# ArchSummit全球架构师峰会北京站2015

## 爱奇艺基于容器技术的大规模弹性转码平台

# Geekbang.
# 极客邦科技

整合全球最优质学习资源，帮助技术人和企业成长

Growing Technicians, Growing Companies

**InfoQ**

专注中高端技术人员的
技术媒体

**EGO** EXTRA GEEKS' ORGANIZATION NETWORKS

高端技术人员
学习型社交网络

**StuQ**

实践驱动的
IT职业学习和服务平台

**GiT** GEEKBANG INTERNATIONAL TRAINING 极客邦培训

一线专家驱动的
企业培训服务

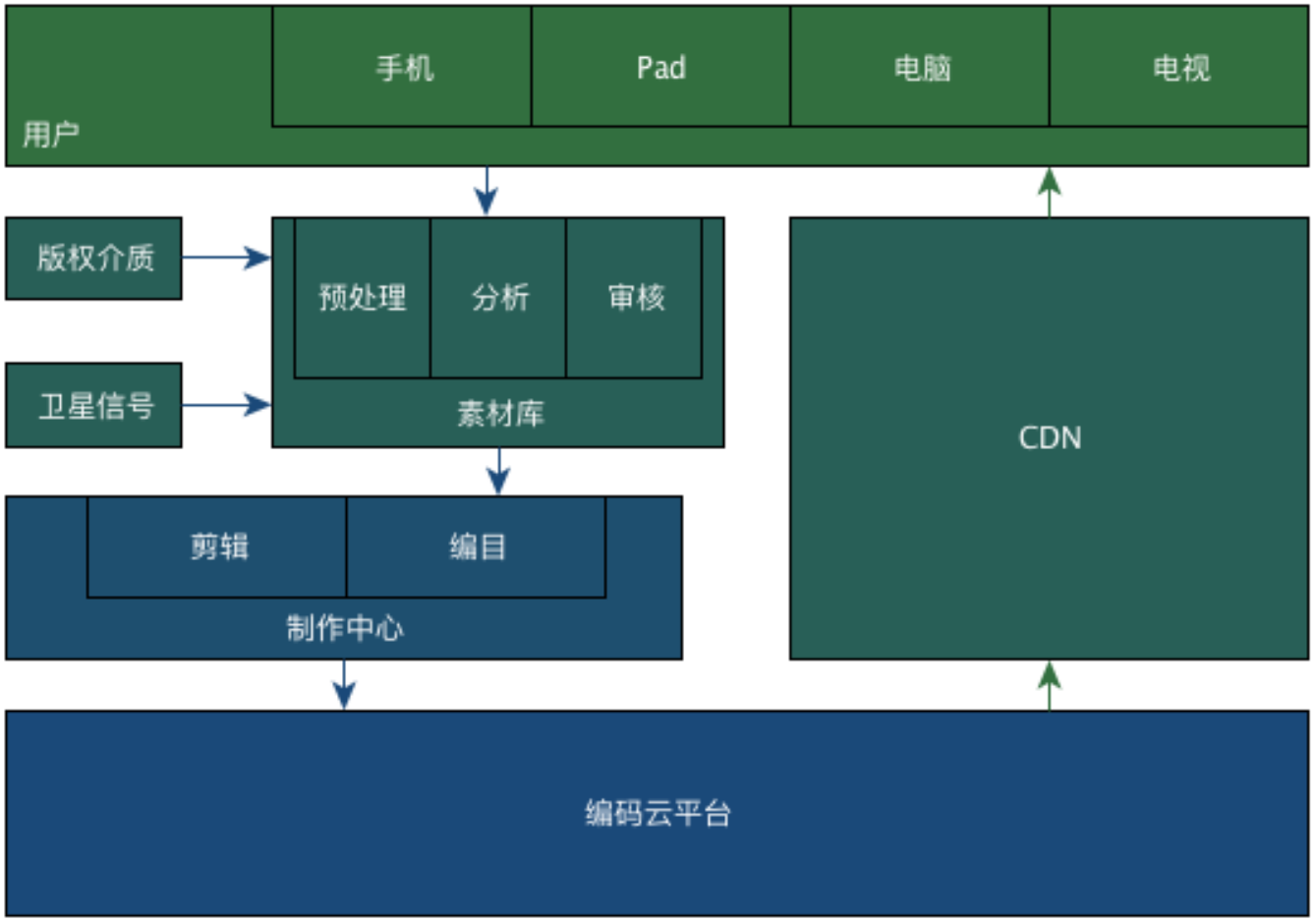
地址：北京市朝阳区洛娃大厦C座6层1607室　　网址：www.geekbang.org

