(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 107733696 A
(43)申请公布日 2018.02.23

(21)申请号 201710881113.0

(22)申请日 2017.09.26

(71)申请人 南京天数信息科技有限公司
地址 210000 江苏省南京市雨花台区软件大道180号5栋4层

(72)发明人 李云鹏 倪岭 任义龙 张建 刘伟佳 赵志强

(74)专利代理机构 南京钟山专利代理有限公司 32252
代理人 戴朝荣

(51)Int.Cl.
H04L 12/24(2006.01)
H04L 29/06(2006.01)
H04L 29/08(2006.01)
G06F 17/30(2006.01)

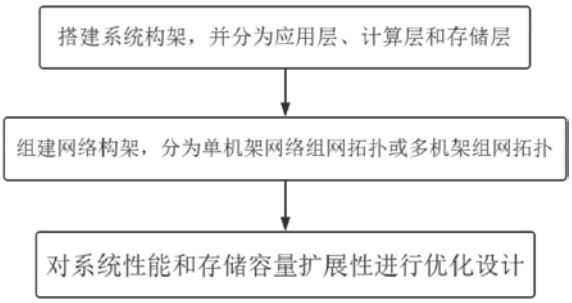
权利要求书2页 说明书6页 附图2页

(54)发明名称

一种机器学习和人工智能应用一体机部署方法

(57)摘要

本发明公开了一种机器学习和人工智能应用一体机部署方法，该方法是通过以下步骤实现的：(1)搭建系统构架，将所述系统架构在逻辑上分为应用层、计算层和存储层；(2)组建网络架构，将所述网络架构在逻辑上划分为外部网、管理网、计算网和存储网；(3)对系统的扩展性进行优化设计，采用横向扩展架构增加计算资源，提高性能，采用分层架构增加存储容量。本发明公开的一体机部署方法通过使用丰富的算法库、高性能计算引擎等，设计出功能集成的系统架构、灵活方便的组网拓扑以及优秀高效的系统可扩展性，从而使一体机加快大数据机器学习速度，提升了人工智能分析程序的运行效率。



CN 107733696 A

1.一种机器学习和人工智能应用一体机部署方法,其特征在于,包括以下步骤:

步骤一,将数据存储和数据处理进行隔离,采用高可扩展性的Shared-Nothing架构搭建整体系统架构,所述系统架构在逻辑上分为应用层、计算层和存储层,并且应用层、计算层和存储层都采用分布式架构;

步骤二,组建网络架构,网络构架分为单机架组网拓扑或多机架组网拓扑,所述网络架构在逻辑上划分为外部网、管理网、计算网和存储网;

步骤三,对系统的扩展性进行优化设计。

2.根据权利要求1所述的一种机器学习和人工智能应用一体机部署方法,其特征在于,步骤一中,所述应用层根据实际需要配置不同数量的应用节点;所述计算层根据实际需要配置不同数量的计算节点;所述存储层根据实际需要配置不同数量的存储节点。

3.根据权利要求2所述的一种机器学习和人工智能应用一体机部署方法,其特征在于,所述计算节点配置如下软件栈:

a.支持多种编程语言;

b.提供用于机器学习及深度学习的API;

c.集成了深度学习框架TensorFlow;

d.集成了优化过的分布式计算框架Spark;

e.集成了优化过的分布式内存文件系统Alluxio来加速数据读写;

f.集成了优化过的RDMA特性。

4.根据权利要求2所述的一种机器学习和人工智能应用一体机部署方法,其特征在于,所述存储节点提供数据库和通用型文件系统两种存储服务;所述数据库包括关系型数据库PostgreSQL和时序型数据库,所述关系型数据库PostgreSQL采用HAWQ分布式构架,所述时序型数据库采用OpenTSDB+Hase分布式构架;所述通用型文件系统采用HDFS+Ceph混合结构,HAQW底层采用HDFS。

