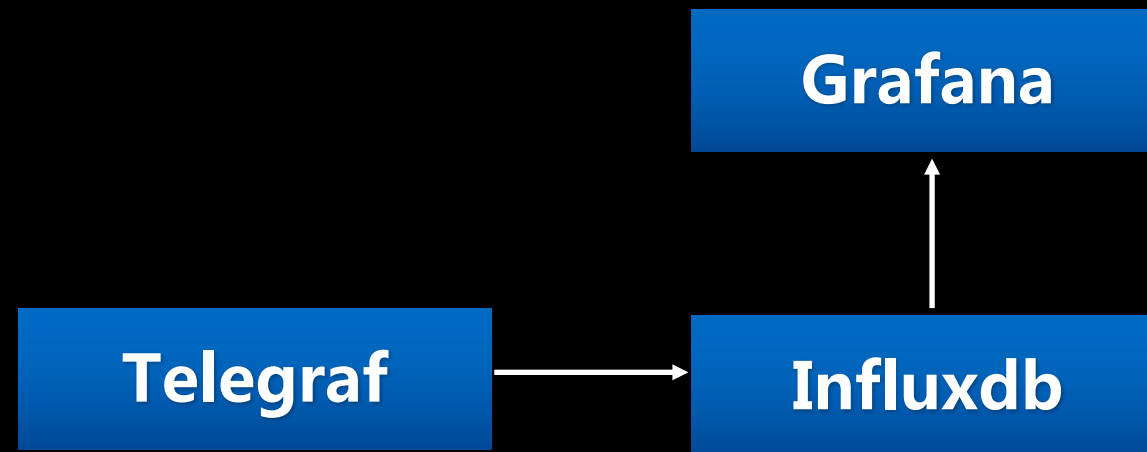
```
docker run -i busybox sh
```



```
docker run --entrypoint /d/dwait -i busybox sh
```



Mesos监控



Plugins

- chronos
- mesos-slave
- mesos-master
- system
- zookeeper
- ...