# 内容提要

- 业务场景介绍
- 技术组件介绍
- 设计与实现
- 踩过、填过的坑
- Sisyphus

# 业务场景介绍

# 视频生产流程

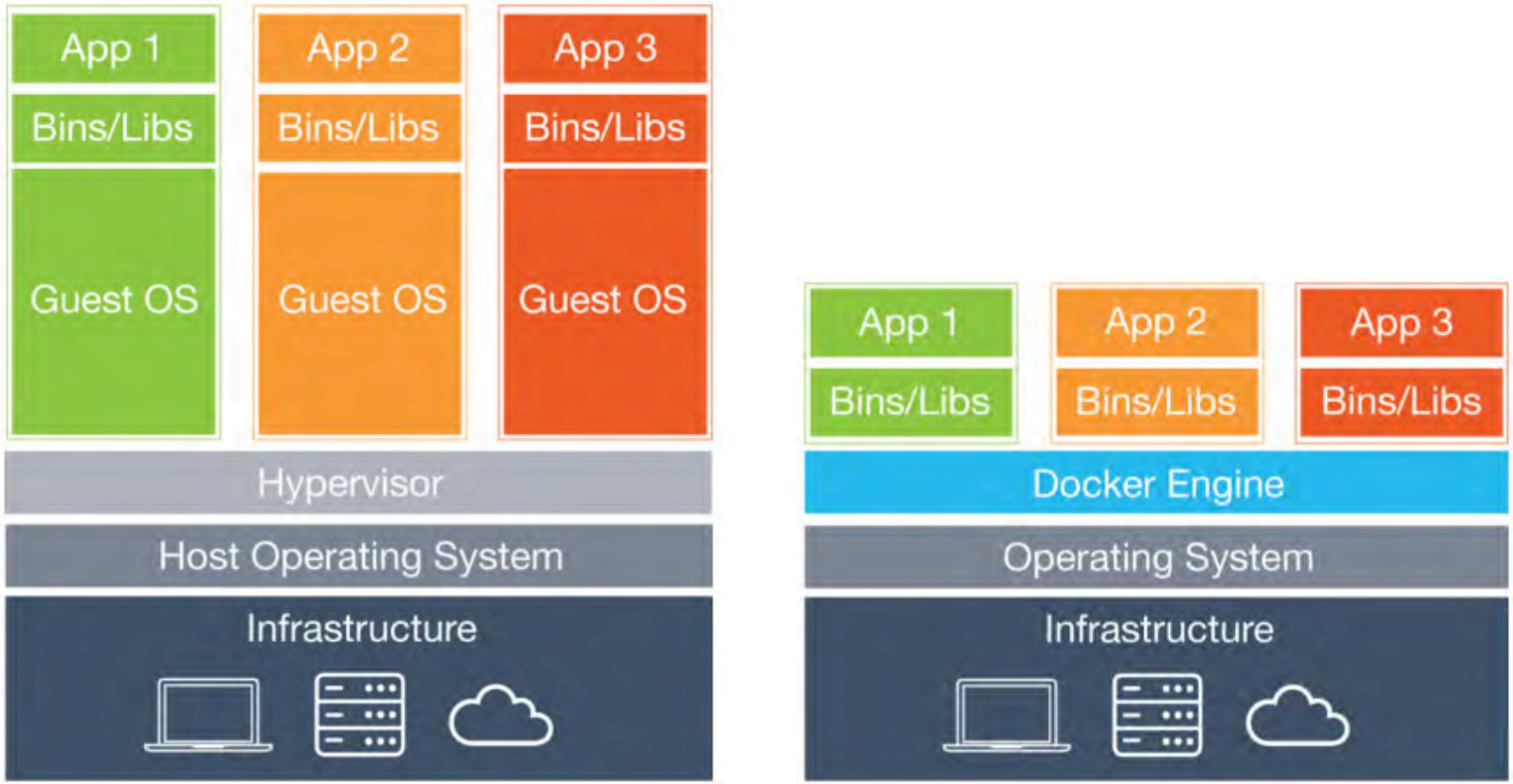


# 业务介绍

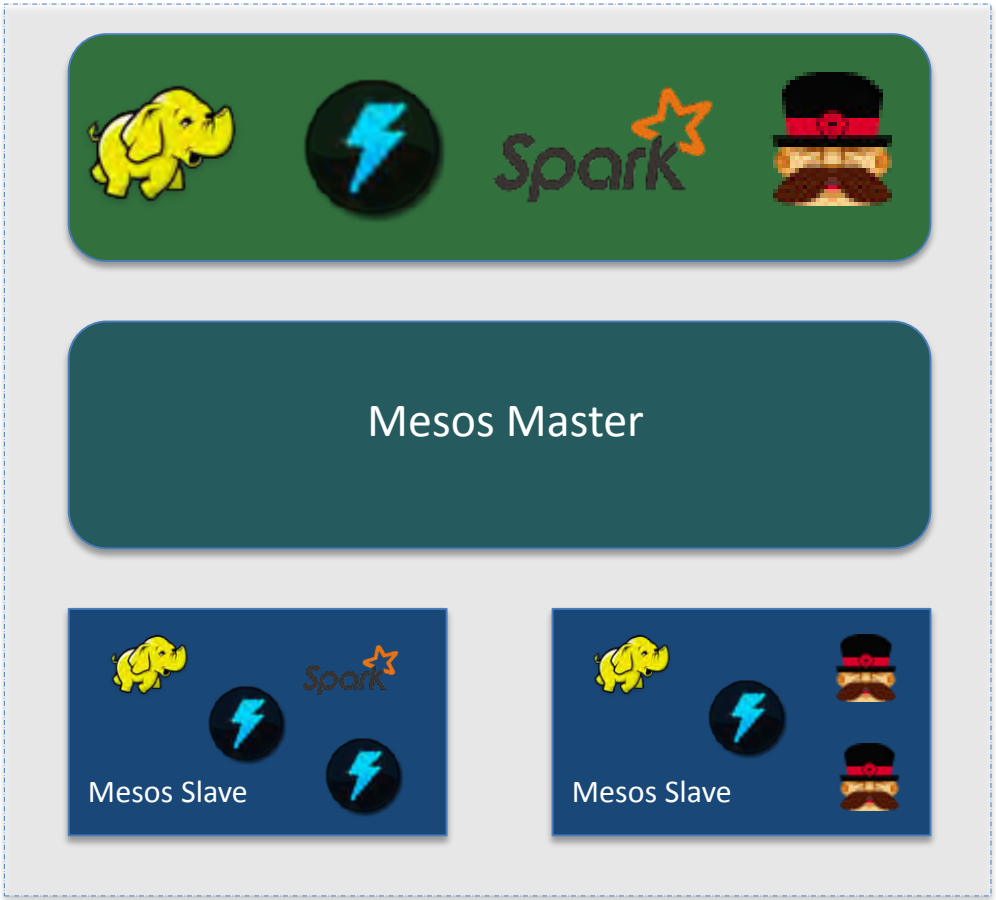
- 任务类型
  - 转码（PPC、UGC、PGC）
  - 视频爬虫
  - 花屏检测
  - 截图
- 任务特点
  - CPU, Network Bound
  - 不同优先级
  - 可能实时、可能离线
  - 并发量大
  - 时长从几秒到几小时
- 频道资源预留
  - 综艺
  - 动漫
  - 电视剧
  - 电影
- 编辑人为干预
  - 《跑男 3》上线
  - 大 V 上传视频
  - 实时新闻

