中国移动
China Mobile

DCOS on YARN 在中国移动的实践

中国移动苏州研发中心
陶捷
2016年12月

www.10086.cn

提 纲

1. Slider on YARN
2. Jenkins in Docker on YARN
3. 未来规划和展望

为什么需要DCOS

- 苏研聚焦大数据的技术研究及产品研发，对内服务中国移动各省公司、专业公司，降本增效；对外助力各行各业提升IT能力，实现价值。
- 大数据平台基于开源Hadoop软件面向公司内外提供DaaS、PaaS和SaaS服务，提供统一的运营管理平台。

DCOS方案：Mesos VS YARN

在Mesos上支持Hadoop任务？

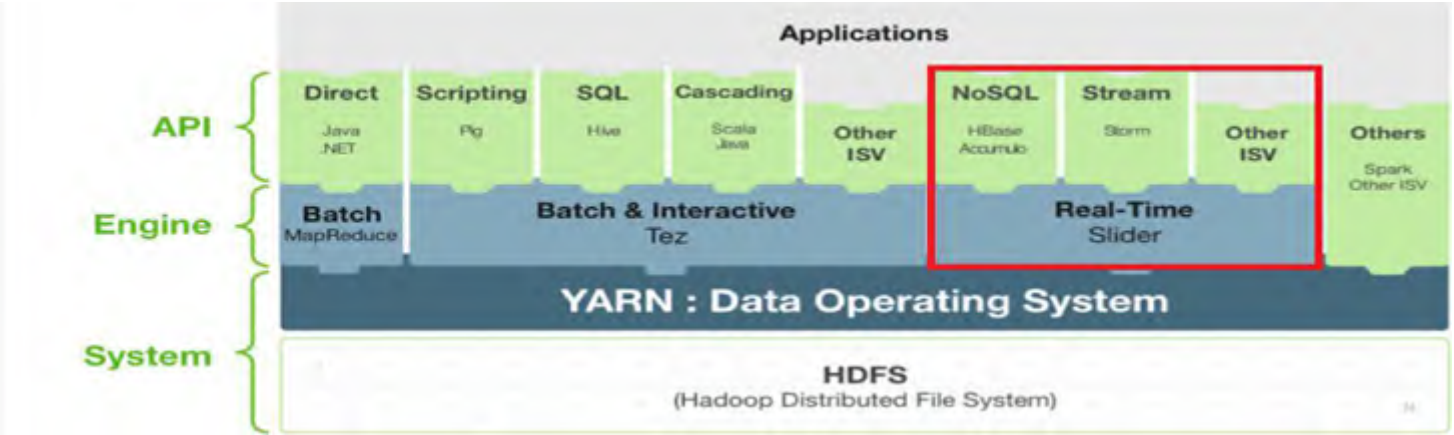
Myriad

在YARN上运行长时服务？

关于Slider

由Hortonworks主导开发，基于Hoya项目

支持long-running服务能够不经改动，直接运行在Yarn上



- Slider成为Apache孵化项目 2014-5
- 2014-5 发布 0.30.0
- 2015-10 发布 0.81.0
- 2016-6 发布 0.91.0

Slider架构

SliderClient

—用户通过SliderClient创建和管理应用生命周期

—应用安装包的管理

SliderAppMaster

— 与YARN交互获取应用的资源

— 接收客户端请求，并分发给Agent

SliderAgent

— 应用的配置和启动，启动的app组件为Comp inst.

— 与AppMaster进行心跳交互，并检查服务的状态

— 应用端口的动态分配

— 执行AM发送过来的应用管理命令

App Package

—应用的定义及管理脚本，其中应用的定义文件包括appConfig.json和resource.json

appConfig.json

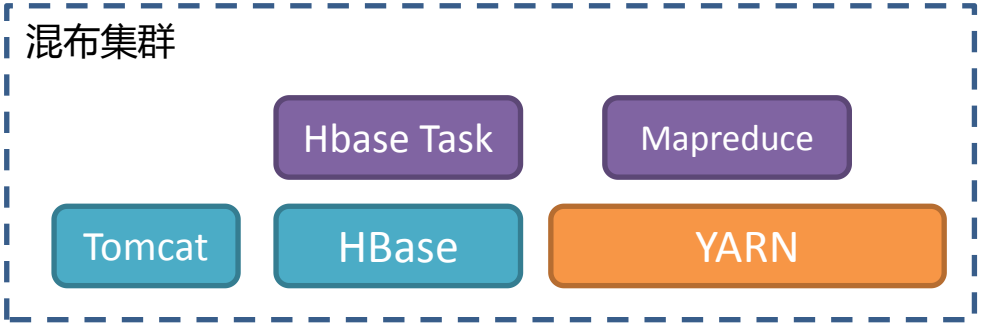
```
{
    "application.def":"/slider/hbase_v096.zip",

    "site.global.app_log_dir":"${AGENT_LOG_ROOT}/app/log",

    "site.global.app_pid_dir":"${AGENT_WORK_ROOT}/app/run",
    "site.global.hbase_master_heapsize":"1024m",
    "site.global.ganglia_server_host":"${NN_HOST}",
    "site.global.ganglia_server_port":"8667",
    "site.global.ganglia_server_id":"Application1",

    "site.hbase-site.hbase.tmp.dir":"${AGENT_WORK_ROOT}/work/app/tmp",
    "site.hbase-site.hbase.master.info.port":"${HBASE_MASTER.ALLOCATED_PORT}",
    "site.hbase-site.hbase.regionserver.port":"0",
    "site.hbase-site.hbase.zookeeper.quorum":"${ZK_HOST}",

    "site.core-site.fs.defaultFS":"${NN_URI}",
}
```

