BDTC 2016中国大数据技术大会
Big Data Technology Conference 2016

腾讯大数据能力输出之路

陈鹏

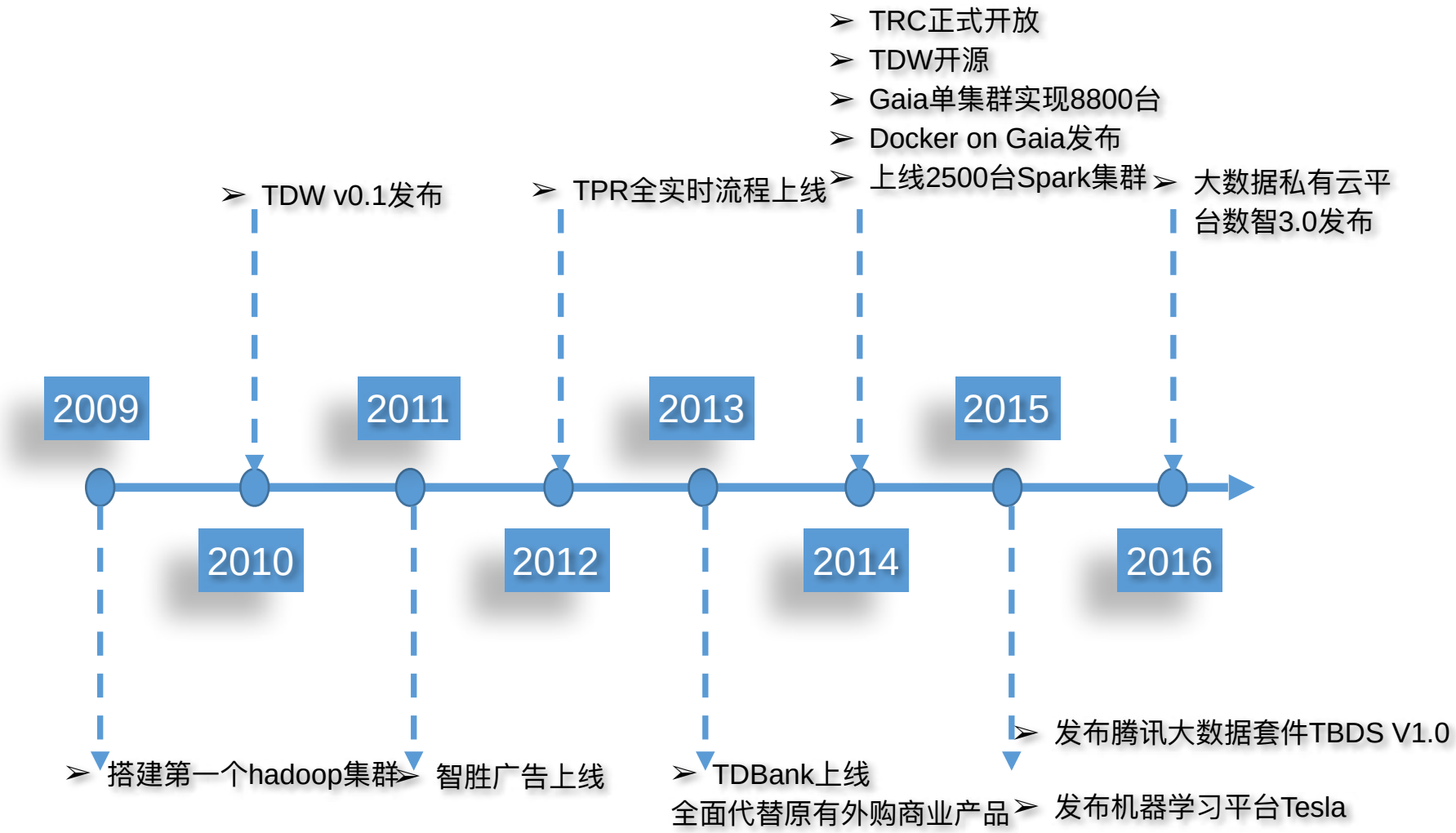
腾讯云数智大数据平台
研发负责人

- 概览
- 基础平台
- 任务调度云化

概览

Part 01

发展历程


TRC正式开放
TDW开源
Gaia单集群实现8800台
Docker on Gaia发布
上线2500台Spark集群
大数据私有云平台数智3.0发布
TDW v0.1发布
TPR全实时流程上线
2009
2010
2011
2012
2013
2014
2015
2016
搭建第一个hadoop集群
智胜广告上线
TDBank上线
全面代替原有外购商业产品
发布腾讯大数据套件TBDS V1.0
发布机器学习平台Tesla

