Brought by Geekbang. 极客邦科技 InfoQ

CNUT
Connect Container Community

全球容器技术大会

剖析容器企业实践 关注容器生态圈开源项目

腾讯游戏的Docker实践

—现状、经验及展望

尹烨 @Tencent

目录

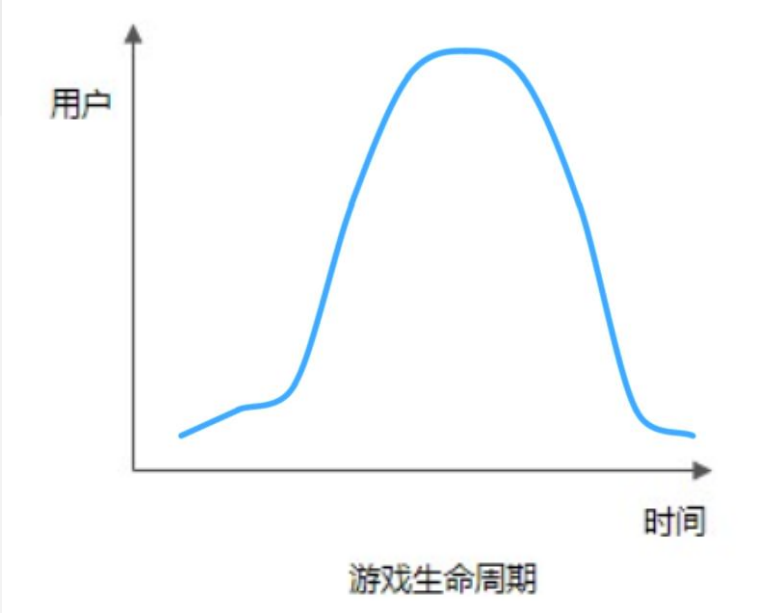
- 腾讯游戏应用现状
- 问题及解决方案
- 总结及展望

目录

- 腾讯游戏应用现状
- 问题及解决方案
- 总结及展望

从游戏业务说起

- 业务周期
- 种类
  - 端游、手游、页游
  - 自研、代理
  - 分区分服、全区全服



Why docker?

- Like VM
- Not only just like VM

Docker vs VM

- 优势
  - 轻量、Image、APP centric、在线调整资源配额
- 劣势
  - 安全、隔离性

现状

- 2014.6 ~ Now
- 业务接入
  - 1000+物理机、~4000个Container
  - 数十个端游、手游、页游
    - 《我叫MT2》、《QQ宠物企鹅》…