# 技术组件介绍

# Docker - Build, Ship, and Run Any App, Anywhere

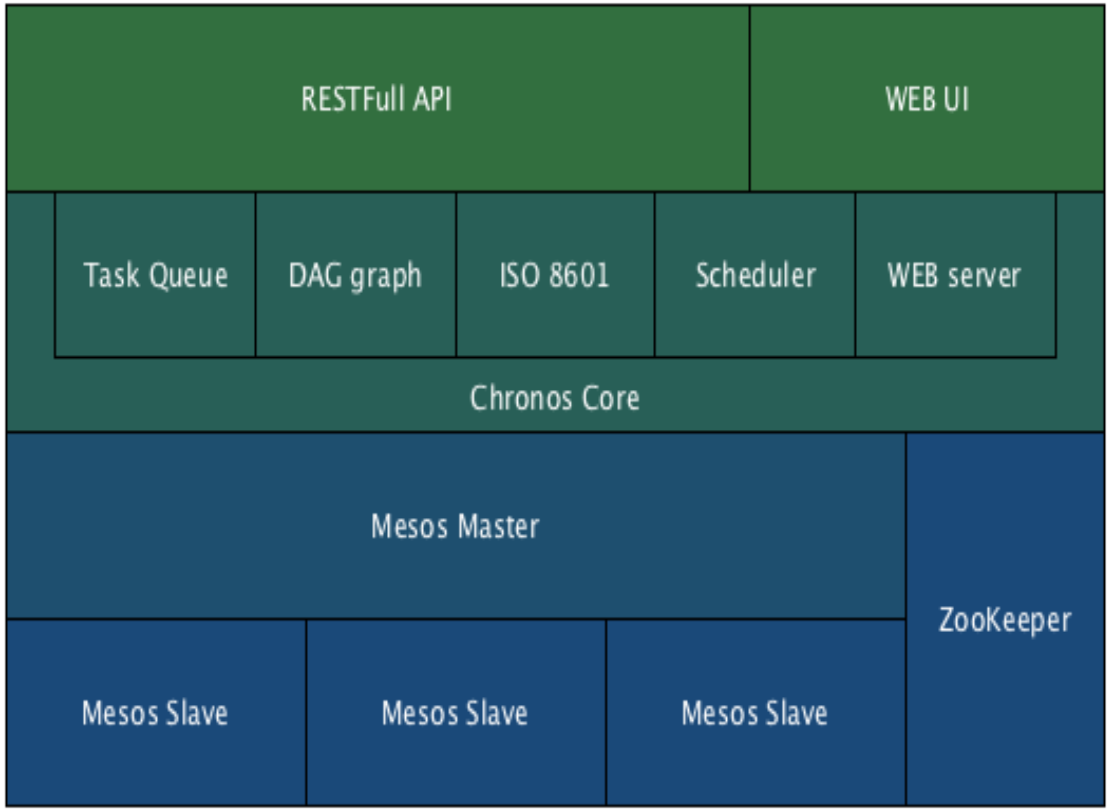


# Mesos – A distributed systems kernel




- Scalability to 10,000s of nodes
- Fault-tolerant replicated master and slaves using ZooKeeper
- Support for Docker containers
- Native isolation between tasks with Linux Containers
- Multi-resource scheduling (memory, CPU, disk, and ports)
- Java, Python and C++ APIs for developing new parallel applications
- Web UI for viewing cluster state

# Chronos - a distributed time-based job scheduler



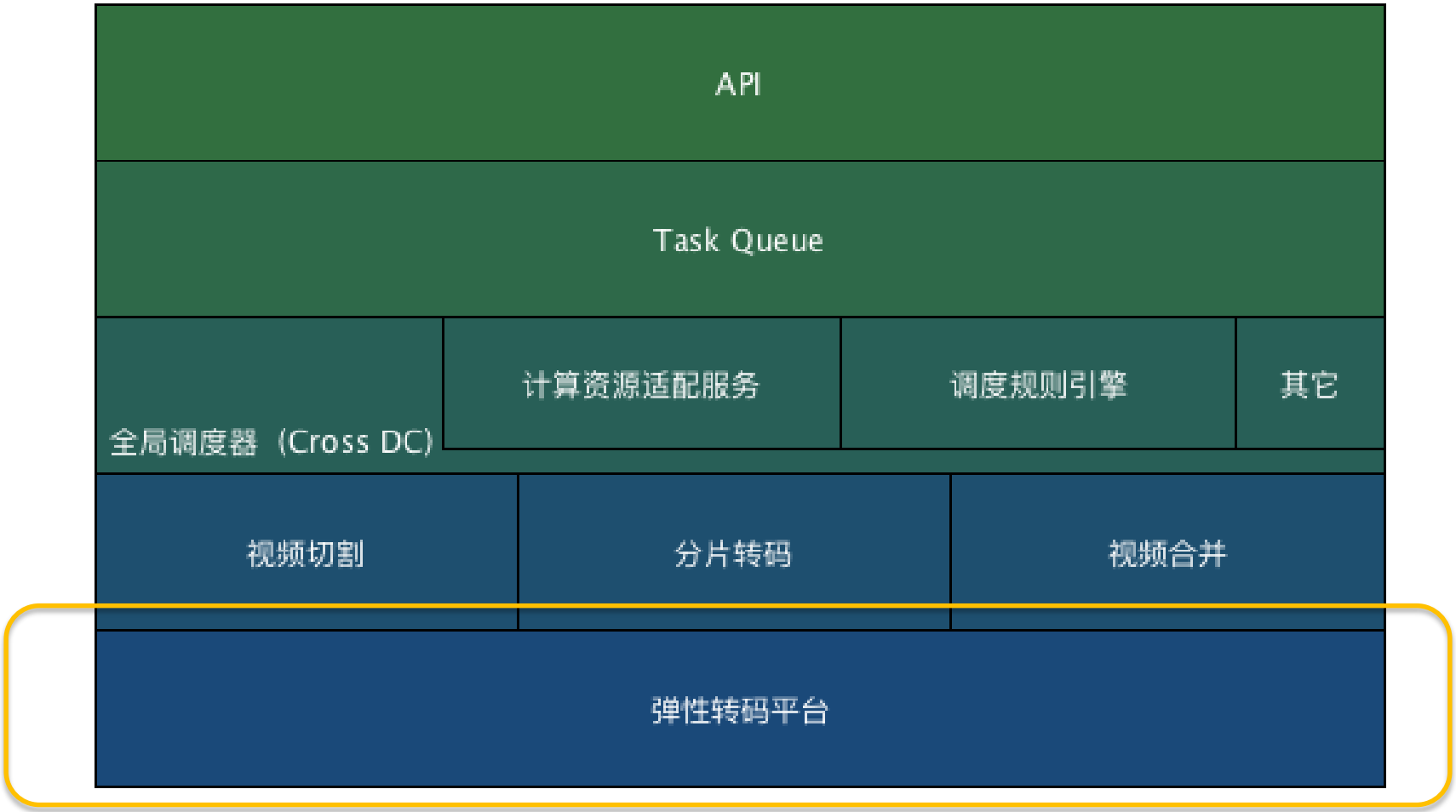

- 一个基于 Mesos 的调度器
- 基于时间的任务（ISO8601）
- 高可用性
- 作业级别的可配置失败重试
- DAG 作业依赖
- 原生 Docker 支持
- RESTFull API & WEB UI
- 作业历史和状态

