

360搜索技术论坛长期交流群



基于Mesos的360搜索私有云平台

谭博侃

01

360搜索面临的挑战

02

Mesos介绍

03

私有云平台架构

04

经验分享和未来展望

海量数据

收录几千亿网页
索引几百亿网页

时效性高

每天上万次服务数据变更
每周PB级别数据更新



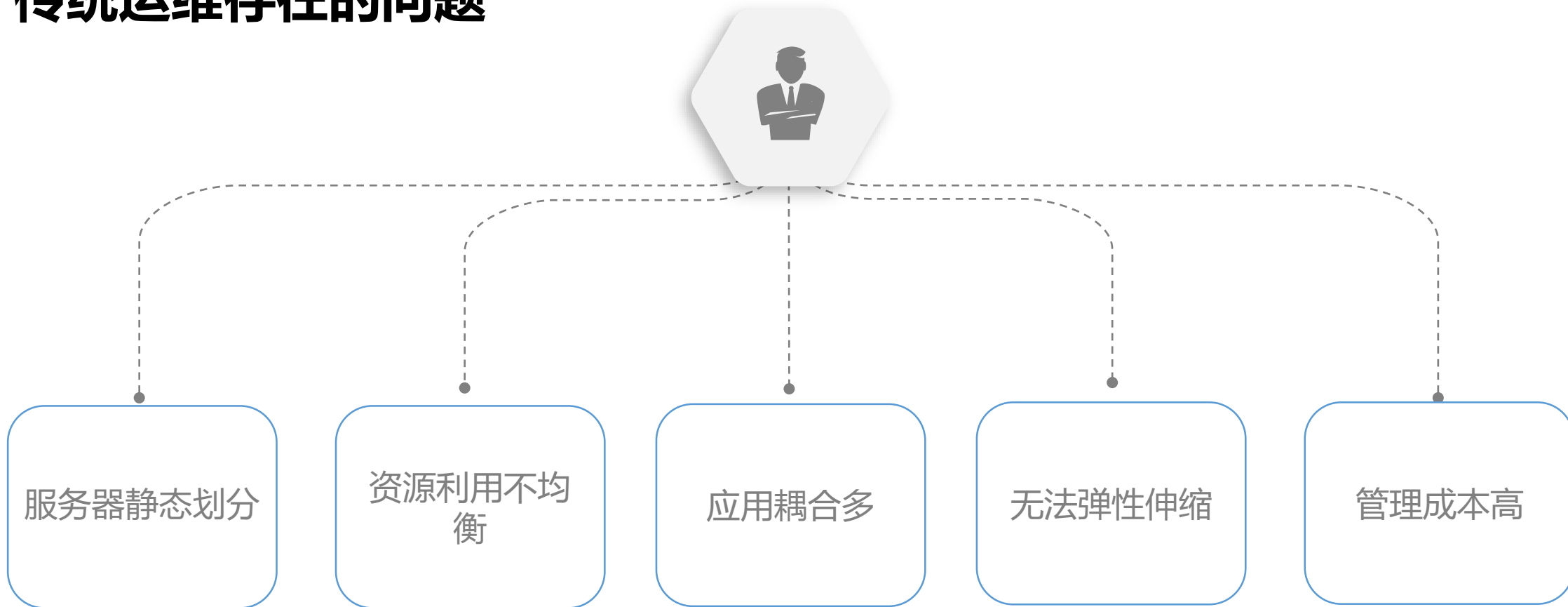