5.根据权利要求2所述的一种机器学习和人工智能应用一体机部署方法,其特征在于,所述应用节点上部署外部网网卡和管理网网卡,所述计算节点上部署管理网网卡和计算存储网网卡,所述存储节点上部署管理网网卡和计算存储网网卡。

6.根据权利要求1～5中任一项所述的一种机器学习和人工智能应用一体机部署方法,其特征在于,步骤二中所述单机架组网拓扑包含一个机架,组建方法为:

配备一台以太网交换机,所述以太网交换机的端口数大于或等于机架内的总节点数;

配备一台计算存储网交换机,所述计算存储网交换机的端口数大于或等于机架内的总节点数;

配备一台外部网交换机。

7.根据权利要求1～5中任一项所述的一种机器学习和人工智能应用一体机部署方法,其特征在于,步骤二中所述多机架组网拓扑包含多个机架,组建方法为:

每个机架配备一台以太网交换机,所述以太网交换机的端口数大于机架内的总节点数,并为连接其他机架预留端口;

每个机架配备一台计算存储网交换机,所述计算存储网交换机的端口数大于机架内的总节点数,并为连接其他机架预留端口;

配备合适数量的外部网交换机;

配备核心交换机，各个机架的管理网交换机采用简单树形连接到所述核心交换机上。

8.根据权利要求7所述的一种机器学习和人工智能应用一体机部署方法，其特征在于，所述计算存储网交换机为InfiniBand交换机，各个机架的InfiniBand交换机连接多个所述核心交换机组成胖树结构。

9.根据权利要求2～5中任一项所述的一种机器学习和人工智能应用一体机部署方法，其特征在于，步骤一中所述对系统的扩展性进行优化设计的具体步骤如下：

采用横向扩展架构提高性能；

采用分层架构增加存储容量。

10.根据权利要求9所述的一种机器学习和人工智能应用一体机部署方法，其特征在于，所述采用横向扩展架构提高性能的步骤为：

增加所述计算层中的计算节点数；

增加适量的网络交换机。

一种机器学习和人工智能应用一体机部署方法

技术领域

[0001] 本发明涉及机器学习和人工智能技术领域，特别涉及一种机器学习和人工智能应用一体机部署方法。

背景技术

[0002] 人工智能早在上个世纪50年代就被提出，它是控制论、信息论、计算机科学、数理逻辑、神经生理学、心理学、语言学、教育学、医学、工程技术以及哲学等多种学科相互渗透的交叉学科。人们梦想着用当时刚刚出现的计算机来构造复杂的、拥有与人类智慧同样本质特性的机器。这个无所不能的机器，它有着我们所有的感知(甚至比人更多)，我们所有的理性，可以像我们一样思考。机器学习是研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。它是人工智能的核心，是使计算机具有智能的根本途径，其应用遍及人工智能的各个领域。机器学习最基本的做法，是使用算法来解析数据，从中学习，然后对真实世界中的事件做出决策和预测。与传统的为解决特定任务、硬编码的软件程序不同，机器学习是用大量的数据来“训练”，通过各种算法从数据中学习如何完成任务。

[0003] 机器学习是人工智能发展中一个十分热门的领域，机器学习的研究目的，是希望计算机具有像人类一样从现实世界获取知识的能力，同时建立学习的计算理论，构造各种学习系统并将其应用到各个领域中去。机器学习研究主要有三个方向，一是以模拟人类的学习过程出发，试图建立学习的认识生理学模型，这个方向与认知科学的发展密切相关；二是基础研究，发展各种适合机器特点的学习理论，探讨所有可能的学习方法，比较人类学习与机器学习的异同与联系；三是应用研究，建立各种实用的学习系统或知识获取辅助工具，在人工智能科学的应用领域建立自动获取知识系统，积累经验，完善知识库与控制知识，进而使机器的智能水平类似于人类。