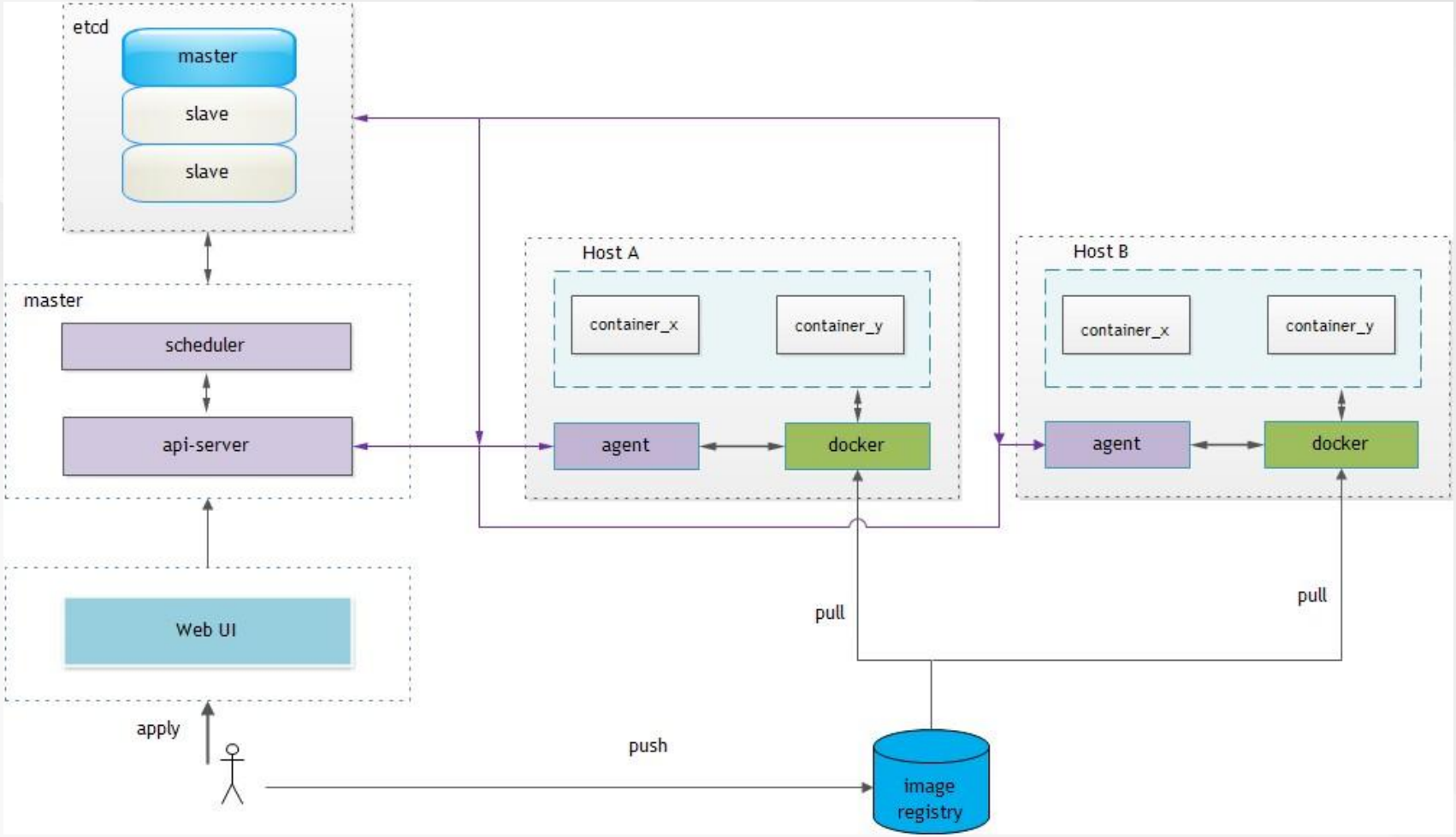
目录

- 腾讯游戏应用现状
- 问题及解决方案
- 总结及展望

问题1—容器集群调度(1)

- Fig？Shipyard？…
- Kubenetes
  - CPU core、 机器类型、内存、磁盘、网络…
  - Black list/White list
  - Specific host
  - Affinity/No Affinity

问题1—容器集群调度(2)

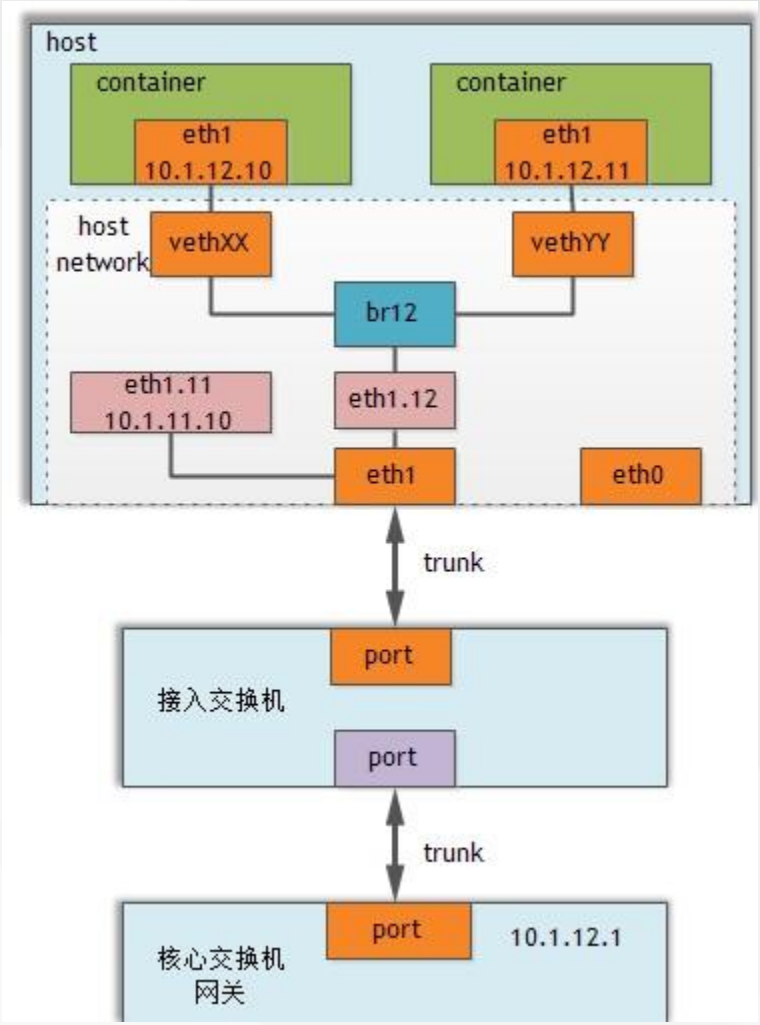

etcd
master
slave
slave
master
scheduler
api-server
Web UI
apply
Host A
container_x
container_y
agent
docker
Host B
container_x
container_y
agent
docker
pull
pull
push
image registry

问题2—Network(1)

- NAT?
  - Performance poor、IP hidden
- Host?
  - No network isolation
- Overlay network?
  - Complex、performance
  - Communicate with physical/virtual machine?