大规模集群

数万台机器, 多个IDC

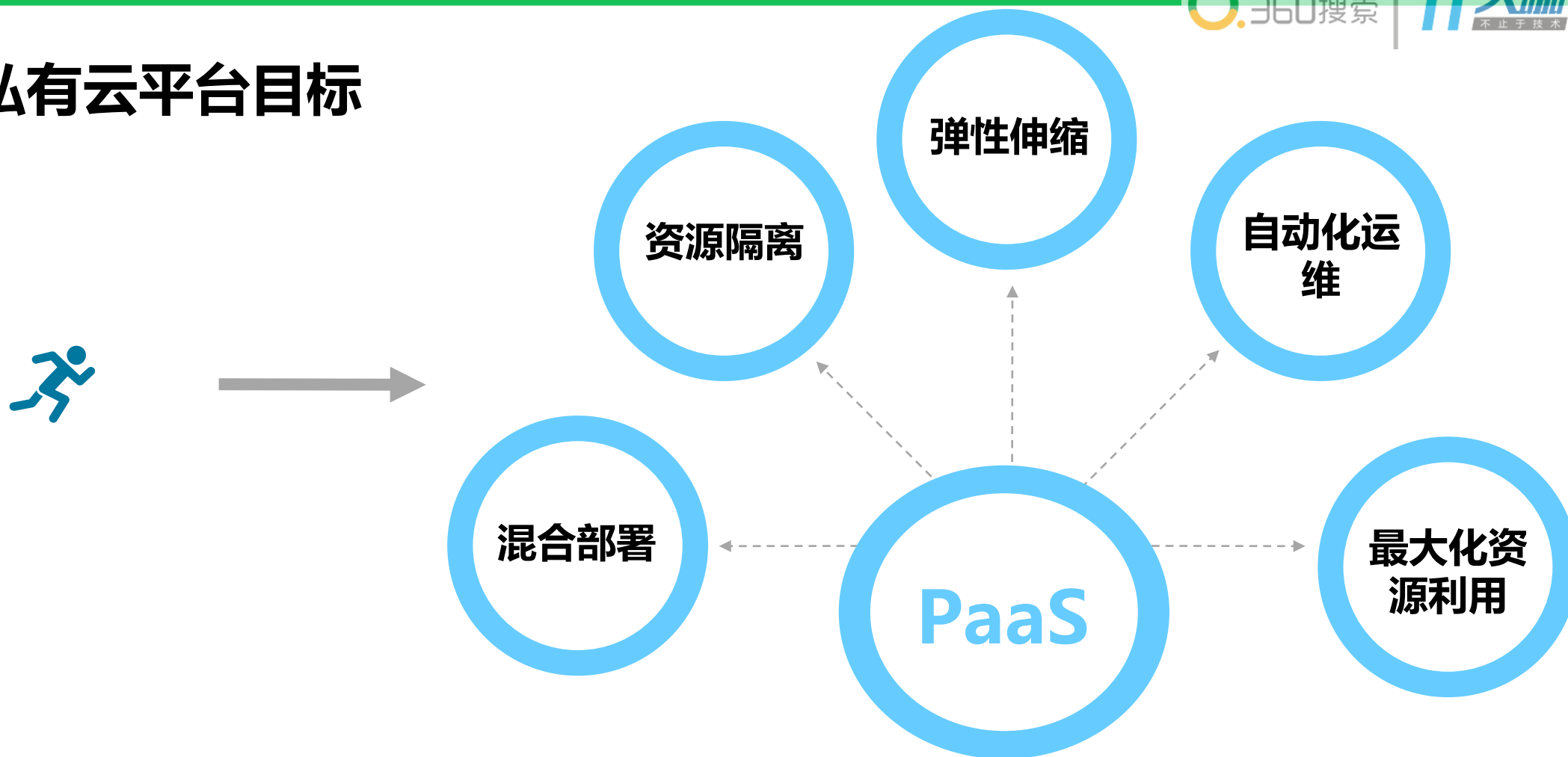
稳定性高

每秒数千次检索请求
可靠性要求 99.99%

传统运维存在的问题



私有云平台目标



最终目的：优化搜索引擎的延迟和吞吐，从而应用更复杂的相关性算法来提高搜索质量

集群管理调度系统:

Google Borg

Apache Mesos

Google Kubernetes

01

360搜索面临的挑战

02

Mesos介绍

03

私有云平台架构

04

经验分享和未来展望

Mesos: A kernel for data center applications

Program against your datacenter like it's a single pool of resources

整个数据中心当成一个资源池来使用

二.Mesos介绍- Why Mesos

- **生产环境验证**

- Twitter/Apple/Netflix

- **扩展性好**

- twitter 3w+
- 接口兼容性好

- **可定制性强**

- Framework/Executor
- Allocator/Container/Network/Storage

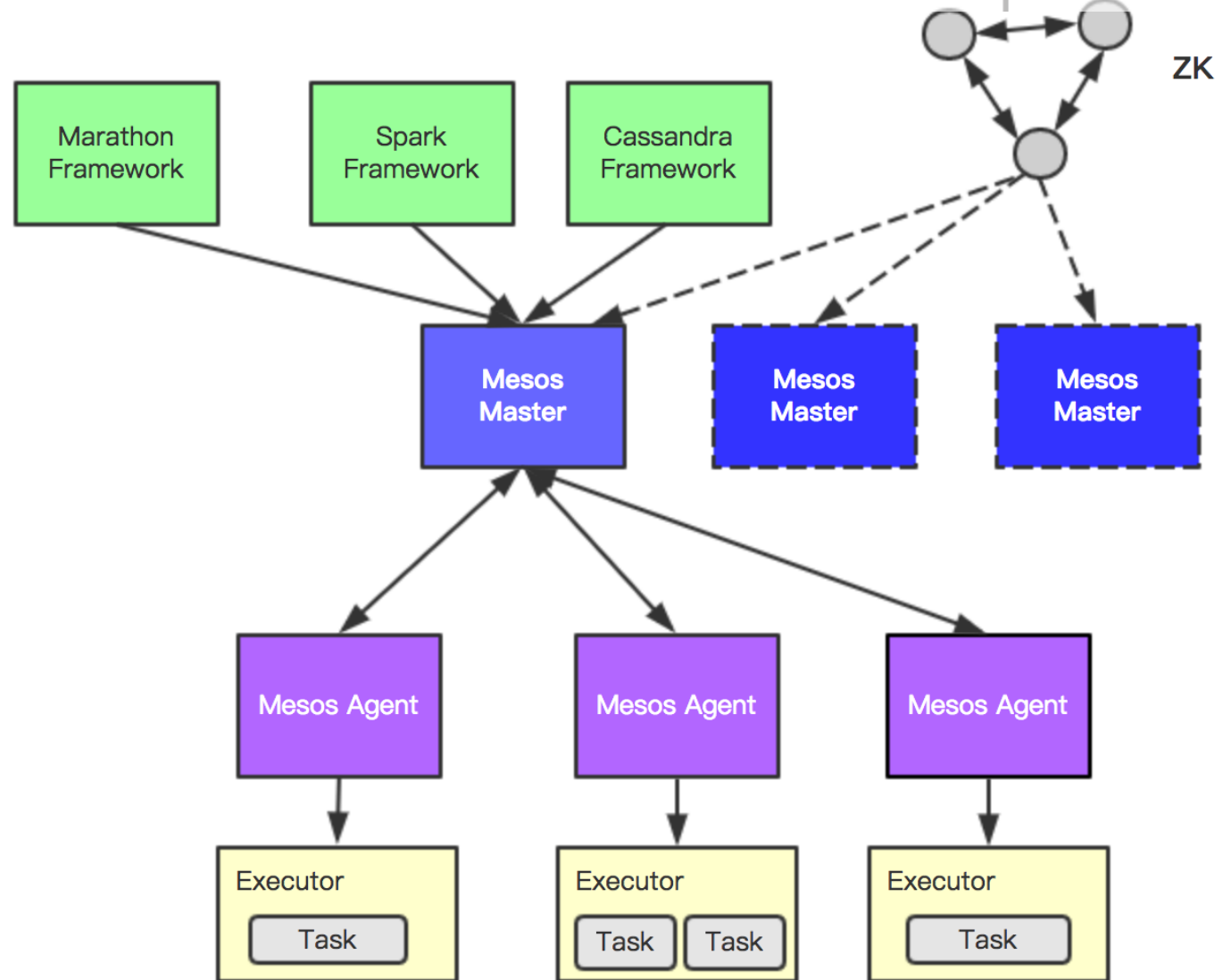
二.Mesos介绍-架构

Zookeeper:
实现Master高可用

Master:
接受Agent上报的资源
分配Offer给上层Framework

Framework:
基于Mesos API实现的调度器
根据调度需求决定是否接收Offer

Agent:
接受并执行Master的命令
通过Executor启动Task



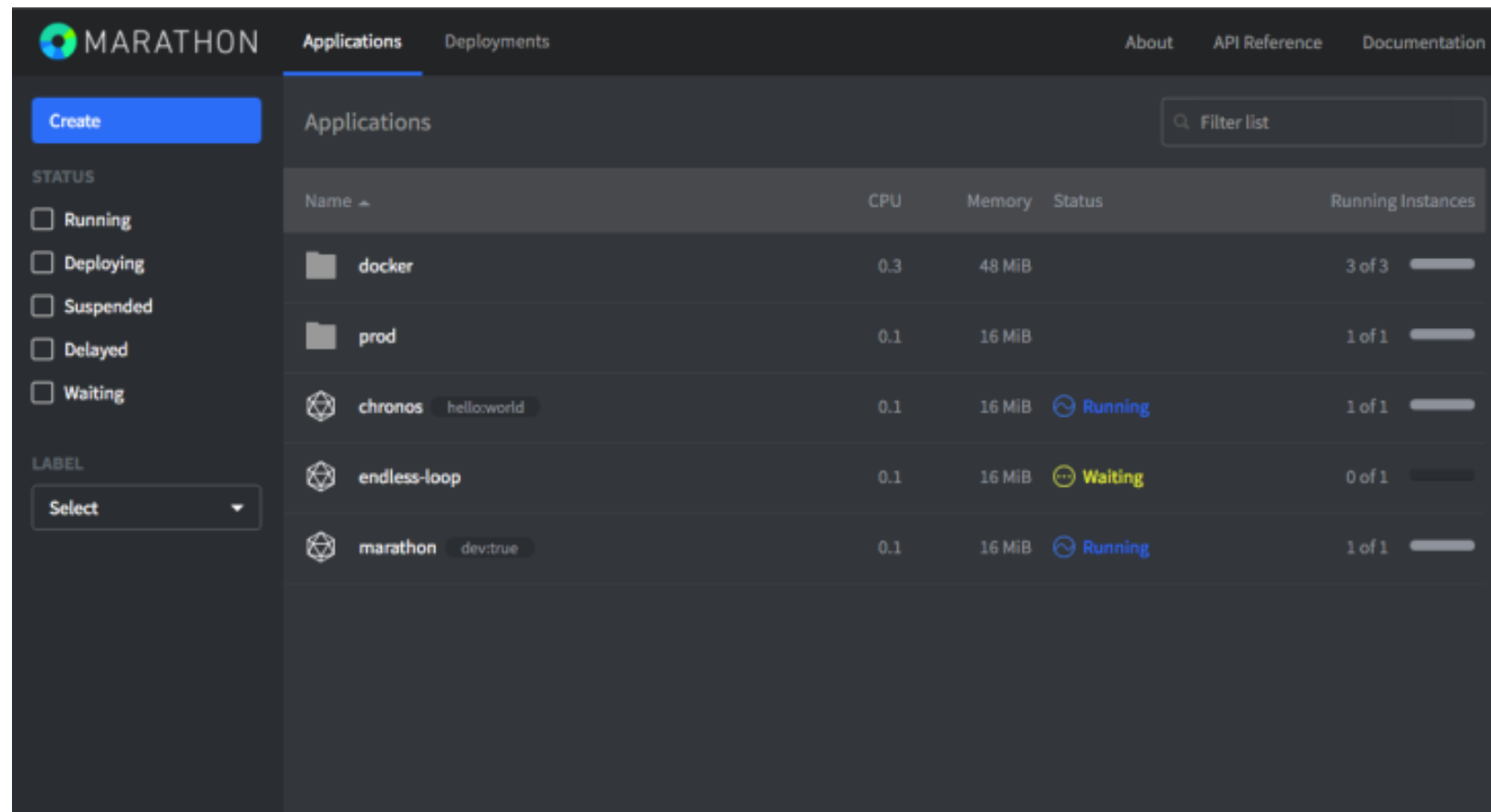
二.Mesos介绍-Framework

Au Aurora	Ku Kubernetes	Dp Dpark	Ex Exelixa	Cc Cray Chapel	JK Jenkins	Tq Torque	Lg Logstash	
Ma Marathon	Si Singularity	Ha Hadoop	Mi MPI	Ck Cook	Ch Chronos	Es Elasticsearch	Ca Cassandra	
Sp SSSP	Fe Fenzo	Sk Spark	St Storm	My Myriad	Ka Kafka	Ht Hypertable	Co Cotton	Ce Ceph

Mesos

Marathon特性

- 长服务
- 约束性条件
- 服务发现
- 健康监测
- Event Bus
- WebUI/RestAPI



The screenshot shows the Marathon web interface. The top navigation bar includes 'Applications' and 'Deployments'. The main content area displays a table of applications with columns for Name, CPU, Memory, Status, and Running Instances. The table lists several applications: 'docker', 'prod', 'chronos' (with label 'helloworld'), 'endless-loop', and 'marathon' (with label 'dev:true'). The 'Status' column shows 'Running' or 'Waiting' with corresponding icons. The 'Running Instances' column shows the number of instances and a progress bar.

Name	CPU	Memory	Status	Running Instances
docker	0.3	48 MiB		3 of 3
prod	0.1	16 MiB		1 of 1
chronos <small>helloworld</small>	0.1	16 MiB	Running	1 of 1
endless-loop	0.1	16 MiB	Waiting	0 of 1
marathon <small>dev:true</small>	0.1	16 MiB	Running	1 of 1

