Docker容器虚拟化技术是一种共享操作系统内核的虚拟化解决方案，基于Docker技术的云平台日益发展，逐步成为下一代云计算的核心。容器云中资源调度对集群的性能和资源利用率起到决定性作用。Kubernetes容器编排引擎以其强大的服务发现、集群监控、错误恢复能力成为当前应用最为广泛的容器云平台调度器。但Kubernetes没有考虑多计算框架资源调度问题，导致容器云集群进行大数据处理时资源利用率低、负载均衡性差。

针对多计算框架下Kubernetes资源调度性能不足，本文提出了一种基于多维资源空闲率权重的调度方法MRWS(Multidimensional Resource Weights Scheduling)，并基于开源容器云OpenShift研发了面向多计算框架的容器云平台Paladin，在该平台上设计和实现新的调度方法并进行调度性能测试。主要贡献如下:

* 提出了面向多计算框架的容器云资源调度方法MRWS。该方法综合考虑待调度容器的资源需求、物理节点的CPU、内存、磁盘、网络带宽空闲率和已部署的容器应用数量等因素，基于多维资源权重参数进行综合评分。
* 设计自动求解MRWS权重参数的方法。针对MRWS调度方法中的权重参数，使用模糊层次分析法FAHP(Fuzzy Analytic Hierarchy Process)对集群资源自动建模并求解容器应用多维资源权重参数。
* 在容器云平台上设计和实现MRWS调度方法并进行性能分析。首先在容器云仿真平台ContainerCloudSim上进行大规模的调度仿真，并与Random、FirstFit、Kubernetes调度算法进行对比，新的调度方法能提升约40%的集群负载均衡。然后基于OpenShift研发了面向多计算框架的容器云平台Paladin。最后在该平台上进行多计算框架资源调度性能测试，新的调度方法约提升20%的任务处理性能。

Docker container virtualization technology is a virtualization solution that shares the operating system kernel. The cloud platform based on Docker technology is gradually becoming the core of the next generation cloud computing. In the container cloud, the resource scheduler plays a key role in the performance and resource utilization. Kubernetes container orchestration engine has become the most widely used container cloud platform scheduler with its powerful service discovery, cluster monitoring and error recovery capabilities. However, Kubernetes does not consider the multi-computing framework resource scheduling problem, resulting in low resource utilization and poor load balance when the container cloud cluster processing big data problems.

For the insufficient performance of Kubernetes resource scheduling under multi-computing framework, this thesis proposed a MRWS(Multidimensional Resource Weights Scheduling) scheduling method, and developed a container cloud platform Paladin for multi-computing framework based on open-source container OpenShift. Designed and implemented new scheduling methods on the platform and performed scheduling performance tests. The main contributions are as follows:

* A container cloud resource scheduling method MRWS for multi-computing framework is proposed. The method comprehensively considers the resource requirements of the container to be scheduled, the CPU of the physical node, the memory, the disk, the network bandwidth idle rate, and the number of deployed container applications, and comprehensively scores based on the multi-dimensional resource weighting parameter.
* Designed a method to automatically solve MRWS weight parameters. For the weight parameters in the MRWS scheduling method, the FAHP(Fuzzy Analytic Hierarchy Process) is used to automatically model the cluster resources and solve the container application multi-dimensional resource weight parameters.
* Designed and implemented the MRWS scheduling method on the container cloud platform and performed performance analysis. Firstly, a large-scale scheduling simulation is performed on the container cloud simulation platform ContainerCloudSim, and compared with the Random, FirstFit, and Kubernetes scheduling algorithms. The new scheduling method can improve the cluster load balance by about 40%. Then based on OpenShift, the container cloud platform Paladin for multi-computing framework was developed. Finally, the multi-computing framework resource scheduling performance test is performed on the platform, and the new scheduling method improves the task processing performance by about 20%.

# 引言

## 云计算资源调度研究概述

随着计算机网络和互联网技术的飞速发展，网络规模不断扩大，各行业中应用业务量和数据都呈现爆炸式的增长，如何存储和高效处理各种应用产生的海量数据已成为各大互联网企业和机构面临的一个巨大挑战。继分布式计算、网格计算和并行计算后，一种新型的将整个互联网资源聚合起来处理数据的计算模式应运而生: 云计算 。这种共享资源、按需付费、统一管理、可伸缩、可度量的计算模式发展迅猛，给信息产业带来了新的变革。云计算服务资源池有基础设施即服务(IaaS，Infrastructure as a Service)、平台即服务(PaaS，Platform as a Service)和软件即服务(SaaS，Software as a Service)三种服务模式。计算类型根据用户对象的不同可以划分为公有云、私有云和混合云，典型的公有云有Google Cloud、Amazon的EC2、微软Azure、阿里云ECS、百度云、腾讯云等。云计算需要底层的虚拟化技术作为支撑，当前维基百科收录的就有超过60种虚拟化技术，其中基于X86体系的虚拟化就超过50种，也有RISC体系的虚拟化。虚拟化主要包括硬件虚拟化、操作系统层虚拟化、桌面虚拟化、应用程序虚拟化以及网络虚拟化等。云计算虚拟化服务通常以虚拟机的形式提供给用户，用户根据自己的资源需求和应用场景申请合适的虚拟机进行服务。

在云计算中，资源调度对集群的性能和资源利用率起到决定性作用，是云计算的核心之一。针对传统云计算系统调度策略的研究 相当广泛，主要集中在降低系统能耗、提高云计算集群资源利用率、集群服务器的负载均衡以及基于成本模式的资源管理等方面。文献 提出一种根据虚拟机负载动态调节处理器电压和频率来降低集群能耗；文献 通过动态分配云计算中心的虚拟机，减少服务器的数量来节约能耗；文献 通过提出一种中心平衡器的平衡算法来实现集群服务器的负载均衡；文献 将应用需求和物理机计算资源建模，基于蚁群算法、粒子群算法等迭代方式求解最佳分配策略，减少服务器的数量，提升集群资源利用率；文献 提出面向市场的体系结构和资源分配调度方法，该体系结构通过SLA资源分配器实现用户和服务商的协商，从而实现资源的优化配置。由此看出，在传统的云计算系统中，人们对云计算资源调度方法进行深入的研究，广泛应用于当前的云计算系统中，对推动云计算普及起到巨大的作用。

## Docker虚拟化技术概述

虚拟机是云计算的核心技术之一，而以Docker 为代表的容器虚拟化技术近几年迅速崛起，逐步成为一种主流的虚拟化技术。Docker是一种操作系统层面的虚拟化技术，其底层以LXC(Linux Container)作为支撑。和传统面向操作系统的虚拟技术和硬件虚拟化不同，Docker是面向进程提供虚拟运行环境，其提供的虚拟环境就是容器。操作系统Linux可以为容器分配资源，如CPU时间、I/O时间、内存并进行外设访问控制，通过内核控制组(cgroups)子系统限定特定进程使用资源的量，然后使用Linux内核的namespace隔离容器间的进程。这样就可以实现一个高级的容器引擎，开发者可以快速构建、部署、测试和发布应用，容器应用具有较好的跨平台性。从资源管理角度而言，Docker依赖于LXC、LXC基于cgroups子系统，Docker主要是对容器进行封装，管理容器的生命周期、查询和控制相关信息、而所有与操作系统的交互都是通过libcontainer容器引擎完成。

容器与虚拟机对比

容器与虚拟机对比

图1-1中基础设施Infrastructure可以是个人电脑、服务器、云主机等，主机操作系统是运行在基础设施上的系统，主要是Linux的各种版本。虚拟机管理系统(Hypervisor)可以实现在主机操作系统上独立运行多个子操作系统，在子操作系统上安装完应用所需的各种依赖后就可以实现应用资源的隔离。相对于虚拟机，Docker要简便很多，当前所有的Linux版本以及MacOS、Windows都能运行Docker，Docker Engine取代了Hypervisor，负责管理Docker容器并与操作系统通信，各种应用直接打包到镜像文件中，实现容器应用的快速启动和重复应用。

对比Docker和虚拟机的架构 发现，Docker直接通过守护进程与操作系统进行通信、容器管理以及资源分配，实现容器与主操作系统的隔离。Docker没有资源开销较大的子操作系统，各容器直接与主操作系统共享资源，其虚拟化开销大幅减少，应用启动时间甚至达到毫秒级。用户可以快速构建、部署和交付应用，并且具有较强的跨平台性。 虽然Docker拥有众多的优势，但其隔离仅仅是在进程层面进行，并不能完全隔离整个运行环境。因此，用户需要根据自己的实际应用场景，在彻底隔离运行环境的需求下选择虚拟机技术，如果仅是应用层面的隔离选择容器技术，如数据库、前端、后端等。

## 容器云资源调度研究概述

虚拟机是当前云计算提供服务的主要实现形式，也是云计算的核心技术之一，除虚拟机外，容器在云计算中应用越加广泛，容器云发展迅猛。当前容器虚拟技术以Docker为典型代表，这种既能与宿主机共享资源又能为用户提供隔离的虚拟化方案迅速受到人们关注。容器虚拟化技术只需较小的虚拟化开销就可以实现应用的快速部署、开发、测试、交付以及具有较好的跨平台性。以Docker为基础构建的CaaS(Container as a Service)应运而生 ，各大互联网公司都投入巨资进行研发，根据451 Research预测，容器作为一种高速成长型的工具，年增幅高达40%。容器将作为应用最为广泛的云工具，超过OpenStack、PaaS以及其他相关的产品，该机构在调研了125家应用容器厂商后预测，应用容器将从2016年的7.62亿美元增长到2020的27亿美元。容器的管理和调度市场也在进行快速的组合并购，Apprenda收购Kubernetes支持者Kismatic，思科收购Docker Swarm支持者ContainerX等，这些活动都加速了容器云的快速发展。当前较为出色的容器云有Google Container Engine、SAE、 Cloud Foundry、AWS ECS、Red Hat OpenShift等。

容器云和以虚拟机与基础构建的传统云计算系统一样需要一个性能强大的容器编排管理器进行容器调度、创建、销毁、监控、重启、错误恢复以及服务组合等工作。容器调度器既是容器技术普及的重要推力也是决定集群性能和资源利用率的关键因素 ，当前有三大主流的容器编排引擎 : Docker Swarm、Apache Mesos以及Google Kubernetes，其中应用最为广泛的要属轻量开源，性能强大的Kubernetes。在调度算法方面Swarm主要包括Random、Spread和Binpack三种；Mesos更多使用传统的调度算法实现资源分配，如DRF(Dominant Resource Fairness)等；Kubernetes使用两阶段算法选取最大评分节点实现容器调度。几种调度模型和算法各有优缺点，用户可以根据自己的需求选取相应的容器调度框架构建其容器云环境，实现更好的资源分配和调度。本文主要对Kubernetes的调度方式进行研究，在多计算框架容器应用进行大数据处理的场景下提升集群资源利用率和负载均衡，使多计算框架容器应用任务执行时间更短。

## 面向多计算框架的容器云资源调度研究

当前，容器云平台的容器编排引擎对容器应用的调度大多使用常规的调度方法，研究也主要针对调度算法进行，容器调度器将计算框架的容器应用分解为不关联的单个容器进行资源分配和任务调度，没有综合考虑计算框架的特性。针对大数据处理场景下容器云如何提升资源利用率和任务执行效率的研究较为分散，往往对单个处理框架进行研究，不存在通用调度方法，也几乎没有针对多计算框架容器应用同时执行任务处理的研究。如Hadoop大数据处理框架用于离线数据批处理，是I/O和网络带宽密集型应用，在进行资源调度时应考虑磁盘I/O和网络带宽的影响，尽可能将容器应用部署到数据存储节点上。Spark大数据处理往往用于处理内存密集型应用，在进行容器应用部署时应尽可能分散部署到内存空闲资源较多的节点上。单个计算框架容器应用在进行大数据处理场景下，其处理框架和处理任务的特性往往对资源需求不同。数据中心在使用计算框架容器应用进行大数据处理时，容器调度器应综合考虑各计算框架的特性和处理任务本身的资源需求特点，不能简单分解为单个不关联容器应用进行资源调度。

容器云中容器编排引擎的调度算法通常是一些常规的调度算法，用于解决单个容器应用的资源调度，并没有专门针对多种计算框架容器应用调度的研究，仅有的部分研究也是对特定的计算框架进行。面向多计算框架的容器云资源调度研究的缺失导致容器云对大数据的处理性能低下，资源利用率不高，任务执行时间过长。为了解决多计算框架容器应用对大数据处理问题，本文提出了新的调度方法，提升容器云资源利用率和负载均衡性，进而缩短任务处理时间。

## 论文主要工作和结构安排

### 论文的主要工作

基于Docker容器虚拟化技术的PasS层OpenShift容器云平台使用Kubernetes进行容器管理和调度，该调度方式通过预选和优选两阶段选取评分最优的节点作为容器调度的目标，在优选阶段仅考虑内存和CPU的影响因素。本文针对其调度方式造成的资源利用率较低和负载不均衡的缺点，在面向多计算框架的容器云平台Paladin上，设计和实现了一个新的调度方案，本文主要工作如下:

1. 深入研究Kubernetes调度核心技术，针对其调度算法不足，提出了一种基于多维资源空闲率权重的评价函数和调度方法MRWS(Multidimensional Resource Weights Scheduling)。该方法综合考虑物理节点CPU、内存、磁盘、网络带宽空闲率和已部署的容器应用数量等因素影响，使用模糊层次分析法FAHP(Fuzzy Analytic Hierarchy Process)对集群资源自动建模并求解容器应用多维资源权重参数，选取最大评分节点进行容器调度。
2. 针对MRWS调度策略，在容器云仿真平台ContainerCloudSim上进行大规模调度仿真，并与Random、FirstFit、Kubernetes默认调度策略在资源利用率和负载均衡性方面进行对比。
3. 基于开源OpenShift Origin和Paladin Storage构建面向多计算框架的容器云平台Paladin，该平台支持Hadoop、Spark、MPI、Storm等十多种分布式处理框架，用户可以快速构建大数据处理环境并实现容器伸缩。
4. 在Paladin上设计并实现MRWS调度策略，针对多计算框架下大数据处理场景，比较MRWS与其他调度策略下的集群性能。新的调度策略无论在单计算框架容器应用还是多计算框架容器应用混合执行效率都有极大的提升。

### 本文结构安排

本文总共分为六个章节，第一章主要阐述传统虚拟机技术构建的云计算和Docker容器技术构建的容器云的区别，云计算底层虚拟化技术和基本架构、云计算服务模式和云计算资源调度方法。接着介绍容器云底层的Docker虚拟化技术，一些典型的容器云产品，容器云中三大主流的容器编排器。针对计算框架容器应用进行大数据处理场景下容器云资源调度器的不足，引出本文需要研究解决的问题。 第二章主要综述容器云资源调度模型，比较分析容器云中典型的集中式调度、两层调度和共享状态三种调度模型，并以Docker Swarm、Mesos和Kubernetes为例阐述三种调度模型的架构和优缺点。最后针对Kubernetes共享状态调度模型容器编排管理器的组织架构、组件、任务处理流程、原理以及其调度的不足进行深入分析。 第三章提出一种全新的综合考虑容器应用和集群物理资源特点的调度方案MRWS，用于解决OpenShift容器云平台下Kubernetes共享状态调度器性能的不足。新方案主要包括调度流程优化、多维资源建模、空闲资源评分，反馈器的设计、权重参数的求解以及负载均衡度评价等。 第四章使用FAHP方法解决新的调度方案中多维资源权重的问题，对容器应用和集群资源进行数学建模、自动构建满足一致性要求的模糊成对比矩阵和判断矩阵，自动求解应用资源权重参数。 第五章首先在ContainerCloudSim容器云仿真平台上进行大规模调度仿真，对比分析MRWS和Kubernetes默认调度方法、FirstFit以及Random调度方法在集群资源利用率、负载均衡度方面的性能。然后在容器云平台Paladin上开发部署十几种分布式计算框架，设计和实现MRWS调度方法，在多计算框架混合部署场景下进行性能测试。最后一章对本文工作进行总结，展望未来的研究方向。

# 容器云资源调度模型综述

Docker容器虚拟化是容器云集群的核心技术，但一个高效的容器编排引擎也是集群不可或缺的部分。一个好的调度器既要提现出作业调度的“公平”性，又要兼顾其性能和鲁棒性，还要能应用于实际生产环境中。在集群数据中心，可以通过应用对资源的需求感知用户部署的应用类型，通常可以划分为CPU密集、内存密集、I/O密集和网络带宽密集型应用 。在集群上通常是多种密集型应用同时部署和运行，调度器如何进行资源分配至关重要，这使调度变得异常复杂和困难，往往不存在最优解决方案。传统虚拟机的云计算中心资源调度已有相当多的研究，针对容器集群，各大容器厂商也相继推出了几款优秀的容器编排引擎。根据其调度架构的不同可以划分为集中式调度模型、两层调度模型、共享状态调度模型，其典型代表分别是Docker Swarm、Apache Mesos、Google Kubernetes。