resources.json

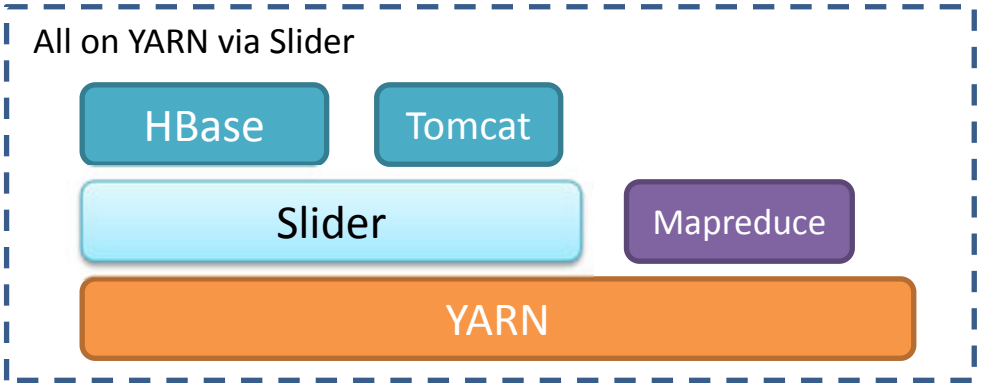
```
{
    "schema":"http://example.org/specification/v2.0.0",
    "global":
    {
        "yarn.memory":"512"
    },
    "components":{
        "HBASE_MASTER":
        {
            "yarn.role.priority":"1",
            "yarn.component.instances":"1",
            "yarn.vcores":"1",
        },
        "HBASE_REGIONSERVER":{
            "yarn.role.priority":"2",
            "yarn.component.instances":"1"
        }
    }
}
```

混布集群 VS Slider部署

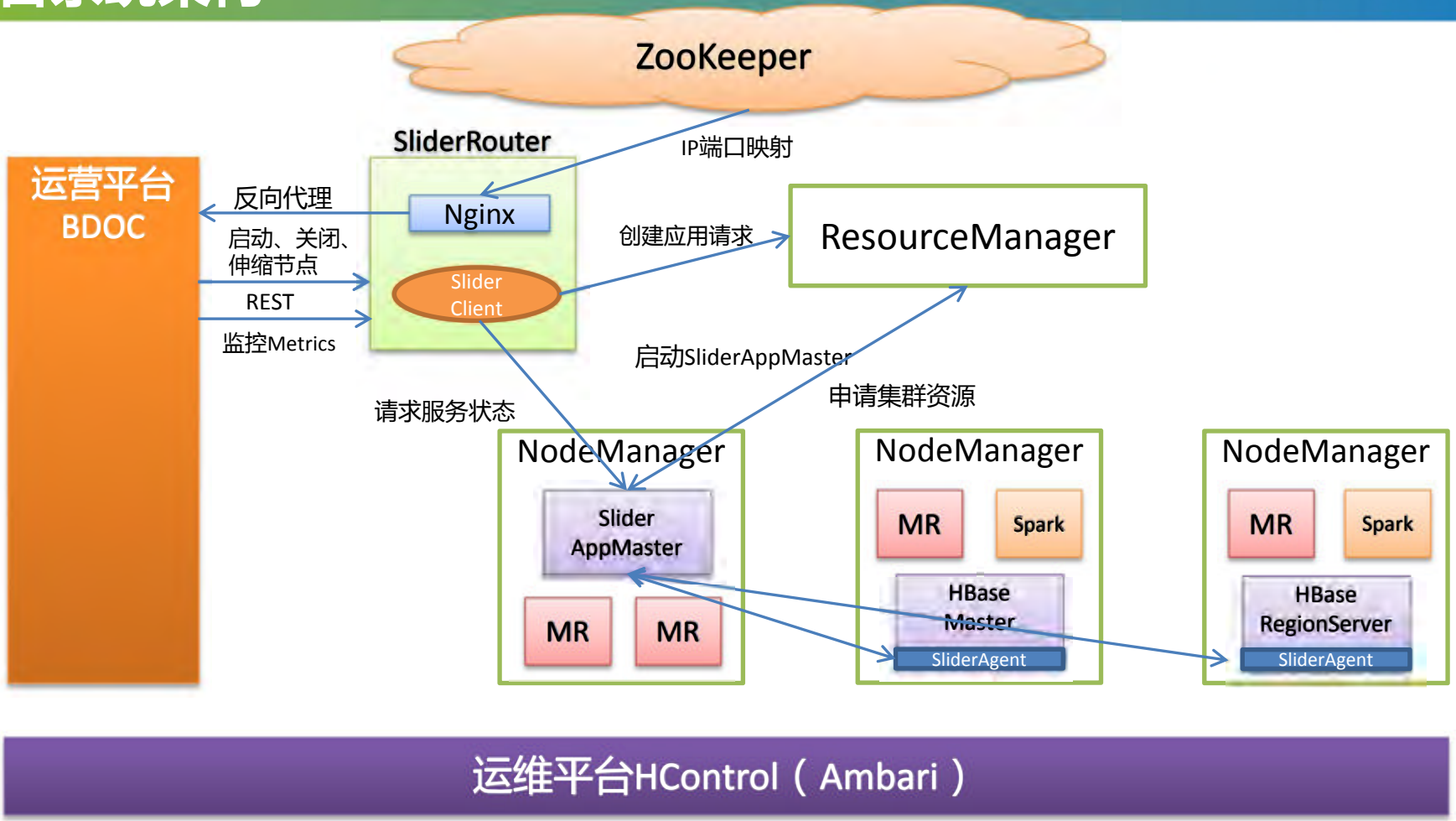
- 传统混布集群
  - 资源无法统一管理
  - 多租户实现依赖应用本身
  - 资源利用率低下



