



GOPS 2016
Shanghai



全球运维大会

2016

重新定义运维

上海站

会议时间： 9月23日-9月24日

会议地点： 上海·雅悦新天地大酒店

主办单位：



指导单位：



多快好省运营大规模弹性云集群

韩建飞 京东商城



目录



1

弹性云v1.0线上运营情况

2

弹性云v1.0挑战

3

弹性云v1.0架构

4

弹性云v1.0 大促备战

5

弹性云v1.0容器性能提升

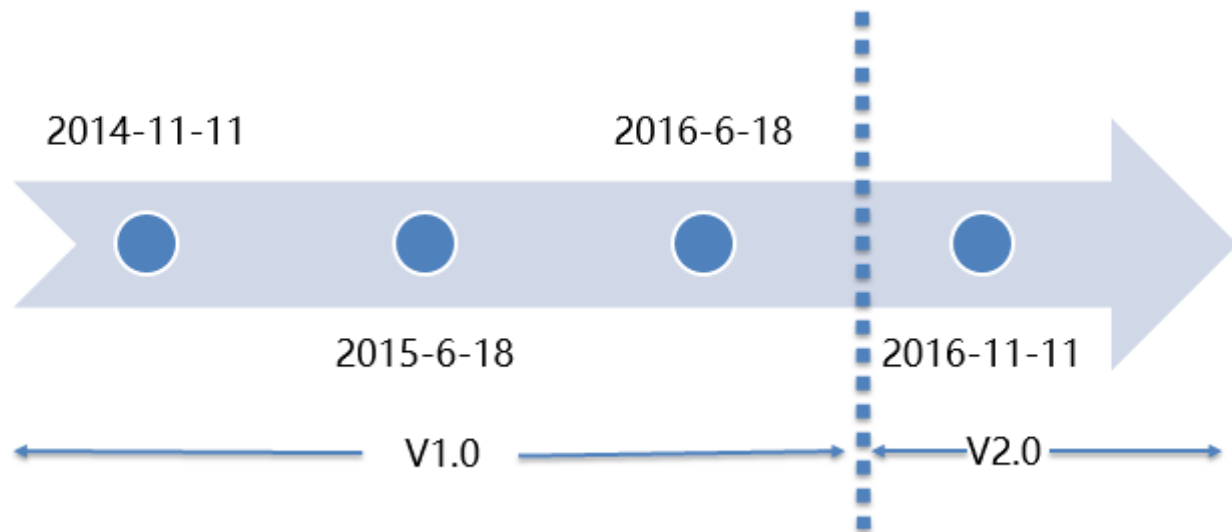
6

弹性云v2.0起航



弹性云v1.0 线上运营情况

- 618大促前统计15W+，JD全部业务均run在弹性云容器之上



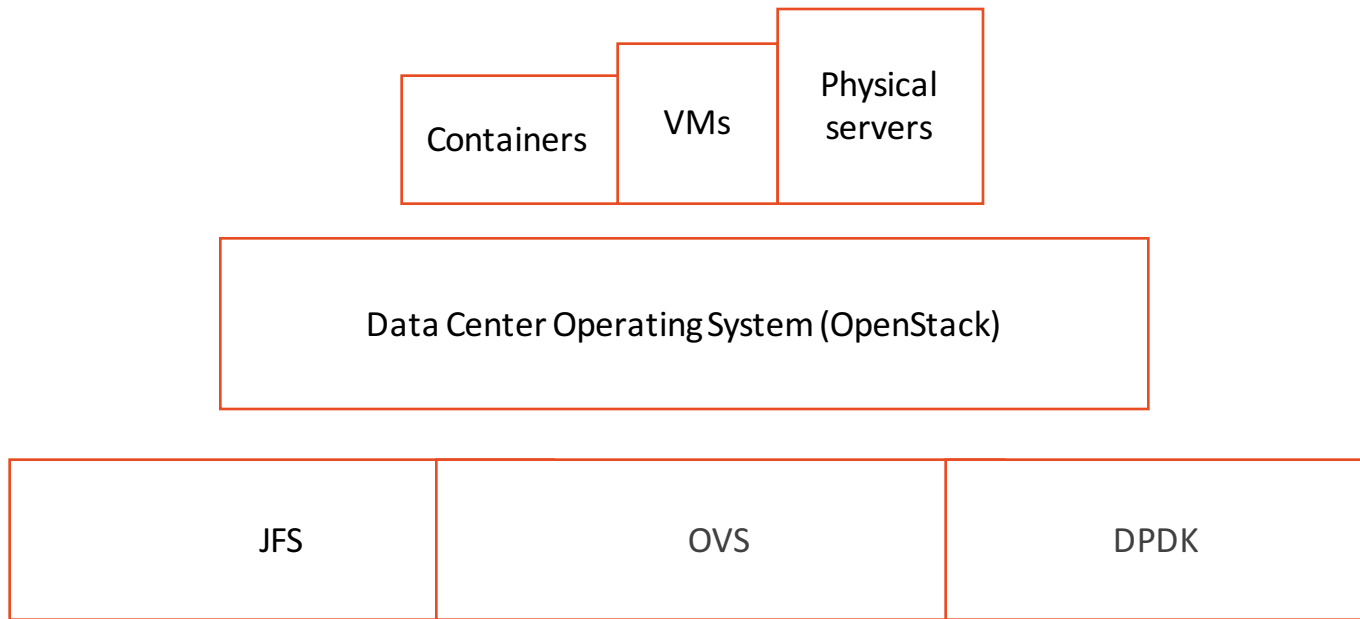
挑战

- 选择 (2014/6/18)
- 向前兼容
- 第一个核心系统 (单品页)
- 容器性能&稳定性
- 大规模
- 运维成本

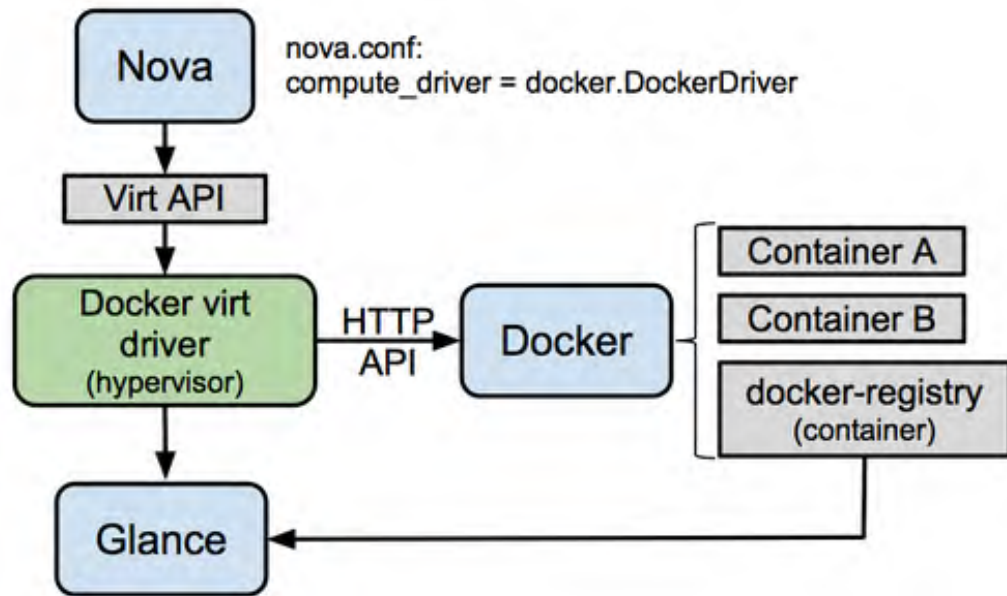


弹性云v1.0架构

Openstack Icehouse + docker1.3 + OVS2.1.3/2.3.2+Centos6.6