## 集中式调度模型

### 集中式调度模型概述

集中式调度通过单个调度节点处理所有任务请求，使用固定的、单一的调度算法调度全部任务。集中式调度模型中存在一个中心调度器，所有的任务都必须经过中心调度器中统一的调度算法进行调度，资源调度和任务管理都由中央调度器负责。首先，集中式调度模型将导致系统扩展性很差，集群规模不能大规模增加，而且所有的调度信息在单节点上运算，不能并行执行，这要求中央调度器的部署节点必须是高性能计算机，否则调度节点会成为性能瓶颈。其次，集中式调度很难定制用户自己的调度策略，针对特定场景下用户自定义的调度策略要融入中央调度器是一项巨大的挑战。针对集中式调度模型的不足，Google的Omega系统也进行了一定的优化，开发多种调度策略存放于中央调度器中，根据不同的调度任务选择相应的调度策略进行任务调度，这种优化在集群规模较小时能够缩短任务执行时间，起到一定的优化效果。在大规模的容器云集群上，调度策略依然存在于单个节点上，任务的精准分类和调度策略的频繁切换得不到有效解决，系统扩展性并没有本质的改善。此外，还有通过集群硬件资源静态分区来优化集中式调度模型的方法，将集中控制的资源静态划分为多个区域，计算框架分配到特定的区域执行，多个计算框架之间从物理上进行隔离。静态划分简单有效，但集群资源利用率低下，经常出现部分计算框架系统资源不足。

### Docker Swarm集中式调度模型

Docker公司2014年发布的容器编排管理工具Docker Swarm  ，无需安装、易于搭建、并且拥有活跃的社区，已应用于实际的生产环境中。

Swarm架构图

Swarm架构图

Swarm集群由管理节点和工作节点构成，其中管理节点可以配置多个，实现高可用的多管理模式，内部通过RAFT算法实现管理节点高可用性。管理节点上除Docker Daemon守护进程、Load Balancing、Scaling组件外，最为重要的是Scheduler和Discovery组件。Discovery负责集群中节点的发现和状态更新，Scheduler先根据用户的资源需求对节点进行筛选，然后使用内置的调度策略进行应用调度。工作节点主要运行Docker Daemon和Load Balancing，根据控制节点的指令运行绑定的容器应用。在调度器的过滤模块主要提供了约束过滤器和健康过滤器，此外还可以配置吸引力过滤器、依赖过滤器和端口过滤器。

约束过滤器是通过一定的约束条件筛选节点，集群中每个节点都带有一个标签Label，对于一些特殊的应用可以指定其Label标签调度到指定的节点上运行；健康过滤器过滤掉不健康的节点，避免容器调度后运行失败；吸引力过滤器是将新的容器链接到已经创建的容器上，实现共同运行和销毁，主要有镜像吸引和标签吸引两种方式。镜像吸引是将容器直接调度到拥有该镜像的节点上，避免镜像下载的重复开销，节约网络资源，标签吸引是通过标签指定链接到已创建的旧容器实现共同工作；依赖过滤器是新容器依赖于其他的容器，可能会共享磁盘卷、或在同一个网络栈上等；端口过滤器将需要特定开放端口的容器运行到开放该端口的节点上，避免容器端口不可用的情况发生 。

Swarm的调度策略的算法主要包括Random、Spread和Binpack三种 ，下面分别对其算法思想进行简单的介绍:

1. Random算法。该算法随机从过滤完的节点中选取一个节点进行调度判断，如果该节点满足条件则将容器调度到该节点上，否则随机选取下一个节点直至找到合适的节点调度或返回调度错误信息。
2. Spread算法是最少容器算法，该算法的初衷是保证容器集群的负载均衡。每次遍历一遍集群中每个节点上运行的容器数量，选择容器数量最少的节点进行容器调度，若该节点不满足条件，则依次从后往前进行调度尝试，直至找到满足条件的节点或返回调度错误。
3. Binpack算法是最多容器算法，该算法的目的在于最大化利用集群中节点资源，和Spread算法相反，每次从集群中选择运行容器数量最多的节点进行容器调度，若满足条件则将容器调度至该节点，否则依次从前往后尝试调度容器到节点，直至找到合适节点或返回错误。

## 两层调度模型

### 两层调度模型概述

两层调度模型是将资源调度和作业调度分开，资源调度层只负责给计算框架分配所需的资源，具体的作业调度由每个计算框架的任务调度器完成。在一些成熟的处理框架中如Hadoop、Spark、MPI等都有相对成熟和高效的调度算法，两层调度模型将这些调度算法集成进来，通过一个轻量的资源共方式来控制资源的分配和访问。如在两层调度模型上执行Hadoop处理框架，资源调度层只负责将系统资源分配给Hadoop调度框架，具体的任务调度由Hadoop自身的调度器执行。两层调度模型分而治之的设计思想极大提升了集群的可扩展性和任务调度的并行执行能力，通过中间的资源管理器为上层任务框架提供系统资源，并且掌握底层资源的全部状态。两层调度模型也存在不小的缺点，首先，上层的单个计算框架无法获取集群的全部状态，只有下层的资源调度器才掌握全部集群资源信息，这就导致上层计算框架在任务运行失败时无法做出重启还是重调度的抉择。其次，两层调度采用“悲观锁”对资源进行控制，相当于资源的一把全局锁，所有资源被推送给某个计算框架，待该计算框架系统资源分配完成后，其他计算框架才能进行资源分配，这造成系统响应慢，并发粒度小，并发性受到极大的限制。

### Apache Mesos两层调度模型

Apache Mesos 是两层调度模型中最为典型的代表，最初是加州伯克利分校的AMPLab在2009年开发的集群管理软件，用于管理快速扩张的各种分布式计算框架，后在Twitter中用于容器云集群管理，获得巨大的成功。

Apache Mesos架构图

Apache Mesos架构图

如图2.2所示，Mesos的总体架构也采用主从设计，Master节点作为集群的管理控制节点，可以有一个或者多个(最好为单数)，为了防止单节点故障，通过ZooKeeper提供一致性服务。每次在多个Master中选举一个作为leader对外提供服务，其他的Master副本随时保证和Master的状态一致，一旦服务出现故障，立即启用新的备份节点，从而实现集群的高可用性。集群的架构可以分为两层: 计算框架层和Master调度层。Master调度层管理众多的计算节点，负责收集各节点的资源情况并作出资源分配决定，计算框架层负责实际任务的调度。整个调度系统分为任务调度层和资源调度层，两个层级互不干扰，分别使用自身的调度算法，提升了集群调度策略的可扩展性。从架构的组成部分来看，主要有五大组成部分，下面分别对组成部分的功能进行介绍:

1. ZooKeeper组件。ZooKeeper作为一款Hadoop项目中分布式系统的协调系统，主要用于解决分布式应用数据管理问题，如应用配置管理、统一命名服务、状态同步服务以及组服务和集群管理等。ZooKeeper非常简单易用、功能丰富可靠、并且提供了通用协议下开源共享的存储库，其核心就是一个精简的文件系统，还能提供一些简单和抽象的文件操作。在Mesos中，ZooKeeper主要用于解决Master节点的状态一致性和高可用问题，避免单节点故障，实现集群持续稳定的对外服务。
2. Mesos Master组件。Master是集群的控制器，是整个集群调度的大脑中枢，既要对底层各Slave节点的资源进行收集和管理，也要通过一定的资源分配策略提供资源给上层的各处理框架Framework。当前对各Framework的资源分配策略使用DRF (Dominant Resource Fairness)算法，这是一种针对多维资源(CPU、内存、I/O、网络带宽等)不同需求设计的公平调度算法。Master还负责资源的访问控制，一旦系统资源被推送给某个Framework，必须等该计算框架资源分配完成后才能再次进行其他框架的资源分配。
3. Mesos Slave组件。该组件作为底层调度实例的具体执行者，接收来自Master的指令，将自身的资源分配给每个执行器，执行器上运行一个或多个任务，并将各任务作为容器运行起来。此外，Slave节点还定期向Master节点汇报资源使用情况作为Master调度器的调度依据。Slave上运行一个containerizer用来管理容器的生命周期，包括容器的创建、更新、监控和销毁。
4. Framework组件。Framework负责将各计算框架如Hadoop、Spark、MPI等注册接入到集群中，Master的调度器负责对其需要的资源进行分配，具体的任务调度则由各计算框架完成。各计算框架通过调用Master的API进行任务创建和调度请求，Master再将任务下发到Slave上执行。
5. Executor组件。Executor负责启动框架内部的Task任务，各种计算框架接入Mesos的方式，接口不同，因此要接入一个新的计算框架就需要编写一个新的Executor，用来通知Mesos如何启动框架中的Task任务。

Mesos作为一款优秀的分布式资源管理框架，采用双层调度机制，资源分配层负责将资源分配给计算框架，计算框架使用自身的任务调度器执行任务调度。Mesos可以对集群的资源进行细粒度的划分，按照计算框架实际任务的需求进行资源分配，极大提升了集群资源利用率。Mesos不需要清楚各Framework的具体调度逻辑，只需要通过API向上提供资源分配即可，具有较强可扩展性。Mesos是模块化的实现，新增一个Framework不需要对Mesos进行重新编码，可以快速实现接入。Master节点使用ZooKeeper保证其状态一致性，能够实现高可用性。但是，Mesos对底层的资源采用“悲观锁”的方式进行控制，相当于资源的一把全局锁，所有资源被推送给某个计算框架，待该计算框架系统资源分配完成后，其他计算框架才能进行资源分配，这造成系统响应慢，并发粒度小，并发性受到极大的限制。独立的调度框架只能访问集群部分状态信息，往往不能进行调度优化。

## 共享状态调度模型概述

在集中式模型中，资源分配和任务管理都由中心调度器负责，并且集成了具体的调度算法，难以扩展调度策略。在两层调度模型中，资源分配由资源调度层完成，任务调度由具体的计算框架自己完成。集中式的调度很好的保证了全局状态的一致性，但是扩展性较差，两层调度虽然可扩展性较好，但并发粒度小、并行性差，并且容易造成资源的竞争和死锁。为了解决这些不足，共享状态调度模型被提了出来，其核心在于所有的调度逻辑共享集群状态，选择最优的节点进行资源分配和任务调度。 其中最为典型就是Google推出的Borg 、Kubernetes以及使用事务方式解决一致性管理问题的Omega ，其中Kubernetes以其轻量开源的特点深受大家喜爱，各大主流的互联公司都加大了对其支持力度。

共享状态资源调度模型

共享状态资源调度模型

在共享状态调度模型中，每一个上层的调度器都可以对整个集群的状态进行访问，资源对所有的调度器都是透明的，可以进行自由竞争，不存在单一的资源分配器，也不存在单一节点访问瓶颈。集群不仅支持节点资源的快速增加，也支持调度器的扩展，用户根据自己的实际应用场景开发特殊的调度器，并且很轻易地集成到系统中。为保证集群的高并发性和可扩展性，共享状态调度模型采用“乐观锁” 对底层资源进行并发控制。具体的实现是通过对集群的所有状态都增加一个版本属性，每次提交时比较提交的版本和当前数据版本号大小，若版本号小于当前的版本号，则不进行任何处理，只有提交版本号大于当前数据版本号的操作才被接受，更新完状态后版本号递增。“乐观锁”并发控制虽然会增加资源访问的冲突数，影响系统的吞吐率，但在实际的系统中依然在一个可以接受范围，下一小节将对Kubernetes调度模型架构和流程做详细的分析。

## Kubernetes共享状态调度模型

Kubernetes是典型的基于共享状态的调度模型，是一个轻量开源的平台，用于容器应用管理和服务，使用Label和Pod的概念将容器划分为逻辑单元，实现关联容器共同部署和调度。本小节对Kubernetes核心组件和整体架构进行深入研究，尤其是分析其调度算法的不足，为下一步提出新的调度方案提供依据。

### Kubernetes简介

Kubernetes源自Google的Borg项目，Borg是Google集群管理工具，一直用于管理全球上百万台服务器，性能稳定可靠。为在容器云竞争中占据主导地位，Google基于Borg的管理经验，研发了Docker的容器编排工具Kubernetes，并在2014年将其开源，逐步发展成为一个大的生态技术圈。作为一个跨主机的应用容器编排引擎，Kubernetes提供了一系列强大的功能，包括应用容器部署、资源调度、服务发现、动态扩容以及错误恢复等。Kubernetes强大的集群管理能力非常适用于分布式系统，实现了多租户应用、服务发现、服务注册、负载均衡、故障发现和自动恢复、在线扩容、细粒度调度、资源配额管理等功能，完美定义了构建业务系统的标准化架构。除集群管理方面强大功能外，Kubernetes还提供完整的开发、测试、部署、发布、运维监控等各种开发管理工具。

Kubernetes逐步发展成一个巨大生态圈，为容器编排提供一种简单、轻量的方式，用户还可以进行功能定制。采用Kubernetes的云计算服务商和用户越来越多，如微软、Yahoo、IBM、华为、VMware、网易、阿里、百度等，还有一些初创公司灵雀云、青云等都采用Kubernetes作为容器云的管理系统。

Kubernetes拥有强大而活跃的社区，众多开发者不断对其进行迭代更新，其代码更加完善。当前Kubernetes社区的支持者有Google、CoreOS、RedHat、华为、网易、阿里云、浙大SEL实验等。Google在2015年联合其他20多家公司成立了开源组织CNCF(Cloud Native Computing Foundation)，加入OpenStack社区，致力于Kubernetes的应用推广，支持其在公有云、私有云等更多基础设施平台上应用，提供更加简便丰富的工具集，方便用户开发和使用。

### Kubernetes架构和组件

Kubernetes是一个主从架构体系，该集群管理器很好地解决了扩容和服务升级两大难题，具有较强的横向扩展能力，Kubernetes的整体架构图 如下。

Kubernetes体系架构图

Kubernetes体系架构图

如图2.4所示，Kubernetes主从架构主要由控制节点Master、工作节点Node以及外部工具集kubectl、web UI等附加依赖组成。Master作为集群的控制节点，主要负责集群的管理调度，由API Server、Scheduler、Etcd、Controller Manager等组成，实现其主要功能。Node节点主要根据控制节点的指令执行具体的任务，实现应用容器的运行，主要由kubelet、kube-proxy、cAdvisor、Container Runtime等组成。外部可以通过kubectl命令工具对集群进行增删查改的操作，也可以使用Web UI与集群进行交互。下面对集群中的重要组件分别进行介绍:

1. API Server组件。API Server是系统管理指令的统一入口，负责对外提供RESTful的API服务功能，所有与集群的交互都需要通过API Server组件完成，是集群外部和内部的通信枢纽。APIServer也是资源配额限制的入口，通过Authentication组件提供集群完备的安全认证机制。API Server接收外部Web browser或kubectl的命令请求，将REST对象和状态持久化到Etcd中存储，同时与Node节点上的kubelet进行交互。
2. Scheduler组件。集群调度组件，负责集群资源调度和Pod分配工作，Scheduler监控集群中未分配的Pod，根据其对资源的约束条件和集群资源可用性，将Pod调度到实际的Node上运行。调度器是一个可插拔的模块，用户可以开发自己的调度算法集成到集群中，其调度流程和调度策略后面会详细介绍。
3. Controller Manager组件。控制管理器提供服务发现、集群管理、Pod扩容、服务绑定、应用生命周期管理等功能。其中Node Controller用于管理节点、Replication Controller用于应用容器管理，保证容器副本和需求一致、Namespace Controller用于命名空间管理、Service Controller提供负载和服务代理、Persistent Controller管理维护Persistent Volume和Persistent Volume Claim等。
4. Etcd组件。一个高可用、强一致性的服务发现键值存储仓库，用于保存集群中所有的网络配置和对象的状态信息，具有中心数据库的地位，进行分布式的部署，通过watch机制进行服务更新支持。
5. Kubelet组件。Kubelet是运行在Node节点上的控制器，用于裁决和驱动容器的执行层，是API Server和Pod的主要实现者。单个Pod中可以运行多个容器和存储数据卷，能够将Pod和相关的依赖项很方便地打包迁移， API Server进行访问控制，Scheduler进行资源的调度，但是最终Pod能否在Node上运行成功是由kubelet决定的。此外，kubelet通过cAdvisor组件对Node节点的状态、资源进行监控，定期汇报给控制节点，存储在Etcd中。
6. Kube-Proxy组件。负责负载均衡和反向代理组件，通过创建Pod的代理服务，用户可以通过IP地址直接访问Pod应用，实现服务到Pod的路由转发。此外，kube-proxy还实现了一个高可用的负载均衡解决方案。

除上述给出的核心组件外，Kubernetes还有负责提供集群DNS服务的kube-dns、提供外网访问入口的Ingress Controller、提供资源监控的Heapster、提供管理控制界面的Dashboard、提供日志采集、存储和查询的Fluentd-elasticsearch组件等。

