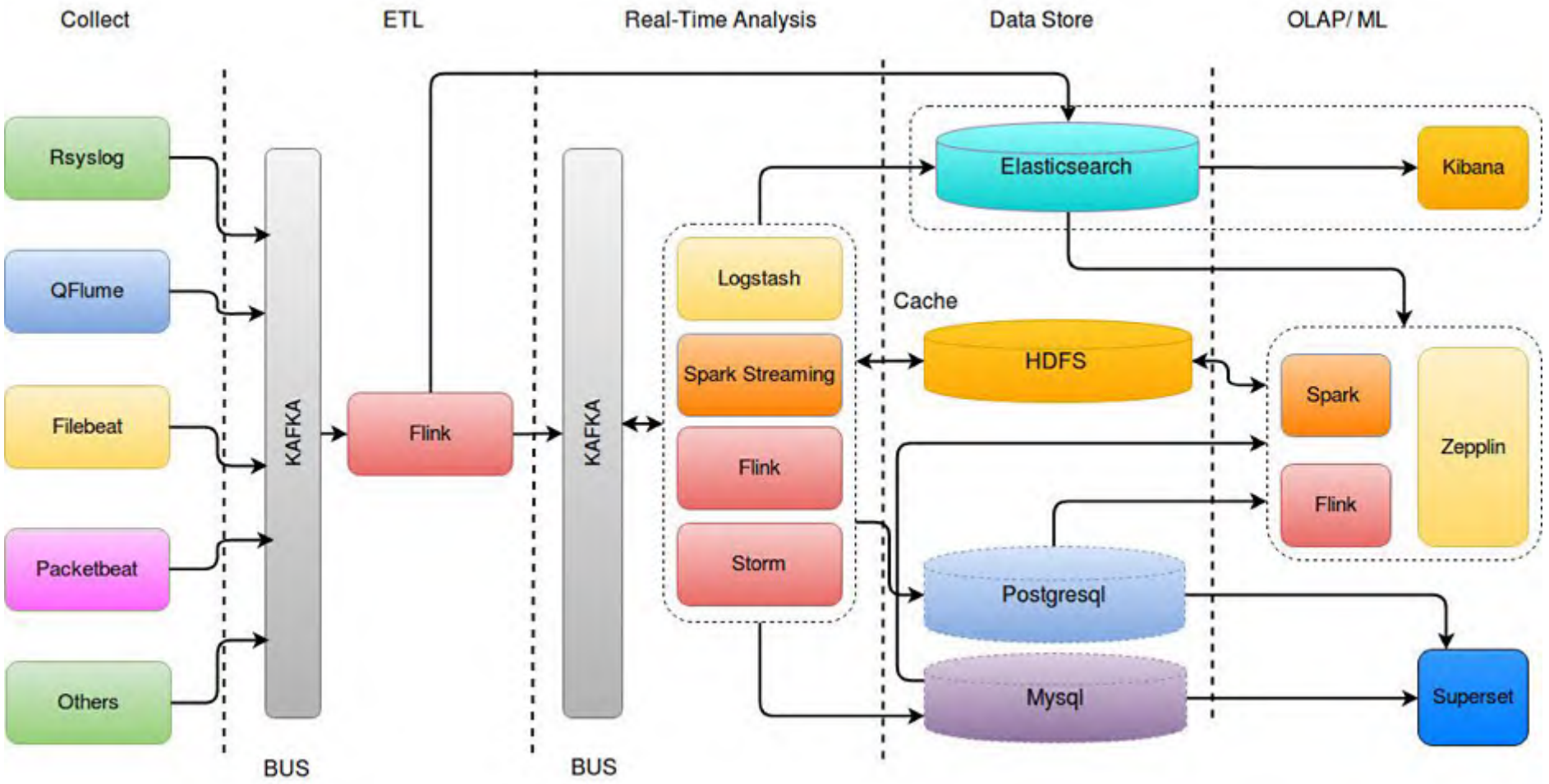
基于Mesos/Docker构建数据处理平台

- 平台概览
- 为什么选择Docker/Mesos
- 组件容器化与部署
- 基于Marathon的Streaming调度
- ELK on Mesos
- 监控与运维

