分布式存储与离线混部弹性计算平台

Sogou　申贤强

About us

- 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

Catalogue

目录

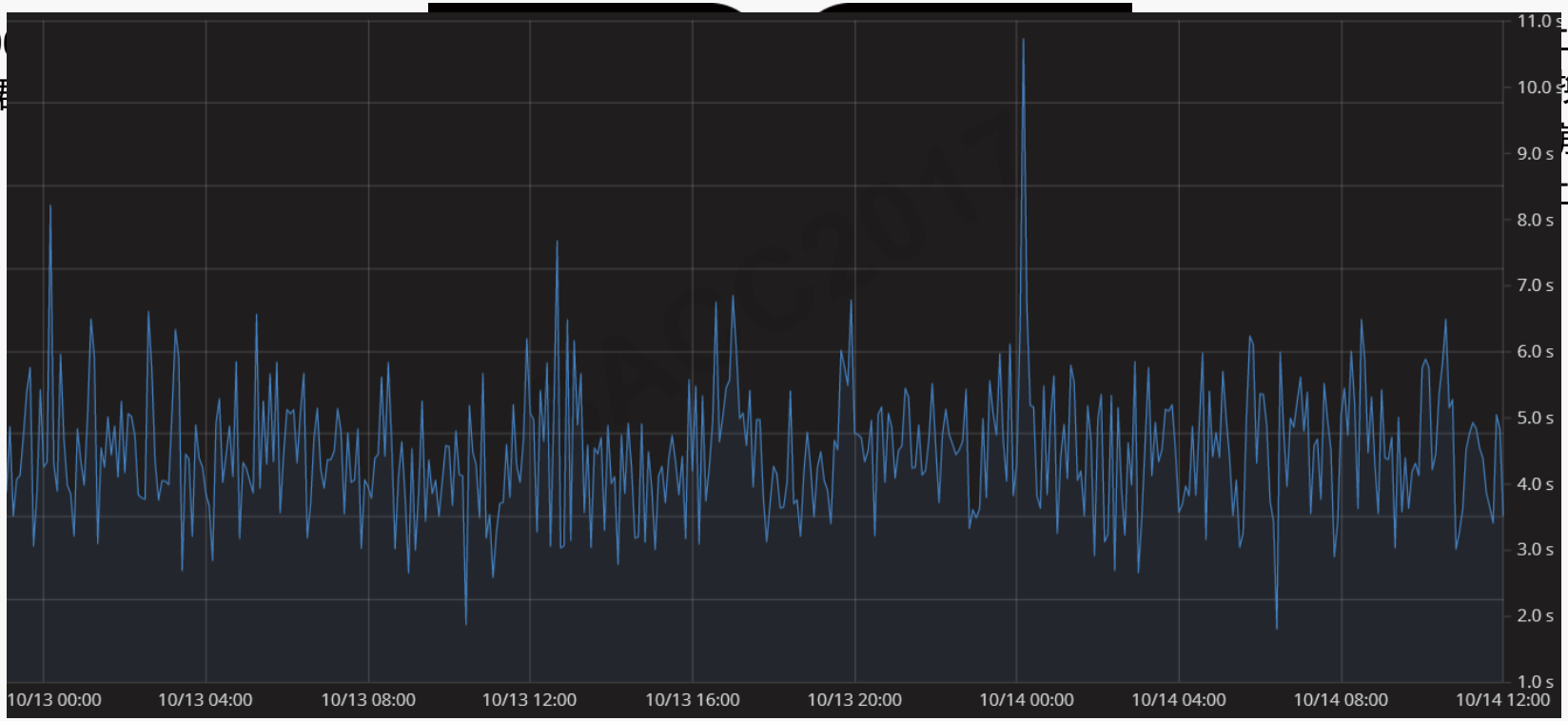
1 背景

Background

Background

背景

10
集群



换

崩溃

2 技术选型

Technology selection

技术选型

Technology selection

- 考察业内的选择
  - 国外主流企业
  - 国内互联网企业
- 目的实现集群的无限扩容，提高性能
- 最终选择

  借鉴和自研的垂直扩展Hadoop元信息的技术，即社区的Fedration方案，将集群的管理能力扩展到理论无上限，且保持高性能，稳定性达到99.99%