# Why Mesos + Chronos + Docker?

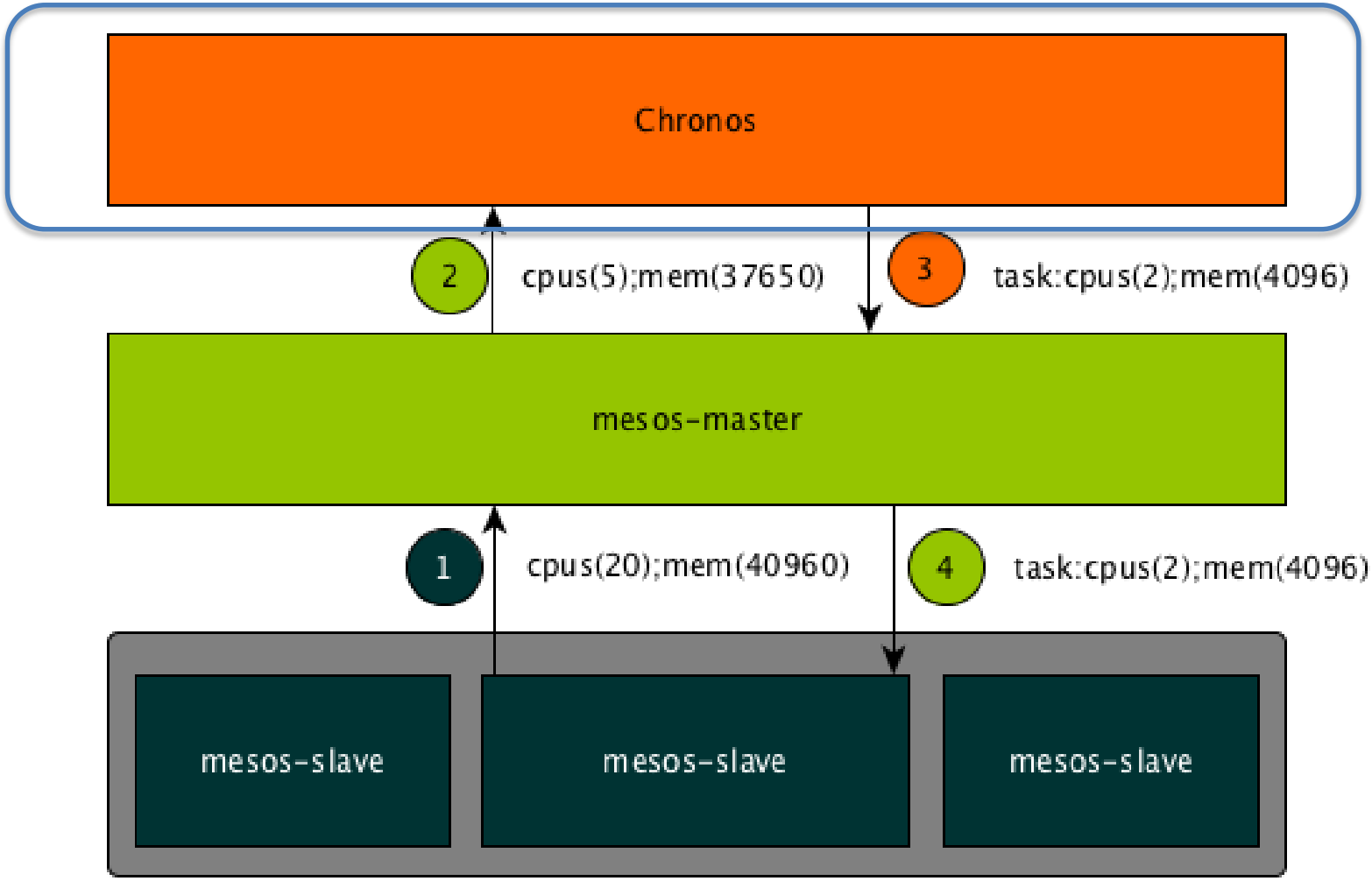
- Why Mesos?
  - 开源数据中心管理系统
  - 资源共享、提高资源综合利用率
- Why Chronos?
  - 基于 Mesos，共享资源
  - 开源，Do not Reinvent the wheel
  - 无状态、简单的高可用模型
  - RESTful API、basic web UI
  - 任务灵活可配置
- Why Docker?
  - Why not Docker?

