基于OGSA的网格工作负载监测系统研究

刘晓明，饶 翚

(解放军理工大学指挥自动化学院 南京 210007)

【摘要】为对网络系统进行实时管理并适应其动态变化，需要对其工作负载进行监测，因此提出网格工作负载监测体系结构。该体系结构以端到端的方式对基于OGSA的网格构件的工作负载进行监测，以便进行分类和确立相互关系。通过设计的监测点算法，建立Petri网模型，该模型能自动收集工作单元的数据，为网格用户提供监控数据信息浏览，有利于用户对系统性能进行有效管理。在建立上述体系结构的基础上，提出了未来扩展系统功能方面的思路，以适应网格环境的复杂性，减少通信延迟和通信代价，为用户更好地监测网格系统提供方便。

关 键 词 网格； 监控； 开放网格服务结构； Petri网； 工作负载
中图分类号 TP393.9 文献标识码 A

Research on OGSA-Based Grid Workload Monitoring System

LIU Xiao-ming，RAO Hui

(Institute of Command Automation, PLA University of Science and Technology Nanjing 210007)

Abstract In order to manage Grid system in real time and adopt its dynamic changes, a workload monitoring infrastructure is proposed. The infrastructure classifies and establishes correlation for workload across components in Grid based on the Open Grid Service Architecture (OGSA) in an end-to-end manner. A Petri net model based on monitor point algorithm is constructed automatically to collect data of work units and provide monitor data information for users to explore. and manage system performance efficiently. On the base of the proposed infrastructure, some ideals to extend system function are put forward to satisfy the complexity of Grid and decrease communication delay and cost.

Key words grid; monitor; open grid service architecture; Petri net; workload

典型的网格环境是由多种元素组成的动态异构分布式环境。网格通过中间件等技术隐藏其复杂性，可为用户的操作带来便利。这种情况在系统运行良好时是可行的。但是，当系统出现故障，用户需要准确了解故障的来源并及时做出补救措施时，则存在弊端。掌握网格工作负载变化是网格调度和应用性能分析与预测的基础[1]，因此网格的监控是非常有必要的。

网格监控用来度量和显示网格组件在某个时刻的状态。为了保证有效监控，监控必须是“端到端”的。在应用端的所有组件都需要监控，监控的内容包括组件的状态信息、运行时间信息等。监控这些信息有助于对网格性能的判断。

基于开放式网格服务体系结构(Open Grid Service Architecture，OGSA)的网格[2]是一种标准的网格体系结构。该体系结构具有基于Web服务的特性，可以实现资源的动态调度和管理。本文基于该体系结构提出一种工作负载监控的体系结构，实现分类和跟踪工作单元，使用收集的数据自动构建响应时间，服务Petri网模型。建立基于OGSA的网格中间件，使用开放标准的监测方法，能在异构环境和多平台环境下使用。

1 监测体系结构的设计

网格监测系统需要解决的主要问题是：工作负载变化频繁，信息更新应能及时反应；由于传递的信息量大，监控信息的系统资源占用率要控制在合理的范围内；性能度量的系统花销要实现最小化；用户要能够通过监测系统实时和直观地了解系统运行状况。

典型的监测系统包括信息生产者、信息中介者和信息消费者三个部分。生产者采集相关信息，发布到具有目录服务设施的中介者。消费者根据不同的需要查询、使用、订阅特定信息。订阅信息可以

收稿日期：2006 – 06 – 20

作者简介：刘晓明(1956 –)，男，教授，博士生导师，主要从事计算机仿真方面的研究；饶 翚(1978–)，女，博士生，主要从事系统仿真、网格理论方面的研究.