### Kubernetes任务处理流程

Kubernetes调度器运行在Master节点上，作为一个可插拔的模块，在默认配置下，调度器可以满足大部分的需求，如特定的Pod分配到指定的节点，相同集合下的Pod分配到不同节点，平衡各节点的资源使用率等。但在一些特殊的应用场景下，尤其涉及多计算框架容器应用对大数据的处理，其默认配置的调度方法显得力不从心，需要用户开发自己的调度策略。

kubernetes任务处理流程

kubernetes任务处理流程

调度器相对于普通用户而言类似一个黑盒，输入Pod的资源需求，输出Pod和节点的绑定策略，起到指挥中枢的作用。但是在很多业务场景下，用户希望自己的Pod调度可控，如定制调度算法、特殊硬件需求的Pod调度到相应的节点、通信量大的计算框架部署在相同区域和机架、数据需求大的Pod部署在数据节点等。Kubernetes的任务处理流程如图2.5所示，命令工具Kubectl或Web Browser向Master上的API Server发送任务请求如创建Pod，API Server对请求做出响应处理并将处理的结果和状态存储在Etcd中，同时设置PodSpec.NodeName为空，加入未调度Pod队列。调度器监控Etcd中未调度Pod队列状态，发现有未调度的Pod时通过调度策略尝试绑定Pod 到满足需求的节点。调度策略将满足资源需求的节点进行综合评分，主要考虑资源使用的均衡性、相同副本的容灾性等因素，最终选取得分最高的节点，将Pod调度到该节点上，并将绑定状态存储到Etcd中。节点上的Kubelet监控Etcd中Pod调度结果，接管任务的后续工作，负责Pod生命周期的管理，一个完整的任务处理流程结束。

## 本章小结

本章主要介绍基于Docker构建的三种典型容器调度模型: 集中式调度模型、两层调度模型和共享状态调度模型，并以Docker Swarm、Apache Mesos和Kubernetes为例对比分析三种调度模型的组织架构和调度优缺点，最后重点分析了容器编排引擎Kubernetes共享状态调度模型的任务处理流程和调度原理，为提出新的共享状态调度方法奠定基础。

# 面向多计算框架的容器云MRWS调度方法设计

Kubernetes作为OpenShift Origin容器云平台的容器编排引擎，其默认配置的调度方案能够满足大部分用户需求，但其调度算法较为简单，集群资源利用率不高，负载均衡性较差，无法满足用户的在特定场景下的调度需求。本章在深入分析默认调度算法不足后，提出了一种新的基于多维空闲资源权重参数的调度方法MRWS (Multidimensional Resource Weights Scheduling)。该方法主要包含MRWS调度流程优化和MRWS调度算法设计，在调度流程优化部分对Kubernetes共享状态调度模型的调度流程进行优化，缩短容器应用调度时间，节约系统资源。在调度算法设计部分，对容器应用资源需求和集群资源建模，提出一种综合考虑容器资源需求、集群资源空闲率、资源重要性和已部署Pod数量等因素的调度策略。MRWS调度方法既能满足普通用户的调度需求，也能极大提升多计算框架容器应用进行大数据处理场景下容器云集群资源利用率和负载均衡性。

## OpenShift 容器云平台

Docker是当前主流容器化技术的代表，Kubernetes作为现阶段应用最为广泛的容器编排引擎，OpenShift 将这两种主流技术相结合服务于企业，是红帽公司提供的一款开源容器云平台。 该平台底层以Docker作为容器引擎驱动，Kubernetes作为容器编排组件，对外提供多种开发语言、中间件、数据库以及极易操作的用户界面、DevOps(Development and Operations)工具等。允许开发者和开发团队在该平台上进行应用的构建、测试、部署以及发布，是一个完善的PaaS层云平台。该平台可以运行和支持有状态和无状态的应用，为容器应用提供较强的安全防护，包括基于用户的访问控制、检查机制以及强制隔离措施。OpenShift平台还实现了多种综合云原生服务，便于快速、灵活地开发应用、构建各种分布式系统，平台支持多种云环境包括Amazon Web Service、Azure、Google云平台以及VMware等，为开发运维团队提供一个通用的环境和工具集，用于持续开发和测试。OpenShift分为开源的社区版OpenShift Origin和收费的企业版OpenShift Enterprise，面向多计算框架的容器云平台Paladin的服务层基于OpenShift Origin研发。

从技术堆栈的角度分析，OpenShift Origin容器云平台自下而上可以划分为基层架构层、容器引擎层、容器编排层、PaaS服务层、界面及工具层 。下面分别对这几个层次功能进行简单介绍：

1. 基础架构层。提供OpenShift运行所需的基础设施和环境，包括物理机、云主机、虚拟机、各种公有云、私有云以及混合云等。OpenShift支持多种操作系统，如CentOS 7.0以上版本、Fedora 21、Red Hat Enterprise Linux等，包括在Linux基础上专门定制和优化容器运行环境的操作系统Atomic Host，该系统可以为应用提供高度一致的运行环境，保证集群的稳定和安全。容器应用虽然具有较强的跨平台性，但容器的隔离依赖于底层操作系统Linux的内核，要求底层系统的内核和配置必须一致。
2. 容器引擎层。以当前主流的Docker作为OpenShift平台容器引擎，Docker已广泛应用于各社区和企业生产环境中，经过了安全、稳定和高可用的检验。OpenShift并未修改任何原生的Docker代码，只是将Docker的开放性和庞大的镜像资源无缝衔接到平台上，Docker的普通用户可以快速整合到平台中，所有的应用最终到底层都生成一个Docker实例。
3. 容器编排层。容器编排系统是影响容器云性能和资源利用率的关键因素，OpenShift最终选择轻量开源的Kubernetes作为其容器编排引擎。Kubernetes已在Google内部使用多年，其诞生初衷就是为解决大规模集群中容器的调度和管理问题。OpenShift平台中很多基本的概念和重要组件如Namespace、Pod、Replication Controller等都继承自Kubernetes。 OpenShift同样只是将Kubernetes进行叠加式的整合，并未修改其原生代码和对象，用户依然可以通过原生的命令操作Kubernetes的组件和对象。
4. PaaS服务层。容器云的目标在于对外提供服务，OpenShift在PaaS服务层提供了多种开发语言、框架、数据库以及中间件，极大提升了上层应用的开发、测试和交付速度。OpenShift有一个专门的社区和Docker Hub提供各种应用的镜像，用户可以快速获取一个应用的基本镜像，构建自己所需的容器应用，Red Hat的JBoss中间件几乎全部实现了容器化。
5. 界面及工具层。OpenShift平台强大的界面及工具集极大帮助普通用户高效完成相关应用业务，用户可以通过Web进行服务交互，平台将自动从Docker Hub中拉取所需的镜像进行应用环境构建，全自动化的服务极大降低了运维成本和提升服务效率。此外，OpenShift平台还提供S2I(Source to Image)服务，用户开发完成后可以提交整合到镜像中，实现快速的应用交付，提升开发、测试、部署效率。针对用户端接入问题，OpenShift平台提供Web-console控制台、IDE集成工具、命令行工具集、以及RESTful API编程接口，是一个完善的企业级PaaS层容器云平台。

OpenShift Origin架构图

OpenShift Origin架构图

如图3.1所示，OpenShift平台的核心组件包括Master、Node、Pod、Scheduler、Service、Storage等。Mater是主控节点，可以配置高可用的多个主控节点，负责管理和维护OpenShift集群状态。Master上运行的API Server是其核心组件，所有用户的Web Console以及RESTful API服务都通过该组件进行访问认证控制，各Node节点也会定期轮询API Server更新节点和容器的状态。Data Store将所有的状态信息存储在分布式的数据库Etcd中，并通过Raft一致性协议保证其数据的一致性。Scheduler调度控制器进行Pod资源的分配和调度，根据各节点收集的资源情况，选择合适的节点作为容器应用的调度目标。Replication Controller是异常自检测和恢复组件，负责监控集群中容器应用的状态和数量是否和用户要求一致，自动启动和关闭容器应用满足用户的需求。Node节点通过接收Master节点指令维护容器应用。Pod是OpenShift平台调度器调度的最小单元，一些容器与容器之间往往存在较大的关联性，将几个联系紧密的容器部署在一个Pod中进行调度是OpenShift平台的主要创新工作，如分布式数据库多个容器放在一个Pod。容器是一个非持久化的对象，一旦容器重启或销毁，其状态信息将会随之销毁，集群每次给Pod分配的IP地址不同。要对外提供统一持久的服务，需要Service组件，该组件能将所有的信息转发到其对应的容器IP和端口上。此外，还有Router、Persistent Storage、Registry、Haproxy、Kubelet等都是集群的重要组成组件。数据可以持久化存储在集群数据库、分布式数据库或者其他数据库中，当前支持NFS、Ceph RDB、GlusterFS等。

## MRWS调度流程优化设计

Kubernetes调度流程分为预选和优选两阶段，在预选阶段过滤掉不满足需求的节点，优选阶段对剩余节点评分，选择评分最高的节点作为调度目标。整个调度器可以分为待调度的Pod列表、满足条件的Node列表以及调度策略三部分。MRWS调度方法将对其调度流程进行一定的优化，增加调度“记忆”模块和信息存储模块，缩短任务调度时间和节约系统资源。

### Kubernetes调度流程

在Kubernetes默认调度流程中，预选阶段解决节点过滤问题，优选阶段解决节点选择问题。用户可以对两个阶段进行简单的配置，调度器将根据用户指定的预选和优选规则进行过滤和评分计算，最终输出满足条件的Node和Pod绑定策略，将Pod调度到合适的Node，针对两阶段的规则，下面进行详细的介绍。

Kubernetes调度器流程

Kubernetes调度器流程

Predicates阶段过滤掉不满足Pod资源需求的节点，依据一定规则进行筛选，包括磁盘卷冲突、端口冲突、资源容量、节点检测、服务占用、亲和性等。1.7版本的筛选条件主要如下，新的版本Kubernetes还在不断更新和完善。

1. NoDiskConflict：检测磁盘卷冲突，在当前的规则中，两个不同Pod不能使用相同的卷。因此，一旦Node上挂载了该卷并被某个Pod使用，则新的Pod不能再调度到该Node上，否则造成卷冲突。不同的系统对该冲突检测范围不同，如Google Compute Engine在只读模式下允许共用卷，Ceph RDB允许两个Pod分享部分资源，Amazon EBS禁止两个Pod挂载同一个卷。
2. NoVolumeZoneConflict：在给定Zone限制的条件下，检测Node上部署的Pod是否存在卷冲突，当前只限定对PV(Persistant Volume)范围进行支持。
3. PodFitsHostPorts：端口冲突和服务占用检测，待调度Pod内所有的容器需要的端口在Node上是否被其他Pod占用。
4. PodFitsResources：根据Pod资源的需求检测节点空闲资源量是否满足其运行需求，主要检测CPU、内存、磁盘等资源。Kubernetes的调度是静态调度，资源的判断是根据分配的资源量而不是资源的实际使用量。
5. HostName：检测Pod是否指定了Node节点，所有不在指定Node集合内的节点都将被过滤掉。
6. MatchNodeSelector：检测Pod是否指定了MatchNodeSelector属性，若指定了该属性，Node的Label必须和该属性匹配。
7. MaxEBSVolumeCount：确保挂载的EBS存储卷总和不超过设置最大值，调度器计算每个Node上直接或间接使用的全部卷总和，一旦超过最大值，Pod不能调度到该节点上。
8. CheckNodeMemoryPressure：判断节点是否存在内存压力，若存在内存压力则标记为1，Pod只能调度到内存标记为0的节点上。
9. CheckNodeDiskPressure：判断节点是否存在磁盘压力，若存在磁盘压力则标记为1，Pod只能调度到磁盘标记为0节点。
10. MatchInterPodAffinity：节点的亲和性过滤，检测Pod是否和已部署的Pod存在亲和性，将有亲和性的Pod调度到相同节点。
11. PodToleratesNodeTaints：判断将Pod调度到节点后是否满足节点容忍的条件，相同的Pod副本为了满足容灾性，一般部署到不同的Node甚至是不同的机架和数据中心。

此外、还有专为Google Compute Engine和Amazon EBS配置的MaxGCEPDVolumeCount、MaxAzureDiskVolumeCount卷检测条件，检测节点上挂载的卷容量总和是否超过了设定的最大值。

Priorities阶段根据一定的评分算法对预选出来的节点列表进行综合评分，评分主要依据节点资源空闲量、资源消耗平衡性、Pod亲和性、Pod与节点的匹配度等因素。Kubernetes使用优先函数集合内的函数对节点进行0-10评分，最终计算评分总和，评分越高表明Pod调度到该节点越适合，同时还可以为集合内的每一个函数设置一个权重值作为函数系数，主要的评分函数如下:

1. LeastRequestedPriority：节点资源空闲率评分函数，即根据节点的空闲资源与节点资源总量的比值((节点资源总量-节点上Pod的资源和-新Pod的需求)/总容量)来计算。设置CPU和内存具有相同权重系数，都设置为0.5，资源空闲率越大，剩余资源越多，节点的评分就越高。
2. BalancedResourceAllocation：该函数必须联合LeastRequestedPriority函数一起使用，用于平衡各维度资源的使用率，资源使用越均衡，节点的评分越高。Kubernetes主要对内存和CPU资源消耗进行平衡，由两者的“距离”决定分值大小，即使用10-abs(CPU空闲率-内存空闲率)\*10进行计算。
3. SelectorSpreadPriority：为了更好应对容灾和宕机风险，同属Service、Replication Controller下的Pod尽可能分散部署到不同的Node上。对于指定Zone的Pod，也要尽量分散到不同区域的主机。调度器统计Node上相同Service、Replication Controller下的Pod，Node上Pod数量越少，评分越高。
4. NodeAffinityPriority：设置调度器的亲和性机制，主要对节点进行精确匹配。有两种选择器匹配模式，选择器"hard"模式下设置的NodeSelector必须和节点的Label匹配，保证选择的Node是完全满足需求规则的，"soft"模式下不能完全保证百分百的匹配，但会尽量满足规则要求。
5. ImageLocalityPriority：为避免镜像的重复下载对网络和磁盘资源的重复消耗，该函数根据Pod中运行容器需要的镜像对节点进行评分，节点上存在的镜像越多，该节点评分越高。若该节点上没有所需的镜像，评分为0，镜像评分的权重可以根据镜像大小按比例决定。
6. MostRequestedPriority：和LeastRequestedPriority相反作用，两者使用一个来计算评分。该函数确保使用资源越多的节点，评分越高，目的在于使用更少的服务器提供服务，节约服务器资源，提升集群资源利用率。
7. TaintTolerationPriority：容忍性评分函数，匹配Pod的TolerationList与节点Taint，匹配项越多，该节点的容忍性越好，从而评分越高。
8. InterPodAffinityPriority：用于迭代WeightedPodAffinityTerm的元素计算和，若该节点同时满足亲和性设置，则将该评分加入节点的整体评分中。
9. NodePreferAvoidPodsPriority：根据节点是否设置Anotation属性进行评分，没有设置该属性则评分为10，设置该属性并且待调度的Pod正好是副本，则该节点评分为0，目的在于避免相同副本调度到同一节点。

Priorities优选阶段定义了一个评分函数集合，各评分函数的权重可以指定，用户根据调度依据的重要程度给各函数赋予不同的权重值，最终所有函数的评分总和就是该节点的最终得分，选择评分最高的节点作为Pod的调度目标。

### MRWS调度流程优化

Kubernetes调度器的Default算法核心在于优选阶段的评分函数，调度策略根据节点评分高低做出调度决策。在所有的评分函数中，除一些特殊的调度规则外，如节点亲和性、节点容忍度、节点上的镜像文件等，最为重要的是根据节点资源空闲率进行评分的LeastRequestedPriority和节点资源消耗均衡性进行评分的BalancedResourceAllocation函数。这两个函数是整个评分函数集合中的核心，在其他外在规则相同的条件下，直接决定节点的评分高低。但是，这两个评分函数仅考虑了内存和CPU的空闲率和消耗平衡性，这种评分会造成节点其他维度资源利用不均，如磁盘、网络带宽、节点上运行的Pod数量等。一旦某个维度的资源过载，节点将不能部署更多的Pod，造成集群资源利用率低下。因此，新的调度方法MRWS将综合考虑集群CPU、内存、磁盘、网络带宽以及节点已部署Pod数量等因素的均衡性，提升集群的负载均衡和资源利用率。