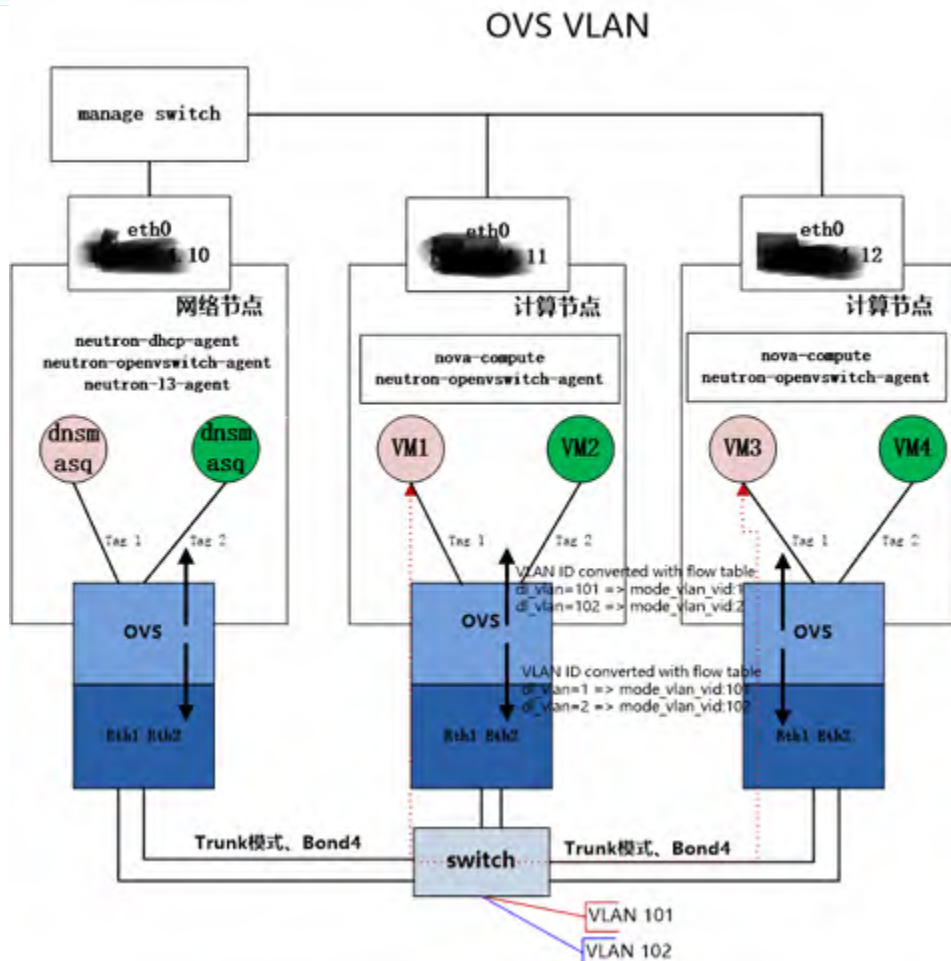


弹性云v1.0架构



弹性云v1.0架构

OVS-VLAN网络模式



弹性云v1.0 业务方研发同事使用

- 对接原有上线系统
- 自助上线

弹性计算平台															团队	使用手册	
申请编号	申请编号	所属应用	所属应用	申请人	申请人	申请类型	全部	申请状态	待审核	Q搜索	工单接入						
申请号	应用	部署环境	镜像	规格	机房	ZONE	数量	描述	申请人	申请时间	已有数	可用数	状态	应答	操作		
18474		线上正式环境	Worker镜像(JD...	4核/8G内存/50G磁盘			5			2015-11-30 14:08:55	7		待审核		操作		
18469		线上正式环境	ECC镜像	16核/64G内存/50G磁盘			1			2015-11-30 10:46:04	26		待审核		操作		
18464		线上正式环境	Web镜像(Tomc...	4核/8G内存/50G磁盘			2			2015-11-27 16:37:40	3		待审核		操作		
18461		线上正式环境	Web镜像(Tomc...	4核/8G内存/50G磁盘			3			2015-11-27 16:30:42	4		待审核		操作		
18457		线上正式环境	Web镜像(Tomc...	8核/16G内存/50G磁盘			4			2015-11-27 15:58:04	10		待审核		操作		
18460		线上正式环境	Web镜像(Tomc...	4核/8G内存/50G磁盘			4			2015-11-27 16:12:21	3		待审核		操作		
18456		线上正式环境	Web镜像(Tomc...	4核/8G内存/50G磁盘			1			2015-11-27 15:21:29	63		待审核		操作		
18455		线上正式环境	Web镜像(Tomc...	4核/8G内存/50G磁盘			4			2015-11-27 15:12:23	0		待审核		操作		
18454		线上正式环境	Worker镜像(JD...	4核/16G内存/50G磁盘			4			2015-11-27 15:10:59	0		待审核		操作		
18453		线上正式环境	Worker镜像(JD...	4核/8G内存/50G磁盘			4			2015-11-27 15:09:36	0		待审核		操作		

