Docker在腾讯游戏的实践之路

腾讯互娱运营部/存储与计算资源中心/刘雷

levyliu@tencent.com





选择docker? Yes or No



- 1、Docker如此多优点,和游戏结合能发生点什么?
- 2、单纯一个docker还远远不够,还需要评估哪些因素?
- 3、当前游戏的业务架构和运维模式能和docker无缝兼容吗?
- 4、新技术的推广过程中势必会导致一些异常,如何可控? (特别是游戏)

1、Docker 容器 & VM & 物理机

2、主机模式 & 集群服务模式

3. build - ship - run run - build - ship

4、兼容运维习惯和原运维平台

- 5、稳定,稳定,稳定。 接入节奏,风险控制
- Docker for Online service

 what

 How

6、运维支撑

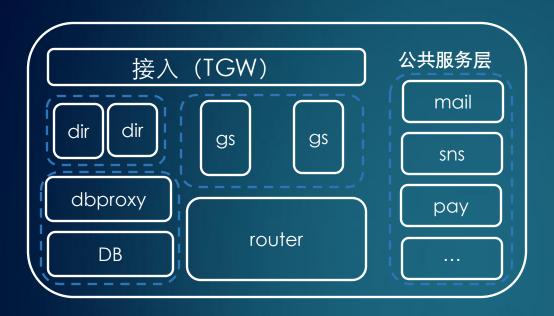
- 7、其他评估要素(包含但不局限于)
- ◆ 编排工具
- ◆文件系统
- ◆监控
- ◆成本管理
- ◆集群管理平台

- ◆网络方案
- ◆隔离性
- ◆ image标准&管理
- ◆性能适配
- ◆容器配置



主机模式的支撑方案

典型页游手游server端架构



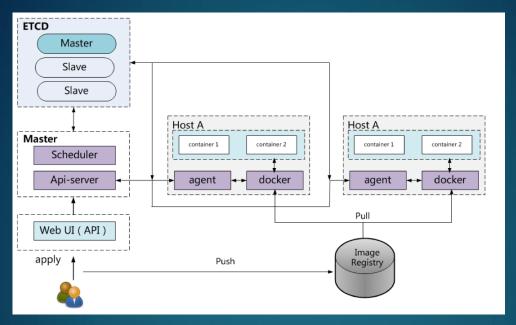


特点: 分区分服类游戏。 滚服策略。



- 1、资源快速调度
- 2、本地垂直升降配
- 3、基于image的标准化

主机模式-调度编排



Kubernetes & Docker









CPU: 容器cpu core绑定;控制高负载情况下的资源强制。

MEM: 直接利用cgroup的memory限制

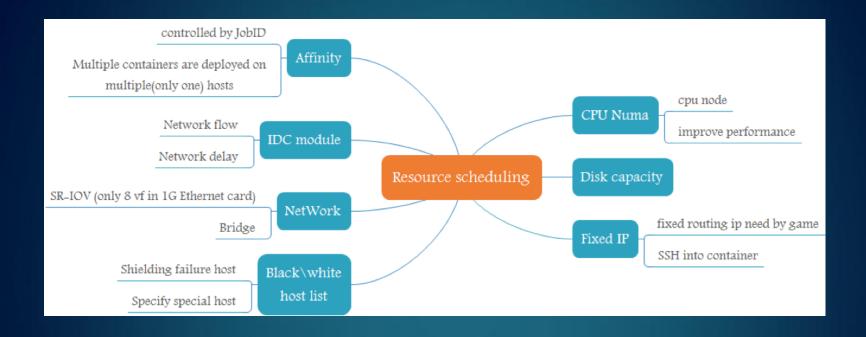
DISK: devicemapper (doing overlayfs in tkernel2.0)

对于log的存储,目前是host上面创建一个目录, xfs的project quota。

Net: 每个容器都配置了一个可路由的独立IP地址;

网络方案采用的是SR-IOV。解决当前游戏需要网络大包量的需求。

主机模式-k8s的scheduler策略定制

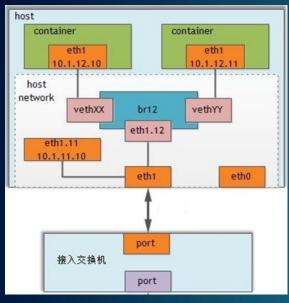


调度策略:

- IDC campus
- ◆ 机架
- ◆ 集中分布 & 分散
- ◆ 网络模式
- ◆ VF选择
- ◆ host黑白名单
- cpu numa node
- ◆ 空闲资源

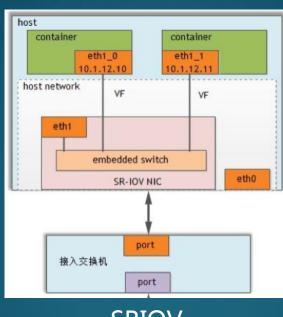
主机模式-网络方案

选择? NAT 、HOST、BRIDGE、SRIOV、VXLAN



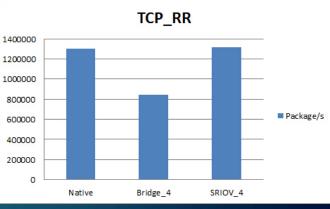
Bridge

- ◆ 大二层网络 + 千兆
- ◆ 三层网络+万兆
- ◆ SRIOV + RPS
- ◆ VF终端绑定
- ◆ 性能好,但受限于VF数

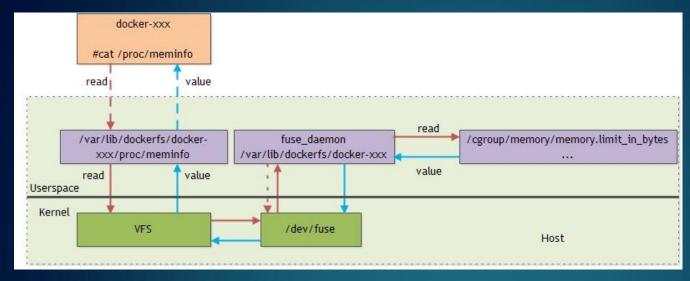








主机模式-容器性能监控(1/2)



```
top - 20:19:17 up 2:19, 1 user, load average: 0.07, 0.04, 0.00
Tasks: 26 total, 1 running, 25 sleeping, 0 stopped, 0 zombie
0սզմ
        0.0%us, 0.0%sy,
                         0.0%ni, 99.7%id,
                                          0.3%wa,
                                                   0.0%hi,
                                                                     0.0%st
                                                            0.0%si,
                0.0%su,
                         0.0%ni,100.0%id,
                                          0.0%wa,
                                                   0.0%hi.
                                                                     0.0%st
Ըքս1
        0.0%us,
                                                            0.0%si,
Cpu2
        0.0%us,
                0.0%sy,
                         0.0%ni,100.0%id,
                                          0.0%wa,
                                                   0.0%hi.
                                                            0.0%si,
                                                                    0.0%st
        0.0%us, 0.0%sy, 0.0%ni,100.0%id, 0.0%wa, 0.0%hi, 0.0%si,
Cpu3
                        91184k used, 1563/456k free,
                                                           0k buffers
     15/28640k total,
      2088952k total,
                          708k used.
                                      2088244k free.
                                                       73464k cached
Swap:
```