[0004] 目前，包括百度和谷歌在内的科技巨头，2016年在人工智能上的投入在200亿至300亿美元之间，其中90％投入研发和部署上，还有10％用于人工智能收购。目前的人工智能投资速度3倍于2013年以来的外部投资增长。人工智能发展领域主要集中在高科技/电信、汽车/组装和金融服务行业。机器学习切实能被用来帮助工业界解决问题，特别是当下的热点，如深度学习、无人驾驶、人工智能助理等对工业界的影响巨大。

[0005] 大数据推动了人工智能的发展，同时，人工智能的发展也让数据产生巨大的价值，成为“智能数据”。人工智能现已应用在各种大数据应用中，如:搜索推荐、购物推荐、语音识别、图像识别、聊天机器人，智能医疗等等。机器学习和人工智能是在大数据的基础上不断发展起来的，为了让杂乱无章的海量数据产生价值，需要使用复杂的网络模型对数据进行大量地分析，才能训练处高准确率的模型，这就需要庞大的计算量，因此计算能力对机器学习和人工智能的发展变的越来越重要。

[0006] 目前的大数据机器学习算法和人工智能分析应用效率较低，而资源占用率较高。对海量数据的处理速度较慢，并且大量数据在处理过程中对硬件的要求极高，无法满足数

据驱动型企业快速增长的智能计算要求。

发明内容

[0007] 为解决现有技术的不足,本发明的目的在于提供一种机器学习和人工智能应用一体机部署方法,该方法采用了的专门设计和多种优化技术,使得一体机具有超高计算性能,能够显著地加快程序的运行速度,适合应用于大数据环境下的机器学习和人工智能应用。

[0008] 为了实现上述目标,本发明采用如下的技术方案:一种机器学习和人工智能应用一体机部署方法,其特征在于,包括以下步骤:

[0009] 步骤一,将数据存储和数据处理进行隔离,采用高可扩展性的Shared-Nothing架构搭建整体系统架构,所述系统架构在逻辑上分为应用层、计算层和存储层,并且应用层、计算层和存储层都采用分布式架构;

[0010] 步骤二,组建网络架构,网络构架分为单机架组网拓扑或多机架组网拓扑,所述网络架构在逻辑上划分为外部网、管理网、计算网和存储网;

[0011] 步骤三,对系统的扩展性进行优化设计。

[0012] 进一步地,所述应用层根据实际需要配置不同数量的应用节点;所述计算层根据实际需要配置不同数量的计算节点;所述存储层根据实际需要配置不同数量的存储节点。

[0013] 进一步地,所述计算节点配置如下软件栈:

[0014] 支持多种编程语言;

[0015] 提供用于机器学习及深度学习的API;

[0016] 集成了深度学习框架TensorFlow;

[0017] 集成了优化过的分布式计算框架Spark;

[0018] 集成了优化过的分布式内存文件系统Alluxio来加速数据读写;

[0019] 集成了优化过的RDMA特性。

[0020] 进一步地,所述存储节点提供数据库和通用型文件系统两种存储服务;所述数据库包括关系型数据库PostgreSQL和时序型数据库,所述关系型数据库PostgreSQL采用HAWQ分布式构架,所述时序型数据库采用OpenTSDB+Hase分布式构架;所述通用型文件系统采用HDFS+Ceph混合结构,HAQW底层采用HDFS。

[0021] 进一步地,所述应用节点上部署外部网网卡和管理网网卡,所述计算节点上部署管理网网卡和计算存储网网卡,所述存储节点上部署管理网网卡和计算存储网网卡。

[0022] 进一步地,所述单机架组网拓扑包含一个机架,组建方法为:

[0023] 配备一台以太网交换机,所述以太网交换机的端口数大于或等于机架内的总节点数;

[0024] 配备一台计算存储网交换机,所述计算存储网交换机的端口数大于或等于机架内的总节点数;

[0025] 配备一台外部网交换机。

[0026] 进一步地,所述多机架组网拓扑包含多个机架,组建方法为:

