Hadoop 2.0基本架构与发展趋势

演讲人：董西成
技术博客：dongxicheng.org
微博ID：西成懂（私信开放）

提纲

- 什么是Hadoop 2.0？
- 什么是YARN？
- YARN的现状？
- YARN发展趋势？
- MapReduce与YARN的关系？

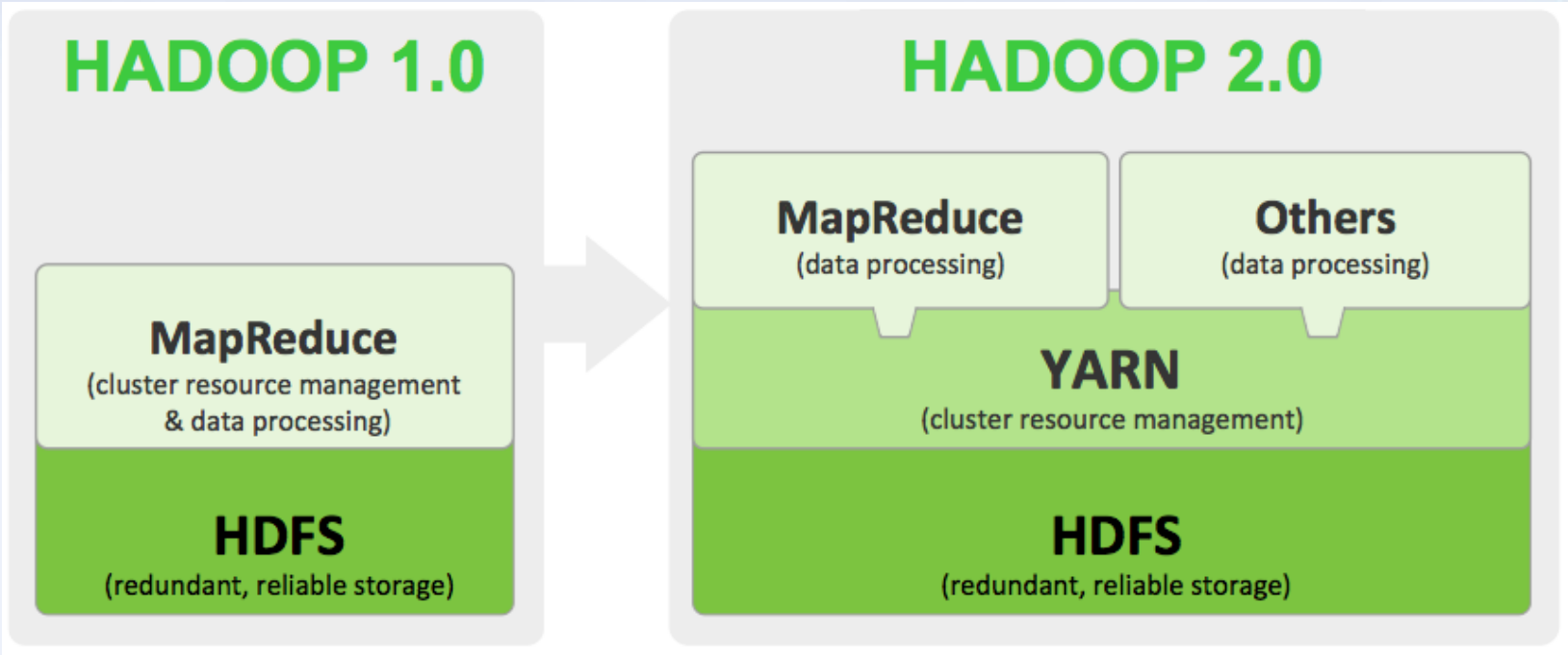
提纲

- YARN产生背景
- YARN基本架构
- 多角度理解YARN
- YARN发展趋势
- 总结

提纲

- YARN产生背景
- YARN基本架构
- 多角度理解YARN
- YARN发展趋势
- 总结

Hadoop 2.0



- ✓ 由HDFS、MapReduce和YARN三个分支构成；
- ✓ HDFS：NN Federation、HA；
- ✓ MapReduce：运行在YARN上的MR；
- ✓ YARN：资源管理系统