云市场细分

数据敏感性客户

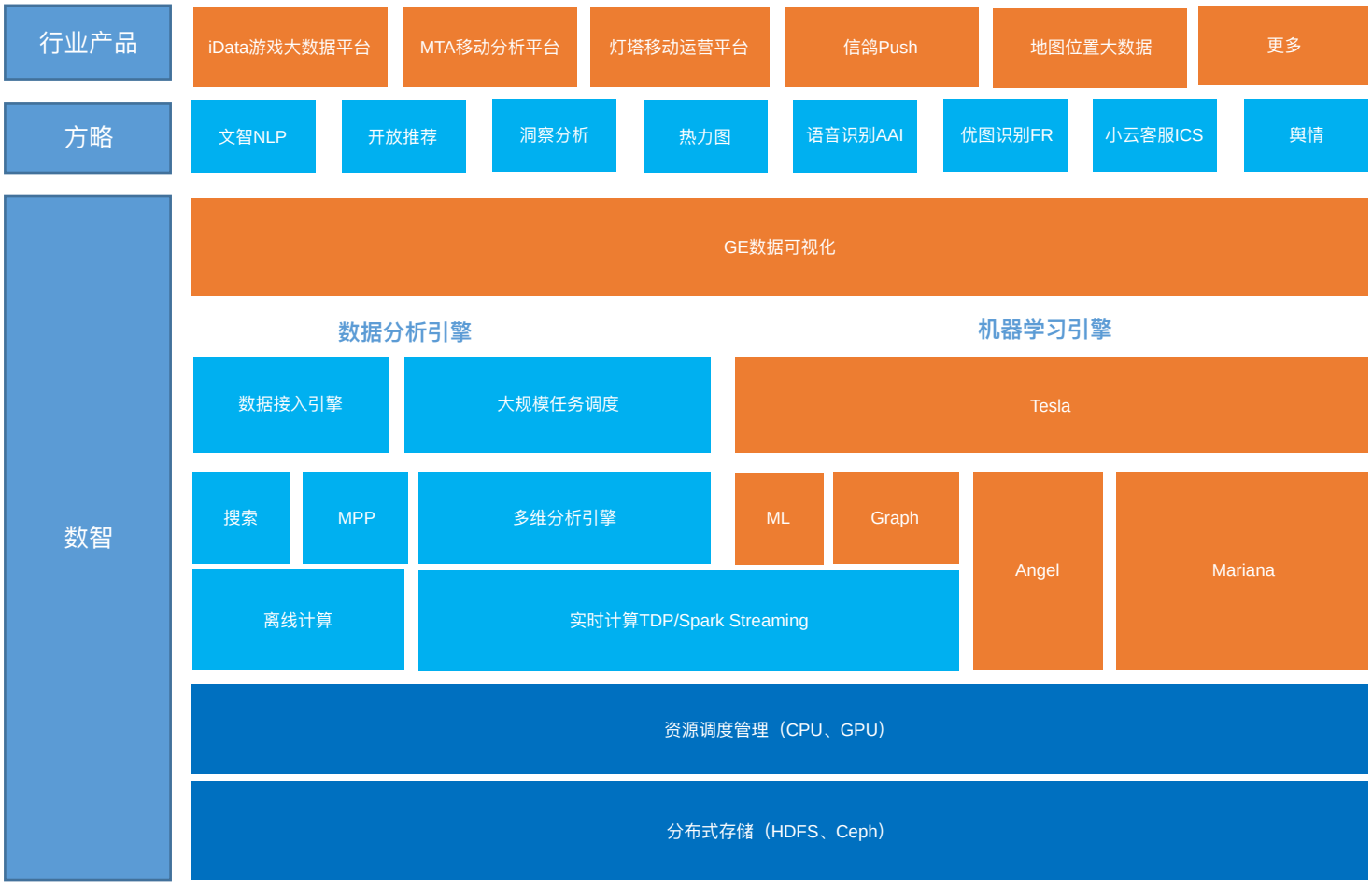
免运维、弹性需求客户

成本极敏感、数据融合诉求

私有云	公有云单租户	公有云多租户
大数据平台能力		

数智方略产品家族图谱

一站式、全流程大数据服务平台



数智产品特点

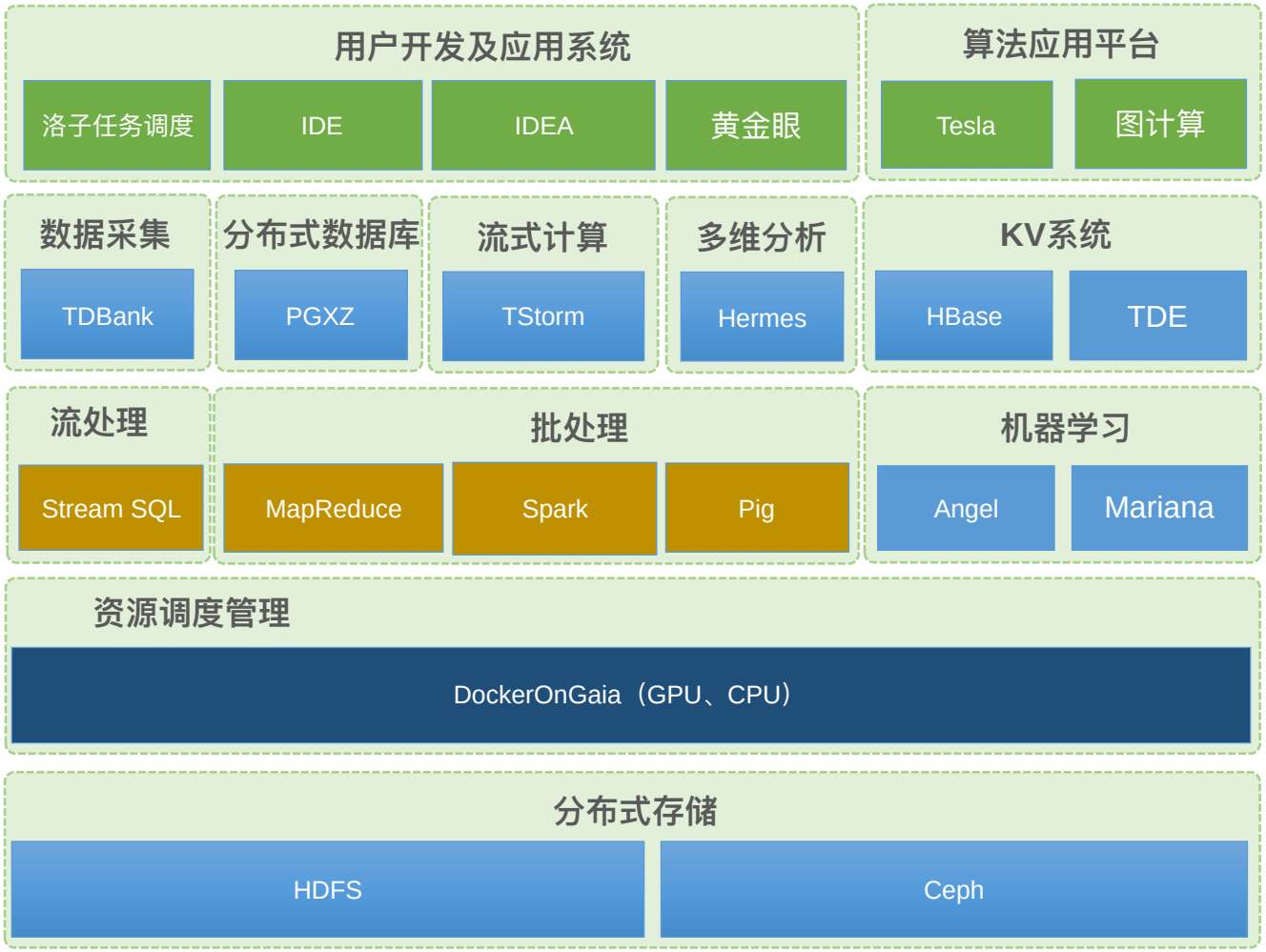
- 高度集成化，接入、存储、离线/实时计算、机器学习、可视化展现服务
- 提供可拖拽式的支持分钟级调度的任务调度系统
- 提供高性能多维分析引擎
- 提供全局设备、组件、任务纬度的运维系统

数智产品优势

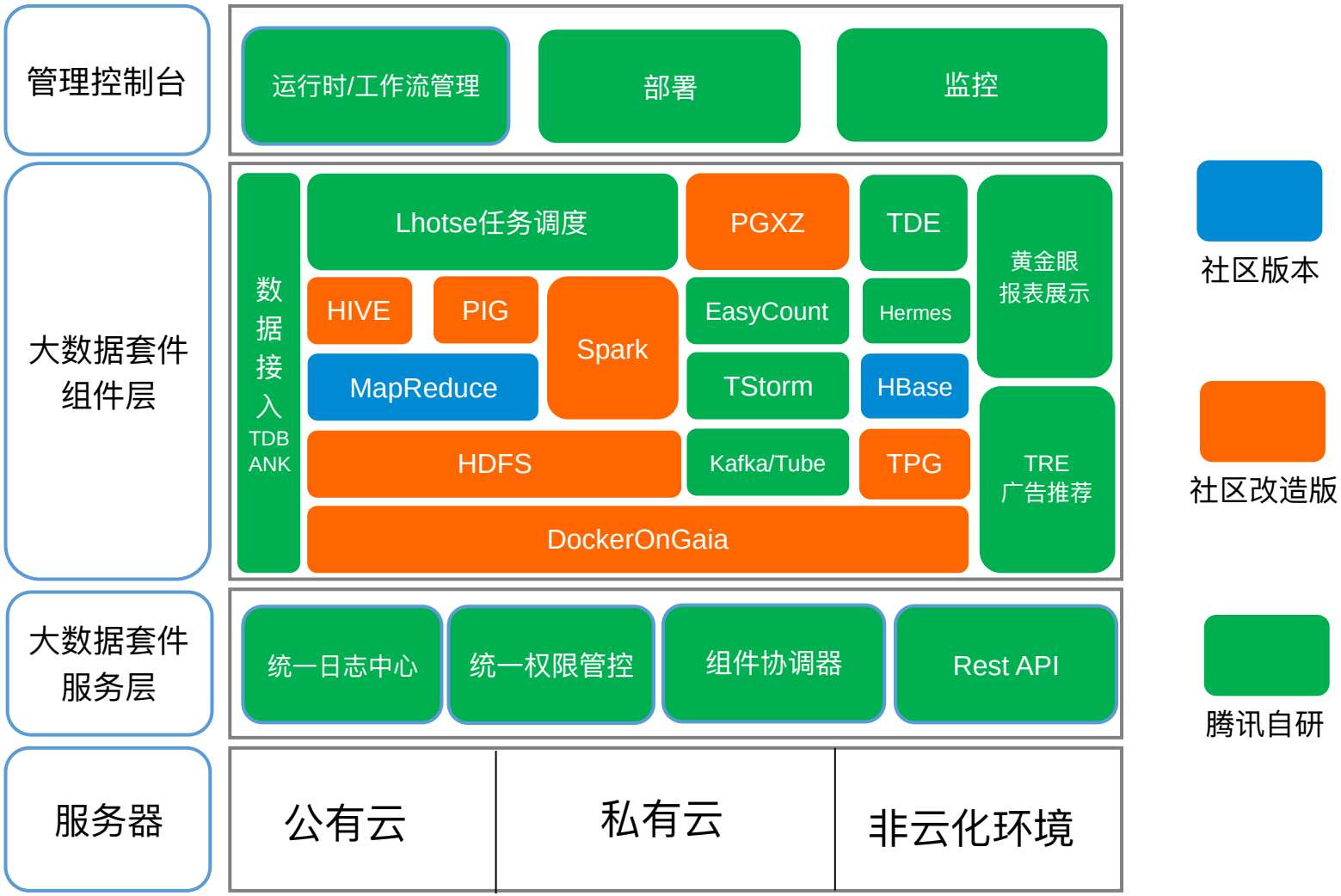
一站式 门户	统一数据 安全管控	30+ 深度优化组件	全开放 API

源于社区，完美兼容，平滑过渡

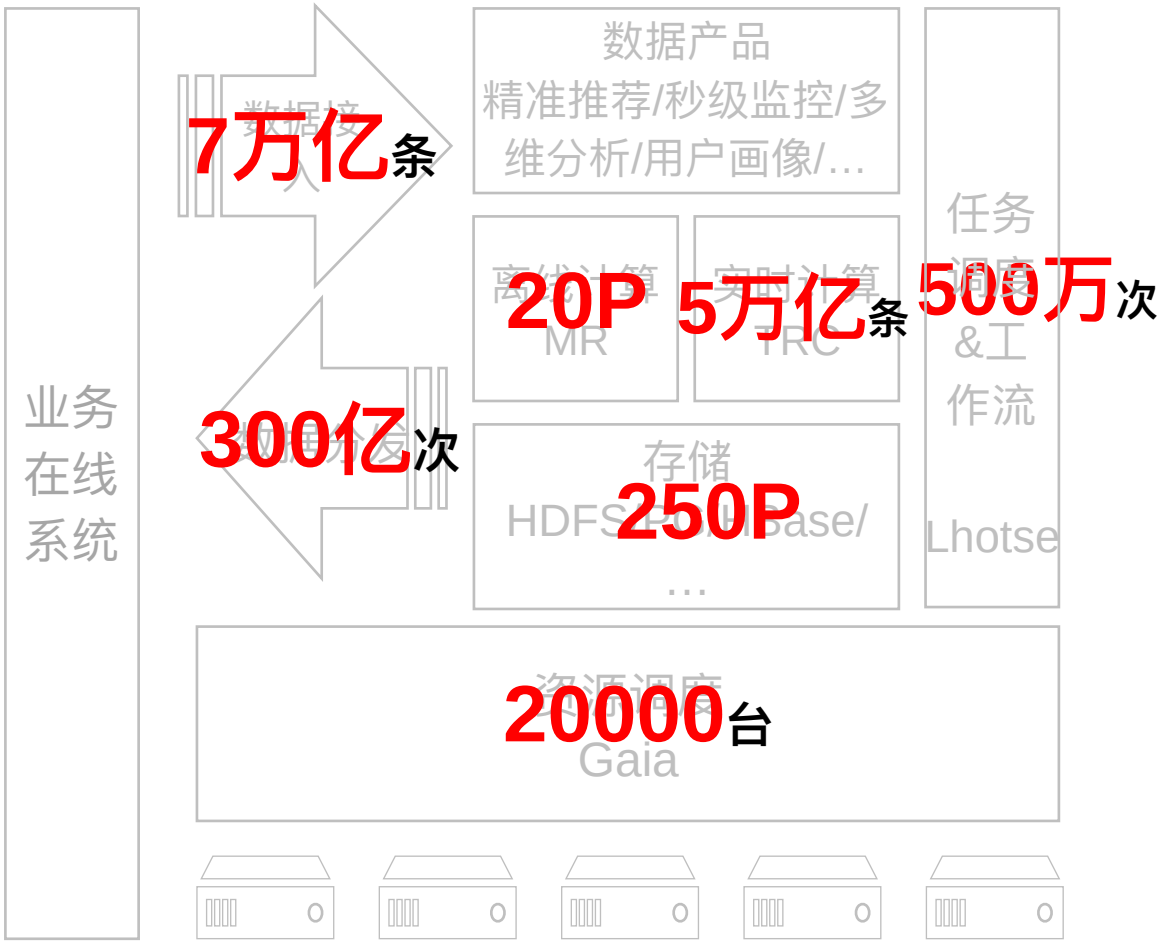
数智家族图谱


用户开发及应用系统
洛子任务调度
IDE
IDEA
黄金眼
算法应用平台
Tesla
图计算
数据采集
TDBank
分布式数据库
PGXZ
流式计算
TStorm
多维分析
Hermes
KV系统
HBase
TDE
流处理
Stream SQL
批处理
MapReduce
Spark
Pig
机器学习
Angel
Mariana
资源调度管理
DockerOnGaia（GPU、CPU）
分布式存储
HDFS
Ceph