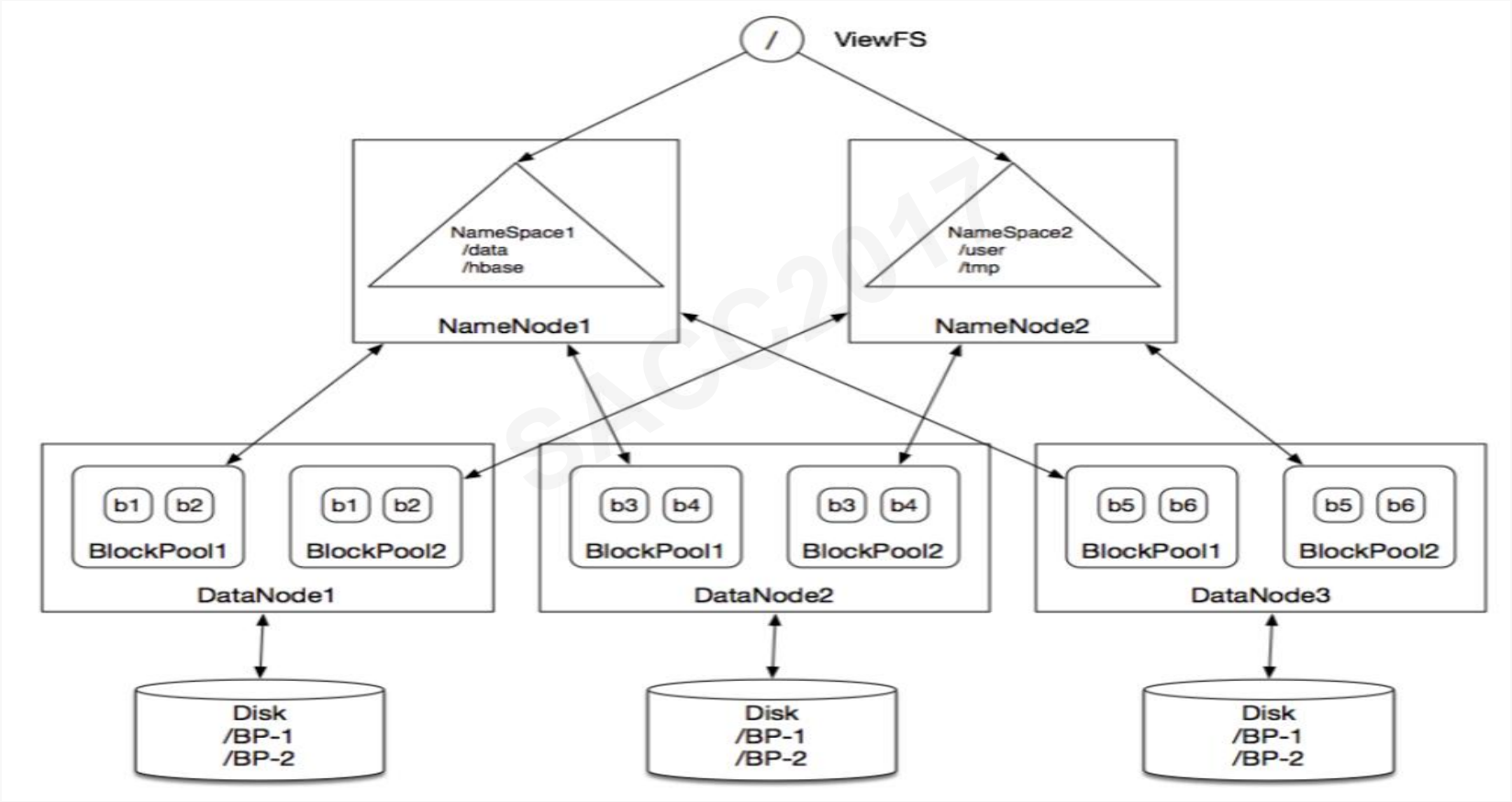
3 分布式存储优化

Distributed storage optimization

Distributed storage optimization

分布式存储优化

1. HDFS Fedration



Distributed storage optimization

分布式存储优化

2. NameService拆分

业务原则建议

- 数据盘：原始日志
- 业务盘：/user目录
- 其他：如mr_history,logs等

重要性原则

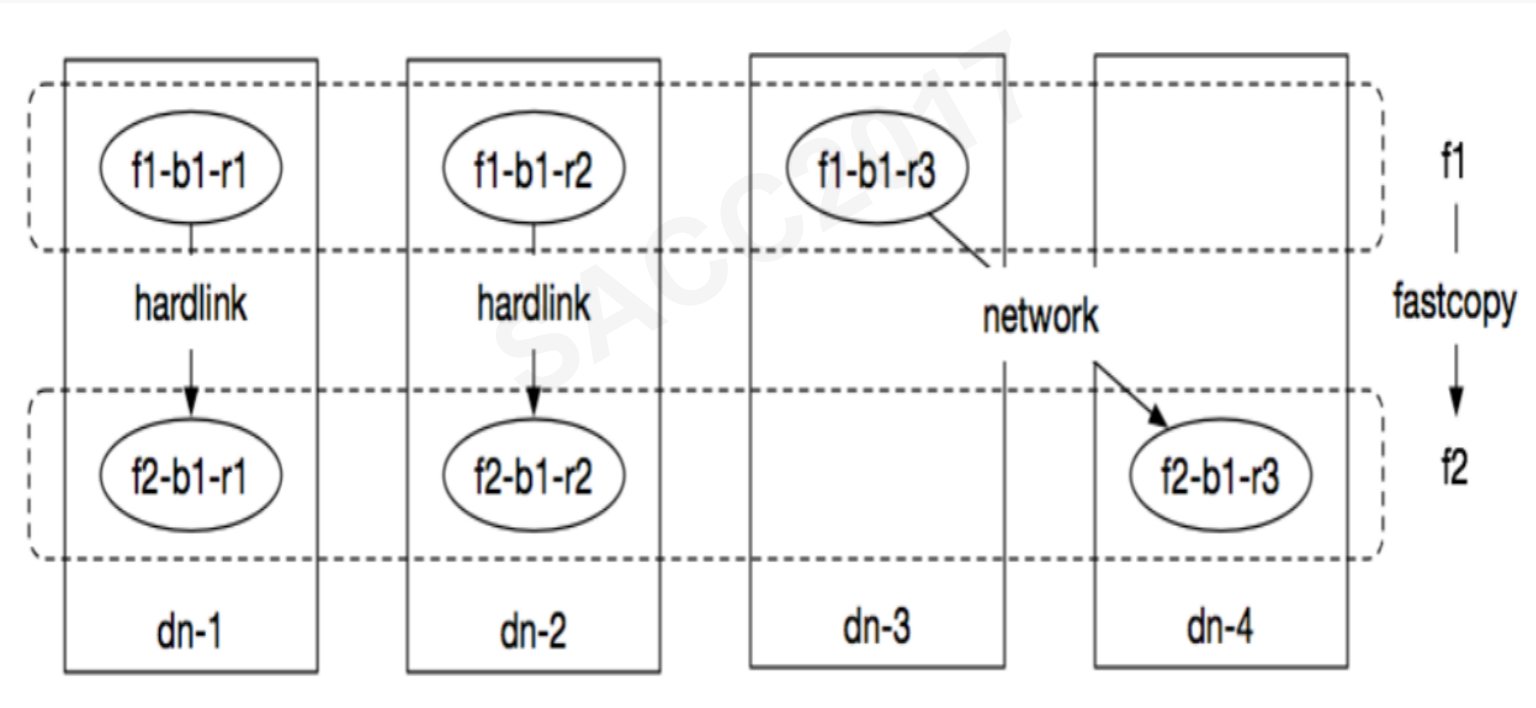
- Online：重要业务，HDFS稳定性要求比较高
- Offline：离线处理业务