[0027] 每个机架配备一台以太网交换机,所述以太网交换机的端口数大于机架内的总节点数,并为连接其他机架预留端口;

[0028] 每个机架配备一台计算存储网交换机,所述计算存储网交换机的端口数大于机架

内的总节点数，并为连接其他机架预留端口；

[0029] 配备合适数量的外部网交换机；

[0030] 配备核心交换机，各个机架的管理网交换机采用简单树形连接到所述核心交换机上。

[0031] 进一步地，所述计算存储网交换机为InfiniBand交换机，各个机架的InfiniBand交换机连接多个所述核心交换机组成胖树结构。

[0032] 进一步地，所述对系统的扩展性进行优化设计的具体步骤如下：

[0033] 采用横向扩展架构提高性能；

[0034] 采用分层架构增加存储容量。

[0035] 更进一步地，所述采用横向扩展架构提高性能的步骤为：

[0036] 增加所述计算层中的计算节点数；

[0037] 增加适量的网络交换机。

[0038] 本发明的有益之处在于：

[0039] (1)一体机将数据存储和数据处理进行隔离，采用高可扩展性的Shared-Nothing架构，将客户端、数据处理、数据存储进行分离，逻辑上分为三个层次：应用层，计算层，存储层。每个层次都采用分布式架构，可以达到较高的计算并发度和数据读写并发度，同时让整个系统具有良好的可扩展性、可靠性和可维护性。

[0040] (2)分布式超融合硬件架构和与软件栈的巧妙搭配，避免存储和计算资源的浪费，保障数据分析流水线的稳定性，提升分析效率。针对硬件架构的各个层面，包括CPU、内存、层次化存储、GPU都进行了专门的优化，充分挖掘了硬件的能力。同时还深度集成了TensorFlow等框架，对分布式机器学习算法和通信机制进行了大量优化。

[0041] (3)通过框架、算法改进以及硬件的充分利用，实现数量级的计算加速，降低企业对大数据基础设施及人力的投入。通过数据清洗、建模分析，获得高质量、有意义的信息，从而挖掘出数据价值。

附图说明

[0042] 图1是本发明流程图；

[0043] 图2是系统的整体构架示意图；

[0044] 图3是系统的组件部署构架示意图。

具体实施方式

[0045] 以下结合附图和具体实施例对本发明作具体的介绍。

[0046] 参照图1所示，本发明一种机器学习和人工智能应用一体机部署方法，包括以下步骤：

[0047] 步骤一，一体机将数据存储和数据处理进行隔离，采用高可扩展性的Shared-Nothing架构，在整体上可分为三层：应用层，计算层，存储层。一体机采用分层架构，拥有完整的冗余的硬件保护，任何一个计算节点或者存储节点出现故障，能够保证数据不会丢失，且一体机仍能够正常工作，极大地提高了系统的可靠性。

[0048] 将客户端、数据处理、数据存储进行分离，逻辑上分为三个层次，每个层次都采用

分布式架构，可以达到较高的计算并发度和数据读写并发度，同时让整个系统具有良好的可扩展性、可靠性和可维护性。

[0049] 其中，应用层主要运行用户接口服务，像处理登录、监控、管理、计算任务编排/提交等工作。要求CPU和内存配置中等，存储容量配置低；计算层用来执行用户提交的计算任务。要求CPU和内存配置高，存储容量配置低；存储层主要为计算节点提供大容量存储。要求CPU和内存配置低，存储容量配置高。如图2所示，一体机还能根据不同的实际需要，灵活地配置不同数量的应用节点、计算节点和存储节点，且具有高度的可扩展性，集成了WebUI、资源管理、系统监控，资源调度，任务管理等功能。

