·专栏· 文章编号:1000-3428(2015)01-0001-05 文献标识码:A 中图分类号:TP391

支持异构集群并行的高能物理数据处理系统

霍 菁[1,2],雷晓凤[1,2],李 强[1,2],孙功星[1]

(1. 中国科学院高能物理研究所,北京 100049;2. 中国科学院大学,北京 100049)

摘 要:传统集群计算系统无法充分利用本地磁盘的存储资源和I/O,大量网络I/O成为系统瓶颈,导致资源利用率降低,并造成高昂的存储和网络成本。使用Hadoop处理分析作业可有效利用本地磁盘存储和I/O资源,而集群资源统一管理工具Mesos则使用轻量化的设计和高效的通信机制,能在不同计算集群之间动态共享集群资源。为此,分析高能物理数据处理的特点,利用Mesos构建异构集群间资源共享的高能物理实验数据处理系统,实现Torque/Maui和Hadoop集群的集成。测试结果表明,该系统能够在集群间动态分配集群资源,并利用本地存储和磁盘I/O显著降低网络I/O,提高集群资源利用率。

关键词:高能物理;集群资源管理;资源共享;Mesos工具;Hadoop平台;Torque/Maui系统

中文引用格式:霍 菁,雷晓凤,李 强,等. 支持异构集群并行的高能物理数据处理系统[J]. 计算机工程,2015,41(1):1-5.

英文引用格式:Huo Jing,Lei Xiaofeng,Li Qiang,et al. High Energy Physics Data Processing System with Heterogeneous Clusters[J]. Computer Engineering,2015,41(1):1-5.

High Energy Physics Data Processing System with Parallel Heterogeneous Clusters

HUO Jing[1,2],LEI Xiaofeng[1,2],LI Qiang[1,2],SUN Gongxing[1]

(1. Institute of High Energy Physics,Chinese Academy of Sciences,Beijing 100049,China;
2. Graduate University of Chinese Academy of Sciences,Beijing 100049,China)

【Abstract】 The traditional cluster computing system can not make best of the local disks and disk I/O resources, therefore the network becomes the bottleneck of the whole system. And this is the reason of low utilization of the cluster resources and high cost on data storage and network equipment. Using Hadoop to process analysis can significantly reduce the pressure on network I/O by using the local disks as a distributed file system. Mesos is a cluster resource manager with light-weight design and efficient communication mechanisms that can dynamically share resources among clusters. This paper introduces the features of High Energy Physics(HEP),data processing,presents a new HEP data processing system by using Mesos to provide dynamic resource sharing among clusters,and implements integration of Toruqe/Maui and Hadoop which can avoid the disadvantages. The test result shows that the new system can dynamic distribute the cluster resource,and reduce the network I/O,improve the resource utilization.

【Key words】 High Energy Physics(HEP);cluster resource management;resource sharing;Mesos tool;Hadoop platform;Toruqe/Maui system

DOI:10.3969/j.issn.1000-3428.2015.01.001

1 概述

高能物理实验是一项庞大的系统工程,数据处理是其中的关键步骤。高能物理实验产生的数据量非常庞大,欧洲核子中心(CERN)使用的大型强子对撞机LHC(Large Hadron Collider),每年产生约25 PB的实验数据;北京正负电子对撞机(BEPCII)和北京谱仪(BESIII)在经过改造以后,2012年产生的数据量超过了过去几年的总和,总规模超过3 PB。

高能物理实验数据处理有多种不同的类型作业。传统集群使用的批作业处理系统Torque[1]/Maui[2]采用计算资源和数据存储相分离的系统架构,适用于蒙特

基金项目:国家自然科学基金资助项目(11375223,11375221);国家自然科学基金A3前瞻计划基金资助项目(61161140454)。
作者简介:霍 菁(1985-),男,博士研究生,主研方向:分布式计算,集群资源管理;雷晓凤、李 强,博士研究生;孙功星,研究员。
收稿日期:2014-02-17 **修回日期**:2014-03-20 **E-mail**:huojing@ihep.ac.cn