弹性云v1.0 运维同事运维系统

- 物理资源管理
- 集群管理
- 异常消息追踪
- 每日巡检

云数据库 计算资源管理 - 消息跟踪列表

request id: request id vm uuid: vm uuid 搜索 【查看详情: 输入消息id或实例id】

timestamp	request id	publisher	event	priority	instance id	reason
2015-11-30 13:11:18	req-0e10b394-2733-4057-6636-642ae0e4e0d	scheduler	scheduler_run_instance	ERROR	7a77e0d4-8d2f-4709-a072-a103f900161	reason



今天 (6 封)

jdossinspector@jd.com	32分钟前
云数据库JDOS巡检邮件	
jdossinspector@jd.com	32分钟前
1区JDOS巡检邮件	
jdossinspector@jd.com	33分钟前
云数据库JDOS巡检邮件	
jdossinspector@jd.com	33分钟前
1区JDOS巡检邮件	
jdossinspector@jd.com	37分钟前
1区JDOS巡检邮件	
jdossinspector@jd.com	38分钟前
1区JDOS巡检邮件	

618备战

- 资源
- 20倍压测
- 弹性 **scale up** & scale out (不是银弹)
- 调优
- 内存即存储
- 区别业务特点
- 细到尘埃的监控系统
- 高峰流量 提前规划资源+弹性scale up



坚如磐石

- 稳定性
- 性能

```

Last login: Tue May 10 09:52:22 2016 from 10.1.1.1
*****
*                                     *
*                !!!!!!!NOTIFICATION!!!!!!                *
*   This is a JD container environment server. Some hardware-related tools   *
*   can not be execute in this environment.Such as hpacucli,lsiutil,hdparm,ipmitool *
*   reboot,megacli,etc!!! *
*   Also the main proc "sleep 9999999d" can not be killed while running!!! *
*   Thanks for your cooperation! *
*                                     *
*                                     JDOS TEAM *
*****

```



硬件环境

- 服务器节能模式的“坑”
- 容器实例密度对local机械磁盘的挑战
- 不同品牌网卡的困扰
- CPU:Memory = 1: ?
- 计量性能的规则和手段
- 不同cpu硬件性能带给业务方困扰



万兆NIC

- 多CPU分摊网卡软中断（网卡队列均衡以及中断散列）
- OVS-VLAN , simple flow , no offload , 性能与物理网卡相当
- kernel 2.6.32的网络namespace对sync报文承受弱 , kernel 3.1 fixed
- 优化OVS对内核锁的使用
- Monitor more



Cgroup : CPU

- CFS (RR/FIFO)
- 流量与CPU%成线性关系吗
- 某些容器Load影响其他容器性能吗
- cpu set & cpu share & 逃逸



Cgroup : Memory

- $\% = \text{RSS} + \text{Share} / \text{Total}$
- OOM
- Cache
- Slab JD分散回收时机及回收阈值设定
- Swap& Swappiness
- Monitor more , 如swap