Distributed storage optimization

分布式存储优化

3. FastCopy

https://issues.apache.org/jira/browse/HDFS-2139



Distributed storage optimization

分布式存储优化

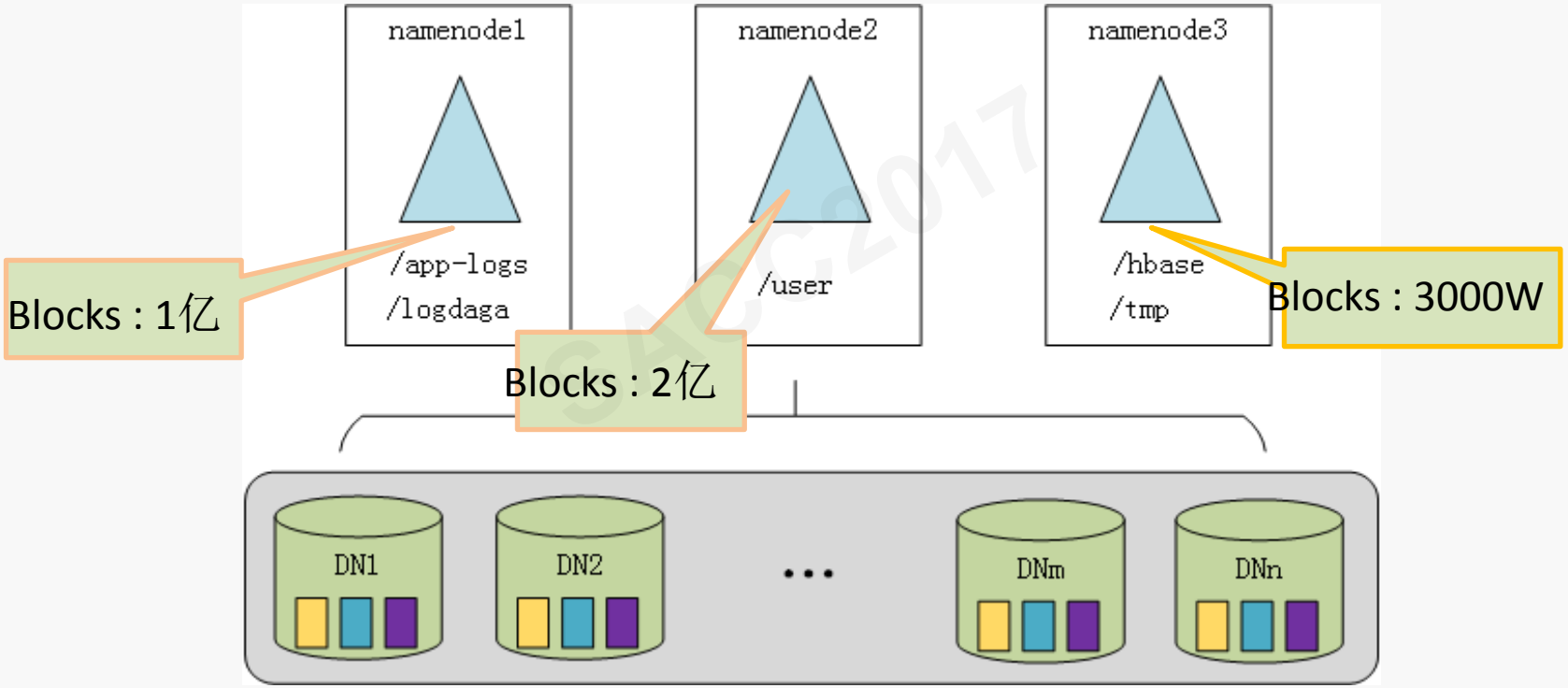
4. 切分基本流程

- 不停服务全量拷贝
  - 关闭HDFS Balancer
  - Split，将输入目录进行切分，建议粒度比如为500
  - distribute fastcp，启动分布式fastcp，建议低并发
  - 权限管理，目录chown/chmod管理
  - Checksum校验，结果进行校验，防止出错
- 增量迁移数据

Distributed storage optimization

分布式存储优化

5. 二级目录拆分 (现实却往往不理想)