数智大数据平台架构


管理控制台
运行时/工作流管理
部署
监控
大数据套件组件层
数据接入 TDB ANK
Lhotse任务调度
PGXZ
TDE
黄金眼 报表展示
HIVE
PIG
Spark
EasyCount
Hermes
MapReduce
TStorm
HBase
HDFS
Kafka/Tube
TPG
TRE 广告推荐
DockerOnGaia
大数据套件服务层
统一日志中心
统一权限管控
组件协调器
Rest API
服务器
公有云
私有云
非云化环境
社区版本
社区改造版
腾讯自研

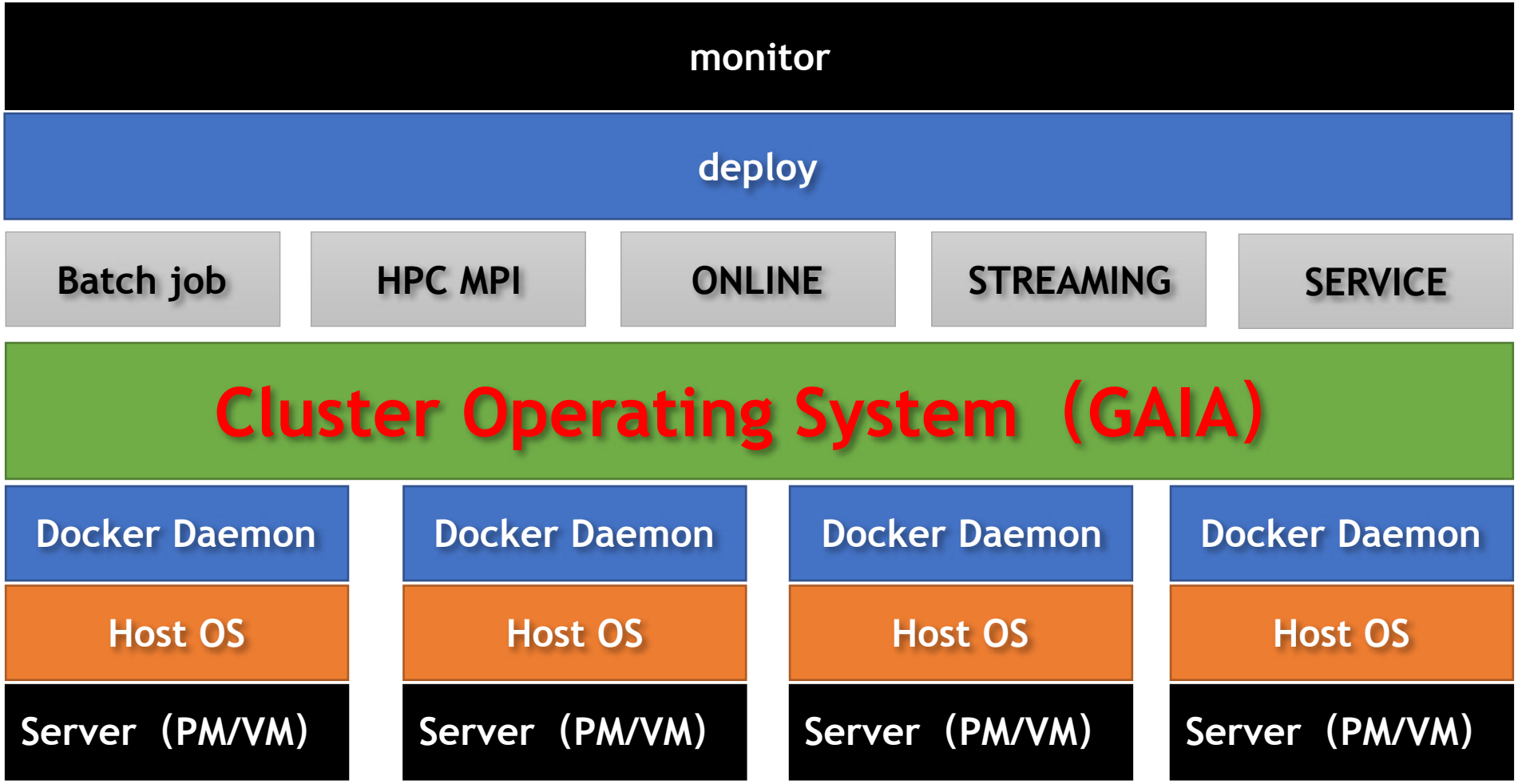
处理能力


业务
在线
系统
7万亿条
300亿次
数据产品
精准推荐/秒级监控/多
维分析/用户画像/…
离线计算
MR
实时计算
TRC
20P
5万亿条
存储
HDFS
250P
…
任务
调度
&工
作流
Lhotse
500万次
20000台
Gaia

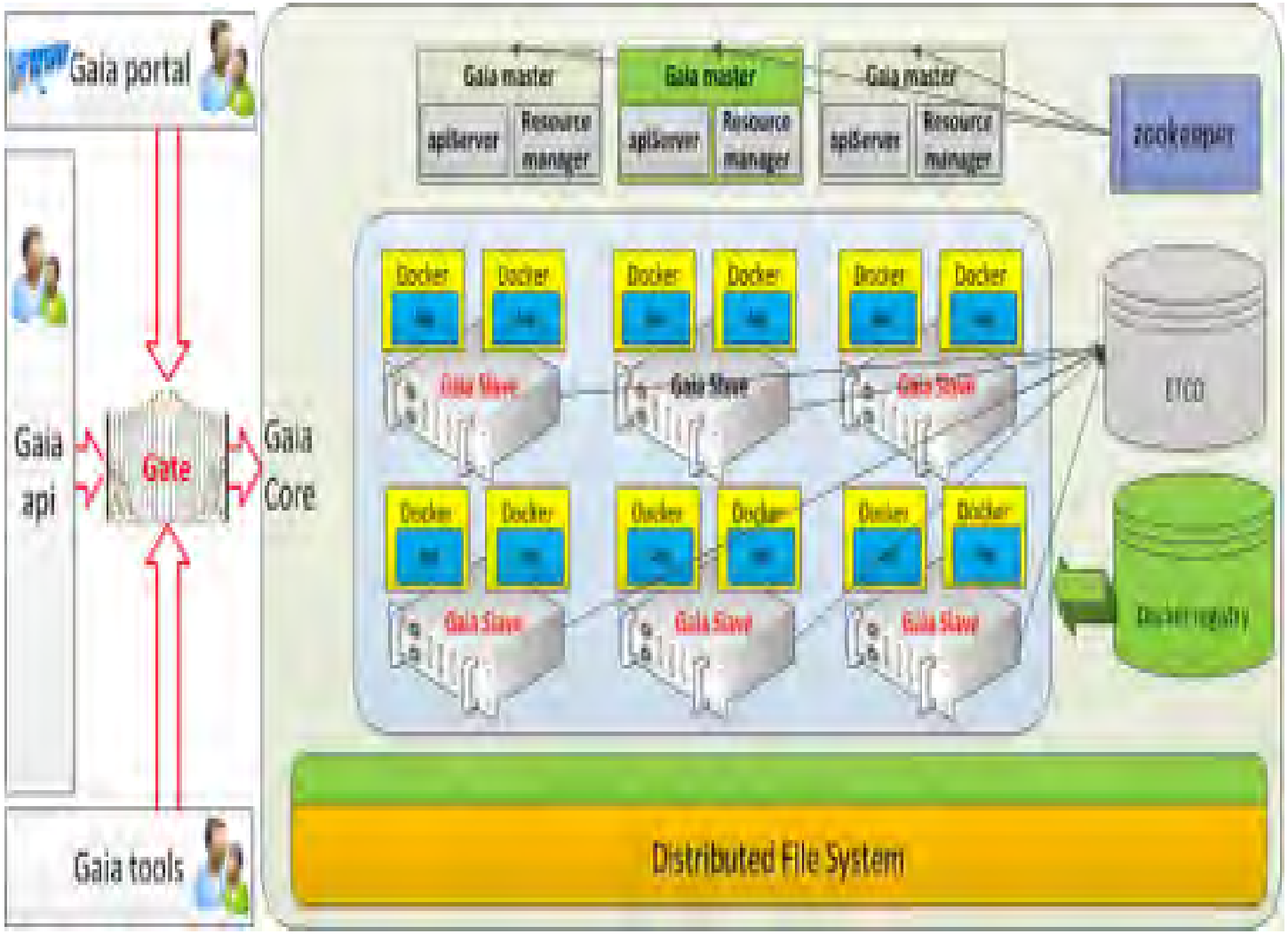
基础平台

Part 02

1. **资源管理**
   - ➢ YARN vs I层(Infrastructure)
2. **数据存储**
   - ➢ HDFS vs I层(Infrastructure)


monitor
deploy
Batch job
HPC MPI
ONLINE
STREAMING
SERVICE
Cluster Operating System（GAIA）
Docker Daemon
Docker Daemon
Docker Daemon
Docker Daemon
Host OS
Host OS
Host OS
Host OS
Server（PM/VM）
Server（PM/VM）
Server（PM/VM）
Server（PM/VM）