MRWS调度流程优化

MRWS调度流程优化

在Default算法中，优选阶段根据评分集合进行节点评分，其评分函数如下：

其中，是第i个节点的总评分，是第j个评分函数权重，是第j个评分函数，每个函数的权重可以根据用户的需求进行指定。

如图3.3所示，新的调度方法MRWS首先对调度流程进行优化，预选阶段从集群上的*P*个节点中根据预选规则筛选出*M*个节点，这些节点满足Pod调度的资源需求和预先设定的过滤条件。*M*个节点列表和Pod的内存、CPU、磁盘、带宽需求作为评分阶段的输入，查找Hash表中该Pod以前是否被创建和调度过，若该Pod之前被调度过并且以前调度的节点同时在预选阶段剩余的节点中，将新Pod直接绑定到节点。通过Hash表的作用，可以避免重复的Pod被调度到不同的节点，导致容器镜像被重复下载，节约磁盘空间和网络资源，缩短Pod应用的调度算法计算时间，使集群具有容器调度记忆功能。若Pod是一个新建的Pod，进入函数评分模块，信息存储模块通过节点上的cAdvisor资源监控代理收集节点资源信息，同时记录各节点上已分配的Pod资源总量和已部署Pod数量，权重参数库存储容器应用资源权重系数。最终通过全新的评分函数体系进行节点评分，该评分函数核心在于对节点资源建模，通过模糊层次分析法FAHP自动求解多维权重参数，使集群中各节点各维度资源利用更加均衡。

## MRWS调度算法设计

Kubernetes调度器评分阶段的核心算法仅考虑内存和CPU空闲率以及消耗平衡性，容易造成其他维度资源过载，从而使集群资源利用率低下。MRWS方法将综合考虑集群中节点的CPU、内存、磁盘、网络带宽以及节点上已部署Pod数量等因素，致力于多维资源的均衡利用，尤其是在多计算框架容器应用同时进行大数据处理的场景下，让各计算框架的应用调度更加合理，缩短任务的执行时间。整个MRWS方法算法部分包括资源空闲率评分模块，平衡评分模块和算法均衡度评价模块三部分，下面分别对其建模和评分计算方法进行详细介绍。

### MRWS空闲资源评分

为了达到集群多维资源利用均衡，提升集群服务性能的目的，某个节点空闲资源越多并且各维度消耗越平均，该节点的评分就应该越高，就越适合新创建Pod容器的部署。针对大规模的集群，每个节点资源消耗均衡性不能完全依赖人为判断，并且评分函数的权重参数也不能全部由人主观给出，这种方式既不科学也不准确。用户应用场景和对资源重要性判断的差异性，将导致相同应用在同一调度算法下的不同调度结果。为了避免此种情况发生，需要对预选阶段剩余的节点建模，使用数学方法对资源的权重参数自动求解。在构建评分算法前，先对集群节点和Pod资源需求进行建模(节点均指预选阶段过滤后的节点)。

1. 定义集群节点和资源的符号表示。假设*P*个节点经过预选阶段筛选后剩余*M*个节点，表示为。单个节点的资源维度为*D*，在该模型中仅考虑节点CPU、内存、磁盘和网络带宽四种资源因素，因此*D*=4。集群中单个节点的资源总量表示为，其中表示第i个节点拥有的第d维资源的总量。同理，可以定义节点各维度资源的使用量，表示为，其中表示第i个节点上第d维资源的使用量。当前需要调度的Pod对资源的需求可以表示为，表示Pod对d维度资源的需求量。
2. 已部署Pod空闲率计算。节点上已部署Pod应用数量可以记录和统计，但将其作为影响调度策略的因素时没法加以限制和度量，也不能指定节点部署Pod总量。针对已部署Pod这个影响调度的因素，需要单独进行处理，使用一定的方式计算其负载和资源空闲率。考虑该因素的目的在于使每个节点部署的Pod数量尽量接近，若某个节点部署少量大资源需求的Pod，另一个节点部署大量小资源需求的Pod，虽然两个节点资源使用率相近，但是大量小资源数量Pod的管理会极大增加节点开销，影响该节点的服务质量。因此，应尽量使不同节点上运行的Pod数量均衡。采用如下的方式计算节点已部署Pod的负载，集群上剩余节点已调度的Pod总数C表示为：
3. 其中，表示第i个节点上已部署的Pod数量，可以用表示第i个节点已部署Pod的资源空闲率，如式(3-3)所示。
4. 多维资源的权重系数。在评分函数模块，给评分函数集合中的每个评分函数人为赋予一个权重，最终按照总得分排名选取最高。在实际数据中心，根据调度Pod请求的资源数量可以感知部署的应用类型，如CPU密集、内存密集、I/O密集以及网络密集型等。从经济效益角度考虑，节点上各种资源重要程度也具有较大差异，CPU和内存较为稀缺，价格相对昂贵一点。定义权重系数分别用于表示节点CPU、内存、磁盘I/O、网络带宽和已部署Pod数量等因素的重要程度，且。权重参数将在下一章通过FAHP进行节点资源数学建模并实现自动求解，用户通过微调权重参数更加灵活使用新调度算法。

根据上述定义的集群符号，可以根据各节点资源空闲率对节点进行评分，资源空闲率模块反映节点空闲资源的数量，空闲资源越多，节点评分越高。在考虑资源空闲量时又充分考虑各维度资源的重要程度，越是宝贵的资源其重要程度越高。节点i空闲模块的最终评分如式(3-4)所示。

其中，表示节点i部署资源需求为的Pod后第d维资源的空闲率。

### MRWS平衡模块评分

在空闲资源模块评分中，各维度的资源根据其空闲量和资源的重要程度决定评分，这会导致各维度资源利用不均衡。比如某个节点CPU资源空闲率很低，但其他维度空闲率很高，尽管赋予了CPU较高影响因子，但其整体评分依然偏高，会继续调度CPU密集的Pod到该节点，造成节点性能下降，资源消耗更加不均衡。为避免上述情况的发生，需要设计一个平衡模块，用于平衡各维度资源消耗。平衡模块用于反映节点各维度资源的均衡状况，各维度资源消耗越均衡，其评分越高。计算节点各维度资源空闲率的均值如式(3-5)、(3-6)所示。

其中，表示经预选阶段筛选后剩余m个节点上第d维资源空闲率的平均值，表示剩余节点上已部署Pod资源空闲率的平均值。因此，可以使用来度量节点i上第d维资源空闲率在集群中的不均衡性。同理， 表示节点已部署Pod数量的不均衡性。该值越大，表明各维度资源利用越均衡，节点评分越高。平衡模块节点的最终评分如式(3-8)所示。

整个评分算法综合考虑CPU、内存、磁盘、网络带宽以及已部署Pod等因素，分值由资源空闲评分和资源均衡评分组成。资源空闲率反映节点可用资源的状况，资源均衡反映节点各维度资源消耗的均衡性，单个节点最终评分如式(3-9)所示。

### MRWS均衡度评价模块

设计完调度算法后需要对调度算法多维资源利用均衡性进行度量，由于资源重要程度不同，不能简单用空闲资源利用率总合进行加权平均计算。在概率论和统计中，通常用方差来度量一组数据的离散程度，计算一个变量与整体均值之间的差异。因此，用预选阶段剩余节点的多维资源空闲率的方差来度量集群的负载均衡性，定义集群的负载均衡度如式(3-11)。

其中，表示节点i上多维资源空闲率的负载均衡度，该值越小，表示多维资源利用越均衡，即各维度资源利用越接近平均值。因此，用集群中满足预选阶段剩余节点负载均衡的均值表示集群的负载均衡度。这种使用方差度量集群均衡性的方法不仅适用于MRWS算法，对于Kubernetes默认的Default算法以及Random、FirstFit算法同样适用。在后面进行算法性能对比实验中，将使用上式的负载均衡度衡量不同调度算法下集群的负载均衡性。

## 本章小结

本章首先详细介绍OpenShift Origin容器云平台的技术架构、技术层次和核心组件。接着介绍该平台的容器编排引擎Kubernetes调度方法的调度流程，该流程分为预选和优选两个阶段，对预选阶段的筛选规则和优选阶段的评分函数集合进行详细分析。在多计算框架进行大数据处理场景下，针对其调度流程和调度方法的不足，提出了一种新的MRWS调度方法，该方法先对调度流程进行适当的优化，使其调度记忆功能，更好地处理重复调度问题。然后根据需要对集群和应用资源需求进行建模，转化为数学表示，为使用FAHP权重参数求解奠定基础。最后设计了MRWS调度算法模块，综合考虑影响调度的各种因素后制定详细的评分规则，根据权重参数和资源空闲率进行资源空闲评分和均衡性评分，资源的权重参数自动求解将在下一章节进行建模和求解。

# FAHP自动求解MRWS调度方法权重参数

在上一章MRWS调度算法的评分函数中，根据资源的重要程度使用系数作为节点CPU、内存、磁盘I/O、网络带宽和已部署Pod数量的重要程度系数。本章需要对筛选后的节点和容器应用资源进行建模，并利用FAHP方法进行Pod应用多维权重参数的自动求解。

## 模糊层次分析法FAHP

模糊层次分析法FAHP (Fuzzy Analytical Hierarchy Process)是在层次分析法的基础发展而来，常用于解决实际问题中影响决策的多因素权重系数。下面从层次分析法开始介绍，为弥补层次分析法的不足引入FAHP方法，然后举例说明如何使用该方法解决实际问题。

### FAHP方法介绍

在许多实际问题中，通常有较多的解决方案，我们需要对影响方案的因素进行比较、判断、评价，然后做出方案选择，这种人为主观做出决策的方式不但效率低下，结果也不够准确。层析分析法AHP (Analytical Hierarchy Process)是由著名的美国运筹学家，匹茨堡大学的T.L.Saaty教授提出的一种层次权重决策分析方法。该方法用于解决复杂的多目标决策问题，影响最终决策的因素被分解为目标层、准则层和方案层等层次，使用数学方法对各因素权重进行精确求解。层次分析法将问题的总目标、各层子目标以及决策因素分解成多个层次结构，然后构建各层次的判断矩阵，求解判断矩阵的特征值。最大特征值对应的特征向量归一化获得上一层各因素对本层目标的权重系数，然后用加权和的方法递阶归并各备选方案对总目标的最终权重，选择最大权重的备选方案作为最终的目标。该方法是一个系统性的分析方法，先将问题分解成多个层次，然后对影响子目标的因素进行比较判断、最终进行综合决策。该方法结构层次清晰、单层的权重参数设置都会影响到最终的决策层，不断分割量化各因素的影响。层次分析法非常实用，将定量分析和定性分析相结合，用于解决许多实际问题如电力分配、旅游决策等，该方法计算简单明了，易于掌握。最后，该方法用于模拟人们解决问题的思维，所需的定量信息较少，仅需要对影响决策因素的重要性做出判断即可。

但是，层次分析法也具有较大的局限性，在一些场景下效果较差甚至无法使用。首先，该方法只能从实际的备选方案中帮助决策者选出较优的决策方案，不能给决策者提供新的解决方案或者反馈合理方案的意见，并且该方法是单目标决策方法，使用者只能从备选方案中选出一个较优的目标。其次，该方法模拟人类大脑决策过程，过于依赖于定性分析，只能进行粗略比较计算，不能完全做到精确模拟。最后，从层次结构转化为成对比矩阵过程中，判断者的主观因素影响过大，使用者不同，得出的最优方案也不尽相同，可重复性较差。并且随着影响决策因素的增加，构建的成对比矩阵阶数增大，计算难度和复杂度随之增加，很难一次性构建出满足一致性要求的成对比矩阵。

为减少人为主观因素对决策结果的影响，模糊层析分析法FAHP在层次分析法的基础上增加人为判断的模糊性。层析分析是通过两两比较构建成对比判断矩阵，而FAHP通过两两比较构建模糊成对比矩阵，提升了层次分析法解决问题的可靠性。在进行问题判断或专家咨询时，通常获得的不是一个具体值，而是一个模糊数，如三值判断(最低可能值、最可能值和最大可能值)、二值区间(上界和下界)等。下面介绍模糊数集的概念，一个明确的集合表示如下:

在明确的集合中，时值为1，时值为0。但对于一个模糊数集，并不能完全明确x是否属于A，只能用[0,1]表示其隶属度。

表示的隶属度，也称为集合A的隶属函数。对于任意的模糊数集，都对应一个隶属函数，隶属函数通常模仿概率论中的分布函数如正态分布、梯形分布、K次抛物线分布、S分布以及柯西分布等，值域在[0,1]上。

1983年，荷兰学者Van Loargoven首次提出用三角模糊数 作为模糊数集的判断标准，并运用三角模糊数的运算和对数最小二乘法获取权重值。设论域R上的模糊数为三角模糊数，其对应的隶属函数满足下列函数:

， a和c分别表示三角模糊数的上界和下界，b是隶属度为1的中间值，x=b时表示x完全属于，x在a和c之外的不属于模糊数。定义一个置信度可以将三元组的模糊数化为一个割集的二元形式，从而构建割集矩阵，设定优化参数后将二元值的矩阵转化成最终的判断矩阵。

根据Arnold J. Kaufmann在文献 中的描述，割集后的基本运算规则如下：

本文采用三角模糊数作为FAHP的隶属函数进行权重参数求解，并将三元组的=(a,b,c)转化成割集形式 。

### FAHP求解权重参数步骤

介绍完层次分析方法和模糊层析分析方法以及模糊数的基本概念后，下面用三角模糊数作为模糊程度的衡量值，FAHP求解权重参数流程如下：

FAHP计算权重参数过程

FAHP计算权重参数过程

如图4.1所示，使用FAHP方法计算权重参数的步骤可以分为构建模糊成对比矩阵、设置转化为割集矩阵、设置优化参数转化为判断矩阵，求解判断矩阵的最大特征值，检验最大特征值是否满足随机一致性比率要求。若通过检测则将最大特征值对应的特征向量归一化作为权重参数值，否则需要调整判断矩阵值，重新进行特征值计算。下面详细介绍求解步骤:

1) 构建模糊成对比矩阵。在FAHP中，对影响决策因素进行两两对比并使用表示其相对重要程度，值越大，表示该因素相对另一个因素对决策目标的影响越大，重要性越高。模糊数、相对重要程度、三元组及其割集如表4.1所示，可以构建最终的模糊成对比矩阵，其中的元素表示因素i相对j的重要程度，模糊成对比矩阵中元素值如式(4-6)所示。

$$\widetilde{a}\_{ij} = \left\{\begin{array}{l}
1 \quad i=j \\ [0.2cm]
\widetilde{1},\widetilde{3},\widetilde{5},\widetilde{7},\widetilde{9}\ or\ \widetilde{1}^{-1},\widetilde{3}^{-1},\widetilde{5}^{-1},\widetilde{7}^{-1},\widetilde{9}^{-1} \quad i\not=j
\end{array}\right.$$

模糊数、相对重要程度和割集关系表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 模糊数 |  |  |  |  |  |
| 重要性 | 同等重要 | 稍微重要 | 重要 | 明显重要 | 非常重要 |
|  | (1,1,3) | (1,3,5) | (3,5,7) | (5,7,9) | (7,9,11) |
|  | [1,3-2] | [1+2,5-2] | [3+2,7-2] | [5+2,9-2] | [7+2,11-2] |

2) 构建割集矩阵并转化为判断矩阵。构建模糊成对比矩阵后，根据表4.1将模糊数转化成三元组的三角模糊数，给定后，将三元组转化成割集，从而构建一个割集矩阵，割集矩阵中的元素如式(4-7)所示。给定一个优化参数表示判断优化程度，可以将二元组割集矩阵转化成判断矩阵A，A是一个方阵，接着求解判断矩阵特的征值和特征向量。设给定的优化参数，将割集矩阵中上下界范围转化成判断矩阵的单值，判断矩阵中的元素，如式(4-8)所示。通过给定获得二元组的割集矩阵，给定割集矩阵一个优化参数可以得到最终的判断矩阵A。

3) 计算判断矩阵最大特征值和特征向量。对判断矩阵特征值的求解是一个纯数学问题，求解方阵特征值和特征向量的方法有几何平均法(方根法)、算术平均平均法(和法)、最小二乘法(幂法)和特征向量法。几种方法求解的特征值近似相等，但也存在细微的差别，这种细微的误差可能会影响实际生产。通常情况下，几何平均法和算术平均获得的结果精确度较差，最小二乘法计算复杂度较高，因此使用特征向量法对判断矩阵的特征值和特征向量进行求解。在本文中，直接使用Python中numpy库eig()函数求解判断矩阵的特征值和特征向量。

4) 一致性检验和归一化权重。通过特征向量法获得判断矩阵的最大特征值及其对应的特征向量后，需要进行层次单排序和层次总排序的一致性检验，用于检测特征向量和真实权值的契合度，即求解的权重值是否合理。判断矩阵一致性检测方法公式如式(4-9)所示。

其中，是判断矩阵最大特征值，n判断矩阵阶数即影响决策因素的数量，CI是一般一致性指标，RI是随机一致性指标，因此，CR是随机一致性比率。若CR<0.1时，判断矩阵满足一致性要求，否则程序自动调整判断矩阵直至满足一致性判断要求。当CR<0.1时，归一化对应的特征向量，其值w就是所需求解的权重系数。n阶随机一致性指标RI的值如表4.2所示。