# 设计与实现

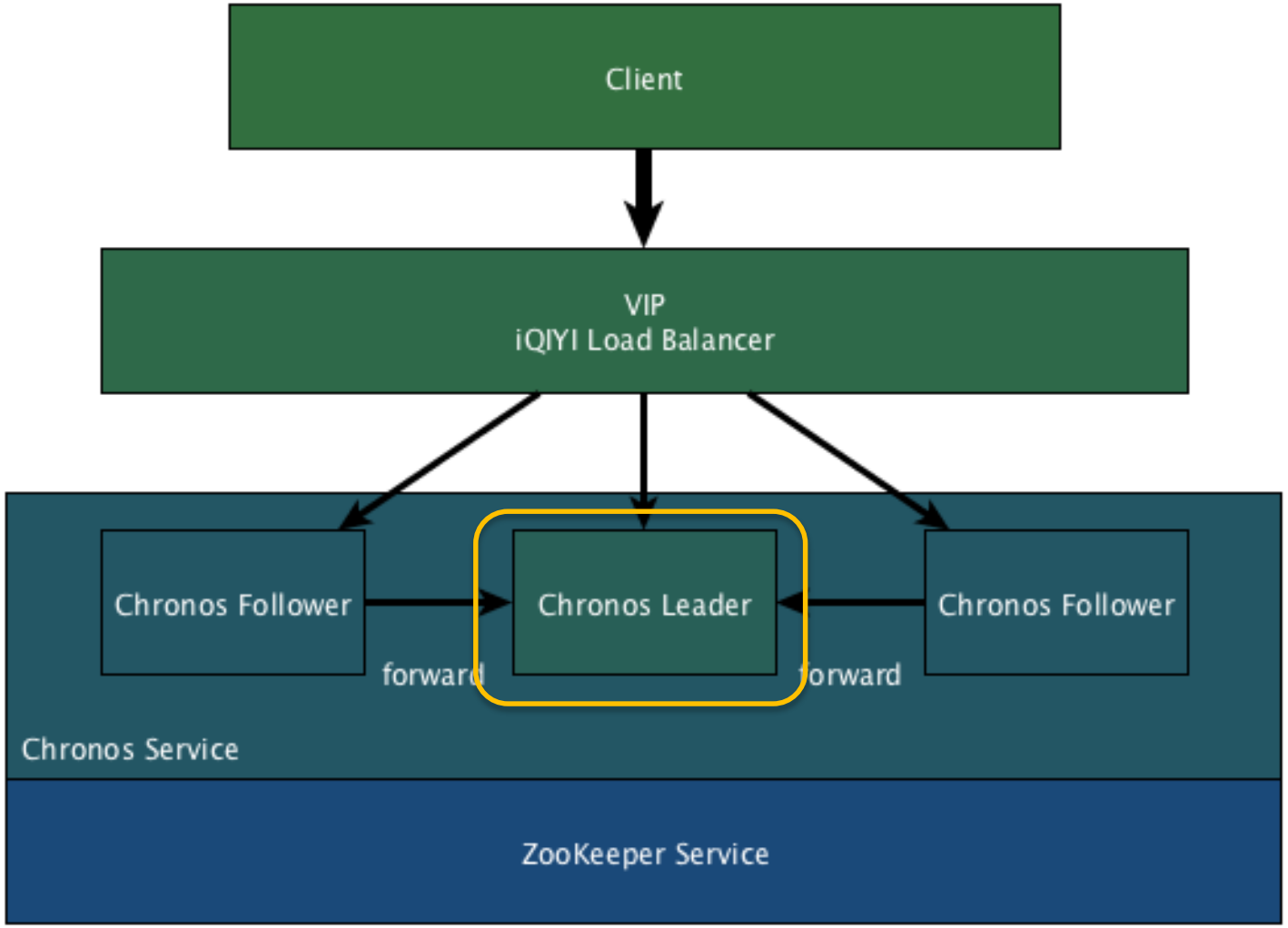
# 视频生产架构


API
Task Queue
全局调度器（Cross DC）
计算资源适配服务
调度规则引擎
其它
视频切割
分片转码
视频合并
弹性转码平台

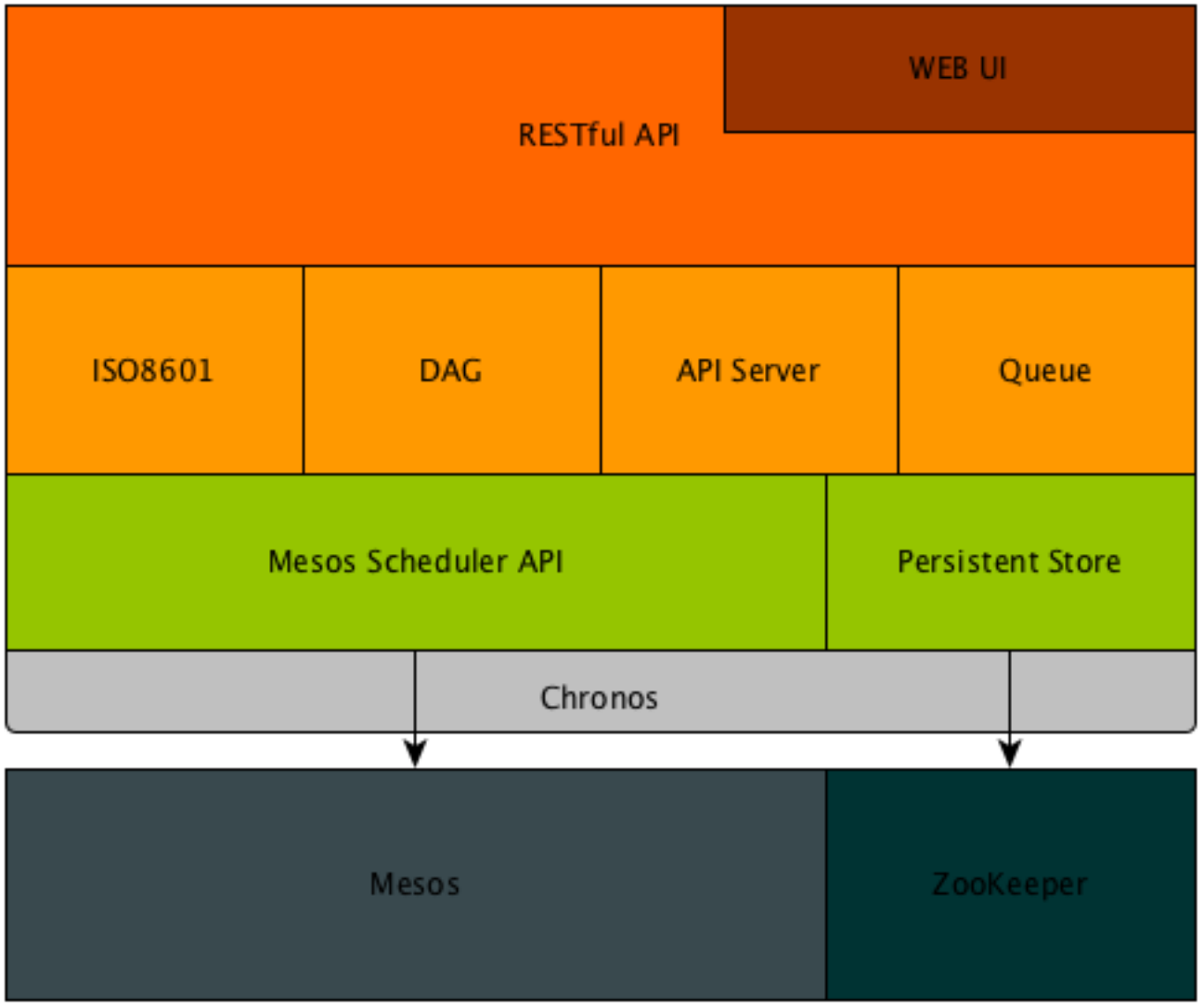
# 弹性转码平台：Mesos + Chronos



# 弹性转码平台：Chronos



# 弹性转码平台：Chronos内部



# 业务需求

- 实时新闻，大 V 上传视频，编辑干预：插队！
  - 支持优先级
- 为了保证每个频道通道的实时性，要预留资源！
  - 先用共享资源！(高级用户)
- 数据统计：跑了多少任务，失败多少，成功多少，超时多少？
  - 新 API 满足统计需求
- 爬虫任务，需要访问外网！
  - OK，支持结点过滤
- 上层调度器要统筹全局，不能总往一个 DC 打任务，打满了，排队了，怎么办？
  - 汇报资源
- 立即给我开始！
  - 添加即时任务支持

# 踩过、填过的坑

# Chronos 被 deactivate，再也收不到 Offer

## Mesos master log

I0729 13:39:55.346318 1759 master.cpp:2933] **Sending 2 offers** to framework 20140722-064549-613093130-5050-1717-0000

I0729 13:39:55.666754 1763 master.cpp:2628] **Status update TASKRUNNING** (UUID: 0303d09b-b37f-494f-93fe-362455dc3492) for task ct:1406612390227:0:Transcoding3399315169476042 of framework 20140722-064549-613093130-5050-1717-0000 from slave 20140723-113920-613093130-5050-1693-0 at slave(1)@10.15.139.38:5051 (mesos-test-dev003-bjdxt.qiyi.virtual)