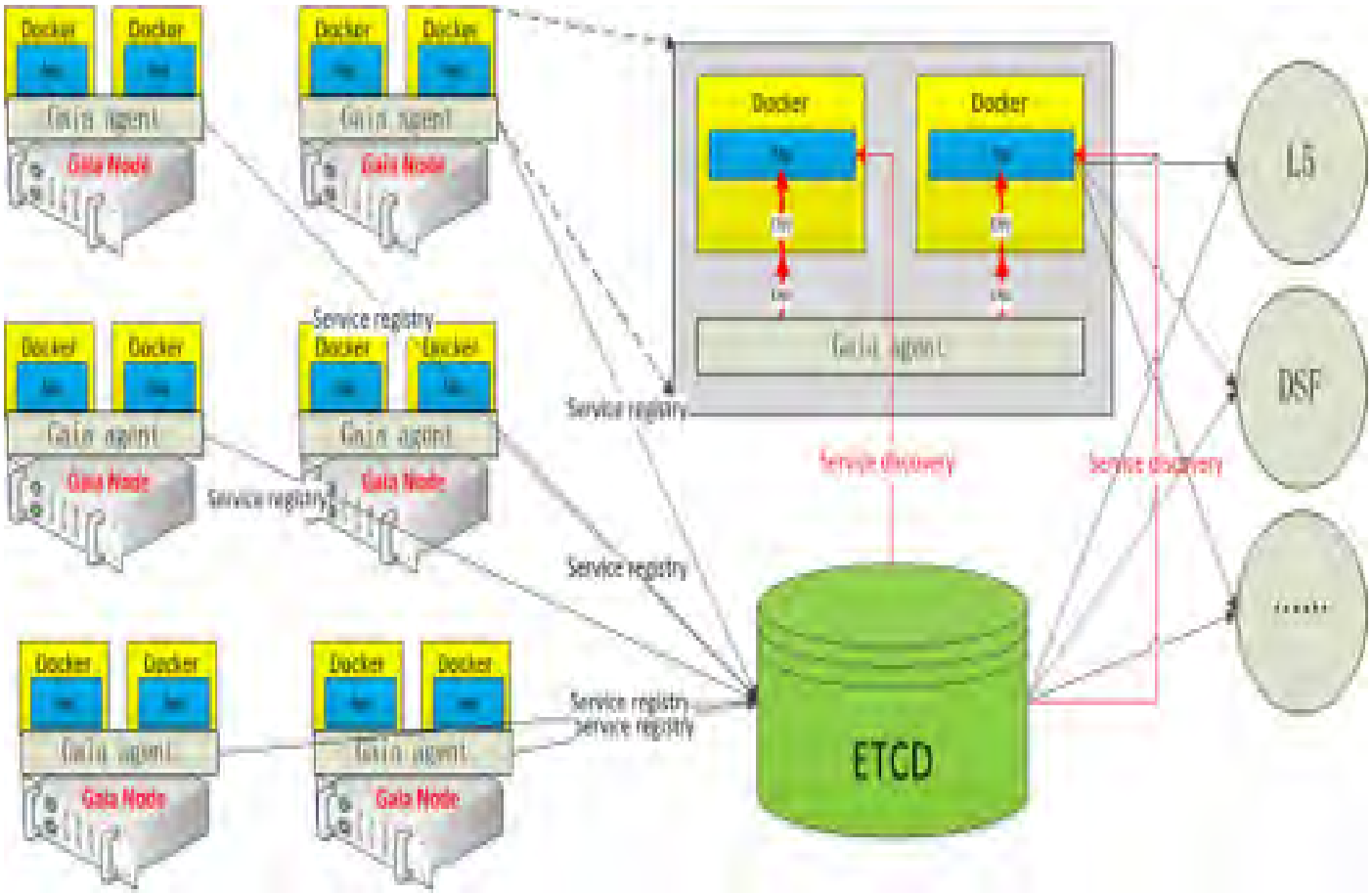
Docker on Gaia

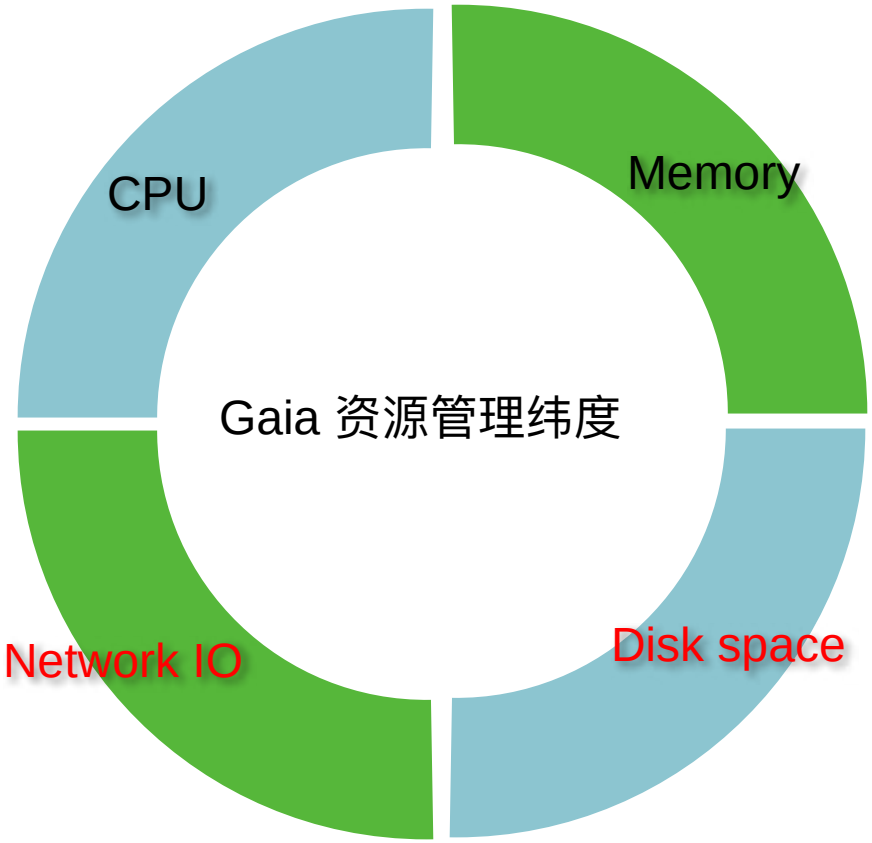


- **8800**：单集群节点数
- **20w+**：调度能力覆盖20w个核
- **8k**：作业并发度数
- **2500**：资源池个数
- **0.2ms**：container平均调度匹配时间
- **1.3亿**：日运行container数
- **120w**：日运行作业数
- **95%**：峰值vcore、memory使用率

- **资源共享**
- **异构环境共存**
- **动态扩缩容**
- **容灾容错**
- **自动化运维**
  - 一键式部署：申请资源后提交app，剩余事情交给dockerongaia
  - 用户聚焦业务
- **灰度运营**
  - 以container为单位进行升级、回滚等操作

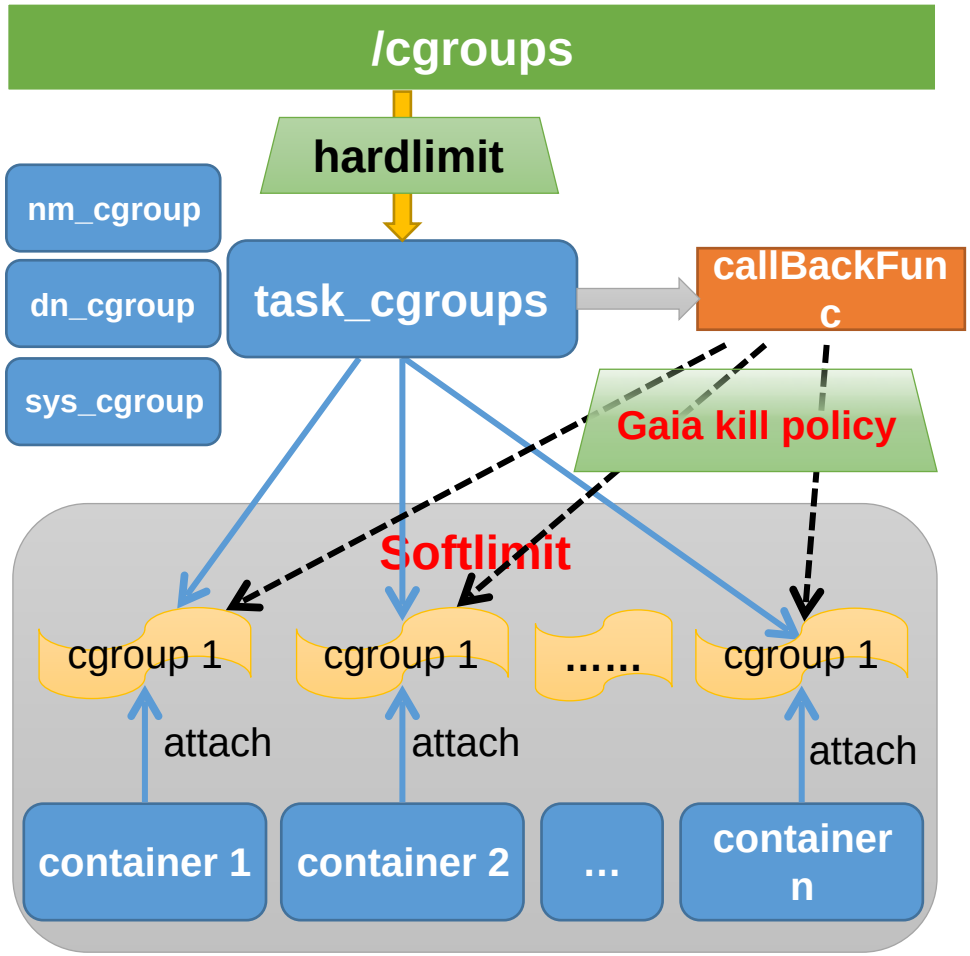
- **多业务共享**
  - 公平的使用集群资源
  - 保证各自业务的quota
- **保证高优先级作业**
  - 抢占
  - Service
  - batch
- **集群整体资源利用率**
  - cpu和memory-intensive的作业混布
  - 大作业和小作业混布
  - Service和batch混布
- **可扩展性**
- **调度吞吐**