定时发送给消费者，在生产者更新信息后，中介者也需做出相应更新，并提供给消费者。也就是说，消费者既可以主动获取信息，也可以被动获取信息。

根据对监测系统的基本要求，对监测系统进行设计。图1是监测体系结构的模型，主要包括代理模块、发布-订阅框架[3]、数据库、Web服务器和客户浏览器五个组成部分。

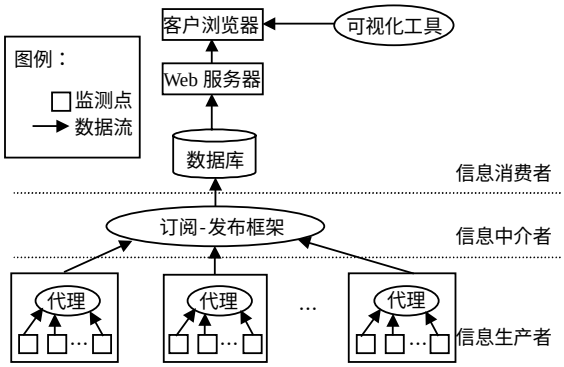

图1　基于OGSA网格工作负载监测的体系结构

在代理模块中，为基于OGSA的网格中间件(如Globus[4]客户端、Tomcat[5]和Globus等)提供监测点MP(Monitor Point)，该监测点位于平台终端，可以监测网络构件的运行，如图2所示。

图2　网格中间件结构

在网格中，每个机器上可能有一个或多个平台，具有一个或多个监测点。每个机器上都有一个代理，监听MP来获取数据，并将它们集中起来。这些数据传送到发布-订阅框架。该框架作为信息中介者，为信息消费者提供网格资源的工作状态信息。每个工作单元为它的登录MP设立唯一的ID。用户或管理者可以查询工作单元。查询可以触发数据库中相关数据的修改。Web服务器完成业务逻辑，实现客户与数据库之间的对话，客户浏览器上的可视化工具方便用户对查询结果的直观监测。

2　监测系统关键技术实现

2.1　监测点算法

监测点收集工作单元的相关数据。工作单元有请求和响应两种状态。在MP监测工作单元的处理中，主要完成服务分类、生成不同服务之间的关系和测算运行时间三个任务。最后将得到的测量数据信息发送给代理。工作单元中监测点算法的流程图如图3所示。

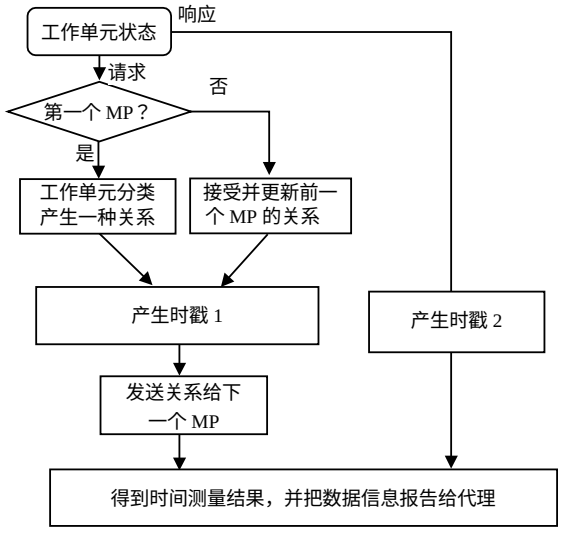

图3　监测点算法流程图

2.2　服务分类

当工作单元进入网格，它们根据各自的属性分成不同的服务类(如所有权、类型、调用参数等)。服务类代表工作的优先级和响应工作单元的时间目标。MP上的服务类能实现自动的服务层次验证。服务类用于为工作单元提供唯一的ID。

2.3　建立相互关系

工作单元的ID是用来通信的。在MP上的信息被收集好之后，要对数据建立相互关系。另一方面也是为了推导服务之间的关系。推导服务关系是自动的，不用知道服务的逻辑知识或源代码。推导调用关系是直接的。

在工作负载的处理中，整个系统的性能不仅由单个服务的运行时间决定，也由它们所组成的整个响应时间决定。前者很容易测定，后者由工作单元的并行性决定。如果两个或两个以上的服务是顺序调用的，整个响应时间就是所有服务花费的时间之和。如果有些服务调用是并行的，总的消耗时间将减少。服务之间的关系主要有两种：