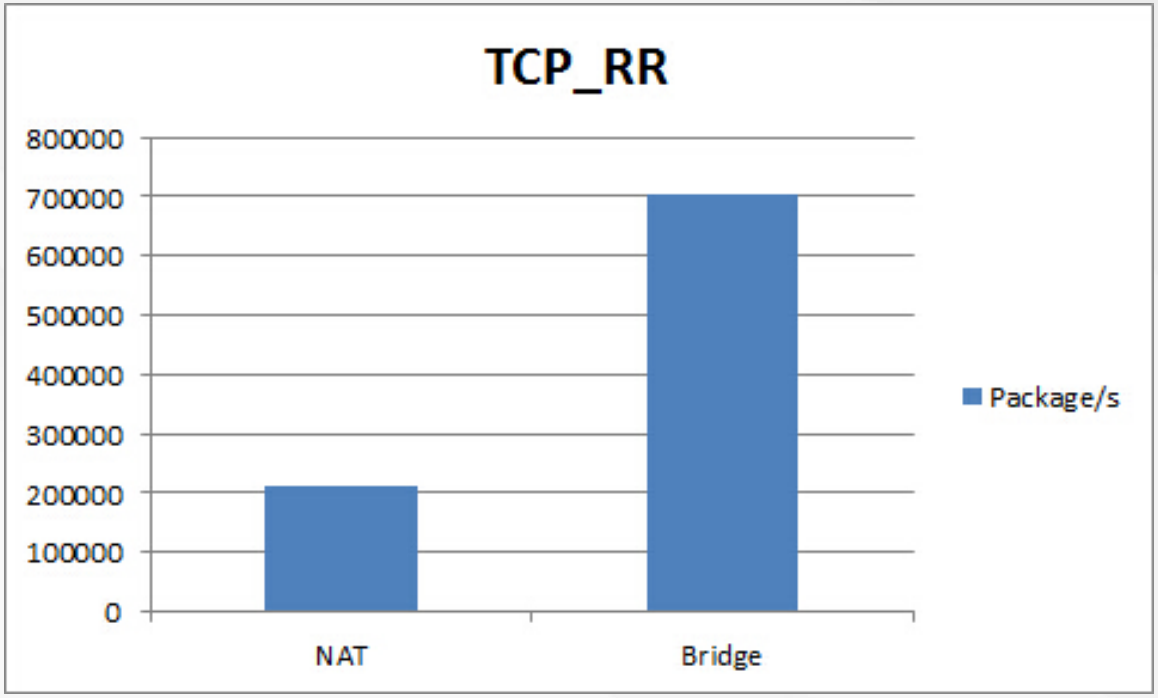
问题2—Network(2)

- Bridge+VLAN
  - Performance loss
  - CPU consumed
- Optimization
  - Set veth txqlen to 0



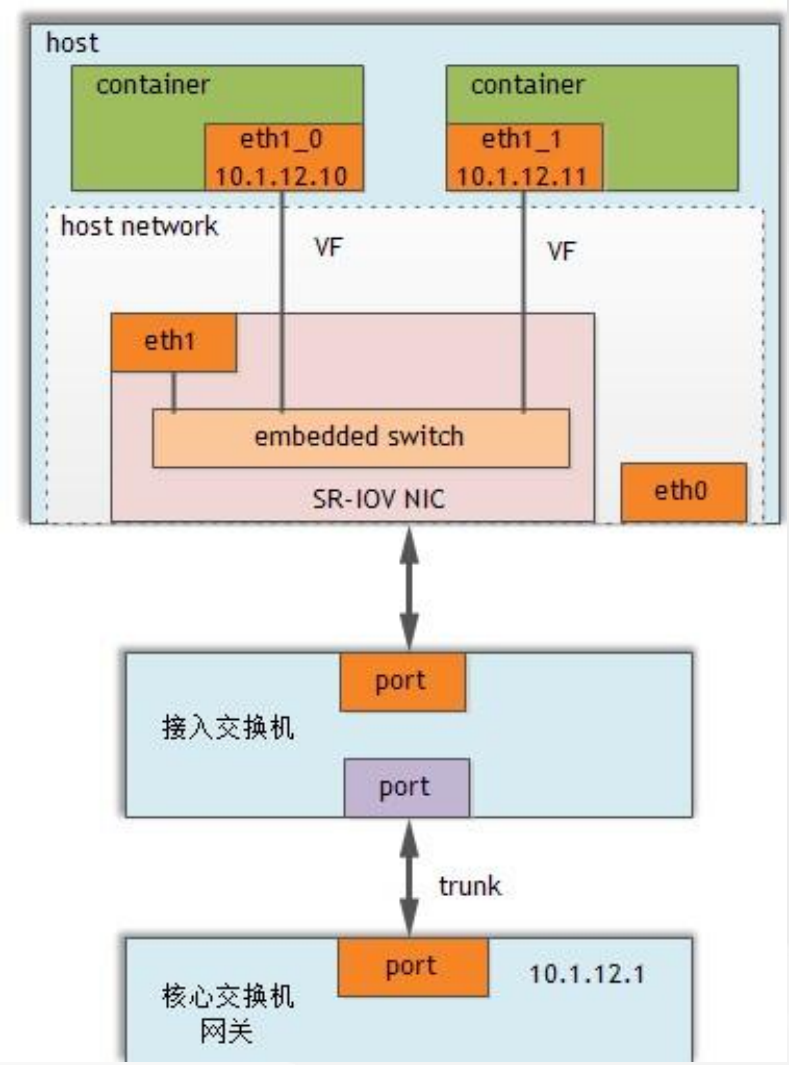
问题2—Network(3)