01

360搜索面临的挑战

02

Mesos介绍

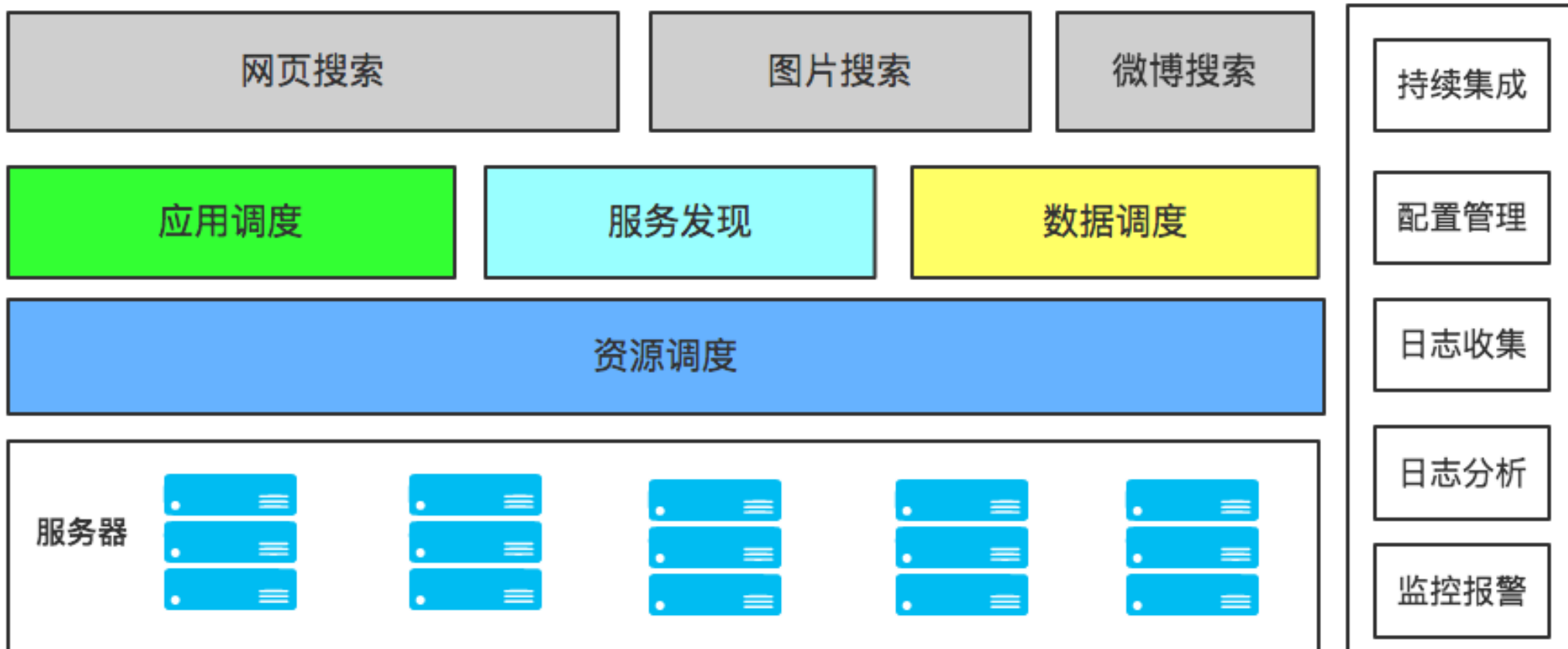
03

私有云平台架构

04

经验分享和未来展望

三.私有云平台架构



三.私有云平台架构-技术选型



	已选方案	其他选择
自动化部署	Ansible	Puppet/SaltSlack
服务调度	Mesos + Marathon	K8S/Swarm
服务发现	Marathon-lb(haproxy)	Nginx/LVS/Zookeeper
配置管理	Consul + Consul-template	Zookeeper/Qconf
数据依赖	自研发数据调度DS	NFS/Ceph/S3/HDFS
日志收集/分析	Filebeat + ELK / QLog	Scribe / Flume
容器	Mesos / Docker	OCI/APPC
网络	Host	CNI
监控报警	InfluxDB/Sensu	Prometheus

