文章编号:1001-2486(2005)04-0075-06

仿真网格原型系统的设计与实现*

张传富,刘云生,张 童,查亚兵,黄柯棣
(国防科技大学 机电工程与自动化学院,湖南 长沙 410073)

摘 要:进行仿真网格原型系统的研究是解决仿真网格关键技术的基础。SGE网格引擎是一个构建本地和集群级网格的工具,通过SGE构建基本的网格平台,根据HLA仿真任务的特点,对分布式仿真网格的原型系统进行了研究和设计。提出了仿真网格的双通道通信机制:一个通道负责仿真网格组件的通信,另一个通道负责HLA仿真任务之间的通信。提出了仿真任务在仿真网格中的三种调度模式:以进程为单位的调度、以联邦成员为单位的调度、以联邦为单位的调度等。通过研究初步实现了仿真网格的原型系统,并对HLA仿真任务的三种调度模式进行了简单的测试。

关键词:SGE;分布式仿真网格;任务调度;双通道通信;联邦生成器;HLA
中图分类号:TP391.9 **文献标识码**:A

The Design and Implementation of the Simulation Grid Prototype System

ZHANG Chuan-fu, LIU Yun-sheng, ZHANG Tong, ZHA Ya-bing, HUANG Ke-di
(College of Mechatronics Engineering and Automation, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: The simulation grid is a distributed simulation platform built through grid technology. On the platform the distributed simulation applications can be executed by utilizing the powerful sharing resource. The research of the simulation grid prototype system is the basis of solving the key technology to build simulation grid. SGE is a tool to construct the local and cluster grids. This paper presents a recent research of building SGE-based distributed simulation grid prototype system. Specifically, its framework provides the dual channels of communication. One channel takes charge of communicating among components of the grid, and the other one takes among HLA federates. In particular, the grid supports three modes of HLA scheduling tasks, which are differentiated according to (1) processes, (2) federates and (3) federations. Results of this study demonstrate the technical feasibility of constructing SGE-based simulation grid, and simply build the prototype system to test the performance of three scheduling task modes.

Key words: SGE; distributed simulation grid; task schedule; dual-channels communication; federation spawner; HLA

仿真网格就是利用网格技术构建的分布式仿真应用平台,使分布式仿真的执行能够利用网格所提供的强大的共享资源。从概念上讲,仿真网格就是执行仿真任务的资源集合;从仿真用户的角度来看,仿真网格就是一个大系统,提供单个入口点,以访问强大而分散的仿真资源。在执行分布式交互仿真应用的过程中,仿真网格接受由用户提交的仿真任务,并根据资源管理策略将仿真任务分配到网格内适当的仿真节点上执行,当仿真任务执行结束以后,将仿真结果返回给用户。通过仿真网格,仿真用户可以一次提交若干仿真任务,而不必考虑它们在何处运行。

目前国内外对仿真网格的研究时间还不是很长,研究仿真网格的思想大都试图使用Globus的网格服务Grid Service,将HLA的服务包装成为网格服务,使其网格化;或者通过Grid Toolkit工具包构建仿真网格[1]。这样构建的仿真网格使得通信频繁的仿真应用在网格平台上,带来了额外的通信开销。本文提出了通过网格引擎SGE(Sun Grid Engine)构建仿真网格的方法,充分利用了SGE在集群中应用的优势,采用双通道通信机制,使仿真应用能够高效地利用网格的共享资源,同时保证了仿真联邦的频繁通

* 收稿日期:2005-05-10
基金项目:国家部委基金资助项目(51404010403KG0155)
作者简介:张传富(1973—),男,博士生。

信需求。

1 SGE网格引擎

SGE是由SUN公司开发的一个本地和集群级的网格框架,由该框架构成的网格环境是由许多协同工作的计算资源组成的,为某一项目或部门的用户提供单一的访问入口点,由网格系统帮助创建、管理、执行应用[2]。此外,系统还要检索用户的身份以及与项目或用户组的从属关系。