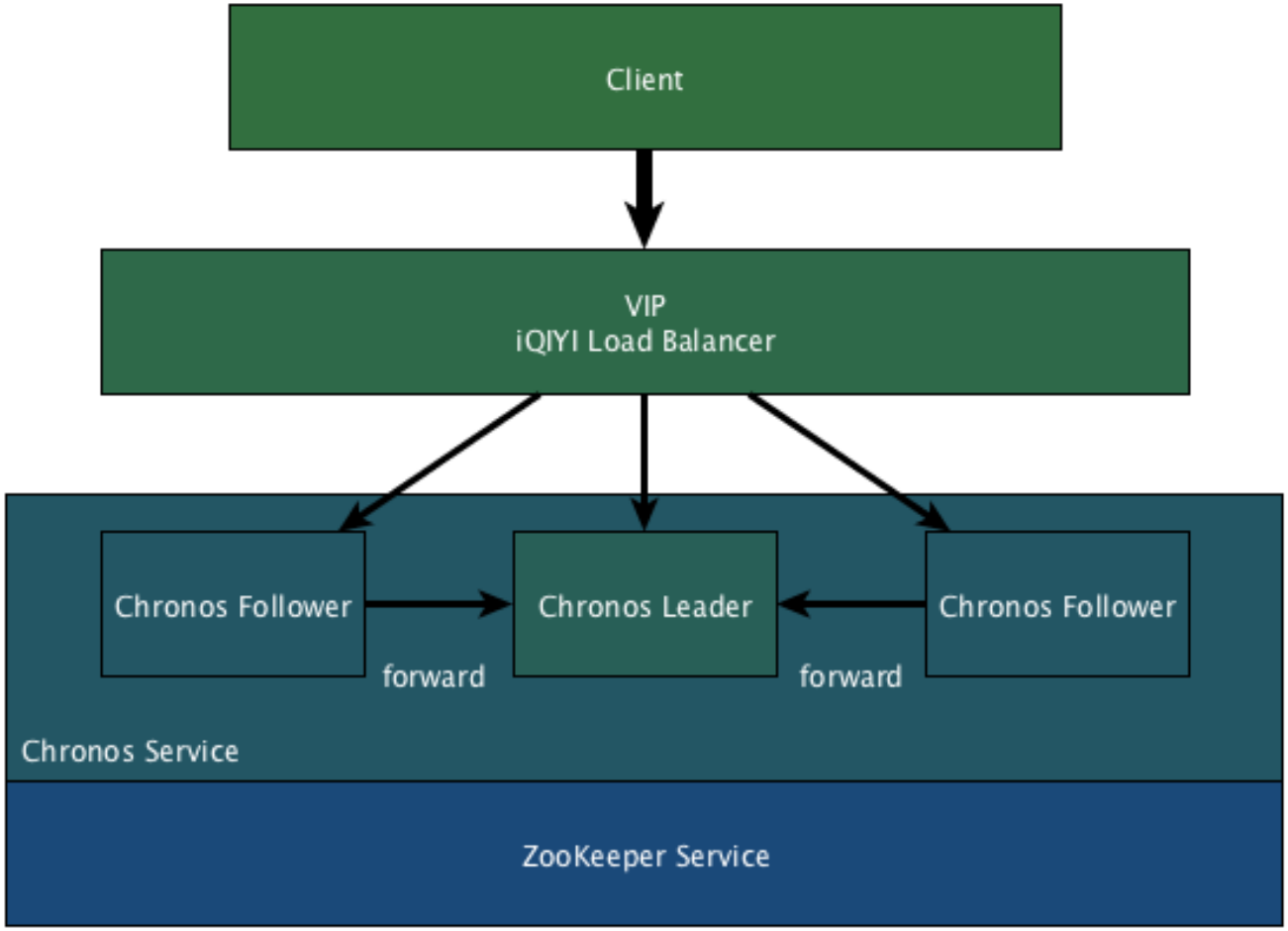
I0729 13:39:55.676386 1761 master.cpp:1319] **Deactivating** framework 20140722-064549-613093130-5050-1717-0000

I0729 13:39:55.676584 1751 hierarchicalallocatorprocess.hpp:407] **Deactivated** framework 20140722-064549-613093130-5050-1717-0000

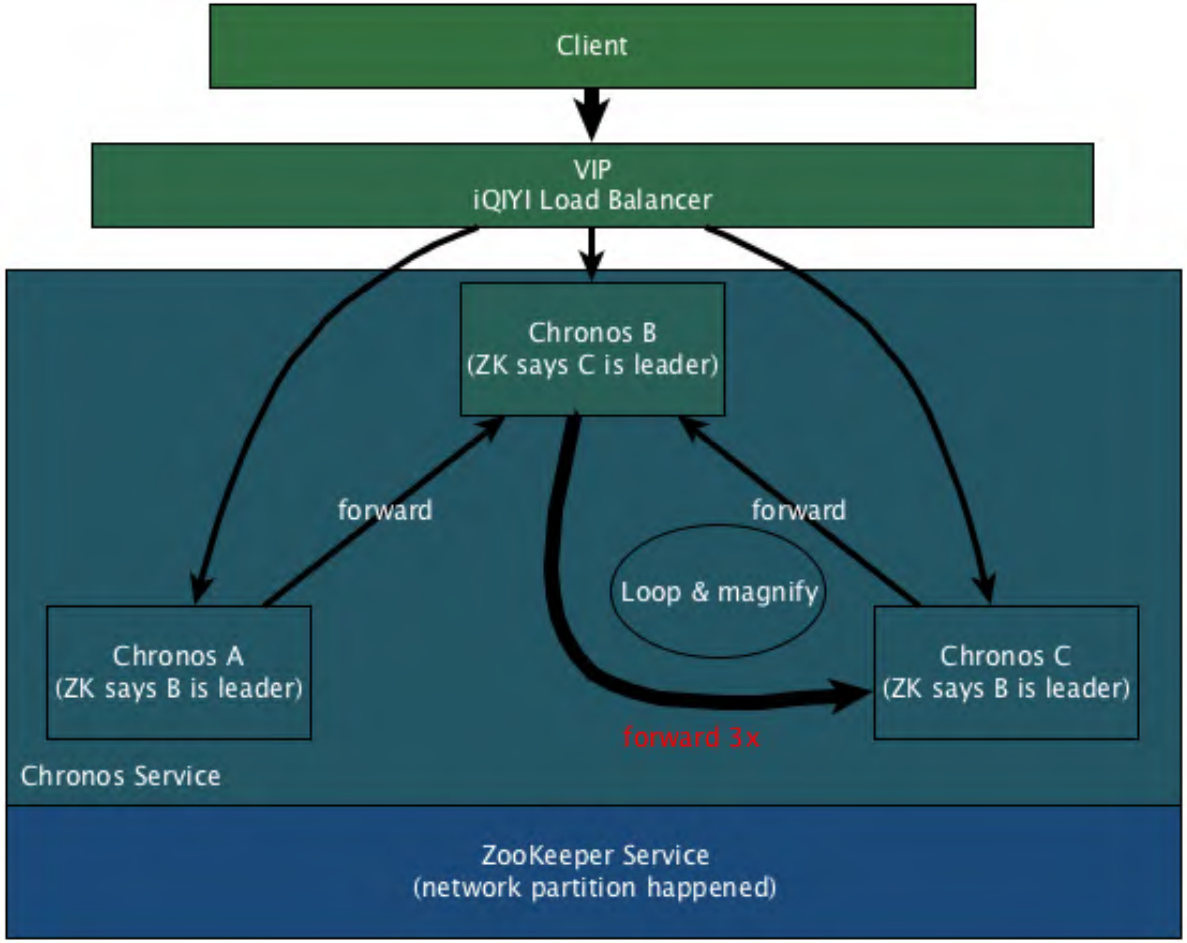
## Chronos log

I’ m hungry, haven’ t received any offer for a long time…

# Chronos 锁死，不能处理任何请求



# Chronos 锁死，不能处理任何请求（续）



# 坑、坑、坑

- Chronos 任务可能丢失
- 从 Zookeeper 加载数据失败导致不释放 offer

- 修复有限次重复任务的状态
- 修复异步任务
- 修复 API 返回码以正确显示状态
- 修复处理失败任务持久化错误
- 修复 memory, disk 资源数据类型