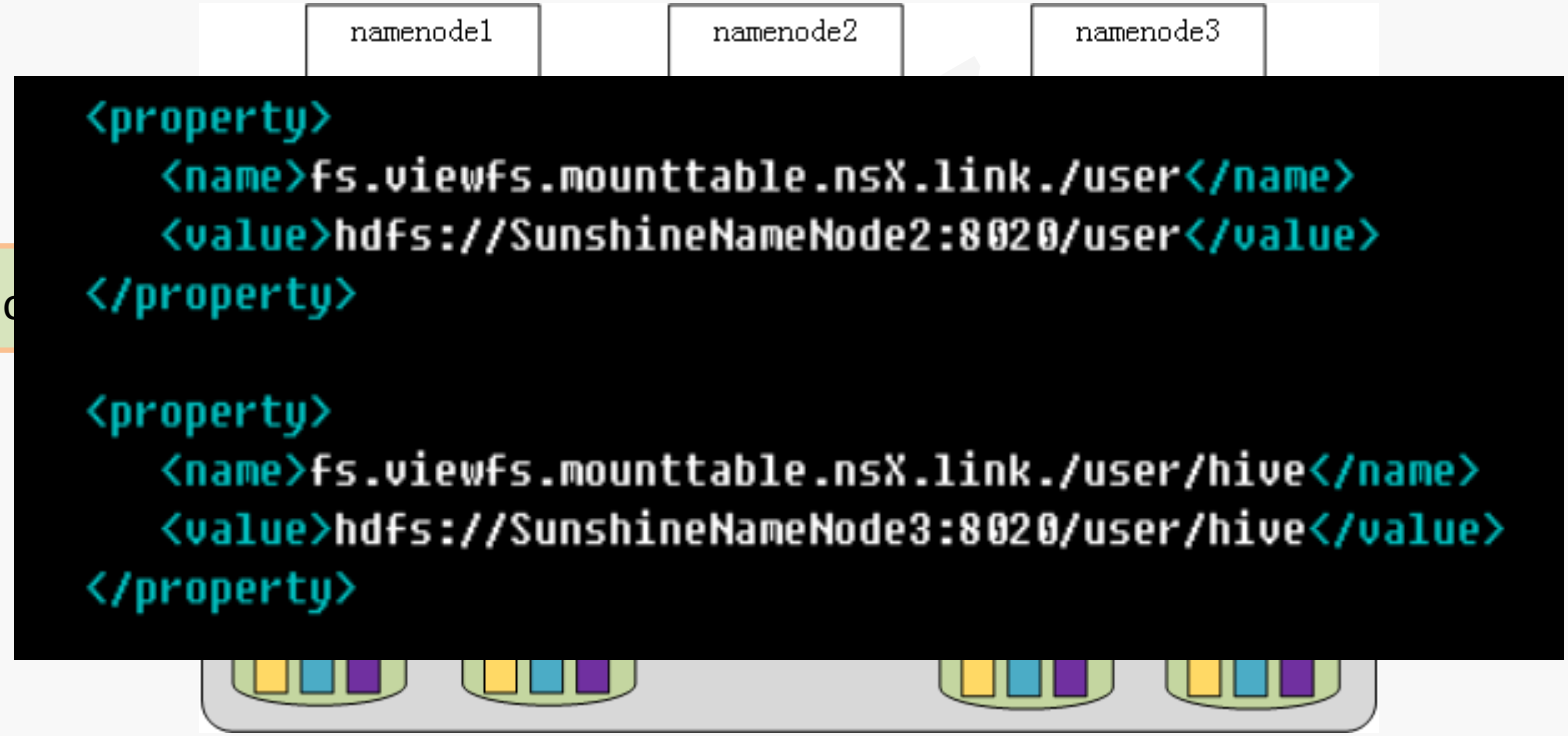


Distributed storage optimization

分布式存储优化

5. 二级目录拆分

https://issues.apache.org/jira/browse/HDFS-12555

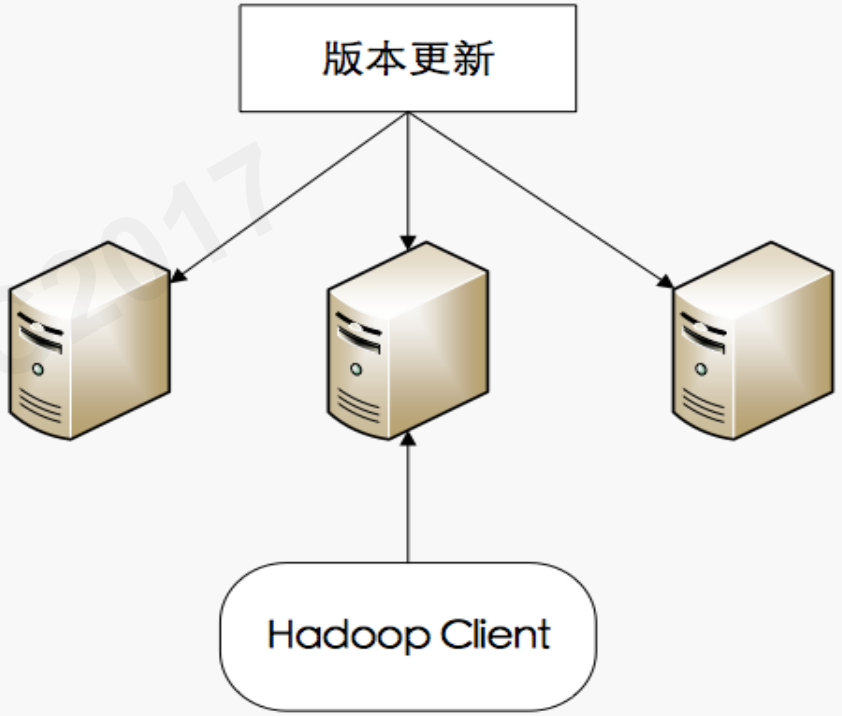


Distributed storage optimization

分布式存储优化

6. 自更新Client Mount table

- 地域特性，就近原则
- 版本控制
- 增量更新需要的Jar
- 更新conf本地优先
- 支持conf强制更新