- 使用Slider管理服务
  - 统一管理计算资源
  - 资源利用率高，更具有弹性
  - 使用统一的权限管理、运维监控服务
  - 更高健壮性

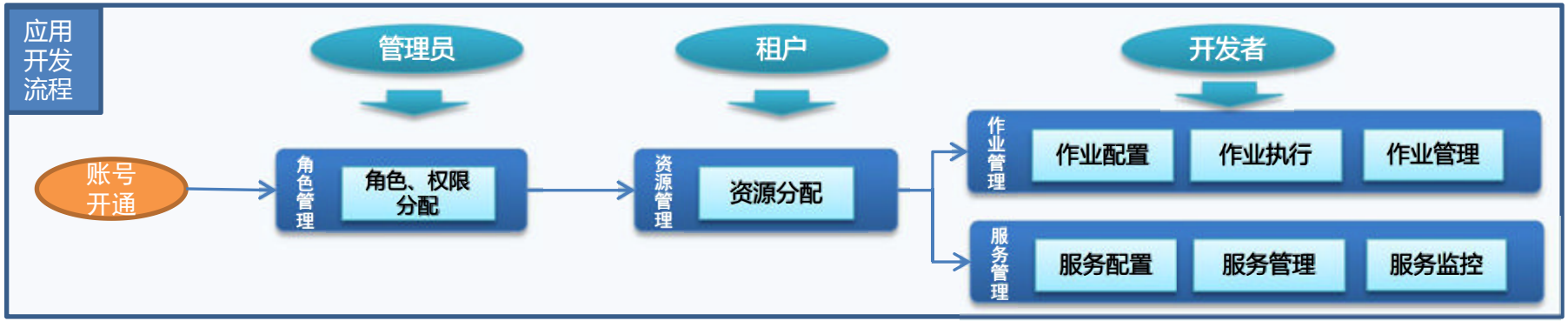


平台系统架构

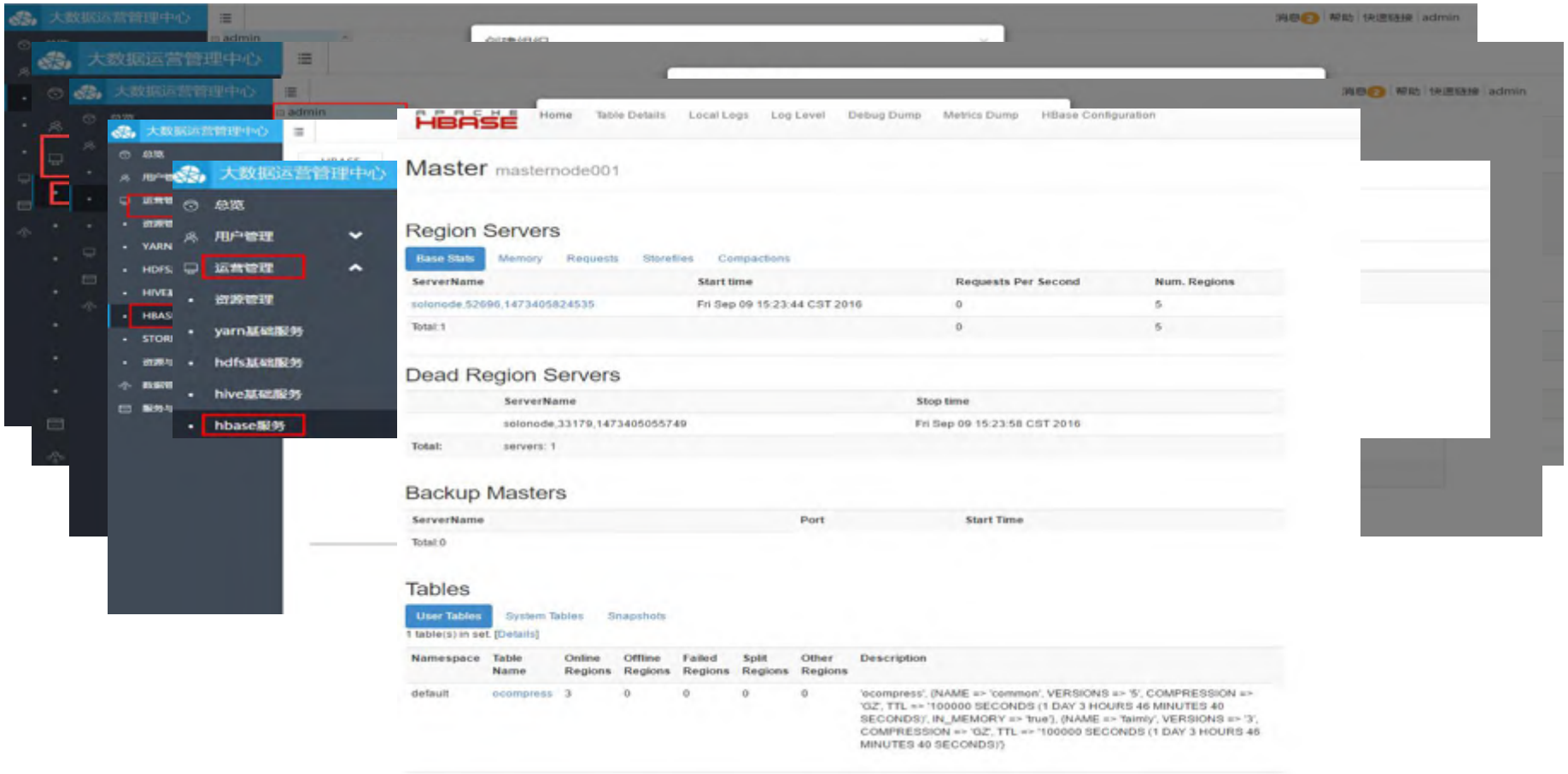

ZooKeeper
SliderRouter
IP端口映射
运营平台
BDOC
反向代理
Nginx
启动、关闭、
伸缩节点
REST
监控Metrics
Slider
Client
创建应用请求
ResourceManager
启动SliderAppMaster
申请集群资源
请求服务状态
NodeManager
Slider
AppMaster
MR
MR
NodeManager
MR
Spark
HBase
Master
SliderAgent
NodeManager
MR
Spark
HBase
RegionServer
SliderAgent
运维平台HControl（Ambari）

