CNUTCon 2017
全球运维技术大会

腾讯游戏容器云平台演进之路

尹烨
高级工程师

主办方 Geekbang 极客邦科技 InfoQ

Geekbang.
极客邦科技

极客搜索

全站干货，一键触达，只为技术

s.geekbang.org

扫描二维码立即体验

有没有一种搜索方式，能整合 InfoQ 中文站、极客邦科技旗下12大微信公众号矩阵的全部资源?
极客搜索，这款针对极客邦科技全站内容资源的轻量级搜索引擎，做到了!

扫描上方二维码，极客搜索!

EGO EXTRA GEEKS' ORGANIZATION NETWORKS

这里只有
技术领导者

EGO会员第二季招募季正式开启

报名时间：9月1日-9月15日

扫描添加E小欧，

邀您进入EGO会员预报名群

立即报名

E小欧

TABLE OF **CONTENTS**

平台概况

- 2014 – now
- 200+ APP、23W+ CPU core、800T+ Mem
- 业务场景
  - 轻量虚拟机
  - 微服务
  - 离线计算（大数据、机器学习）

技术栈

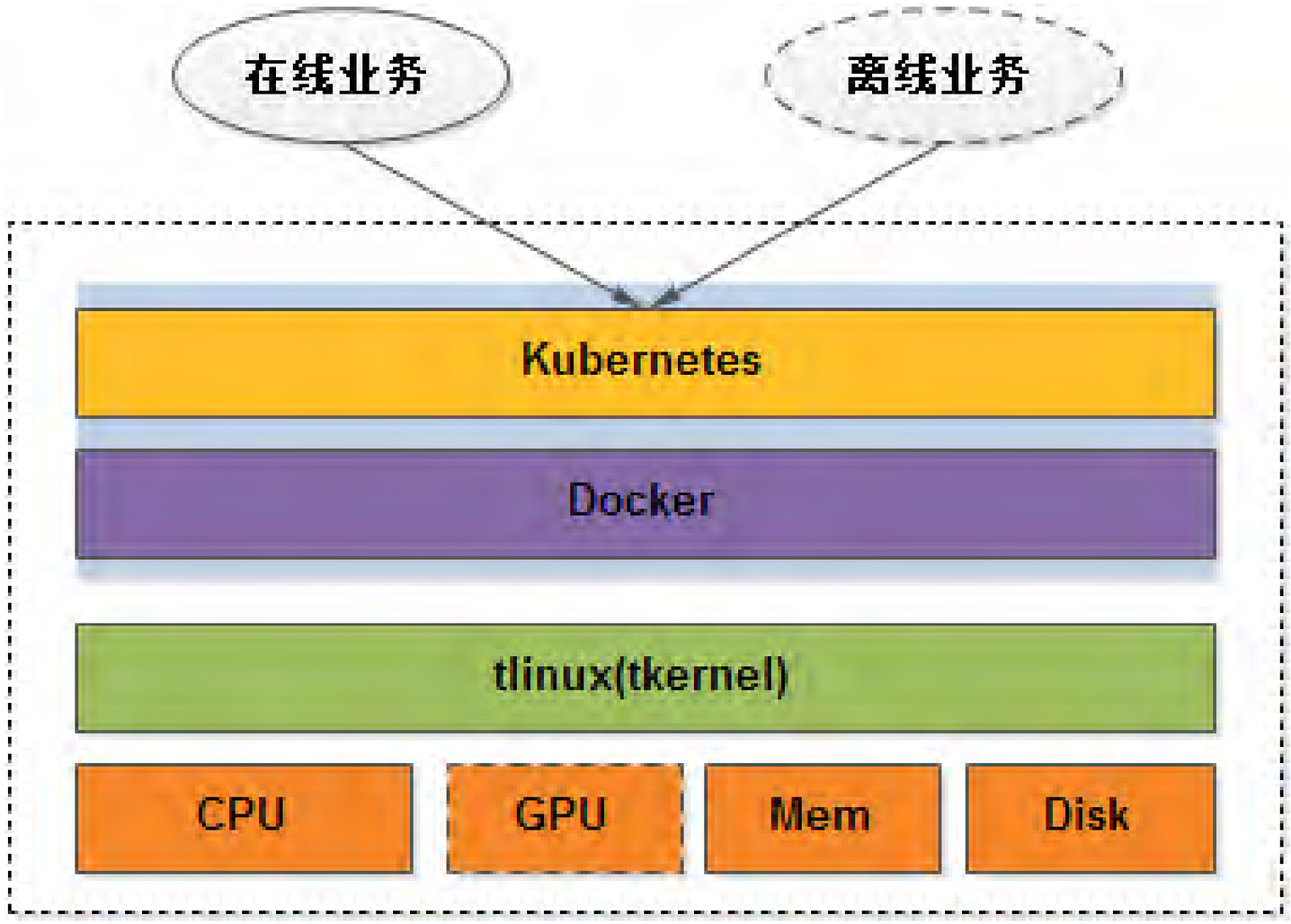

在线业务
离线业务
Kubernetes
Docker
tlinux(tkernel)
CPU
GPU
Mem
Disk

TABLE OF **CONTENTS**

轻量虚拟机

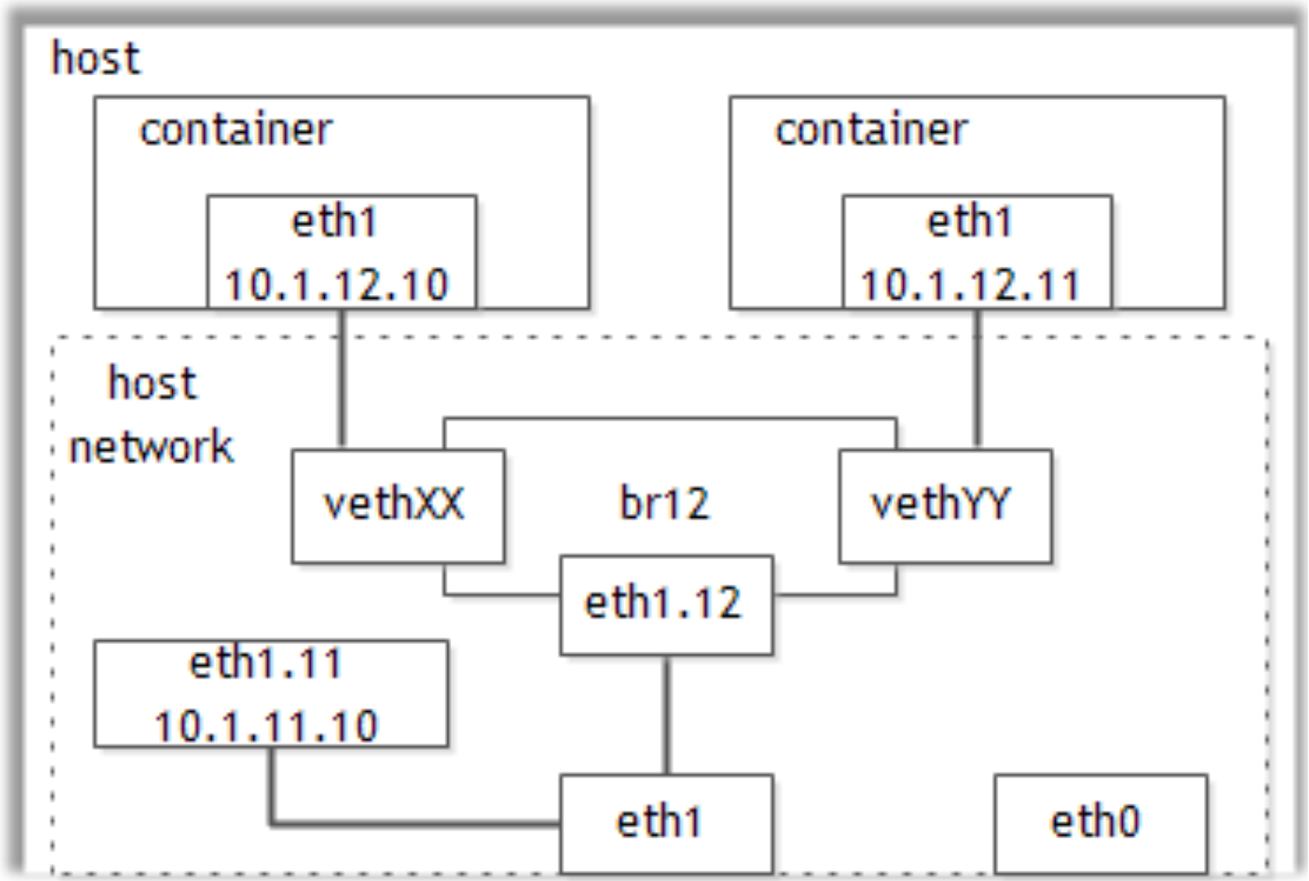
- System Init（ sysvinit /systemd ）+ SSH
- IP per light-VM
- Run monitor agent in light-VM

systemd

- Container Interface
  - container=docker
- Cgroup is needed
- udev is not available when mount /sys read-only
- Systemd defines that shutdown signal as SIGRTMIN+3
- ...

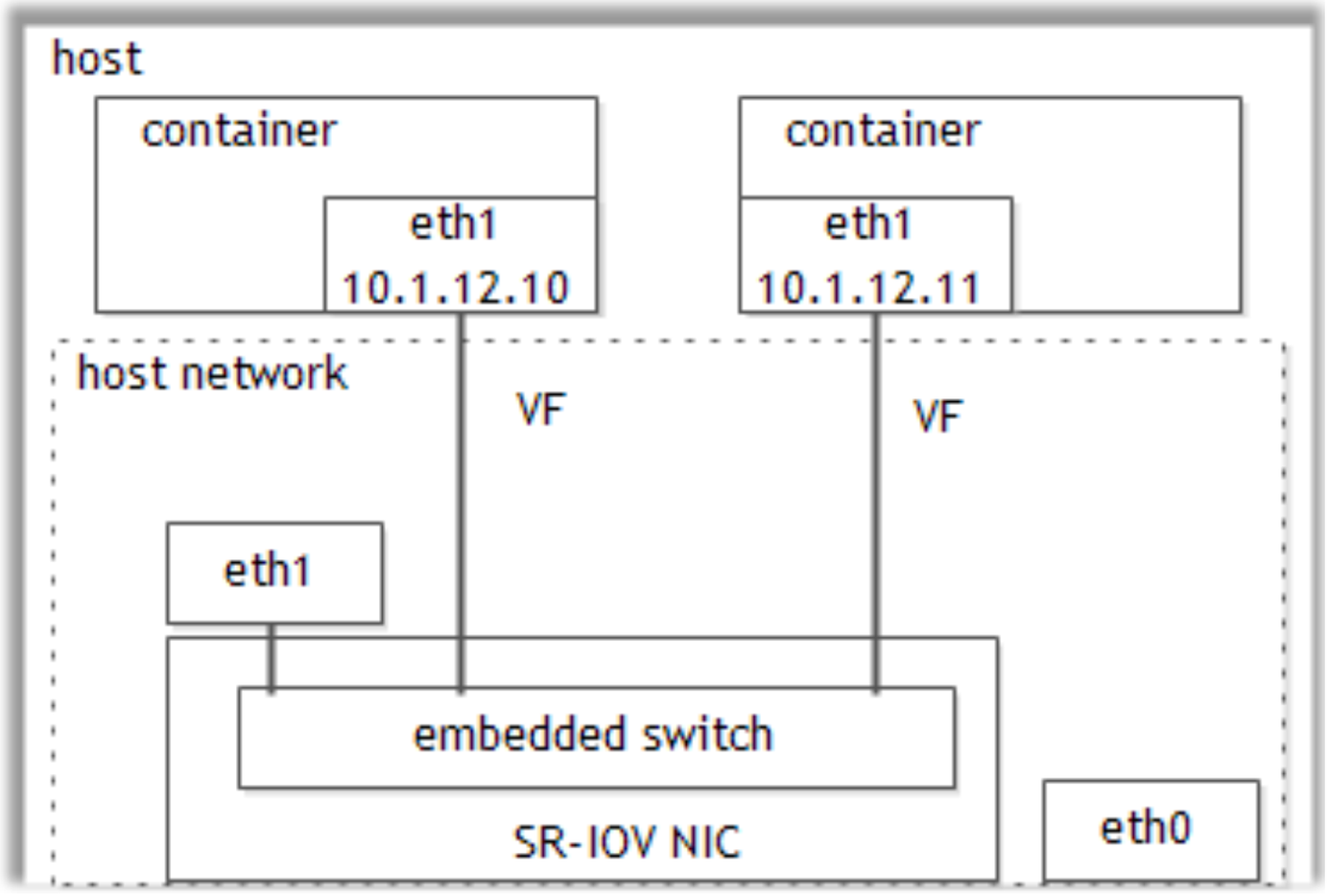
Network(1)