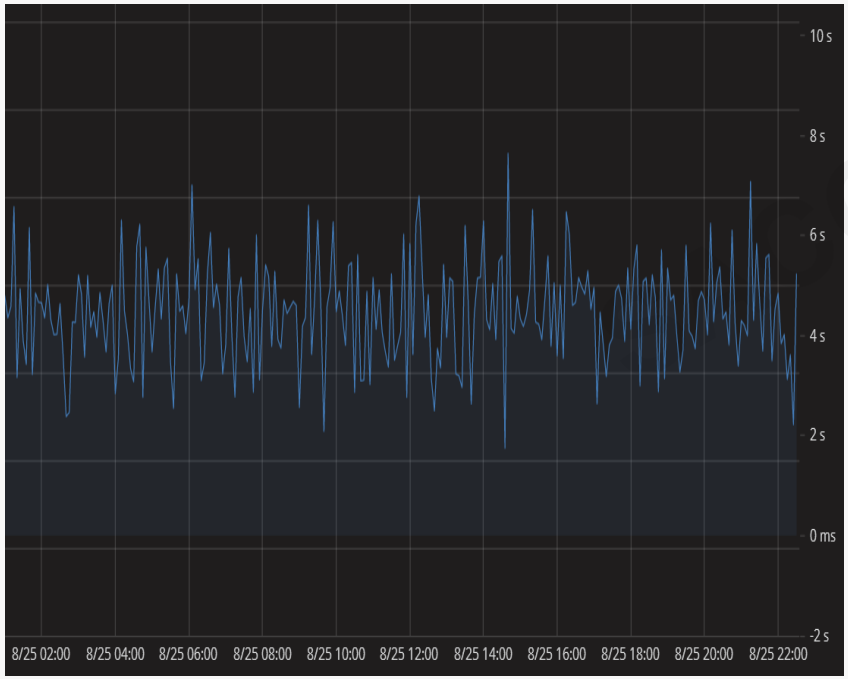


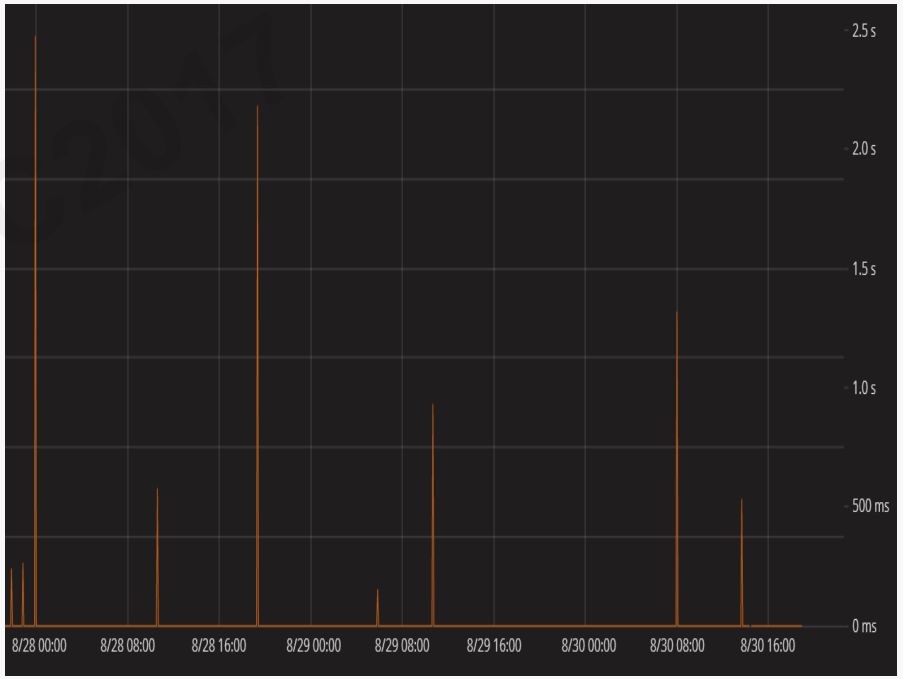
Distributed storage optimization

分布式存储优化

优化后的性能

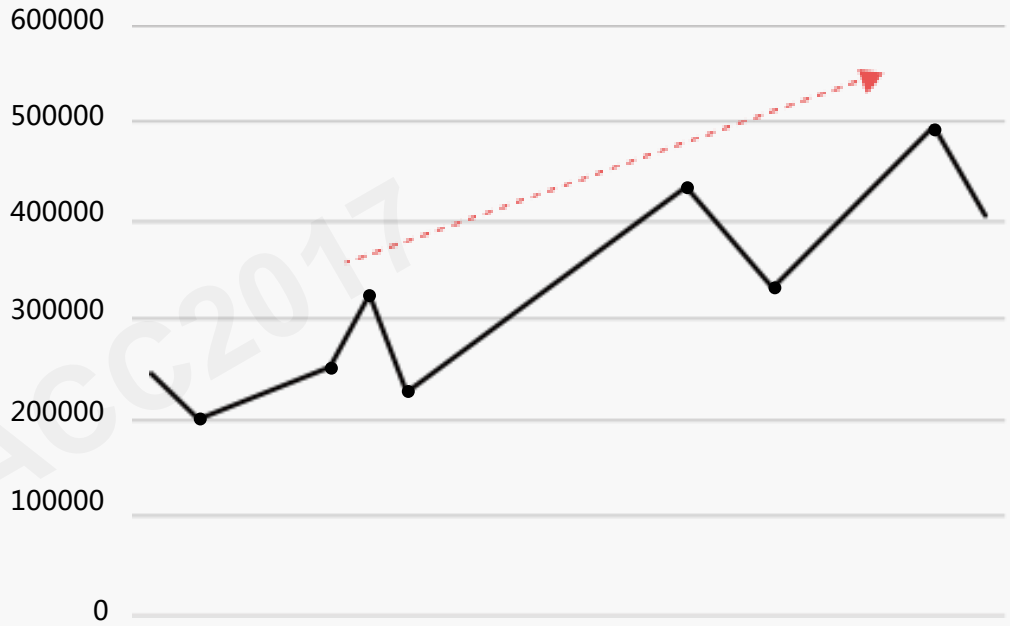
- NameNode的响应时间由平均秒级降低到毫秒





分布式存储优化

Distributed storage optimization



优化后的性能

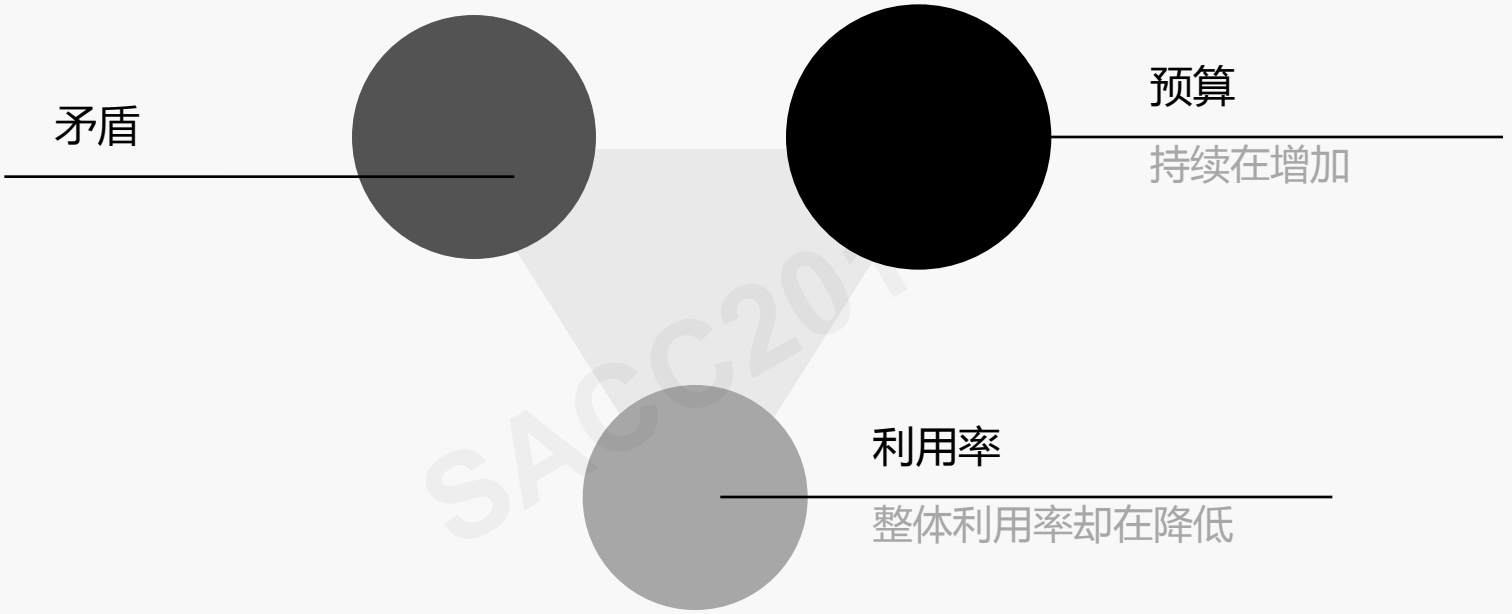
- Master的性能和吞吐提升3倍
- 计算性能提升12%以上
- SLA水平到达99.99%