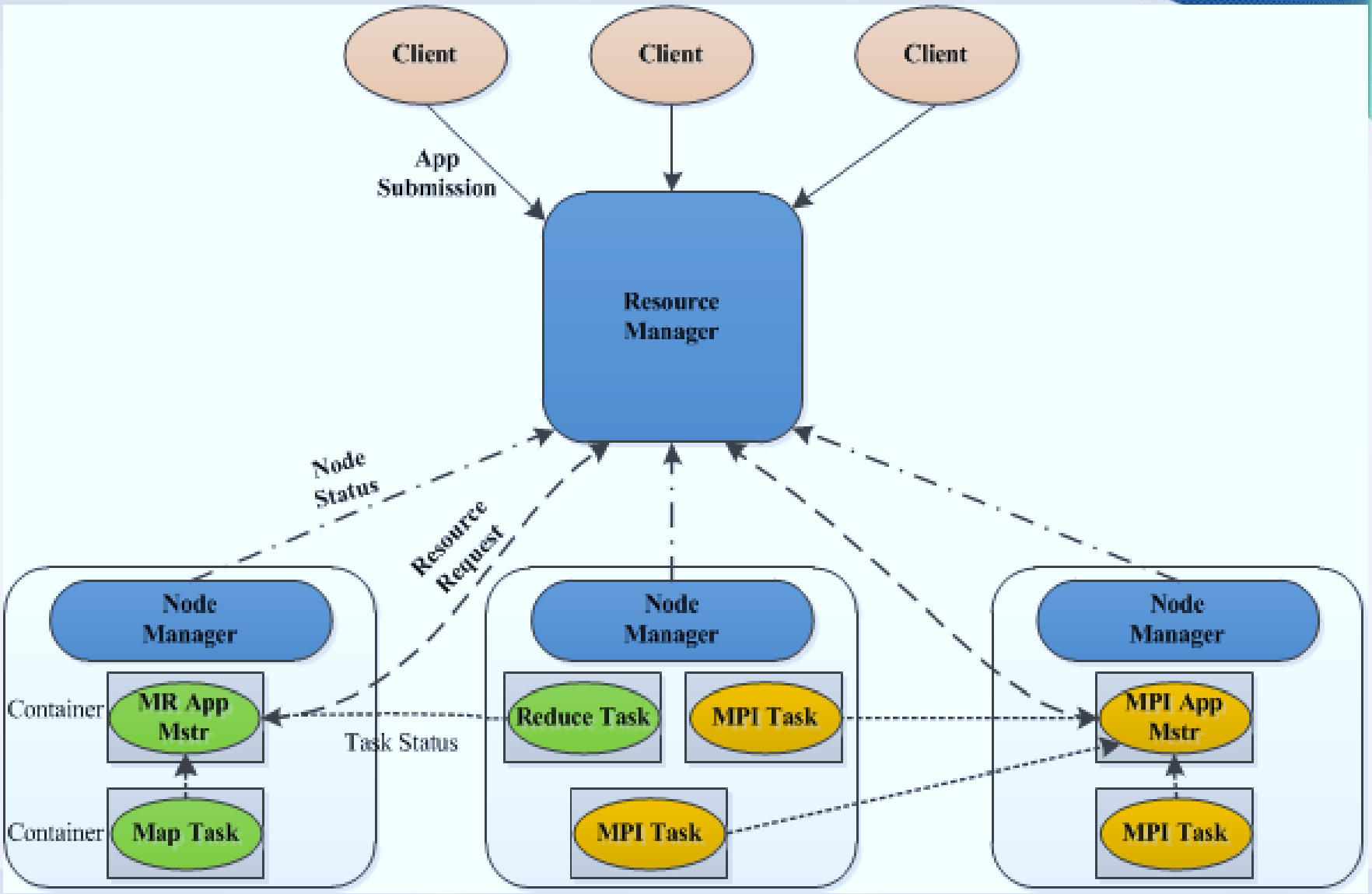
YARN产生背景

- 直接源于MRv1在几个方面的无能
  - ✓扩展性受限
  - ✓单点故障
  - ✓难以支持MR之外的计算
- 多计算框架各自为战，数据共享困难
  - ✓MR：离线计算框架
  - ✓Storm：实时计算框架
  - ✓Spark：内存计算框架

