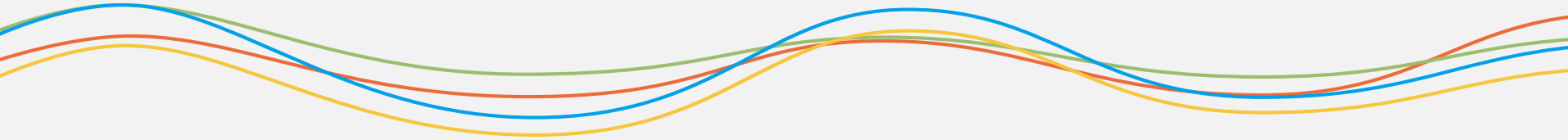
APMCon

Docker on Yarn微服务实践

搜狗 申贤强



关于我们

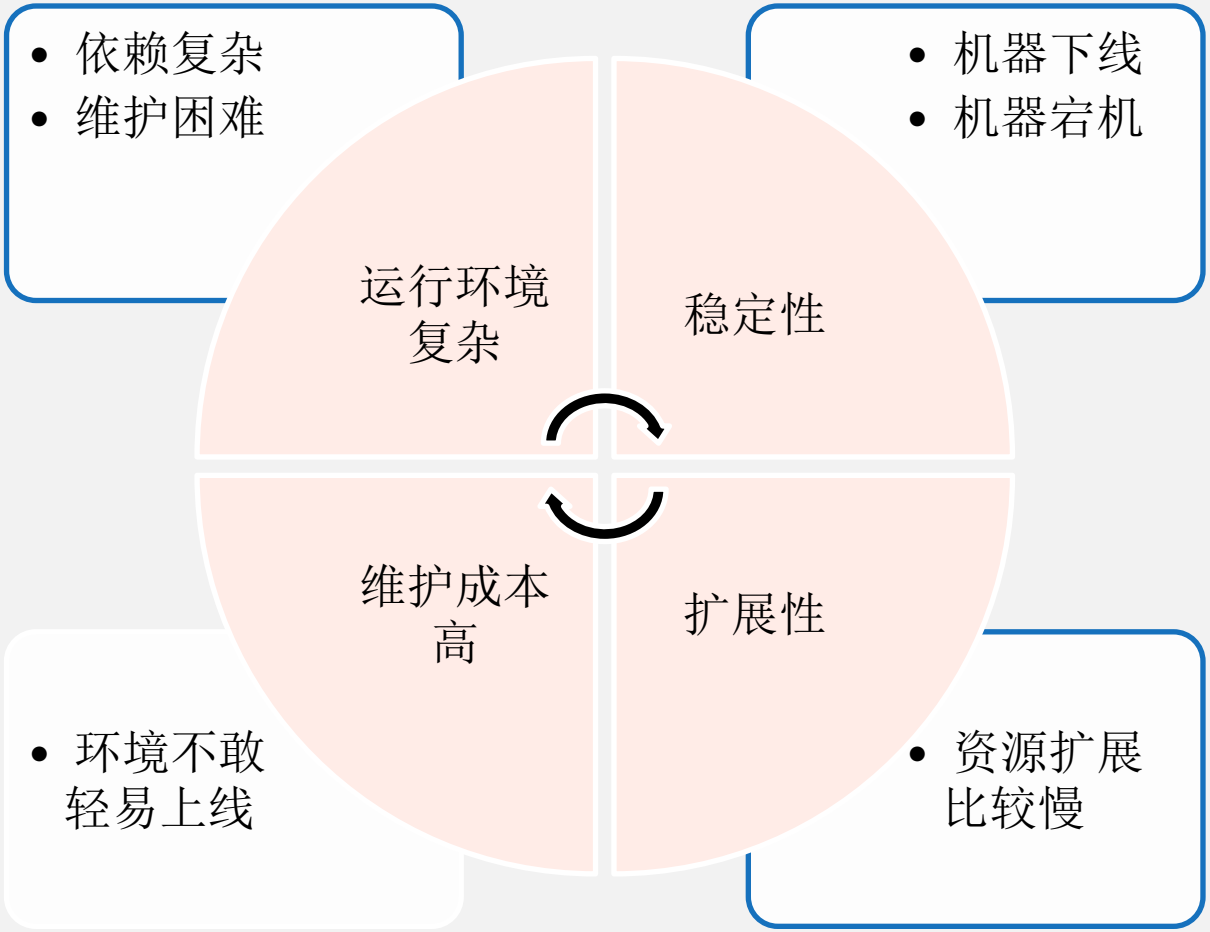
APMCon

- 来自搜狗大数据平台部
- 基于Apache Hadoop生态，建设搜狗海量数据存储和计算平台
- 提供稳定高效的数据分析系统，为搜狗各类型大数据应用，提供一站式数据处理服务
- 每天数十亿的数据增量，数以万计的数据计算流程，使数据的价值得到充分利用
- 最前沿技术落地及推进开源技术的发展

