开放式计算管理系统研究*)

冯 萍 刘君瑞 孙 蓬
(西北工业大学计算机学院 西安710072)

摘 要 对于大规模科学计算来说,高可用系统仍然是目前需要解决的关键技术。文章介绍了一个开放式计算管理系统 (Open Calculating Management System,OCMS)。OCMS 体系结构是基于 Web 技术的、遵循开放式网格协议(Open Grid Services Architecture,OGSA)的计算服务的集合。在网格结点中的集群系统采用 BrowserS / Server / ServerS 体系结构实现,其服务支持多用户操作。基于 Java 技术的 BrowserS / Server / ServerS 体系结构具有单一映像功能。最后文章介绍了在 OCMS 系统运行中尺度大气数值模型计算的性能报告。

关键词 开放式计算管理,BrowserS / Server / ServerS,Java,集群,OGSA

An Implement of Open Calculating Management System

FENG Ping LIU Jun-Rui SUN Peng
(College of Computer,Northwestern Poly-technical University,Xi'an 710072)

Abstract The need to conduct and manage the largescale scientific calculating dramatically increased over the last decade. However,the high availability calculating system is a key problem for largescale scientific calculating. This paper introduces an open calculating management system (OCMS). OCMS infrastructure is based on the Web Services technology and Open Grid Services Architecture (OGSA)specification. For the sake of high availability,the OCMS is designed as a collection of Grid services,and the cluster of the Grid site has BrowserS / Server / ServerS architecture which can be realized the performance of Single System Image Clustering easily. The OCMS services can be accessed concurrently by multiple users. The OCMS has been implemented based on Java distributed technology. One results of using OCMS for Model calculating of Mesoscale Model (MM5)is report.

Keywords Open calculating management,Browsers / server / serverS,Java,Clusters,OGSA

1 引言

随着网格技术的发展,大规模科学计算从集群系统迁移到网格环境是发展的必然趋势,但是,目前成熟的集群系统管理系统均为 C/S 模式,例如,挪威 Scali 公司的 Scali 集群管理系统。对于大规模科学计算来说,高可用系统是目前需要解决的关键技术。我们研究的 OCMS 体系结构支持机群和网格环境。为了提供高可用的网格计算环境,对于集群系统管理我们提出了一种新型的 BrowserS / Server / ServerS 体系结构,其特点是,既支持 Internet 的 Browser/ Server 模式,又易于实现集群的单一映像功能。

OCMS 体系由服务器端和用户端的一系列服务协议的集合组成,包括操作命令(operate command,OC)、用户接口(User Portal,UP)、注册服务(Registry Service,RS)、计算准备服务(Calculating Preparation,CP)、计算执行服务(Calculating Executor,CE)、计算管理(Calculating Monitor,CM)和应用数据可视化(Application Data Visualize,ADV)。

2 OCMS 体系结构

OCMS 体系结构是基于 Web 技术和 OGSA 协议的[1]。OCMS 支持用户在集群和网格环境下进行大规模的科学计算。而 OCMS 的功能主要是由 OC 具体操作实现的。OC 指定问题、系统或机器参数的参数范围,包括程序变量、文件名、目标机、机器规模、调度策略、数据分发等。OCMS 收集源结点的申请和输入的文件,确定实施计算的目标结点名和全部输出文件名字。当一个计算完成后,结果文件包括输出文件和性能数据存储在数据仓(Calculating Data Repository,CDR)中,并通过可视化技术展示在浏览器窗口。

网格点的集群结构参见图1,其体系结构是基于 BrowserS/Server/ServerS 的。OCMS 系统作为一个开放式网格结构,由分布于若干网格点的若干网格服务组成。一个网格点是一个能够访问 Internet 的

*)该课题得到国家863计划引导项目(编号:2003AA001018);航空科学基金项目(编号:02F53031)资助。冯 萍 副教授,主要从事计算机网络系统结构,网格计算等方面的研究。

高速集群。网格服务是一个可以运行在网格点的应用，可以通过预定义的接口远程调用该服务。用户可以调用服务，特定的服务也可以作为用户申请服务。

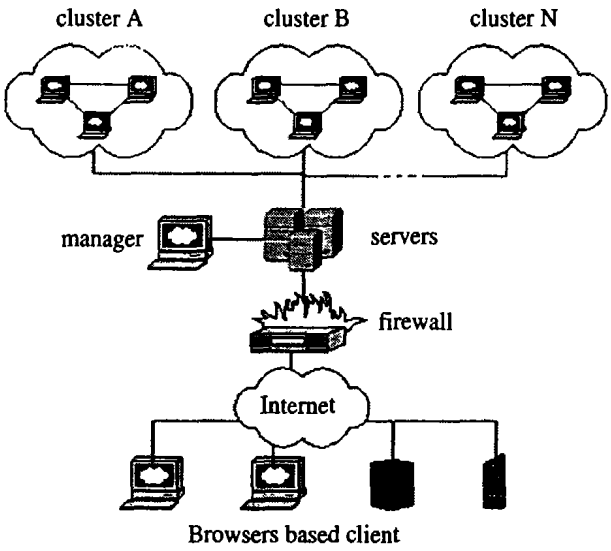

图1 基于 BrowserS/Server/ServerS 的集群体系结构

2.1 操作命令(operate command,OC)

