

ECUG Con 十周年盛会



携程容器云实践

吴毅挺

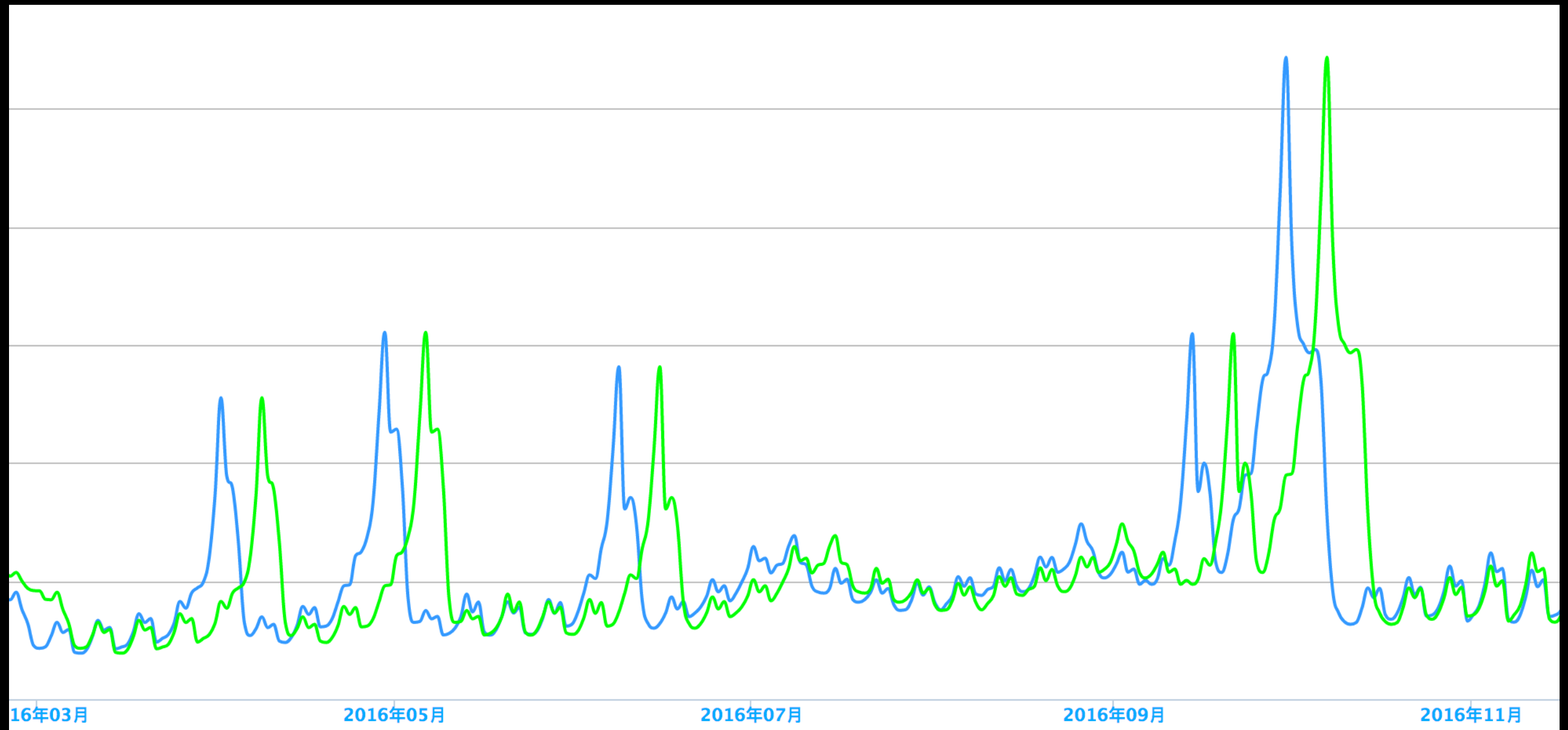
携程系统研发部高级总监

Agenda

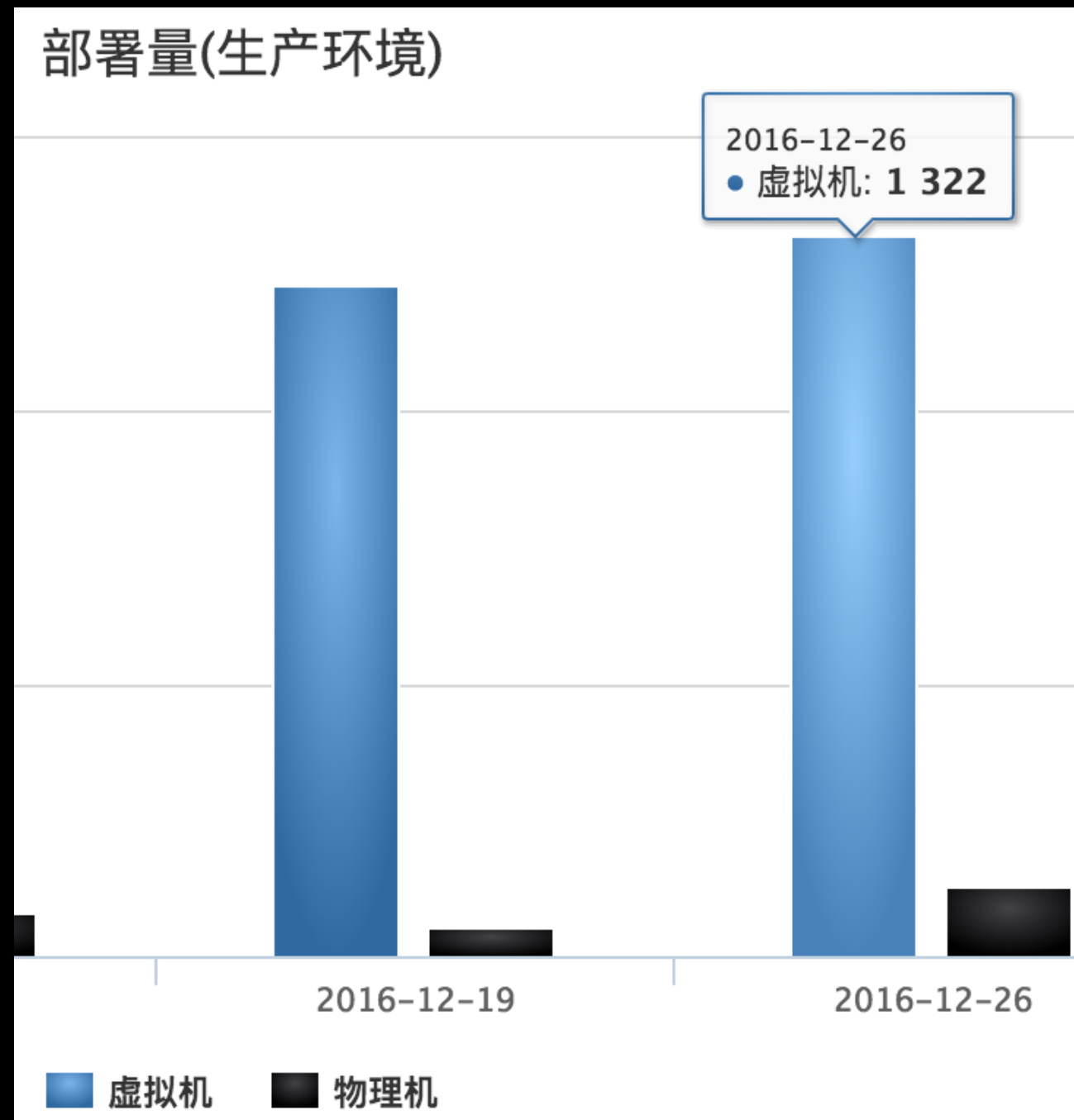
- 在线旅游与弹性需求
- 携程容器云定位