平台概览

平台架构

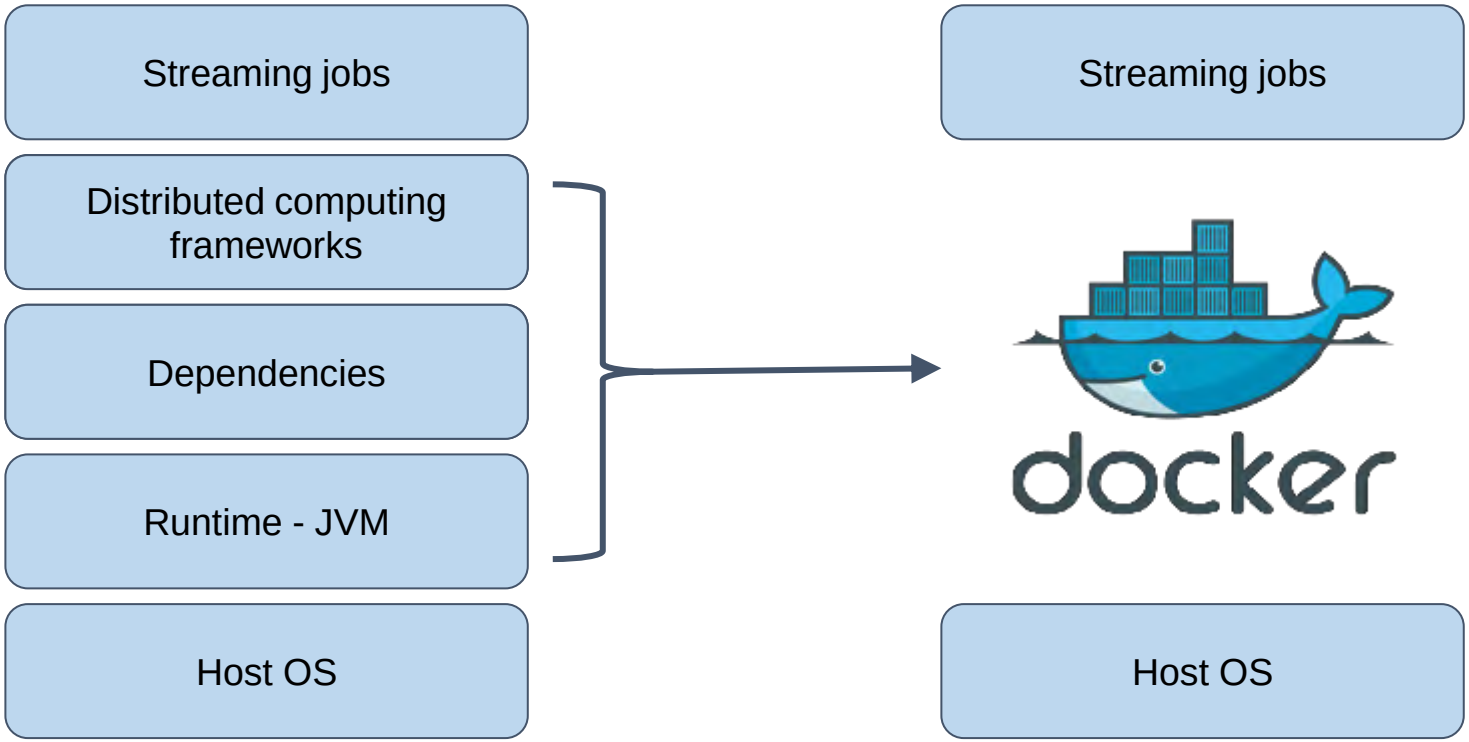

Collect
ETL
Real-Time Analysis
Data Store
OLAP/ ML
Rsyslog
QFlume
Filebeat
Packetbeat
Others
KAFKA
BUS
Flink
KAFKA
BUS
Logstash
Spark Streaming
Flink
Storm
Elasticsearch
Kibana
Cache
HDFS
Spark
Flink
Zepplin
Postgresql
Mysql
Superset

- 每天处理约340亿/25TB的数据
- 90%的数据在100ms内完成处理
- 最长3h/24h的数据回放
- 私有的Elasticsearch Cloud
- 自动化监控与报警

为什么选择Docker/Mesos

为什么选择Docker

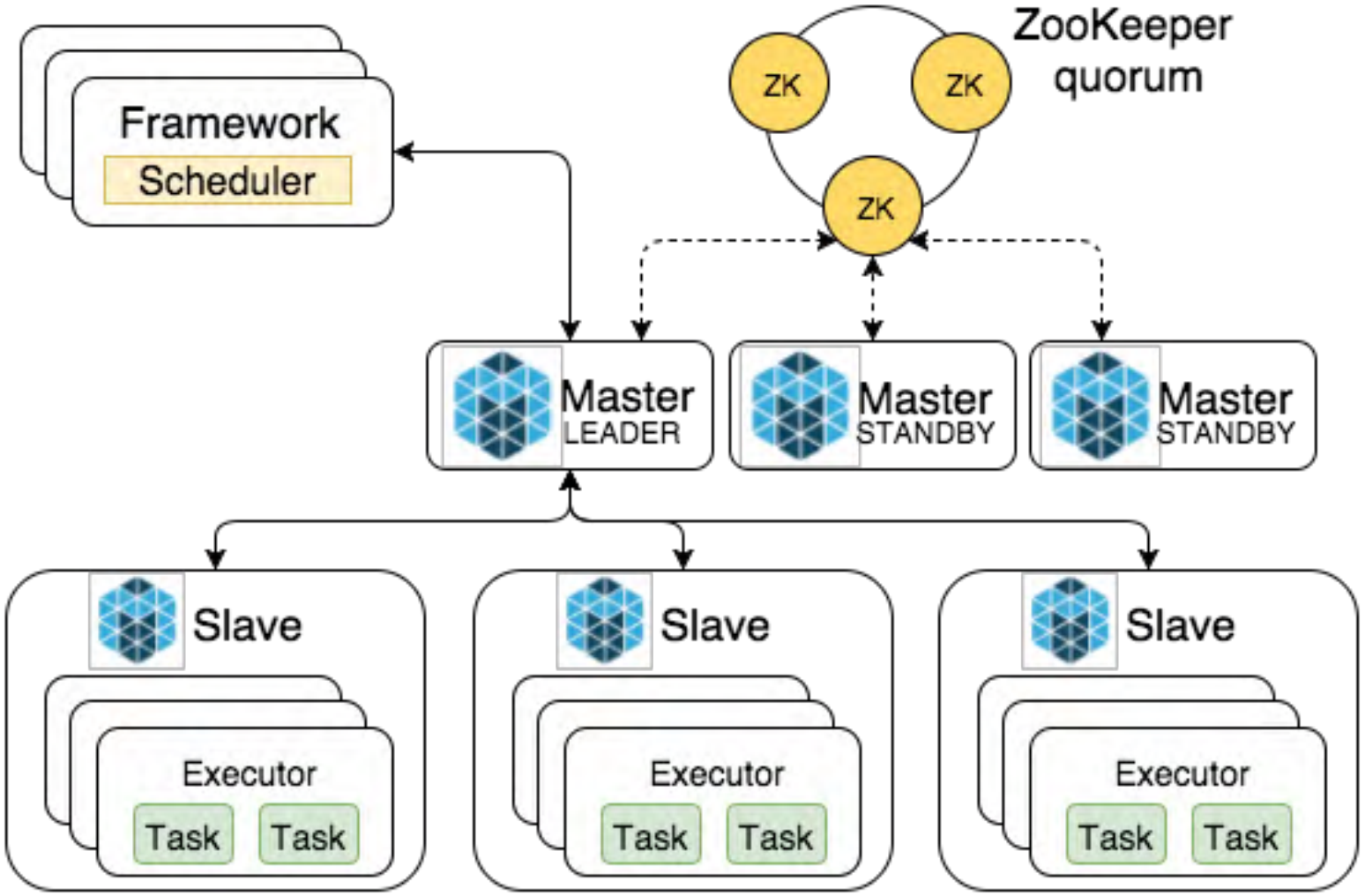
- 打包
  - runtime的一致性
  - runtime的分发
- 运维
  - 资源限制
  - 不再关心依赖
  - 简单的清理机制


Streaming jobs
Distributed computing frameworks
Dependencies
Runtime - JVM
Host OS
Streaming jobs
docker
Host OS