卡洛模拟等计算密集型应用。而 Hadoop[3] 平台使用 MapReduce[4] 的方式,能够利用计算节点本地磁盘组建 HDFS[5],使分析作业读取本地磁盘数据,减轻网络 I/O 压力,更适合数据密集型的分析作业[6]。由于高能物理数据处理具有随时间分布的特征,2 个集群的作业分布密度不同,在集群间动态共享资源能够有效提高集群资源利用率。集群资源统一管理工具 Mesos[7],使用轻量化的设计和高效的通信机制,可以在不同的集群之间动态共享资源,比虚拟机技术更加高效。目前 Mesos 已支持多种应用,例如 Hadoop, Spark, Hypertable 等。已有多家著名公司使用 Mesos 管理集群资源,包括国外著名公司 Twitter 和国内著名视频网站爱奇艺等,但在高能物理实验中还没有应用。

本文分析高能物理实验数据处理的不同作业类型和特点,以及 Torque/Maui 集群和 Hadoop 集群的特性,以提高集群资源利用率和利用本地存储空间为目标,利用集群资源管理工具 Mesos 集成 Torque/Maui 和 Hadoop 框架,实现一个支持混合集群并行的高能物理实验数据处理平台,并应用 BESIII 高能物理实验数据进行初步评估。

2 研究背景

2.1 BESIII 实验数据处理的流程和作业类型

BESIII 数据处理的作业类型主要有模拟作业、重建作业和分析作业 3 种。

实验数据由对撞机产生,生成原始数据(RAW Data)后存储在 Lustre[8] 中。为了验证数据的正确性,需要一定量的模拟数据来进行对比,因此,使用蒙特卡洛模拟产生模拟数据。原始数据和模拟数据在经过重建(Reconstruction)后,生成可以供物理学家进行分析的 DST 数据。对 DST 数据的分析是典型的一次写多次读的应用。不同的物理学家使用自己编写的分析程序分析 DST 数据,挑选自己感兴趣的物理事例,最后生成图表等结果。本文设计的混合集群的目的是使用传统集群计算系统处理主要消耗 CPU 资源的模拟和重建作业,使用 Hadoop 处理主要消耗 I/O 资源的分析作业,如图 1 所示。

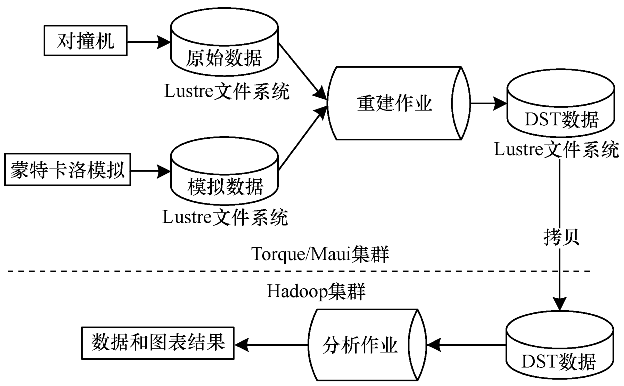

图 1 BESIII 实验数据处理的作业类型和步骤

2.2 Mesos 集群资源统一管理框架

Mesos 同大多数分布式系统一样,采用的是 Master-Slave 的架构,如图 2 所示。该架构主要由 4 个部分构成:Master, Slave, Framework 和 Executor。Master 负责资源管理和分配, Slave 负责汇报资源和管理本地资源,并负责启动作业执行器 Executor。在 Mesos 中,所有接入系统的分布式计算框架都成为 Framework,这些框架在 Slave 节点上用来运行作业的执行器称为 Executor。