SGE系统框架是由四种类型主机构成,分别是:主控主机、执行主机、管理主机、提交主机等[3]。框架图如图1所示。

(1)主控主机是所有网格节点活动的中心,进行网格的管理和调度活动。它主要运行网格主控守护进程sge_qmaster和调度守护进程sge_schedd。

(2)执行主机是有权执行SGE作业的节点。主要运行网格执行守护进程sge_execd。

(3)管理主机是由主控主机赋予管理权限的主机,其执行任何种类的SGE系统管理活动。

(4)提交主机是由主控主机赋予的作业提交权限的主机,通过该主机用户可以向网格提交执行作业。

(5)影像主控主机是在主控主机或主控守护程序出现故障时,行使主控主机功能的主机。

在SGE网格中,一台主机可以作为多种类型的主机使用,例如主控主机也可以作为管理、提交、执行主机使用。

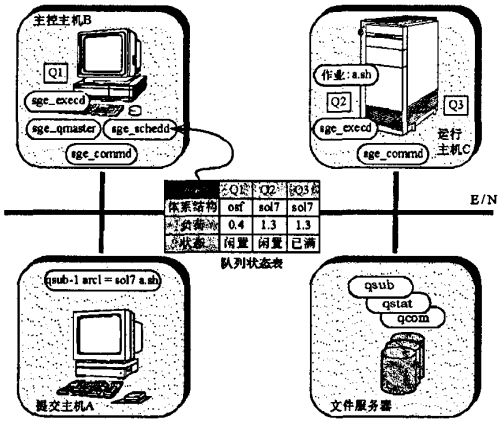

图1 SGE网格系统框架示意图
Fig.1 Framework of the sun grid engine system

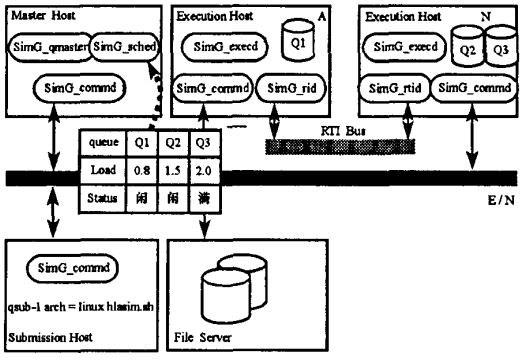

图2 仿真网格框架示意图
Fig.2 Framework of the simulation grid

2 基于SGE的仿真网格原型系统设计

基于SGE的仿真网格原型系统是以SGE为基础根据分布式仿真应用的特点进行设计的。在系统中对SGE的框架进行了改进和扩展,利用SGE集群管理的优势,协同使用网格中共享的计算资源,为用户提供单一入口点,负责创建、管理和执行用户提交的作业;同时对SGE的功能进行了扩展,使得网格平台适合仿真任务的执行。主要扩展部分是对基于HLA/RTI的分布式仿真任务和并行仿真任务的支持。同时考虑到了仿真网格平台的高可靠性,仿真网格也提供了对用户级检查点仿真程序的支持,和针对某些操作系统的内核级检查点仿真作业的支持,使得点检查作业可从一个仿真节点迁移到另一个仿真节点,而不需要用户干涉。此外,仿真网格还提供了监视和控制仿真任务的工具,使管理员能够对仿真任务的运行进行管理。

仿真网格框架主要是对SGE原有的框架进行了改进和扩展,相应地由五个守护进程构成,其他的命令都是从这些守护进程衍生出来的,客户端的命令也主要是与这些守护进程进行交互,完成作业"提

交-调度-执行-返回"整个过程。如图2所示。

2.1 SimG_qmaster 守护进程

该进程模块负责整个仿真网格环境中的所有仿真节点的管理和调度,是整个仿真网格的中心,主要维护仿真节点的主机表、仿真队列表、仿真作业表、仿真节点系统负荷表以及仿真用户权限表等。该进程对 sge_qmaster 的改进主要是根据 HLA/RTI 仿真任务的特点对仿真主机信息和队列信息进行了扩展。该守护进程与 SimG_schedd 互相配合,完成对仿真任务的调度。