Docker
Gaia agent
Gaia Node
Service registry
Service discovery
L5
DSF
ETCD



增加资源维度

更多的资源管理纬度
弹性的CPU控制
弹性的内存控制
弹性的磁盘容量控制
弹性的网络出带宽控制
弹性的网络入带宽控制
弹性的Disk IO控制
Buffer IO控制



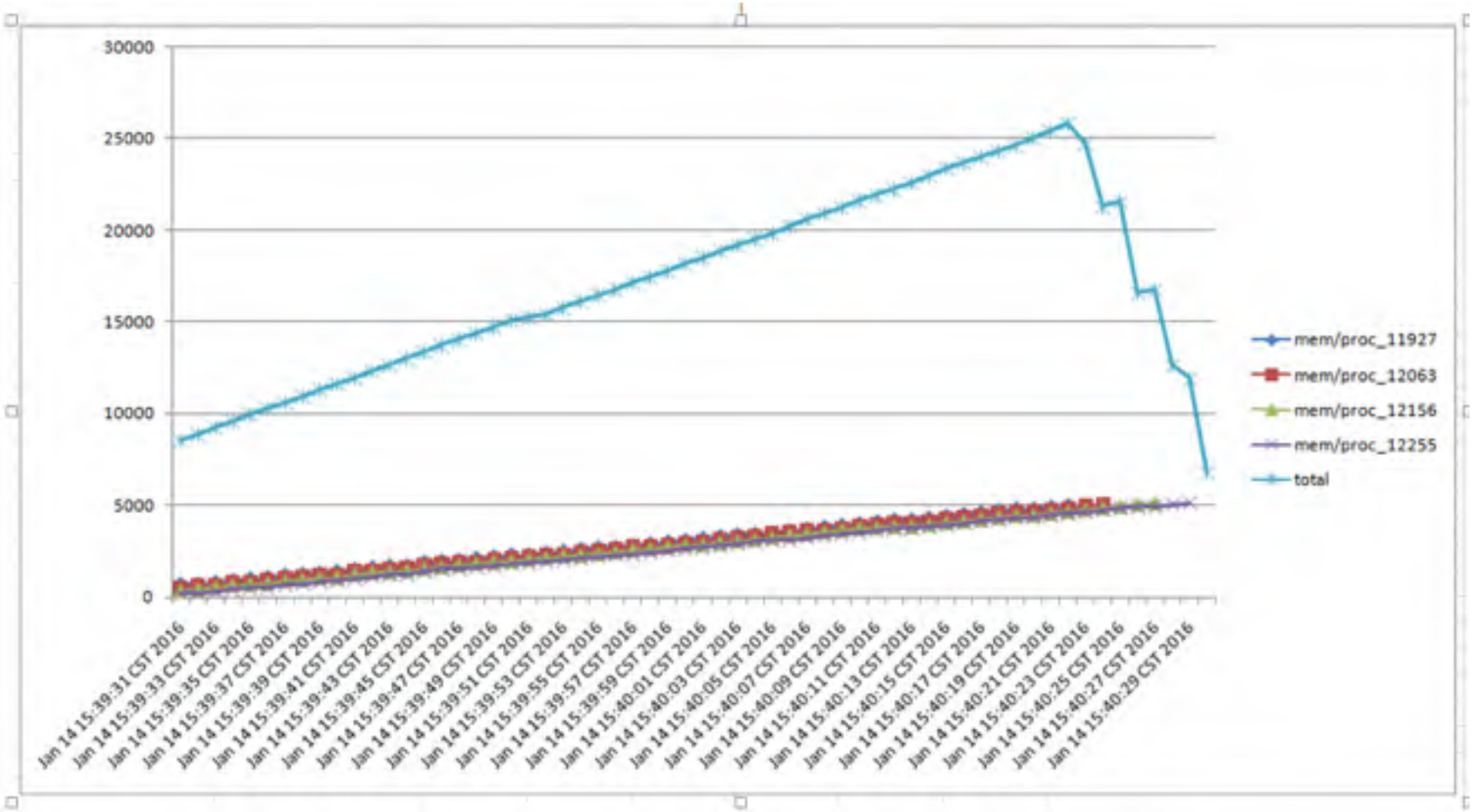
EMC Elastic Memory Control（弹性内存控制）

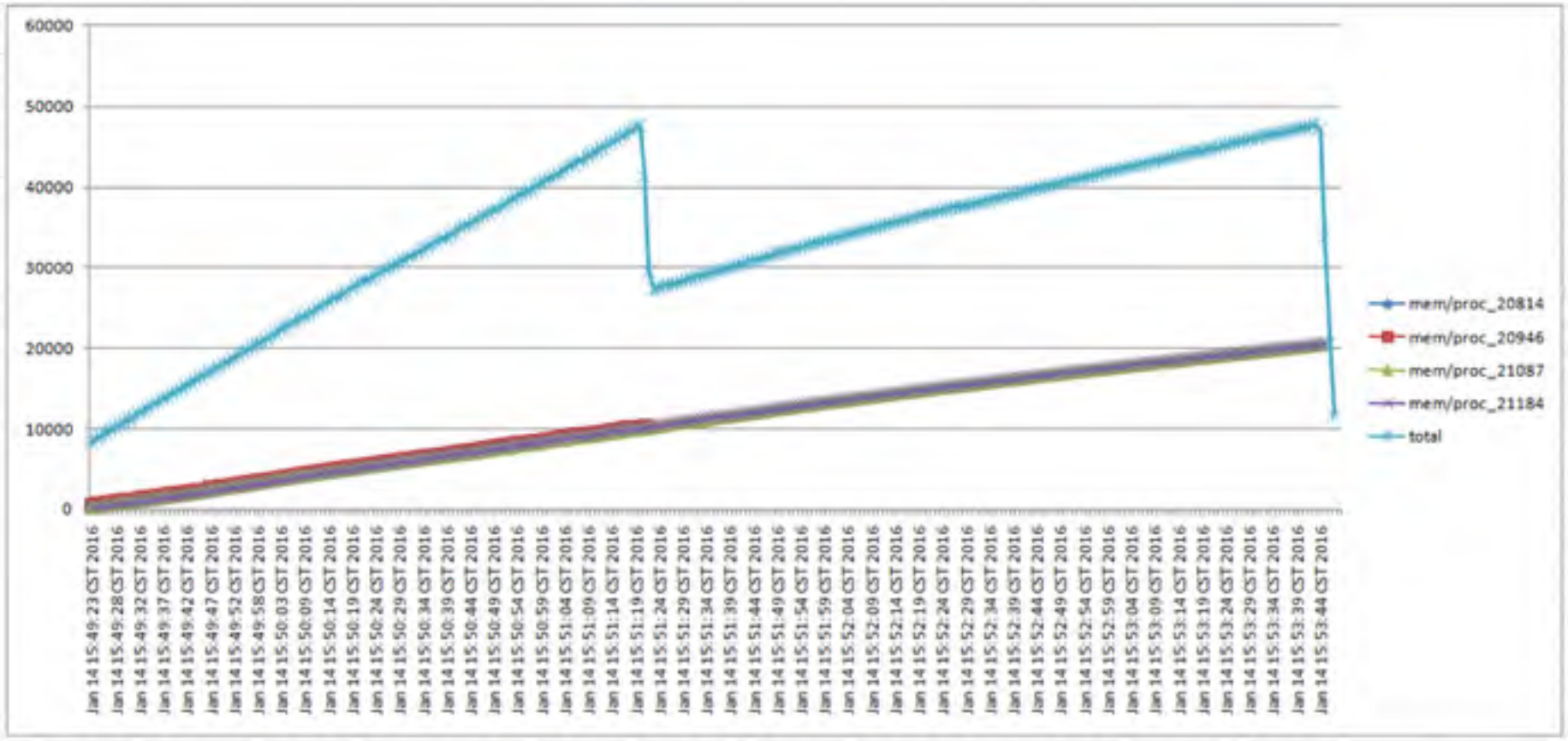
1）不会触发系统oom kill：使用了container机制，且task_cgroup是hardlimit。