# 性能优化

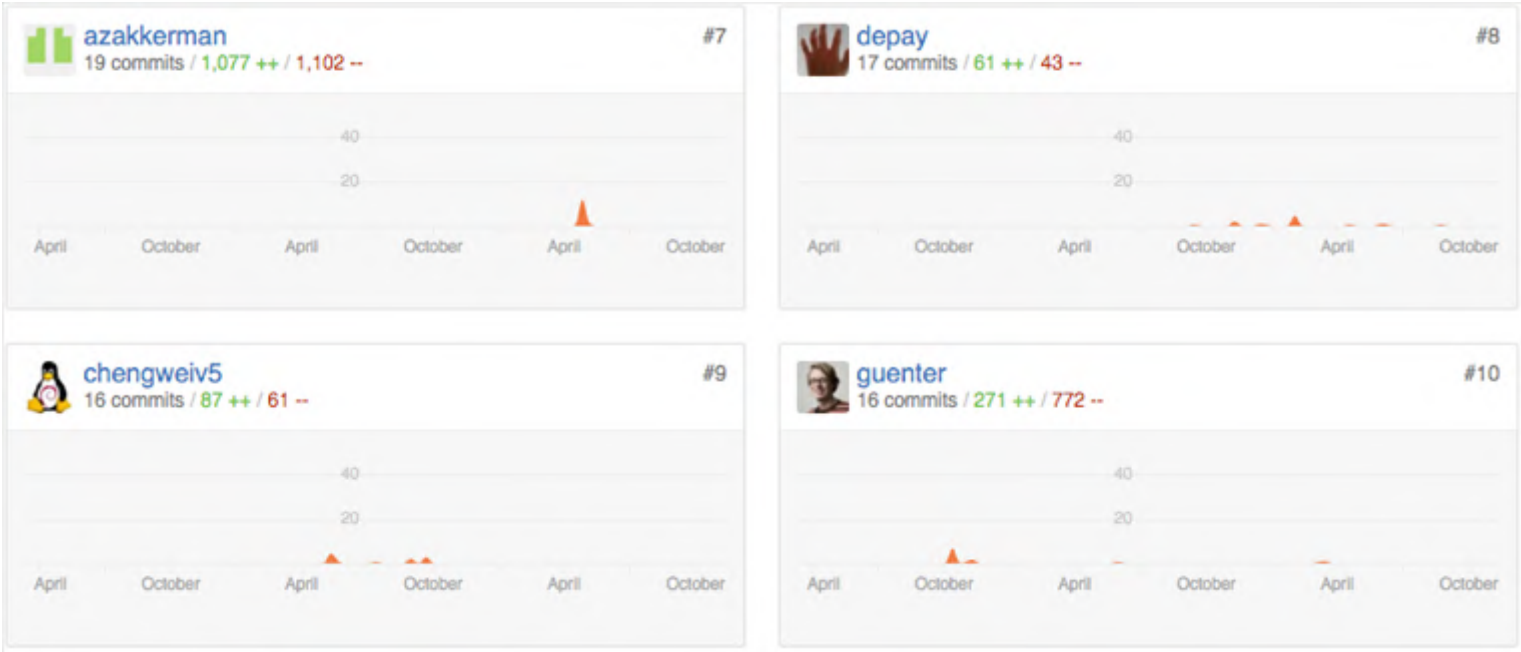
- 调度算法
  - 均衡调度，避免将多个任务调度到同一个结点
  - 避免失败作业重试时调度到以前的结点
  - 优先使用共享资源
- 任务队列优化
  - 读写分离：bank switch，消除读写锁

# 最终结果

- Chronos 并发能力提高为至少原来的 两倍
- 服务可用性 超过 99.99%
- 稳定性提高 数十倍
- 每周启动 200 万 ~ 300 万 个 Docker 容器，峰值近 350 万
- Mesos 集群峰值在线（RUNNING）任务数超过 4000
- 集群资源（CPU）利用率平均约 50%，峰值利用率超过 90%

# We Do Community

- 33 commits 被社区接受
- 2 top10 贡献者
- 所有填坑的 patch 都反馈给 upstream，在生产环境中，这些重要补丁非常重要



# Go Beyond

# Sisyphus – A Mesos batch job framework

- Fork from Chronos
- No DAG dependency jobs
- No ISO8601 jobs
- Only support immediately jobs
- More simpler data structure
- Merge Job & Task into one
- Twice performance as Chronos
- Stable than ever

