www.qconferences.com
www.qconbeijing.com

QCon

QCon北京2014大会 4月25—27日

伦敦 | 北京 | 东京 | 纽约 | 圣保罗 | 上海 | 旧金山

London · Beijing · Tokyo · New York · Sao Paulo · Shanghai · San Francisco

QCon全球软件开发大会

International Software Development Conference

www.infoq.com/cn

InfoQ

@InfoQ

infoqchina

软件
正在改变世界！

Alibaba
technology
Association

Hadoop2.0应用
基于Yarn的淘宝海量数据服务平台

曹龙@封神
阿里巴巴集团-海量数据
微博：封神无度

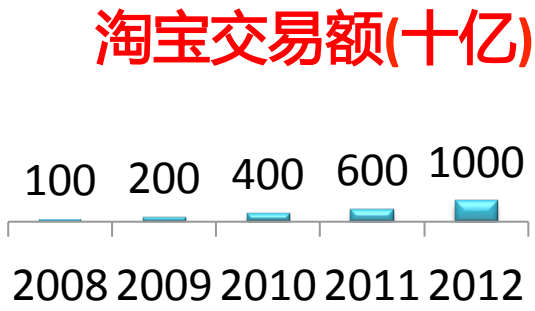
Alibaba
technology Association

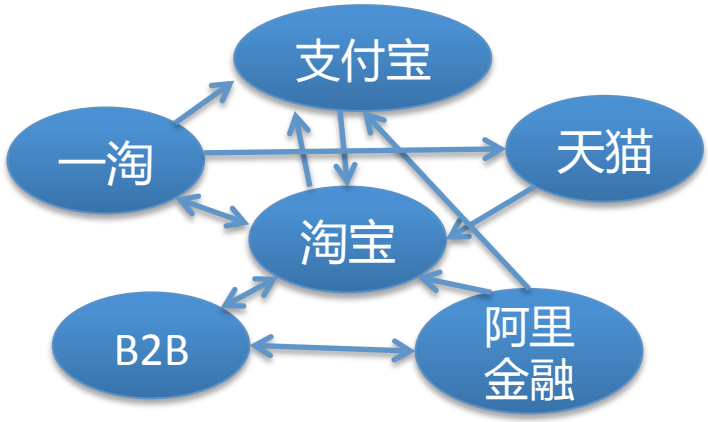
演讲提纲

- 大数据
- 目前的云梯现状
- YARN的介绍
- YARN在云梯的应用
- 未来展望

大数据

- 数据的价值
  - 阿里的三个发展阶段: 平台、金融、数据
- 数据增长趋势
  - 用户、商品、交易
- 数据的复杂度
  - 子公司众多
  - 业务逻辑复杂并相互依赖





Alibaba
technology Association

演讲提纲

- 目前的云梯现状
- YARN的介绍
- YARN在云梯的应用
- 未来展望

Alibaba
technology Association

演讲提纲

- 目前的云梯现状
- YARN的介绍
- YARN在云梯的应用
- 未来展望

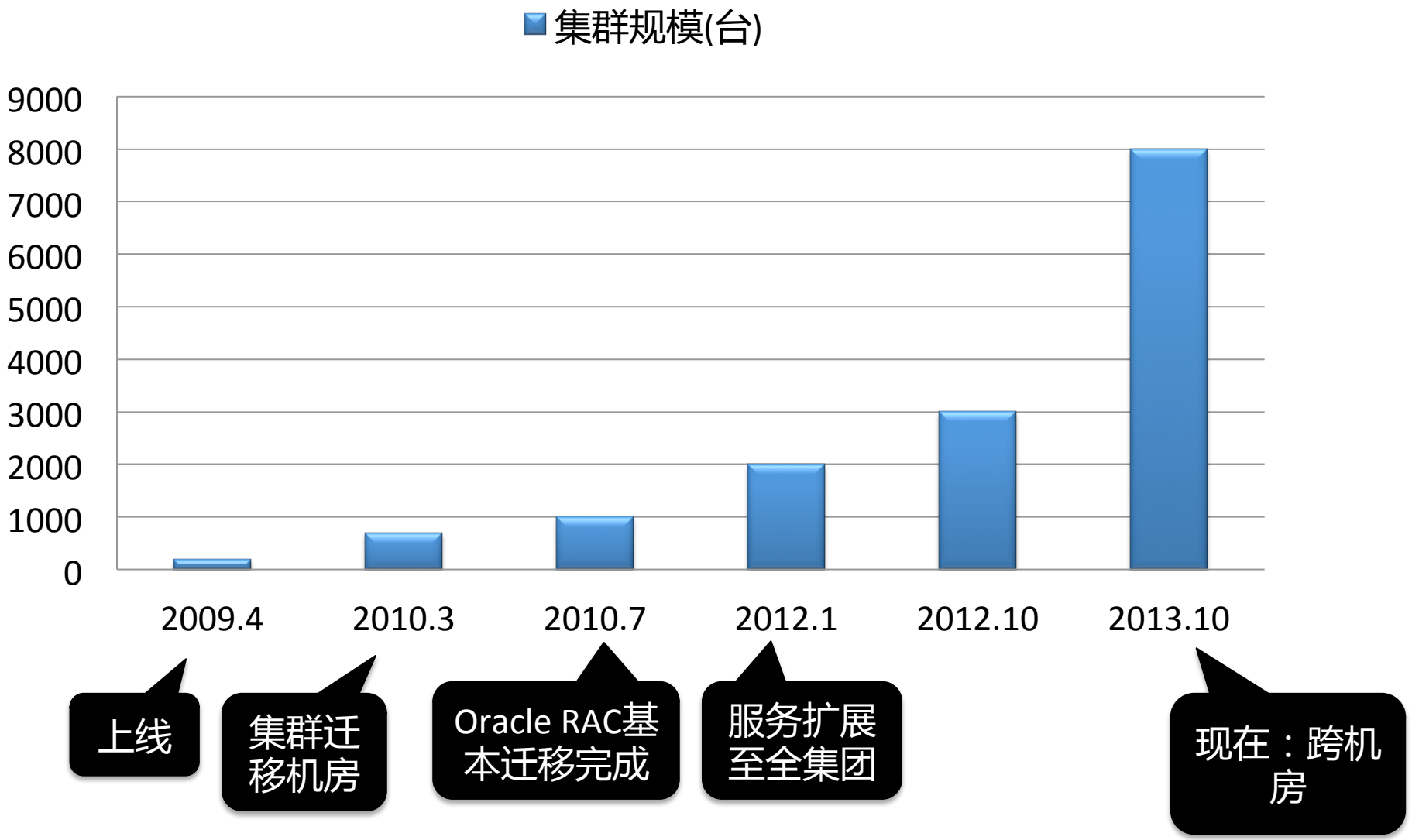
Alibaba
technology Association

阿里大数据发展历程

- 云梯
  - 一个集群
  - 一项服务
- 为阿里集团提供海量数据的存储和计算服务
- 为何选择 Hadoop？
  - MapReduce 和 HDFS 能满足大部分离线业务的需求
  - 商业公司 Yahoo / Facebook 支持，工业级应用
  - 可扩展，大规模
  - 开源软件，社区活跃