- 容器编排选型&取舍
- 容器网络选型&取舍
- 监控方案
- CDOS Overview
- Q&A

节假日业务高峰

业务较前一天突然增长 5x



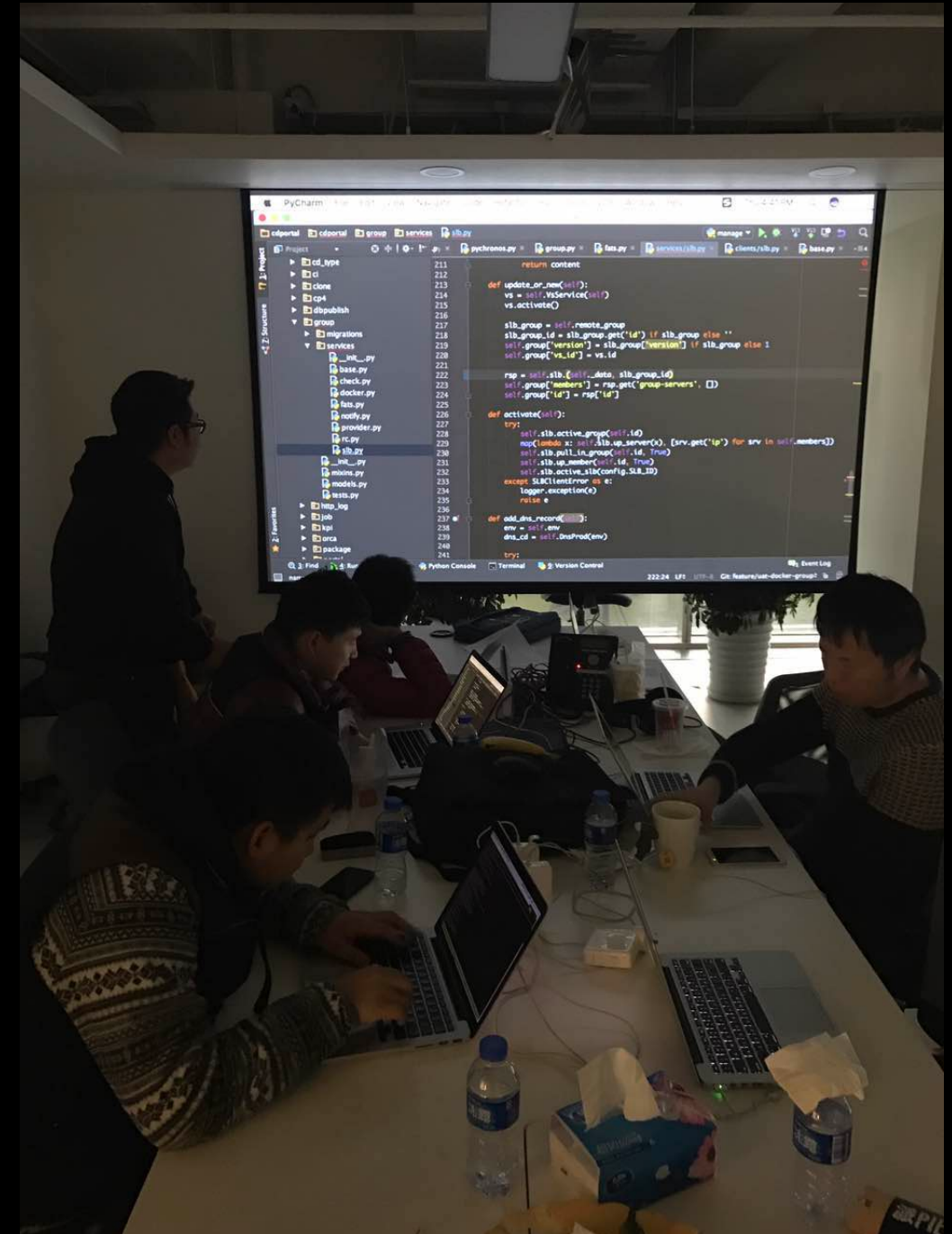
每逢佳节”倍”扩容



vm 扩容 10 - 20分钟还能更快吗?