- Bridge
- Bad performance
- [Set veth txqlen=0]



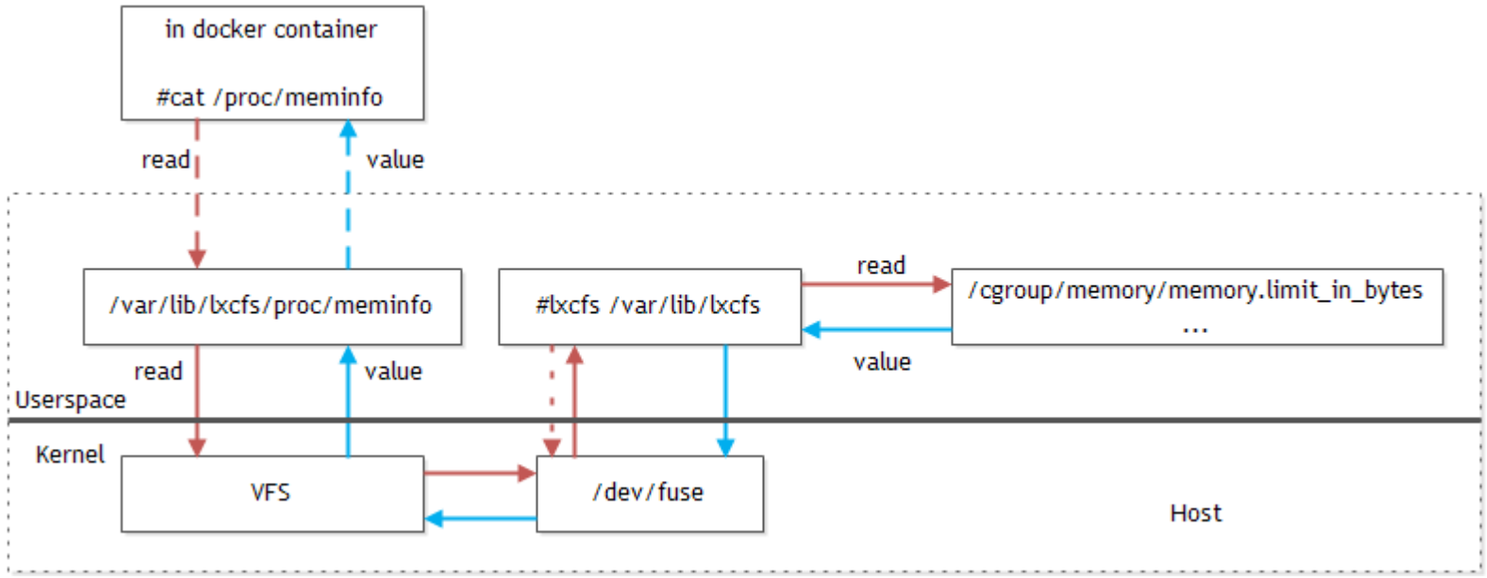
Network(2)