Alibaba technology Association

云梯集群发展历程



Alibaba
technology Association

演讲提纲

- 目前的云梯现状
- **YARN的介绍**
- YARN在云梯的应用
- 未来展望

计算的框架

Spark
Lightning-Fast Cluster Computing

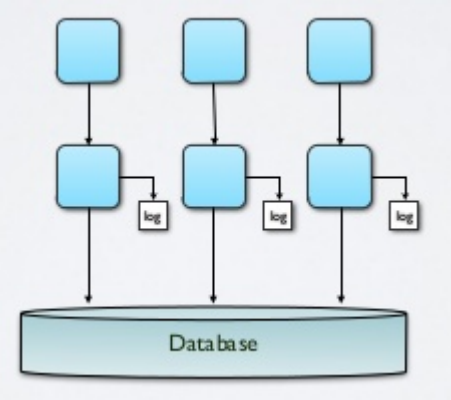
MPI

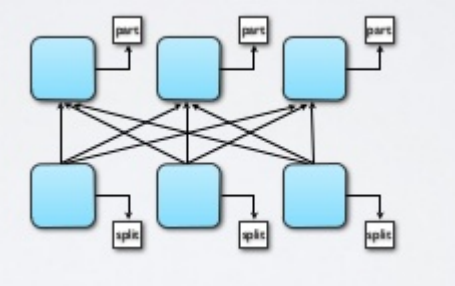
APACHE
HBASE

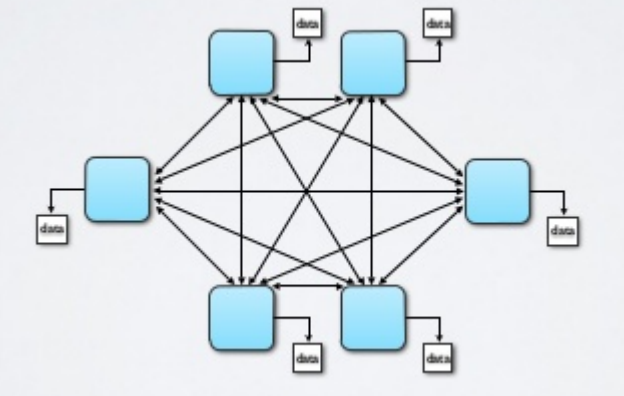
Storm
Distributed and fault-tolerant realtime computation

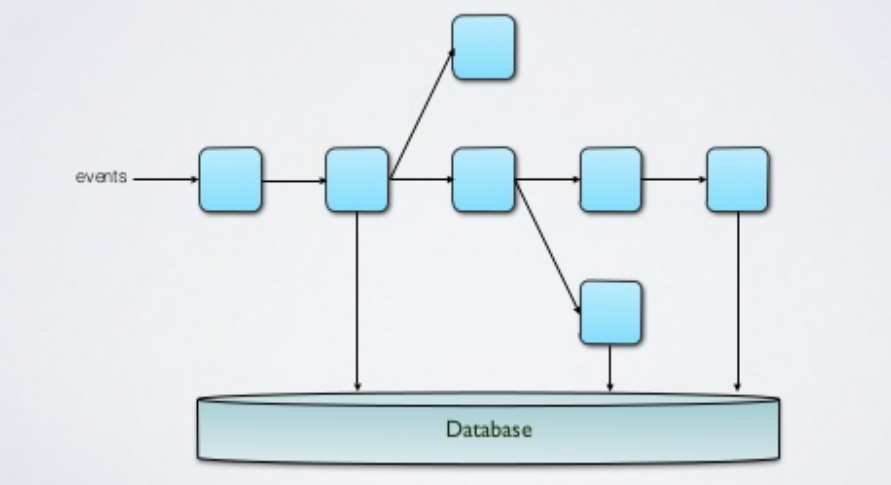
Apache Tez
A framework for near real-time big data processing