某神秘深度学习项目急需 1000 core...

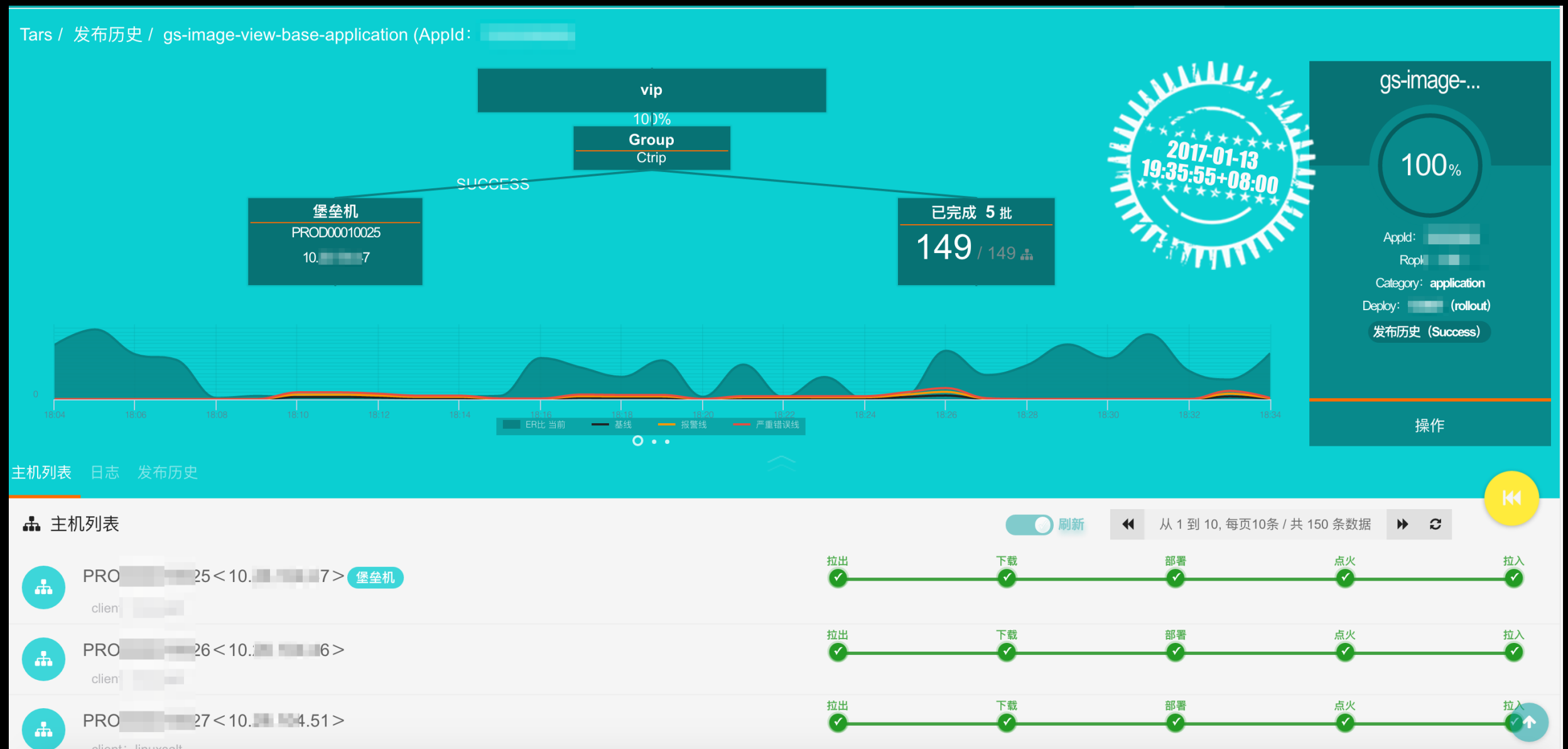
“关小黑屋”



某神秘深度学习项目需要1000core

5分钟 完成扩容 150 docker instance扩容

如果容器调度是异步&并发...



Tars: 携程新一代发布系统 开源

<https://github.com/ctripcorp/tars>

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

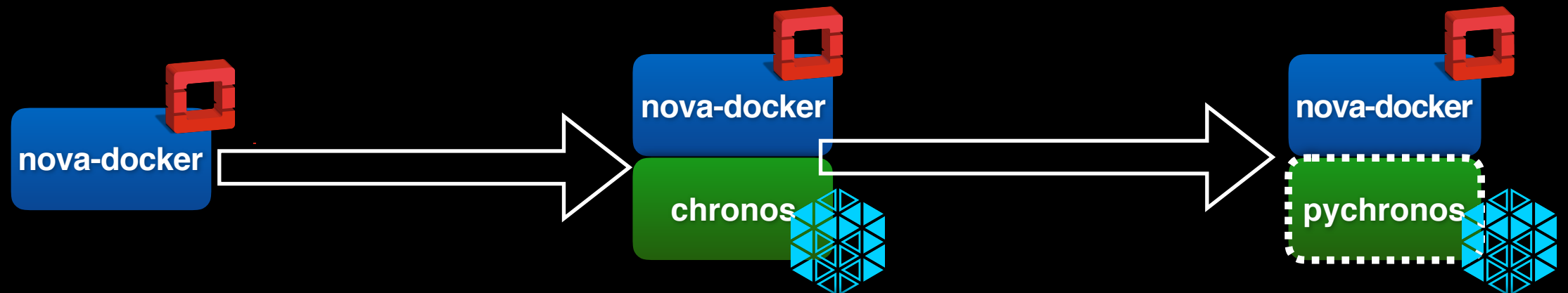
- 单容器单应用
- 单容器单IP，可路由的IP
- 容器镜像发布
- immutable infrastructure
- 容器内部只运行App，所有agent部署在host层面
 - monitoring/ES/salt/...