BDOC：集群运营管理系统

- 大数据中心全流程统一管理门户
- 支持多种服务能力和异构集群
- 统一管理用户、资源和权限

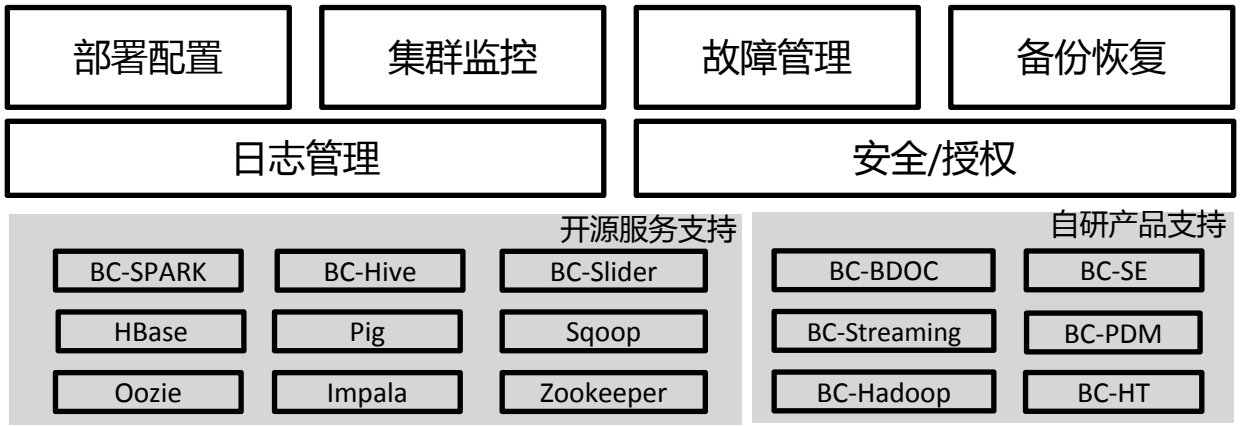


BDOC管理服务生命周期



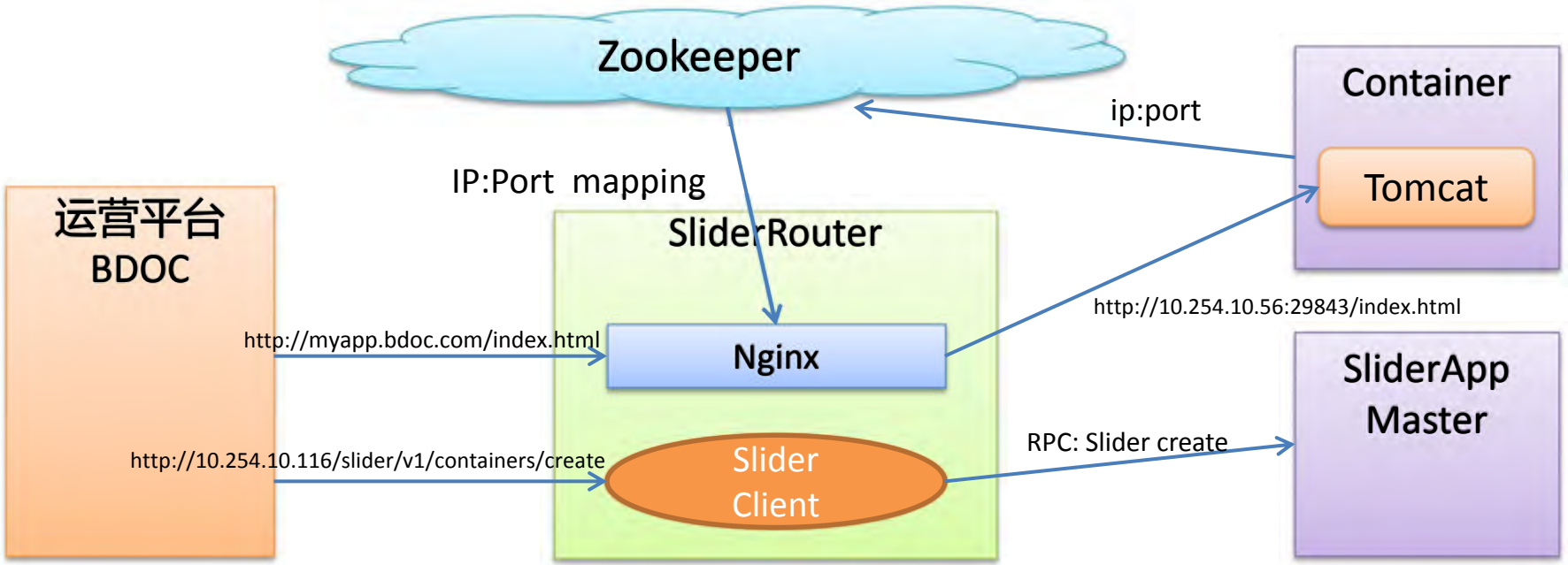
HControl：集群运维管理系统

- HControl基于hortonworks主导的开源项目Apache Ambari
- 部署、管理和监控集群中Hadoop平台各组件
- HControl改进功能：
  - 增加对NTP、LDAP、Impala等支持
  - 增加对自研产品的支持
  - 服务配置自动优化
  - Web界面汉化以及优化



Slider-Router服务

- 封装了SliderClient，对外提供REST接口服务。
- 维护应用节点IP端口信息，提供nginx反向代理服务



Slider服务状态监控

原生Slider对于服务状态的监控，只是由SliderAgent监控服务进程。
进程存在不等于服务状态正常

改进：

- 增加HealthCheck机制，强化对服务状态的监控
- 通过appConfig.json进行配置
- 支持服务端口监听、HTTP请求两种方式
- 对外提供REST接口查询

实践中遇到的问题

Slider应用资源被抢占问题？

FairScheduler支持配置特定队列中资源不被抢占的特性（YARN-4462）

Slider应用Container重启到了其他节点？

YARN支持节点资源预留机制：Slider在启动的Container时会对这个资源标记一个label。Container结束后，YARN会在这个节点上对Container资源锁定一段时间，在此期间，只有原先的应用才能调度该Container资源。（YARN-5636）