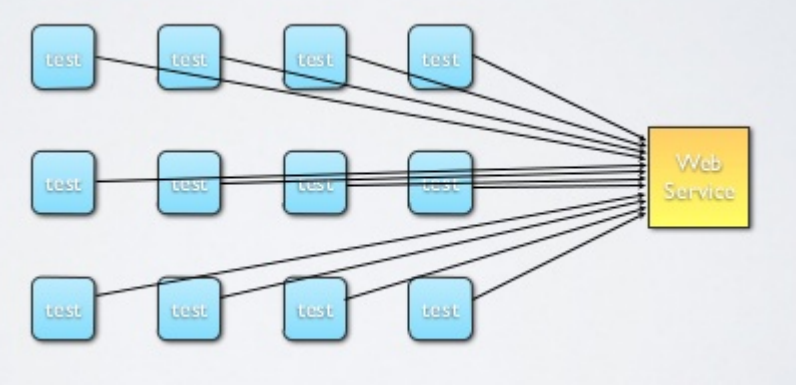
计算模型





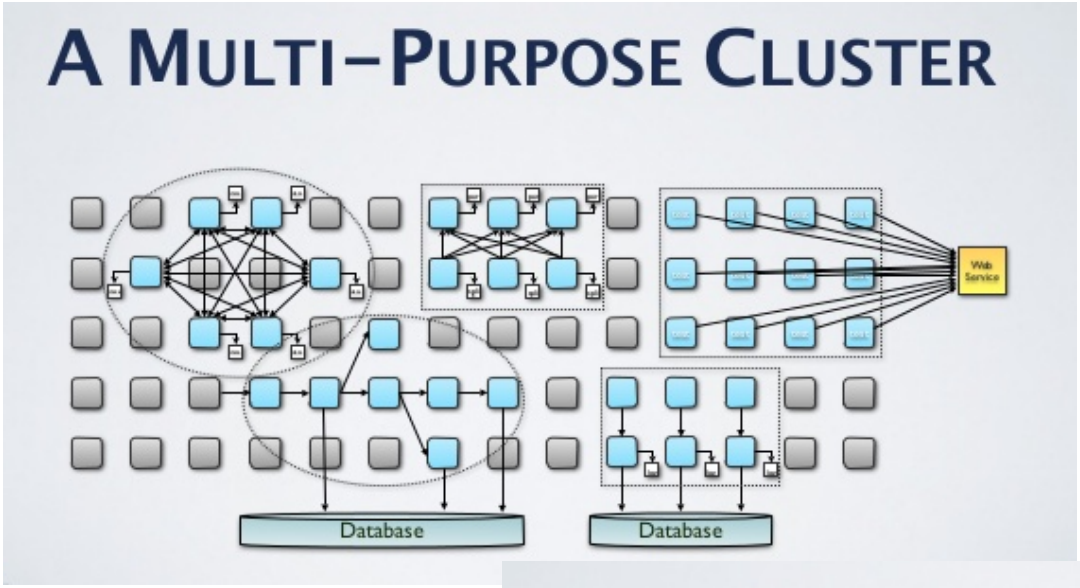


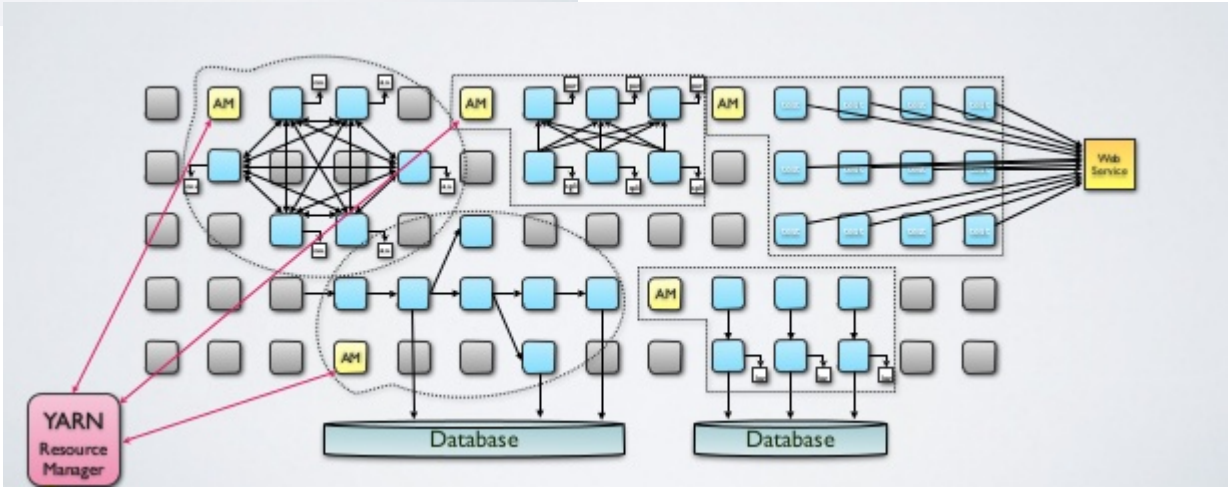




资源来自：YARN: Hadoop Beyond MapReduce - Andreas Neumann

多计算模型框架并存

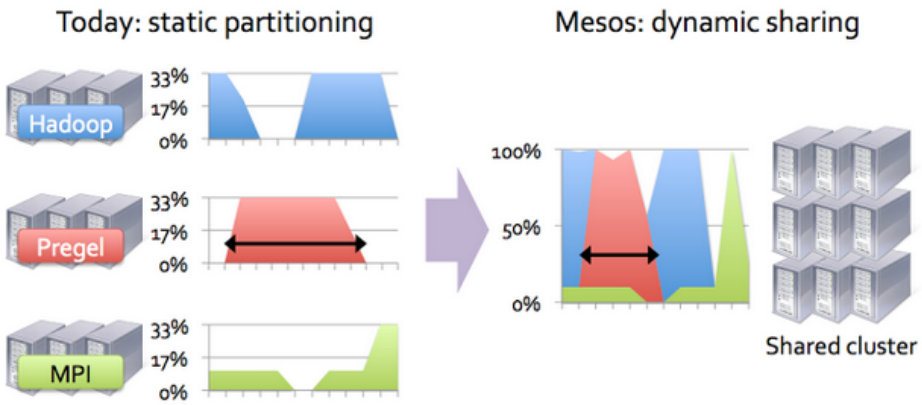




资源来自：YARN: Hadoop Beyond MapReduce - Andreas Neumann

YARN的优势

- 支持多种计算框架在同一个集群中并存，错峰节约硬件资源。



以下数据是虚拟的
假设：
Hadoop 1000台 综合平局利用率50%
Storm 500台 综合平局利用率50%
MPI 200台 综合平局利用率50%
合并后 综合平局利用率可以为80%
则：
节约637.5台机器，37.5%的成本
一年节约大约 3000W 人民币

- 在资源控制方面，突破传统的slot的概念，细化到具体的资源，如：内存、CPU。可以使利用率更高，进一步节约资源，还可以隔离影响。

Yarn的优势

- Master(RM)节点压力小

hadoop jar hadoop-tools-0.0.1.jar org.alibaba.hadoop.tools.SmallJobBench -numJobs 10000 -maxRunningJobs 1 -maps 1000 -reduces 1000 -mt 2000 -rt 2000

94.52%

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	Active Nodes
159	59	100	0	86540	100.8 TB	106.64 TB	4.9 TB	5460

User Metrics for dr.who

94.91%

Cluster Metrics

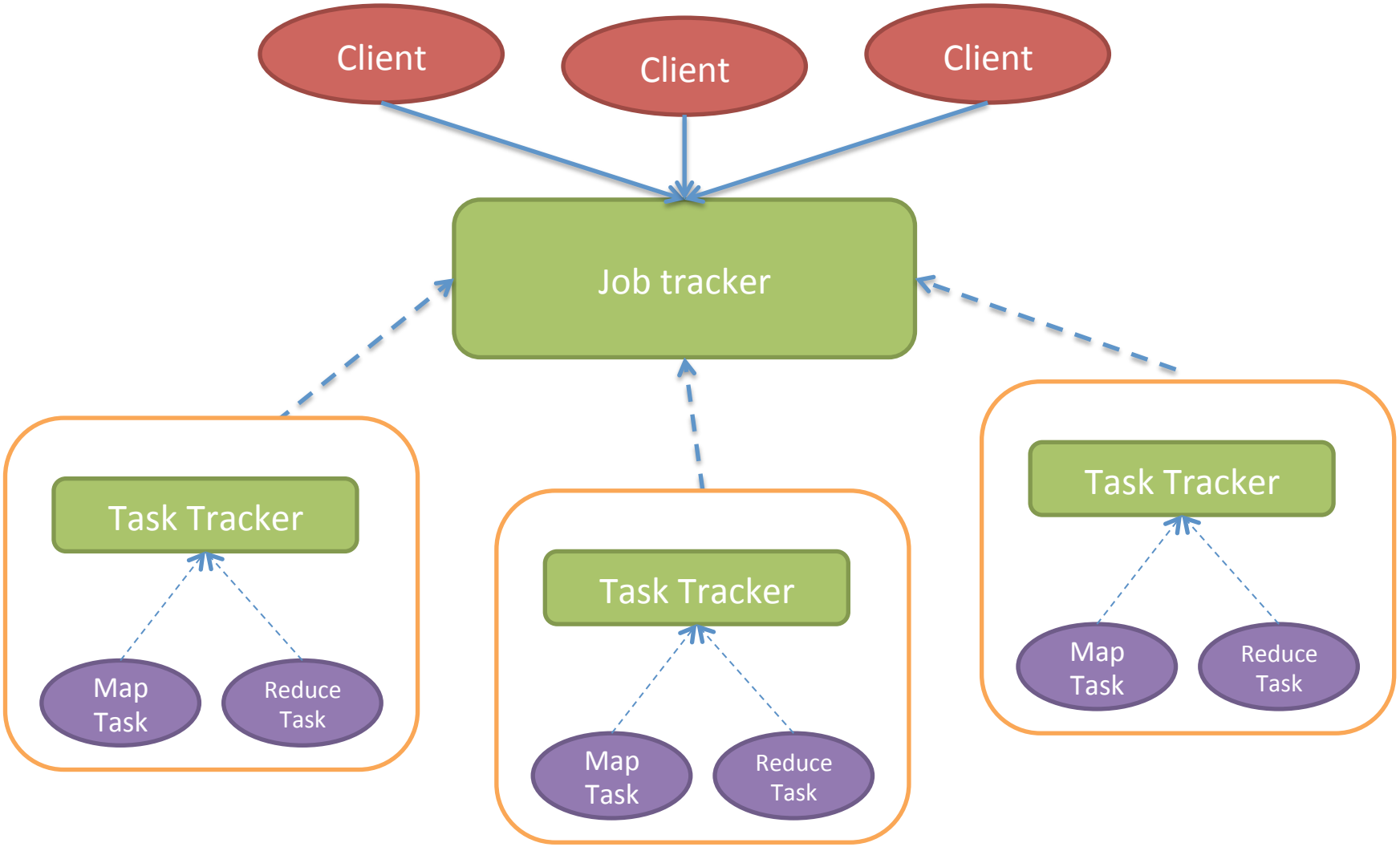
Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	Active Nodes
162	62	100	0	144501	168.24 TB	177.73 TB	8.97 TB	9100