# Sisyphus – future plan

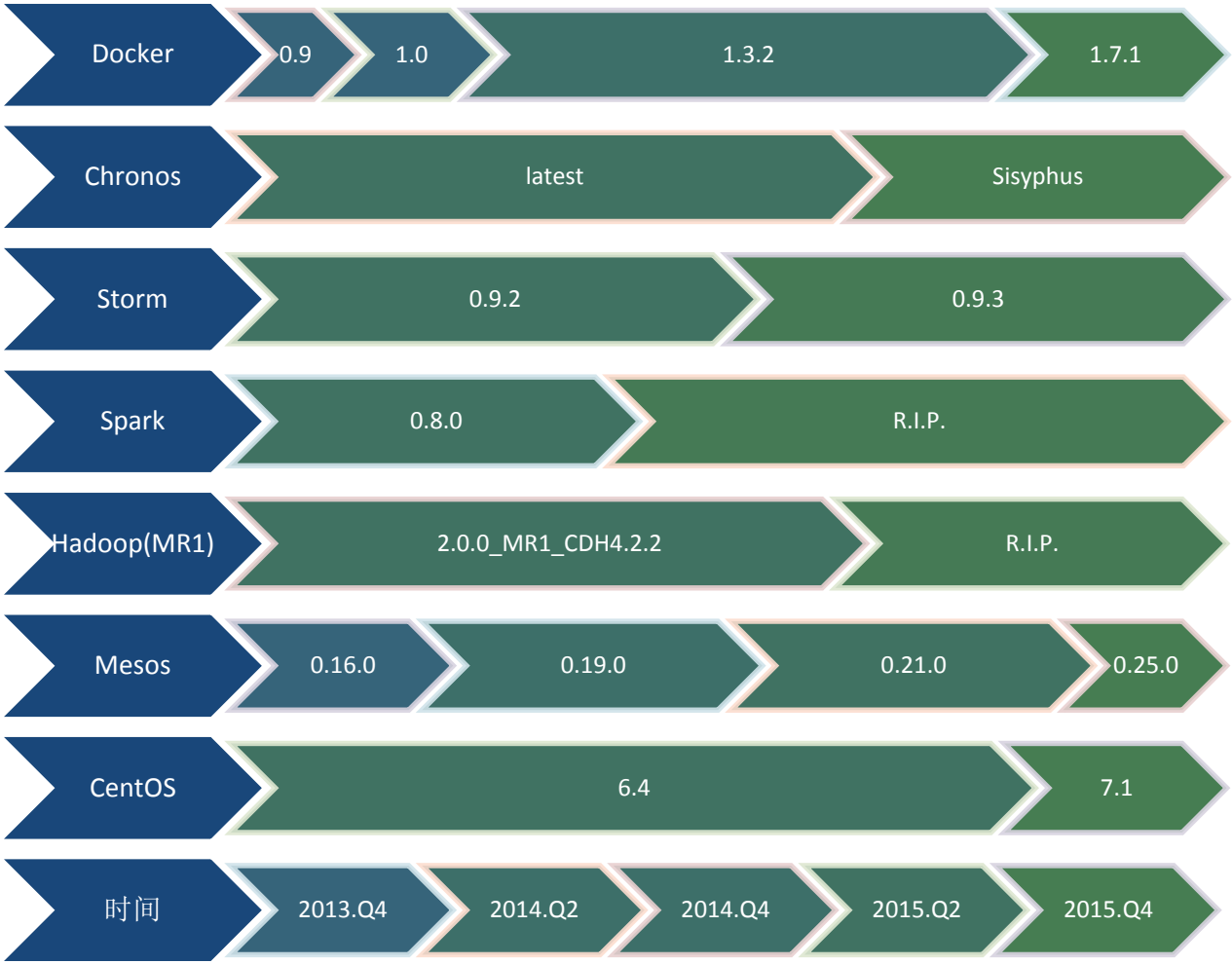
- Useful metrics
- Fancy WEB UI
- Job preemption
- and etc.
- Planed to OPEN SOURCE

# Q & A

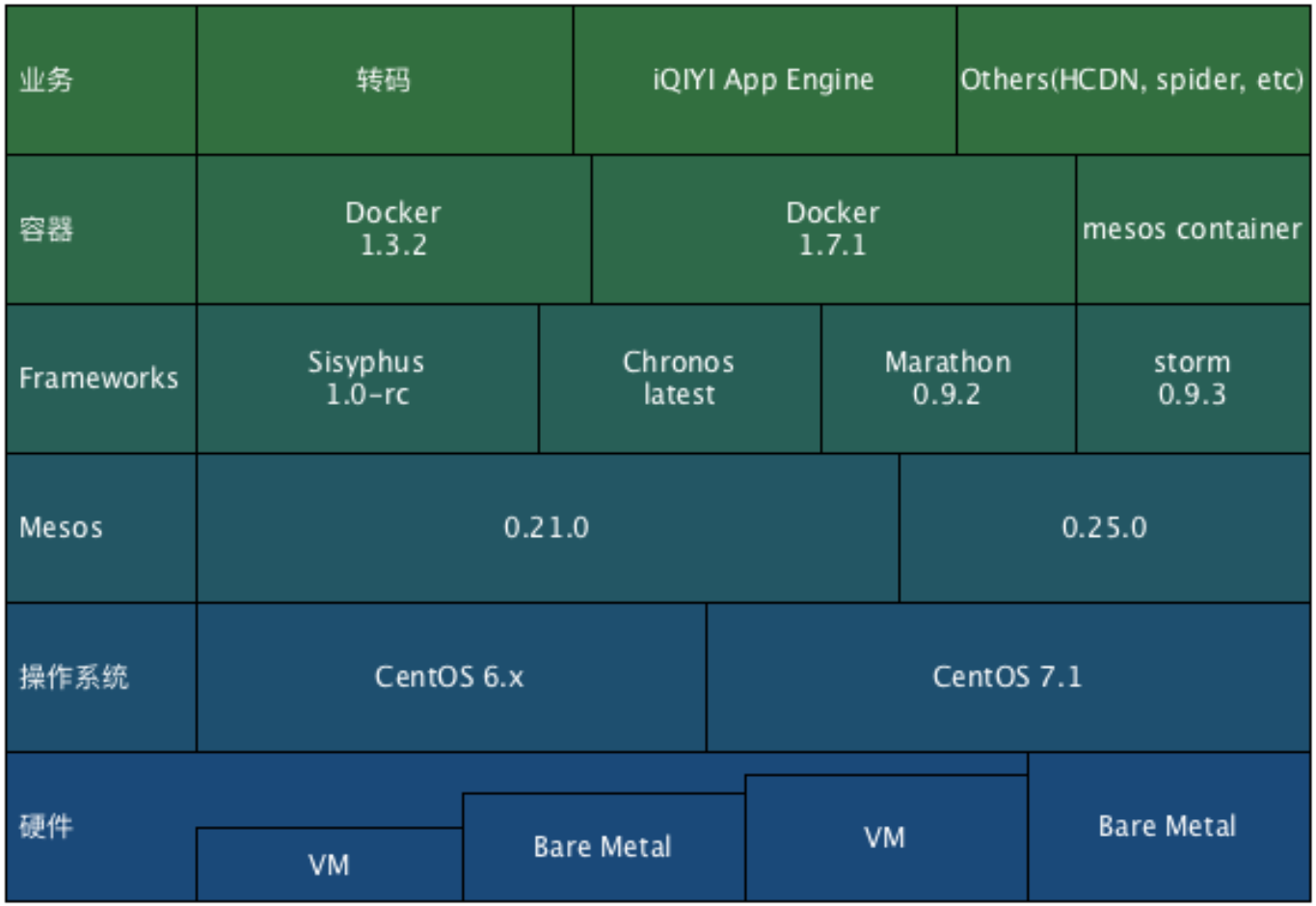
欢迎加入我们！ yuke@qiyi.com

# Thanks!

# The history(backup)



# Now(backup)


业务
转码
iQIYI App Engine
Others(HCDN, spider, etc)
容器
Docker 1.3.2
Docker 1.7.1
mesos container
Frameworks
Sisyphus 1.0-rc
Chronos latest
Marathon 0.9.2
storm 0.9.3
Mesos
0.21.0
0.25.0
操作系统
CentOS 6.x
CentOS 7.1
硬件
VM
Bare Metal
VM
Bare Metal