(1) 调用关系 $\mapsto^{*}=\{\langle a,b\rangle \mid a\in S\wedge b\in S\wedge a\neq b\}$，其中$S$为服务集合；$\langle a,b\rangle$ 为在一个工作单元中a调用服务b；a为b的父服务；b为a的子服务。

(2) 依赖关系 $\Rightarrow^{*}=\{\langle a,b\rangle \mid a\in S\wedge b\in S\wedge a\neq b\}$，其中$S$为服务集合；$\exists c\in S:\langle c,a\rangle\in\mapsto^{*}\wedge\langle c,b\rangle\in\mapsto^{*}$ 为服务c调

用服务a和服务b；$\langle a,b\rangle$ 为服务a依赖于服务b，b的输出是a的输入，b为a的先决服务，a为b的依赖服务。

如果服务调用一个或多个子服务，该服务应该在所有的子服务启动前启动，在所有子服务终止之后结束。子服务可以是并行的，也可以是顺序的。如果一个服务依赖一个或多个先决服务，那么该服务要等所有的先决条件都完成了再启动。

在定义了服务之间的实时关系之后，可以为工作单元采用Petri网[6]来编码和表示这些关系，其原因在于：

1) Petri网能建模各种并发关系，如并行、顺序和同步等关系。

2) Petri网的相关理论完善。如能够从一个随机产生的Petri网[7]中推导端到端的响应时间。

3) Petri网标记能自然映射到工作单元，Petri网就可以直观的显示工作单元在网格中的工作状态。假设每个工作单元需要响应时间和服务尽可能快地调用，定义Petri网为四维空间(P,Γ,Δ,W)，其中：

(1) P是库所(space)的集合，表示服务前部和后部，一个工作单元的服务前部表示服务请求，服务后部表示服务响应。

(2) Γ 是一个函数，$P\to N$ (正数)表示在库所中的时间标记成为有效的标记，从而计算花费在服务前部/服务后部的时间。

(3) Δ 表示迁移集合，运行一个迁移，除了表示进入/离开Petri网的源/目的的迁移，还能产生四种信息：(1) 没有子服务，仅返回服务的调用信息。(2) 没有先决服务的子服务同时调用信息。(3) 没有依赖关系的子服务同步调用信息。(4) 有依赖服务并同时调用这些服务的子服务的同步调用信息。

(4) $W\subseteq(P\Delta)\cup(P\Delta)$是连接迁移和相关库所的有向弧的集合。

图4描述了两个用户竞争资源的Petri网图。状态用圆圈表示，变化用↓表示。标记是包含在状态中的一种标志(用黑点表示)，用来描述Petri网的状况，通过标记的流动来模拟实际系统的动态运行行为。资源有busy和idle两种状态，并在两个状态之间转换。

用户有active、requesting和accessing三种状态。用户周期性地重复这三种状态，对三个用户状态的修改分别为request、start和end。根据用户状态的变化，资源也发生相应变化，其状态的变迁由箭头的走势来表示。

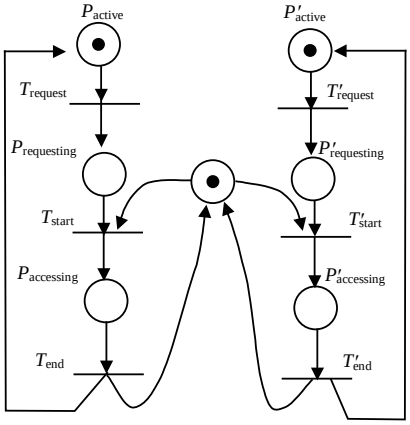

图4 两个用户竞争资源的Petri网图

3.4 测算响应时间

联合P和 Δ，Petri网可以形成一个加权的有向图。在有向图中的最长的加权路径决定整个响应时间，该路径称为关键路径。在路径比较复杂的情况下，可以使用Dijkstra 算法[8]确定关键路径C。算法的复杂度为$O(|P\times\Delta|^{2})$。整个运行时间的计算公式为$t=\sum_{i=1}^{[C]}t_i=\sum_{i=1}^{|C|}(t_i^F+t_i^R)=\sum_{i=1}^{|C|}\sum_{j=1}^{|M_i|}(t_{ij}^F+t_{ij}^R)$，其中 t_i 是第i个服务的运行时间；t_i^F 和t_i^R 表示在第i个服务前部和服务后部的运行时间；t_{ij}^F、t_{ij}^R 与t_i^F 和t_i^R 具有相似的意义，但是针对的是某个服务由多个平台共同完成的情况；j 表示第 j 个平台；M_i 是服务 i 的平台集合。