User Metrics for dr.who

53.24%

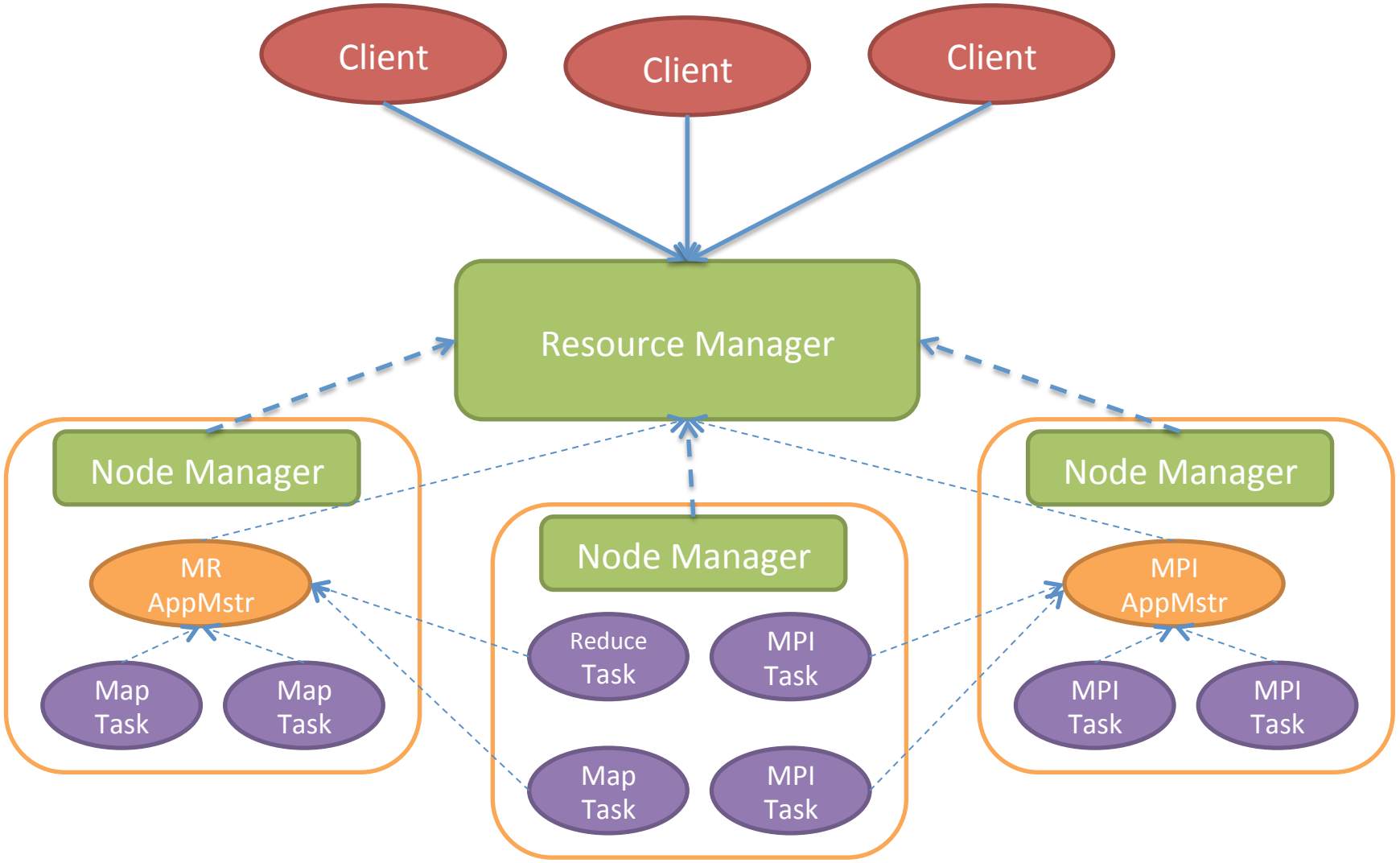
Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	Active Nodes
163	63	100	0	162909	189.66 TB	355.47 TB	0 KB	18200

以前的基本架构

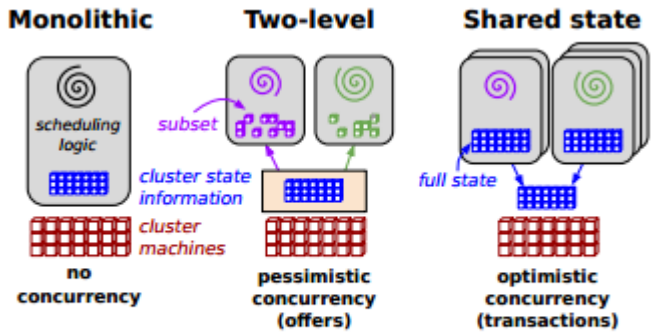

Client
Client
Client
Job tracker
Task Tracker
Map Task
Reduce Task
Task Tracker
Map Task
Reduce Task
Task Tracker
Map Task
Reduce Task