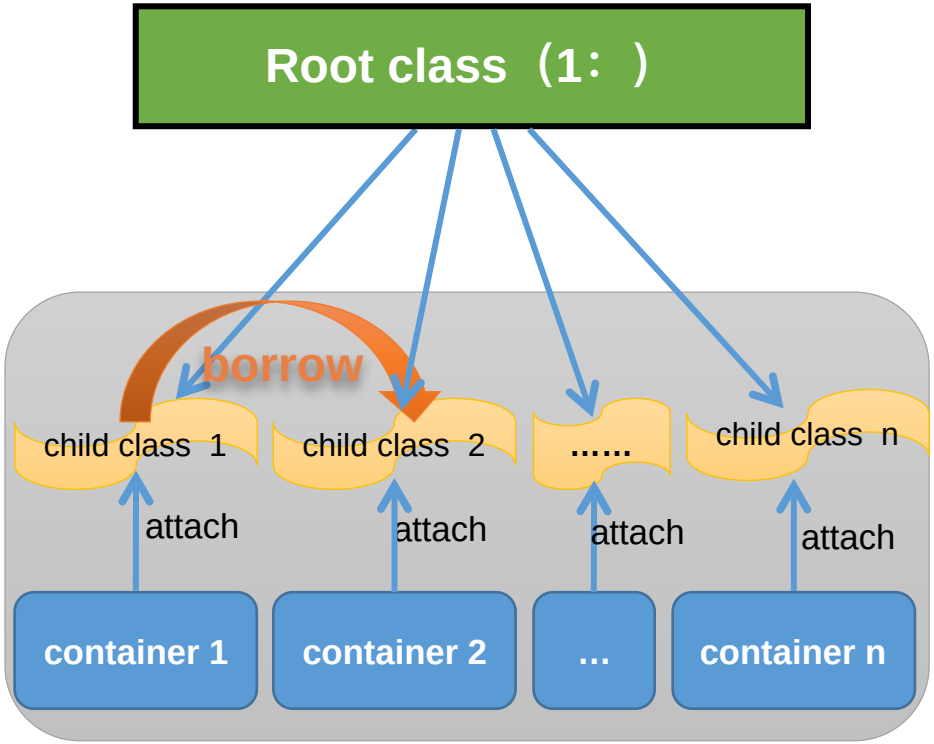
2）可以容纳更多container：可按照平均值分配container。

3） 作业失败率大大降低：container之间是softlimit机制。

4） 对用户资源评估能力要求降低


30000
25000
20000
15000
10000
5000
0
mem/proc_11927
mem/proc_12063
mem/proc_12156
mem/proc_12255
total
Jan 14 15:39:31 CST 2016
Jan 14 15:39:33 CST 2016
Jan 14 15:39:35 CST 2016
Jan 14 15:39:37 CST 2016
Jan 14 15:39:39 CST 2016
Jan 14 15:39:41 CST 2016
Jan 14 15:39:43 CST 2016
Jan 14 15:39:45 CST 2016
Jan 14 15:39:47 CST 2016
Jan 14 15:39:49 CST 2016
Jan 14 15:39:51 CST 2016
Jan 14 15:39:53 CST 2016
Jan 14 15:39:55 CST 2016
Jan 14 15:39:57 CST 2016
Jan 14 15:39:59 CST 2016
Jan 14 15:40:01 CST 2016
Jan 14 15:40:03 CST 2016
Jan 14 15:40:05 CST 2016
Jan 14 15:40:07 CST 2016
Jan 14 15:40:09 CST 2016
Jan 14 15:40:11 CST 2016
Jan 14 15:40:13 CST 2016
Jan 14 15:40:15 CST 2016
Jan 14 15:40:17 CST 2016
Jan 14 15:40:19 CST 2016
Jan 14 15:40:21 CST 2016
Jan 14 15:40:23 CST 2016
Jan 14 15:40:25 CST 2016
Jan 14 15:40:27 CST 2016
Jan 14 15:40:29 CST 2016

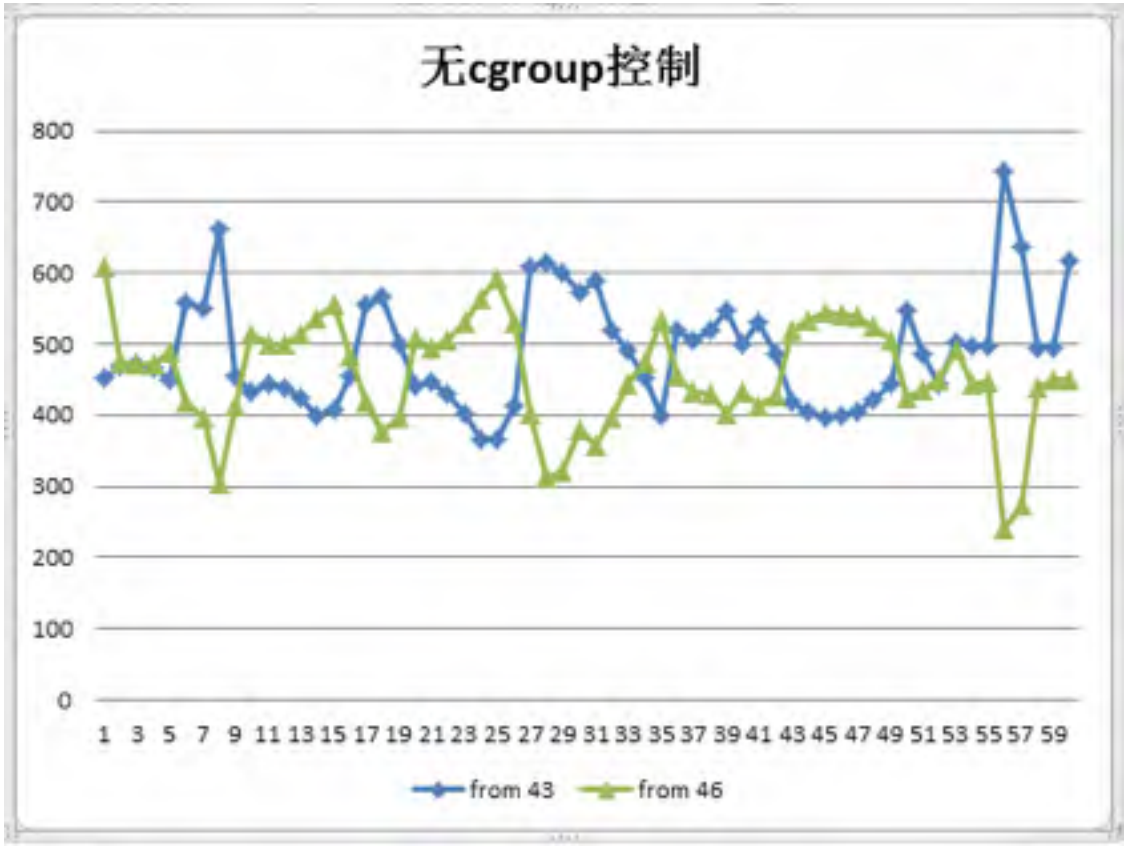

60000
50000
40000
30000
20000
10000
0
mem/proc_20814
mem/proc_20946
mem/proc_21087
mem/proc_21184
total



1）TC+cgroups相结合的方式控制。

2） container之间的网络带宽可以相互borrow，可以充分利用网络资源。

3）内核实现专门控制网络入带宽的cgroup controller，增加netrx subsystem



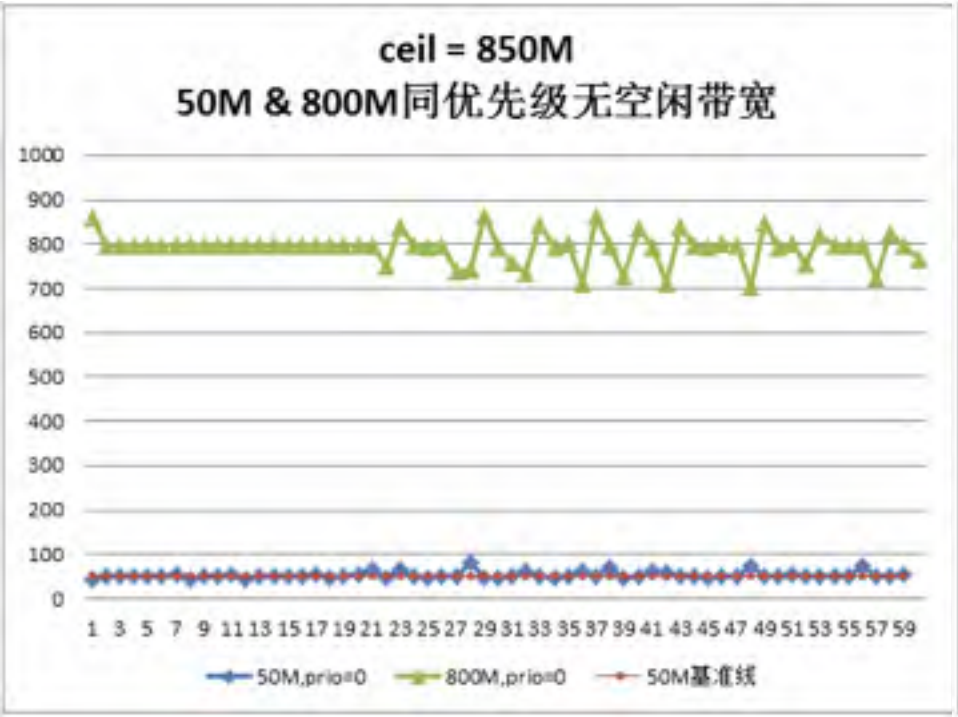
设计目标

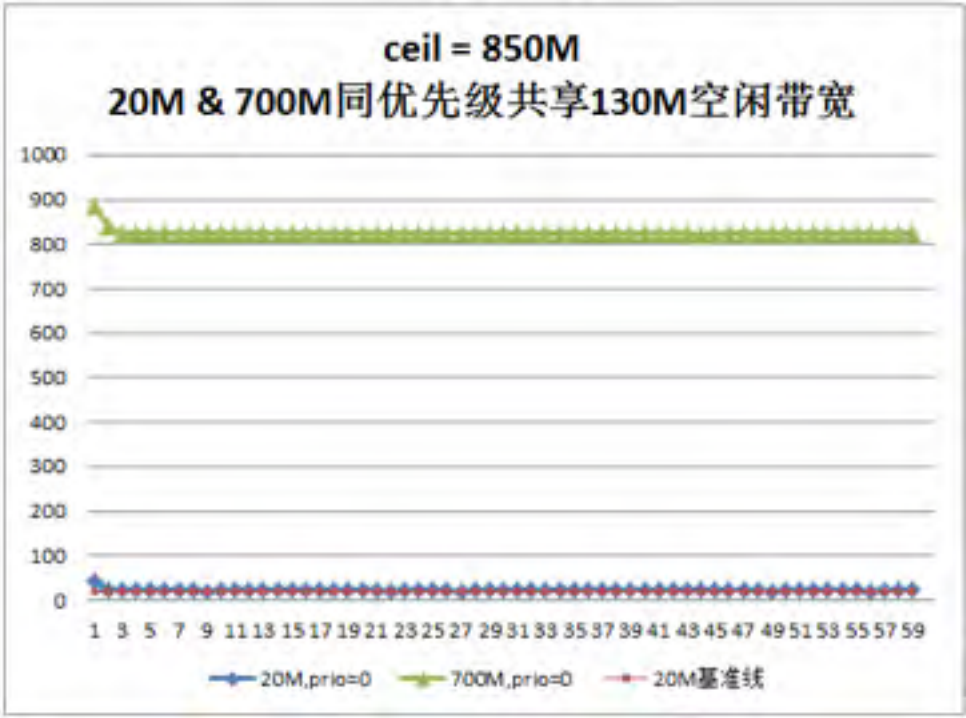
▪在某个cgroup网络繁忙时，能保证其设定配额不会被其他cgroup挤占

▪在某个cgroup没有用满其配额时，其他cgroup可以自动使用其空闲的部分带宽

▪在多个cgroup分享其他cgroup的空闲带宽时，优先级高的优先； 优先级相同时， 配额大的占用多，配额小的占用少

▪尽量减少为了流控而主动丢包





- 队列： 不增加队列， 对每个报文直接在正常代码路径上进行决策
- Cgroup区分(标记)： 在正常处理流程中，报文查找到目标socket结构之后，根据socket的owner process来确定cgroup
- 报文决策： 令牌桶 + 共享令牌池 + 显式借令牌