01

360搜索面临的挑战

02

Mesos介绍

03

私有云平台架构

04

经验分享和未来展望

私有云平台三点经验分享

服务发现

静态主机和端口变成动态
上游需要能感知下游服务

数据调度

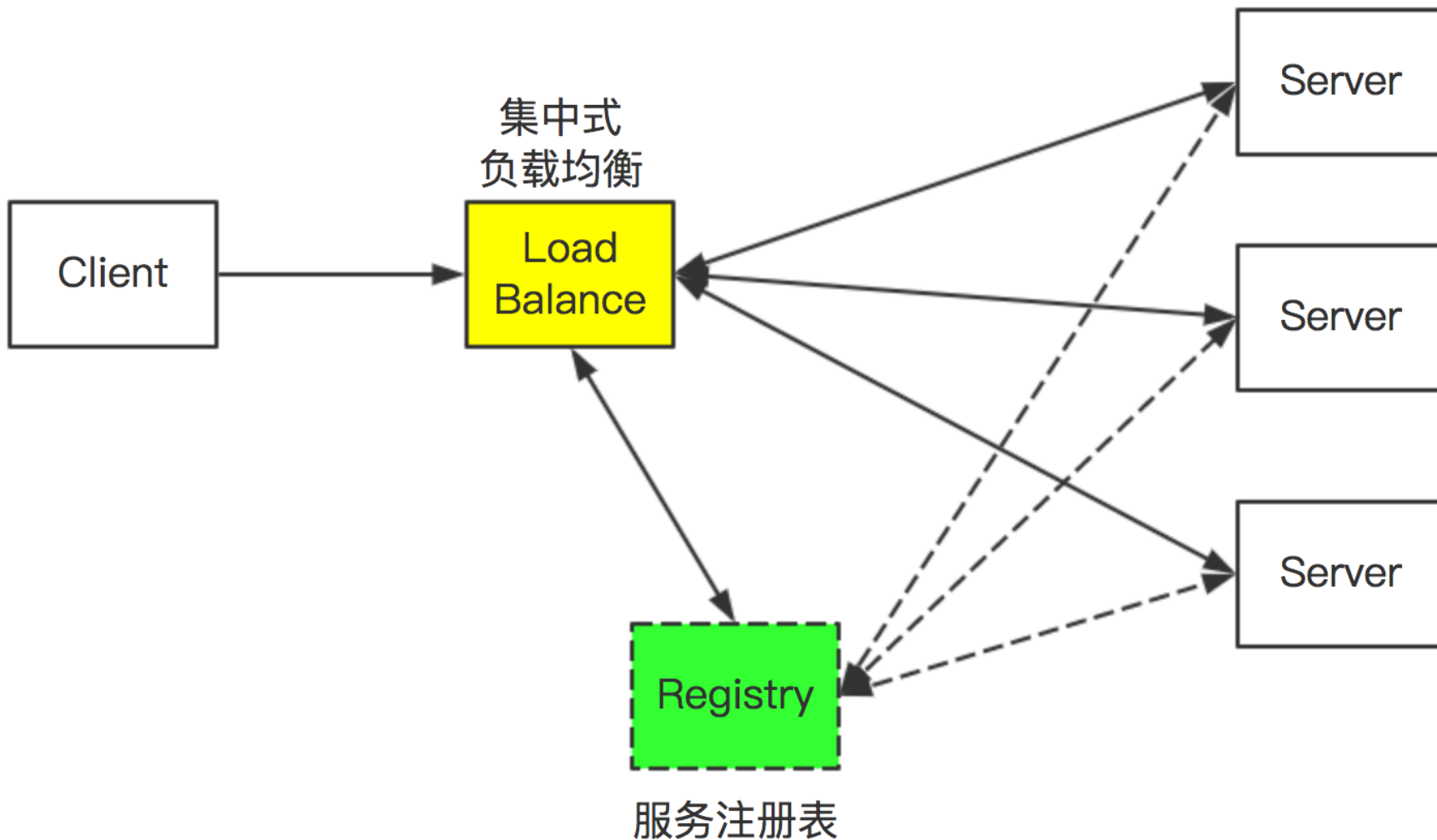
有状态业务迁移云平台难
自研发的DS数据调度系统
解决有状态服务的迁移问题

高可靠保证

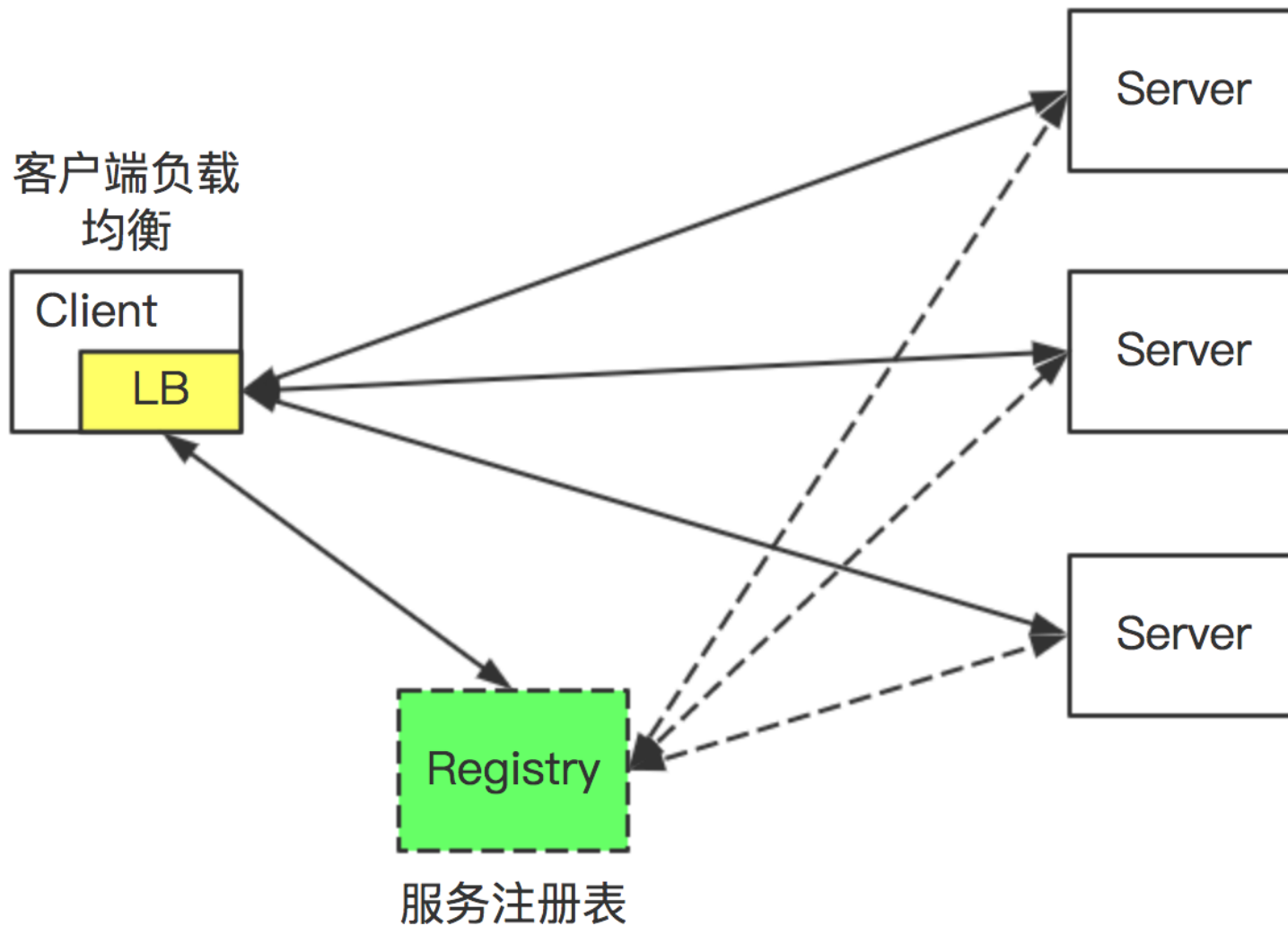
开发者担心迁移后稳定性
如何保证服务可靠性

集中负载均衡

- LVS
- HAProxy
- Nginx



智能客户端



Mesos周边解决方案

- Mesos-DNS
- **Marathon-Ib(集中负载均衡)**
- Zookeeper + SDK(智能客户端)
- Traefik
- Linkerd

四.经验分享-数据调度



无状态的服务

有状态的服务

有状态服务:

- 约束性条件固定节点 (维护成本高)
- Mesos动态预留和持久化卷 (数据无法迁移)
- 分布式文件存储(网络中断,性能差)
- Framework和应用可以进行数据备份和数据迁移(类似Cassandra,改造成本高)