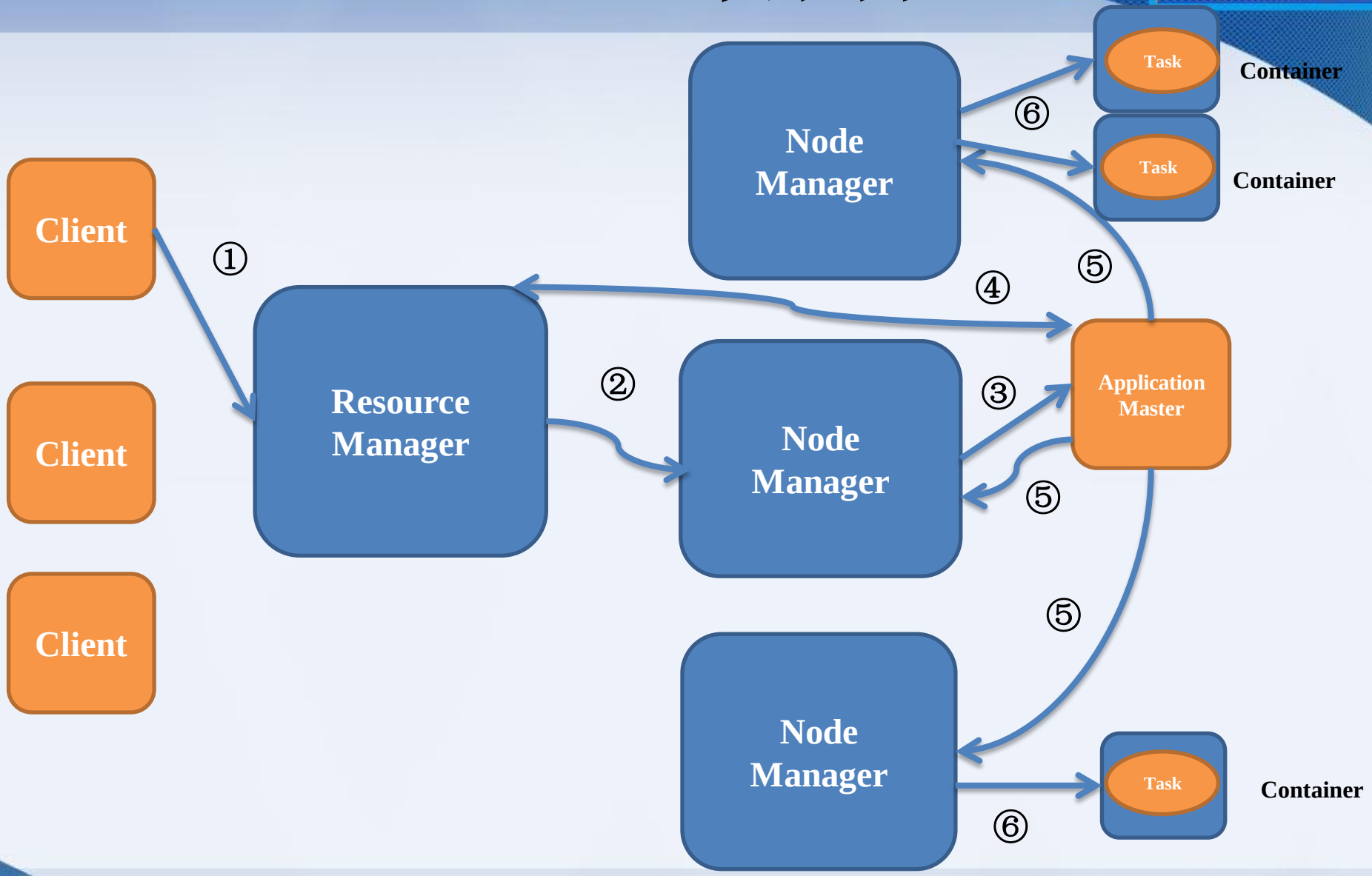
提纲

- YARN产生背景
- YARN基本架构
- 多角度理解YARN
- YARN发展趋势
- 总结

YARN基本架构



YARN基本架构



YARN基本架构

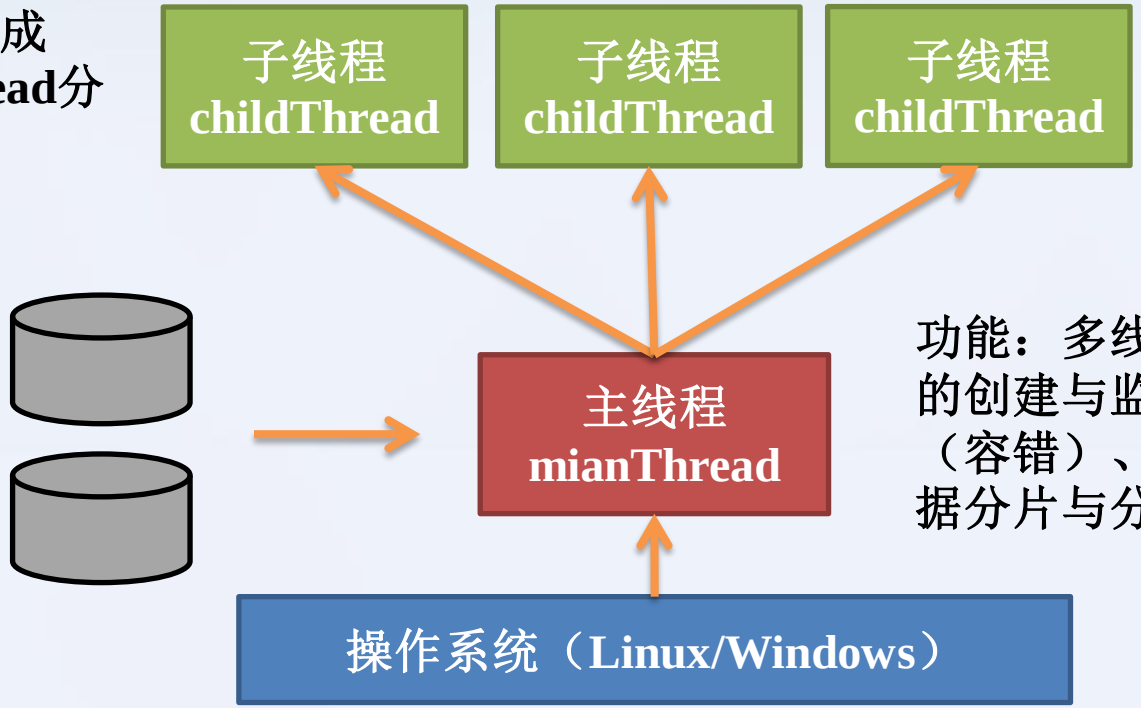
- ResourceManager
  - ✓处理客户端请求
  - ✓启动/监控ApplicationMaster
  - ✓监控NodeManager
  - ✓资源分配与调度
- NodeManager
  - ✓单个节点上的资源管理
  - ✓处理来自ResourceManager的命令
  - ✓处理来自ApplicationMaster的命令
- ApplicationMaster（以MRAppMaster为例）
  - ✓数据切分
  - ✓为应用程序申请资源，并分配给内部任务
  - ✓任务监控与容错