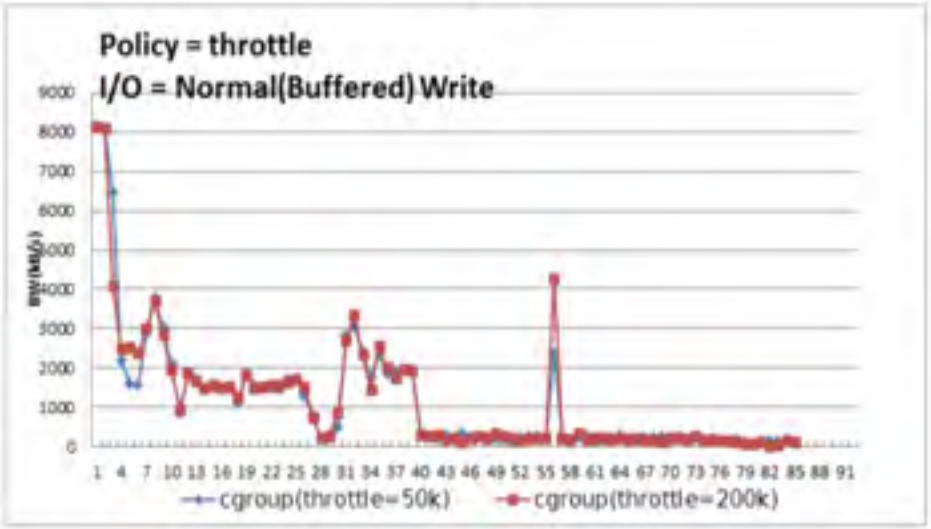
  [专利2013107167896] - 一种保证速率和充分利用空余带宽的流量调度方法
- 限速方式： ECN标记 + TCP滑窗 + 丢包

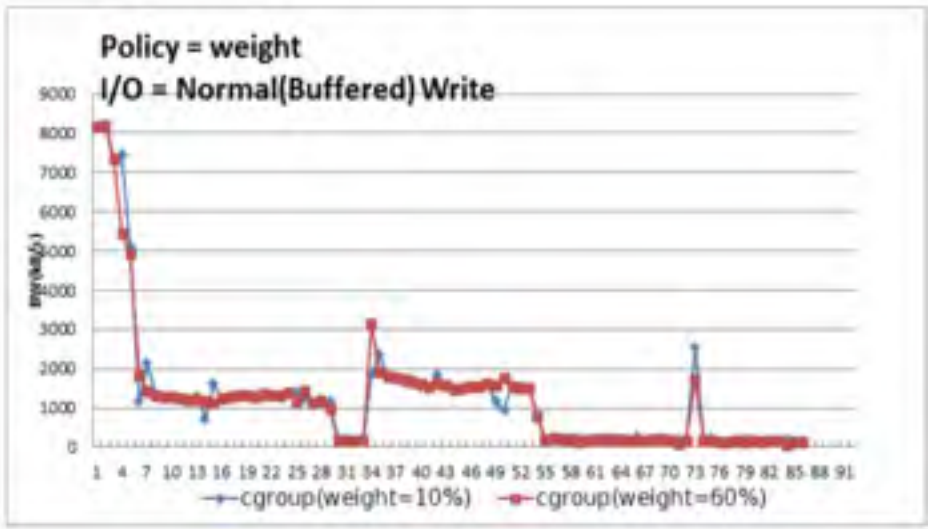
  [专利201310743471.7] - 通过接收端主机标记ECN进行网络入流量限速的方法

  [专利2013107175144] - 根据令牌桶的水位调整TCP通告窗口的网络入流量主动限速方法

对buffer io失控。cgroup通过识别pid，控制磁盘io。但在buffer io中，失去了原有的pid信息，导致不可控。

Normal Write





throttle=50k	
最小	90.0
最大	8185.0
平均	1154.1
方差	2368974.4

throttle=200k	
最小	17.0
最大	8155.0
平均	1171.8
方差	2282716.1

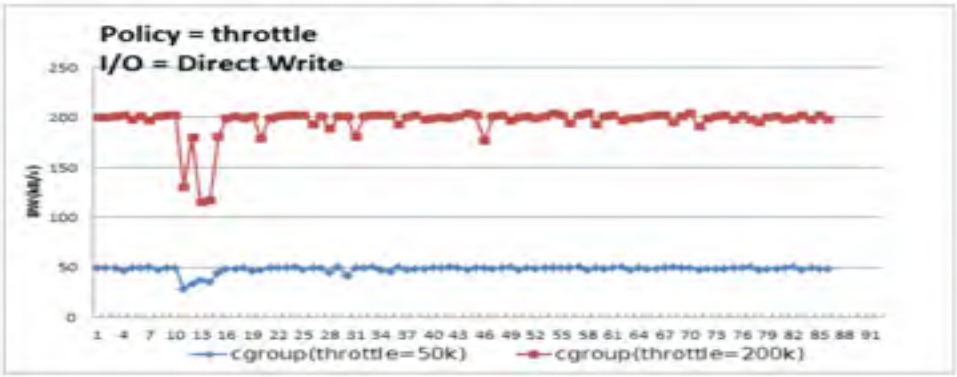
期望比例	0.2500
实际比例	0.9849

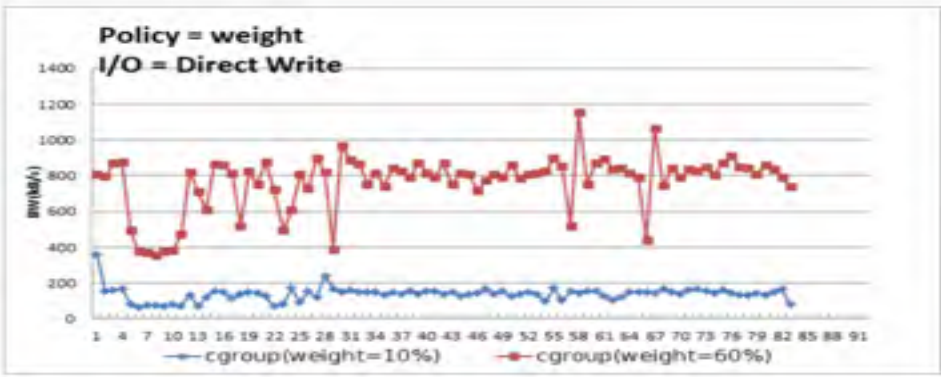
weight=10%	
最小	125.0
最大	8206.0
平均	1264.1
方差	2718006.0

weight=60%	
最小	37.0
最大	8206.0
平均	1244.0
方差	2483813.4

期望比例	0.1667
实际比例	1.0162

Direct Write

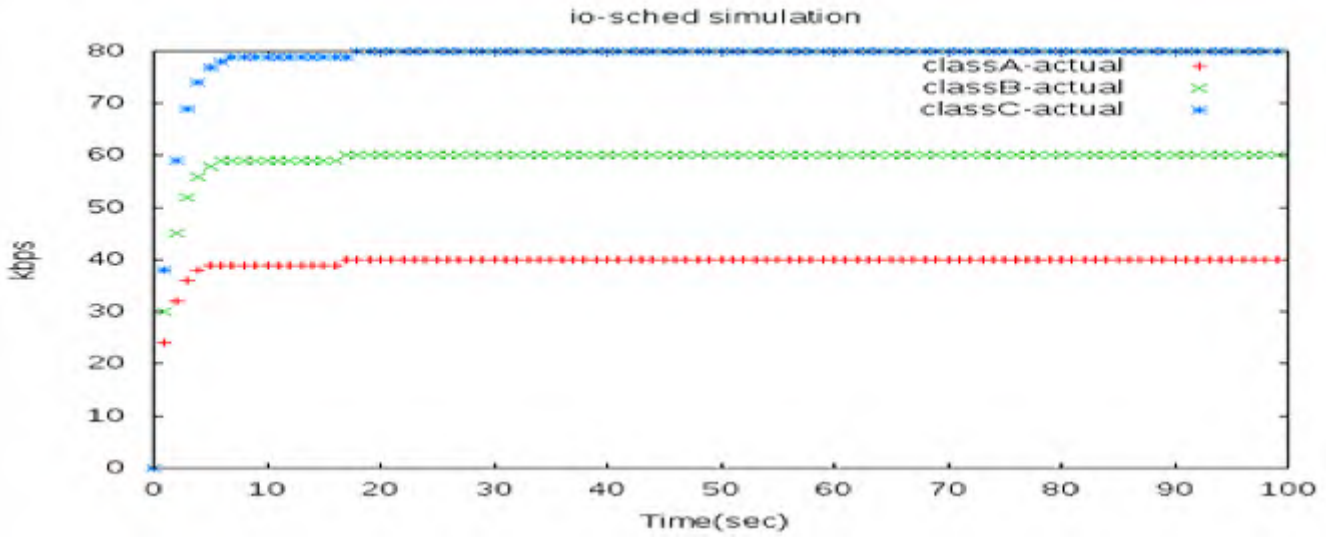




throttle=50k		throttle=200k			
最小	29.0	最小	115.0	期望比例	0.2500
最大	51.0	最大	204.0	实际比例	0.2482
平均	48.6	平均	195.9		
方差	12.7	方差	234.8		

weight=10%		weight=60%			
最小	67.0	最小	353.0	期望比例	0.1667
最大	358.0	最大	1154.0	实际比例	0.1808
平均	138.5	平均	766.2		
方差	1536.4	方差	24653.2		