- NAT vs Bridge



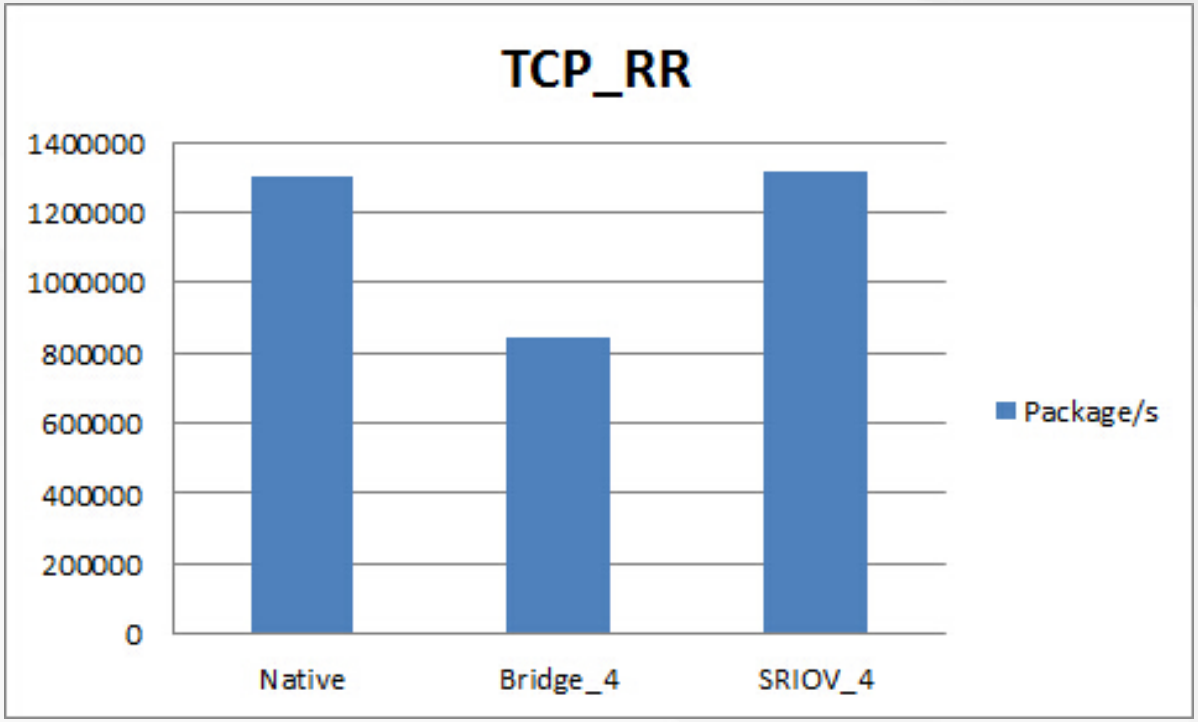
问题2—Network(4)

- SR-IOV
  - Good performance
  - Limited by VF numbers
- Optimization
  - Bind VF interrupt to CPU
  - Enable RPS