Alert

- smoke
- email
- TTS oncall

Mesos 集群监控



CLUSTER

docker103. ~~docker101~~

3.57 day

docker101

0

docker102

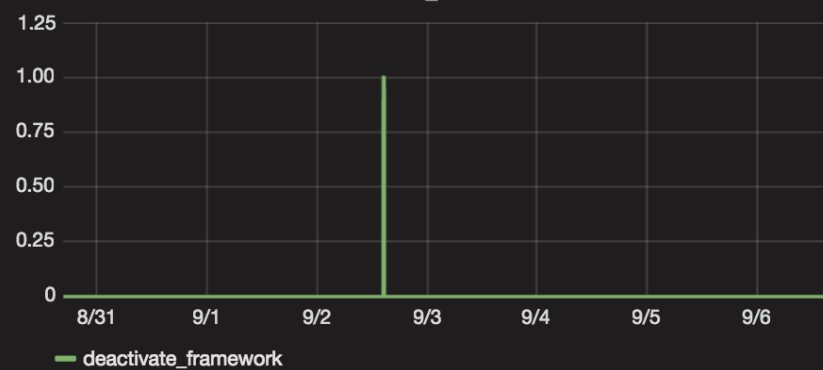
0

docker103

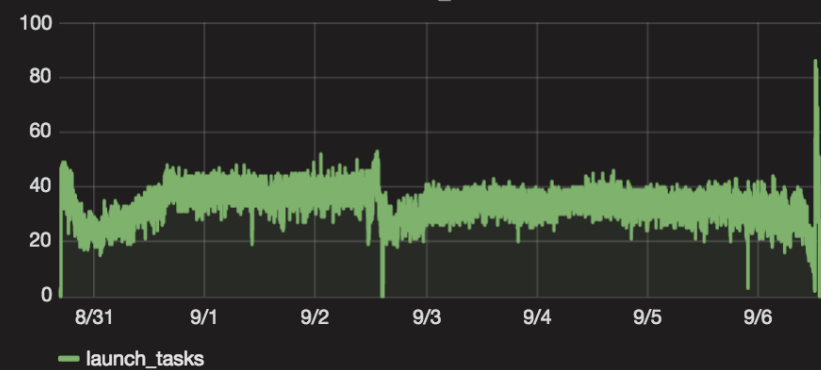
1

MESSAGES

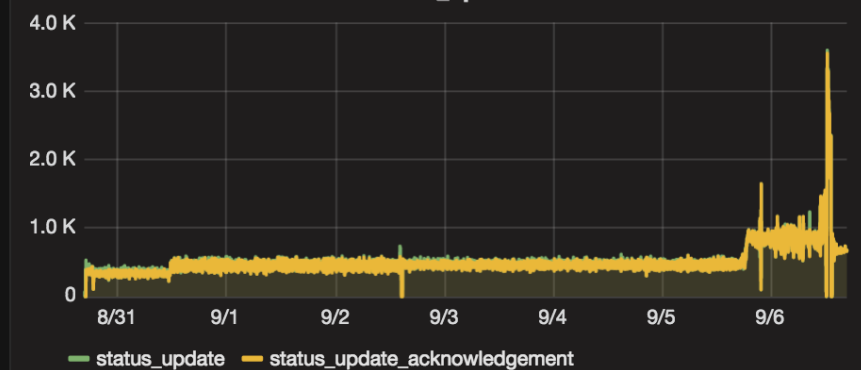
deactivate_framework



launch_tasks

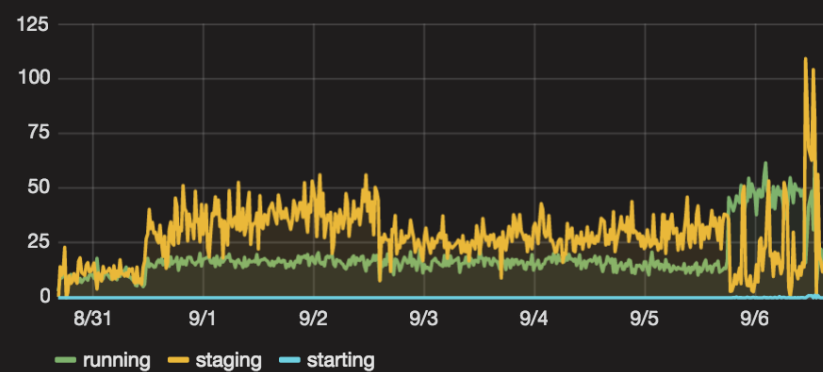


status_update

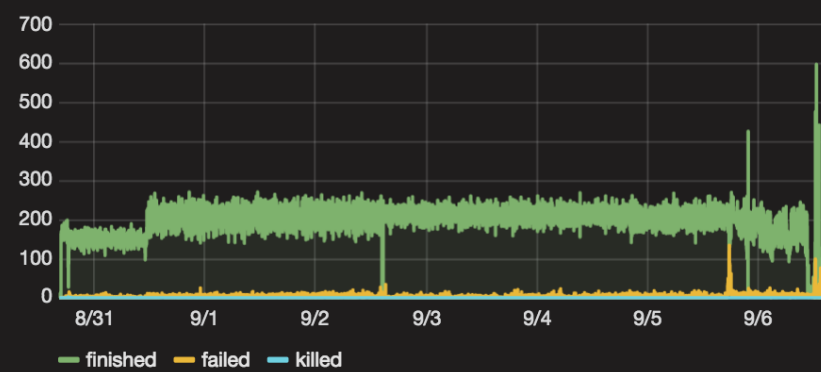


TASKS

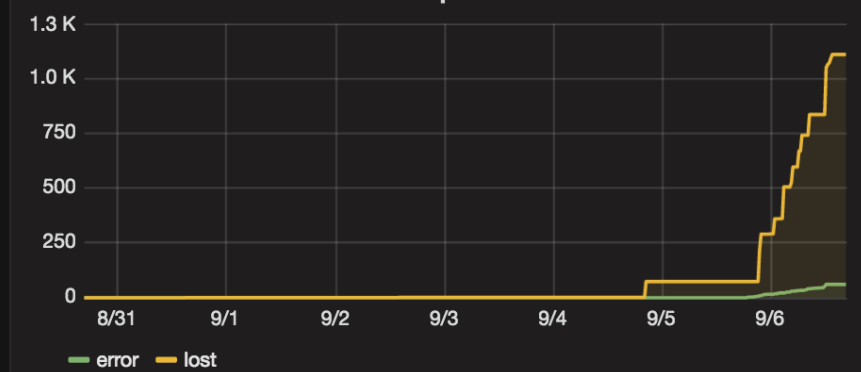
Status



Results



Exceptions



面向应用的监控



hostname: All

appid: 100007307

Alert

Release

CAT

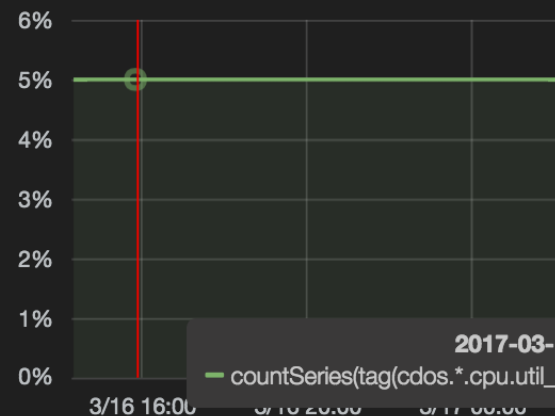
Clog

Public

在线容器个数

5

在线容器数量



CPU 平均使用率

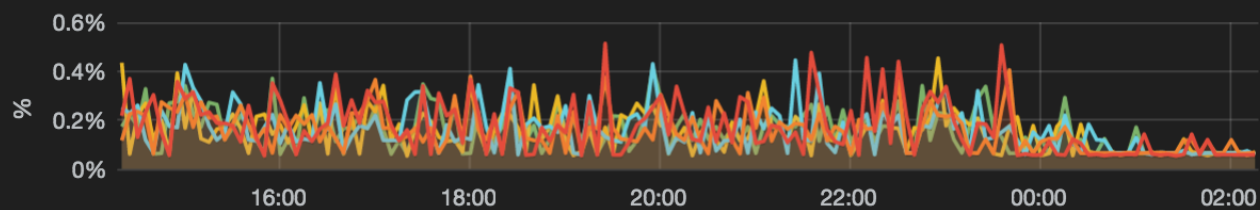
0.16%

内存平均使用率

26.35%

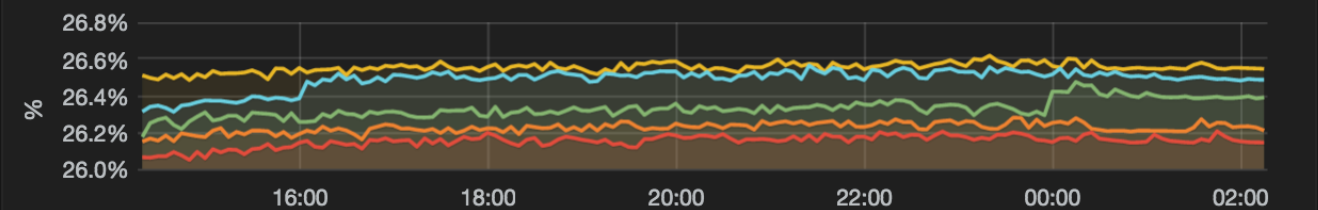
系统监控指标

CPU Util %



name	min	max	avg	current
5m cdos.SVR13812HP360.100007307.10_28_105_79.cpu.util_percent	0.05%	0.51%	0.18%	0.06%
5m cdos.SVR13737DE630.100007307.10_28_105_77.cpu.util_percent	0.05%	0.45%	0.15%	0.07%
5m cdos.SVR13740DE630.100007307.10_28_105_76.cpu.util_percent	0.06%	0.44%	0.16%	0.06%
5m cdos.SVR13728DE630.100007307.10_28_105_75.cpu.util_percent	0.06%	0.42%	0.16%	0.06%
5m cdos.SVR13746DE630.100007307.10_28_105_78.cpu.util_percent	0.06%	0.40%	0.16%	0.07%

Used Percent %



name	min	max	avg	current
5m cdos.SVR13728DE630.100007307.10_28_105_75.mem.used_percent	26.17%	26.47%	26.33%	26.39%
5m cdos.SVR13737DE630.100007307.10_28_105_77.mem.used_percent	26.48%	26.62%	26.55%	26.55%
5m cdos.SVR13740DE630.100007307.10_28_105_76.mem.used_percent	26.31%	26.56%	26.49%	26.49%
5m cdos.SVR13746DE630.100007307.10_28_105_78.mem.used_percent	26.15%	26.28%	26.22%	26.21%