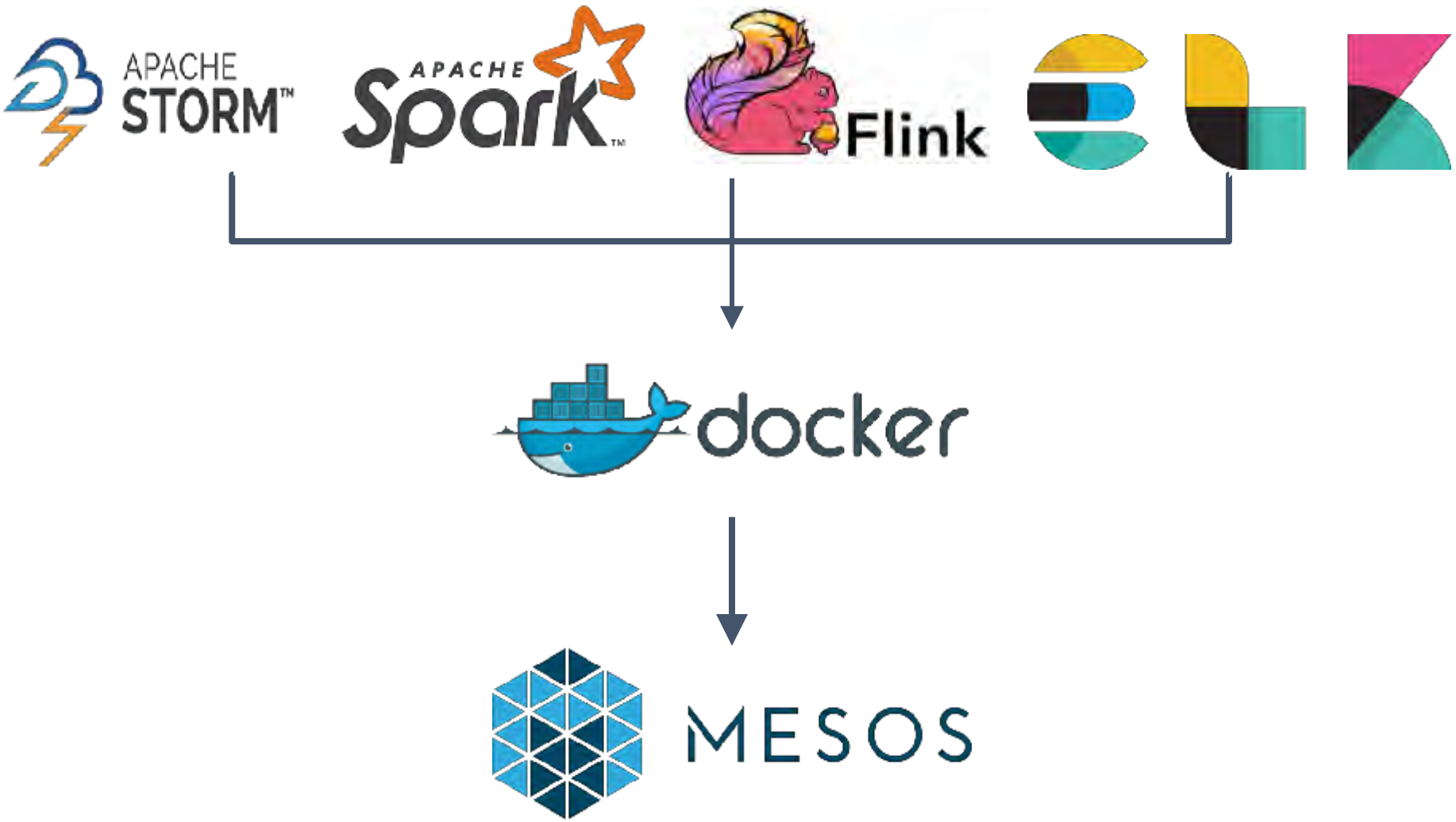
为什么选择Mesos

- 足够简单稳定
  - 大规模调度的成功案例丰富
  - 方便的定制化能力
  - 多种容器
- 较成熟的调度框架
  - Marathon
  - Chronos

Mesos结构


Framework
Scheduler
ZK
ZK
ZK
ZooKeeper
quorum
Master
LEADER
Master
STANDBY
Master
STANDBY
Slave
Executor
Task
Task
Slave
Executor
Task
Task
Slave
Executor
Task
Task

目标



组件容器化与部署

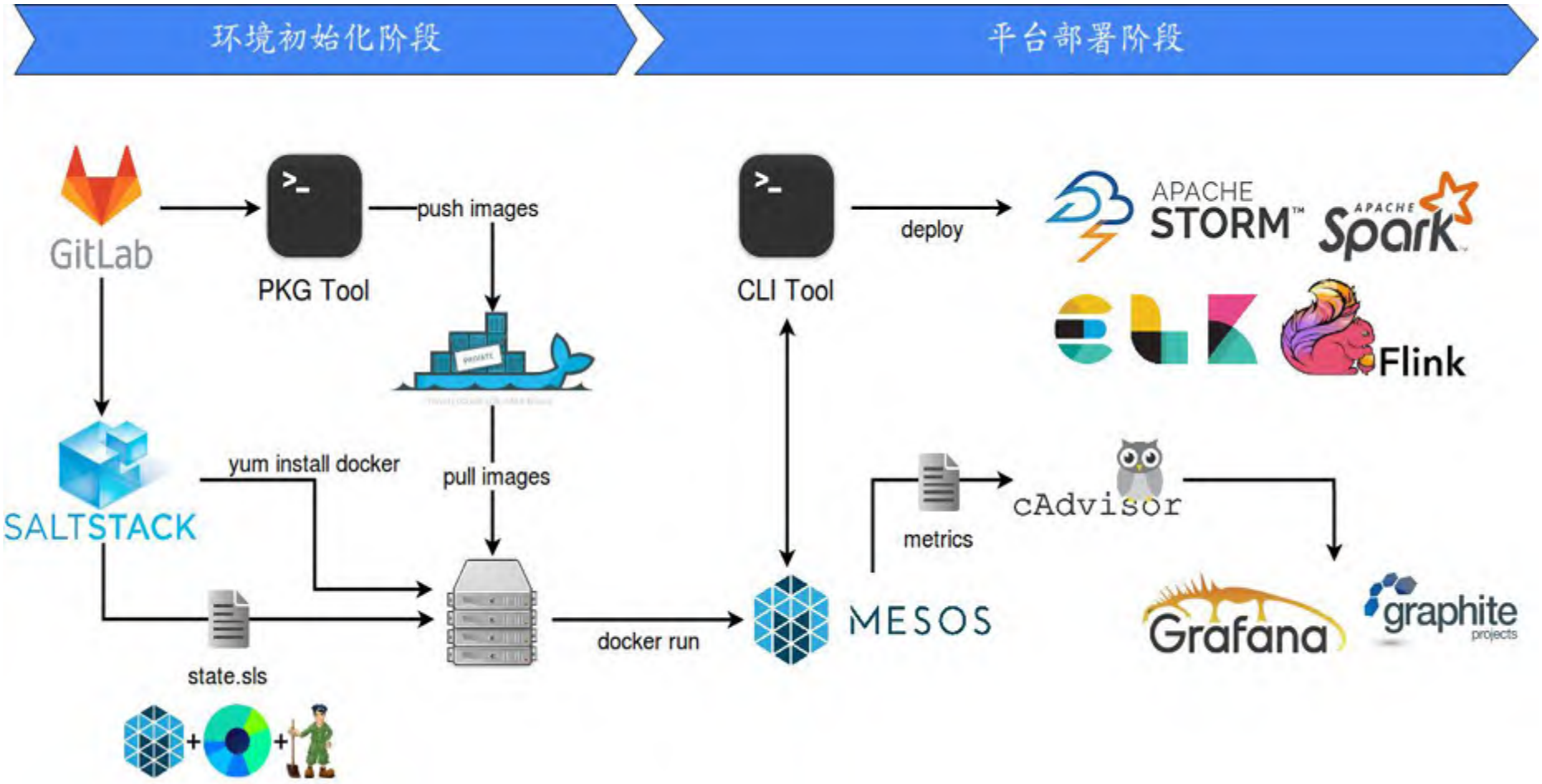
- 潜在创建文件的配置都要注意
  - java.io.tmpdir
  - -XX:HeapDumpPath
  - -Xloggc
- 时区与编码
  - --env TZ=Asia/Shanghai
  - --volume /etc/localtime:/etc/localtime:ro
  - --env JAVA_TOOL_OPTIONS="-Dfile.encoding=UTF-8 -Duser.timezone=PRC