Slider应用不受NodeManager管理？

使用前台脚本启动服务进程，从而保证所有服务进程都是NodeManager的子进程。（SLIDER-1055）

Slider-优势

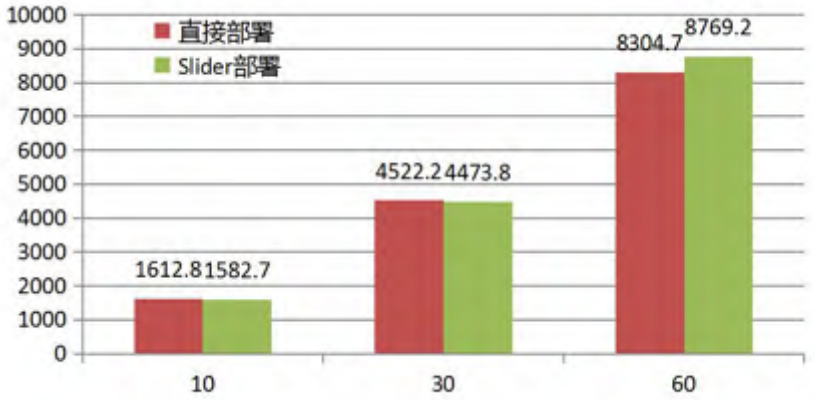
- 支持新应用成本低
- Yarn框架提供了资源管理和隔离
- Yarn框架实现应用的自动恢复
- 支持同一集群内多应用实例，支持不同版本