目 录

APMCon

背景



Docker的优势

APMCon

	传统方式	使用Docker
开发	搭建开发环境，解决环境依赖与冲突，重复造轮子	开发者复用Docker Image，减少重复工作
测试	搭建测试环境，偶尔需要开发协助解决环境依赖与冲突	测试人员直接Pull Image
上线	Svn /git tag -> DailyBuild -> OP部署线上Server环境	开发交付上线Image，OP直接拉取上线
总结	• 容易产生系统相关的BUG，不易追查难以定位 • 版本依赖复杂	• 开发、测试、上线环境一致 • 应用使用的库文件，不依赖宿主操作系统

技术选型

APMCon

微服务设计原则

小

- 按照业务职责设计
- 高内聚

轻

- 接口管理
- 数据协议

松

- 耦合性低

技术选型

APMCon

编排系统的选择

Mesos

- 较为完善的长服务解决方案，包括服务发现，负载均衡，资源调度等
- 满足特定业务需要二次开发成本高
- 整套解决方案的开发语言较多

Kubernetes

- 理念先进复杂，功能相对完善复杂
- 满足特定业务需要二次开发成本高

Yarn

- 一定的技术积累，Hadoop集群结合支持统一集群调度
- 功能不完善，需要开发服务发现，Load Balance等基础功能

Docker在Sogou的应用

APMCon

线上业务

- 企业搜索核心模 块，搜索相关核心，以及前端系统，索引建库等

线下业务

- 调研流程，各类线下实验平台，GPU深度学习平台，VR展现等

Clotho

- Hadoop离线任 务管理，主要涉及相关性rank，意图识别反作弊，数据统计等
- Long Running Service，如ELK，Grafana Sensu监控报警系统，ganglia等

同时运行的Docker Container达到万级以上，每天完成近百万离线Job，托管了上千类型的Long Running Service

Clotho系统的功能与总体框架

APMCon

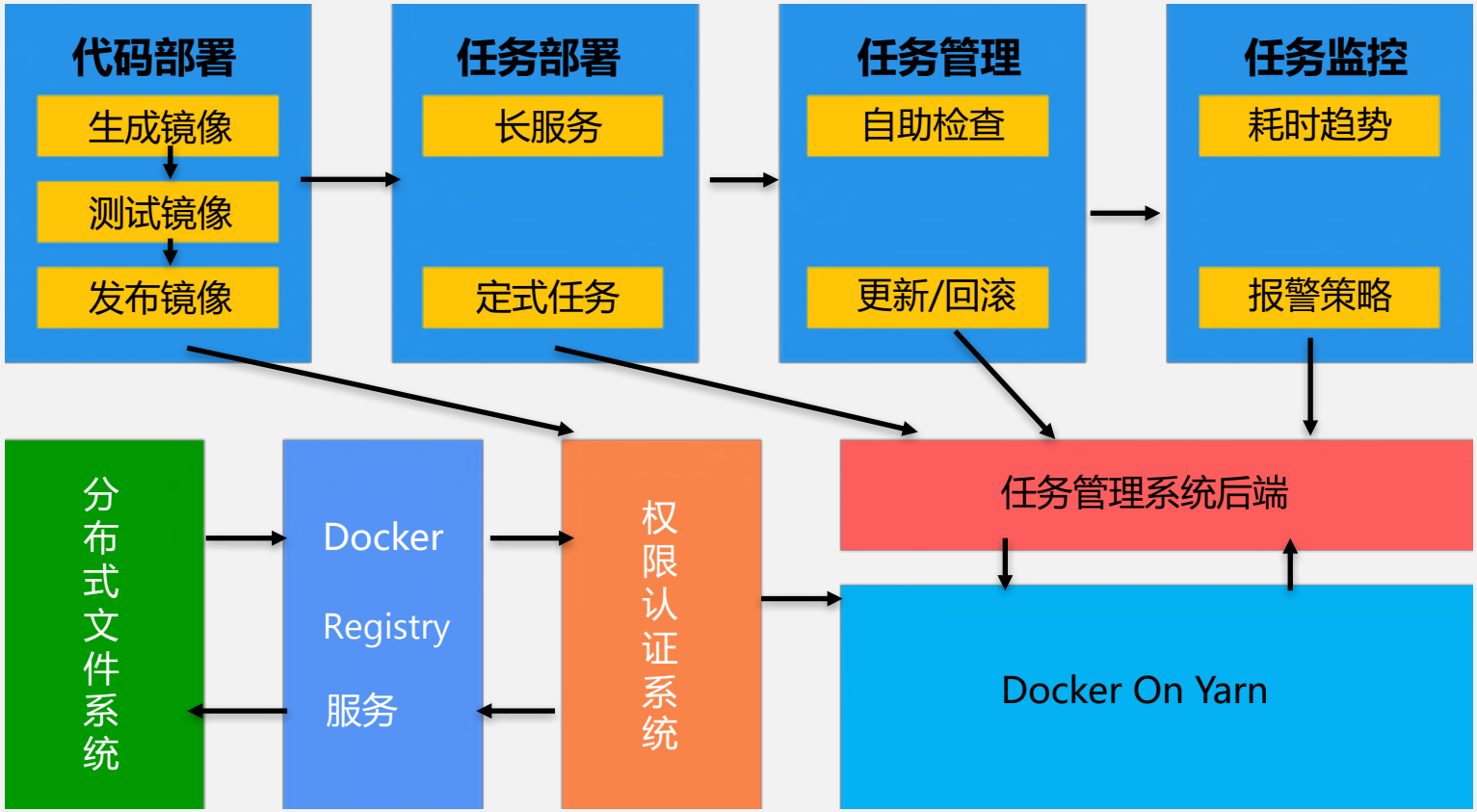
Clotho系统的总体功能

```
usage: cloudtask.py [-h] {hdfs,run,get,init,debug,commit,login} ...

positional arguments:
  {hdfs,run,get,init,debug,commit,login}
    commit              更新本地代码到任务执行代码(upload the
                        code)
    init                创建/初始化任务(Create a new task/Init a exists
                        task)
    login               用户登录(login)
    get                 获取任务管理系统代码到本地目录(Get code
                        from remote)
    run                 执行本地任务(Run the task once)
    hdfs                显示hdfs路径(list hdfs path)
    debug               远程调试任务运行环境(Debug the runtime of
                        task)

optional arguments:
  -h, --help            show this help message and exit
```