Alibaba
technology Association

YARN基本架构



YARN双层调度

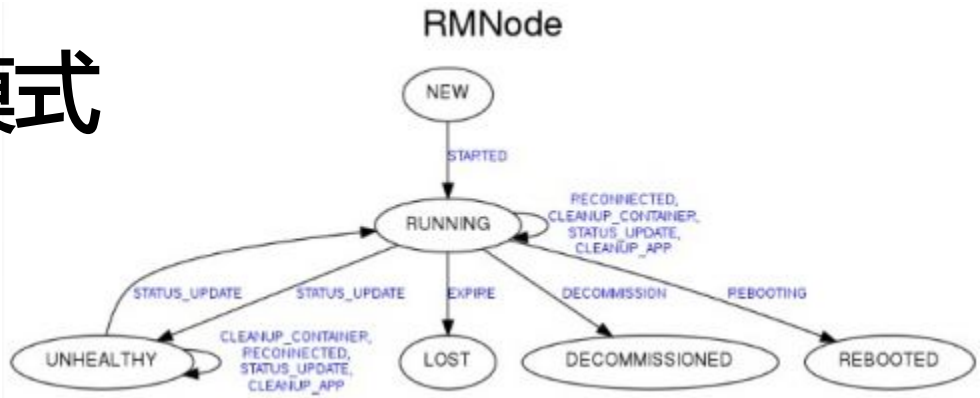


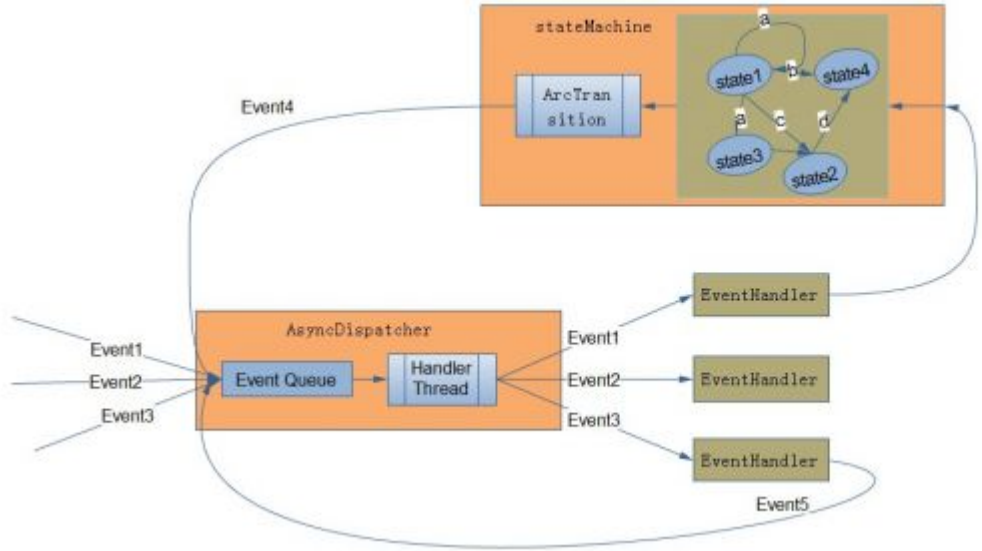
	特点	缺点	典型代表
Monolithic scheduler	结构简单，实现难度相对较低	集群规模受限、很难引入新的调度策略	JobTracker
Two-level scheduler	扩展性好1w-10w，支持多种调度模式	各应用无法感知集群整体的使用状况	YARN mesos
Shared state Scheduler	应用感知集群的整体的使用状况	复杂	Omega

Omega: flexible, scalable schedulers for large compute clusters

YARN的设计

- 服务生命周期管理模式
- 事件驱动模式
- 状态驱动模式





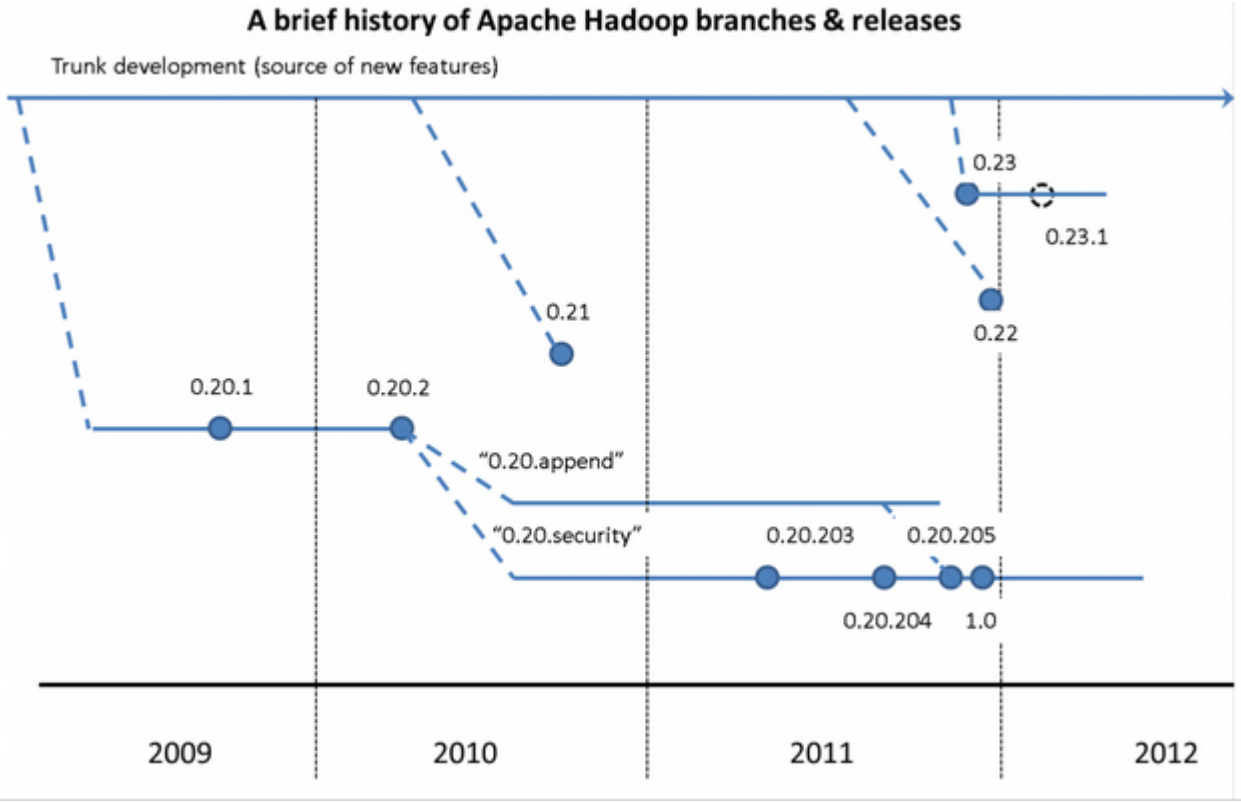
Alibaba
technology Association

演讲提纲

- 目前的云梯现状
- YARN的介绍
- **YARN在云梯的应用**
- 未来展望

云梯选择的版本

[Charles Zedlewski An update on Apache Hadoop 1.0]