- 主动设置heap
  - 防止ergonomics乱算内存
- CMS收集器要调整并行度
  - -XX:ParallelGCThreads=cpus
  - -XX:ConcGCThreads=cpus/2

- 需要关注的配置参数
  - MESOS_systemd_enable_support
  - MESOS_docker_mesos_image
  - MESOS_docker_socket
  - GLOG_max_log_size
  - GLOG_stop_logging_if_full_disk

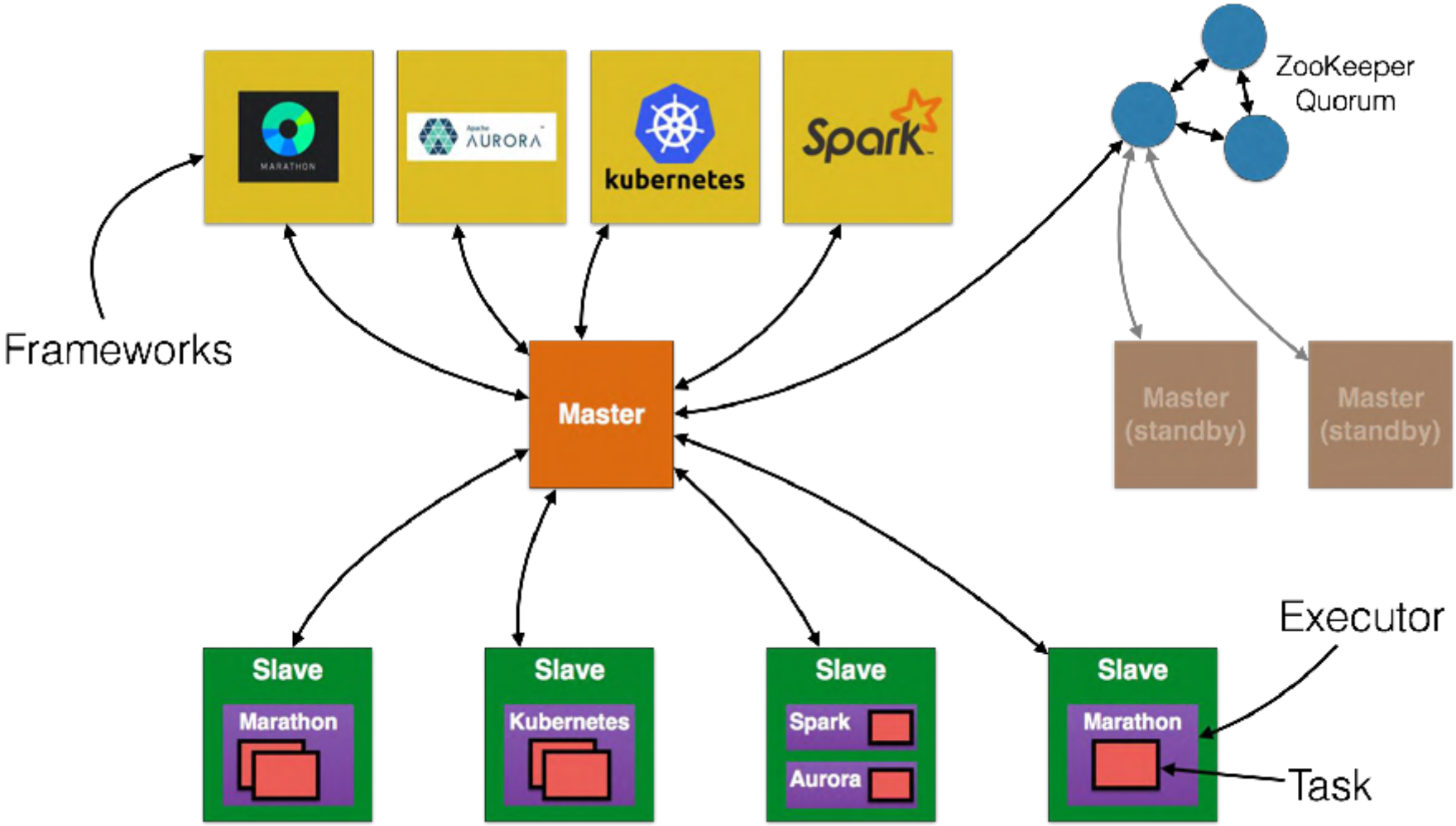
- 需要关注的run参数
  - --pid=host
  - --privileged
  - --net=host (optional)
  - root user

平台部署


环境初始化阶段
平台部署阶段
GitLab
PKG Tool
push images
CLI Tool
deploy
APACHE STORM
APACHE Spark
ELK
Flink
SALTSTACK
yum install docker
pull images
metrics
cAdvisor
state.sls
docker run
MESOS
Grafana
graphite projects