Clotho系统的功能与总体框架

APMCon

Clotho系统总体框架

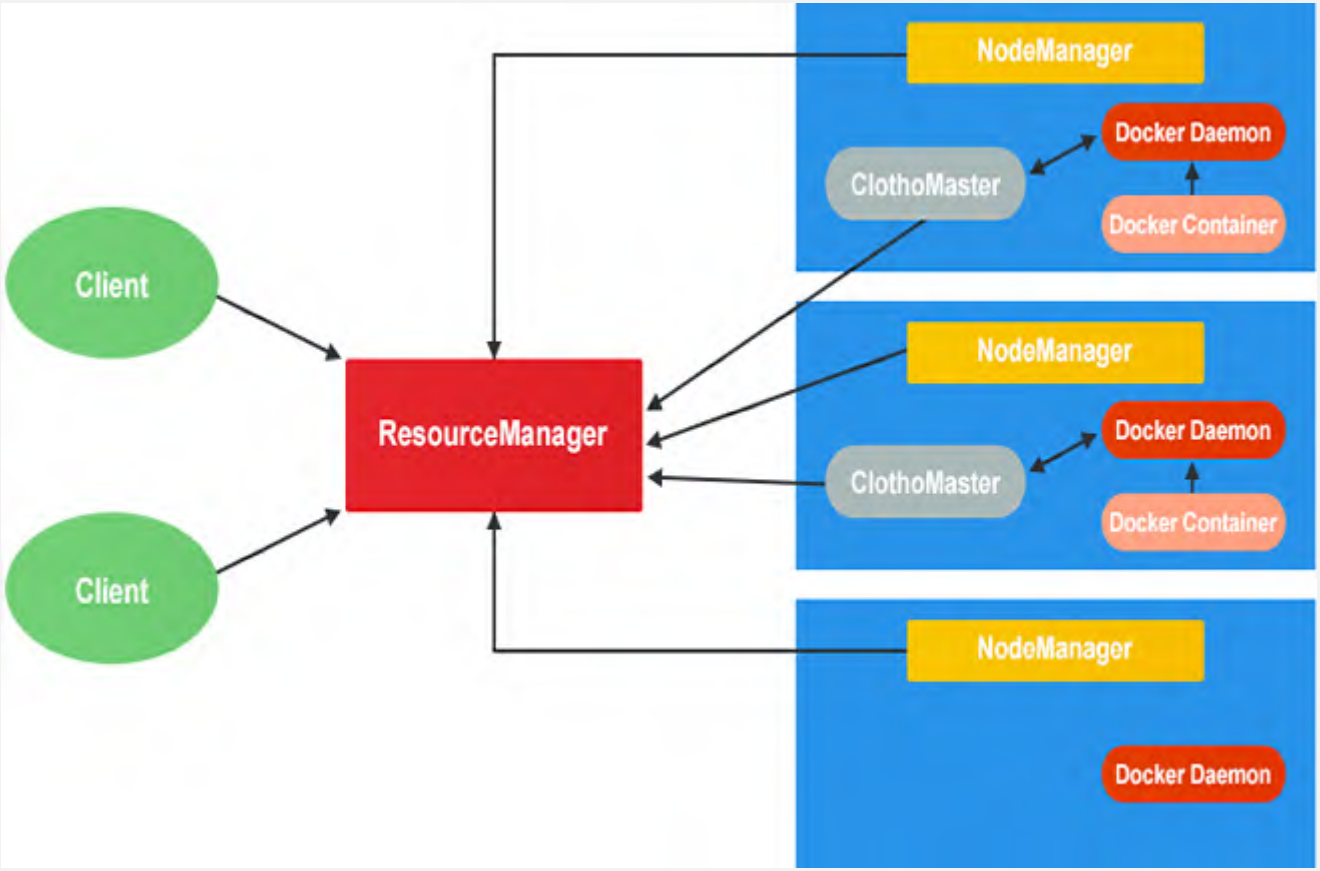


Clotho系统的功能与总体框架

APMCon