SimG_qmaster 进程包含了当前的最新数据,包括仿真作业的状态、队列状态,以及其他网格运行所必需的对象状态等。网格中与 qmaster 的通信是通过 Grid Engine Database Interface(GDI)实现的。schedd 通过使用 GDI 事件驱动的更新协议从 qmaster 进程获取所有的信息。在整个过程中,qmaster 进程与 schedd 进程协调工作,为仿真作业定制所需要的资源,使作业顺利执行。

2.2 SimG_schedd 守护进程

该进程主要负责仿真网格中的任务调度。由于仿真网格平台主要是对基于 HLA/RTI 仿真任务进行调度,因此该守护进程也主要是针对这种应用进行设计实现的。在该守护进程中,实现了三种针对 HLA/RTI 仿真应用的调度,包括:以进程为单位的调度、以联邦成员为单位的调度和以联邦为单位的调度。完成调度以后,该守护进程将调度信息转发给 SimG_qmaster,并请求适当仿真节点的 SimG_execd 进程进行处理。该守护进程在 SimG_qmaster 的帮助下,维护仿真网格所有节点最新状态视图。它所进行的调度主要是决定将仿真作业分配到合适的仿真队列中,然后,守护程序将这些决定转发至 SimG_qmaster,后者将启动所需的操作。

2.3 SimG_execd 守护进程

该守护进程主要负责仿真节点上仿真任务的执行。该进程通过 SimG_shepherd 进程启动 HLA/RTI 仿真任务的执行,并在执行过程中接受仿真网格主控主机的控制,使仿真作业可以暂停/重启、取消、重新调度等,甚至根据仿真用户的需要进行仿真作业的迁移等操作。在该进程中包含了两个模块 PDC(Portable Data Collector)和 PTF(Priority Translation Facility)。PDC 是 SimG_execd 内部的模块,负责收集运行作业的信息,例如 CPU 的使用情况、内存的使用情况等等,按照每一个作业的相同标准从作业的所有进程中收集数据。PTF 模块的作用就是根据仿真作业的情况设置仿真作业的优先级,并重新设置作业所有进程的优先级。

2.4 SimG_commd 守护进程

由于 HLA 仿真任务之间的通信都是通过 RTI 进行的,仿真网格的通信模块需要兼顾网格通信与仿真任务通信两个方面。因此,在仿真网格的设计中提出了双通道通信的方式,一个通道由该守护进程负责,另一个通道由 SimG_rtid 负责。

SimG_commd 守护进程运行在每台执行主机和主控主机上,通过 TCP/IP 端口进行通讯。所有的 SimG_commd 网络组成了仿真网格控制调度信息通信的主干。在仿真网格中采用了 SGE 将网格组件与通信层分离的方式,通过两个独立的通道完成仿真网格的通信。仿真网格的其它组件与 SimG_commd 的通信是通过 SimG_commlib 库实现的,SimG_commd 与 SimG_commlib 一起实现了仿真网格中的组件与通信层的分离,并负责网格中组件之间的通信。

2.5 SimG_rtid 守护进程

该守护进程负责仿真网格双通道通信的另一个通道,主要是提供 HLA 仿真任务之间的互联、互通和互操作。该守护进程由执行主机节点上的 SimG_execd 守护进程负责启动,主要包括一个符合 HLA 规范的 RTI 进程和一个 RTI 进程管理器。RTI 进程要求在支持传统的联邦成员之间的通信交互之外,还要支持集群或高速网络中并行联邦成员的通信交互。RTI 管理器负责收集 HLA 任务执行过程中 RTI 的信息,以及 RTI 的错误报告。RTI 进程负责用户提交的 HLA 任务之间根据配置文件的连通。