- SR-IOV
  - Good performance
  - Binding VF interrupt
  - Enable RPS



/proc

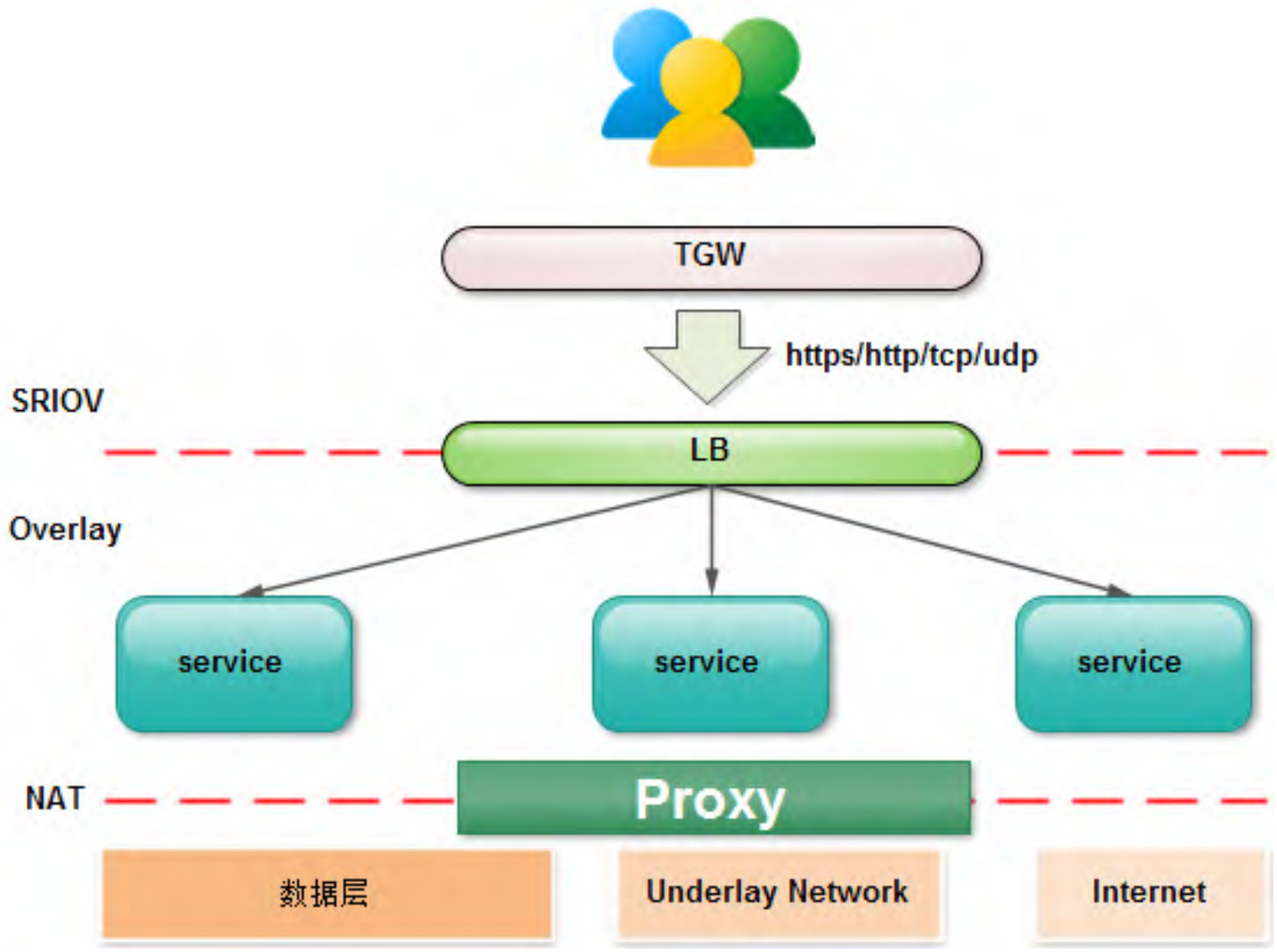
- Lxcfs
- Kernel support



微服务

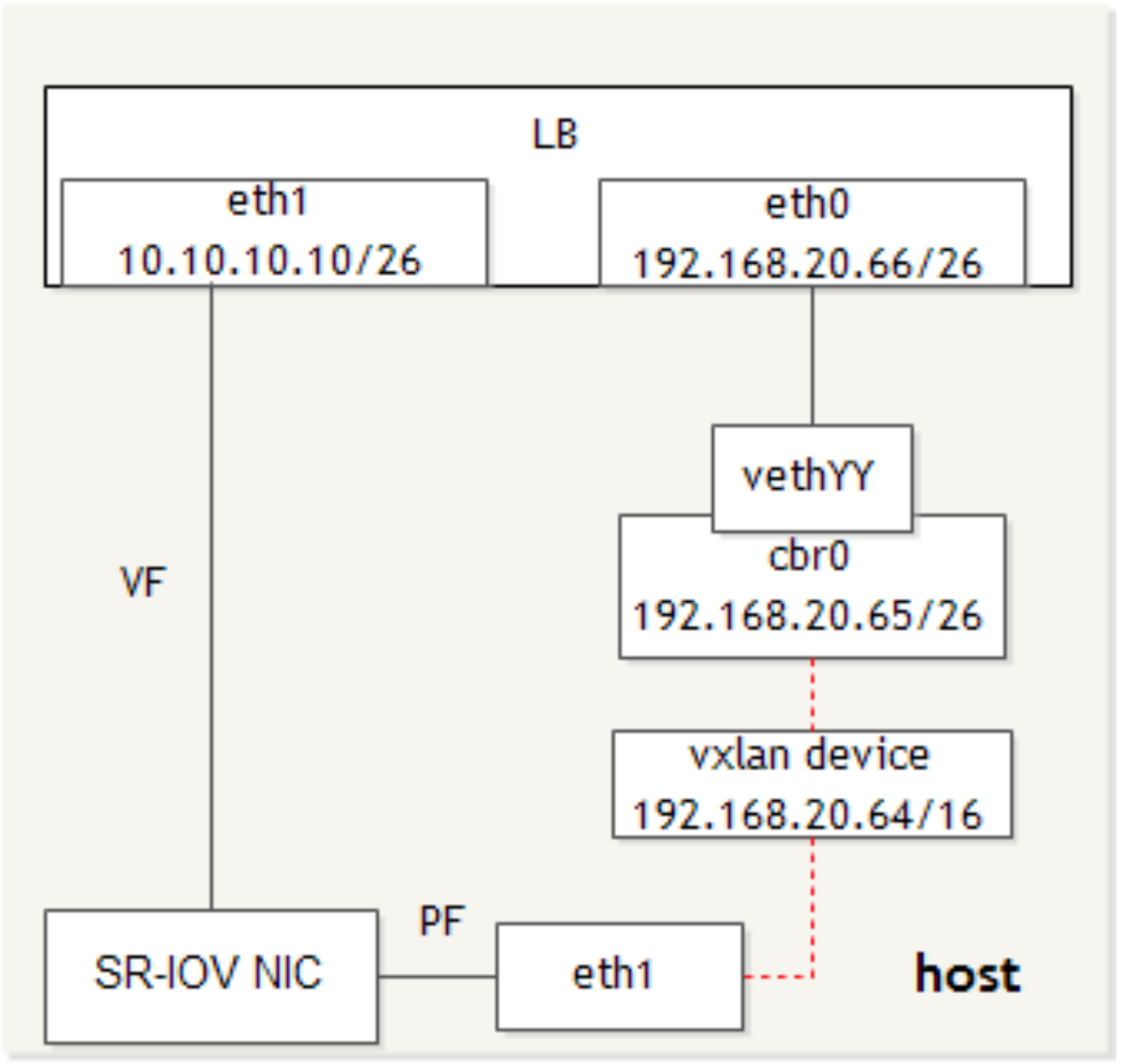
- Only app in container
- IP per container ?
- Monitor