Docker On Yarn组件

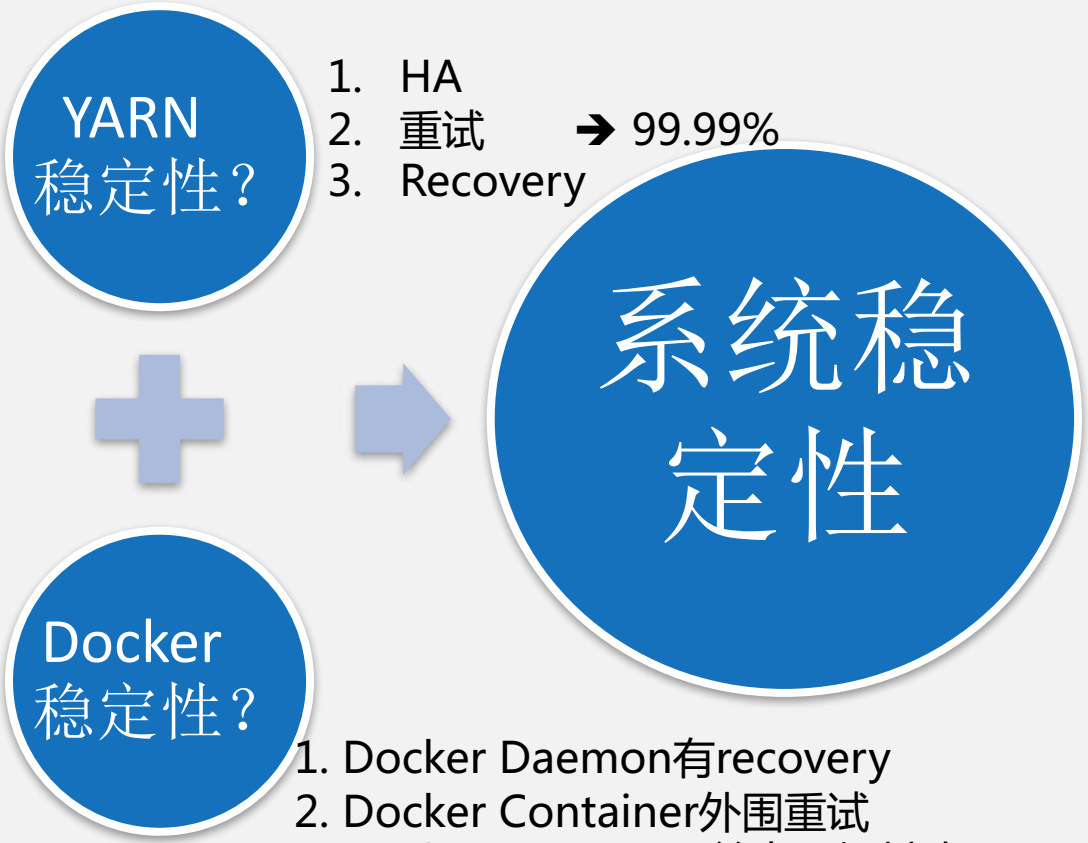
https://github.com/sogou/docker-on-yarn.git