4 弹性计算平台背景

Elastic computing platform background

Elastic computing platform background

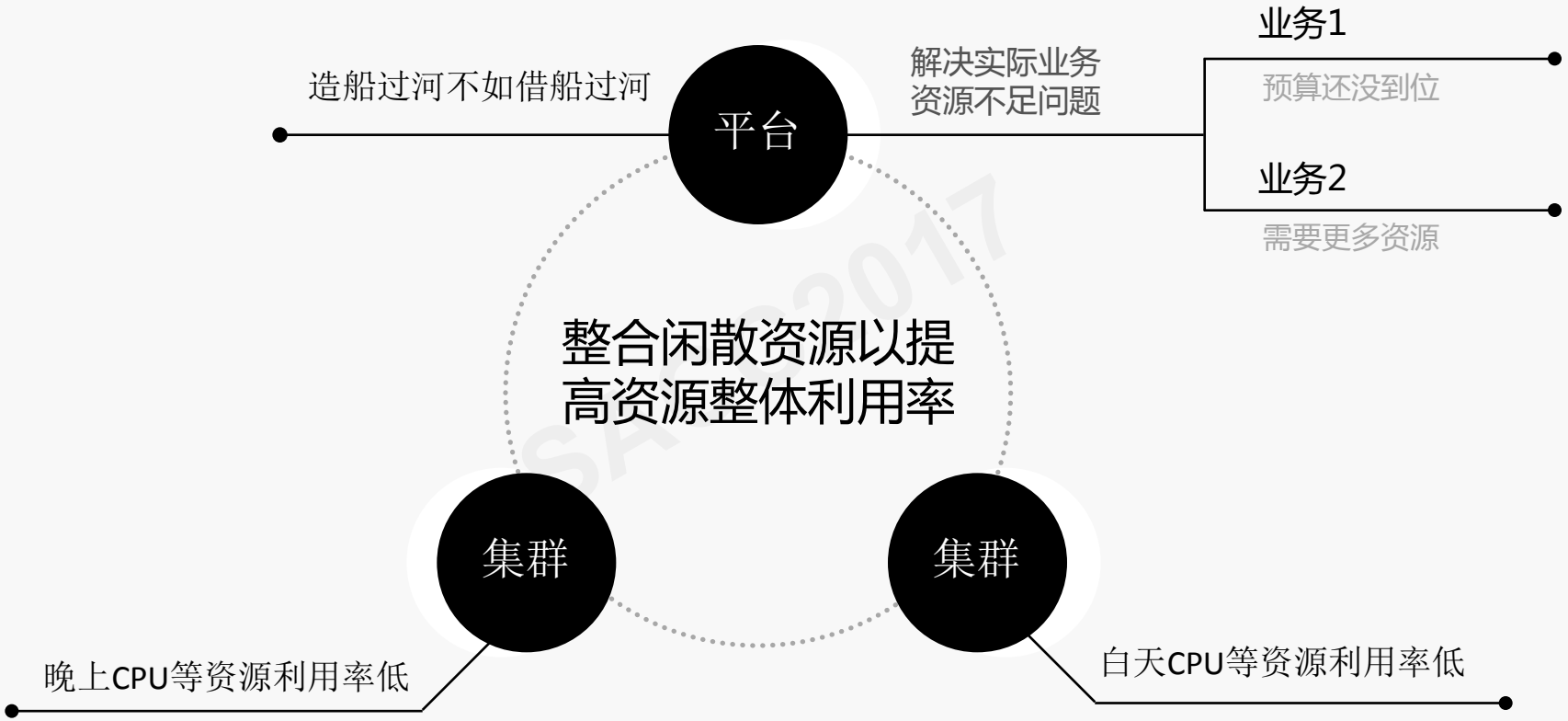
弹性计算平台背景



资源需求在增加，但资源的整体利用率却不高

Elastic computing platform background

弹性计算平台背景



弹性计算平台—Pythagoras

Elastic computing platform—Pythagoras

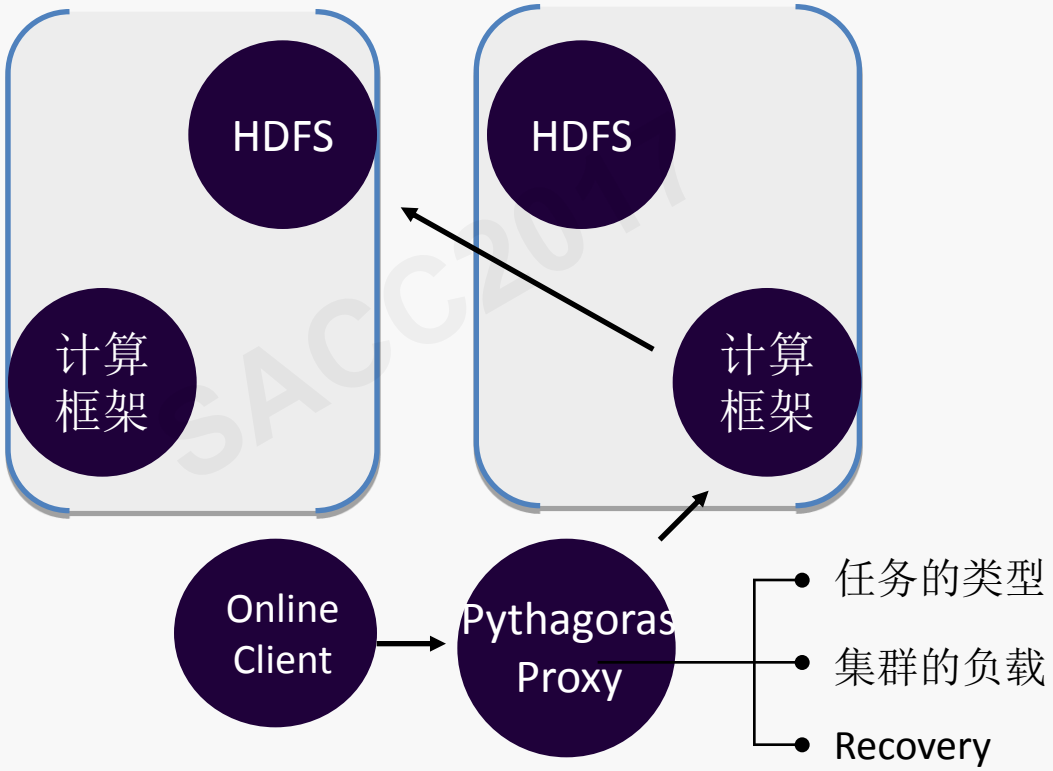
1. Hadoop任务弹性计算

- Hadoop资源隔离性差
  - 任务间的影响造成高优先级业务SLA无法得到保证
- 集群按重要性划分多个
  - online重要业务集群
  - offline离线业务集群
- 集群间HDFS不共享
- 根据offline集群的资源负载弹性计算
- 对用户是透明的