随机一致性指标RI值

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| n | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| RI | 0 | 0 | 0.58 | 0.90 | 1.12 | 1.24 | 1.32 | 1.41 | 1.45 | 1.49 |

上述是用FAHP方法求解权重参数的全部过程，首先对影响决策的因素进行两两比较，构建模糊成对比矩阵，用表示相对重要程度。针对模糊成对比矩阵，给定获得割集矩阵，设定优化参数获得判断矩阵，对判断矩阵特征值和对应的特征向量进行求解，检测判断矩阵是随机一致性比率值是否满足要求，最大特征向量归一化获得权重系数。

### FAHP求解权重参数示例

上一小节介绍了FAHP求解权重参数的详细流程和计算方法，下面通过一个简单的例子展示该方法求解权重参数的计算过程。假设在一个物理问题中力学指数的评价由巴氏硬度、耐荷重性、耐冲击性和满水变形四个因素决定，其影响力学指数的判断因素层次如图4.2所示。

力学指数的评价因素

力学指数的评价因素

某专家通过对影响力学指数的因素进行两两对比后给出其三角模糊判断数，根据其给出的模糊数构建模糊成对比矩阵如式(4-10)，从左至右依次为巴氏硬度、耐荷重性、耐冲击性和满水变形。

给定和优化参数，根据式(4-7)和(4-8)可以获得割集矩阵和判断矩阵，判断矩阵A如下:

使用特征向量法求解判断矩阵进行特征值和对应的特征向量，最大特征值，其对应的特征向量值，使用最大特征值对判断矩阵A的一致性进行判断，CR=0.082<0.1满足一致性要求。对其最大特征值对应的特征向量进行归一化，其权重w=[0.558,0.265,0.057,0.120]分别表示巴氏硬度、耐荷重性、耐冲击性和满水变形四个因素对力学系数影响的权重系数，用于对后面的目标层做出决策。

## 容器应用多维资源权重自动求解

上一小节详细介绍了FAHP方法求解权重系数的流程，并用一个力学系数的例子展示其计算过程。在MRWS调度算法中，需要使用容器应用各维资源的权重参数进行节点评分，由于需要调度的容器应用数量庞大，不能依赖人工给每一个容器应用的资源需求赋予一个权重值，这样既不准确也不现实。因此，下面将介绍使用FAHP方法对容器应用的多维资源进行建模并自动求解。

### 容器应用多维资源建模

在MRWS调度方法中，集群中节点CPU、内存、磁盘和网络带宽以及已部署的Pod数量作为影响容器应用调度决策的因素，预选阶段筛选后的节点作为其调度的备选节点。不同的用户和应用场景对资源需求不同，容器云中允许用户配置容器应用的资源量既能节约成本，又能实现按需服务。首先对影响调度的因素进行分层建模，如图4.3所示，整个层次可以分为目标层、准则层和方案层。目标层是选择一个预选阶段过滤的后的节点作为容器应用调度的目标；准测层是影响调度决策的因素，包括节点CPU、内存、磁盘、网络带宽以及已部署Pod数量；方案层是所有满足容器资源需求的主机。从该层次模型中可以发现，每一个因素都对方案层直接施加影响，是一个全层次的模型结构。

调度器的速度是衡量服务的重要指标，一种快速的调度算法对容器云至关重要。MRWS调度方法在进行调度流程优化时，增加一个应用类型和对应的参数存储模块，初始可以通过数据中心感知的应用类型，分级建立部分权重系数库。一旦有新的容器应用请求，将资源需求和库中的资源层级进行对比，如果资源差值在可接受范围内，直接使用库中应用的系数作为待调度容器的权重参数，实现快速调度。若系数库中不存在该类型应用，需要重新计算该类应用的资源权重系数，并保存到系数库中，不断扩充应用参数库，也可以人工调整参数，提升精确度。在本文的方法中，权重参数根据容器应用资源需求与单位资源组的比值实现自动求解，无需用户给出模糊成对比矩阵，用户也可以对系数库加以修正。

多维资源递阶模糊层次结构

多维资源递阶模糊层次结构

在云计算数据中心，CPU和内存一般是相对稀缺的资源，容器应用中CPU和内存相对其他资源重要性一般要高。单个节点上部署Pod数量较大，已部署Pod这个影响决策的因素相对其他因素的重要性较弱，但是大量小资源需求Pod会增加节点的管理成本。在构建模糊成对比矩阵时，该因素的模糊数较小，只是作为平衡各节点已部署Pod数量，容器应用的资源需求才是首要考虑因素。根据上述的层次结构图，下面介绍权重参数的自动求解方法。

### 多维资源权重参数自动求解

容器应用多维资源模糊层次结构如图4.3所示，大量容器应用进行模糊成对比矩阵构建时完全依赖人工判断既不准确也不适用于生产环境。在实际系统中，需要对每一个容器应用的多维资源进行权重参数求解，自动构建模糊成对比矩阵，然后转化为割集矩阵和判断矩阵，实现最大特征值的自动求解和判断矩阵的一致性检测。若判断矩阵不满足一致性要求，程序需要能自动调整判断矩阵直至其满足一致性检验。因此，如何对影响决策的CPU、内存、磁盘、网络带宽和已部署Pod等几个影响决策的因素构建模糊成对比矩阵成为解决问题的关键，整体的解决思路是用待调度容器资源需求占单位资源组的比例进行两两对比。设某个容器应用对各维度资源的需求和节点单位资源组如表4.3所示，需要求解容器应用各维度资源的权重系数且。

某容器应用和节点单位资源组

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Resources | CPU | Memory | Disk | Bandwidth | Pod |
| U-Node | C | M | D | B | .. |
| Pod |  |  |  |  | 1 |
| Weight |  |  |  |  |  |

在构建模糊成对比矩阵时，根据容器应用资源需求和单位资源组的比值之间的差值确定模糊数。如在对比CPU和内存的重要性时，采用作为判断依据，使用表示相对重要程度的大小，在[0,1]范围内将其切分为5等份，根据比值差值自动构建模糊数。模糊成对比矩阵的模糊数、容器应用各维度资源需求和单位资源组的比值差值对应如表4.4所示。在数据中心的集群中，各节点的资源构成通常是异构的，即单个服务器各维度资源不同，不能直接使用容器应用与各节点资源比值的差值获取模糊数，需要抽象出单位资源组概念作为构建依据。如1个core的CPU、1G内存、50G磁盘、10M带宽作为一个单位资源组，单位资源组的大小根据集群实际情况确定，将容器应用和单位资源组的比值差值作为模糊数的依据。最后使用表4.4的方法构建模糊成对比矩阵。

模糊数和资源比值差值对应表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 差值 | (0, 0.2] | (0.2, 0.4] | (0.4, 0.6] | (0.6, 0.8] | (0.8, 1) |
| 模糊数 |  |  |  |  |  |
| 差值 | (-0.2,0] | (-0.4, -0.2] | (-0.6, -0.4] | (-0.8, -0.6] | (-1, -0.8] |
| 模糊数 |  |  |  |  |  |

针对已部署Pod这一影响调度的因素，其重要性相对其他资源要低，容器应用对该资源的需求以及节点上能部署的Pod总数都无法度量，不能像其他资源一样直接通过比值的差值作为模糊数的获取依据。因此，将CPU、内存、磁盘和网络带宽四种资源的比值从小到大排序，相较于Pod这一因素的重要性依次获得模糊数为()，反之，Pod相对于其他资源的模糊数为()。

容器应用资源需求和单位资源组比值的差值可以获取各因素之间两两对比的模糊数，从而构建模糊成对比矩阵，最终自动获取割集矩阵和判断矩阵。求解出判断矩阵的特征值和对应的特征向量，判断矩阵满足一致性判断后，归一化出各因素的权重值，实现权重参数的自动求解。下面用一个实际的例子详细展示模糊成对比矩阵的构建以及最终权重参数的求解过程。

若某个集群的单位资源组配置如表4.5所示，集群中各节点各维度资源的配置都是单位资源组的整数倍，单位资源组是构成集群资源的最小组成单元。提出单位资源组的概念主要是为解决异构集群的场景，通常数据中心的集群都是异构的，尤其是在云计算中心中，需要将互联网上大量闲置的计算机作为服务计算单元，这种整合的闲置资源，并不是同构服务集群。

物理集群单位资源组配置

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 类型 | CPU(M) | MEM(M) | Disk(G) | BW(M) |
| 数量 | 1200 | 8000 | 500 | 50 |

在该数据中心集群上进行应用容器的调度，设某个应用容器对各维度资源的需求和集群单位资源组的比值如表4.6所示。

某应用容器的资源需求

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 类型 | CPU(M) | MEM(M) | Disk(G) | BW(M) | Pod |
| 数量 | 300 | 1100 | 210 | 40 | 1 |
| 占比 | 0.25 | 0.1375 | 0.42 | 0.8 | - |

根据各维度资源所占比值的差值作为模糊数的依据，可以构建出该容器应用的模糊成对比矩阵，已部署Pod因素影响较弱，其他因素与该因素对比获取的模糊数较大。自动构建的模糊成对比矩阵如式(4-12)所示，行和列依次为CPU、内存、磁盘、网络带宽以及已部署Pod等影响决策的因素，通过两两对比确定模糊数后获得一个5\*5的模糊成对比方阵。

给定割集参数，优化参数，参数反映模糊数中最大可能值的模糊程度，即真实值与最大可能值的距离，当时，模糊数的真实值就是最大可能值，时模糊程度出现最大范围，真实值和最大可能值接近程度最低。作为优化参数，在模糊给定后反映上界和下界的取值，时取上界值，时取下界值。两个参数对权重参数的影响可以参考文献[26]，该文献对两个参数对权重产生的影响进行了详细的分析。将上述的模糊成对比矩阵转化为割集矩阵，然后再转变为实际判断矩阵如(4-13)所示。

使用特征向量法求解判断矩阵的特征值和对应的特征向量，获得最大特征值，计算一般一致性指标CI=(5.2498-5)/4=0.0623，查找随机一致性RI表在n=5时其值RI=1.12，因此随机一致性比率CR=CI/RI=0.0558<0.1，判断矩阵满足随机一致性比率CR<0.1的要求。

满足一致性的判断矩阵A最大特征值对应的特征向量的值为(0.2291, 0.1452, 0.3607, 0.8903, 0.0600)，将该特征向量归一化获得容器应用CPU、内存、磁盘、网络带宽以及已部署Pod几个因素的权重系数。从权重参数可以看出，该容器应用是一个网络密集型的应用，该应用对网络带宽资源需求更多，其权重参数最大，因此，权重参数也能反映容器应用对某个维度资源的需求量。至此，一个完整的自动构建模糊成对比矩阵并实现各维资源权重参数自动求解过程完成。

MRWS调度算法中，容器应用在进行调度时需要对集群中节点进行评分，选择评分最高的节点作为容器调度目标，在对节点空闲资源评分时需要使用容器应用资源权重系数作为乘积系数。在式(3-4)和(3-8)中，系数就是待调度容器应用各维资源的权重系数，该系数大小反映待调度容器对某个维度资源需求量。如一个CPU密集型应用，系数较大，节点中若某个节点CPU资源空闲较多，则该节点评分相对较高，Pod调度到该节点的几率就越大。使用FAHP对待调度容器Pod各维资源进行层次建模后，通过容器应用资源需求和集群单位资源组比值之间的差值构建模糊成对比矩阵，最终自动求解的各维资源权重值就是空闲资源评分系数。其中单位资源组大小、割集系数和优化参数的选取对调度评分都至关重要。使用系数调节后的MRWS算法可以实现容器云集群中各维资源的均衡利用，将CPU密集型应用调度到CPU剩余资源较多并且各维资源使用较为平衡的节点上，其他密集型的应用调度相似。因此，这种调度方法将各种密集型容器应用尽可能分散混合部署到节点上，避免大量相同密集型集中调度到相同节点上，造成某一维度的资源不足，尤其是在多计算框架容器应用调度的场景下，能够提升集群资源利用率和负载均衡性，极大增强集群的服务性能。

## 本章小结

在MRWS调度算法设计中，需要求解容器应用各维度资源的权重参数作为节点评分乘积系数，本章应用FAHP求解容器应用各维资源权重参数。首先从层次分析法AHP开始，详细介绍该方法的历史背景和解决问题的能力。为减少专家判断的主观性和模糊性，使用各种分布函数作为模糊数，选定三角模糊数作为判断依据，从而引入了模糊层次分析法FAHP，并使用表示因素之间相对重要程度。接着详细阐述应用该方法求解权重系数的步骤，构建模糊成对比矩阵，给定置信度将其转化为割集矩阵，设置优化参数后转化为判断矩阵，求解判断矩阵最大特征值和特征向量，检测该判断矩阵是否满足一致性要求。通过调整判断矩阵，满足一致性判断要求后，归一化最大特征值对应的特征向量获得各维度资源的权重值，接着通过一个力学系数的例子展示各计算步骤的详细计算过程。

最后，使用FAHP求解容器应用各维度资源的权重参数，先对各资源需求和备选节点进行层次建模，获得一个全层次的模型结构。将容器应用各维度资源和单位资源组的比值之间的差值作为自动构建模糊成对比矩阵的依据，获得模糊成对比矩阵，并转化为割集矩阵和判断矩阵，实现权重参数的自动求解。使用一个具体的容器应用展示该应用各维度资源的权重系数自动计算过程，最终获得的权重参数就是MRWS调度算法中用于各节点空闲资源和资源平衡性的评分乘积系数。

# 面向多计算框架的容器云平台研发与调度实验

提出MRWS调度方案优化OpenShift容器云平台底层容器编排引擎Kubernetes调度流程和调度算法后，需要对其资源利用率和负载均衡性进行评估，测试新调度方案下集群的调度性能。本章实验主要包括如下几个部分:

1. 在容器云仿真平台上ContainerCloudSim进行大规模容器应用调度仿真、对比Kubernetes默认的Default算法、Random、FirstFit和MRWS调度算法在资源利用率和负载均衡性方面的优劣。
2. 基于开源OpenShift Origin和存储系统Paladin Storage研发了面向多计算框架的容器云平台Paladin，该平台是一个集海量数据存储管理、多计算框架快速部署、调度优化、用户注册、按需服务等多功能的PaaS平台。
3. 在Paladin上支持十多种大数据处理框架如Hadoop、Spark、Storm等，将其打包成镜像文件，Push到仓库中，用户可以快速构建大数据处理环境。
4. 在Paladin容器云平台上设计并实现MRWS方法的调度器，使用多计算框架容器应用混合部署进行调度测试，对比几种调度方案的任务处理性能。

## ContainerCloudSim仿真MRWS调度方案

自研的面向多计算框架的容器云平台Paladin部署在小规模物理集群上，不能进行调度方案的大规模调度测试和分析。为评估MRWS调度方案的性能和可靠性，需要在容器云仿真平台ContainerCloudSim 上进行大规模的仿真实验。首先构建仿真实验环境，然后开发Kubernetes默认的Default算法、Random、FirstFit和MRWS调度算法，使用大规模负载进行测试分析。

### ContainerCloudSim容器云仿真平台

#### CloudSim云仿真平台

容器和容器编排技术的逐渐成熟推动了容器云的飞速发展，在容器云计算系统中，为评测资源调度策略和集群服务性能，一个容器云仿真平台变得异常重要。调度方案经过仿真平台大量测试和对比分析，不仅可以节约开发时间，也能避免资源浪费和减少试错成本。针对不同的应用场景和新调度方案，如果实际部署大规模容器云平台进行性能测试分析，绝大部分小公司和研究人员并不具备这种条件。在传统的云计算模式下，应用服务的组成、供应、配置和部署条件较为复杂，当用户需求和系统配置动态变化时，评估一个调度策略以及工作负载是相当困难的，一个优秀的云计算平台模拟器可以很好解决这个问题。云计算模拟器通过控制环境变量和重复试验可以加速理论研究和开发过程，根据需求和应用场景不同，各大公司和研究机构推出了一系列云计算仿真平台。

MDCSim 是一个全面、灵活、可扩展的多层云计算仿真平台，整个仿真平台分为通信层、内核层和用户层。通信层用于模拟集群内部通信、内核层用于模拟调度和分析系统性能、用户层用于模拟各种实际应用。该模拟器能够根据底层硬件的不同特点进行混合建模，用于评估数据中心的能耗，让用户在保持低功耗的同时实现集群服务性能的提升。GroudSim 是一个基于Java的模拟器，用于模拟科学应用在网格和云设施上执行问题，应用执行完成后给用户提供了基础的数据统计和分析功能。NetworkCloudSim 用于解决网络模拟问题，弥补其他模拟器对网络细节关注不足，支持MPI和工作流。TeachCloud 用于对MapReduce应用建模并集成一个负载生成器，提供图形化的接口和实现定制的网络拓扑结构。CloudSim 是墨尔本大学Gridbus项目推出的云计算仿真软件，既能对系统性能和应用服务模拟、仿真、测试，也能用于评估资源调度策略的优劣。