Mesos or K8S or OpenStack



- Openstack
 - 技术&人员储备、生产大规模部署
 - 是否是容器的正确玩法？
 - 调度效率、整体的复杂度
- K8S
 - 先进的理念和设计， replication controller/Pod/Yaml/...
 - 缺乏大规模部署案例；网络尚无比较成熟的方案(iptables)
- Mesos
 - 全新的资源调度模式

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cron job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

Why Mesos?

公司应用标准切换到pod模型成本较高
服务发现\域名服务等与公司已有的基础组件功能重叠
网络模型改造代价大



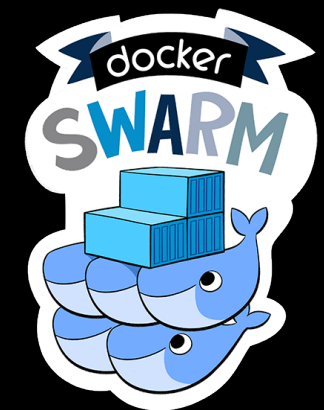
scala 学习成本
chronos不稳定, 官方维护并不积极



大规模部署、有技术储备
切入CAAS领域太重、性能问题



docker-compose+cluster-as-single-node的模型不便于承载复杂多变的应用部署需求



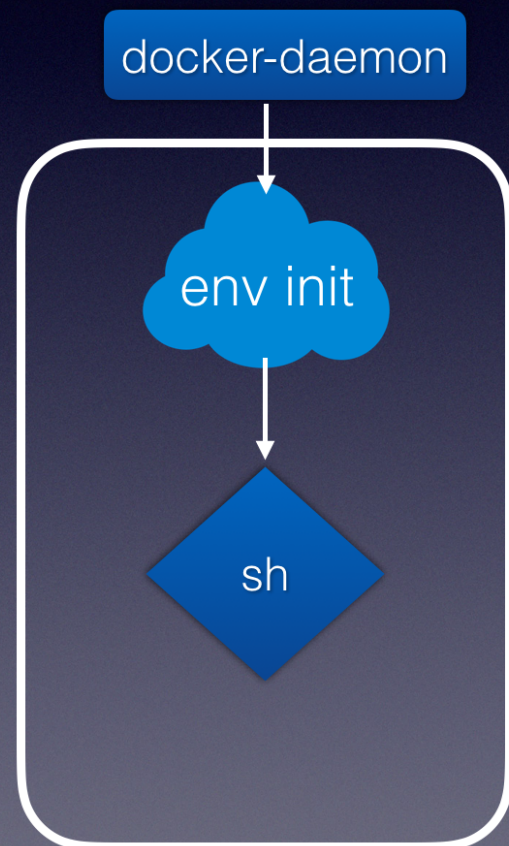
容器网络选型

- OpenStack + OVS + Vlan
 - DPDK
 - https 硬件加速
- overlay
 - flannel
 - 网段划分，每个主机独立的子网 or vlan
- vxlan + SDN controller
 - vxlan offload TOR解封装

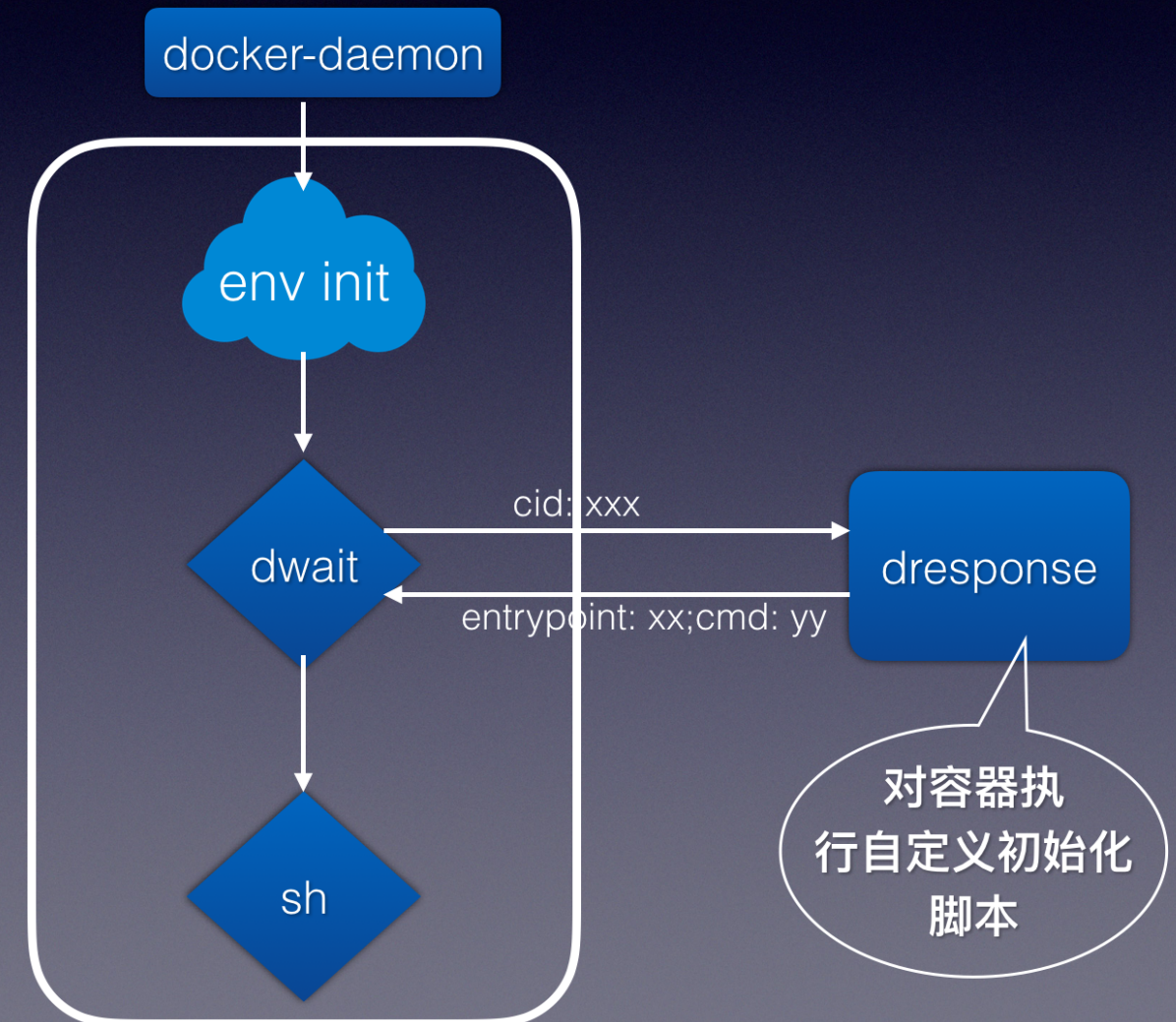
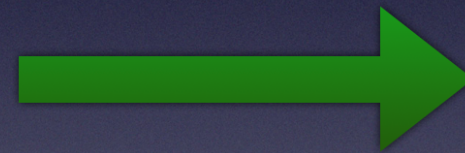
如何配置容器网络