[0050] 参照图3所示。应用节点通过提供一个Web UI，让用户方便地进行任务管理，系统监控，资源管理等管理方式，应用节点作为一体机的最外层，暴露给用户操作。应用节点，具体将提供如下功能接口：应用管理(应用提交/应用删除/应用状态查询)、数据储存和查询(结构化存储接口/非结构化存储接口)、文件管理(拷贝/粘贴/上传/下载/创建/移动/删除)、资源监控(GPU/CPU/Memory/Network/Disk/Others)、管理(资源管理/角色管理/用户管理/包管理/节点管理)。

[0051] 计算节点针对计算资源的大量耗费进行优化，采用专门设计的软件栈：

[0052] a.提供了多种编程语言的支持，如Python、R、Java、Scala等；

[0053] b.提供了用于机器学习以及深度学习的API，同时也支持一些其它的通用计算API；

[0054] c.集成了深度学习框架TensorFlow。Tensorflow作为深度学习框架，大量的应用基于此框架上进行开发，一体机集成了该框架，使得基于此框架开发的应用能够直接在其上运行。

[0055] d.集成了优化过的分布式计算框架Spark。Spark是一个高效的分布式计算系统，在此基础上，优化了Spark的底层算法库，使得分布式任务在每个计算节点具有更快的运行速度。

[0056] e.集成了优化过的分布式内存文件系统Alluxio来加速数据读写。Alluxio是一个分布式内存文件系统，允许文件以内存的速度在集群框架中进行可靠的共享，在此基础上，进一步进行了优化，使得调度框架能够更好的利用Alluxio的分布式内存特性。

[0057] f.集成了优化过的RDMA特性：JXIO。RDMA(Remote Direct Memory Access)技术可以解决网络传输中服务器端数据处理的延迟问题。RDMA通过网络把数据直接传入计算机的存储区，将数据从一个系统快速移动到远程系统存储器中，而不对操作系统造成任何影响，这样只需要用到很少的CPU资源。它消除了外部存储器复制和文本交换操作，从而释放了内存带宽和CPU周期用于提供应用程序性能。

[0058] 另外根据一体机分布式框架的特点，优化的计算平台会被部署到每个计算节点中，由于上层使用mesos进行任务调度和资源管理，因此每个计算节点的角色是完全相同的，不区分master和worker节点的概念。

[0059] 一体机的存储节点负责提供存储功能，主要提供数据库和通用型文件系统两种存储服务。数据库可以分为两类，分别为关系型数据库PostgreSQL和时序型数据库。分布式集群方案分别采用PostgreSQL的HAWQ和时序数据库的OpenTSDB+Hase。参照图3所示，文件系统同样采用分布式结构，使用HDFS+Ceph混合结构，HAQW底层采用HDFS，其它组件均使用

Ceph进行存储。

[0060] 因此，上层数据管理工具会根据文件的存储方式，自动选择数据是存储在HDFS上还是Ceph上。数据库软件层部署在计算集群，文件系统软件部署在存储集群，各个集群数据管理功能定位如下：

[0061] 1)应用集群：

[0062] 提供数据的统一访问接口；

[0063] 提供大规模数据的import/export接口；

[0064] 部署数据库的管理软件客户端；

[0065] 部署数据库状态的监控工具。

[0066] 2)计算集群：

[0067] 部署数据库管理软件；

[0068] 提供SQL/REST API接口。

[0069] 3)存储集群：

[0070] 采用混合的HDFS、Ceph分布式文件系统；

[0071] 支持块存储、对象存储。

[0072] 步骤二，组建网络架构，分为单机架组网拓扑和多机架组网拓扑两种，在逻辑上划分为外部网、管理网、计算网和存储网。

[0073] 外部网：用于连接用户的交换机，对外提供访问一体机服务的网络。对外的网络接口采用普通1Gbps以太网即可，只在应用节点上部署外部网网卡。

[0074] 管理网：用于监控、管理一体机的各个节点，以及向计算节点提交计算任务等。这些任务对网络带宽和延时要求不是很高，同时为避免影响计算网和存储网，采用独立于计算网和存储网的普通1Gbps以太网即可，需要在每个节点上都部署管理网网卡。