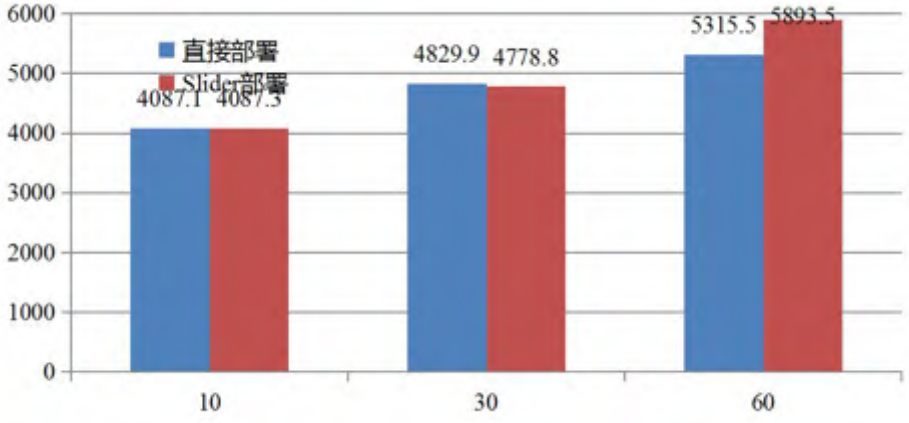
Slider-缺陷

- 客户端单一，仅提供了Shell CLI
- 服务发现机制不够友好
- 应用的监控、日志管理功能、应用配置管理薄弱
- 应用访问本地数据问题

性能对比测试：HBase

写入性能测试：
10亿条数据PUT操作
10/30/60客户端并发

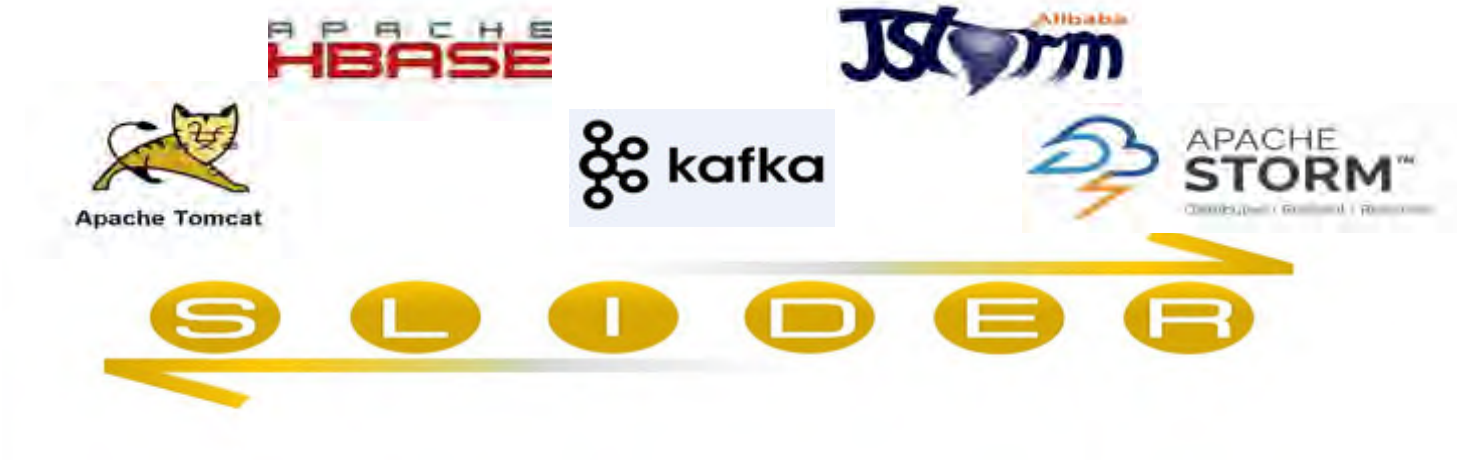




读取性能测试：
3000w用户数据的条件查询
10/30/60客户端并发

实际应用情况

已在生产环境部署的应用：



Yarn+Slider部署的集群

- 上海移动　　Storm/Jstorm　91节点
- 内蒙古移动　HBase/Storm　121节点

提 纲

1. Slider on YARN
2. Jenkins in Docker on YARN
3. 未来规划和展望

Jenkins持续集成的问题