云梯的选择：

09年	0.19.1
12年10月	0.23.X
13年10月	2.2.X

Alibaba
technology Association

Yarn在云梯的状态

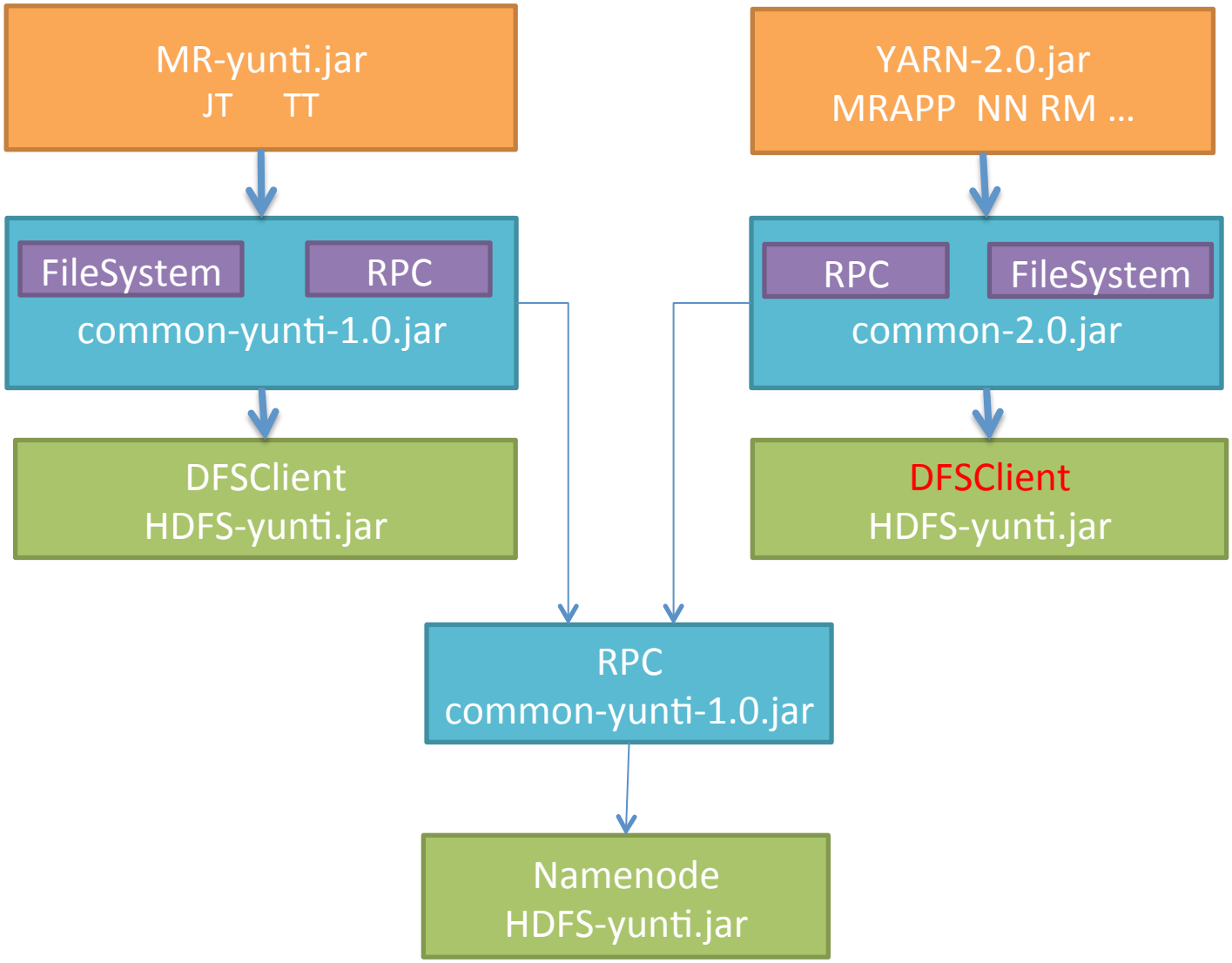
- 目前yarn在云梯还是验证阶段
- 目前150台机器的规模,双机房
- 每日JOB 几K左右
- 已经在线稳定运行4个月左右
- 计划在不久将来增加到几K左右

Alibaba
technology Association

阿里的改动

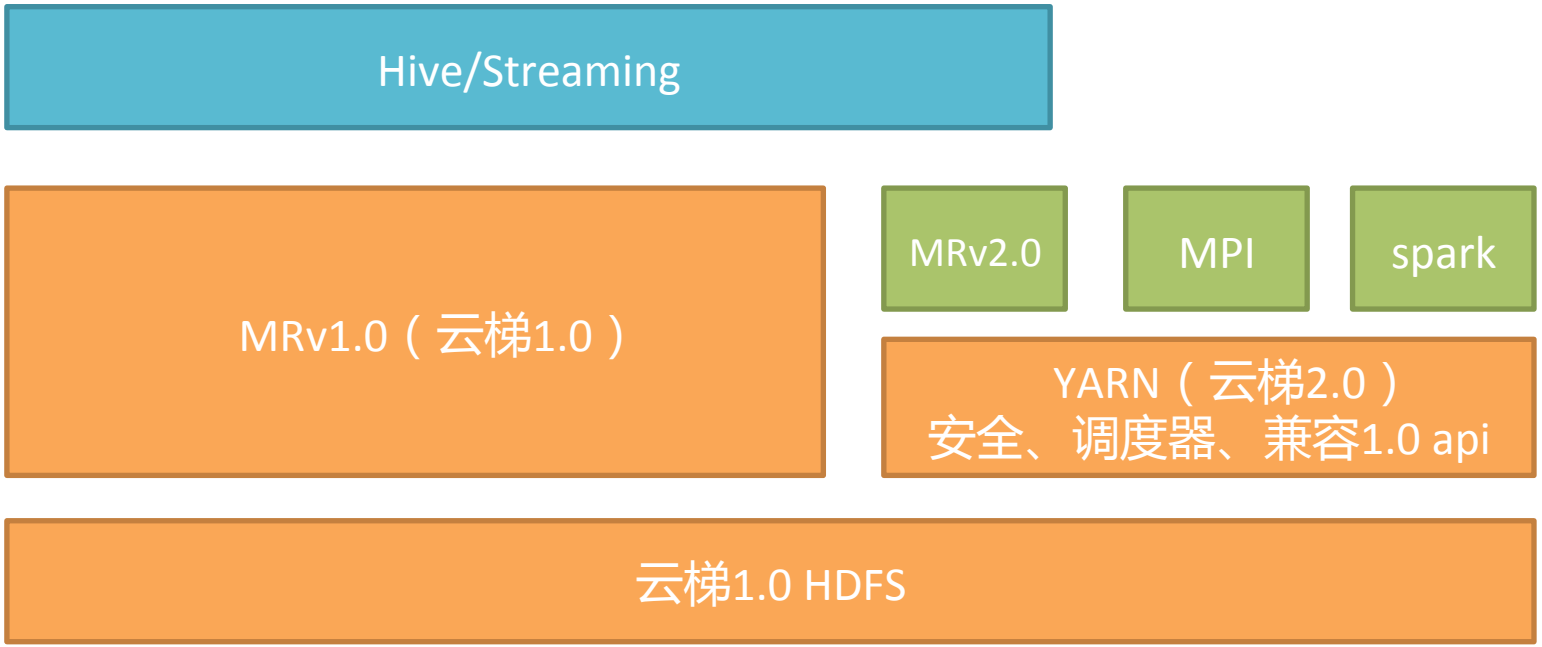
- 兼容阿里0.19.1的HDFS
- 调度器的改动
  - 引入提交App时间点的限制
  - 同组内绝对优先级
  - 跨机房调度
  - 适配安全的一些改造
- 提供一个统一的查询log界面
- 性能优化
- 再集成LZO解压缩算法