提纲

- YARN产生背景
- YARN基本架构
- 多角度理解YARN
- YARN发展趋势
- 总结

单机并行计算角度

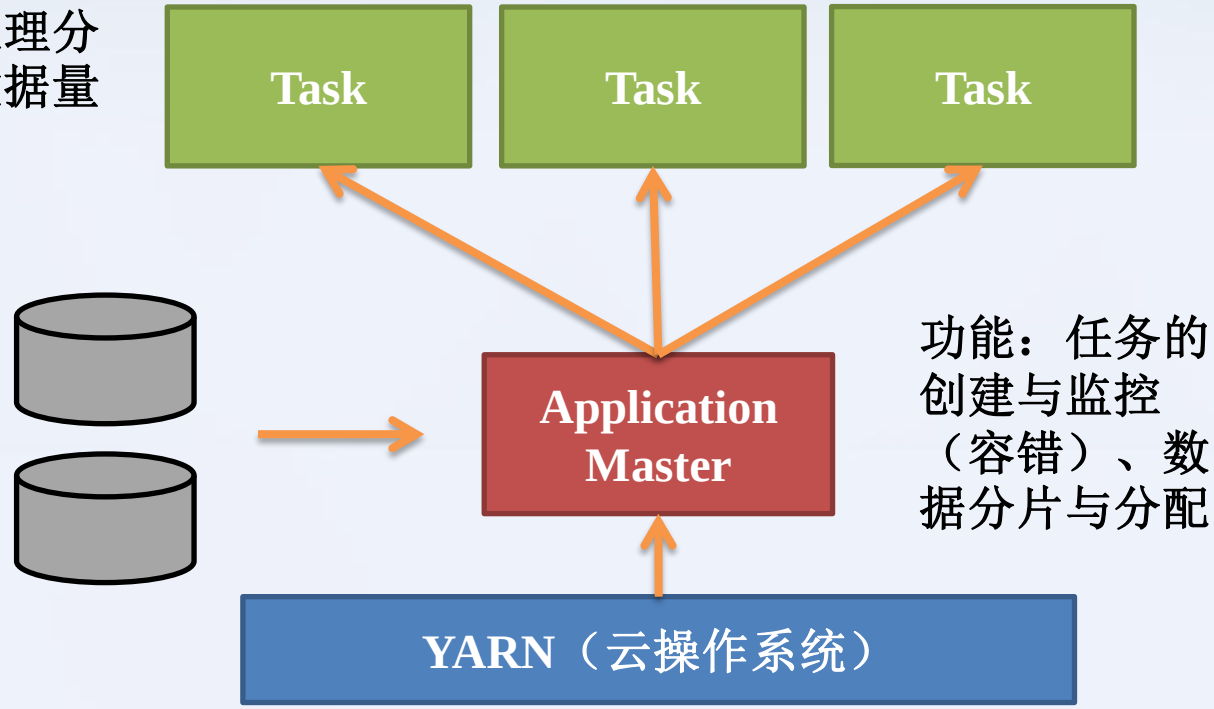
功能：完成mainThread分配的任务



功能：多线程的创建与监控（容错）、数据分片与分配

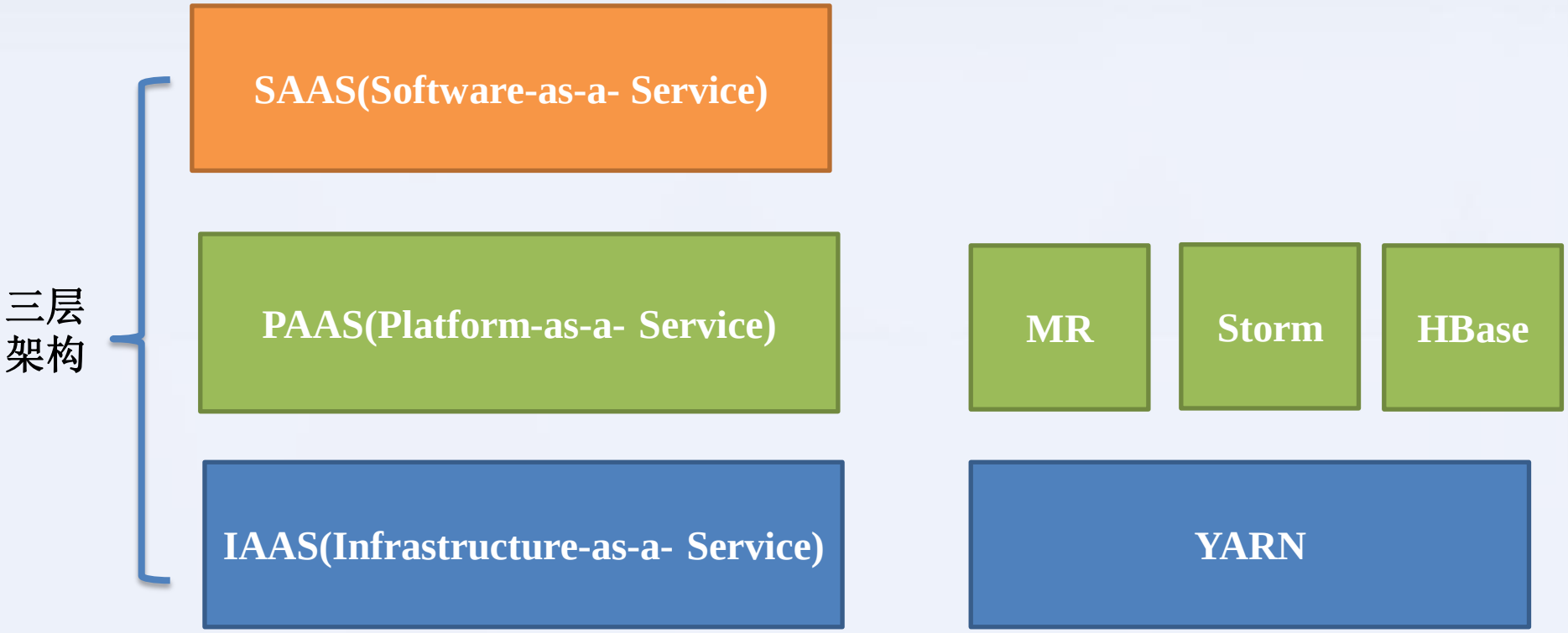
单机并行计算角度

功能：处理分配的任数据量

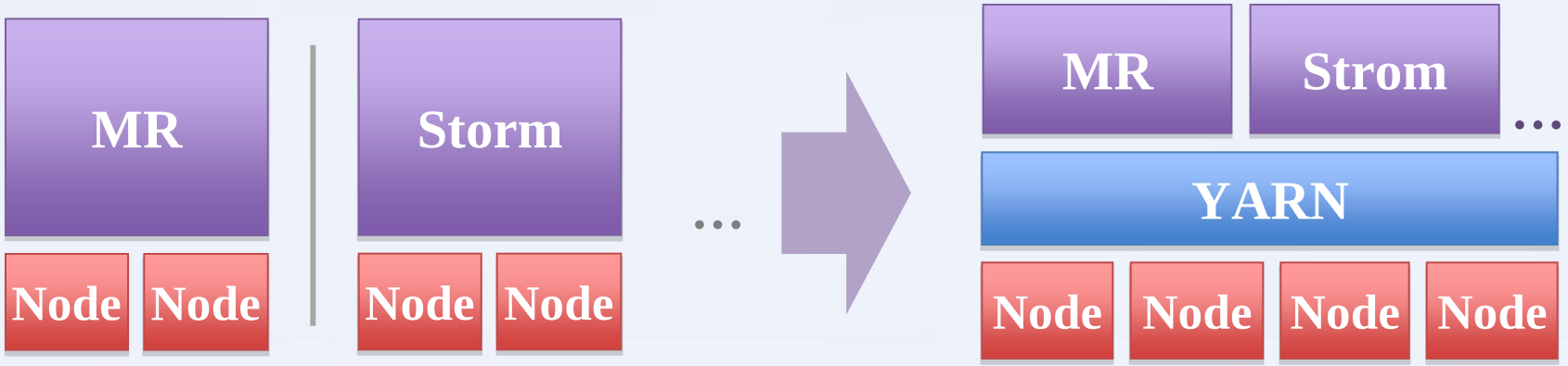


功能：任务的创建与监控（容错）、数据分片与分配

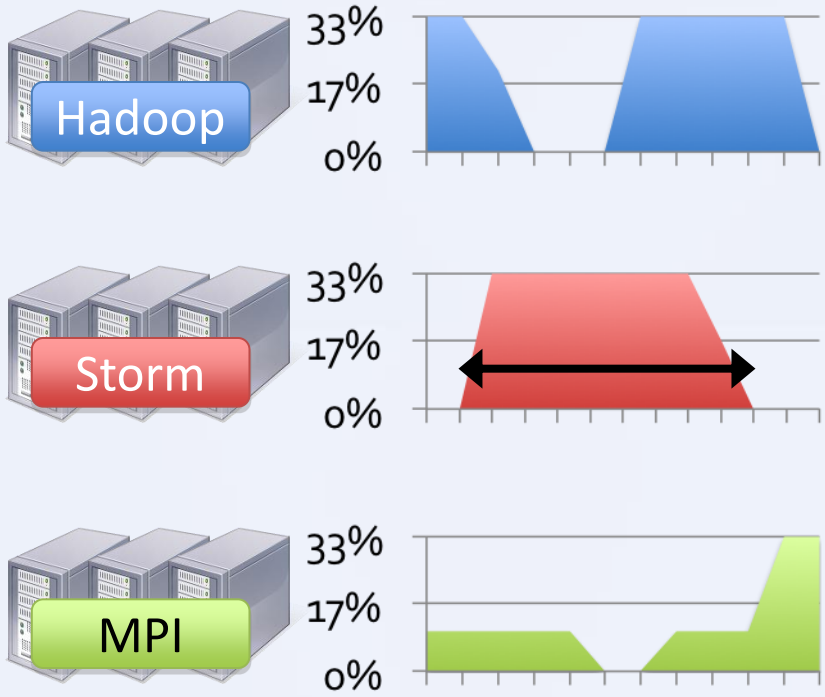
云计算角度



集群管理系统角度



集群管理系统角度



集群管理系统角度（好处）

- ✓资源利用率高
- ✓运维成本降低
- ✓数据共享