dwait/dresponse 用Go开发

`docker run -i busybox sh`



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



Docker Issue

- 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 设备删除异步

- docker run --rm -> device EBUSY -> daemon retry 最多200次, 每次sleep 100ms

- --storage-opt dm.use_deferred_removal=true 容器删除异步

- <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

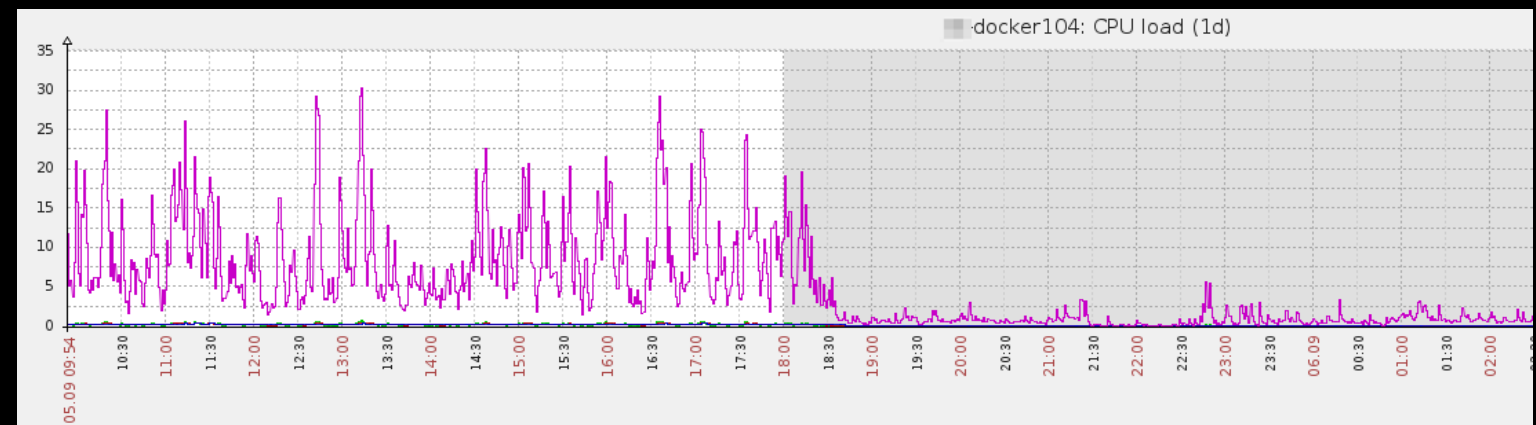
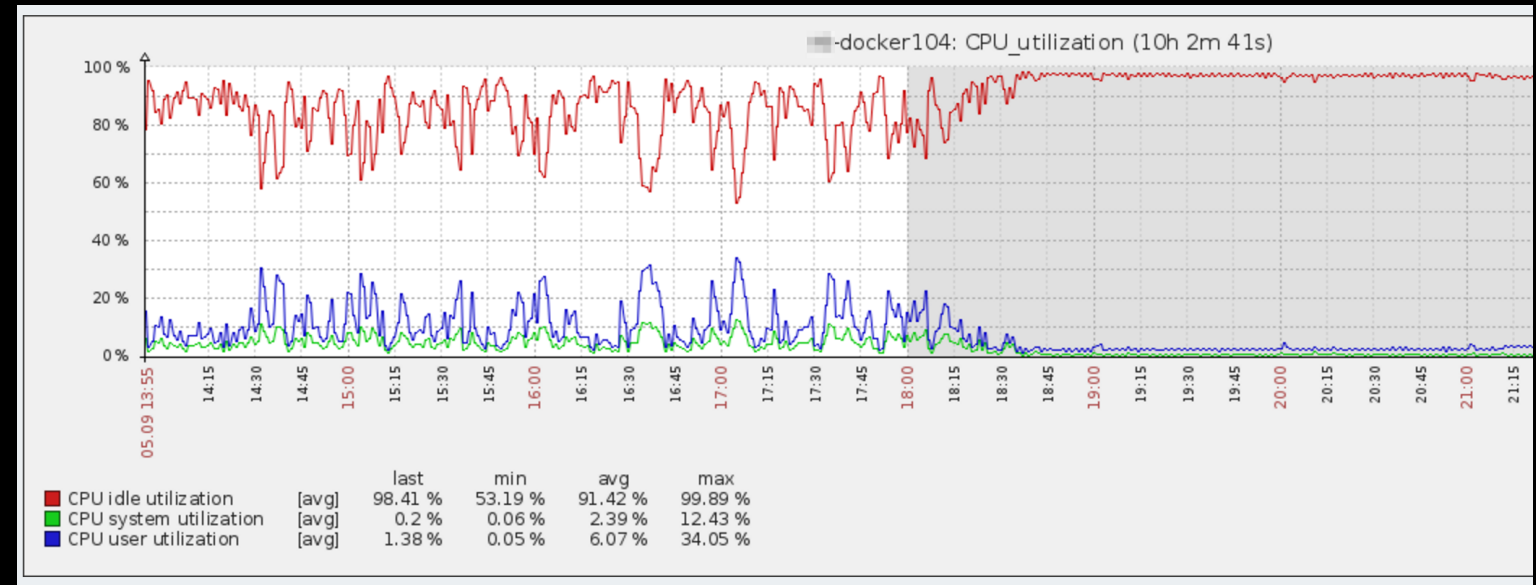
- Shared Memory (Systemd-Journald)