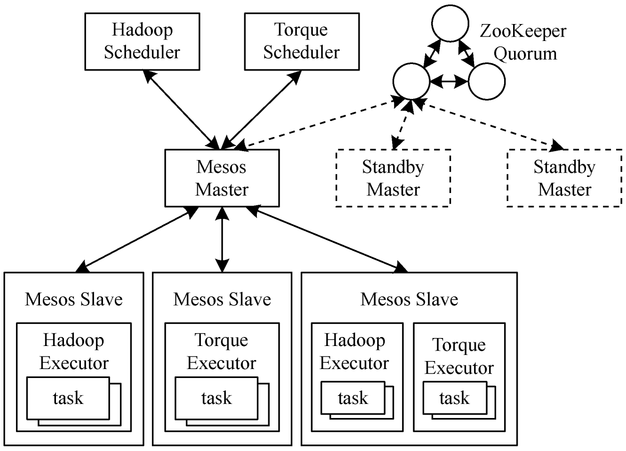

图 2 Mesos 架构

Mesos 的 Master 在设计上实现了轻量化,仅保存 Framework 和 Slave 的部分状态信息,这些信息通过注册的方式由 Framework 和 Slave 汇报给 Master,因此 Mesos 使用 ZooKeeper[9] 解决 Master 的单点失效问题,提高系统容错能力。Mesos 的各组成部分之间通过 LibProcess 和 Protocol Buffers 组成的高效的通信系统。

在 Mesos 中,资源以 Resource Offer 的形式进行调度。Resource Offer 是一种基于资源数量的调度机制,不同于 Hadoop 中的基于 Slot 的调度机制, Mesos 中的资源可以根据不同应用的资源需求灵活的进行资源分配。同时, Mesos 采用一种支持多维资源向量的 Max-Min Fair 资源分配算法 Dominant Resource Fairness[10],使不同应用的不同作业需求能够得到公平的资源分配。

实现集群间资源动态分配需要建立一个全局的资源和作业信息系统,因此,任何一个分布式计算框架,如果想接入 Mesos 获取资源运行作业,需要开发 2 个组件:用于获取 Mesos 分配的资源信息的资源调度器 Framework Scheduler 和用于管理作业并向 Mesos 同步作业信息的作业执行器 Executor。

3 系统设计与实现

经过调研,本文设计了一种新的系统架构,在同一个物理集群上并行运行 Torque/Maui 集群和

Hadoop 集群,使用 Mesos 进行集群资源管理和分配。系统架构如图 3 所示。

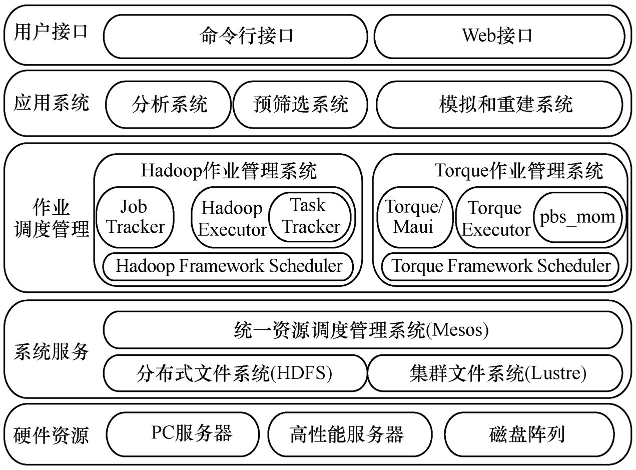

图 3　高能物理数据处理系统框架

该系统使用 PC 服务器和高性能服务器作为计算节点,使用由 Lustre 管理的磁盘阵列和由计算节点本地磁盘构建的分布式文件系统 HDFS 提供数据存储。使用 Mesos 作为系统资源调度管理器。

该系统在应用层提供模拟和重建作业,分析作业,事例预筛选等应用。另外,提供给用户命令行接口和 Web 接口来提交和查询作业。本文使用 AFS 进行用户认证与权限管理,使用 Puppet[11] 和 Ganglia[12] 进行集群管理和监控。

系统的设计和实现的关键技术和本文的主要工作集中在 Torque/Maui 和 Mesos 的集成部分。Mesos 和 Hadoop 的集成和资源分配策略采用了开源社区提供的代码。

3.1　Torque/Maui 与 Mesos 的集成

Torque/Maui 与 Mesos 的集成主要分为 2 个部分:(1)资源调度器 Framework Scheduler 的设计和实现;(2)作业执行器 Executor 的设计和实现。