Network - Overview



Underlay to overlay

- LB
  - http/https/tcp/udp



VXLAN optimization

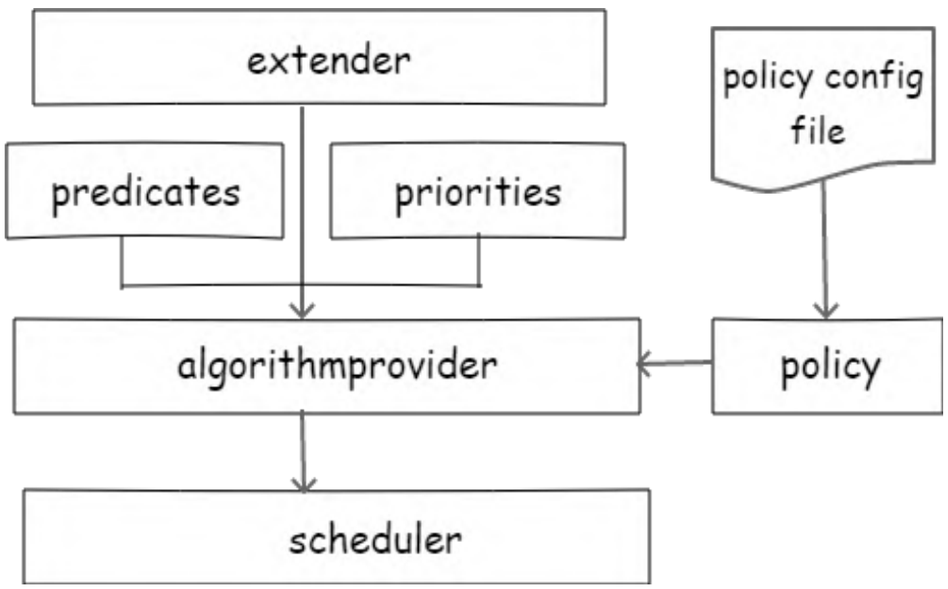
- UDP RSS
  - ethtool -N eth10 rx-flow-hash udp4 sdfn
- VXLAN offload
- VXLAN GRO
  - Kernel 3.14 (net: Add GRO support for vxlan traffic)

CNI

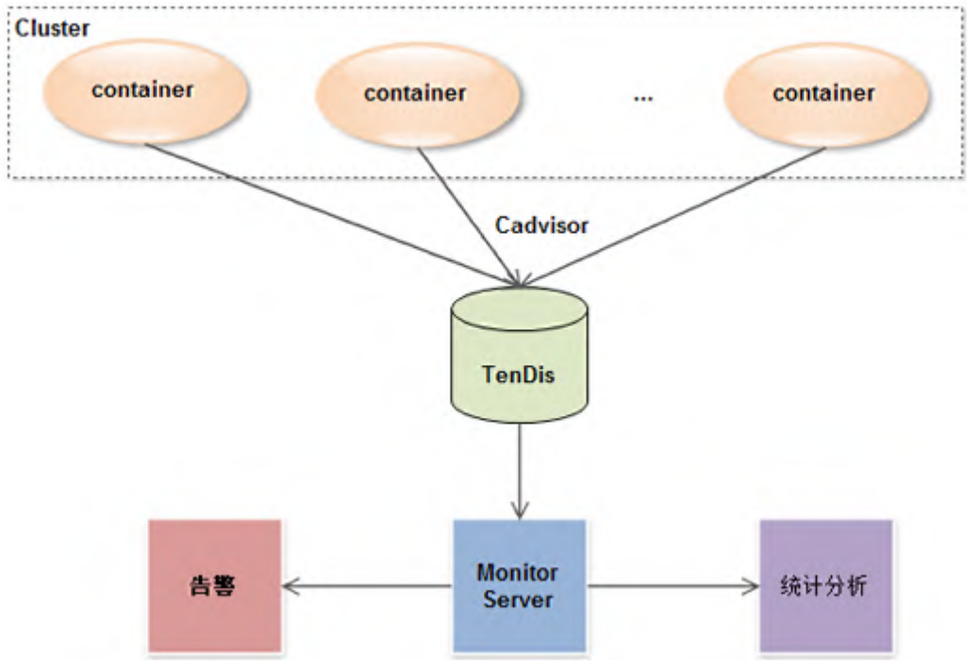
- Simple
- Plugins (macvlan , ipvlan , bridge , multus , …)
- Container runtimes (k8s , rkt , mesos , …)
- SR-IOV CNI (github.com/hustcat/sriov-cni)
  - High performance (NFV , Proxy , LB , …)
  - VF interrupt CPU binding
  - DPDK supported

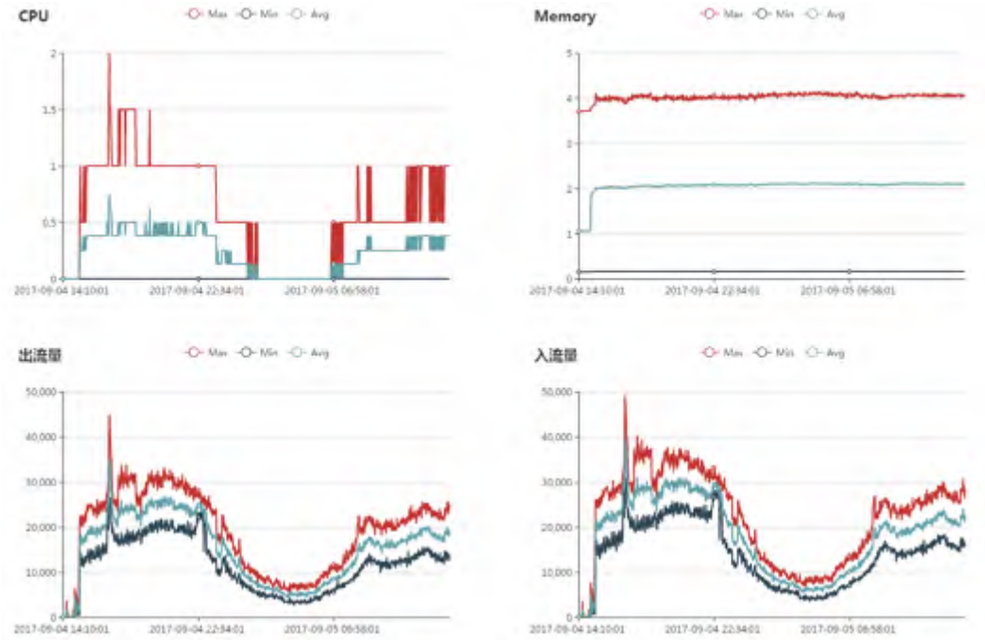
K8S extensions

- Scheduler plugin
- Cpuset and NUMA
  - kubernetes#49186 (v1.8?)