三个cgroup,分别配置“保证带宽”为40, 60, 80 kB/s, 模拟磁盘的带宽为180kB/s

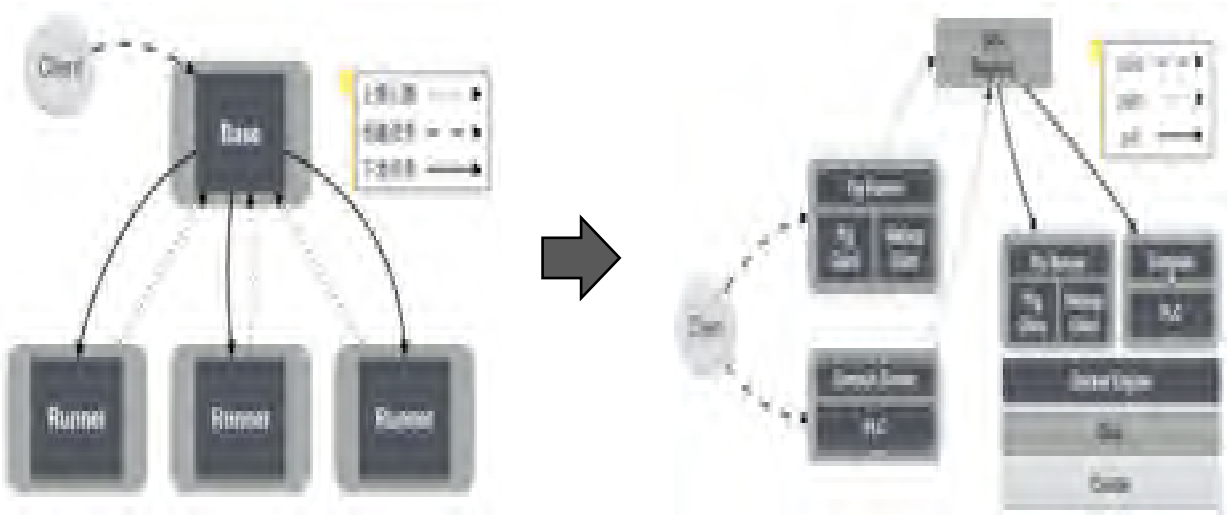


- **CPU管控**
  - Cpu share+cpuset结合管控
  - NM和DN进程纳入container管理
- **容器中资源显示问题**
  - 通过FUSE实现用户态的文件系统
  - 使用cgroup的数据统计container实际资源使用生成仿真的meminfo、stats、diskstats、cpuinfo等文件，并绑定mount到container中
- **Container数据存储**
  - 使用hostvolume存储不需要保留的数据；
  - 使用Ceph RBD存储需要保留的数据：使用Ceph volume plugin 为每个container分配一个RBD存储目录
- **Docker Registry改造**
  - 基于hdfs的存储，实现无限容量
  - 基于tpg修改registry为无状态的
  - 多registry server的负载均衡
- **Docker热升级功能**

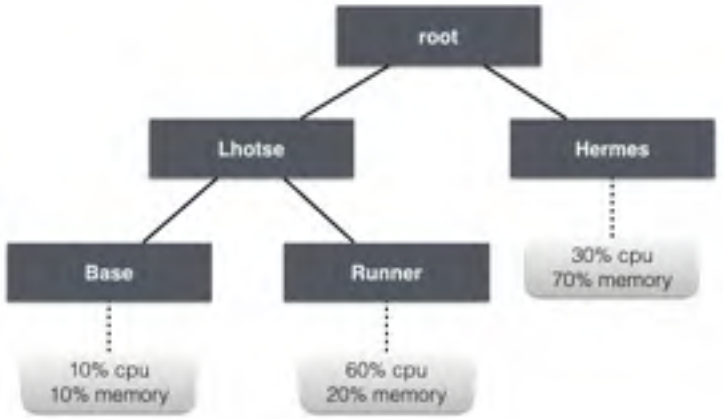
任务调度云化

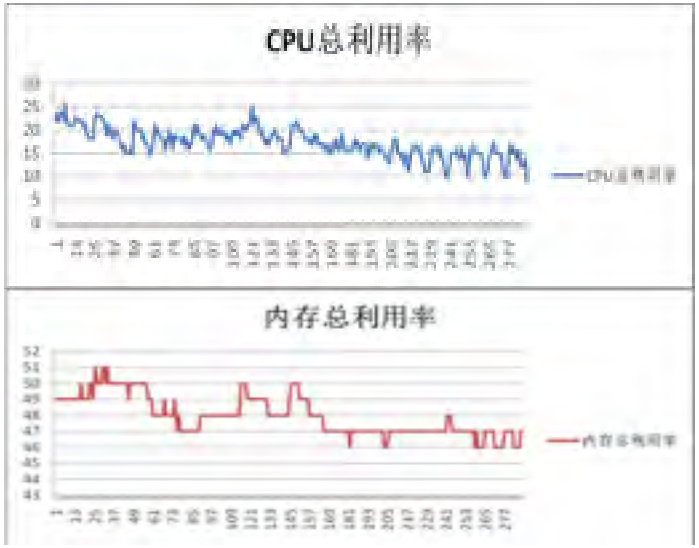
Part 03

云化收益一



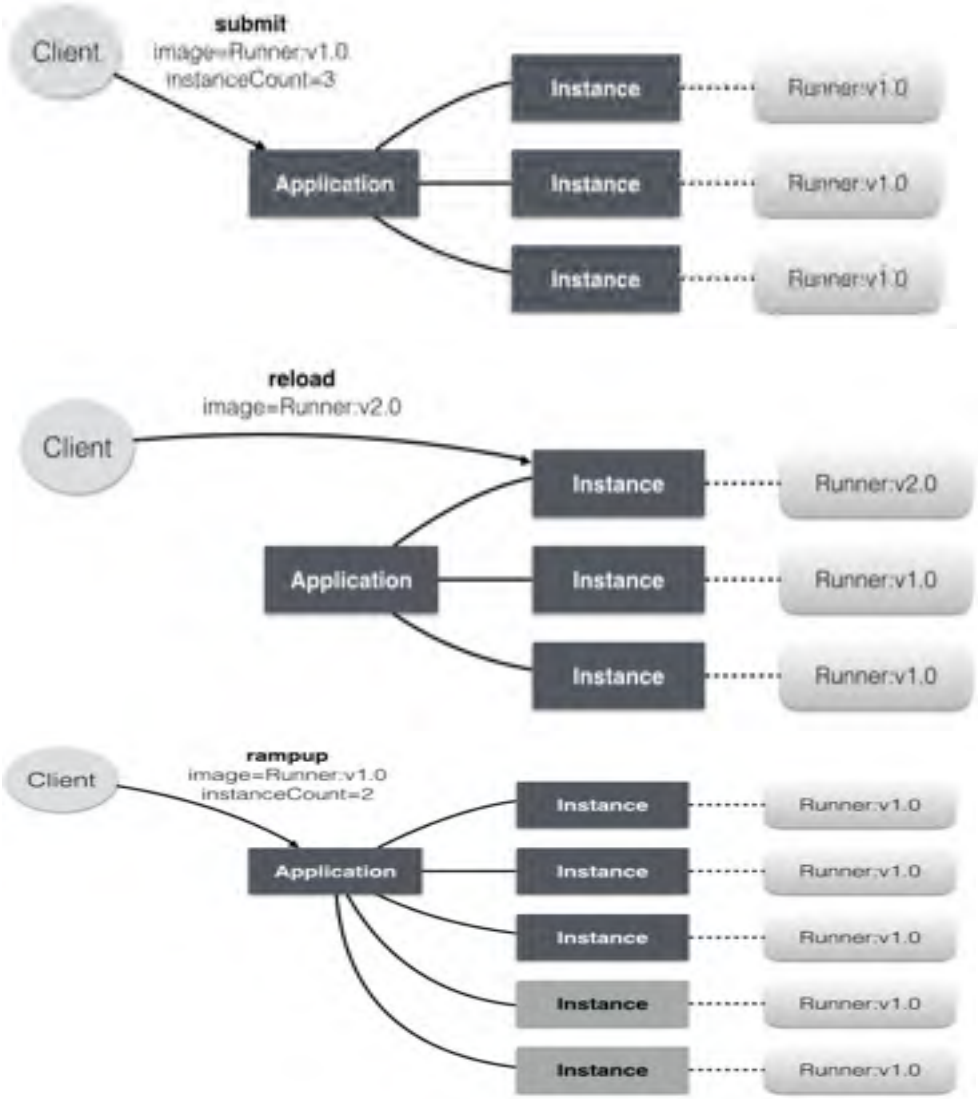
- 环境一致
  - 运行环境镜像化
- 环境隔离
  - 容器间运行环境隔离
- 版本管理
  - 通过构建新镜像升级
- 快速部署
  - Push image





横坐标表示Runner类型编号，纵坐标表示利用率数值

- 调度
  - 粗粒度：
    机器级别资源分配
  - 细粒度：
    Cpu、内存等
- 容错
  - 单机容错：
    - 监控脚本
    - 本地重试
  - 集群容错：
    - 自动重试
    - 自动屏蔽



- runner部署抽象
  - Base-> Application
  - Runner -> Instance
- 灰度升级
  1. push
  2. reload
- 扩缩容
  1. 重置实例个数
  2. Rampup Application

THANK YOU

BDTC 2016中国大数据技术大会
Big Data Technology Conference 2016