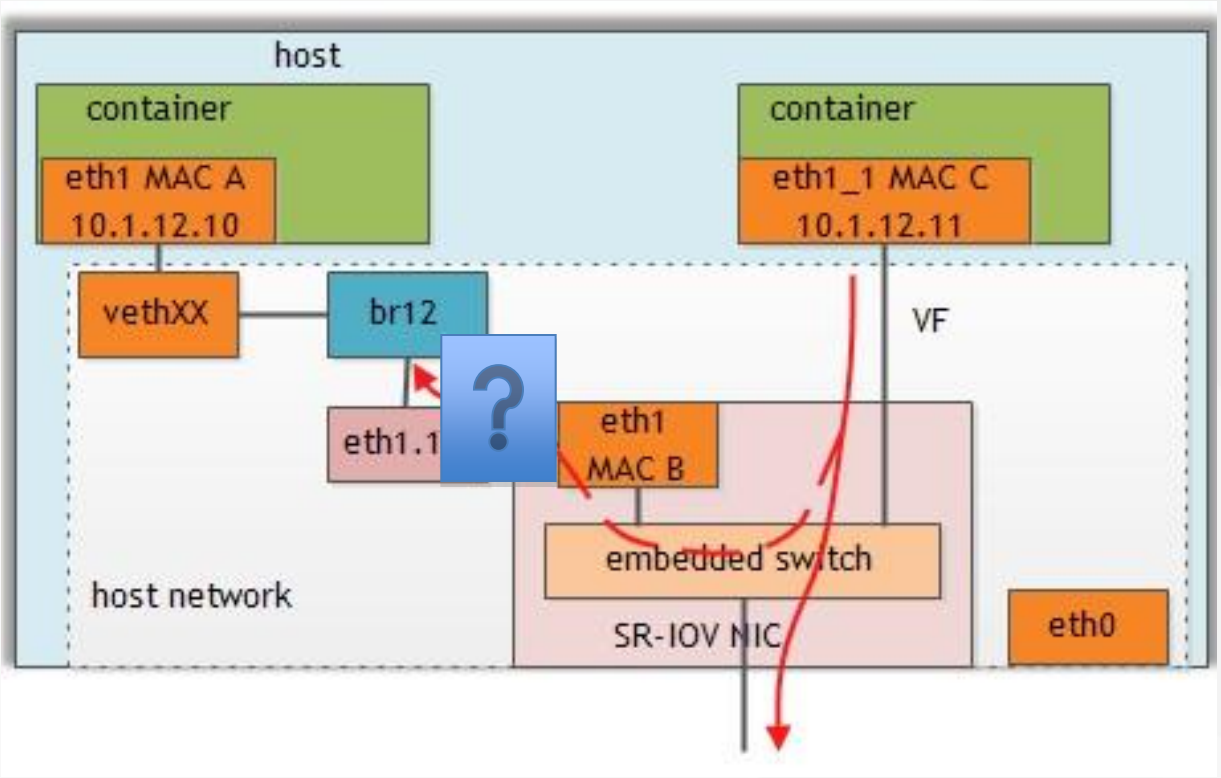
问题2—Network(5)