- 1、复用原有的TMP性能和告警系统
- 2、兼容运维的习惯
- 3、减少不必要的重复建设工作

```
0 : 6

1 : 7

2 : 8

3 : 9

4 : 10

5 : 11

6 : 18

7 : 19

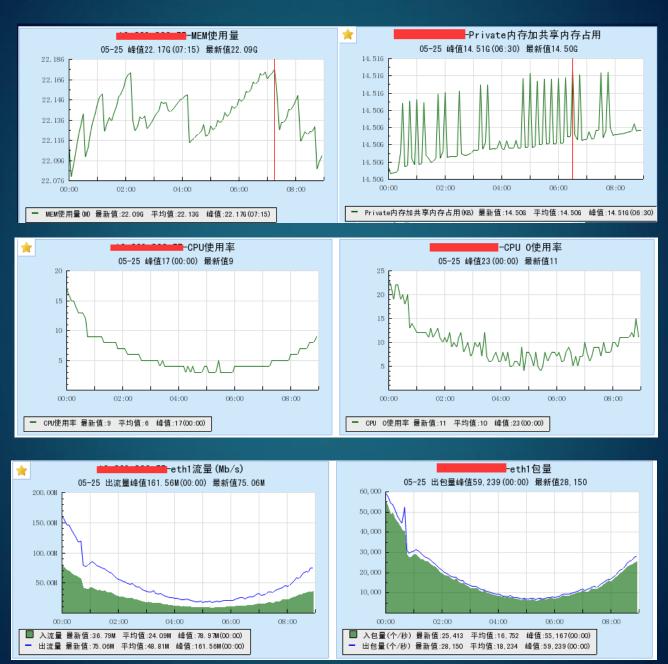
8 : 20

9 : 21

10 : 22

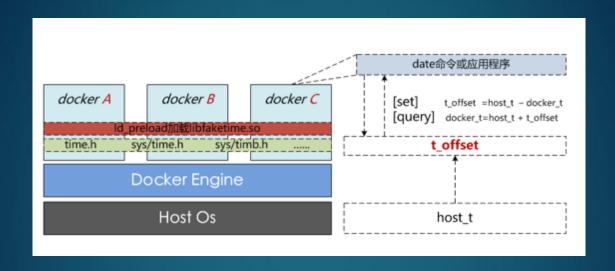
11 : 23
```

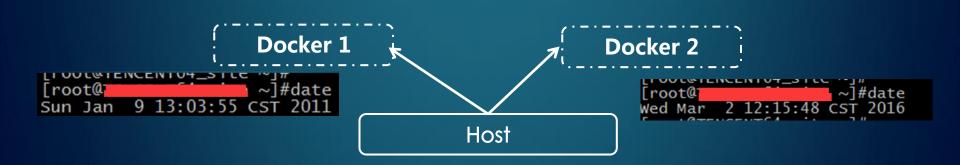
主机模式-容器性能监控(2/2)



主机模式-容器时间

游戏开发阶段的模块单元测试,功能测试,对系统时间的调整是必要需求。





主机模式-容器本地垂直升降配



CPU: 4000 CORE

MEM: 15T DISK: 100T

沙级

半容器

CPU : 4core →1core

MEM : 15G → 8G DISK : 100G → 50G

游戏Docker集群



CPU : 1000 CORE

MEM: 8T DISK: 50T

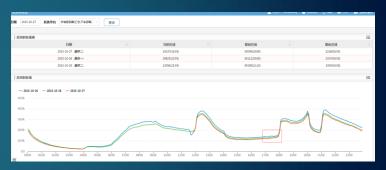


Cpu & M e m: docker cgroup ***

DISK: xfs_quota

案例:XX游戏业务:

在线降配430台120秒;在线平稳,无波动。



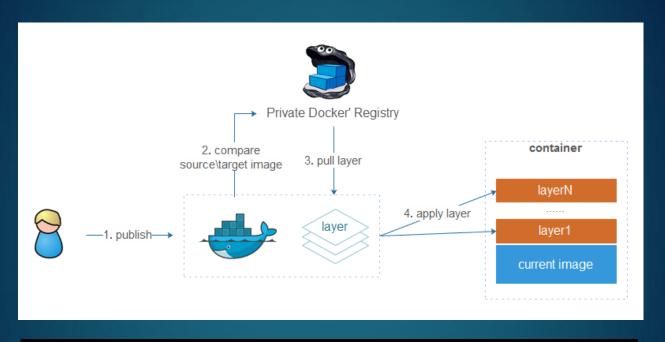
效率: 秒级在线缩(扩)容

体验:玩家毫无感知

成本:资源快速复用

主机模式-image热发布

Docker: img热发布,保留内存信息。



主机模式-总结

中途切入,用传统的滚服类游戏业务将docker容器作为主机来使用, 在不影响原有运维模式的情况下, 为游戏业务带来一些增值收益!