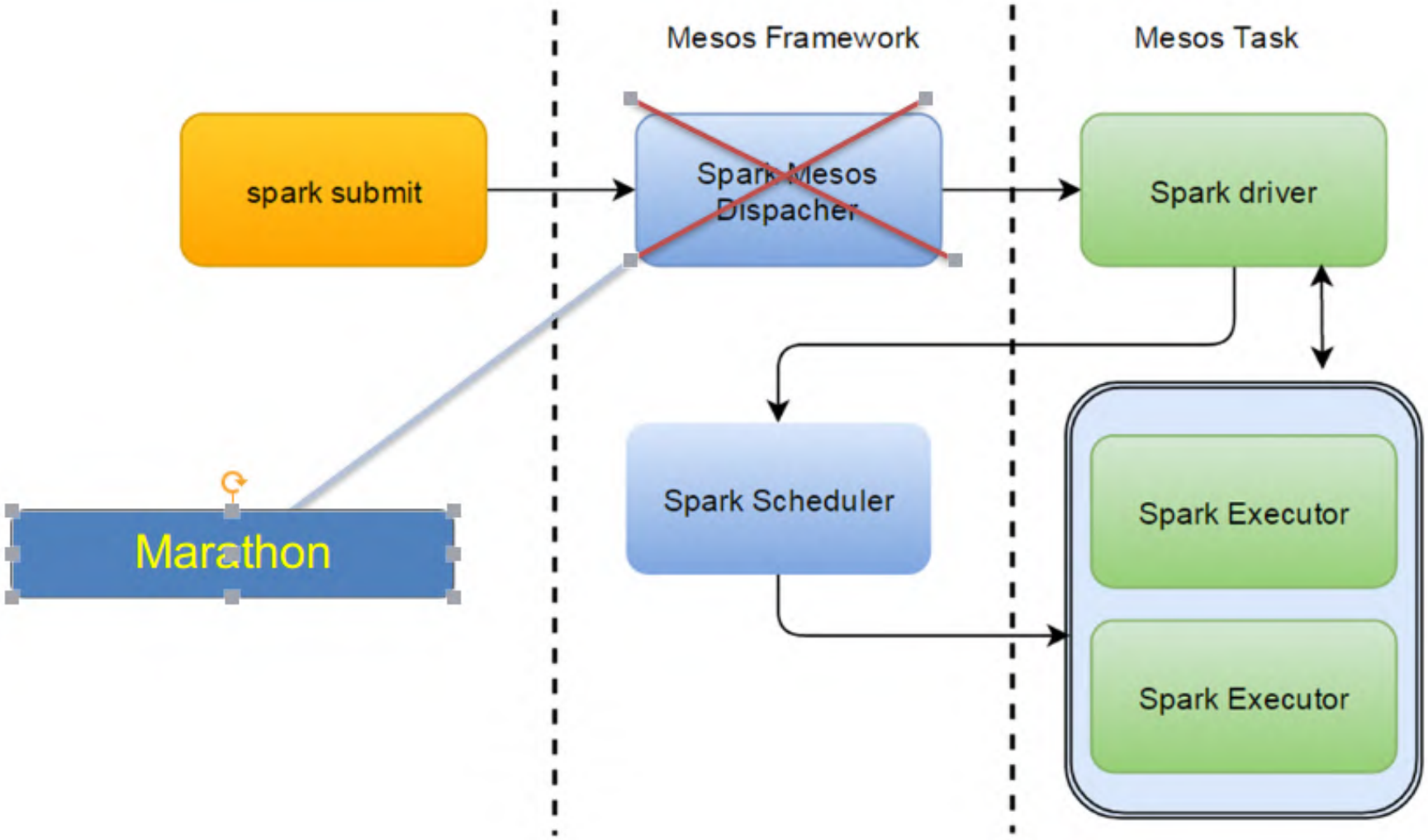
基于Marathon的Streaming调度

Marathon简介

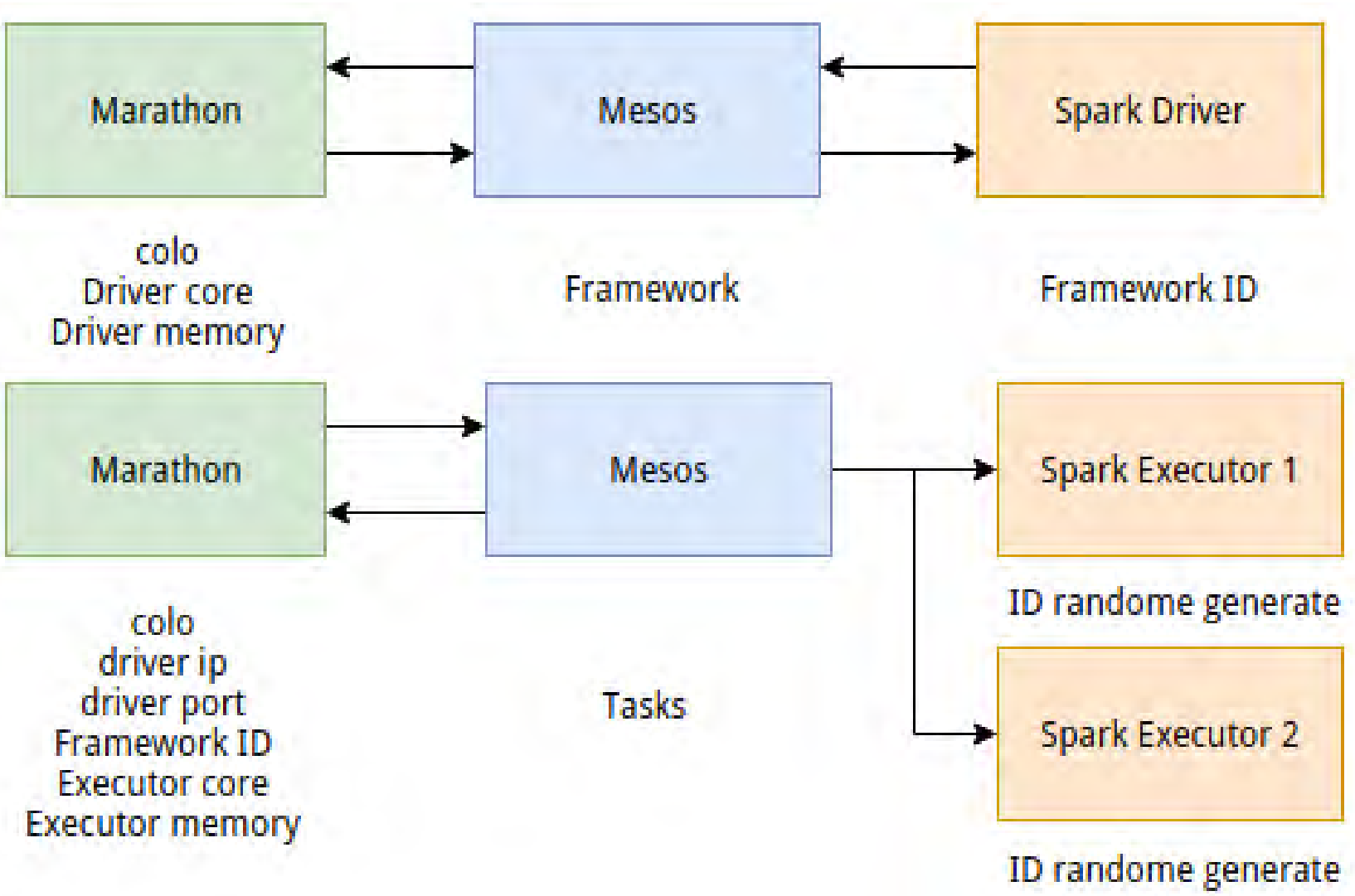

MARATHON
AURORA
kubernetes
Spark
ZooKeeper
Quorum
Frameworks
Master
Master
(standby)
Master
(standby)
Executor
Slave
Marathon
Slave
Kubernetes
Slave
Spark
Aurora
Slave
Marathon
Task