5m cdos.SVR13812HP360.100007307.10_28_105_79.mem.used_percent	26.05%	26.20%	26.15%	26.14%

Docker高频创建删除的问题

- Soft lockup

- 升级内核3.10.0-327.22.2

```
[172277.750801] BUG: soft lockup - CPU#10 stuck for 23s! [docker:3679]
[172277.888953] BUG: soft lockup - CPU#16 stuck for 23s! [migration/16:185]
[172305.781592] BUG: soft lockup - CPU#10 stuck for 23s! [docker:3679]
[172305.919745] BUG: soft lockup - CPU#16 stuck for 22s! [migration/16:185]
[172319.051168] INFO: rcu_sched self-detected stall on CPU { 10} (t=3840106 jif
fies g=20206822 c=20206821 q=0)
[172319.054169] INFO: rcu_sched detected stalls on CPUs/tasks:
[172345.825579] BUG: soft lockup - CPU#10 stuck for 22s! [docker:3679]
[172345.963732] BUG: soft lockup - CPU#16 stuck for 22s! [migration/16:185]
```

- `__dm_destroy()` takes `io_barrier SRCU` lock (`dm_get_live_table`) and `suspend_lock` in reverse order. That can cause AB-BA deadlock:
- `--storage-opt dm.use_deferred_deletion=true`
- `--storage-opt dm.use_deferred_removal=true`
- <https://github.com/docker/docker/issues/24671>
- devmapper: prevent libdevmapper from deleting `/dev/mapper/docker-xxx` symbolic links in `RemoveDeviceDeferred` #24740
- 高频次创建删除: device mapper -> overlay FS

容器宿主机调优



- 容器数量多了之后，需要调整各类系统参数
 - kernel.keys.root_maxkeys
 - kernel.keys.root_maxbytes
 - kernel.pid_max
 - fs.inotify.max_user_instances
 - fs.inotify.max_user_watches
 - vm.min_free_kbytes
- 部分需要内核调用的命令，不能在容器中使用
 - ss
 - sendmail

Mesos相关的一些问题



- mesos docker inspect 执行低效，尤其是单机容器数量大
- MESOS_GC_DELAY : 6h 20k -> 1h
- MESOS_DOCKER_REMOVE_DELAY 1m
- API性能差，功能不完善、获取异步event困难



Q&A