CloudSim是一个开源的仿真软件，最大特点就是提供一个虚拟化引擎，帮助数据中心建立和管理各种虚拟化服务。支持大规模云计算资源管理和调度模拟，将数据中心的资源虚拟化为资源池，CloudSim的分层体系架构图如5.1所示。

CloudSim分层架构图

CloudSim分层架构图

#### ContaienrCloudSim容器云仿真平台

随着容器技术的迅猛发展，CaaS(Congtianer as a Service)作为一种新型的服务模式变得越来越普遍。上述介绍的各种云计算仿真平台以及早期的CloudSim版本并不支持容器仿真，一个容器应用仿真平台变得越加紧迫。为促进容器云发展，缩短新方法研发时间，墨尔本大学的研究人员利用CloudSim虚拟化的特点，在其基础上开发了ContainerCloudSim，专门用于数据中心容器应用的仿真。在最新的CloudSim-4.0上已经集成ContainerCloudSim组件，提供Docker容器应用仿真支持，ContainerCloudSim与CloudSim生态圈关系如图5.2所示。

ContainerCloudSim与CloudSim生态圈

ContainerCloudSim与CloudSim生态圈

集成ContainerCloudSim的CloudSim-4.0云平台模拟器已完整支持容器应用的云计算仿真，新版本仿真平台特点如下:

1. 支持大规模云计算建模和仿真。
2. 支持虚拟机调度策略，提供服务器和主机虚拟化。
3. 支持大规模容器应用的建模和调度仿真。
4. 支持计算资源能量感知的建模和仿真。
5. 支持网络拓扑结构和消息传递应用建模和仿真。
6. 支持元素的动态增加、停止和恢复的仿真。
7. 支持混合云的建模和仿真。

在ContainerCloudSim部分，提供容器、VM、Host以及数据中心CPU、内存和存储等资源的管理。还实现了系统动态监控、容器应用执行控制以及容器虚拟机的供应等功能。容器模拟器需要给研究人员提供容器调度方案接口，以及各种调度算法之间的对比和评测，容器调度策略决定容器如何被调度到虚拟机上。调度算法的能耗问题也是容器模拟器关注的重点，需要提供各种算法的能耗度量，此外，还有容器的合并和迁移等功能。最后，模拟器要能够支持容器的扩展，在CaaS中，容器的数量要远远多于虚拟机。ContainerCloudSim是在CloudSim基础上开发而来，也是一个分层的架构，整体架构如下所示。

ContainerCloudSim分层架构图

ContainerCloudSim分层架构图

从下至上依次分为数据中心管理服务层、能耗监控服务层、资源分配服务层、资源管理服务层、虚拟机生命周期管理服务层、容器生命周期管理服务层和负载管理服务层等。每个层次的大致作用如下:

1. 负载管理服务层关注于客户端应用的注册、部署、调度、应用层的性能以及容器应用的健康监控。
2. 容器生命周期服务管理层负责容器生命周期管理，包括创建容器、注册容器到系统中、启动、停止、重启、从一个主机迁移到另一个主机以及容器销毁。除此之外，还负责执行和管理容器任务，监控资源利用率。
3. 虚拟机生命周期管理服务层负责虚拟机的管理包括创建、启动、停止、重启、销毁、迁移以及资源利用率监控。
4. 资源管理服务层负责容器在满足资源需求和软件环境的虚拟机上创建，由容器调度、虚拟机调度和容器合并组成。容器调度根据容器调度策略调度容器到虚拟机、虚拟机调度根据虚拟机调度策略调度虚拟机到主机，合并策略通过合并容器减少主机需求，减少资源碎片。
5. 资源分配服务层负责管理虚拟机和容器的资源分配，由容器分配服务和虚拟机分配服务构成。容器分配服务负责虚拟机资源分配给容器，虚拟机分配服务负责主机资源分配给虚拟机。
6. 能耗监控服务负责数据中心主机能耗监控，构建必要的能耗模型。
7. 数据中心管理服务负责管理数据中心资源，主机开关机和资源监控。

至此，ContainerCloudSim容器云仿真平台介绍完毕，具体的仿真执行流程和代码实现可以参见其开源的源代码。

### MRWS资源利用率实验

在ContainerCloudSim容器云仿真平台上的ContainerPlacementPolicy库下开发MRWS调度算法、Kubernetes默认的Default算法，该类自带Random和FirstFit算法，对比四种调度算法下集群资源的利用率。该容器云仿真平台影响调度因素较多，需要根据实际需要进行一些必要的条件控制。

MRWS算法和Kubernetes的Default算法在预选阶段使用的筛选规则相同，评分阶段除空闲资源评分函数和平衡函数外，其他评分函数相同。因此，假设其他评分函数和外部条件都相同的情况下，只需模拟比较MRWS算法和Default算法的空闲资源评分函数和资源平衡评分函数。实验目的在于对比大规模容器调度下各种调度算法的集群资源利用率和负载均衡性，云计算中虚拟机的调度已有相当多的研究，不是本文的研究方向。首先，需要对仿真平台的部分条件加以限制，假设ContainerCloudSim数据中心的单个Host上只运行一个虚拟机，且虚拟机的资源配置和Host主机相同。本文不对Container迁移算法进行研究，在模拟器上设置容器禁止迁移，即不触发迁移阈值。由于只对容器应用调度研究，不执行具体的云任务，不设置容器、虚拟机和主机的PE数。测试负载是基于PlanetLab 负载的改进，让CPU、内存、磁盘和带宽的利用率从10%到90%随机变化。虚拟机、主机的配置如表5.1所以，CPU的单位是Mips，容器应用根据负载情况随机构建。在云数据中心中，集群通常是异构的，节点拥有资源数量不同，假设存在三种主机，用于模拟异构集群，处理器MIPS的值通过查询处理器文档获得。

主机和虚拟机配置表

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Host类型 | VM类型 | CPU型号 | MIPS | 内存(G) | 磁盘(G) | 带宽(M) |  |
| #1 | #1 | i7 7500U | 49360 | 16 | 1000 | 100 |  |
| #2 | #2 | i5 8200U | 65770 | 32 | 1000 | 100 |  |
| #3 | #3 | X6 1100T | 78440 | 16 | 1000 | 100 |  |

容器应用的配置根据模拟的负载情况自动生成，并对生成的负载自动求解权重参数，参数设置部分代码片段如下：

[ language=Java]  
public static final int VM\_TYPES = 3;  
public static final double[] VM\_MIPS = new double[]{49360, 65770, 78440};  
public static final int[] VM\_PES = new int[]{};  
public static final float[] VM\_RAM = new float[] {(float)16000, (float) 32000, (float) 16000};//\*\*MB\*  
public static final int VM\_BW = 100;  
public static final int VM\_SIZE = 1000000;  
  
public static final int HOST\_TYPES = 3;  
public static final int[] HOST\_MIPS = new int[]{49360, 65770, 78440};  
public static final int[] HOST\_PES = new int[]{};  
public static final int[] HOST\_RAM = new int[]{1600,3200,1600};  
public static final int HOST\_BW = 100;  
public static final int HOST\_STORAGE = 1000000;  
  
public static final int NUMBER\_HOSTS = 30;  
public static final int NUMBER\_VMS = 30;  
public static final int NUMBER\_CLOUDLETS = 200;

模拟实验中设置资源阈值为0.85，节点上的某维度资源使用率超过该阈值后将不能部署新的容器应用，模拟实验主要测量在相同数量容器应用下各调度算法需要服务器的数量。若节点资源利用越均衡，该节点可部署的容器应用数量越多，从而需要的服务器数量越少。实验对比MRWS调度算法、Kubernetes的Default算法简称为KUB算法、Random算法以及FirstFit算法对资源的需求，后两种算法是ContainerCloudSim自带算法，KUB算法只需模拟内存和CPU均衡利用函数即可。MRWS是综合考虑CPU、内存、磁盘、网络带宽以及已部署Pod应用因素的综合评分算法。模拟实验中单位资源组，单个应用容器的负载是单位资源组的随机变化的值，构建出的应用容器负载如表5.2所示。

N个应用容器的资源配置

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 应用容器 | CPU(MIPS) | 内存(M) | 磁盘(G) | 带宽(M) | Pod |
| 1 | 8500 | 680 | 34 | 9 | 1 |
| 2 | 4000 | 840 | 12 | 16 | 1 |
| 3 | 2000 | 3400 | 20 | 6 | 1 |
| 4 | 3400 | 1200 | 80 | 10 | 1 |
| ... | ... | ... | ... | ... | ... |
| N | 7600 | 600 | 30 | 4 | 1 |

使用FAHP自动建模并根据容器应用资源需求和单位资源组比值的差值构建模糊成对比矩阵，自动求解容器应用各维度资源的权重参数，上述容器列表各维度资源对应的权重参数如表5.3所示。

应用容器资源对应的权重参数表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 应用容器 | CPU(MIPS) | 内存(M) | 磁盘(G) | 带宽(M) | Pod |
| 1 | 0.528 | 0.086 | 0.136 | 0.214 | 0.036 |
| 2 | 0.185 | 0.128 | 0.082 | 0.570 | 0.035 |
| 3 | 0.092 | 0.596 | 0.119 | 0.158 | 0.035 |
| 4 | 0.140 | 0.095 | 0.508 | 0.221 | 0.036 |
| ... | ... | ... | ... | ... | ... |
| N | 0.596 | 0.092 | 0.158 | 0.119 | 0.035 |

测试在相同容器应用数量的条件下所需虚拟机的数量，主机Host和虚拟机配置相同，单个Host上只运行一个虚拟机，虚拟机的数量也是Host主机的数量。实验开始时设置一定数量的Host，三种类型Host数量相同，一旦出现所有主机都无法满足新的容器资源需求，同时增加三种类型Host各一台。如200个应用容器时先给定18台主机，以后每次各类型增加一台，即每次增加三台主机，避免某一类型的主机数量过多，测试结果如表5.4所示。

相同容器数量下各算法所需主机数

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 200 | 400 | 600 | 800 | 1000 | 1200 |
| MRWS | 24 | 45 | 63 | 81 | 96 | 111 |
| KUB | 24 | 48 | 69 | 87 | 105 | 123 |
| Random | 24 | 48 | 72 | 93 | 117 | 135 |
| FirstFit | 27 | 54 | 81 | 108 | 138 | 159 |

从上表5.4可以得出，在应用容器数量和集群规模较小的情况下，各种算法对主机数量需求相差不大，MRWS调度算法的优势并不明显。这是由于集群中节点各维度资源过载现象较少，整体集群的资源利用率不也高。随着容器数量和主机数量的增加，Random和FirstFit算法劣势越加明显，相同数量下所需的主机数量更多。这是由于随机调度和首个合适节点调度方法容易造成大量相同密集型的应用部署到同一个节点，从而使某一维度资源过载，不能部署更多的应用。Kubernetes的Default算法考虑了CPU和内存的均衡性，其效果比其他两种算法要好，MRWS调度算法综合考虑了各维度资源的均衡性，其效果最佳，所需主机数量最少。MRWS还考虑节点已部署Pod因素，在资源利用率相差不大的情况下避免某个节点部署大量较小资源需求的容器应用，给节点容器管理造成过大开销。容器应用应该分散和均衡地部署到集群的各节点上，提升集群服务性能。

### MRWS负载均衡实验

集群服务性能的关键在于集群的负载均衡性，而调度策略是影响负载均衡的核心因素。优秀的负载均衡策略不仅可以提升集群负载处理能力，缩短任务执行时间，也能有效避免单点因过载发生故障。当前服务器的负载均衡算法一般有随机算法、轮询和加权轮询、最小连接和加权最小连接、哈希算法、IP散列法以及URL散列等。一些常用的负载均衡组件有Apache、Nginx、LVS(Linux Virtual Server)、HAProxy、KeepAlived、Memcached等。Apache和Nginx是一个HTTP的服务器，具有反向代理能力，通过对用户请求分流实现负载均衡；LVS是一个虚拟的服务器集群系统，可以实现Linux平台下简单的负载均衡；HAProxy提供高可用、负载均衡以及基于TCP和HTTP的代理，适用于负载较大的web服务站点；KeepAlived也是一个组件，用于检测服务集群中故障节点，及时处理不健康的节点；Memcached是内存缓存系统，对于业务查询数据进行缓存，降低节点的负载，也是一个负载均衡组件。本文主要研究集群中各节点多维资源负载情况，目的在于研究新的调度方法是否具有更好的资源负载，各维度资源利用是否均衡。实验将从集群节点单维度资源利用率、单节点负载均衡度、集群整体负载均衡性三个方面进行对比MRWS算法、Kubernetes的Default算法、Random和FirstFit算法的负载均衡度，负载均衡度衡量指标如式(3-10、3-11)所示。模拟实验在假定集群数量和应用容器数量相同的条件下进行，对比各调度算法性能。

首先测试集群中各节点某个维度资源利用均衡情况，希望集群中各节点某个维度资源如CPU利用率趋于平滑，若节点上该维度资源波动过大则表明其负载不均，服务性能可能会降低，下面以CPU为例，其他维度资源类似。

FirstFit算法CPU利用率

FirstFit算法CPU利用率

四种调度算法下节点CPU利用率波动图

四种调度算法下节点CPU利用率波动图

kubernetes算法CPU利用率

kubernetes算法CPU利用率

四种调度算法下节点CPU利用率波动图

四种调度算法下节点CPU利用率波动图

从图5.4看出，MRWS调度算法下CPU利用率波动性较小，Kubernetes调度算法次之，Random和FirstFit算法节点CPU利用率波动较大。CPU波动性表明集群中节点CPU利用不均，有的节点CPU利用率过高，部分节点CPU利用率过低，集群的服务性能不稳定，高CPU利用率的节点容易出现资源过载。MRWS和Kubernetes算法都考虑了CPU利用的均衡性，其中Kubernetes在进行节点评分时考虑了CPU和内存平衡利用，MRWS调度算法综合考虑了CPU、内存、磁盘、网络带宽和已部署的Pod等因素的平衡性。有效避免单节点CPU维度资源利用过高，出现过载现象。一旦某个节点出现CPU利用过高，新的容器应用将不能调度到该节点上，造成其他维度的资源浪费，调度相同容器数量需要更多服务器。因此，MRWS算法能够有效提升集群的服务性能和资源利用率。

接下来的实验使用之前定义的均衡度对集群中单个节点上负载均衡度进行测量，实验中假设应用容器数量相同，分别用四种算法对其进行容器应用调度，计算单个节点的负载均衡度，部分节点的负载均衡度如下。

FirstFit算法节点负载均衡度

FirstFit算法节点负载均衡度

四种调度算法下节点负载均衡度

四种调度算法下节点负载均衡度

kubernetes节点负载均衡度

kubernetes节点负载均衡度

四种调度算法下节点负载均衡度

四种调度算法下节点负载均衡度

如图5.5所示，四种调度算法中MRWS算法的负载均衡度最小，且波动也最小，Kubernetes算法次之，Random和FirstFit算法节点的负载均衡度波动最大并且节点负载均衡度较大。负载均衡度反映节点上各维资源空闲率的均衡程度，负载均衡度越大，各维资源消耗越不均衡，某个维度资源过载现象越严重。由于Random算法随机选择可用节点，FirstFit算法每次选择第一个满足资源需求的节点，完全不考虑各维资源的空闲情况，导致其负载均衡度较大，节点资源利用不充分。Kubernetes平衡CPU和内存利用率，其效果较另外两种要好，MRWS综合了各维度资源空闲率进行评分，其效果最佳，不仅负载均衡度整体偏小，其波动性也小。

下面进行四种调度算法下集群负载均衡度的对比，假定在容器应用数量不同的情况下，比较整个集群的负载均衡度情况。容器应用数量相同，集群负载均衡度越小，集群中过载的节点数量越少，相同情况下集群服务性能会越好，各维度资源消耗更加均匀，容器应用的调度策略更加合理。

FirstFit算法节点负载均衡度

FirstFit算法节点负载均衡度

四种调度算法下集群负载均衡度

四种调度算法下集群负载均衡度

kubernetes节点负载均衡度

kubernetes节点负载均衡度

四种调度算法下集群负载均衡度

四种调度算法下集群负载均衡度

如图5.6所示，MRWS调度算法下集群的负载均衡度最小，大概为Kubernetes算法的二分之一，Ramdom和FirstFit算法的四分之一。因此，在MRWS调度算法下集群中各节点相同维度资源消耗趋于平衡，单节点上各维度资源利用也较为均衡，出现资源过载现象的节点较少，集群的服务性能更优秀。集群的服务性能将多计算框架容器应用混合部署在Paladin容器云平台上进行测试，测试几种调度算方法下混合应用的执行时间。

## Paladin容器云平台研发部署

由于不具备大规模实际容器云环境对几种调度算法进行测试，将在小规模容器云平台Paladin上实现集群性能对比，实验中使用多计算框架容器应用混合部署下同时进行数据处理来测试集群性能。因此，首先需要搭建一个Paladin实验环境，Paladin是自研的一个集海量数据存储管理，面向多计算框架资源调度，实现存储和计算分离的容器云平台。