搜索服务的状态特点

数据量大

离线构建

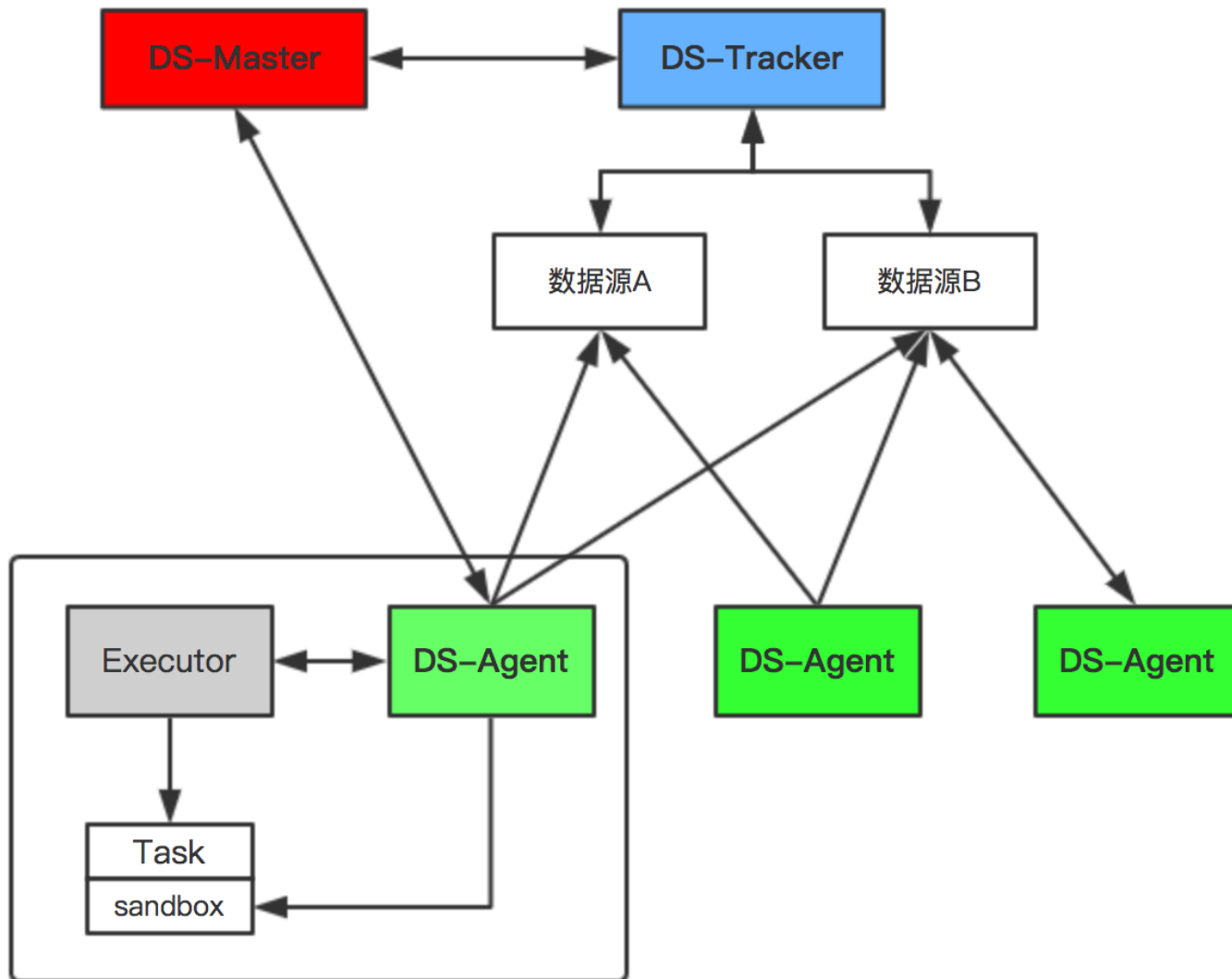
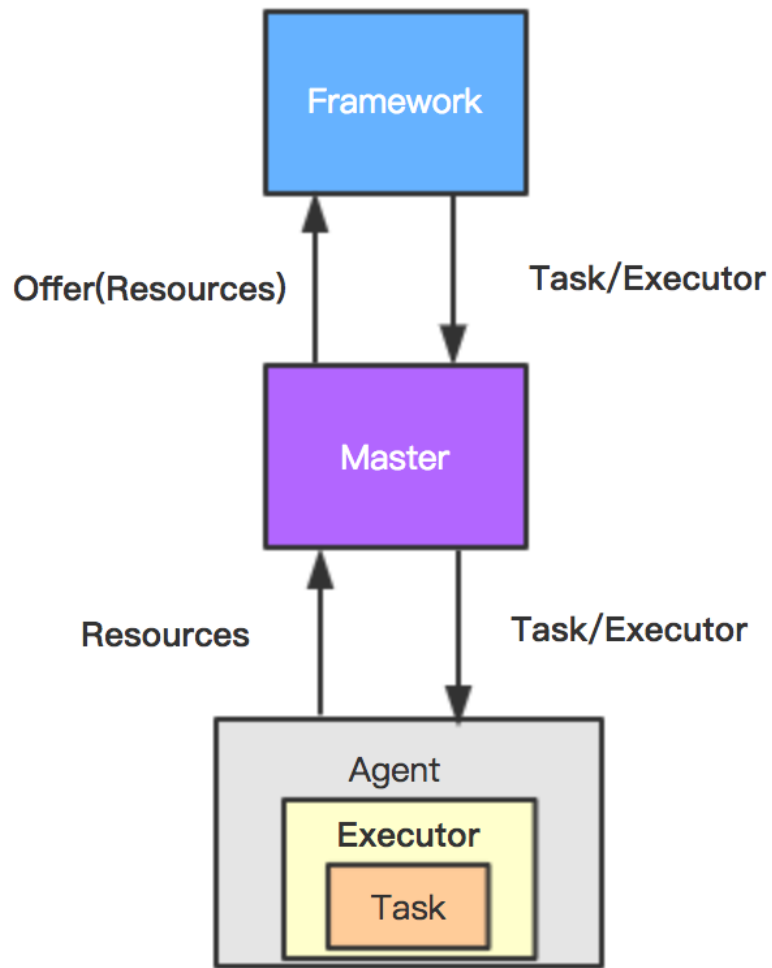
增量更新

应用通知

本地读

数据跨机房
同步

四.经验分享-数据调度 DS架构



架构高可用:

Mesos Master 无状态,通过ZK保证的高可用

Mesos 的Agent支持Recovery

Marathon通过ZK选主和持久化

服务高可用:

Marathon支持健康检查和滚动升级策略

Marathon + Mesos + UCS框架支持优雅退出机制

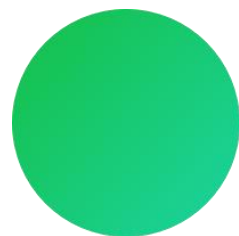
Marathon-lb支持灰度发布

索引服务调度到Mesos私有云平台

定制Framework和DS系统结合，支持数据亲和性调度

DS系统支持HDFS，P2P，跨机房的数据调度

更精准的资源调度，保障业务稳定的前提下最大化利用集群资源



QA