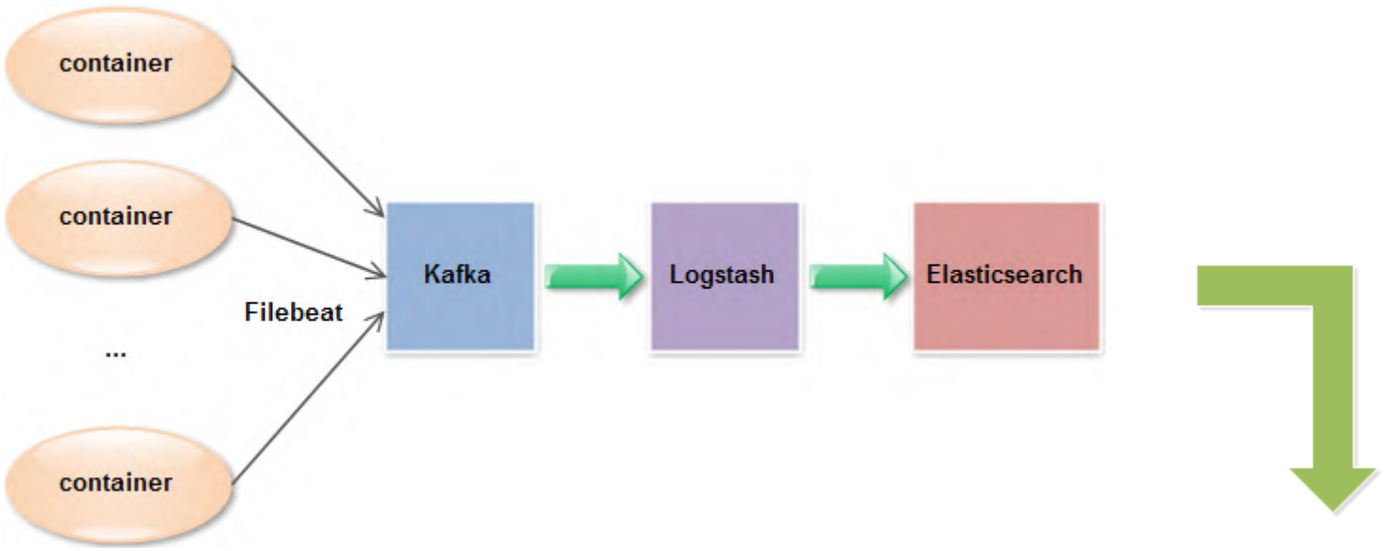
Monitor





服务	规则	通知人	告警方式
[redacted]	AVG(内存) > 50	[redacted]	RTX 邮件 微信
[redacted]	AVG(CPU) > 50	[redacted]	RTX 邮件 微信
[redacted]	AVG(内存) > 80	[redacted]	RTX 邮件 微信

Log



详细信息 端口映射 性能监控 事件 业务日志

选择Pod: All 日志类型: 标准输出 查询内容: 开始时间: 2017-09-03 15:47:16

结束时间: 2017-09-04 15:47:16 查询

Pod	详细信息	时间
[redacted]1592300993-8qmq4	read file ../conf/routetb_info.route, errno=2,No such file or directory	2017-09-04T14:36:09.667084367+08:00
[redacted]1592300993-8qmq4	[2017-09-04 14:36:09][BOOT][15][regist_config.cpp][249][main]	2017-09-04T14:36:09.667081361+08:00
[redacted]1592300993-8qmq4	read file ../conf/routetb_info.weight.route, errno=2,No such file or directory..ignore	2017-09-04T14:36:09.667078639+08:00
[redacted]1592300993-8qmq4	[2017-09-04 14:36:09][BOOT][15][regist_config.cpp][246][main]	2017-09-04T14:36:09.667075201+08:00