操作命令 OC 提供给用户，用于指定计算问题的参数，包括程序变量、文件名、目标机、调度策略和数据分发等。OC 指令还可用于申请特定的性能，例如，负载平衡、执行、通信和同步时间等。

OCMS 的应用文件可以通过使用 OC 指令来指定以自动实现数据计算。当计算完成，输出文件和结果数据存储在数据仓中，多种可视化工具提供给用户对可视化计算结果进行研究。

2.2 用户接口(User Portal,UP)

用户接口 UP 是 OCMS 中位于用户端的部分，是基于浏览器端的代理，向用户提供一个基于浏览器的接口，用于网格和集群环境下提交、管理、控制，分析和可视化计算。用户通过浏览器可以控制计算，UP 向用户描述系统的入口点，使用户可以向目标机加载 OC 应用、提交、控制和管理计算，以及指定显示计算过程中的性能参数和输出数据。用户操作窗口参见图2。

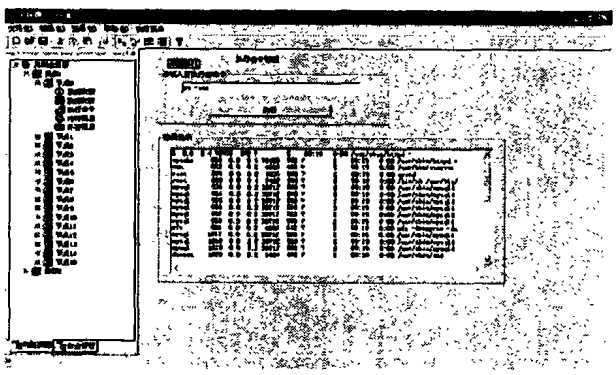

图2 用户操作窗口

2.3 注册服务(Registry Service,RS)

注册服务 RS 作为公共服务实现全部服务的注册。由于 OGSA 环境下主要是临时服务，存在生命周期的问题，必须有生命周期的管理。系统要求在正常情况下，要定期申请服务延续，RS 允许在生存周期内申请服务延续。在延续失效之前如果服务未更新其延续，RS 从注册表中删除这个服务。该系统使用一个通知机制来通知用户端一个新的服务已经被 RS 注册，以及存在的服务更新延续失败等状态。服务延续是一个有效的处理网络失败和保证网络恢复的方法，即使系统某些部分出现故障，不会影响该服务占用资源的回收，用户端确知该服务的生存期，可以指定服务延续。通知机制是通过发送方接口和接受方接口传递实现的，通知机制允许用户通过注册来获取特定的消息。因此，系统始终给用户提供一个网格服务环境的动态更新的视图。

2.4 计算准备(Calculating Preparation,CP)

CP 负责计算准备，其功能主要是根据用户输入来组织计算环境，其功能包括：(1)直接执行和执行 OC 的命令；(2)指定执行计算的目标结点；(3)指定调度；(4)指定输出文件。CP 接收到输入文件后，为了产生相应的计算，它自动通知 RS。在计算环境产生后，计算任务提交到 CE 以便执行。

2.5 计算执行(Calculating Executor,CE)

CE 服务具有在网格点执行计算的功能。CE 作为公共服务，用户可以动态地调用这些服务，并在安装了服务的站点执行计算。操作根据调度策略可以分为4种模式。

1)单处理机作业；

2)集群作业；该模式允许定义资源的类型和使用方法，保证一个作业可以获取所需的资源，用户通过浏览器向服务器提出批作业请求，用户依靠服务器管理各类对象，服务器根据用户的请求工作，完成一个批作业。

3)单 Globus 作业；

4)多 Globus 作业。

在单 Globus 或多 Globus 作业模式，使用局部资源分配管理(Globus Resource Allocation Manager,GRAM)，GRAM 由 gatekeeper、任务管理者、资源管理者等组成，可以为一个或多个计算资源服务，其功能为，处理资源描述语言 RSL 表述的资源请求，针对可用资源等情况，对请求做出拒绝这个请求或创建一个或多个满足这个请求的进程。对创建的作业进行远程监控和管理。根据所管理资源的信息周期地更新元计算目录服务(Metacomputing Directory Service,MDS)。