[0075] 计算网：用于连接各个计算节点，对网络延时要求很高(高配版采用InfiniBand网卡)。

[0076] 存储网：用于连接各个存储节点，对网络带宽和延时要求很高，这里采用高带宽和低延时的56Gbps InfiniBand(标准版可以采用10Gbps的RoCE网卡)将存储网和计算网融合为一个网络。除了在计算节点和存储节点都部署InfiniBand网卡，考虑到应用节点也可能需要访问存储节点的数据，所以应用节点可以考虑也部署InfiniBand网卡。

[0077] 一体机安装方式以机架为单位，每个机架内可包含若干个应用节点、计算节点和存储节点。每个机架配备一台以太网交换机(管理网)、一台InfiniBand交换机(计算存储网)，每台交换机的端口数应该不小于该机架内的总节点数。如果需要扩展多个机架，交换机还要预留一定的端口数以连接到其他机架。至于外部网交换机，考虑到应用节点相对较少，可以考虑多个机架共用一个交换机。对于多个机架的组网，需要增加额外的核心交换机来连接各个机架，各个机架的管理网交换机，可以采用简单的树形汇聚到一个核心交换机。而采用InfiniBand的计算存储网交换机需要采用多个核心交换机来组成胖树结构以保证任意两个节点间都有全带宽通道。

[0078] 步骤三，一体机的设计部署使其具有良好的系统可扩展性，主要分为性能扩展和存储容量扩展。

[0079] 一体机采用横向扩展的架构，可以通过增加计算层中的计算节点，从而增加整体

计算资源(GPU/CPU/Memory)，进而提高应用程序运行速度。当在大量增加计算节点时，可能导致网络成为瓶颈，为了保持计算资源(GPU/CPU/Memory)、存储和网络处在一种平衡模式上，需要增加适量的网络交换机去解决网络瓶颈问题。一体机采用分层架构，将数据存储单元和计算处理单元分离，因此，当需要大量存储的情况下，可直接横向地增加存储容量，非常便捷。

[0080] 以上显示和描述了本发明的基本原理、主要特征和优点。本行业的技术人员应该了解，上述实施例不以任何形式限制本发明，凡采用等同替换或等效变换的方式所获得的技术方案，均落在本发明的保护范围内。