- Native vs Bridge vs SR-IOV

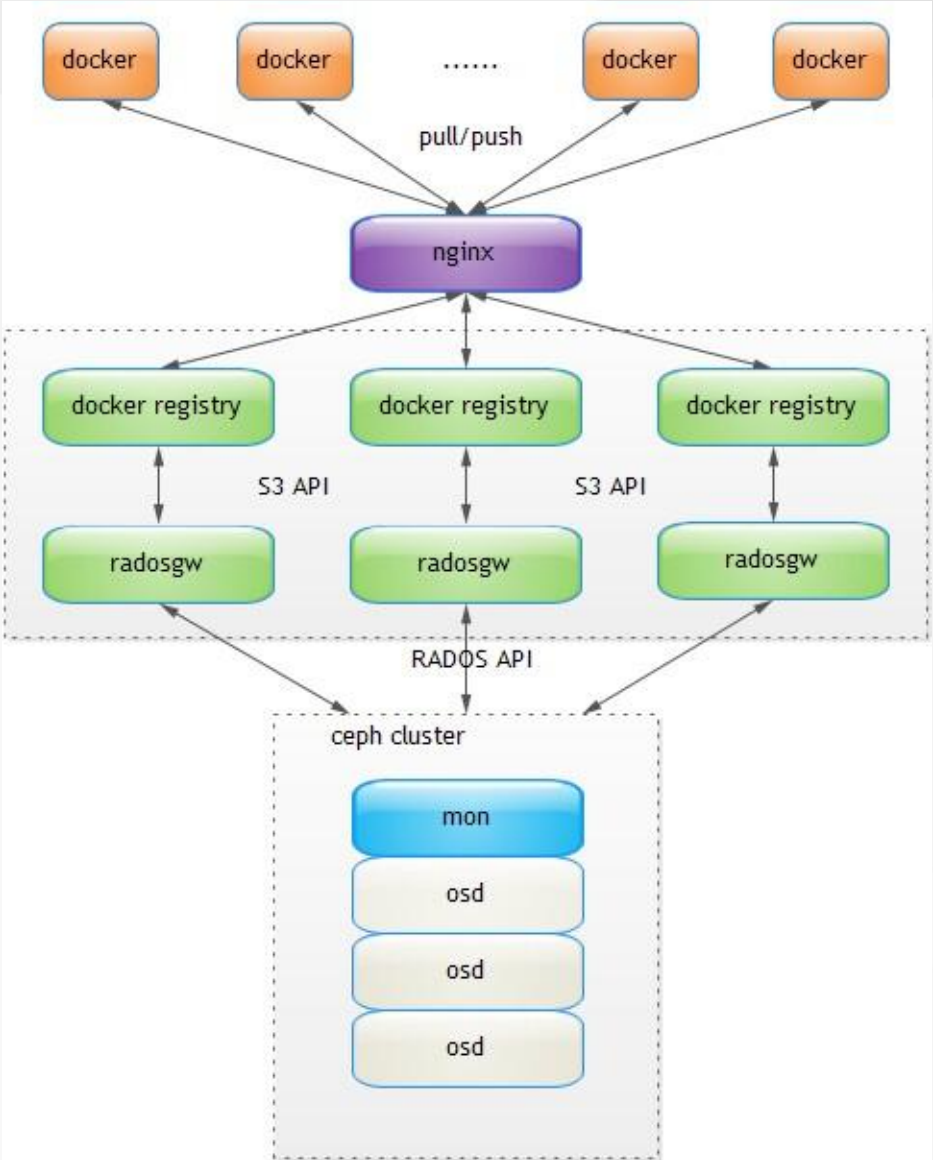


问题2—Network(6)

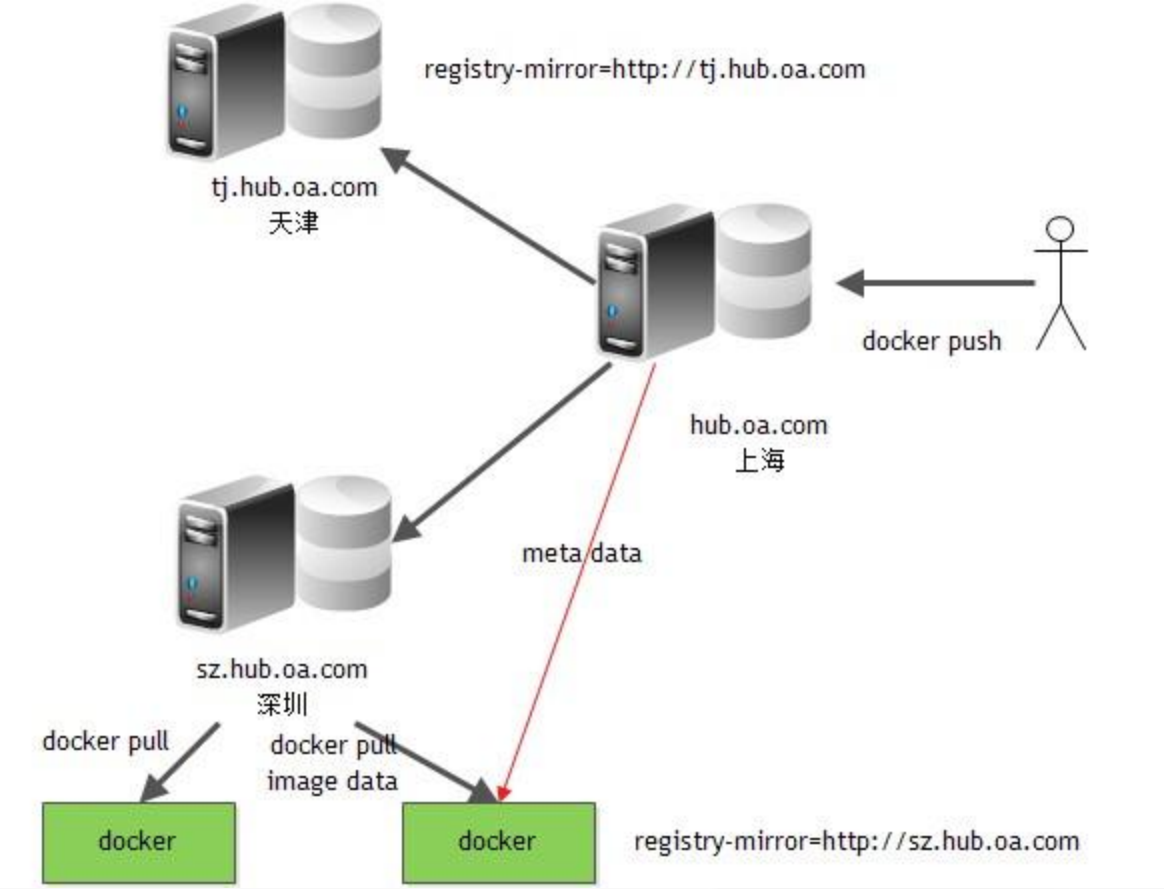
- Mix bridge with SR-IOV?