- 运维标准化&自动化
- 解决Mesos-Dispatcher的不足
  - 配置不能正确同步
  - 基于attributes的过滤功能缺失
  - 按role/principal接入Mesos
  - 不能re-registery
  - 不能动态扩容executor

替代Spark Mesos Dispatcher

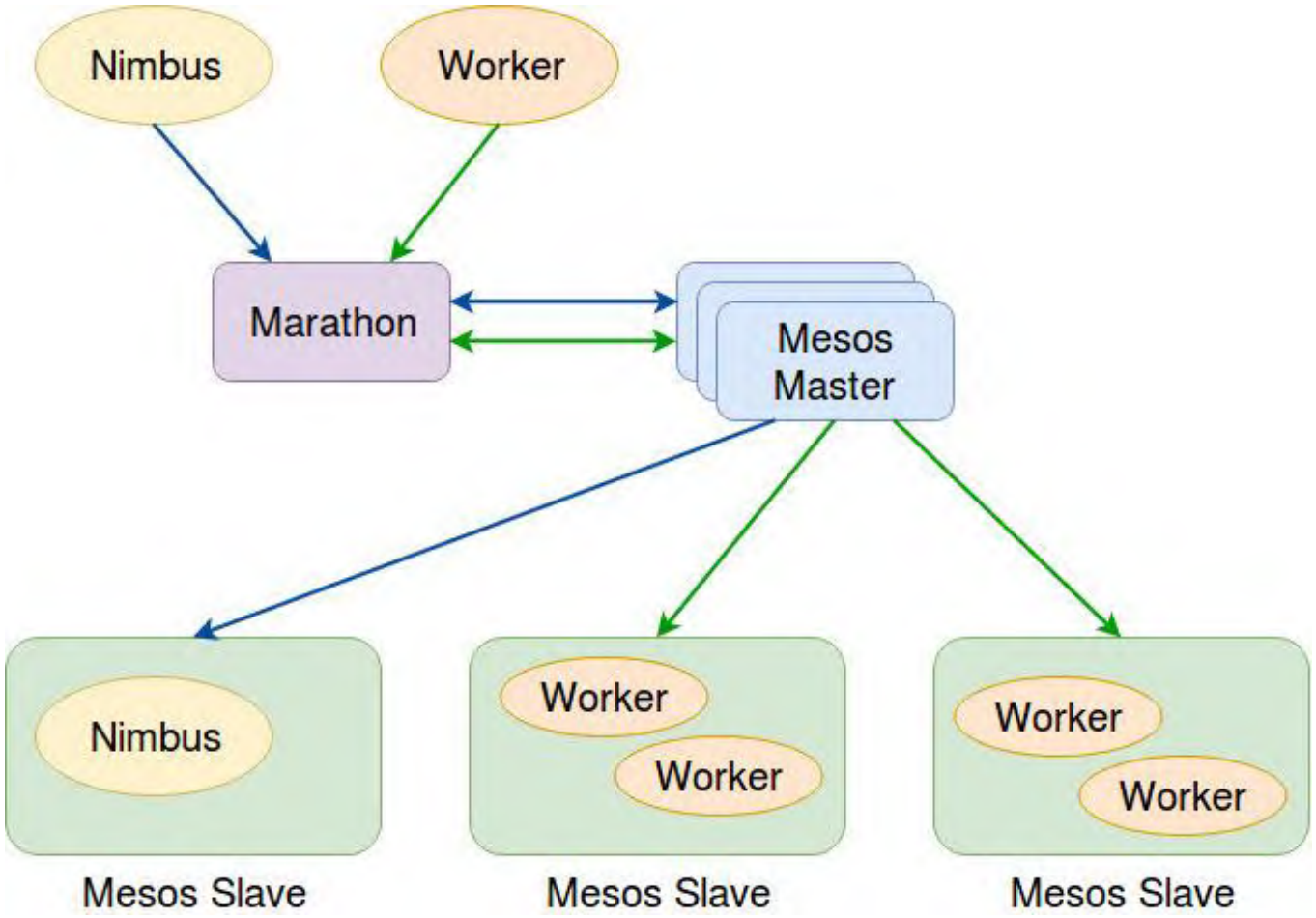


替代Spark Mesos Dispatcher



- Checkpoint & Block
  - 动态预留 & 持久化卷
  - setJars
  - 清理无效的卷
- 临时文件
  - java.io.tmpdir=/mnt/mesos/sandbox
  - spark.local.dir=/mnt/mesos/sandbox
- Coarse-Grained