Mesos Issue

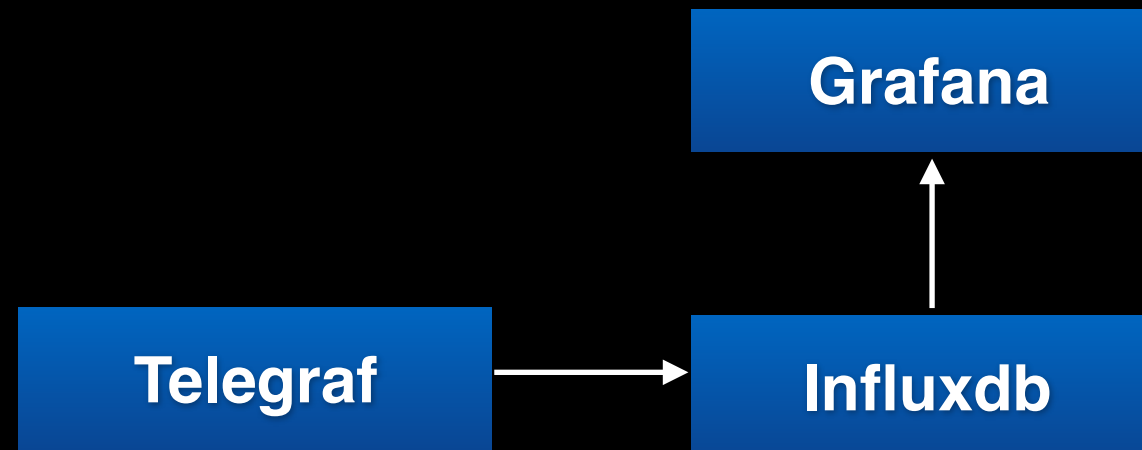
- running的容器数量较多以后，无法再启动新的容器
 - kernel.keys.root_maxkeys debian default 200, centos default 1M
- mesos docker inspect 执行低效，尤其是单机容器数量大
- MESOS_GC_DELAY : 6h 20k -> 1h
- MESOS_DOCKER_REMOVE_DELAY 1m
- docker force pull false
- API性能差，功能不完善、获取异步event困难
- overall, 很稳定, 调度性能足够