Cgroup：资源及隔离

- 用户隔离（uid/线程数）
- 文件句柄限制（file-max）
- 虚拟内存（线程数及overcommit）



大集群

- openstack单集群6K台计算节点

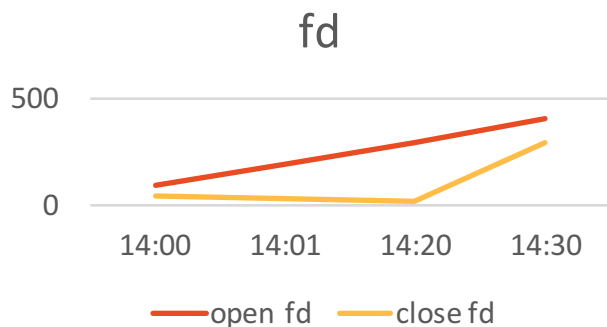
- ❑ 自研brooder服务，代替社区直连和目前的nova-conductor，不依赖MQ，带redis缓存
- ❑ 管理容器生命周期消息tracking，集群任何动作均做tracking，告警
- ❑ Neutron-openvswitch-agent update 平滑，确保容器不中断，稳定
- ❑ 控制节点服务，计算节点服务，依赖DB，内核日志均做服务进程探活监控
- ❑ 避免使用ceilometer采集simple指标走MQ

- 自研调度系统