### Paladin容器云平台介绍

Paladin是一个支持用户构建多种大数据处理框架、并提供海量数据存储与管理的容器云平台。用户在平台上选择所需的计算与存储框架，如批处理(Hadoop、Spark)，流计算系统(Storm、Flink)、分布式的MPI、机器学习框架(Mahout、Tensorflow)、图计算系统（GridGraph、ReGraph)以及数据库系统(MongoDB、Mysql、Redis)等进行数据处理。平台能够快速构建相应的多计算框架运行环境、伸缩容器、并根据具体的需求提供计算任务调度、海量数据存储，使用优化的容器云多维资源利用率均衡调度算法提升集群资源利用率和服务性能。

Paladin平台架构图

Paladin平台架构图

从Paladin架构图可以看出，最底层是基础设施部分，即整个集群的物理机和云主机部分，中间是支持上层的软件服务如分布式文件系统、Docker容器服务、OpenShift Origin容器云等，第三层提供海量数据存储管理和多种数据处理计算框架，顶层是Web服务层。整个架构中最为重要的两个部分数是据存储管理模块和数据处理框架模块，数据存储管理模块是一个为用户提供数据上传、下载、删除、浏览、分享、磁盘挂载等分布式大数据管理存储系统。数据处理框架为用户提供快速构建各种计算框架容器集群环境，配置资源需求、增减容器数量以及应用权限管理等功能。该部分基于开源OpenShift Origin开发，在任何一个容器中都可以挂载用户存储的数据，实现存储与计算分离。当前Paladin容器云平台除支持单应用的Nginx、Apache、PHP、MongoDB等，还支持大量分布式数据处理框架如Hadoop、SPark、Storm、Spark、TensorFlow以及图计算项目Regraph、GridGraph等数十种计算框架运行环境的快速部署。用户既可以用该系统进行大数据存储管理，也能快速构建大数据处理框架，实现海量数据处理。系统管理员对基础设施、存储节点以及系统权限等进行管理。

### Paladin容器云平台部署

首先需要构建小规模面向多计算框架的容器云平台Paladin，该平台集成了大数据存储和管理，实现了存储与计算分离功能。实验中采用6台物理机进行部署，一个主节点Matser，四个Node节点和一个存储S-Master节点，主节点同时也是Node节点。由于该存储需要对用户提供Web Service的服务，单独使用一个节点进行部署。将用户的存储管理服务和计算框架服务分离，分别部署在两个不同的物理机上，提升集群吞吐率，各物理节点的配置如下:

Paladin平台物理机配置

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | CPU核数 | 内存(G) | 磁盘(G) | 带宽(M) | IP |
| Master | 4 | 16 | 1000 | 100 | 192.168.1.100 |
| Node1 | 4 | 16 | 1000 | 100 | 192.168.1.101 |
| Node2 | 4 | 16 | 1000 | 100 | 192.168.1.102 |
| Node3 | 4 | 16 | 1000 | 100 | 192.168.1.103 |
| Node4 | 4 | 16 | 1000 | 100 | 192.168.1.104 |
| S-Matser | 4 | 8 | 500 | 100 | 192.168.1.111 |

各节点上需要安装的主要软件和组件图5.8所示，部署一个完整的大数据存储与处理容器云平台Paladin。在分布式计算框架部分，Master节点上主要部署openshift-ansible及其依赖的各种软件包，用于安装对外提供大数据处理框架的openshift-origin服务集群、底层的Docker及其私有registry仓库、openshift-origin的键值存储系统Etcd、大数据存储管理需要的客户端paladin-client、经过优化改进的origin-web、主操作系统Centos 7.5。Master节点既作为容器云的主控节点，同时也是Node节点，在大数据存储系统中还作为数据存储节点。S-Master是大数据存储和管理的主控节点，使用Centos 6.9系统，安装Nginx服务、研发paladin-web-console对普通用户提供数据存储管理的web-console服务，同时也是数据存储节点。Node节点作为计算节点，安装Docker、openshift-origin集群子节点的需要的依赖组件如kubelet、HAProxy等、还有Etcd和paladin-client等。Node节点都是存储集群的数据节点，系统为Centos7.5。

Paladin平台节点分布图

Paladin平台节点分布图

安装完成后，Paladin平台的计算框架部分，查询集群活跃节点如下:

Paladin安装成功计算节点部分

Paladin安装成功计算节点部分

paladin-web-console存储系统交互部分，用户可以进行存储管理，实现海量数据的上传、下载、删除、修改等操作。还能通过paladin-client将数据以磁盘的形式挂载到物理节点或者容器中，在分布式计算框架容器应用中以文件系统形式直接使用数据，实现存储和计算分离。

Paladin数据存储管理服务

Paladin数据存储管理服务

整个Paladin平台部署完成后，在该平台上开发部署各种数据处理框架和容器应用，平台origin-web-console用户服务界面如下:

Paladin数据处理框架部署服务

Paladin数据处理框架部署服务

至此，整个Paladin数据存储管理和计算框架快速部署容器云平台部署完成，实现了存储计算分离。用户既能进行数据存储管理，实现普通的数据云存储功能，又能快速构建数据处理容器应用框架，实现海量数据处理，为普通用户提供优秀的大数据容器云配套服务。

## 多计算框架容器应用开发

在构建分布式计算应用时，容器数量动态变化需要处理框架及时发现并进行相应的调整，时刻保持计算环境可用。如在Hadoop容器应用框架下，有10个处理Pod，减少4个后，处理框架要能够快速发现并自动缩减为6个容器应用的Hadoop批处理环境。因此，分布式计算框架容器应用需要内嵌一个服务自发现组件如Zookeeper、Etcd、Serf等，处理框架用Serf作为服务注册和自发现组件。

### Serf服务发现

在分布式计算框架中，有的是主从结构如Hadoop、Spark等，有的只是简单的计算框架如MPI等，并无主从之分，需要一个去中心化的服务解决方案。HashiCorp开源的Serf 是一个去中心化的服务发现和节点编排项目，具有成员管理、失败检测、错误恢复、高可用、分区容错等功能，是一个轻量级的服务发现组件。Serf底层实现Gossip协议，一种类似病毒感染的传播协议，能够快速自动感知节点上线和下线。构成分布式计算框架的所有容器应用上都维护一个成员列表，并将节点列表随机发送给周围节点，发送方式有Push和Pull两种，即主动发送和被动接受方式。Serf有一个时效检测器，定期轮询节点列表，给列表中的节点发送心跳包，剔除不在线的节点，确保所有的节点可用。

Serf在分布式处理框架的每个应用容器上运行一个Serf Agent，在第一个容器应用上创建一个聚簇，后面创建容器的代理将会自动连接已存在的聚簇，形成一个集群。Serf是一个去中心化的服务方案，没有一个集中注册的节点，是一个弱一致性的方案，但具有较高的可用性和容错性。在常见的中心化集群中，集群的状态由Master维护，通常为了可靠性需要双备份甚至多备份主节点，主节点会成为整个集群的性能瓶颈。在集群规模较大时，信息的传播给网络带来巨大的负载，集群管理效率和正确性受到挑战，集群规模存在一个上限。去中心化的集群可以进行大规模扩展，其效率和可靠性几乎不受集群规模的影响，Kubernetes底层用Etcd做服务发现，其规模上限通常在一万左右，而Docker Swarm底层用Serf，其规模可以更大，Kubernetes的Etcd将会是制约其集群规模上限的重要因素。

Serf从功能上而言，自下而上可以分为三层：Gossip协议层、消息中间件、客户端接口。Gossip协议层封装Gossip协议以及集群管理的相关操作；消息中间件对传播消息进行封装，包括管理的指令、UserEvent、Query等各种处理逻辑，其中UserEvent是一个单向无需应答的消息，Query是需要应答的双向消息；客户端接口处理集群的输出输出格式，提供rpc接口。Serf是一个弱一致性的解决方案，提供一个简单、高效、可靠的去中心集群管理和服务发现系统实现。

### Hadoop处理框架容器应用构建

在Paladin容器云平台计算框架部分，平台需要支持多种分布式处理框架，方便用户快速构建数据处理环境。用户通过web-console与平台进行交互，只需使用鼠标操作就能构建所需的计算环境并配置容器资源，实现快速自动化部署。计算框架最终以模板Template存在集群中，所需的镜像文件存储在Docker私有仓库Registry中，所有用户都能重复使用该模板，其构建过程图5.12所示。

Hadoop容器处理框架构建流程图

Hadoop容器处理框架构建流程图

首先构建Hadoop容器应用的镜像文件，编写Hadoop Master和Slave的DockerFile文件，该文件在操作系统镜像ubuntu16.04的基础上安装Hadoop所需的环境，主要包括openjdk-8-jdk、hadoop-2.7.2、wget等。还需配置免密的dnsmasq、openssh-server以及服务自发现Hashicorp Serf，将Hadoop所需的配置文件进行修改，整体打包成镜像，部分DockerFile源码如下:

[ language=csh]  
FROM ubuntu:16.04  
MAINTAINER gongkunjxl <gongkunjxl@163.com>  
RUN sed -i 's/archive.ubuntu.com/mirrors.ustc.edu.cn/g' /etc/apt/sources.list  
RUN apt-get install -y unzip curl dnsmasq openssh-server  
RUN apt-get install wget vim openjdk-7-jdk fuse libglib2.0-0  
# dnsmasq configuration  
ADD dnsmasq/\* /etc/  
# install serf  
RUN curl -Lso serf.zip https://releases.hashicorp.com/serf/0.5.0/serf\_0.5.0\_linux\_amd64.zip  
RUN unzip serf.zip -d /bin   
RUN rm serf.zip  
#ADD hadoop-2.7.2.tar.gz /root/  
RUN wget XXX/hadoop-2.7.2.tar.gz(根据实际需要选择版本)  
RUN tar -xzvf hadoop-2.7.2.tar.gz  
RUN mv hadoop-2.7.2 /usr/local/hadoop  
RUN rm hadoop-2.7.2.tar.gz

此外，还有一些修改后的配置文件和样例文件，以及挂载文件系统paladin-client的客户端等。使用docker build命令打包成Image文件，将文件push到私有仓库中，平台私有仓库的IP为172.30.7.23:5000，接下来的Template文件中将使用该镜像文件源。为重复使用该镜像文件以及用户通过web-console快速构建Hadoop计算框架，需要生成处理框架模板放置在平台上，其模板小部分代码如下:

[ language=csh]  
{  
"kind": "Template",  
"apiVersion": "v1",  
"metadata": {  
"name": "hadoop-master",  
"annotations": {  
"description": "Hadoop Master",  
"iconClass": "icon-hadoop",  
"tags": "hadoop"  
}  
 "objects": [  
{  
"kind": "Service",  
"apiVersion": "v1",  
"metadata": {  
"name": "${APP\_SERVICE\_NAME}",  
},  
.......

该模板包含一个Service、一个DeploymentConfig，部署完成后可以实现Container的自动伸缩。根据容器编排引擎Kubernetes底层网络原理，两个Pod之间可以实现IP直接互访，同一个Namespace下的Pod可以通过网桥直接通信。为了更好的实现对外服务，Pod的Service服务将其IP地址内嵌到后创建的同一Namespace下的所有Pod中，但在创建Service之前的Pod中不会嵌入IP，这是实现Pod构成集群的最好方式。利用网络管理的这一特性，使用Serf对第一个容器应用创建聚簇，后面容器的Serf代理通过第一个容器的内嵌IP实现集群发现服务。如第一个容器应用是hadoop-master，由于其创建了Service服务，后面创建的所有hadoop-slave容器中都会有$HADOOP\_MASTER\_SERVICE\_HOST环境变量为hadoop-master的IP地址，以此实现集群的构建。通过web-console可以快速构建Hadoop处理环境，并根据资源需求实现容器的自动伸缩，按需配置服务。

Hadoop集群成员

Hadoop集群成员

同样可以通过hadoop-master Pod的IP地址进行访问，在浏览器中访问Pod的50070和8088端口，查看Hadoop集群状态和运行任务:

Hadoop集群状态信息

Hadoop集群状态信息

至此，一个完整Hadoop容器应用处理框架构建完成，用户可以非常方便地在web-console上实现Hadoop环境的快速部署，并根据资源需求进行Pod数量动态伸缩，实现容器云平台的按需服务，并通过web查看器任务和集群状态。

### 其他计算框架容器应用构建

上一小节详细介绍了Paladin平台上Hadoop分布式处理框架的构建流程，类似上述的构建流程，构建多计算数据处理框架。原始的Docker hub中镜像文件大多只支持单个容器应如PHP、Go、Java、Ruby、Python等语言类容器应用，以及Mysql、MongoDB、MariaDB等数据存储类容器应用。分布式的数据处理框架和计算框架几乎没有，在Paladin上构建了离线数据批处理Hadoop、Spark、流计算Storm、Flink、分布式MPI、机器学习框架Mahout、Tensorflow、图计算系统GridGraph、Regraph等。众多的分布式处理框架可以满足绝大部分用户的需求，该平台将大数据存储管理和数据处理框架有机集合，并实现存储和计算分离。基于Docker技术构建的容器云平台整体性能比虚拟机式的云计算要优秀得多，容器技术减少了虚拟机系统隔离的开销。Paladin平台上构建的分布式应用模板如下:

Paladin平台支持的处理框架模板

Paladin平台支持的处理框架模板

用户选择需要构建的处理框架，创建第一个容器应用及应用的Service，容器应用内部的Serf同时创建集簇，后创建的容器应用通过Service提供的IP用其Serf代理自动连接集簇，构建一个处理框架。根据实际的资源需求，用户动态调整Pod数量，伸缩其资源，实现按需服务。容器应用中的Serf代理会自动监控容器应用的伸缩，用户调整完Pod数量后无需重启或重新构建集群，Serf服务发现解决方案很好地解决了分布式框架容器应用自动伸缩问题。Paladin平台给用户提供一个web-console的可视化服务，用户无需了解部署原理和细节，甚至无需关心其构建过程，只需根据实际需求配置处理框架的容器数量，几秒内就可以搭建好开发测试环境。用户把所有的精力集中在应用的开发上，完全从集群和环境配置中解放出来。该平台同时还提供数据存储管理可视化，可以将数据挂载到所有的容器应用中，无需用户手动分配数据节点。

## 多计算框架应用下MRWS性能测试

部署完面向多计算框架的容器云平台Paladin后，需要对MRWS调度性能进行测试，并与Kubernetes、Random和FirstFit算法性能进行对比。由于是小规模的容器云集群，无法对资源利用率进行大规模测试，调度仿真在ContainerCloudSim已经完成，只进行多计算框架容器应用混合部署下集群性能的对比测试。

### 混合部署多计算框架

Paladin容器云平台只有5个实际的物理节点，在此小规模的平台上测试4种调度算法的性能，不进行大规模的调度测试。因此，构建几种较为典型的密集型应用容器，通过密集型应用任务完成时间测试4种调度算法的优劣。Hadoop作为离线数据批处理，在海量数据处理map阶段完成后不能将所有的数据都放入内存做shuffle，数据需要写入磁盘，其大部分应用都是I/O和网络密集型应用。用其数据读写样例作为I/O和网络密集型容器应用，即Hadoop框架作为I/O和网络密集型调度用例。Spark处理框架的特点是将大部分数据放入内存中运算，每次将map输出的结果保存在内存中，大部分应用是内存密集型应用。Spark的PageRank计算用作内存密集型调度用例。MPI多用于分布式计算，网络和计算资源通常是其瓶颈，MPI的PI计算是一种CPU密集型应用，可以用作CPU密集型调度用例。Paladin集群上将混合部署Hadoop、Spark和MPI三种计算处理框架，每个计算框架的容器数量相同，Hadoop上运行I/O和网络密集型应用TestDFSIO、Spark上运行内存密集型应用PageRank、MPI上运行CPU密集型应用PI计算。

用于调度算法测试的计算框架

|  |  |  |  |
| --- | --- | --- | --- |
| 处理框架 | 类型 | 运行程序 | 容器数量 |
| MPI | CPU密集 | PI计算 | 15 |
| Spark | 内存密集 | PageRank | 15 |
| Hadoop | I/O密集 | TestDFSIO | 15 |

Paladin平台的服务层是基于OpenShift Origin构建的，OpenShift Origin是一个开源的容器云平台，其底层的容器编排引擎是Kubernetes。该平台只是集成在Kubernetes上，并未修改其源码，因此，开发该平台的调度算法和Kubernetes平台上完全一致，最后只需要将调度算法的名称注册到平台的配置文件origin-master.conf中即可替代原有的调度方案。虽然Kubernetes提供了大量的API，但其使用较为繁琐，整个引擎的语言都是go开发，可以使用client-go进行快速开发。client-go是调用Kubernetes集群资源的客户端，对Kubernetes前置API的二次开发都可以使用client-go这个第三方包实现，能够对集群资源对象进行增删查改。调度算法都是使用client-go第三方包开发完成。