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



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

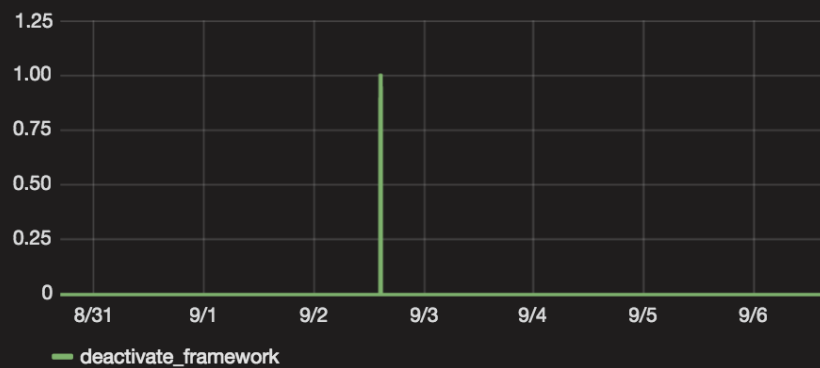
0

docker103

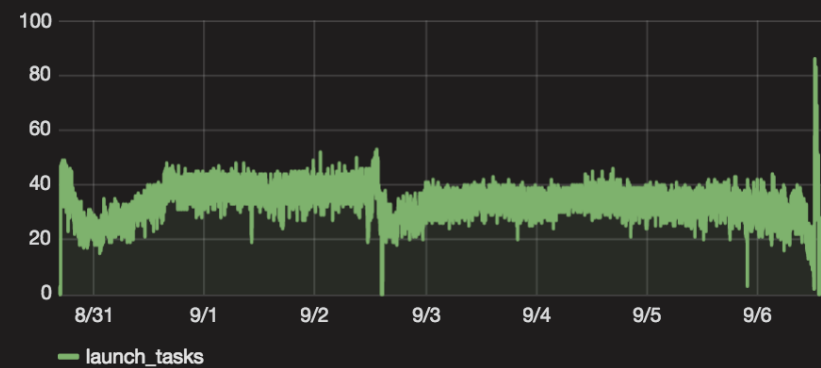
1

MESSAGES

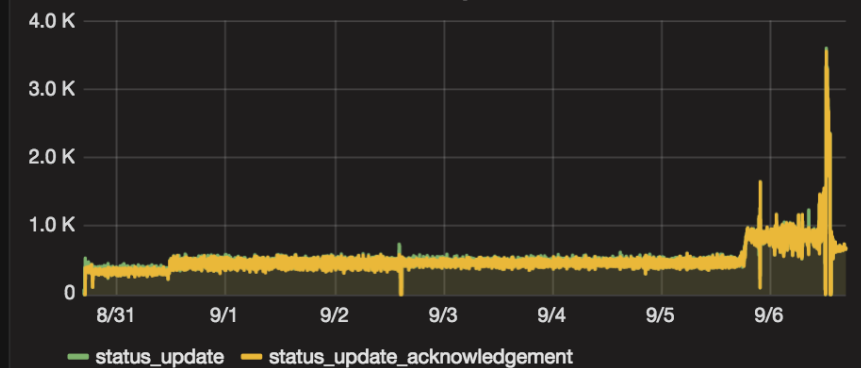
deactivate_framework



launch_tasks

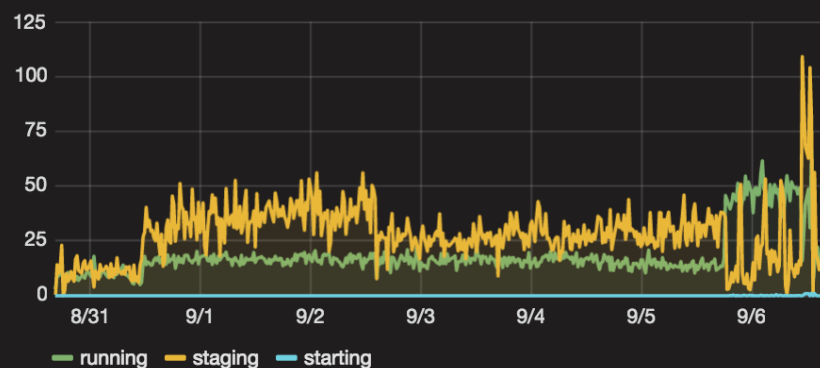


status_update

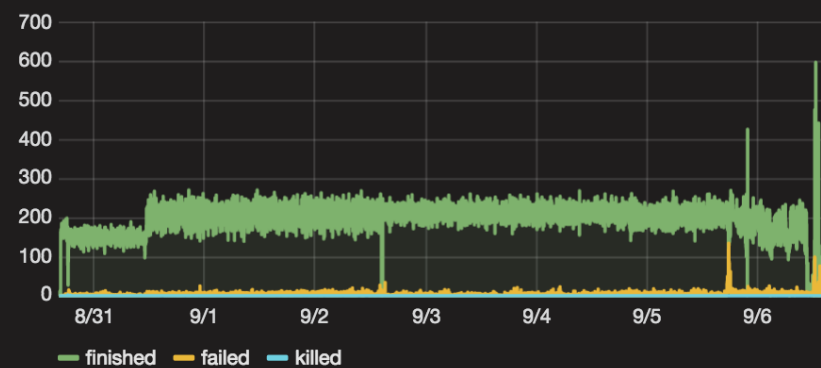