Storm on Marathon



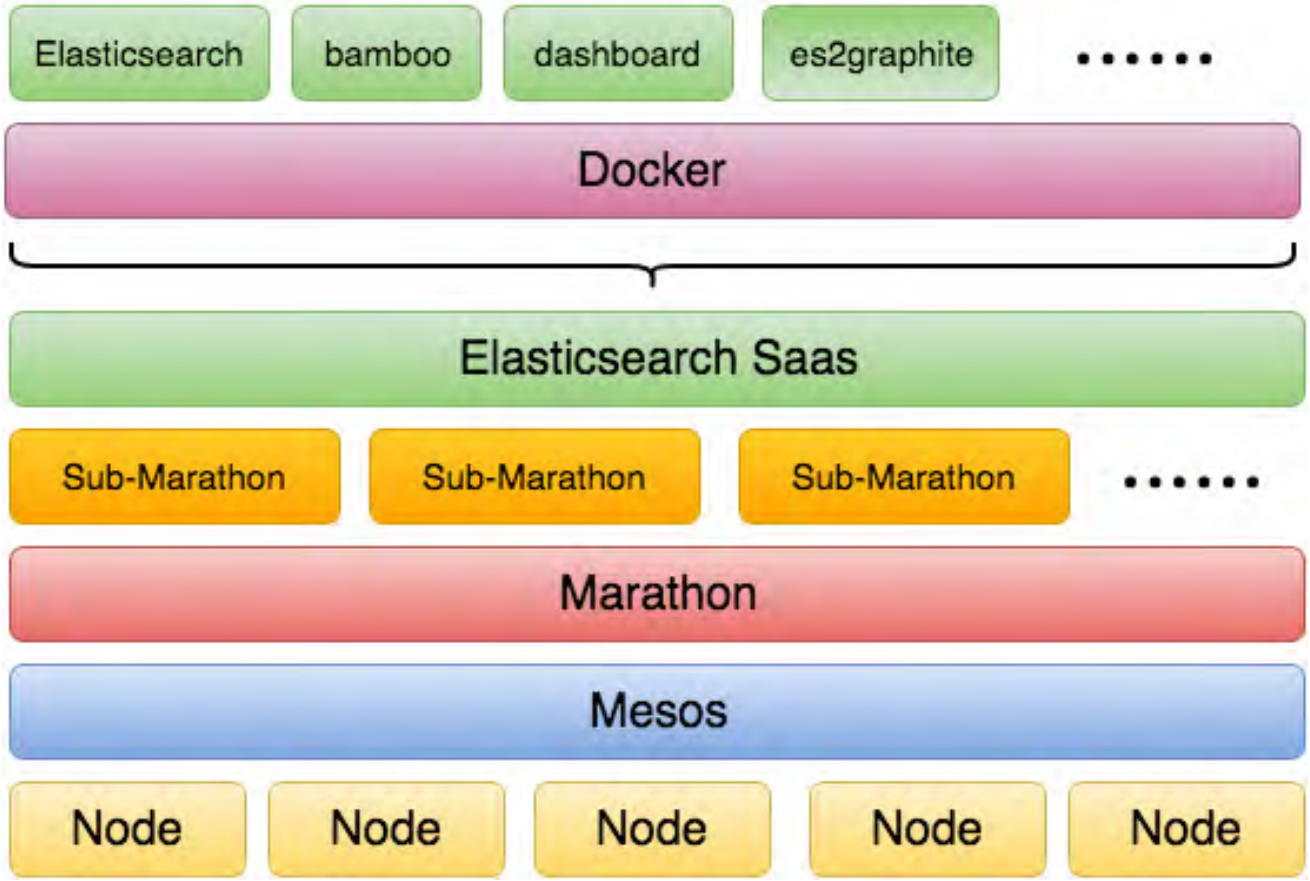
- 源生Web Console
  - 随机端口
  - openresty配合泛域名
- Filebeat + Kafka + ELK
  - 多版本追溯
  - 日常排错
  - 异常监控
- Metrics