问题3—镜像存储及传输(1)

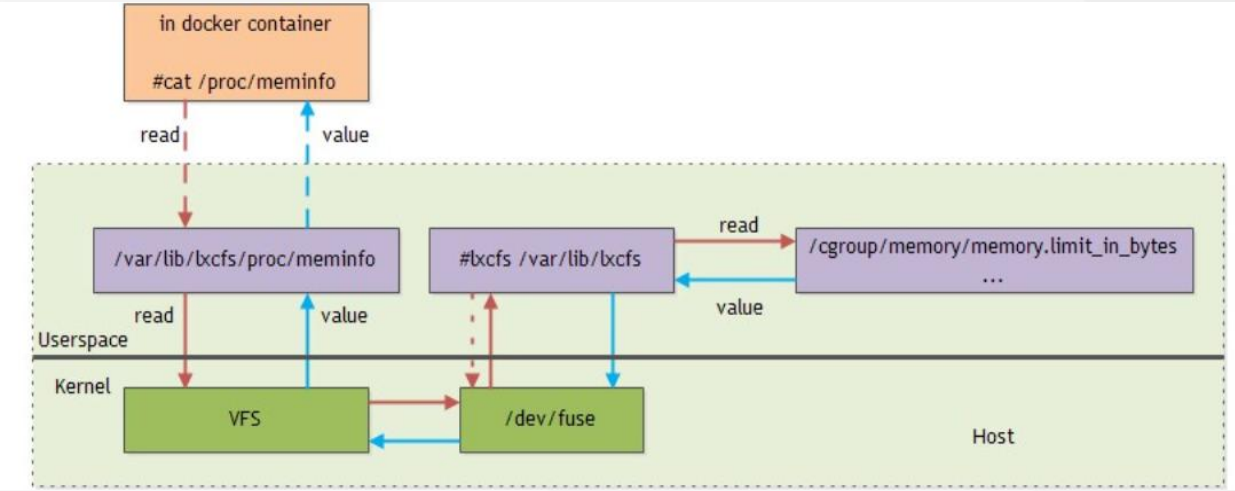


问题3—镜像存储及传输(2)