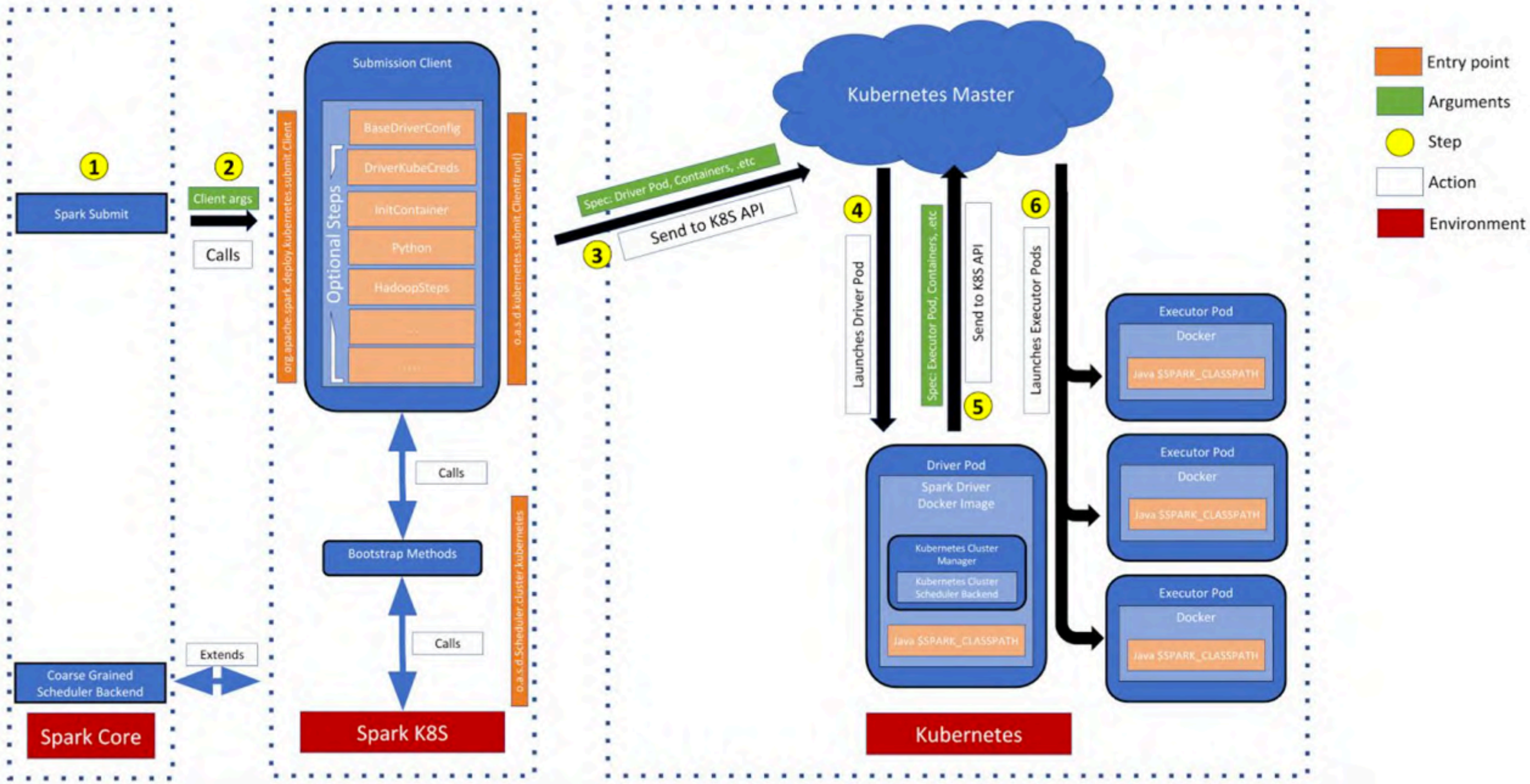
离线计算

- Tensorflow + GPU
  - NVIDIA/nvidia-docker（GPU device、CUDA library）
- Spark

Spark on K8S

- Native support for submitting Spark applications to a kubernetes cluster.
- The submitted application runs in a driver executing on a kubernetes pod, and executors lifecycles are also managed as pods.
- SPARK-18278
- https://github.com/apache-spark-on-k8s

Architecture



Comparison with Spark Standalone on K8S

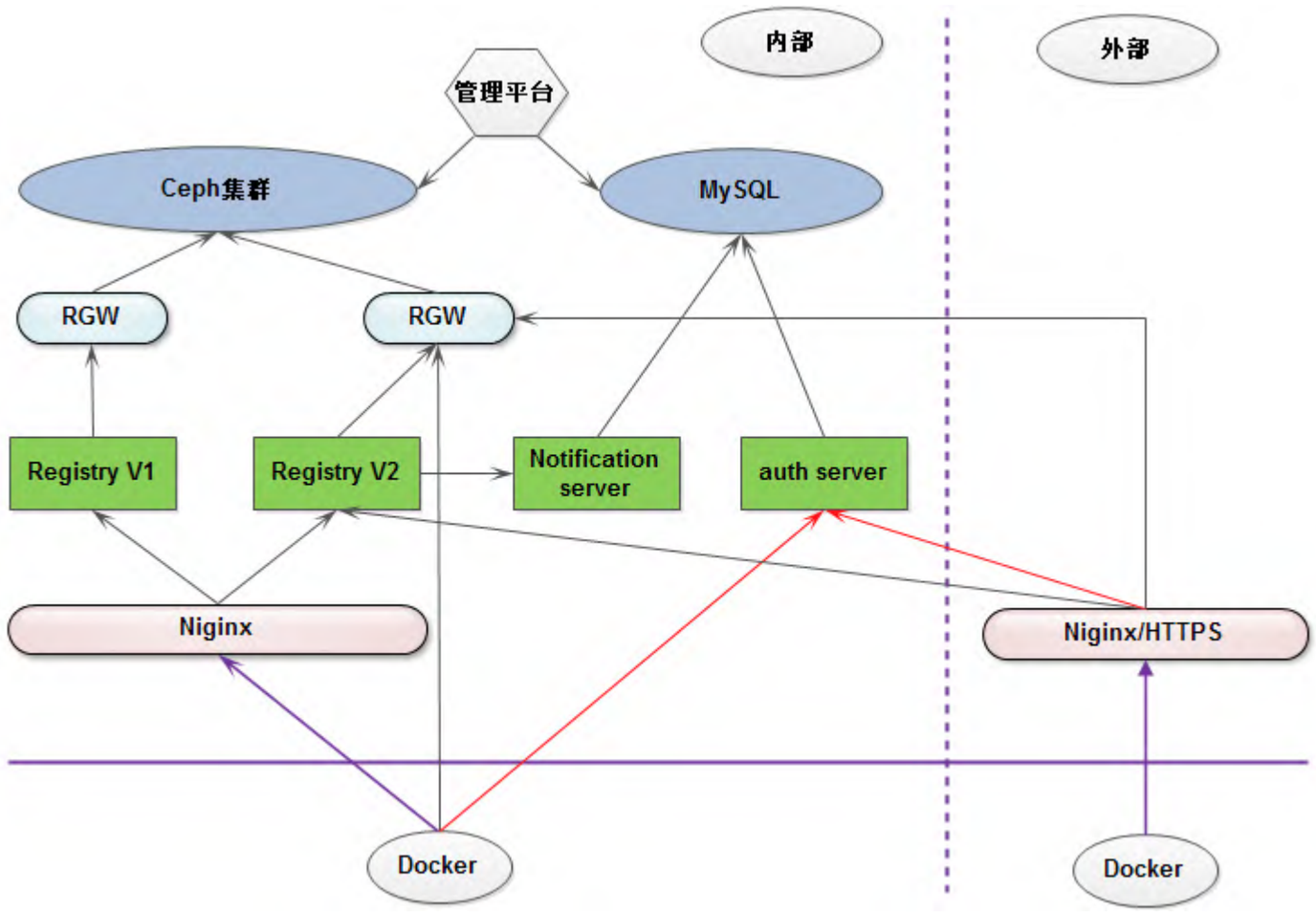
- Elastic
  - Spark executors can be elastic depending on job demands
- Simple
  - Simplifies the process of running Spark jobs
- Efficient
  - Only k8s-based resource scheduler