ELK on Mesos

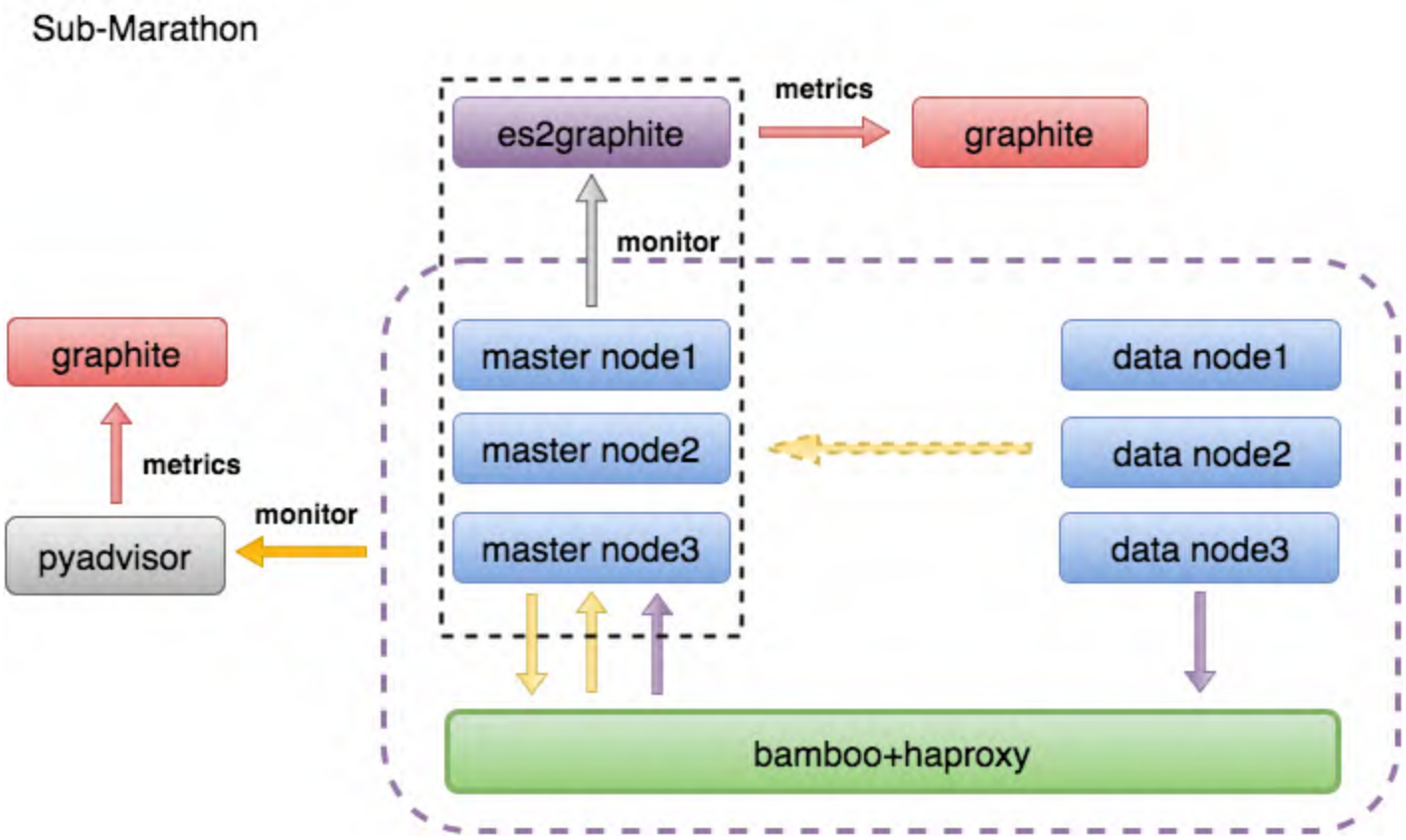
Elasticsearch需求逐步增多

- 目前托管了40+集群
- 100TB+业务数据
- 高峰期 1.2k QPS
- 约110个节点
  - SSD vs HDD

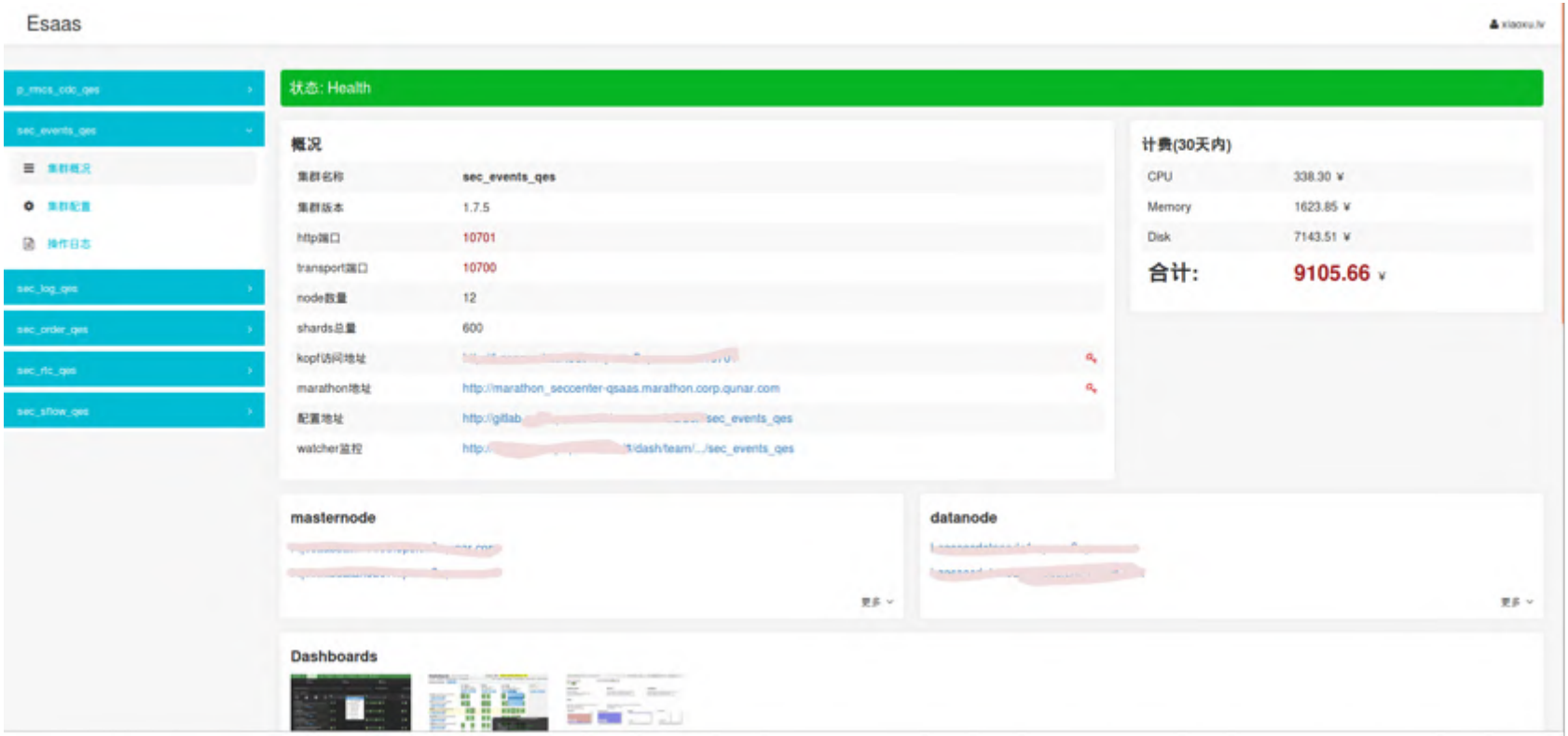
ELK on Mesos结构



Haproxy自组集群



Dashboard


Esaas
状态: Health
概况
集群名称 sec_events_qes
集群版本 1.7.5
http端口 10701
transport端口 10700
node数量 12
shards总量 600
kopf访问地址
marathon地址 http://marathon_seccenter-qsaas.marathon.corp.qunar.com
配置地址
watcher监控
计费(30天内)
CPU 338.30 ¥
Memory 1623.85 ¥
Disk 7143.51 ¥
合计: 9105.66 ¥
masternode
datanode
Dashboards

监控与运维

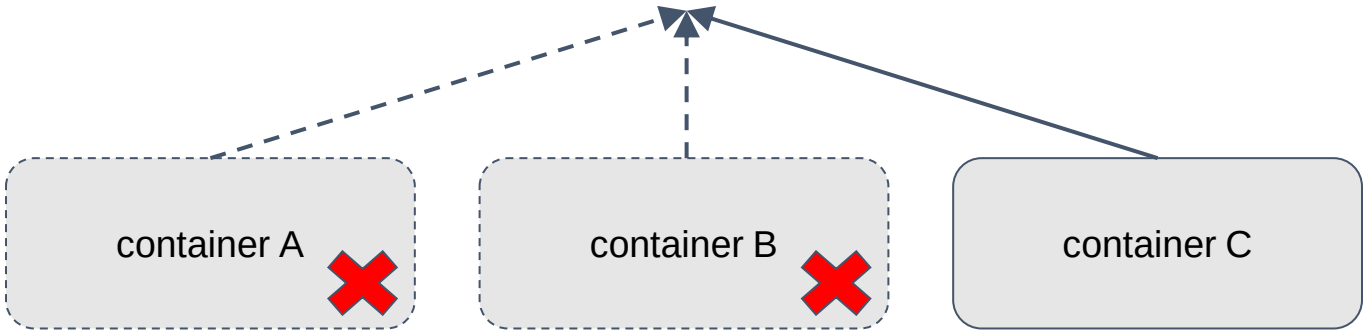
- Streaming拓扑监控
- 业务监控
  - Kafka Topic Lag
  - 处理延迟mean90/upper90
  - Spark scheduler delay/process delay
  - Search Count/Message Count
  - Reject/Exception
  - JVM

- Google cAdvisor足够有效
  - mount rootfs可能导致容器删除失败 #771
  - --docker_only
  - --docker_env_metadata_whitelist
- Statsd + Watcher
  - 基于Graphite的千万级指标监控平台
- Nagios

- 基础监控压力
  - 数据膨胀
  - 垃圾指标增多
  - 大量的通配符导致数据库压力较高
- 单个任务的容器生命周期
  - 发布
  - 扩容
  - 异常退出

- per-host => per-container
- 易变
- 多维度的聚合

<prefix>.<hostname>.<task>.<container>.cpu_usage



- Marathon event
  - TASK_FAILED
  - TASK_LOST
  - deployment -> scale
- Docker event
  - oom
  - delete image/container
  - start container

谢谢大家