Clotho系统的功能与总体框架

APMCon

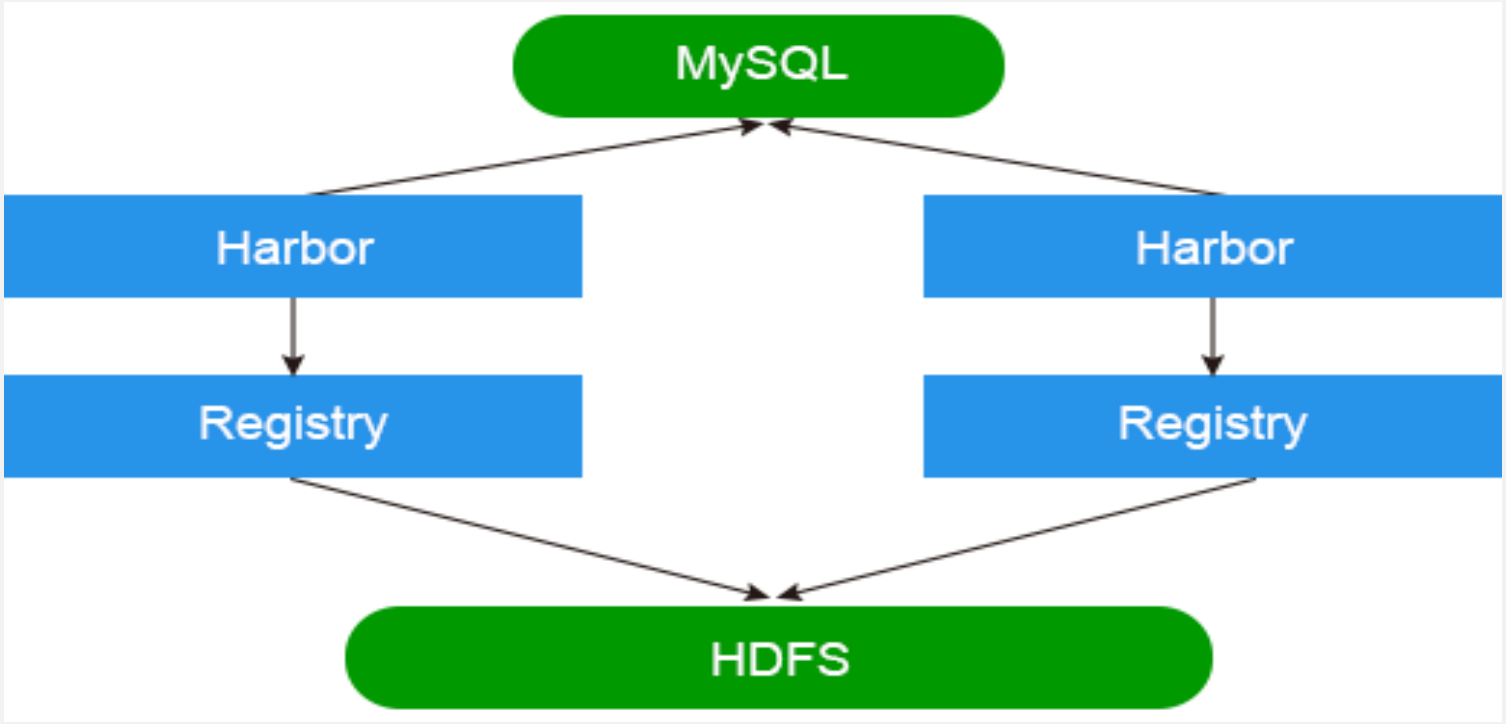
稳定性讨论



稳定性讨论

APMCon

Docker Registry改进



- 支持HDFS持久化存储，保证数据的多备安全性
- 支持分布式Harbor集群，支持LDAP权限认证
- Docker Registry分布式化，且支持负载均衡

稳定性讨论

APMCon

其他改进

- Docker Container随机端口管理，解决冲突问题
- Docker Daemon升级到1.13.1解决假死问题
- YARN解决机器负载/权限问题
  - 对机器进行区分，保证Master机器的稳定性，开发指定机器调度
  - 指定一组机器进行调度，开发指定Label的调度模型
  - 解决网络传输问题，开发指定rack调度
  - 解决YARN recovery/AM黑名单等问题