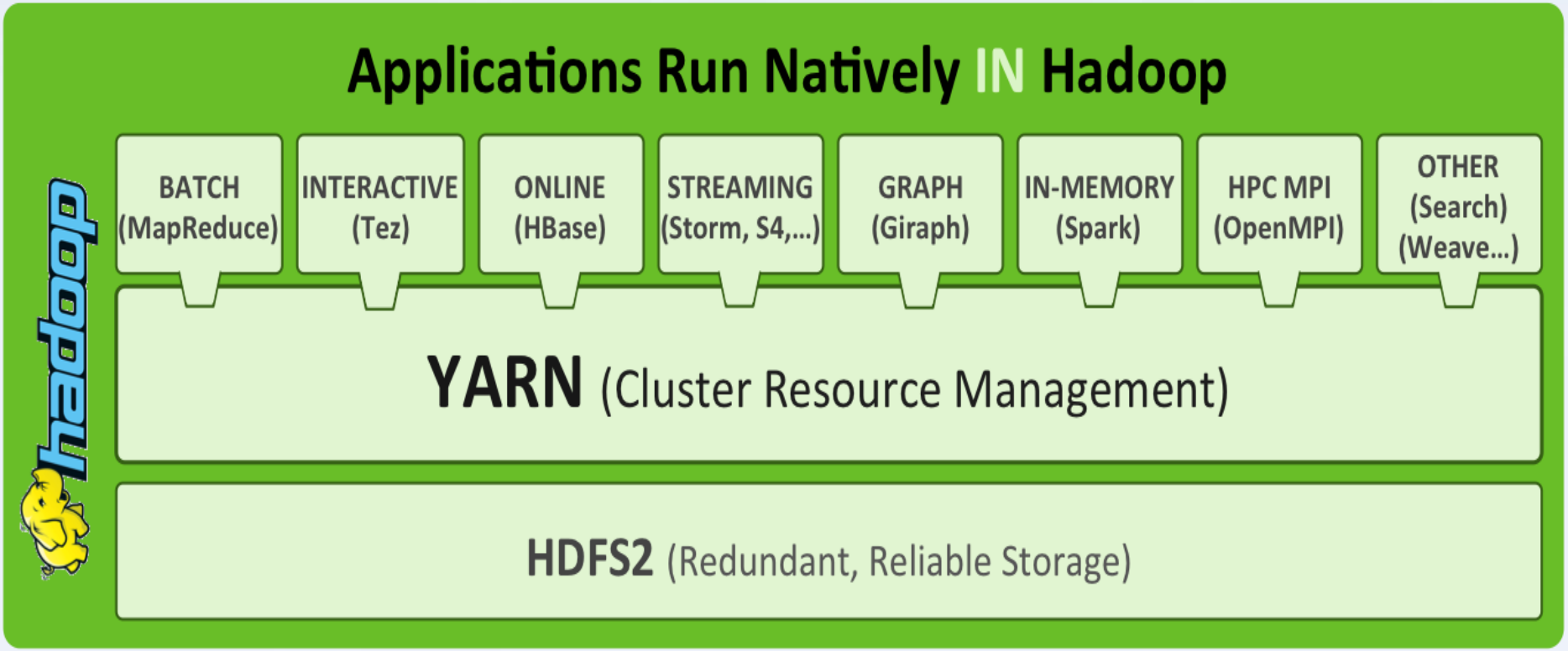
提纲

- YARN产生背景
- YARN基本架构
- 多角度理解YARN
- YARN发展趋势
- 总结

YARN发展现状与趋势

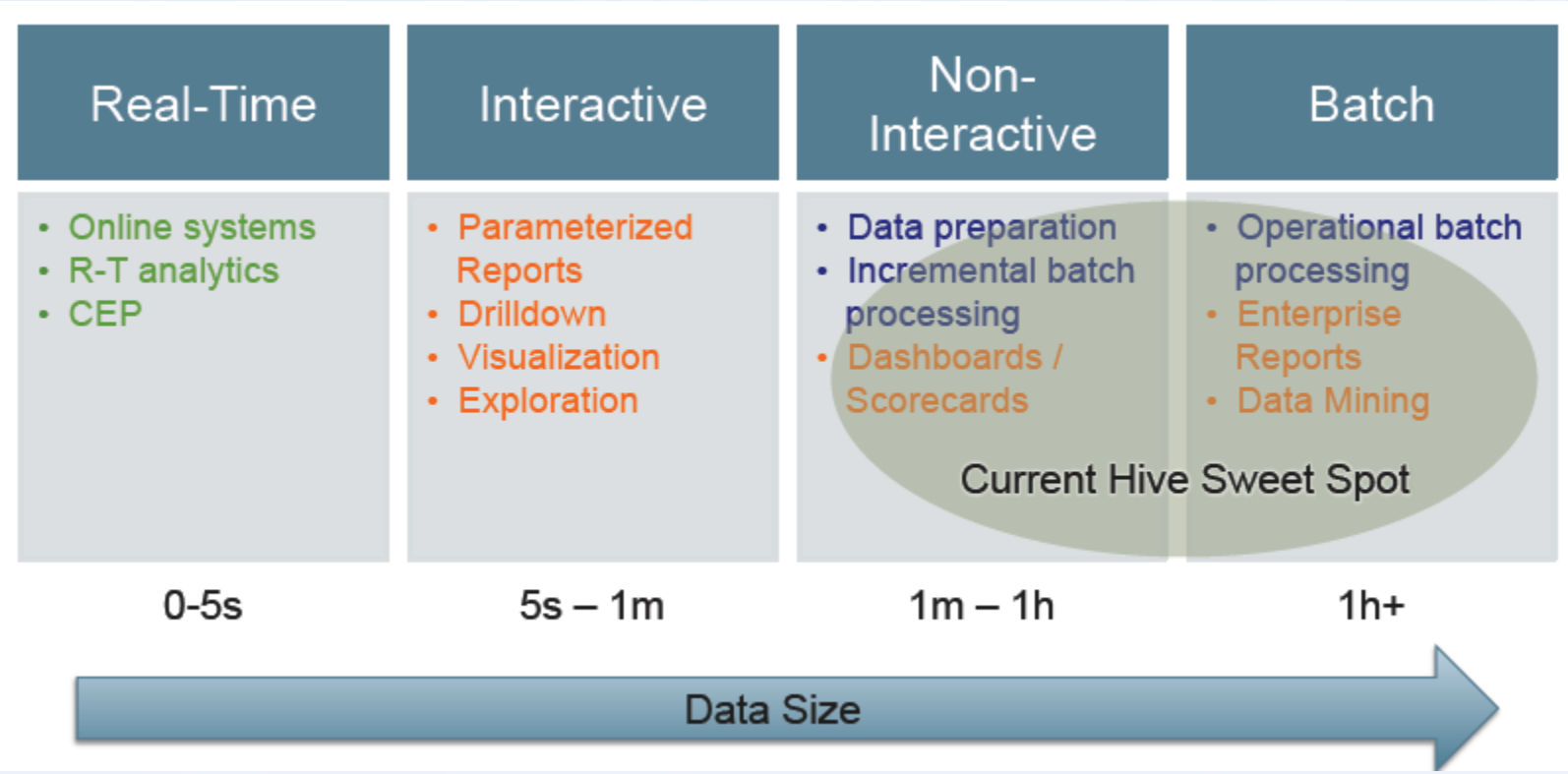
- ✓目前位于alpha版，下个月发布2.1.0-beta版
- ✓多种系统正在往YARN上转移
- ✓调度模型和资源隔离等方面存在不足，多种系统不易运行在YARN上

运行在YARN上的软件



系统链接阅读我的博客文章：“汇总运行在Hadoop YARN上的开源系统”
http://dongxicheng.org/mapreduce-nextgen/run-systems-on-hadoop-yarn/

应用场景分类



【注】摘自Hortonworks PPT：“Stinger Initiative: Deep Dive”