资源快速调度

在线本地扩缩容

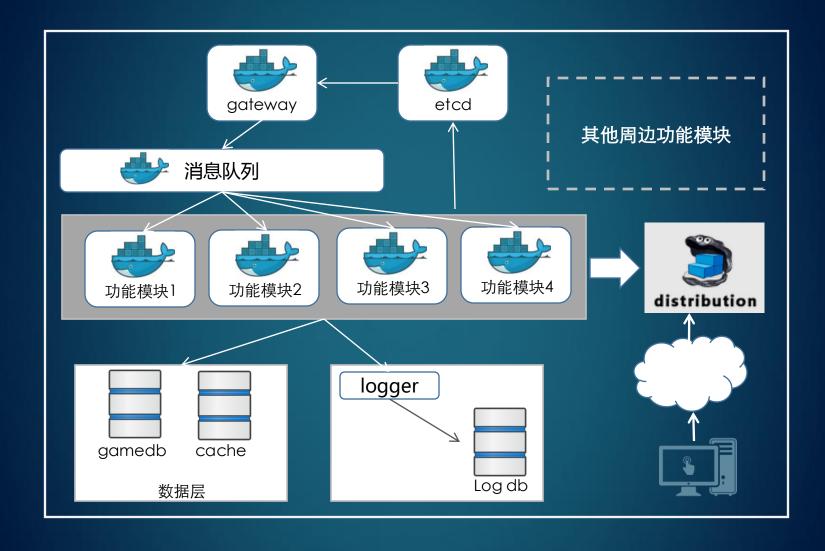
Image标准化操作



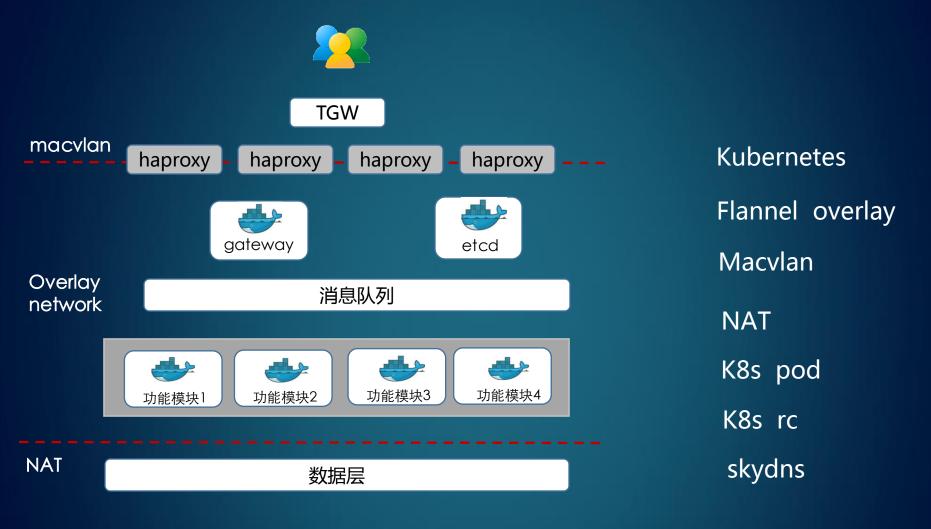


集群服务模式的支撑方案

集群服务化游戏后端架构框图



集群服务模式-网络方案



集群服务模式-优势和挑战

优势

- 1、游戏业务本身的集群服务化架构,弱化单点故障
- 2、基于image的标准化操作,极大的简化了运维部署
- 3、扩缩容简便,一键完成

挑战

- 1、众多的开源组件的集成和定制,稳定性;
- 2、对基础运维的要求更高
- 3、运维支持系统的全新构建 性能,监控告警,日志分析;