经过针对Docker和YARN的改进后，Clotho系统的稳定性达到99.9%

Clotho微服务管理平台

APMCon

主要功能

- 服务SPEC配置管理

  AppName：唯一标识

  Image：实际运行的Image信息

  Tag：标识服务状态

  Port：发布的端口

  Instances：启动的Service数目

- 负载均衡

  多实例情况下简单随机，含地域特性

- 服务发现

  服务定位LocateServer功能

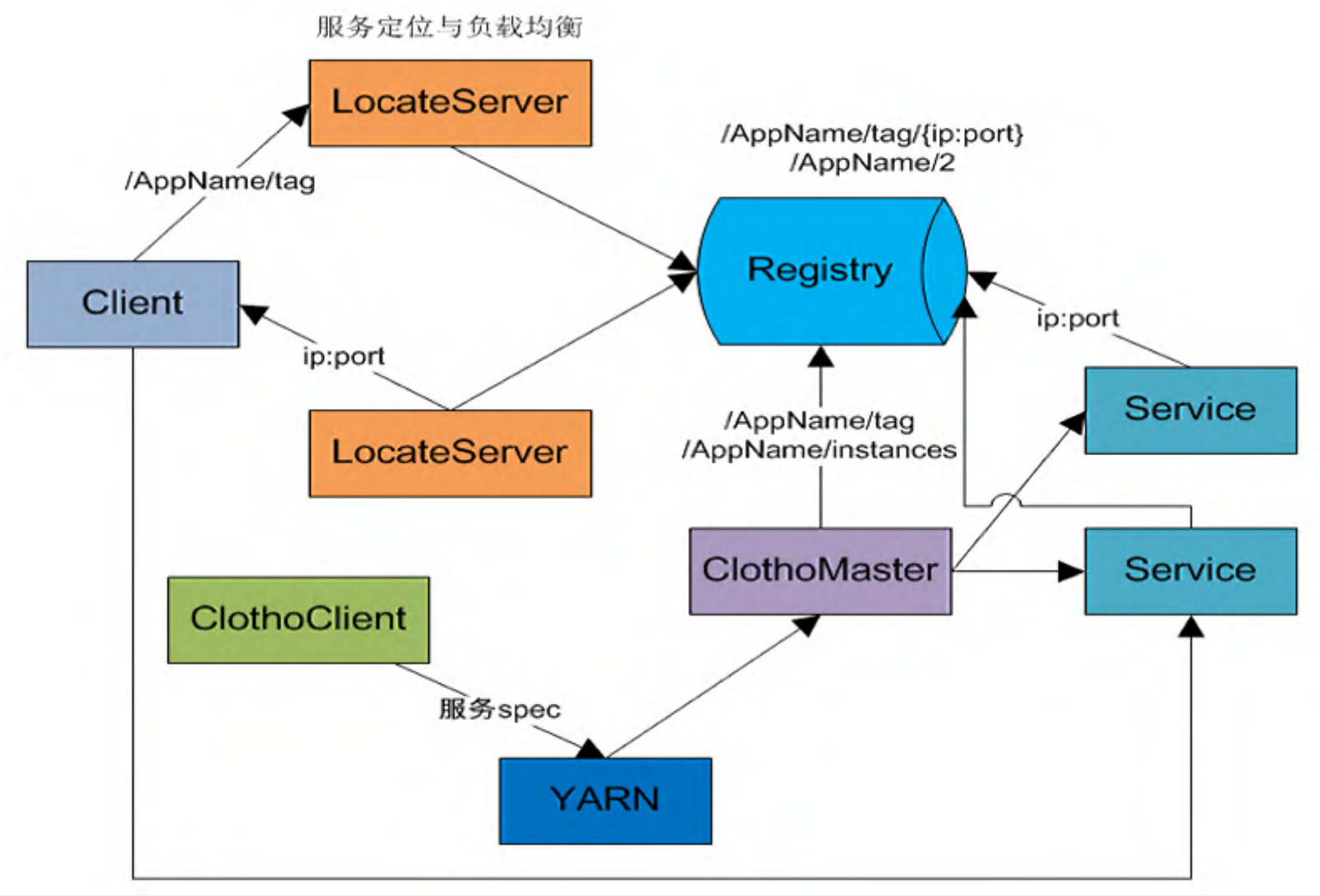
- 弹性伸缩

  直接修改Registry Service里 /AppName/instances，ClothoMaster根据instances动态调整申请的资源

Clotho微服务管理平台

APMCon

Client与Server的交互流程



Clotho微服务管理平台

APMCon

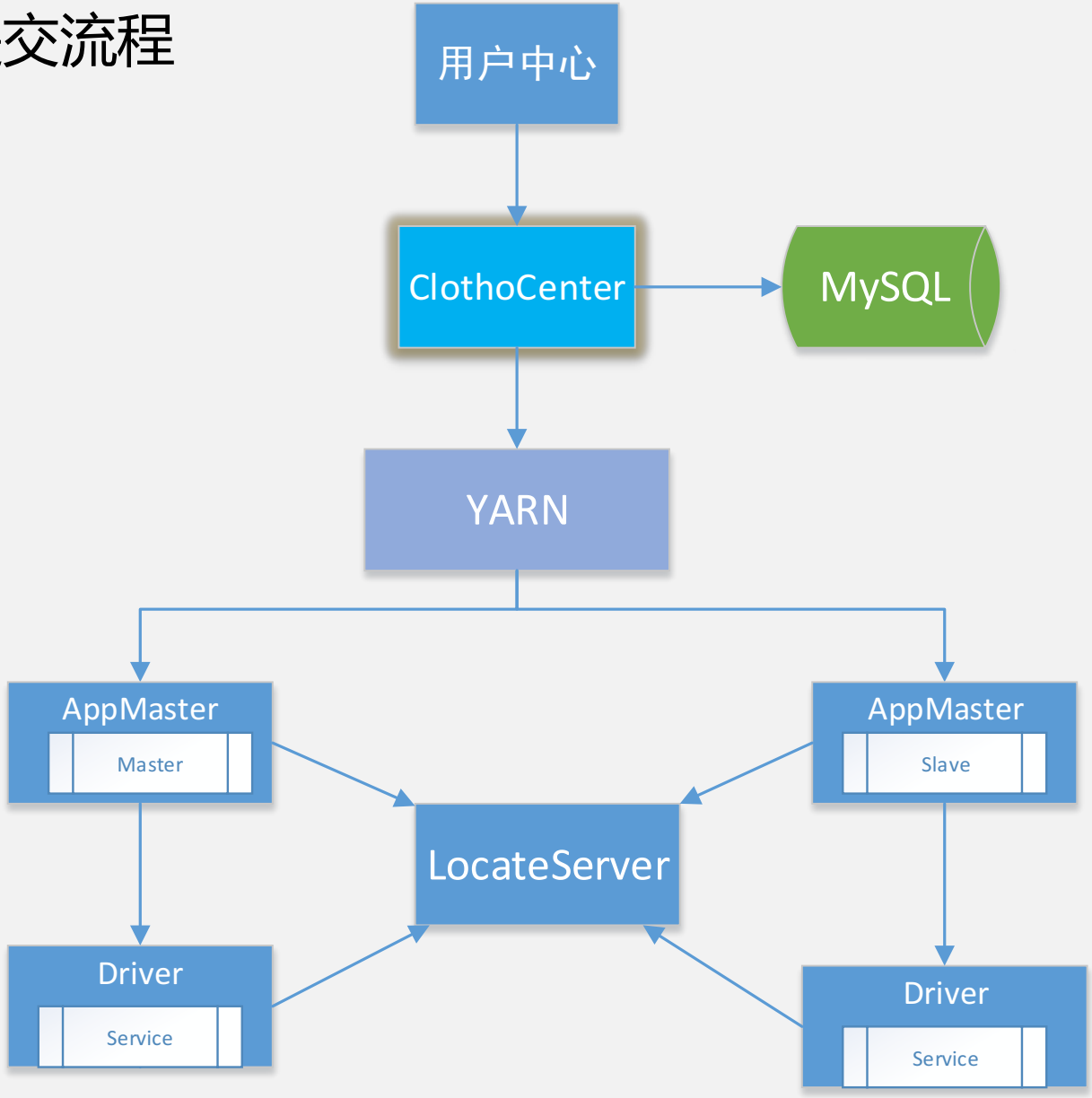
```
appName：SERVICE_MASTER_NAME
containers:
      - image: master-example:1.0
        memory: 2g
        cores: 2
        command: sh bin/start.sh
        ports:
           - port: CLOTHO_RANDOM_PORT
        hosts: master01,master02
        tag: MASTER
        files:
            - /etc/example/master.conf
   instances:
     - num: 1
```

```
appName：SERVICE_SLAVE_NAME
containers:
      - image: slave-example:1.0
        memory: 1g
        cores: 1
        command: sh bin/start.sh
        ports:
          - ports: 80
        tag: SLAVE
 instances :
      - num: 1
    dependency:
      - name: SERVICE_MASTER_NAME
        type: RESTART_IMMEDIATELY
        files:
          - /etc/example/slave.conf
```