- ❑ 自研调度系统，依赖计算节点资源状态信息存放在redis，确保无状态调度进程原子互斥调度
- ❑ 根据应用app-id，夸交换机/机架/物理机；根据业务系统级别，资源特性调度到不同的zone
- ❑ 不同的隔离特点，调度到不同的network zone
- ❑ 某个容器失效，自动触发RC（保持IP不变“迁移”）
- ❑ 应用缩容/扩容，自动CRUD操作LB，DB授权，服务框架注册



细到尘埃的监控平台获得技术尊重

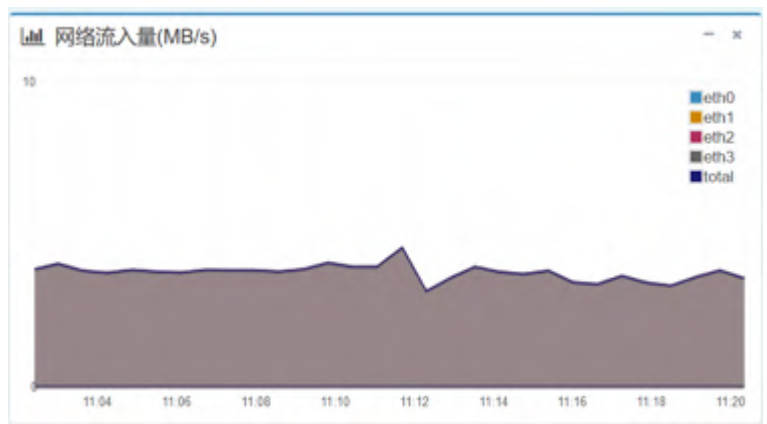
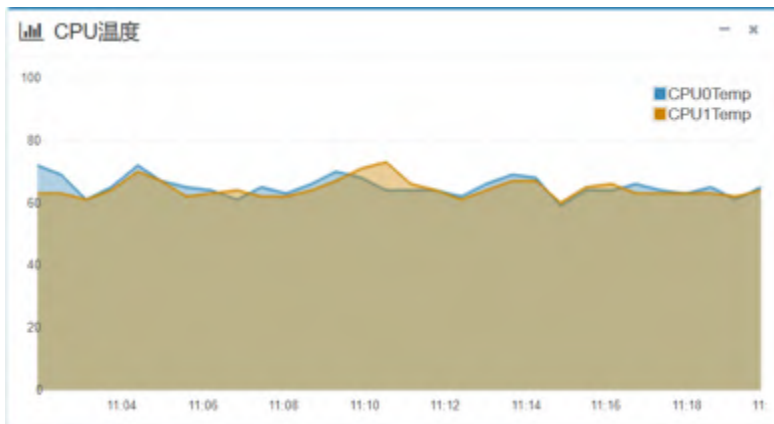
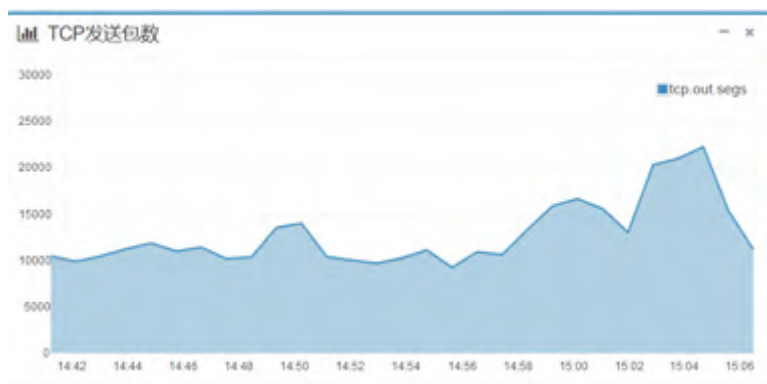
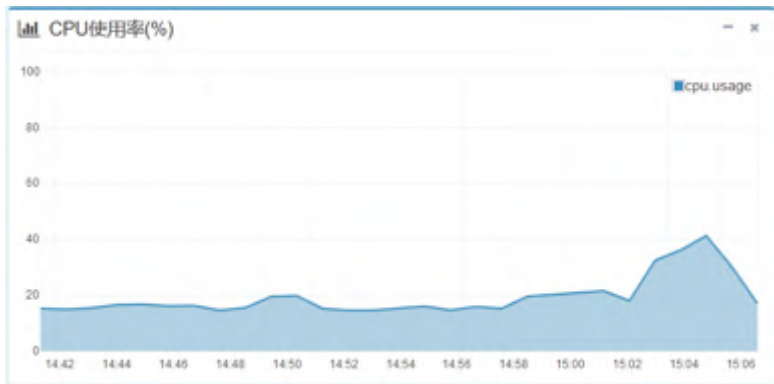