YARN 与 现有HDFS融合



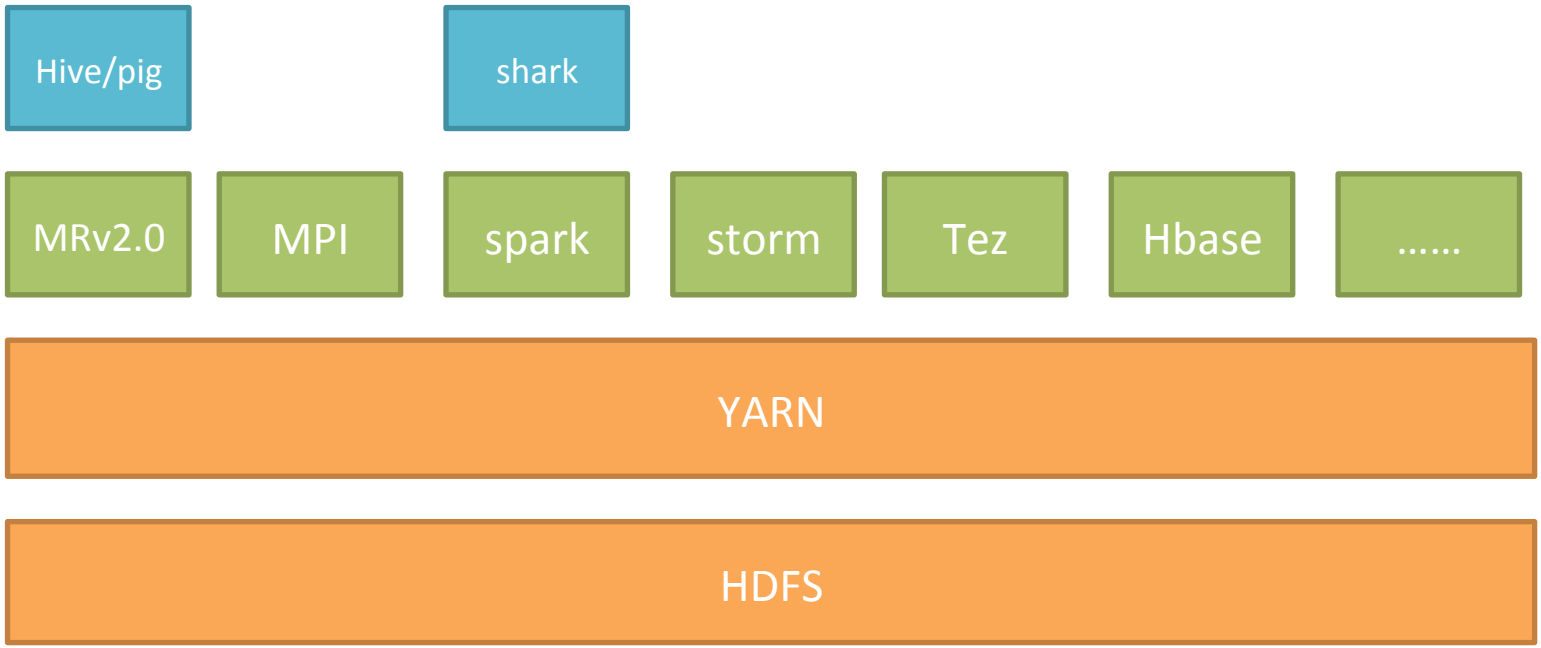
Alibaba
technology Association

目前云梯的现状

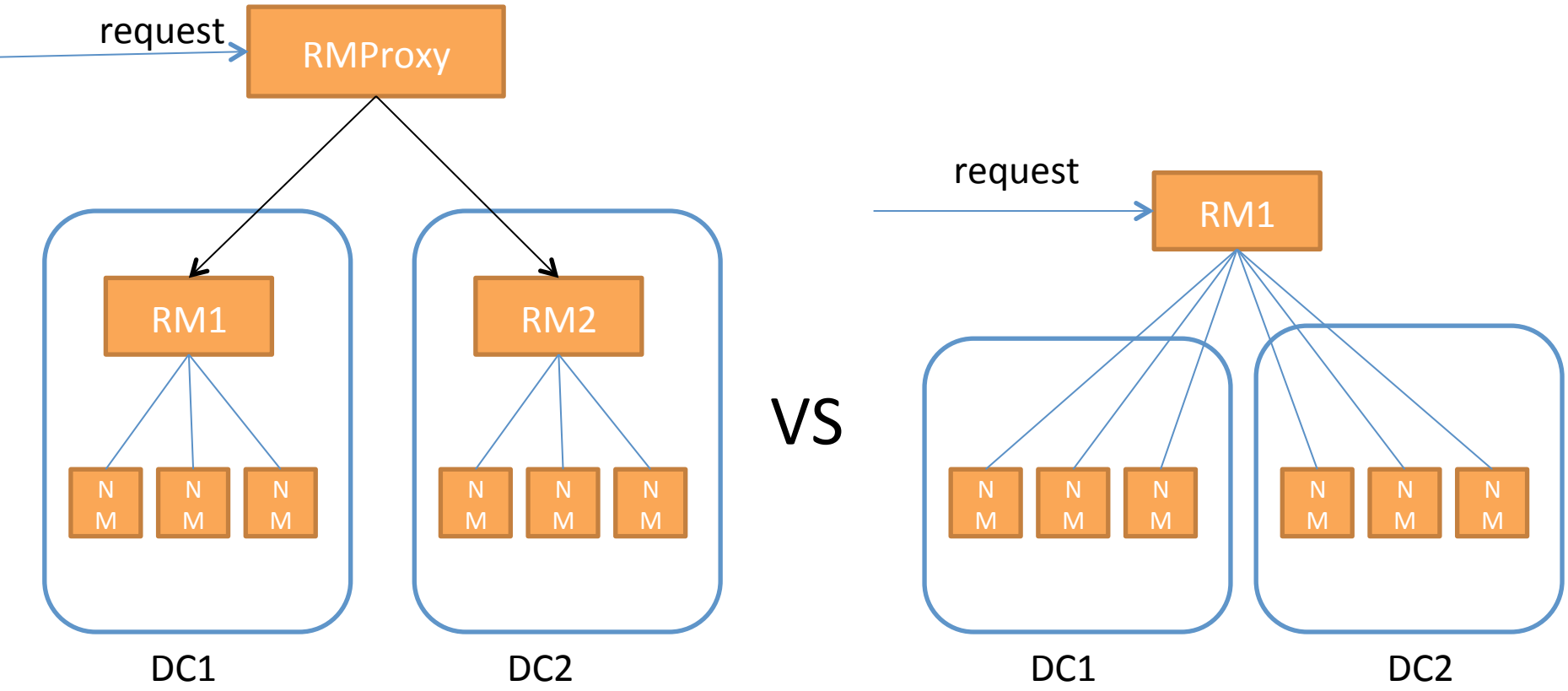


Alibaba
technology Association

不久后的云梯



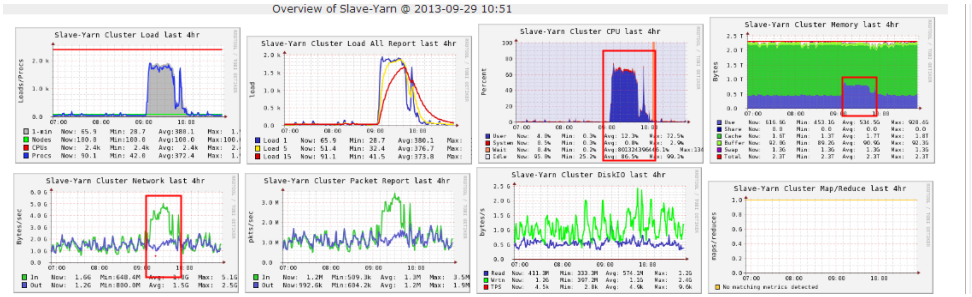
云梯YARN的跨机房

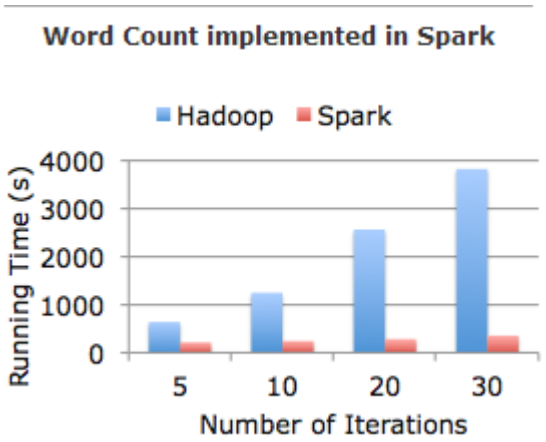

request
RMProxy
RM1
RM2
N M
N M
N M
N M
N M
N M
DC1
DC2
VS
request
RM1
N M
N M
N M
N M
N M
N M
DC1
DC2

Alibaba technology Association

Spark on yarn

目前阿里有几个团队在使用spark，
共享云梯YARN集群，目前spark每天的job大约为100+
有时候单个job对资源的利用还是很多的，
如下：

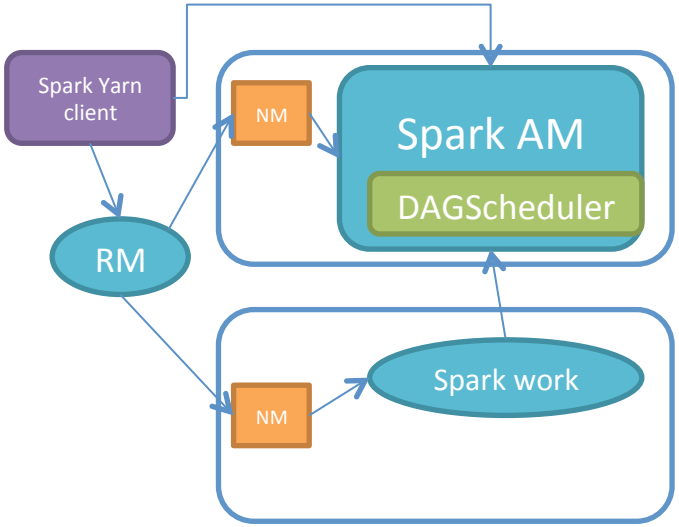




对spark应用的门槛：

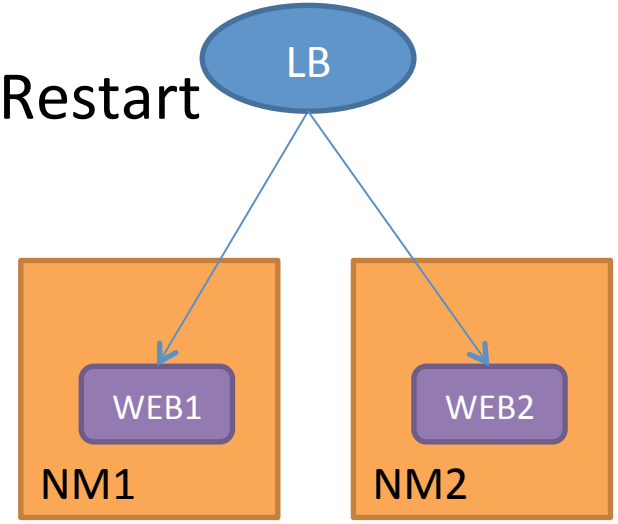
1. 目前还没有应用shark，不能直接写SQL
2. 用spark基本还需要学scala，有一定的语言门槛

跟spark的同质产品也有很多：
如：MPI、 Impala、 Strom



online on yarn

- RM的高可用性
  - YARN-128 The aim of the initial phase is to enhance the RM to be able to continue running existing applications on cluster after the RM has been restarted