RTI 进程支持集群或高速网络的通信交互,主要提供执行节点多 CPU 的共享内存的通信、高速网络

的通信(例如 Myrinet 网络)等[4]。

3 仿真网格中的任务调度

仿真网格中存在不同类型的资源,包括软硬件资源、网络带宽资源,以及各种仿真数据资源和模型资源等等。仿真网格的主要目的是使仿真应用能动态地管理和利用网格资源以提高仿真性能。仿真网格除了支持 SGE 固定的批处理作业、交互式作业、阵列作业以外,主要是对基于 HLA 的仿真任务应用提供支持。基于 HLA 的仿真任务都是以联邦(Federation)的形式进行的,联邦又由若干相互作用的联邦成员(Federate)构成。由联邦成员构建联邦的关键是要求各联邦成员之间可以互操作。因此针对仿真网格中仿真用户提交的仿真任务,需要针对 HLA 仿真应用的特点结合网格资源的管理情况进行调度。根据仿真用户提交任务的粒度不同,可以进行三种方式的调度。

3.1 以进程为单位进行调度

这种调度模式是仿真作业在网格调度中的特殊应用,主控主机以进程为单位对仿真作业进行调度。根据进程对执行节点的软性资源和硬性资源需求(包括内存、CPU、I/O 带宽、文件空间、软件许可证),将仿真进程调度到不同的执行主机节点上,由主机中相应的队列负责仿真进程的执行,实现仿真网格负载平衡和运行管理的功能。当某个节点性能下降到一定程度或是不能提供仿真进程运行所满足的资源条件时,主控守护进程和调度守护进程负责对该进程的状态进行保存,并通知属于同一个联邦成员的其他进程暂停运行,同时通知其他联邦成员,然后暂停整个仿真的执行。等待将该仿真进程迁移到其他符合要求的节点上之后,再使整个仿真在原来的状态下继续运行。

这种调度模式适合用户以仿真进程为单位进行提交执行的情况。将属于同一个仿真联邦成员的进程由主控调度守护进程进行编组,设定整个联邦运行的时间段,使得同属于一个联邦成员的进程在联邦运行之前做好准备,在规定的时间段内使整个联邦成员一同运行。以进程为单位的调度示意图如图 3 所示。

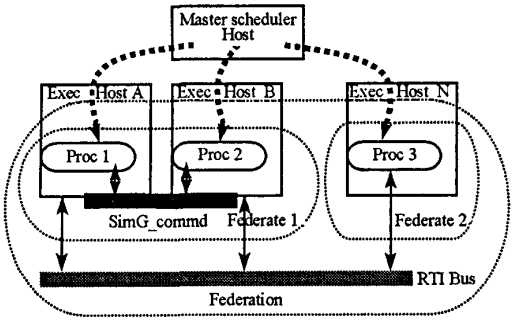

图 3 以进程为单位的调度示意图
Fig.3 Process-level task scheduling

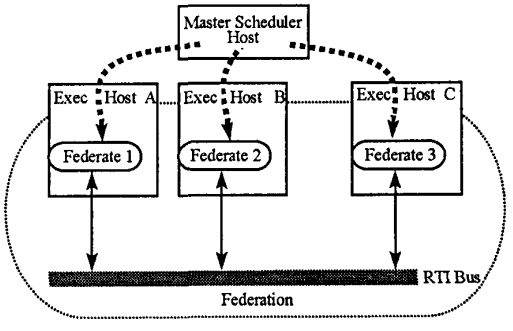

图 4 以联邦成员为单位的调度示意图
Fig.4 Federate-level task scheduling

3.2 以联邦成员为单位进行调度

这种调度模式中,仿真联邦成员与传统意义上的联邦成员一致。仿真网格中的主控主机将仿真任务以联邦成员进程为单位进行调度,根据成员对执行节点的软性资源和硬性资源需求,将仿真联邦成员调度到不同的执行节点的成员执行队列上,实现仿真网格的负载平衡和运行管理的功能。当某个节点性能下降到一定程度或是不能提供联邦成员运行所满足的资源条件(例如 RTI 的运行出现了异常)时,主控守护进程和调度守护进程负责该成员和本地 RTI 的状态保存,并通知其他一起运行的联邦成员,然后暂停整个仿真的执行。等待该成员迁移到其他符合要求的节点上后,再使整个仿真在原来的状态下继续运行。