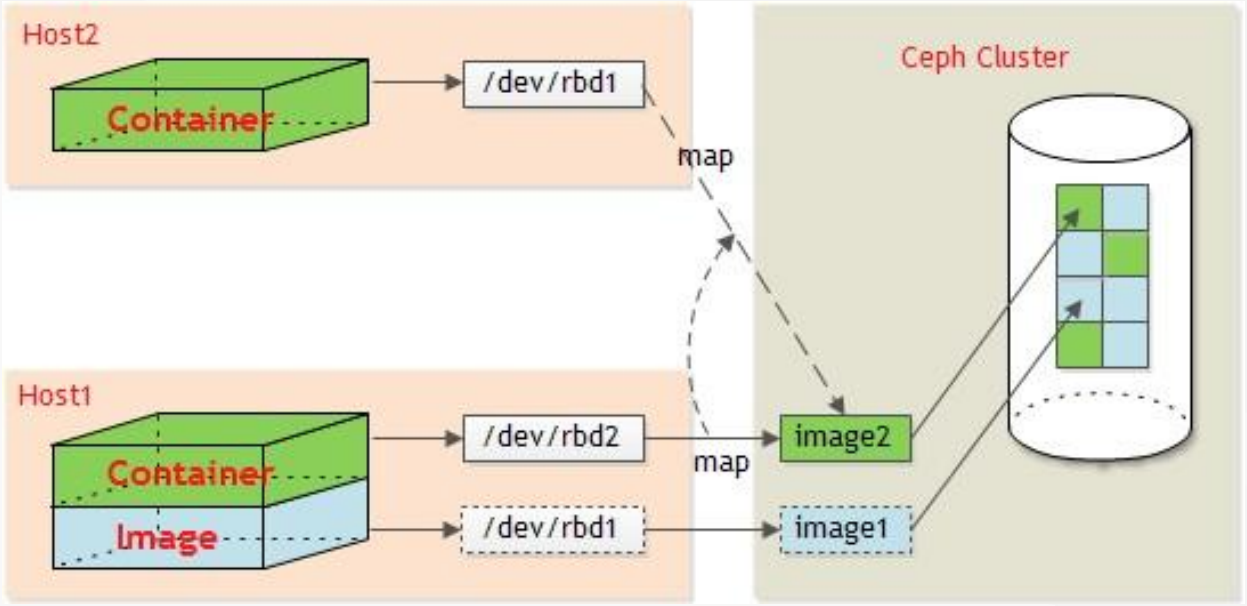
问题4—容器监控

- cAdvisor?
- lxcfs+agent



问题5—容器故障迁移

- Image＋IP漂移
- 网络存储(Ceph)
  - Data volume
  - RBD graph storage driver

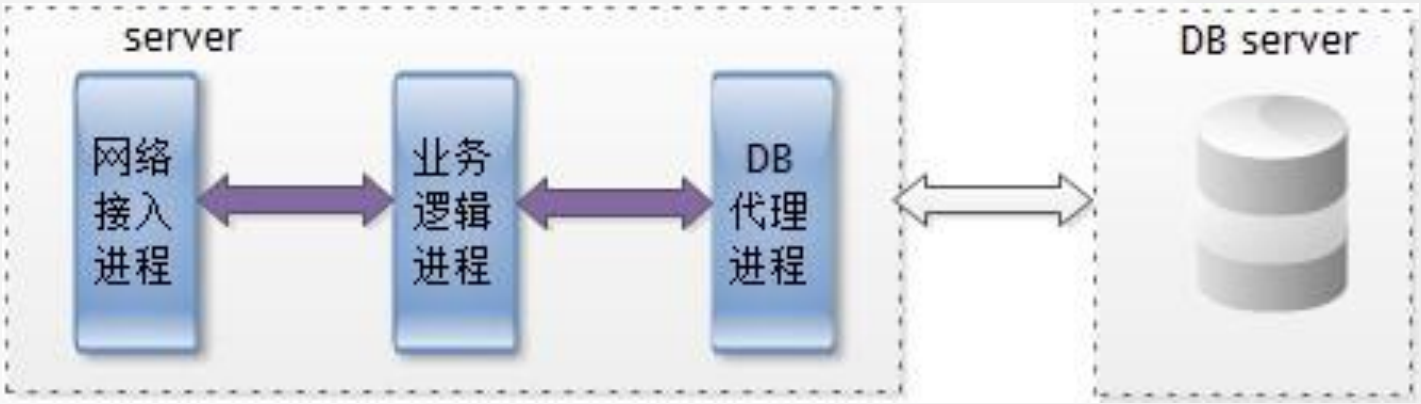


问题6—资源配额(1)

- cpuset.cpus
- memory.limit_in_bytes
- IO throttle
  - Buffer io throttle?
- volume quota
  - XFS project quota
- change quota online
  - docker cgroup cpuset.cpus='0-11'

问题6—资源配额(2)

- What’s meaning of change quota online?



问题7—Kernel

- CentOS6.5?
  - cgroup spin lock
  - VLAN device lock less
  - Kernel variable
- Device Mapper

```
Samples: 1M of event 'cycles', Event count (approx
-  91.02%  [kernel]                [k] _spin_lock
   - _spin_lock
      - 99.39% css_get_next
         - mem_cgroup_iter
            - 99.68% shrink_zone
               do try to free pages
```

```
Samples: 189K of event 'cycles', Event count (appro
-  72.45%  [kernel]                [k] _spin_lock
   - _spin_lock
      - 98.49% dev_queue_xmit
         - 99.77% ip_finish_output
```

```
redistribute3+0x143/0x190 [dm_persistent_data]
rebalance3+0x1d4/0x270 [dm_persistent_data]
rebalance_children+0x19a/0x1b0 [dm_persistent_data]
remove_raw+0x63/0x1a0 [dm_persistent_data]
dm_btree_remove+0xb0/0x150 [dm_persistent_data]
dm_thin_remove_block+0x87/0xb0 [dm_thin_pool]
process_prepared_discard+0x22/0x60 [dm_thin_pool]
process_prepared+0x87/0xa0 [dm_thin_pool]
do_worker+0x49/0x60 [dm_thin_pool]
process_one_work+0x17a/0x480
worker_thread+0x11f/0x3a0
? manage_workers+0x120/0x120
kthread+0xce/0xe0
? kthread_freezable_should_stop+0x70/0x70
ret_from_fork+0x58/0x90
? kthread_freezable_should_stop+0x70/0x70
00 00 00 00 e8 4b fc ff ff 89 de 4c 89 e7 e8 51 fe ff
```

目录

- 腾讯游戏应用现状
- 问题及解决方案
- 总结及展望

Summary

- 弹性的资源交付
- 统一的部署方式
  - 资源交付即部署
- 简单、易用
  - 技术门槛低、社区活跃、庞大的生态圈

Future

- Problems
  - Buffer IO throttle problem
  - Docker daemon hot upgrade
- Docker
  - Network plugin(overlay network)
  - Graph/Volume plugin(ceph)
- K8S
  - Make writing BigTable a CS 101 Exercise - Brendan Burns

THANKS

全球容器技术大会

剖析容器企业实践　关注容器生态圈开源项目