各个项目组都需要使用Jenkins来做持续集成，经常情况是多个项目用一台物理机。

存在问题：

- 环境问题

  环境变量、依赖库、端口冲突……

- 资源竞争

  一些项目持续集成时间长，代码编译、运行单元测试的时间长达多个小时，并发运行机器负载过大。

  另一些项目组机器资源利用率不高

YARN上运行Docker的方案

中国移动 China Mobile

方案一：DistributedShell

分发shell脚本到各个节点，Shell脚本启动和维护Docker实例。
逻辑简单，对Docker的管理弱，缺乏监控、日志等功能。

方案二：DockerContainerExecutor

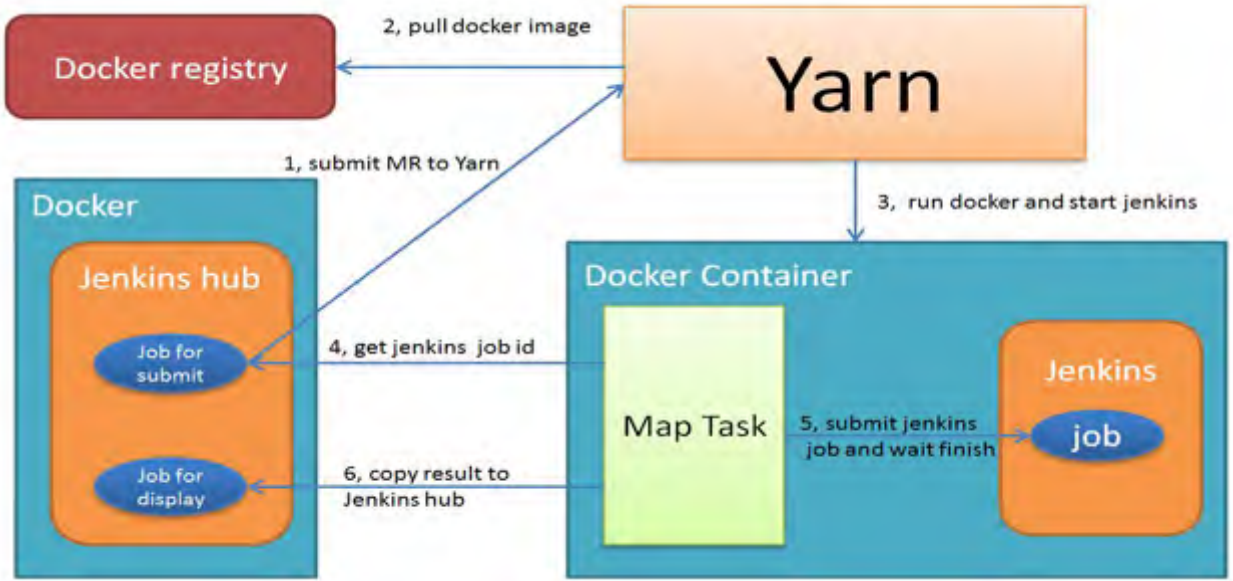
通过DockerContainerExecutor启动MapReduce任务，MR任务运行在Docker中，
并负责与外界交互逻辑。
具有一定Docker管理能力，主要支持MR计算框架