运行在YARN上带来的好处

- ✓一个集群部署多个版本
- ✓计算资源按需伸缩
- ✓不同负载应用混搭，集群利用率高
- ✓共享底层存储，避免数据跨集群迁移

调度模型

- ✓支持CPU和内存两种资源调度；
- ✓Resource-centric Scheduling（NOT Task-centric Scheduling）

调度场景	是否支持	应用举例
任意K个Container（位置没有要求）	√	MR、Tez等
K个独占K个节点的container	√	HBase、MPI等
K个来自同一机架的container	×	Storm等
Container可绑定到固定CPU上	×	MPI等

资源隔离

- ✓支持CPU和内存两种资源隔离
- ✓采用Cgroups对CPU隔离
- ✓采用线程监控内存方案（借鉴MRv1）

MapReduce与YARN

- ✓目前MapReduce有两个版本：独立版和YARN版；
- ✓独立版运行时环境由JobTracker、TaskTracker、MapTask、ReduceTask等组成；
- ✓YARN版MapReduce只能运行在YARN上，不能独立部署运行。

MapReduce与YARN

- ✓MRv1由编程接口，调度环境（JobTracker和TaskTracker）和任务处理引擎（MapTask和ReduceTask）三部分组成；
- ✓YARN重用了MRv1的调度器，黑白名单机制、部分资源隔离机制；
- ✓YARN版MapReduce（代码级）重用了MRv1的编程接口和任务处理引擎；
- ✓旧API编写的MR作业可直接运行在YARN版MapReduce上，但新API则不可以。

如何快速学习MapReduce？

总结

- ✓ 从一定程度上可认为，YARN是“互联网界的OpenStack”
- ✓ YARN时代将要来临

提问

Questions?