进行调度算法测试前，需要构建三种计算框架的容器应用，并对容器应用进行资源配置。根据其运行应用的类型，在相应维度上进行资源配置，如CPU密集型容器应用对CPU资源多配置。实验主要测试多计算框架混合部署下，几种调度算法对容器应用进行调度，集群完成全部调度任务的时间，用于衡量集群性能。

Random调度算法应用容器分布

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Master | Node1 | Node2 | Node3 | Node4 |
| MPI-pod | 3 | 5 | 2 | 4 | 1 |
| Spark-pod | 3 | 2 | 3 | 2 | 5 |
| Hadoop-pod | 1 | 4 | 5 | 2 | 3 |
| Total | 7 | 11 | 10 | 8 | 9 |

FirstFit调度算法应用容器分布

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Master | Node1 | Node2 | Node3 | Node4 |
| MPI-pod | 5 | 2 | 5 | 2 | 1 |
| Spark-pod | 2 | 6 | 1 | 5 | 1 |
| Hadoop-pod | 1 | 5 | 3 | 5 | 1 |
| Total | 8 | 13 | 9 | 12 | 3 |

Kubernetes调度算法应用容器分布

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Master | Node1 | Node2 | Node3 | Node4 |
| MPI-pod | 3 | 3 | 3 | 3 | 3 |
| Spark-pod | 3 | 3 | 3 | 4 | 2 |
| Hadoop-pod | 1 | 3 | 5 | 1 | 5 |
| Total | 7 | 9 | 11 | 8 | 10 |

MRWS调度算法应用容器分布

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Master | Node1 | Node2 | Node3 | Node4 |
| MPI-pod | 3 | 4 | 3 | 3 | 2 |
| Spark-pod | 2 | 4 | 3 | 3 | 3 |
| Hadoop-pod | 3 | 2 | 3 | 3 | 4 |
| Total | 8 | 10 | 9 | 9 | 9 |

通过4种调度方案下应用容器的分布可以看出，Random随机调度三个计算框架的容器应用，不做任何资源空闲率负载均衡，CPU密集型容器应用MPI-pod在Node1上有5个，而Node4上只有1个。I/O密集型容器应用Hadoop-pod在Master上有1个，但Node2上有5个，各种容器应用随机调度分配，但是各节点上容器应用数量相差不大。这种调度方式必然会导致Node4上CPU资源紧缺，Node2上I/O资源紧缺，影响集群服务性能。FirstFit也不做资源使用率平衡，每次选择第一个满足资源的节点进行调度，导致前面的节点负载过高，后面节点负载较低，从整体容器应用分布来看，Node4上仅分配了三个容器应用，而Node1节点高达13个，导致部分节点维度资源负载都过高，部分节点资源利用率不足。Kubernetes对CPU和内存的资源空闲率做了平衡，尽量使内存和CPU资源平衡使用，从分布看出CPU密集型MPI-pod和内存密集型Spark-pod在各节点上的分布较为均匀，都在个容器应用，大部分都是3个，但是I/O密集型应用Hadoop-pod分布不均，Node1和Node3上只有1个Hadoop-pod，Node2和Node4上有5个Hadoop-pod，这种只注重CPU和内存均衡的调度方式导致集群各节点I/O密集型分布不均，影响集群性能。MRWS调度方式可以看到几种密集型应用分布较为均衡，各节点的各种密集型容器应用都在个，单个节点容器数量也在个之间。整个集群各维度资源利用率较为均衡，集群负载均衡性较好。

### 单个计算框架服务性能

针对四种调度方案下混合部署多计算框架容器应用，首先进行单个计算框架的性能测试，在容器应用中设置资源的下限，不限制应用容器使用资源的量，同一个节点上的容器应用可以对节点资源进行自由竞争。对比四种调度算法下单个计算框架应用执行的时间，在Hadoop-pod中运行TestDFSIO程序，分别读写15个128M的文件，记录读写总时间。Spark-pod中运行PageRank程序，测试文件大小约3.2M，轮数为2000轮迭代。MPI-pod中运行PI计算程序，15个节点进行110000轮迭代。每个调度算法分别运行三次，取三次运算时间的平均值，各计算框架容器应用单独执行时间如下(单位:秒)。

四种调度算法应用容器执行时间

|  |  |  |  |
| --- | --- | --- | --- |
|  | Hadoop-TestDFSIO | Spark-PageRank | MPI-PI |
| Random | 255.92 | 267.33 | 387.19 |
| FirstFit | 276.48 | 274.67 | 390.69 |
| Kubernetes | 250.44 | 252.00 | 214.42 |
| MRWS | 228.33 | 250.52 | 212.88 |

对比四种调度算法单计算框架下容器应用的执行时间，可以很明显地发现进行I/O密集型容器应用Hadoop-TestDFIO测试时，Random、FirstFit以及Kubernetes调度算法没有考虑I/O资源均衡利用，导致多个I/O密集型应用被调度到同一节点。在容器中运行实际应用时，大量I/O密集型应用竞争I/O和网络资源，导致应用效率低下，执行时间过长。MRWS调度算法考虑I/O和网络带宽因素，尽量分散到集群各节点上，其性能最佳，Hadoop-TestDFSIO执行时间最短，FirstFit效果最差，Random和Kubernetes相差不大。在单独运行内存密集型应用Spark-PageRank容器应用下，Random和FirstFit都不考虑内存利用率均衡性，其运行时间最长，尤其是FirstFit将大量内存密集型调度到了同一个节点，使得内存成为该节点瓶颈，运行时间最长。Random随机分配节点，其效果一般，Kubernetes和MRWS均做了内存利用率均衡，其效果较好，各内存密集型容器大致可以分散到集群各节点上。与内存类似，CPU密集型应用表现最为明显，PI运算是一个典型的计算密集型应用，大量的时间花费在三角函数的运算上，Random和FirstFit均不做CPU利用率均衡，导致大量的CPU密集型应用调度到同一个节点，该节点的CPU成为其性能瓶颈，运行时间较长。Kubernetes和MRWS调度可以将CPU密集型应用尽可能分散到各节点上，对CPU空闲率进行均衡，使得任务运行时间大大缩短。

### 混合部署多计算框架下的服务性能

混合部署三种密集型容器应用框架，同时运行Hadoop-TestDFSIO、Spark-PageRank和MPI-PI容器应用程序，对比几种调度算法的执行效率。混合部署和同时运行不同计算框架类似于多用户在容器云集群上部署多种计算框架容器应用的场景，不仅相同密集型容器应用需要对节点资源进行竞争，其他计算框架也会参与资源竞争。比如同一节点上MPI-pod对CPU进行竞争，该节点上部署的Spark-pod也会参与竞争CPU资源，调度方案不合理将导致整个集群性能低下，所有用户应用执行效率都将受到影响。每种调度算法执行3次，取3次执行时间的平均值，四种调度下同时运行多计算框架容器应用执行时间如下(单位:秒)。

同时执行多计算框架应用容器执行时间

|  |  |  |  |
| --- | --- | --- | --- |
|  | Hadoop-TestDFSIO | Spark-PageRank | MPI-PI |
| Random | 316.46 | 584.00 | 418.43 |
| FirstFit | 355.81 | 603.33 | 454.71 |
| Kubernetes | 299.87 | 503.33 | 271.33 |
| MRWS | 269.10 | 491.66 | 256.62 |

首先将执行时间和单个计算框架容器应用单独执行时间进行纵向对比，在混合部署同时执行多计算框架下所有应用执行时间都变长。如Hadoop-TestDFSIO在MRWS调度下单独执行时间是228.33，在混合部署执行下269.10，这是由于更多的容器应用参与资源竞争导致整体执行时间变慢。Spark-PageRank执行时间变化最为明显，因为PageRank的运算不仅需要大量的内存资源，还需要一定的CPU，而CPU资源被MPI-PI计算框架占用，导致其执行时间变慢。

横向对比各调度算法混合部署多计算框架下同时运行容器应用执行时间，FirstFit任务处理时间最长。这是由于其每次选择第一个可用节点导致其资源利用不均，部分节点资源利用不足。Random和Kubernetes相较于MRWS在Hadoop-pod的性能较差，MRWS调度算法性能最优，单个任务的执行时间相较于其他调度算法执行都较短。因为MRWS调度算法综合考虑了容器应用资源需求以及节点的CPU、内存、磁盘、网络带宽和节点已部署Pod等因素，利用FAHP自动建模和求解容器应用多维资源权重参数，利用权重参数对节点综合评分，选取最优的节点作为容器应用调度目标。这种调度方式在多计算框架下性能表现最为突出，通常在用计算框架进行数据处理时Hadoop需要大量I/O和网络带宽，Spark需要大量内存资源，MPI需要计算资源较多，混合部署下需要平衡各维度资源的利用率。各计算框架处理任务不同，资源需求也不尽相同，在混合部署时需要充分考虑各计算框架的特点才能充分利用容器集群资源，缩短多计算框架容器应用在容器云平台上的任务处理时间，提升集群服务性能。

## 本章小结

本章从实验的角度对比四种调度算法性能优劣，首先详细介绍云计算仿真平台CloudSim和在此基础上研发的容器云仿真平台ContainerCloudSim。在容器云仿真平台上对比几种调度方法的集群资源利用率和负载均衡度，MRWS相较于其他几种调度方法具有更低的负载均衡度和更高的资源利用率。接着介绍和部署面向多计算框架的容器云平台Paladin，该平台是一个存储与计算分离的容器云平台。然后在该平台上构建数十种数据处理框架，并以Hadoop批处理框架为例详细介绍其构建流程，用户在几秒内实现大数据处理框架快速构建并进行容器伸缩扩展。最后，在Paladin平台上进行多计算框架容器应用调度分析，对比几种调度算法任务处理性能。分别以I/O和网络密集型应用Hadoop-TestDFSIO、内存密集型应用Spark-PageRank和CPU密集型应用MPI-PI计算框架容器应用为调度用例，在混合部署几个计算框架时单独和同时进行任务处理，对比各计算框架容器应用的执行时间，几种算法中MRWS的任务处理时间更短。

# 总结与展望

本文在进行“多元化大数据高效能存储与处理”项目研究时，发现在面向多计算框架的容器云平台Paladin上进行多计算框架容器应用混合部署下集群性能较低。Paladin的服务层是基于开源OpenShift Origin容器云平台构建，而OpenShift Origin的容器编排引擎是Kubernetes。接着对主流的容器编排引擎Kubernetes、Mesos以及Docker Swarm进行分析，尤其针对Kubernetes的调度流程和调度算法进行深入研究，发现其调度流程不够优化，并且调度算法在评分阶段仅考虑内存和CPU的空闲率平衡情况，导致其他维度资源利用率不足和负载不均衡，从而使集群效率低下。针对其调度方案不足，本文首先优化其调度流程，通过一个Hash表管理重复部署的容器应用，避免镜像重复下载，节约磁盘和网络资源。

针对Kubernetes容器调度算法不足，本文提出一个基于多维资源权重参数的调度方案MRWS，该方法综合考虑集群中节点CPU、内存、磁盘、网络带宽和已部署Pod数量等因素，赋予每个影响调度因子一个权重值进行综合评分。权重参数根据容器应用对资源的需求和节点资源空闲率进行建模，使用模糊层次分析法FAHP自动构建模糊成对比矩阵和判断矩阵，计算出满足一致性要求的权重参数作为待调度容器应用各维度资源的权重值。对于新提出的调度算法，在容器云仿真平台ContainerCloudSim上进行大规模调度仿真，并与Kubernetes、Random、FirstFIt调度算法在集群资源利用率和负载均衡度方面进行对比。最后在面向多计算框架的容器云Paladin平台上实际开发调度方案，进行多计算框架混合部署实际调度性能测试。实验表明新的调度算法无论在单计算框架容器应用执行效率还是同时运行多计算框架容器应用执行效率方面都有较大的提升。

总体而言，本文主要围绕面向多计算框架的容器云平台Paladin上多计算框架容器应用混合部署下执行效率过低的问题做了以下几个方面工作:

1. 调研当前容器编排引擎集中式调度、两层调度和共享状态调度模型的典型架构，并以实际例子分析各自的优缺点。深入分析Kubernetes调度流程和算法不足，对其调度流程进行优化并提出了一种新的综合考虑容器应用CPU、内存、磁盘、网络带宽以及已部署Pod等因素的调度方案MRWS。
2. 针对调度方案MRWS中多维资源权重问题，使用模糊层次分析法自动建模和构建满足一致性要求的模糊成对比矩阵和判断矩阵，实现权重系数自动求解。
3. 在容器云仿真平台ContainerCloudSim上针对新的调度方案进行大规模调度仿真，并与常见的Random、FirstFit以及Kubernetes默认的Default调度方案进行资源利用率和负载均衡度对比。
4. 在Paladin容器云平台上构建数十种多计算框架，用户可以实现快速构建大数据处理环境和容器伸缩。接着开发新的调度方案MRWS，并在该平台上进行多计算框架混合部署，测试对比几种调度算法的性能，实验表明新的调度方案极大提升了集群的服务性能，缩短多计算框架任务的执行时间。

本文主要围绕提升面向多计算框架的容器云平台Paladin服务性能开展研究工作，容器资源调度是提升容器云集群的关键因素，要使一个容器云平台具有更好的性能，在后续的研究工作中可以从以下几个方面进行:

1. 数据亲和性调度。Paladin是一个集数据存储与处理于一身的容器云平台，底层是一个分布式的文件系统，通过客户端磁盘挂载形式实现数据访问，并提供web-console的大数据上传、下载、删除、修改等管理操作。因此，在数据处理框架进行数据处理时，数据的位置对其性能存在重大影响，需要研究以数据为中心的调度方案，实现数据亲和性调度，将数据存储和处理容器尽可能调度到相同或近距离的节点，减少网络传输和磁盘I/O。
2. 优先抢占式调度。在大数据批处理中有的任务无需及时处理，如Hadoop进行离线数据处理，有的需要进行及时快速地处理如流计算等。调度系统需要实现优先抢占式调度，在紧急任务模式下暂停优先级较低的任务，给予优先级高任务更多的计算资源。
3. 容器迁移策略研究。多计算框架混合部署场景下资源的分配是静态的，有的用户在部署完计算框架后并不进行任务处理。根据这一特点可以实现集群资源超卖，在保证性能情况下实现经济利益最大化。一些不太活跃用户的容器可以迁移到性能较低的节点上，在用户使用时再迁回高性能的节点。
4. 任务特征调度。根据积累数据中心的数据提取更多用户计算框架处理任务的特征，根据特征制定合适的调度方案，最大化的利用集群的资源。让集群调度更加智能，实现按需服务，提升用户的服务体验。

光阴似箭，岁月如流水，转眼已在园子里度过了三年时光，在三年硕士期间，老师们的谆谆教诲，同学们的热情帮助，让我不仅在学业上进步，也让我度过人生中一段快乐的时光。

首先，我要感谢我的导师武永卫教授，在硕士三年期间不但给予我专业的学术上指导，在生活中也给了我很多帮助，更教会了我许多做人的道理。武老师是一个和蔼可亲的人，他治理下的实验室宽松、活跃以及浓厚的学术氛围让我收获颇丰。我还要由衷感谢陈康副教授和姜进磊副教授，陈康老师不仅与学生亦师亦友，更是一本计算机的百科全书。他知识面广而深，在学术问题上严谨、认真的态度深深地影响着实验室的每一个人，尤其是在我小论文投稿方面给予了很多专业的建议。再次感谢三位老师在硕士期间给予的指导和帮助，祝各位老师阖家幸福！

其次，我要感谢实验室同学的照顾，艾志远是第一个给予我指导和帮助的师兄，他工程能力极强，对我帮助很大。马腾、刘梦醒、杨珂、李雪、夏鑫、李志涛、何东标等实验室的同学不仅让我感受到Madsys实验室的温暖，也常常帮我答疑解惑，与大家在一起的时光很开心，衷心祝愿大家学业进步，事业有成！

最后，我要感谢我的父母和女朋友，你们在我困惑和无助的时候给予我很大鼓励和安慰，一直在背后默默的付出和支持我。感谢你们辛勤的付出和爱，让我能够在硕士期间安心完成学业，你们的爱是我前进的不竭动力。谨向所有在我硕士期间帮助我和指导我的老师、同学、朋友表示衷心的感谢！

1991 年 04 月 27 日出生于 贵州 省 铜仁 市。

2010 年 9 月考入 清华 大学 计算机 系 ，2014 年 7 月本科毕业并获得 工 学士学位。

2016 年 9 月免试进入 清华 大学 计算机 系攻读 硕士 学位至今。

龚坤, 武永卫, 陈康, 等. 容器云多维资源利用率均衡调度研究. 计算机应用研究(已录用, 2020第37卷第4期)