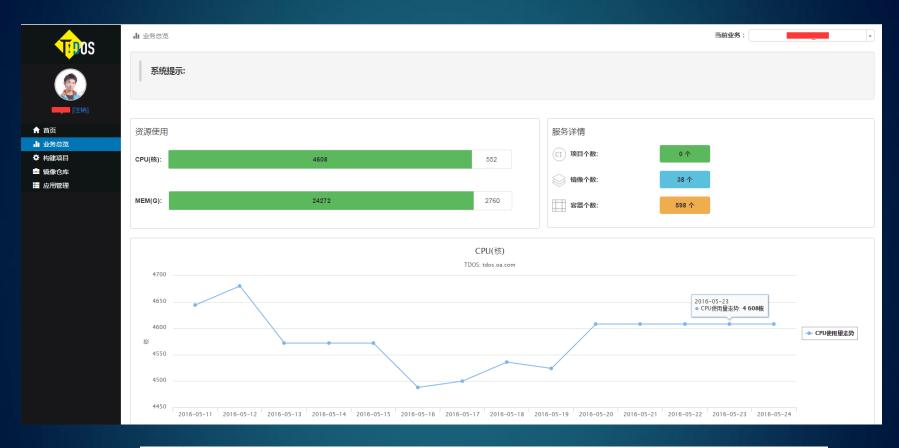


容器化实践进展

Docker在游戏上实践的节奏

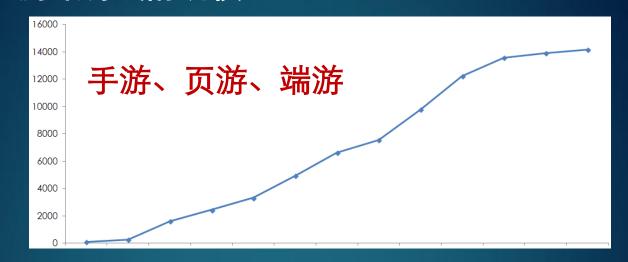


TDOS Tencent Docker for Online Service





Docker目前在腾讯游戏的规模







基于tdos-api的业务功能app







实践过程中的典型问题分享

Devicemapper

问题:

thin-provisioning discard kernel crash

解决:

- (1) nodiscard --storageopt dm.blkdiscard=false
- (2) Tkernel 2.0内核已经修复

Kubernetes

问题:

kubernets 1.1.3在创建容器失败或在容器内部没有找到有效的网络接口时,会将host所有网络接口的hairpin开关强制置为ON,造成arp包回弹,网络时断时续。

解决:

禁止这个逻辑。

Kubernetes & etcd

问题:

Kubernets 是以etcd数据为中心,假设etcd集群的数据发生了错乱,结果如何?

解决:

- 1) Node index一致性监控
- (2) Etcd数据备份,版本升级
- (3) 优化k8s的逻辑

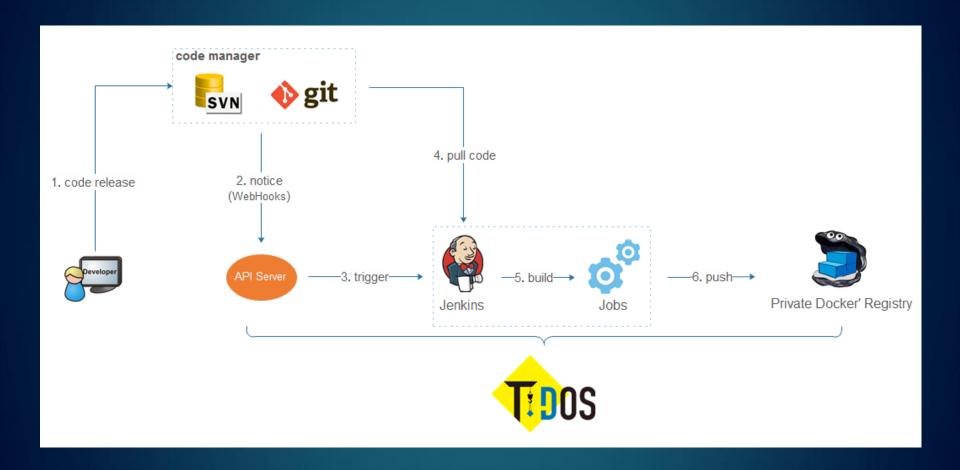






计划中的一些事情

基于docker的持续集成



Something else!

- 1 devicemapper -> overlays
- 2、镜像分发加速 (P2P)
- 3, docker on ceph
- 4、基于持续集成,完成开发-测试-灰度-现网发布的全流程构建!





谢谢!