CE 提供的功能有：(1)增加和取消计算；(2)执行计算；(3)取回计算的状态信息；(4)定期通知回收

状态信息，以避免客户端投票表决；(5)终端计算；(6)传递计输出结果；(7)重新获取属于一个应用的全部计算结果。

CE 方法提供同步(阻塞)和异步(非阻塞)模式。通过通知回收，异步模式可以随机地请求答复。监控主要使用异步模式，获取信息避免了开销大的投票表决。此外，多用户可以并发地访问 CE 服务，满足可扩展网格的功能需求。

在执行脚本中可以给出计算的最大执行时间。如果执行过程中，没有明确的终止消息，则计算属于异常终止(例如，程序或状态错)或者其超过了最大执行时间。为了处理偶然的工作调度错误，或者某些非确定应用因素，或者系统行为，在向 UP 报告失败之前，用户可以指定 CE 重新运行在一个计算的最大时间数内。用户也可以人为地重新启动(失败的)计算。

计算执行服务 CE 负责管理目标结点的计算。CE 起动计算的执行控制。一旦完成，CE 随机地存储计算结果和性能数据进入计算数据仓 CDR，用户可以通过 UP 并发访问 CDR。

2.6 计算管理(Calculating Monitor，CM)

计算管理 CM 使用 CE 服务在线执行和监控计算。通过选择相应的 OC 指令，全部指定应用的计算可以被递交执行。在 CM 中，一个活动的监听者(线程)从 CE 接收关于全部计算状态转换的通知。当执行计算的时候，计算状态变化通过 UP 显示。

OCMS 定义了两种输出结果，分别是可视化性能显示和数据输出文件。可视化性能显示，根据 OC 指示执行性能分析(例如，执行时间，负载，通信等)。在计算准备阶段需要输出文件。一旦计算终止，性能分析和输出文件被存储在 CDR 中。可视化性能结果显示参见图3。

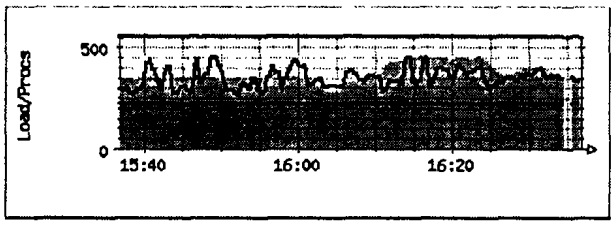

图3 网格负载的性能分析(以小时为单位)

3 实验

为了评价 OCMS 的原型实现，我们进行了中尺度大气数值模型计算(MM5)[2]。在该系统上进行了一个34×37×23的小规模气象算例的测试。在计算中我们使用 MPICH-G2作为计算工具，网格点集群是由8台 PC 机组成的集群系统，Intel Pentium III 800 MHz CPU 和256MB 内存。使用 OGSA 协议提交计算，使用 MPICH-G2库函数进行通信。计算结果完全正确，系统运行的执行时间以及并行效率的测试数据参见图4。

结论 根据机群和网格环境下支持科学计算的需求而开发的 OCMS 具有高性能、高可用性等特点。

采用 Java 分布式技术来设计和实现基于网格的计算服务。通过预定义的接口，可选择多种工作调度，适用于多种运行环境，支持对多用户的服务，其开放式结构，允许添加新的目标对象。

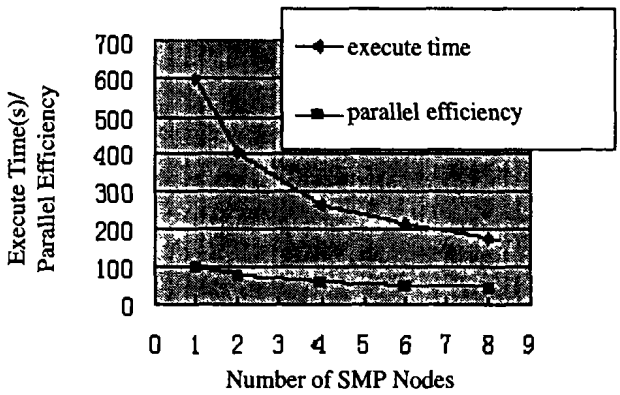

图4 性能测试数据

在网格点中的机群采用 BrowserS / Server / ServerS 体系结构，易于实现集群的单一映像功能[4]。系统管理控制灵活、执行效率高；用户界面实用方便、应用面广。

参 考 文 献

1 Foster I, Kesselman C, Nick J, Tuecke S. The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration. The Globus Project and the Global Grid Forum, Jan. 2002. http://www.globus.org/research/papers/OGSA.pdf

2 MM5 Modeling System Overview. http://www.mmm.ucar.edu/mm5/ MM5 Community Model Homepage.htm

3 MPICH-G2. http://www3.niu.edu/mpi/

4 Buyya R, Cortes T, Jin H. SINGLE SYSTEM IMAGE (SSI). The International Journal of High Performance Computing Applications, 2001, 15(2): 124～135