Elastic computing platform —Pythagoras

弹性计算平台—Pythagoras

1. Hadoop任务弹性计算

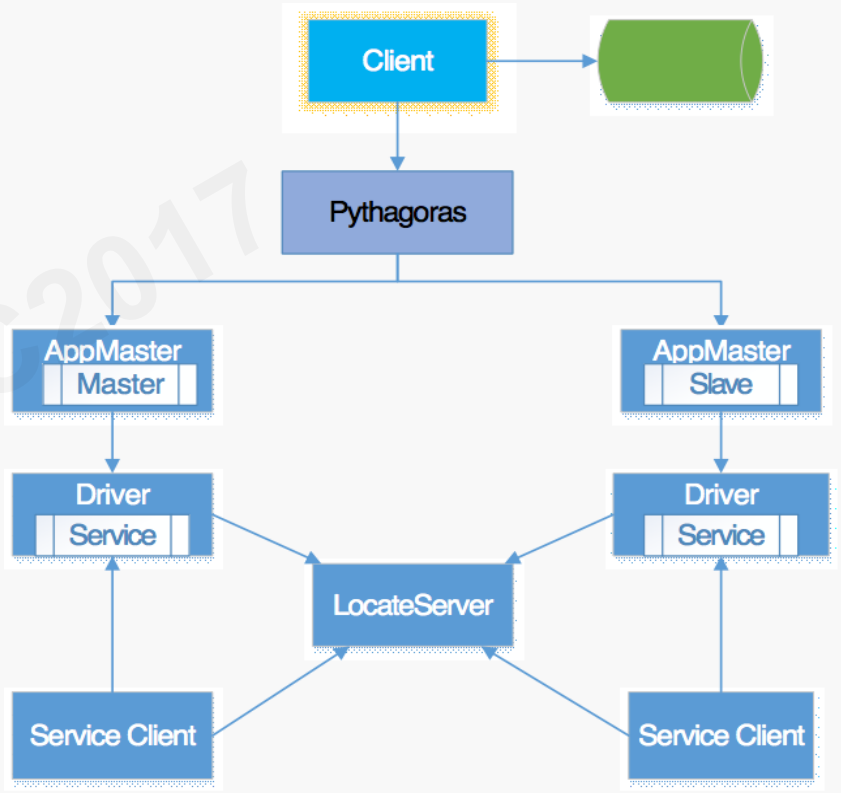


Elastic computing platform —Pythagoras

弹性计算平台—Pythagoras

2. C/S业务弹性计算—总体流程

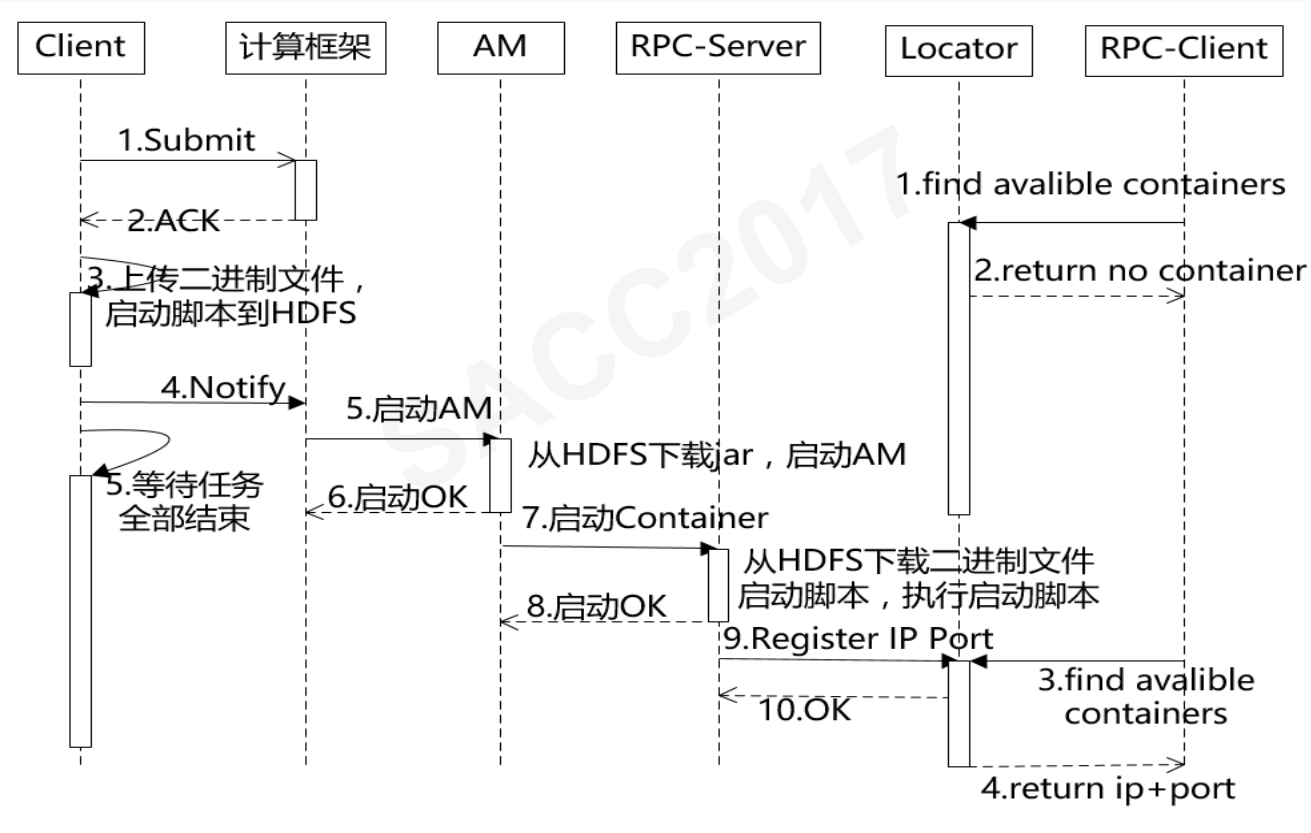
- Driver精细的资源控制
- Docker环境隔离
- YARN自动化资源控制
- 基于时间/负载的资源调度