3 提高监测能力的途径

本文所建立的模型，实现了网格监测体系结构的基本功能，为提高系统的健壮性和实用性，还要考虑以下的方面，以提高监测效率，减少通信延迟。(1) 在Petri网模型中，需对收集数据进行分析处理，运用数据挖掘和机器学习技术，监测各种模式(如路径、构件等差异)产生的延迟，使Petri网模型能运用于随机网络。(2) 为增强监测的可适应性，对不同的服务和平台能够动态调整监测策略，需采用不同的监测粒度和测量方法，改变采用统一策略带来的资源浪费，实现监测点的负载平衡和优化调度。(3) 在网格环境中，工作单元可能发生迁移，因此需要建立工作单元和监测点之间的映射，保证建立正确的上下文关系。(4) 在网格系统出现故障或错误时，能实时保存工作单元的监测点状态，当故障或错误恢复后，监测点读取保存信息，继续收集工作单元信息。

(下转第857页)

安交通大学出版社, 2000: 116-119.
[9] ANDREW S T. Computer network[M]. 4th ed. 北京: 清华大学出版社, 2005.
[10] DOUGLAS E C. Internetworking with TCP/IP VOL I: principles, protocols and architectures[M]. 4th ed. 北京: 电子工业出版社, 2005.

编　辑　张　俊

(上接第826页)

4　总　结

本文提出了一种基于OGSA网格的工作负载监测体系结构。在网格的运行中，通过对工作单元的分类和端到端的跟踪监控，实现对工作单元状态的报告和运行时间的计算。为了增强可移植性和减少应用程序代码的改变，采用了基于OGSA的网格中间件结构。

在该体系结构中，还建立了Petri网模型，用于理解各服务之间的关系，便于准确计算运行时间。该模型能自动收集工作单元的数据，为网格用户提供监控数据信息浏览，有利于用户对系统性能进行有效管理。

在建立上述体系结构的基础上，提出了未来扩展系统功能方面的思路，以适应网格环境的复杂性，减少通信延迟和通信代价，为用户更好地监测网格系统提供方便。

参 考 文 献

[1] FOSTER I, KESSEHMAN C, NICK J M, et al. Grid services for distributed system integration[J]. Computer, 2002, 35(6): 37-46.
[2] FOSTER I, KESSEHMAN C. The grid: Blueprint for a new computing infrastructure[M]. San Francisco: Morgan Kaufmann Publishers, 1999.
[3] SANDHOLM T, GAWOR J. Globus toolkit 3 core-a grid service container framework[EB/OL]. http://www-unix.globus.org/ toolkit/ 3.0/ogsa/docs/gt3_core. PDF, 2005- 05-19.
[4] DAVID R, ALLA H. Petri net and grafcet: Tools for modelling discrete event systems[M]. New Jersey: Prentice Hall, 1992.
[5] DINGLE N J, HARRISON P G, KNOTTENBELT W J. Response time densities in generalized stochastic Petri net models[C]//In WOSP ’02: Proceedings of the Third International Workshop on Software and Performance. Rome: ACM Press, 2002.
[6] BERTSEKAS D. Dynamic programming and optimal control[M]. Massachusetts: Athena Scientific, 1995.
[7] GOODWILL J. Apache jakarta tomcat[EB/OL]. http: // jakarta.apache.org/tomcat/, 2005-06-10.
[8] HAPNER M, BURRIDGE R, SHARMA R, et al. Java messaging service specification[R]. Sun Microsystems, 2002.

编　辑　熊思亮