如果联邦成员由多个进程构成,则将该联邦成员调度到执行节点的联邦成员队列中,由该联邦成员的队列负责派生联邦成员进程,使得本联邦成员进程之间保持协调一致,本联邦成员通过本地执行节点

的 RTI 与其他联邦成员实现交互。

这种调度模式适合用户以联邦成员的形式提交的仿真任务的情况。用户提交的多个联邦成员可以分别通过这种模式进行调度,然后在规定的时间段,属于同一个联邦的联邦成员一起执行,完成仿真任务,并将仿真结果返回用户。调度过程如图 4 所示。

3.3 以联邦为单位进行调度

在这种调度模式下,包含多个联邦成员的联邦作为调度单位,由主控调度守护进程进行调度。在仿真网格中,以联邦为单位的调度是由专门的联邦执行队列负责执行的。

联邦队列通过联邦生成器(spawner)[5],进行联邦成员的调度。

首先联邦队列启动联邦生成器(spawner),该 spawner 称之为 master _ spawner。然后 master _ spawner 在其他适合执行联邦成员的节点复制自己,其他节点的 spawner 被称为 slaver _ spawner。在联邦中由 spawner 根据联邦执行过程中需要的信息和数据,设置本地联邦成员的执行环境和和 RTI 执行环境。设置完毕以后,由 spawner 负责启动和监控本节点联邦成员的执行状况,并将执行状况汇集到 master _ spawner,由 master _ spawner 将联邦的执行情况和执行结果通过主控调度进程返回给用户。在联邦执行的过程中 master _ spawner 将一直等到所有的 slave _ spawner 都执行结束后才结束执行。

这种调度模式适合于用户以联邦为单位提交仿真网格运行的情况。由于联邦包含多个联邦成员,因此,一般是将联邦成员数据事先存放到网格的文件服务器中,由用户提交联邦执行脚本,调度程序从文件服务器中获取数据并将其调度到指定的执行主机。调度过程示意图如图 5 所示。

4 仿真网格原型系统实现

在仿真网格原型系统中的 SimG _ rtid 守护进程使用了两种类型的 RTI,一个是开放源码的 CERTI,中心 RTI 是通过网关 RTI rtig 来实现的。另一种是采用了支持并行和集群仿真的 FDK(Federated Simulations Development Kit)中的 BRTI 和 DRTI 来实现[5,6]。

当用户需要在网格平台上执行基于 HLA 的仿真任务时,任务执行过程如图 6 所示。

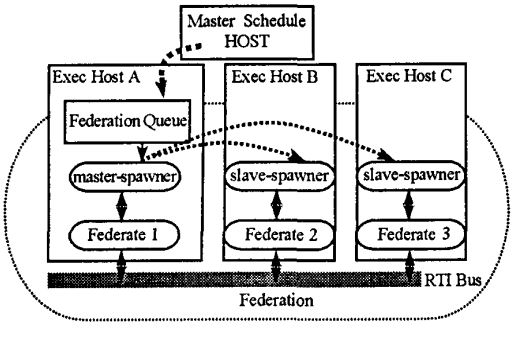

图 5 以联邦为单位的 spawner 调度示意图

Fig.5 Spawner federation-level task scheduling

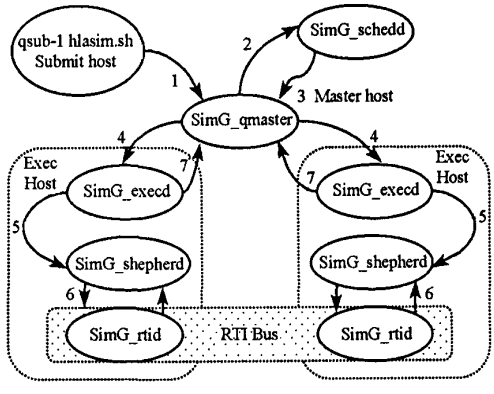

图 6 仿真网格作业执行过程

Fig.6 Job execution process in simulation grid prototype system

(1)用户通过提交主机向仿真网格提交仿真任务,例如:qsub - larch = linux hlasim.sh;