Clotho微服务管理平台

APMCon

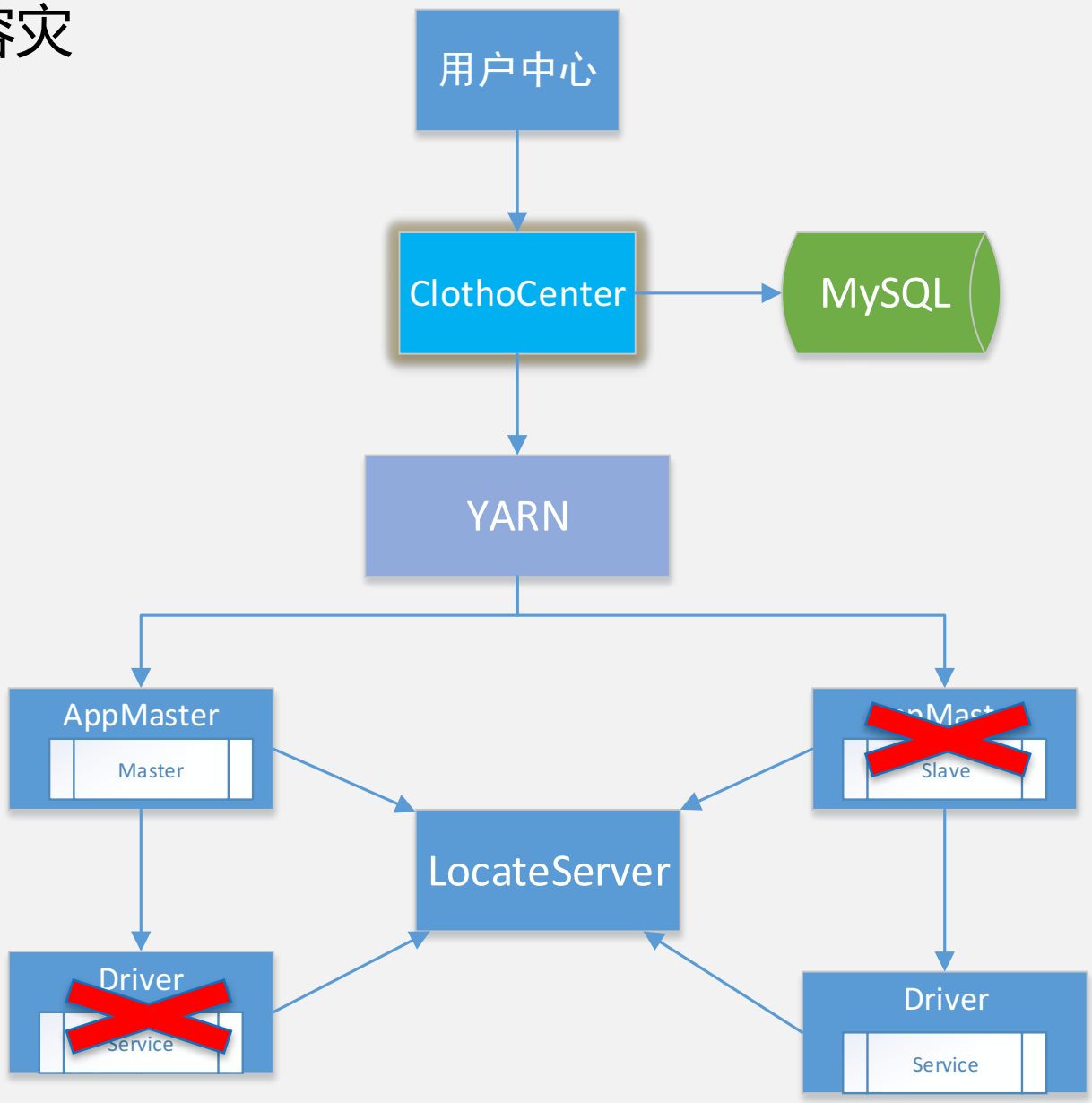
任务提交流程



Clotho微服务管理平台

APMCon

服务容灾



服务容灾

服务实例service异常：

1. YARN通知AppMaster Driver已经退出
2. AppMaster重新申请资源启动Driver
3. 新Driver启动后向LocateServer更新服务地址
4. 下游服务AppMaster发现依赖服务变动，依据重启策略向Driver通知重启服务

AppMaster服务异常：

1. 上游AppMaster异常退出，YARN重新调度新实例new-AppMaster
2. new-AppMaster重新托管Driver
3. new-AppMaster确认所有Driver正常后向LocateServer更新信息
4. 下游AppMaster发现上游任务变动，依据策略通知Driver是否重启

TODO

APMCon

I. 支持复杂DAG

II. 支持更好的资源隔离策略

III. 与Hadoop集群结合，支持离线和在线业务

THANK YOU