3.1.1　Framework Scheduler 的设计与实现

在资源调度器 Framework Scheduler 的设计中,主要实现 2 个功能:(1)从 Mesos 处获得可用资源信息,并将这些信息传递给 Torque 的作业调度器 Maui;(2)Maui对分配到的资源进行处理后,将需要执行的作业信息传递给 Mesos,以便分配资源和启动作业。

在原系统中 Maui 使用 Torque 提供的 API(定义在头文件 pbs_ifl. h 中)从 pbs_server 中获取进行调度所需要的信息,包括:队列信息 QueueInfo,节点信息 NodeInfo,作业信息 JobInfo。然后对根据节点状态,队列和作业的优先级等信息,把作业跟可用资源进行匹配,并将匹配的结果返回给 pbs_server,由 pbs_server 发命令到相应 slave 节点上的 pbs_mom 来执行作业,流程如图 4 中粗实线和细实线所示。

为了将 Torque/Maui 与 Mesos 集成,必须建立一个全局的资源及作业信息系统,因此,修改 Maui 中的资源查询函数,使其改向资源调度器 Framework Scheduler 查询可用的资源信息,并在作业调度结束后,将作业的调度结果反馈给资源调度器。资源调度器接收到作业信息后,由 Mesos 确定其需要的资源数量,为各节点上的 pbs_mom 分配资源。由于 Maui 是使用 C 语言编写的,而 Mesos 使用 C + + 编写,因此本文使用 Socket 在 2 个程序间交换信息。

新系统中 Maui 和 Framework Scheduler 的架构设计和资源、作业的调度流程和信息交互如图 4 中细实线和虚线所示。

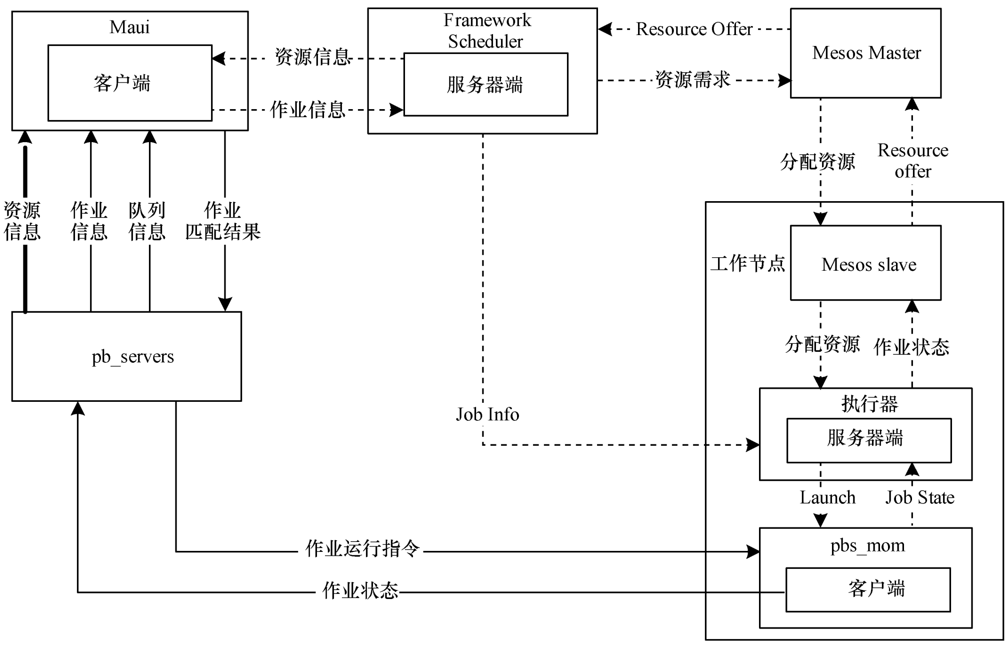

图 4　Torque/Maui 与 Mesos 的集成架构

由于 Mesos 资源分配拉(pull)的模式和 Maui 作业分配推(push)的模式有冲突,本文设计了一个 Hashmap 作为缓冲数组来接收 Mesos 分配的资源,在资源接受完毕后,由 Maui 查询并进行作业分配,将每个资源分配到的作业信息写入数组中,作业信息用邻接链表来保存。对于没有分配作业的 offer,调用 offer. decline()方法回收资源,分配给另外的框架使用。资源数组和作业信息结构如图 5 所示。