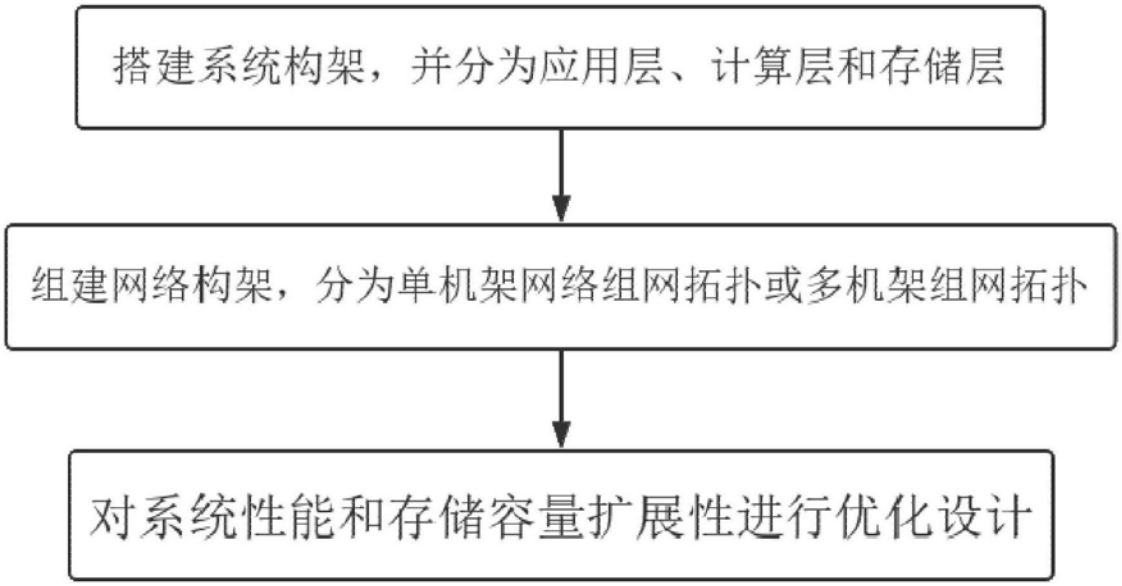

搭建系统构架，并分为应用层、计算层和存储层
组建网络构架，分为单机架网络组网拓扑或多机架组网拓扑
对系统性能和存储容量扩展性进行优化设计

图1


应用层
应用节点1
应用节点1
…
计算层
计算节点1
计算节点2
…
存储层
存储节点1
存储节点2
…

图2

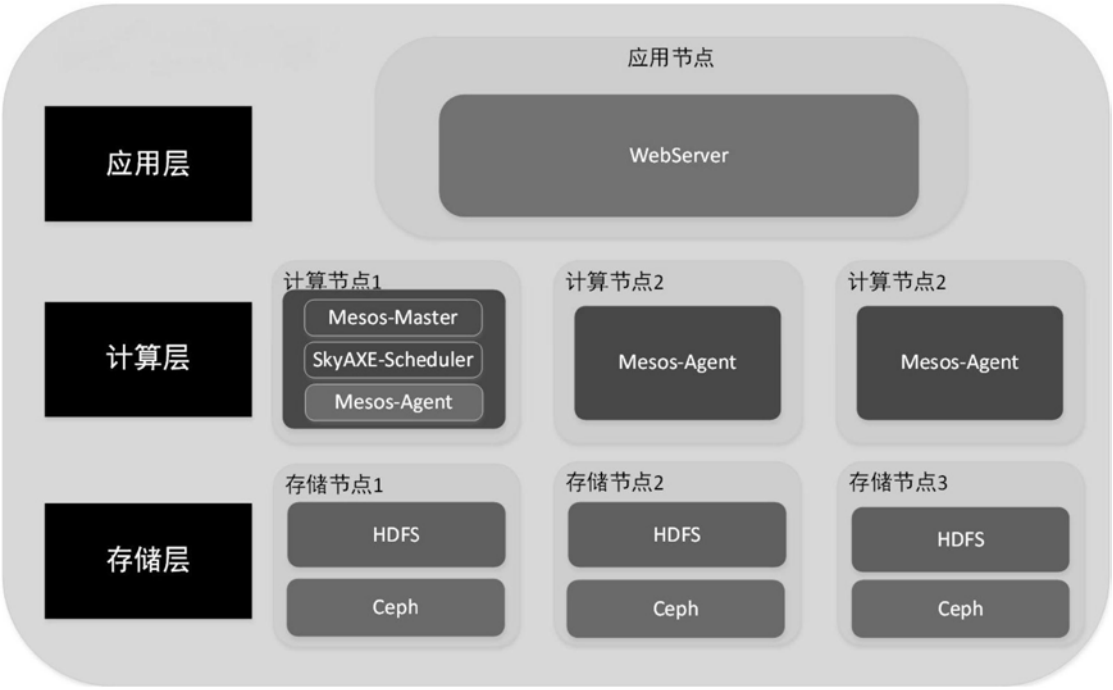

应用节点
应用层
WebServer
计算节点1
计算层
Mesos-Master
SkyAXE-Scheduler
Mesos-Agent
计算节点2
Mesos-Agent
计算节点2
Mesos-Agent
存储节点1
存储层
HDFS
Ceph
存储节点2
HDFS
Ceph
存储节点3
HDFS
Ceph

图3