  - YARN-149 RM HA
  - YARN-556 Work Preserving RM Restart
- AM的高可用性
- 升级
  - Yarn升级
  - Application代码升级
- 调度



Alibaba
technology Association

演讲提纲

- 目前的云梯现状
- YARN的介绍
- YARN在云梯的应用
- **未来展望**

Alibaba
technology Association

未来展望

- 服务类型扩展
  - 支持多种计算模型，比如Spark/MPI/Storm/Tez等，超越Hadoop MapReduce (Hadoop 2.0 Yarn)
- 期望和开源社区结合更加紧密
- 对调度器的改造
  - 支持机器打标签
  - 支持申请单台机器
- 更细粒度控制资源的利用情况
  - 磁盘IO、网络等

未来展望

- 服务质量提升
  - 节点HA
    - NN RM HA (Hadoop 2.0)
    - RM服务恢复
    - 做到不停机升级，加快软件的进化速度
  - 实时化
    - M/R调度性能的深度优化
    - 接入新的一些计算模型，如：spark、shark、storm等
  - 安全性
    - 接入阿里的认证体系，更好的授权体系。
    - Hive表的权限控制，对MR/Pig程序的等访问控制

Alibaba
technology Association

总结

- 目前的云梯现状
- YARN的介绍
- YARN在云梯的应用
- 未来展望

Alibaba
technology Association

QA!

下载来往：

找到我：

特别感谢
QCon上海合作伙伴