(2)主控主机的守护进程 SimG _ qmaster 首先检查所提交用户的权限,查看该用户可以使用的网格资源,并将该用户所提交的作业交给 SimG _ schedd 守护进程进行调度。

(3)调度守护进程 SimG _ schedd 根据作业的资源需求和网格系统的可使用资源状况,以及所有执行主机的负载报告对用户提交的仿真作业进行调度。如果当前没有满足该作业运行的队列,则暂时将该作业置于等候区进行等待,当有满足作业需求的队列空闲时,根据队列所在的执行主机的负载状况,将该作业分配给负载最轻的队列。

(4)主控守护进程 SimG_schedd 将根据调度信息,将用户提交的仿真作业发送至队列所在的执行节点上(HLA 仿真应用一般包括多个联邦成员,可能需要多个仿真执行节点),并进行作业执行前的配置,包括环境变量、执行环境、配置文件等。

(5)执行节点的守护进程 SimG_execd 根据 SimG_qmaster 的作业分配和作业设置,调用 shipherd 进程执行仿真作业,并随时将该节点的负载情况向主控守护进程 SimG_qmaster 报告。

(6)作业执行进程 SimG_shepherd 首先进行该仿真应用的 RTI 配置(主要是在给定的目录中设置 HLA 仿真应用的联邦数据执行文件),然后启动 HLA 仿真作业的执行。

(7)执行节点的守护进程 SimG_execd 将仿真结果返回给主控守护进程 SimG_qmaster,由主控守护进程将结果递交给用户。

仿真网格原型系统采用了一台主控主机同时作为提交主机和管理主机,五台执行主机节点构成的,并对 HLA 仿真应用的三种调度方式进行了简单的实现。主机节点所使用操作系统均为 RedHat Linux 9.0。

5 结 论

以 SGE 框架为基础,进行了仿真网格的原型系统的设计和实现。提出了仿真网格的双通道通信机制和对 HLA 仿真任务的三种调度模式,并初步实现了仿真网格的原型系统。通过该原型系统,仿真网格能够接受仿真用户的仿真任务,根据仿真作业的组织调度,完成仿真应用。下一步的工作是对并行 RTI 进行进一步的开发,使其满足集群或高速网络上运行 HLA 仿真联邦的通用性。进一步对 HLA 仿真任务的调度模式进行深化。另一方面需要对仿真网格中数据存取和仿真模型资源的共享进行研究。

参 考 文 献:

[1] Zaajac K, Tirado-Ramos A, Zhao Z, et al. Grid Services for HLA-based Distributed Simulation Frameworks[A], First European Across Grids Conference, Santiago de Compostela, Spain, Springer-Verlag, Heidelberg, 2003:147-154.

[2] Sun Microsystem Inc., Sun(tm) Grid Engine 5.3 Administration and User's Guide[R], Santa Clara, CA 95054 U.S.A. 2002.

[3] Sun Microsystem Inc., Sun Grid Engine 5.3 and Sun Grid Engine[R], Enterprise Edition 5.3 Manual Pages, Santa Clara, CA 95054 U.S.A. 2002.

[4] Perumalla K S, Park A, Fujimoto R M, et al. Scalable RTI-based Parallel Simulation of Networks[A]. Parallel and Distributed Simulation, 2003.

[5] Steinman J. The Standard Simulation Architecture[A]. In proceedings of the 2002 Summer Computer Simulation Conference, 2002.

[6] Fujimoto R. Parallel and Distributed Simulation Systems[M]. John Wiley & Sons, Inc., 605 Third Avenue, New York, NY 10158-0012, 2000.

[7] Fujimoto R, Hoare P. HLA RTI Performance in High Speed LAN Environments Technical Report[R], College of Computing, Georgia Institute of Technology, 1998.