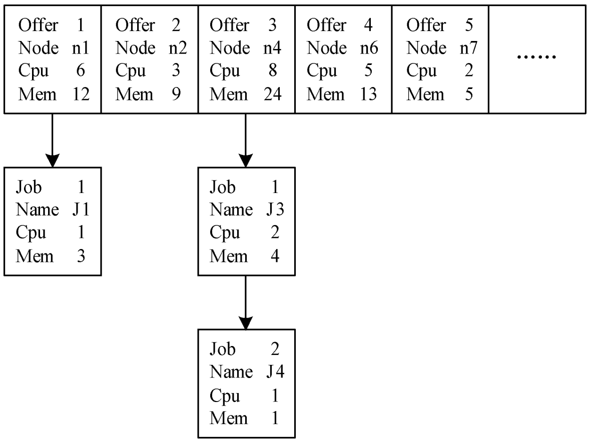

图 5 Framework Scheduler 中的资源和作业信息

3.1.2 Executor 的设计与实现

Executor 的功能是获取节点上 Mesos 分配给它的资源,对作业进行操作和更新作业状态。在作业执行的过程中,Mesos 需要监控作业的运行状态,并根据作业状态决定对该作业的操作和释放资源。

在 Executor 的设计中,只需要让 Executor 启动pbs_mom,具体作业的执行和操作由 pbs_mom 来进行。pbs_mom 在作业执行的过程中,会向 pbs_server 汇报作业状态信息,本文利用这个动作来更新 Mesos 中的作业状态。在这里仍然使用 Socket 进行进程间的通信。

Executor 在启动作业时,为每一个作业都创建了一个线程负责更新该作业的状态,并通过一个全局变量的数组来保存正在运行的作业的状态。pbs_mom 在修改作业状态的同时,会通过 Socket 把作业的状态发送给本地的 Executor。Executor 在接收到作业状态信息后,更新数组中对应作业的状态,并触发相应的作业线程去更新 Mesos 中的作业状态,或触发完成作业释放资源、杀掉作业等操作。Executor 的结构设计如图 4 右下角所示,这样的设计可以把 2 个系统之间的通信分散在各个计算节点上,保证系统的稳定性。

3.2 信息交互格式定义

为简化消息处理机制,使 2 个系统之间的信息交互更简单高效且易读,本文定义了一个结构体 request 来传递消息:

```
struct    request
{
    int       req_type;
    string    hostname;
    string    message;
    int       value;
}
```

结构体 request 中各字段的名称和含义如表 1 所示,其中,req_type 指定消息类型;hostname 指定消息相关的节点名称;message 指定消息内容;value 指定消息的值。表 1 中的示例为:更新节点 host01 上的作业 job0103 的状态为 6。

表 1 结构体 request 各字段定义及示例

字段名称	字段含义	示例
req_type	信息类型	update_job_state
hostname	节点名称	host01
message	消息内容	job0103
value	值	6

4 系统评估测试

系统的实现基于 HDFS2.0,MapReduce0.20,Torque2.5.5,Maui3.2.6 和 Mesos0.14.0。通过部署在 X86 架构的服务器,每个节点的配置为 8 × 2.4 GHz CPU,24 GB 内存,千兆以太网卡,2 ×2 TB 的 SATA 硬盘,形成一个由 1 个服务器节点、6 个计算节点组成的混合集群。