Elastic computing platform —Pythagoras

弹性计算平台—Pythagoras

2. C/S业务弹性计算—总体流程

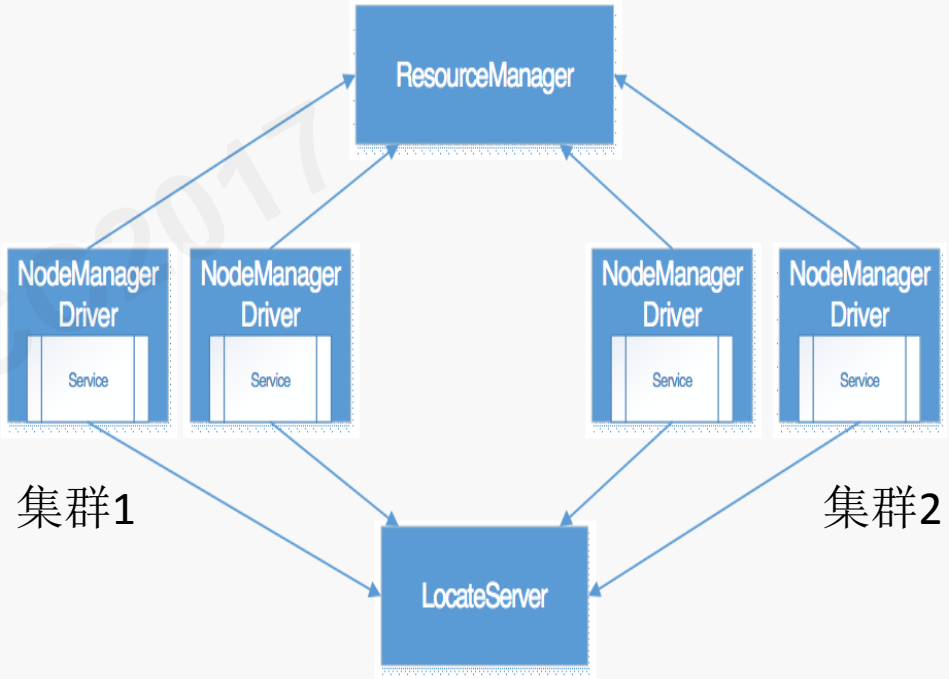


Elastic computing platform —Pythagoras

弹性计算平台—Pythagoras

2.C/S业务弹性计算—基于负载的调度

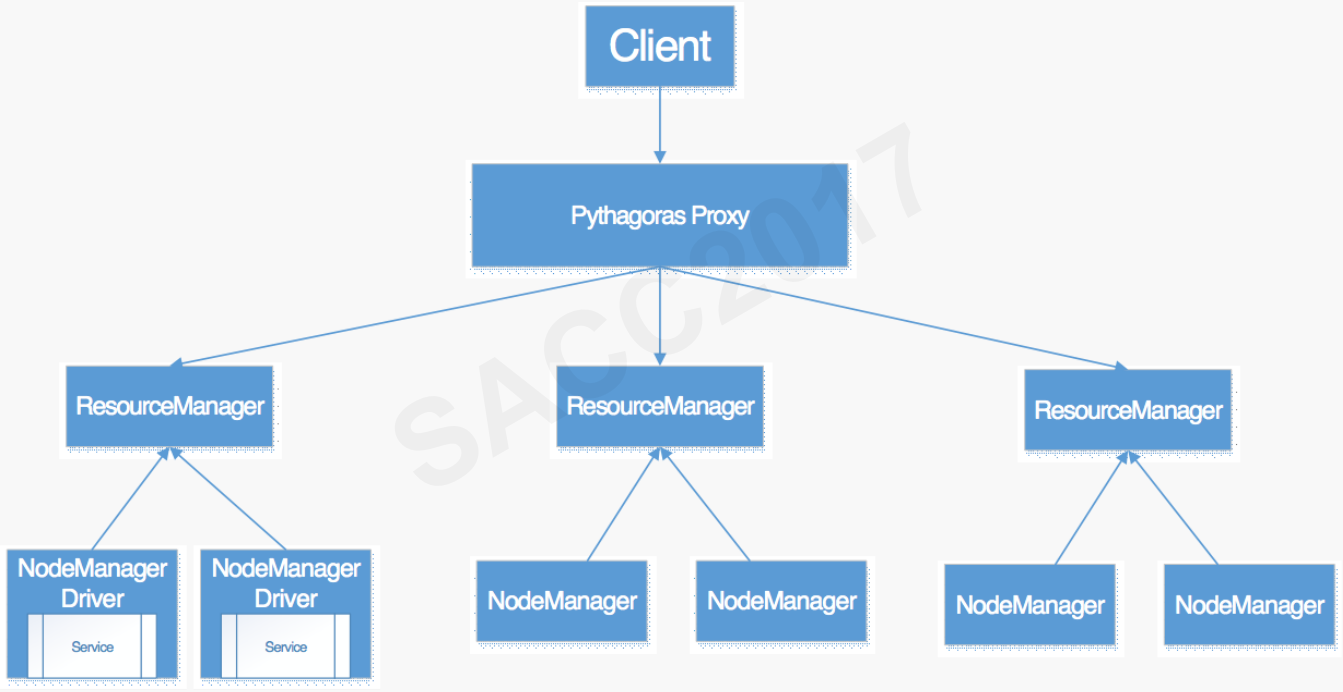
- Driver收集CPU/Mem/Net/Disk负载
- 负载超过阈值Driver Kill Service
- Driver将负载信息上报给LocateServer
- NodeManager将负载上报给RM
- RM根据负载进行资源调度
- LocateServer根据负载返回Service Client
- 不影响集群原有服务
- 提高集群的利用率



Elastic computing platform —Pythagoras

弹性计算平台—Pythagoras

3.集群管理



弹性计算平台—Pythagoras

Elastic computing platform—Pythagoras

集群统一管理优势

- 提高并均衡集群利用率
- 解决业务方资源不足的问题
- 节约成本
- 提高online业务的SLA水平

6 TODO

TODO

Docker支持核心搜索服务

支持合理的资源隔离策略

支持更加合理的基于负载调度策略

支持Yarn Fedration

TODO

TODO


THANKS