TASKS

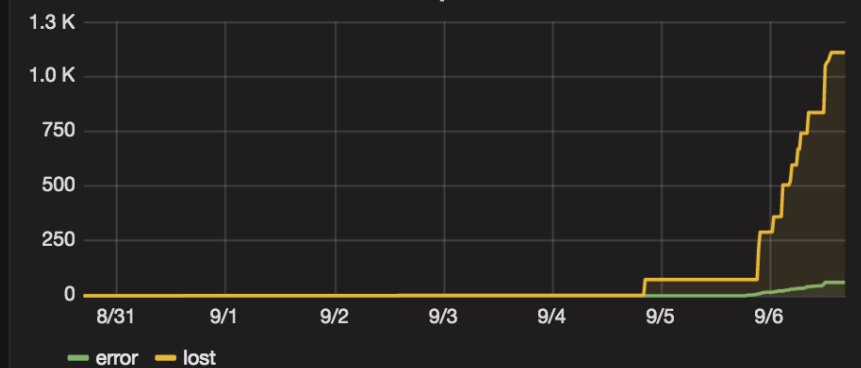
Status



Results



Exceptions



容器监控



系统研发部

appid: 100 39

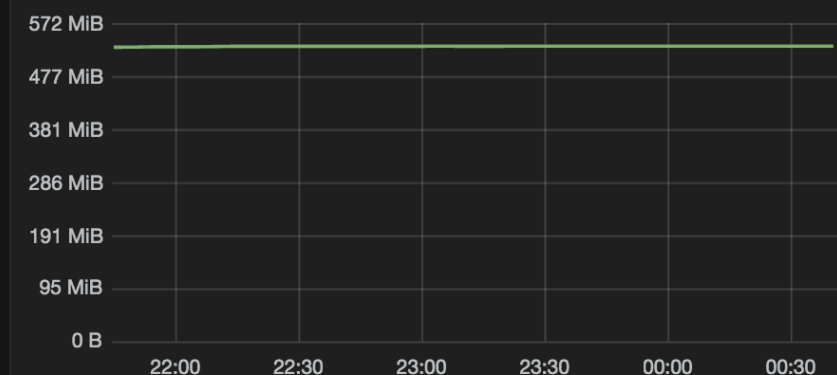
ip: 10_2 21

10_ 120

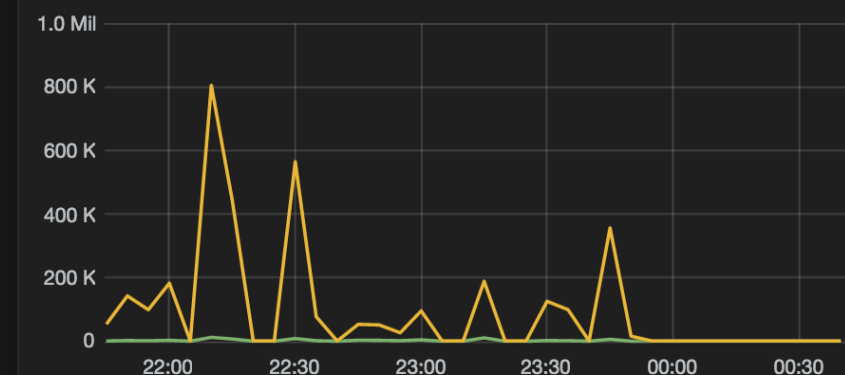
cpu



mem



net



10_ 121

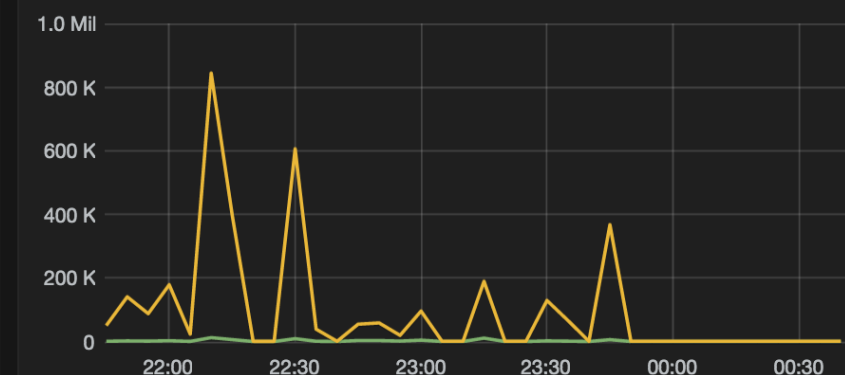
cpu



mem

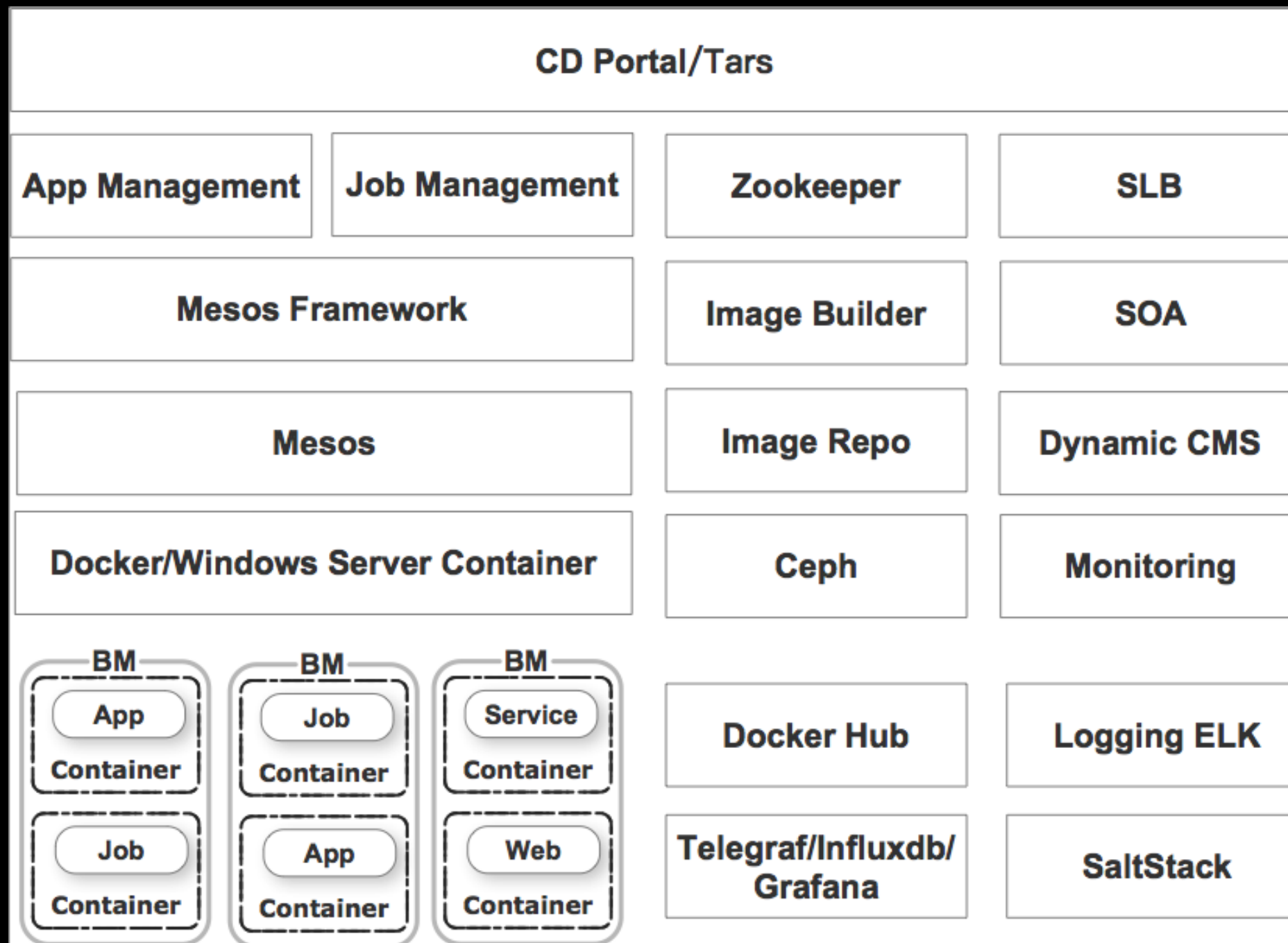


net



+ ADD ROW

CDOS整体架构



Q&A