系统测试采用真实的 BESIII 实验数据和真实的高能物理实验数据处理软件 BOSS[13] 和 BEAN,运行真实的模拟作业、重建作业和分析作业作为测试程序。

图 6 显示的是 Torque/Maui 集群在运行真实物理作业的 CPU 资源使用情况,可以看到,CPU 核平均有约 20% 左右的空置。

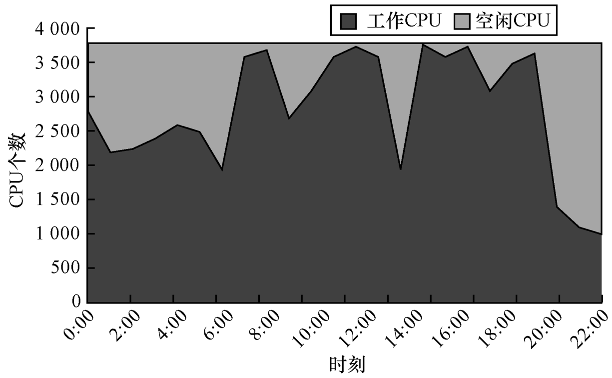

图 6 BESIII 集群 CPU 使用情况

图 7 显示的是采用新的系统架构后集群的整体 CPU 使用情况和 CPU 在 2 个框架间的动态分配情况。本文设置 2 个集群的资源使用比例为 Hadoop:Torque =5:7。由图 7 可见,当没有 Hadoop 作业时,Torque 集群可以独占集群的 CPU 资源,当 Hadoop 中有作业时,资源占用比例逐渐变为 5:7。当 Hadoop 中作业数量减少时,Torque 占用资源增加,当 Torque 作

业数量减少时,Hadoop 资源占用量增加,集群资源可以根据 2 个集群的作业数量动态地进行调整。

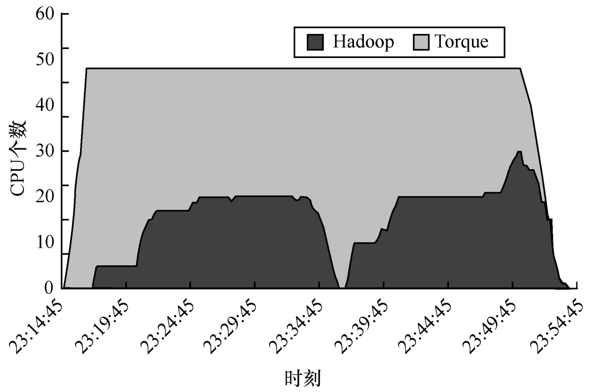

图 7　Hadoop 与 Torque 占用资源比例

图 8 显示的是测试集群分别运行 Torque/Maui 和混合集群时的网络 I/O 速度。由图 8 可见,仅运行 Torque/Maui 时,6 个计算节点的总 I/O 速度约为 200 MB/s,平均每个节点 33 MB/s。如果以现在 BESIII 集群 1 000 个节点的规模估算,总带宽约为 33 GB/s,所以 BESIII 集群的带宽压力很大。而运行混合集群时,由于 Hadoop 集群基本上都从本地读取数据,混合集群的网络 I/O 速度下降了 40% ,因此大大降低了带宽压力。