方案三：Docker on Slider

由Slider负责管理Docker
适合长时服务，只能支持已有应用。

Jenkins in Docker on Yarn流程

- 一个常驻的Jenkins服务，用于任务提交和结果展示
- Map-only的MR任务，监控任务运行状态和拷贝运行结果



优点：

- 对用户透明
- 运行完即释放资源
- 运行环境隔离

实践中遇到的问题

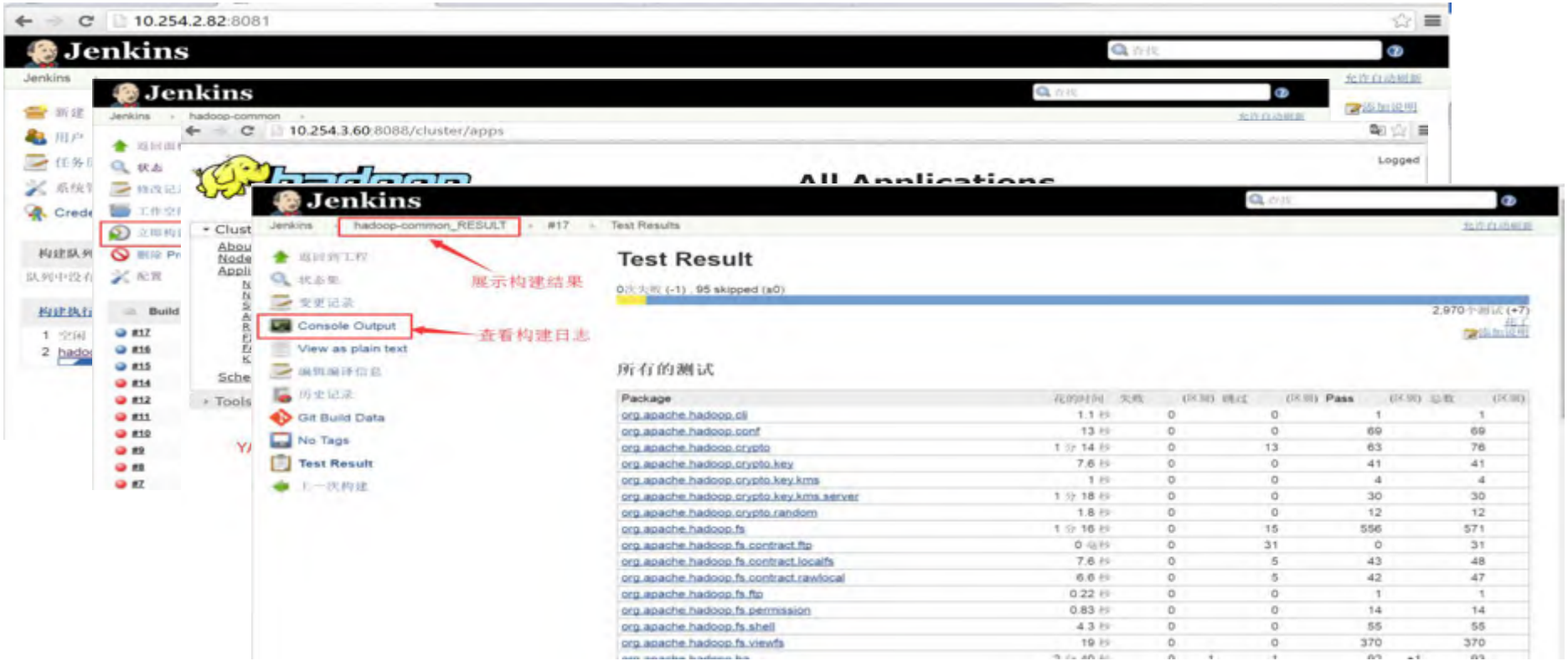
Jenkins端口冲突？

Host网络模式→Bridge网络模式

Docker容量不足？

支持容量自动调整

效果展示


10.254.2.82:8081
Jenkins
10.254.3.60:8088/cluster/apps
All Applications
Logged
hadoop-common_RESULT
#17
Test Results
展示构建结果
查看构建日志
Console Output
View as plain text
Git Build Data
No Tags
Test Result
Test Result
0次失败 (-1) , 95 skipped (a0)
2,970个测试 (+7)
所有的测试
Package
org.apache.hadoop.cli 1.1 秒 0 0 1 1
org.apache.hadoop.conf 13 秒 0 0 69 69
org.apache.hadoop.crypto 1 分 14 秒 0 13 63 76
org.apache.hadoop.crypto.key 7.6 秒 0 0 41 41
org.apache.hadoop.crypto.key.kms 1 秒 0 0 4 4
org.apache.hadoop.crypto.key.kms.server 1 分 18 秒 0 0 30 30
org.apache.hadoop.crypto.random 1.8 秒 0 0 12 12
org.apache.hadoop.fs 1 分 16 秒 0 15 556 571
org.apache.hadoop.fs.contract.ftp 0 毫秒 0 31 0 31
org.apache.hadoop.fs.contract.localfs 7.6 秒 0 5 43 48
org.apache.hadoop.fs.contract.rawlocal 6.6 秒 0 5 42 47
org.apache.hadoop.fs.ftp 0.22 秒 0 0 1 1
org.apache.hadoop.fs.permission 0.83 秒 0 0 14 14
org.apache.hadoop.fs.shell 4.3 秒 0 0 55 55
org.apache.hadoop.fs.viewfs 19 秒 0 0 370 370

提 纲

1. Slider on YARN
2. Jenkins in Docker on YARN
3. 未来规划和展望

社区相关进展

- Assembly：支持在YARN上一次部署一整套服务，例如Storm+Kafka+Solr+Hbase...（YARN-5079）
  - 未来Slider将合并入Yarn
- YARN-DNS：服务发现，container->IP的映射（ YARN-4757）
- Affinity/anti-Affinity: 指定服务节点调度到同一节点/不同节点(YARN-1042)

SDA架构

- 一整套服务部署
- 服务独立的运维监控
- 更完善的资源隔离