监控

- 资源监控：
CPU，负载，内存，SWAP，网络出入项流量、报文个数，网卡错误统计，磁盘使用率、读写流量、读写iops、繁忙程度等。
- 网络协议监控：
tcp 丢包、重传、收发报文，链接数等。
- 其他监控：
内存申请失败，进程/线程数，句柄数等。
- 业务监控：定制的部分业务监控。
- 服务器硬件监控：功率，风扇转速，温度等监控。

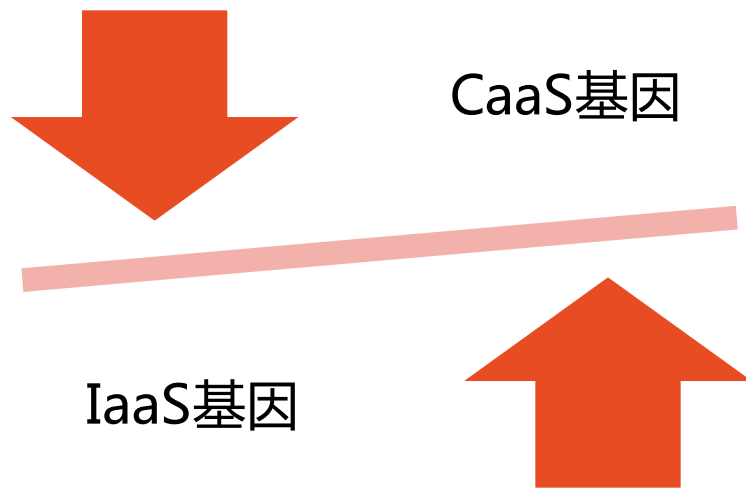


监控



未来

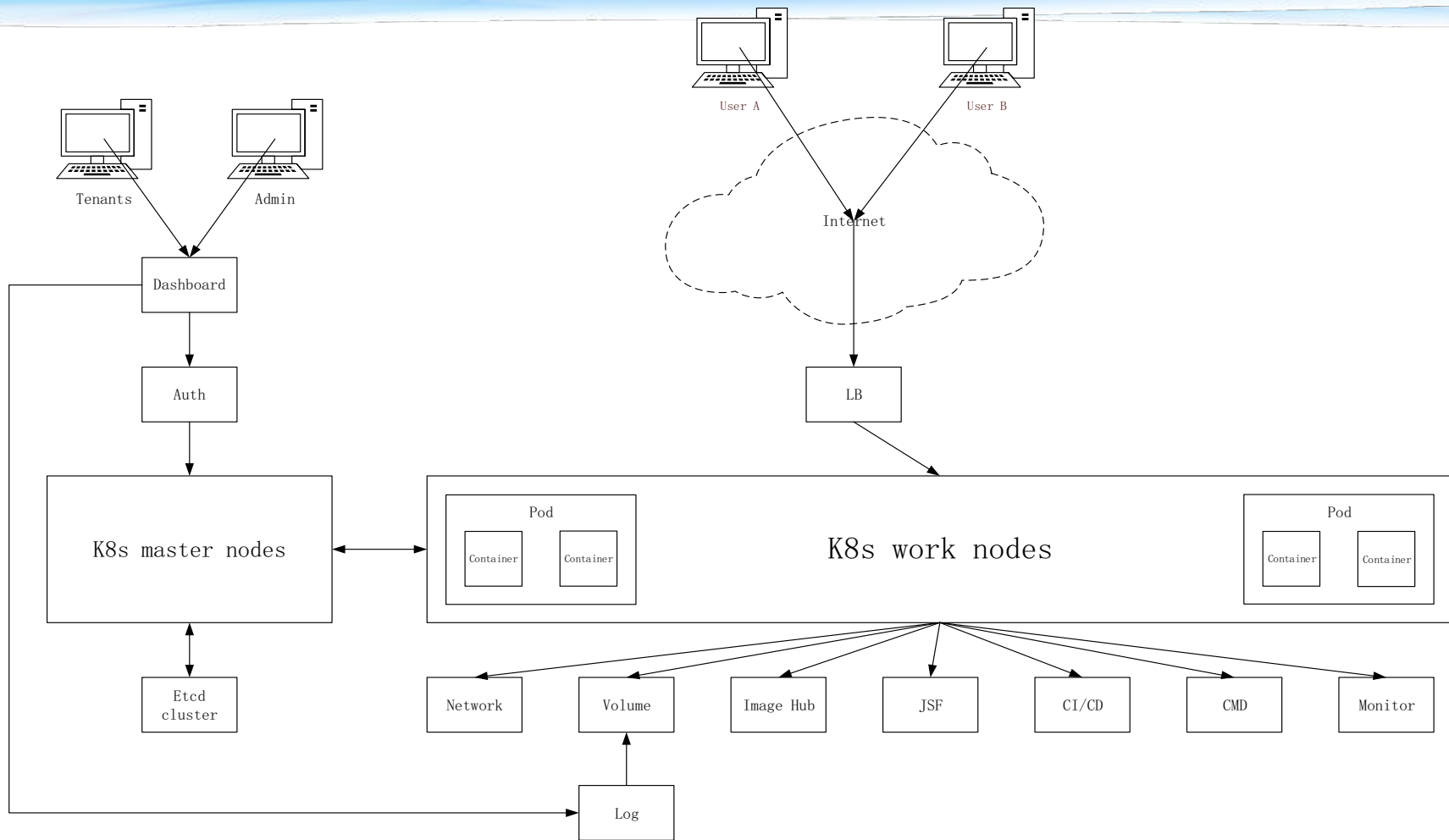
- 弹性云v1.0 所得所失
- 弹性云v2.0 承上启下



弹性云v2.0

K8s + Docker + Neutron + OVS + DPDK + JFS







Thanks

高效运维社区
开放运维联盟

荣誉出品





想第一时间看到高效运维公众号的好文章么？

请打开高效运维公众号，点击右上角小人，并如右侧所示设置即可：



GOPS2016 全球运维大会更多精彩

GOPS2016 全球运维大会·北京站

2016年12月16日-17日
北京国际会议中心