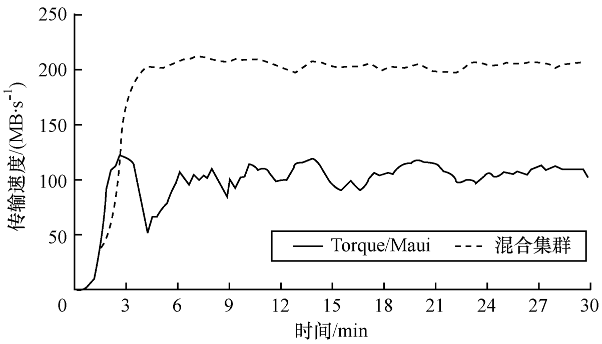

图 8　测试集群网络 I/O 速度

图 9 显示的是测试集群分别运行 Torque/Maui 和混合集群时的磁盘 I/O 速度。

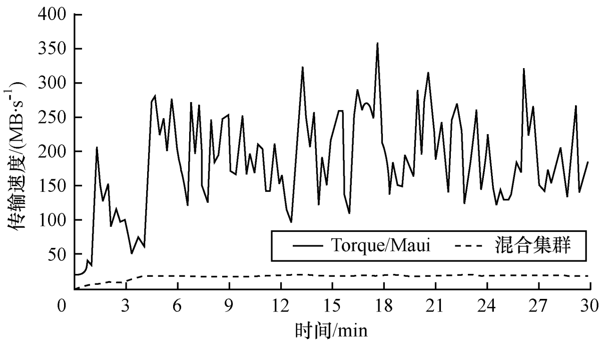

图 9　测试集群本地磁盘 I/O 速度

可以看到,因此 Torque 集群运行时,作业数据全部从网络读取,所以磁盘 I/O 速度很低,共 20 MB/s 左右。但是 Hadoop 集群由于需要从本地磁盘读取数据,因此磁盘 I/O 速度较高,平均有 200 MB/s 左右。这个数值是图 8 中网络 I/O 下降数值的 2 倍,也说明了使用 Hadoop 处理分析作业时的数据读取效率比 Torque/Maui 集群要高。

5　结束语

本文在分析高能物理数据处理的实际需求和集群特点后,利用 Mesos 构建了一个支持集群间资源动态分配的高能物理实验数据处理框架。经过初步测试,证明该系统架构可以在不同的框架间动态地分配集群资源,有效降低网络 I/O 压力,利用计算节点本地磁盘扩充系统存储容量,显著提高了资源利用率。

在实际应用中会遇到更多问题,在资源分配上还需要进一步改进,例如同一集群中的机器性能和配置上可能有较大差异,如何为不同的应用分配合适的机器是资源分配时需要考虑的因素之一。

参考文献

[1] Staples G. TORQUE Resource Manager[C]//Proceedings of 2006 ACM/IEEE Conference on Supercomputing. New York,USA:ACM Press,2006.

[2] Adaptive Computing. Maui[EB/OL].[2014-02-15]. http://www.adaptivecomputing.com/products/open-source/maui/.

[3] Yahoo. Apache Hadoop[EB/OL].[2014-02-15]. http://hadoop.apache.org/.

[4] Dean J, Ghemawat S. MapReduce: Simplified Data Processing on Large Clusters[J]. Communications of the ACM,2008,51(1):107-113.

[5] Shvachko K, Kuang H, Radia S, et al. The Hadoop Distributed File System[C]//Proceedings of the 26th IEEE Symposium on Mass Storage Systems and Technologies. Incline Village,USA:IEEE Press,2010:1-10.

[6] 臧冬松,霍　菁,梁　栋,等. 基于 MapReduce 的高能物理数据分析系统[J]. 计算机工程,2014,40(2):1-5.

[7] Hindman B, Konwinski A, Zaharia M, et al. Mesos: A Platform for Fine-grained Resource Sharing in the Data Center[C]//Proceedings of NSDI'11. Berkeley,USA: USENIX Association,2011:22-22.

[8] Schwan P. Lustre:Building a File System for 1000-node Clusters[C]//Proceedings of 2003 Linux Symposium. Ottawa,Canada:[s. n.],2003:380-386.

[9] Hunt P,Konar M,Junqueira F P,et al. ZooKeeper: Wait-free Coordination for Internet-scale Systems[C]//Proceedings of 2010 USENIX Conference on USENIX Annual Technical Conference.[S. l.]: USENIX Association,2010:11.

[10] Ghodsi A, Zaharia M, Hindman B, et al. Dominant Resource Fairness: Fair Allocation of Multiple Resource Types[C]//Proceedings of NSDI'11. Berkeley,USA: USENIX Association,2011:323-336.

[11] Puppet Labs. What is Puppet?[EB/OL].[2014-02-15]. https://puppetlabs.com/puppet/what-is-puppet/.

[12] IBM. Ganglia Monitoring System[EB/OL].[2014-02-15]. http://ganglia.info/.

[13] Li Weidong, Liu Huaiming, Deng Ziyan, et al. The Offline Software for the BESIII Experiment[C]//Proceedings of CHEP'06. Mumbai,India:[s. n.],2006.

编辑　金胡考