镜像传输

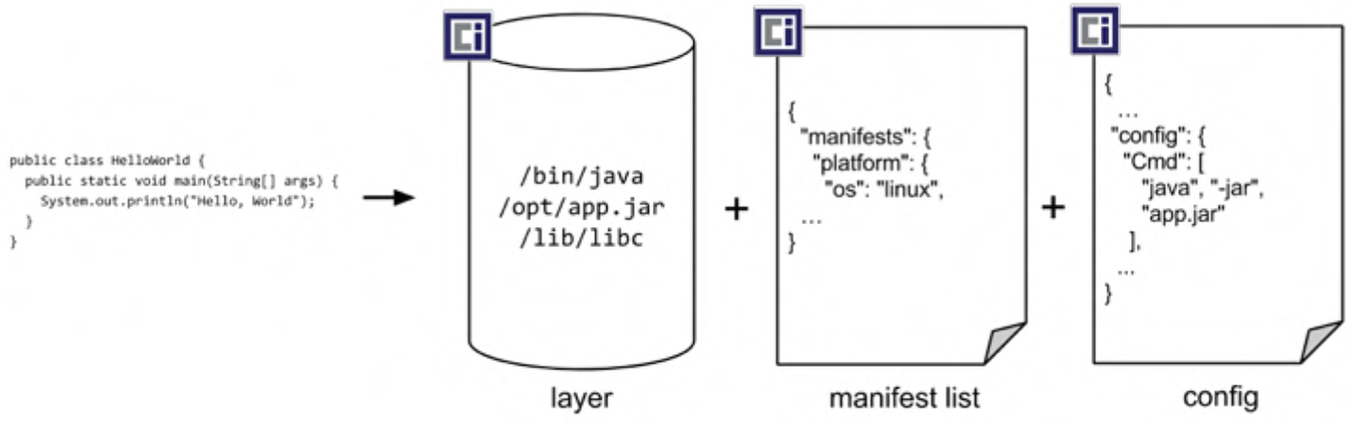
- 自研企业级镜像仓库
- P2P传输

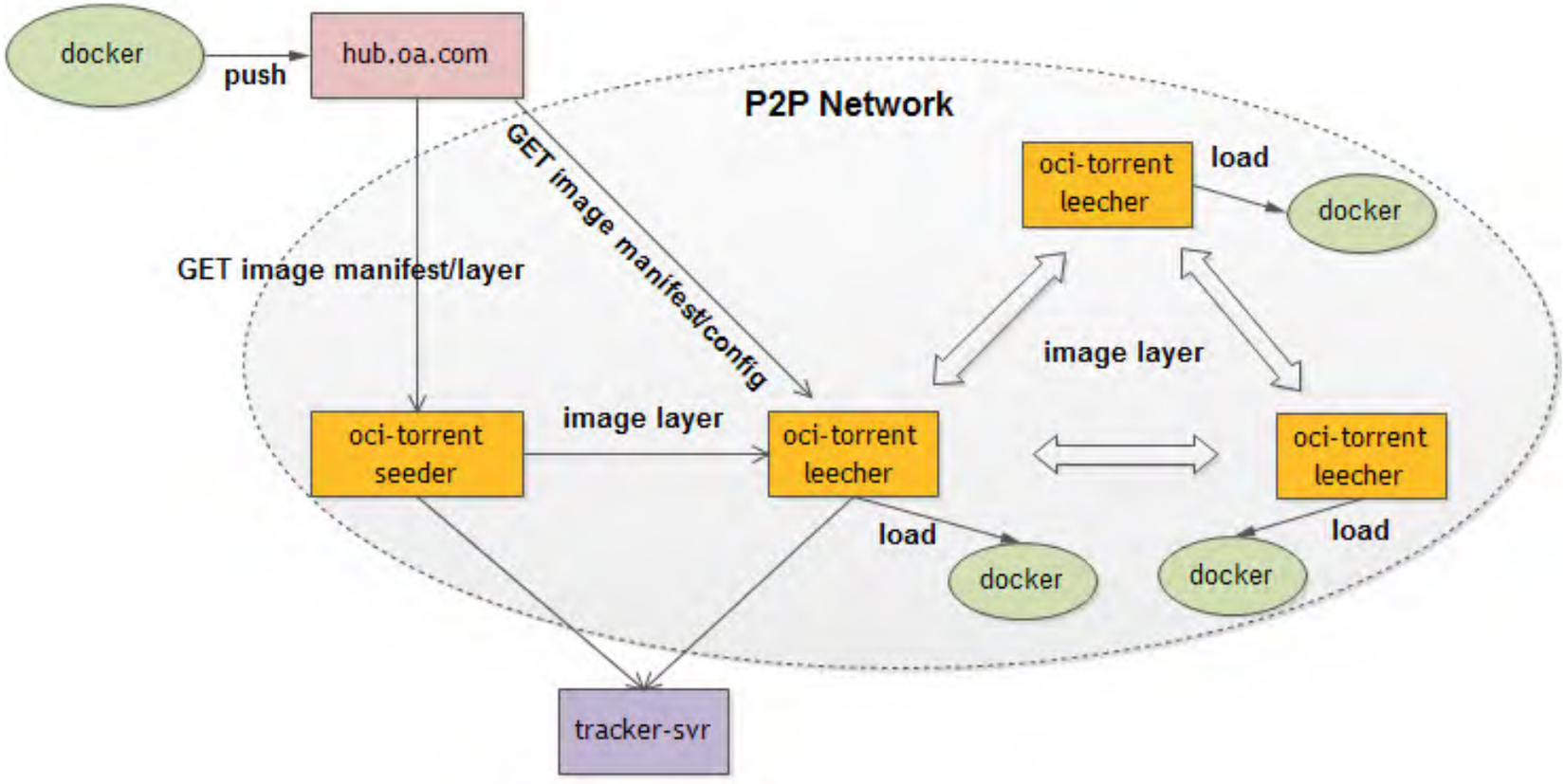
镜像仓库

- Token认证
- 权限控制
- 操作日志及审计
- 分布式存储



P2P镜像传输


public class HelloWorld {
public static void main(String[] args) {
System.out.println("Hello, World");
}
}
/bin/java
/opt/app.jar
/lib/libc
{
"manifests": {
"platform": {
"os": "linux",
...
}
{
...
"config": {
"Cmd": [
"java", "-jar",
"app.jar"
],
...
}
layer
manifest list
config


docker
push
hub.oa.com
P2P Network
GET image manifest/config
GET image manifest/layer
oci-torrent leecher
load
docker
image layer
oci-torrent seeder
image layer
oci-torrent leecher
oci-torrent leecher
load
load
docker
docker
tracker-svr

Kernel

- Overlayfs + XFS
- Buffer IO throttle
- Cgroup namespace
- 网络sysctl内核参数隔离
- Bugfix

Overlayfs + XFS

- Advantage
  - Simple
  - Good IO performance
    - XFS（project quota，inode limit）
- Some problems
  - Inotify（#11705）
  - Unix socket（#12080，Kernel 4.7）

TABLE OF **CONTENTS**

总结

- 容器重新定义业务部署和资源交付方式

THANKS!

智 能 时 代 的 新 运 维

CNUTCon 2017