基于网络断层扫描的网格网络性能测量分析

王 伟 蔡皖东 李勇军
(西北工业大学计算机学院 西安 710072)

摘 要 网格计算通过网络连接来获得一个高性能和高效的计算平台。网格网络的监测和性能测量为网格性能分析、负载平衡、任务调度等提供了重要的科学依据,而成为大规模网格服务的关键组件。现有的几种网格监测方法因缺乏对监测数据的推断分析而无法对网格网络的性能进行测量。通过对网格网络性能测量的特点、GloPerf及传统网络测量技术的分析,提出了基于网络断层扫描的网格网络性能测量方法。研究结果为网格网络性能的测量提供了新的途径。

关键词 网格网络,性能测量,网络断层扫描,推断分析

Analysis of the Grid Network Performance Measurement Based on Network Tomography

WANG Wei CAI Wan-Dong LI Yong-Jun
(School of Computer Science, Northwestern Polytechnical University, Xi'an 710072)

Abstract Grid Computation is a distributed parallel computing system, acquired a high performance, high throughout and high efficiency computing platform by network connecting. As a large-scale grid service's important component, Grid network's monitoring and performance measurement play a major role for grid performance, load balance and task scheduling. While there are several approaches for grid monitoring, surprisingly, a very little research effort aimed at analyzing the monitored data for getting Grid Network performance measurement. On the basis of analysis of characteristic of the grid performance and the drawbacks of the GloPerf tool and the traditional network measurement technologies, this paper proposes a novel Grid Network performance measurement theory based on Network Tomography. The results of this paper present new theoretic foundations for measurement of grid network's performances.

Keywords Grid network, Performance measurement, Network tomography, Inference analysis

1 引言

网格是继万维网之后出现的新型网络计算平台。网格计算是一种分布式的并行计算系统,它将不同地理位置的普通计算机通过网络组织成一个虚拟的超级计算机。网格计算具有灵活而可扩展的体系架构,它提供了解决计算密集型问题经济而有效的方法[1],Globus Toolkit已成为网格计算事实上的标准[2]。

尽管网格环境在理论上具有"无限"扩展性,但当网格的规模扩大时,性能恶化是任何网格服务所难以回避的。在数据传输期间,网络链路的负载、容量和可用性对网格应用程序的性能具有重要影响。如果网格环境缺乏网络性能数据,网格中间件和应用程序就不能适应变化的网络状况,致使网格难以支持其计算所需的服务层协议。尽管数据在网络上传输决定着应用程序的效能,但网格网络(Grid Network,GN)的性能及其可靠性具有极端的异构性。这使得网格已有的几种监控的方法[3]难以对监测数据进行有效的分析,如何测量GN性能就变得尤为重要:首先,GN系统是提供高性能通信的必要手段,通信能力的好坏对网格计算提供的性能影响极大;其次,用户要获得延迟小、可靠的通信服务离不开高速网络;最后,由于缺乏对GN性能的监测手段,高性能计算系统的实际运算速度往往与峰值速度有很大的距离。

本文在对GloPerf及传统网络测量技术分析基础上,提出了基于网络断层扫描的GN性能测量方法,为测量GN性能提供了新途径。

2 GN性能测量的复杂性

尽管目前的系统主要针对计算资源(CPU周期+内存),但网格系统可运行于更广的资源诸如存储器、网络等之上。因而GN中的虚拟资源池是动态多样的,其中的资源随用户的意志可在任意时刻加入或退出,资源的数量可达上千个或更多,其性能或负载能够随时间动态地变化。而用户(或用户Agent)几乎无法预知这个资源池的实际类型、状态和特点,需要根据计算环境的改变在运行期间动态地进行调整。随着资源和相关策略的动态变化,观察、比较和分析性能要比普通的分布式系统复杂得多。另外,由于网格系统本质上的不可重复性,导致用户难以完全控制所有可用的资源和任务。为了给网格中间件和用户提供一个抽象的、同构的场景,全局网络的链路必须要用简单的、相关的度量及其基本的特性来测量。这就使GN的性能评估也随之变得复杂[3]:(1)网格应用程序在一个事先并不存在的虚拟机上执行,其确切的特性无法预知;(2)由于不断涌现的网格应用程序需要不同的性能要求,传统的度量和特性参数不能够或不足以表示GN性能;(3)GN的动态性具有不可再现性,性能调试和优化变得异常艰

王 伟 博士研究生,主要研究方向为计算机网络、网络信息安全等;蔡皖东 博士,教授,博士生导师,主要从事计算机网络、分布式计算、多媒体通信、网络信息安全与对抗等学科方向的研究。李勇军 博士研究生,主要从事计算机网络、网络信息安全等学科方向的研究。

难;(4)GN资源的多样性和异构性会对网格的观察、比较和分析带来显著的复杂性。

3 GloPerf和传统网络测量技术

GloPerf[4]能够灵活地部署并进行简单的端到端TCP测量,它在成对的IP地址之间进行周期性的网络性能测试,Globus利用netperf提供的库函数实现网络性能测试的功能。它提供了两种方式的测试:(1)使用netperf的TCP_STREAM测试获得两台主机间的通信带宽;(2)使用netperf的TCP_RR测试获得两台主机间的通信延迟。显然,GloPerf对GN的测量存在下面的缺陷:其一,对GN性能测量带来很大的额外负载。网格环境中存在成百上千个节点,测量过程会产生大量的通信和数据信息,无疑给网格应用产生极大的额外负载。其二,由于网格的动态性,这种测量的误差和测量的频率很难保持平衡:要提高测量的准确性,就要频繁地对GN进行测量,但随之产生繁重的网络负载会使测量结果失真。其三,由于GloPerf本身不包含任何推断模型,测量的数据缺乏分析依据,而建立基于GloPerf的数学推断模型是其面临的主要困难[5]。文[5]试图采用所谓的"上次测量"分析模型,分别通过计算GN性能测量数据MAE(平均绝对误差)和MSE(均方误差)来评估GN的带宽。这种方法的显著缺点就是个别的测量值对实际的真实带宽不具有代表性(如上次测量值有可能正好是实际带宽曲线的峰值)。

传统的网络测量技术有不同的分类方法[6,7]:根据测量方式,分主动测量和被动测量;根据测量点多少,分单点测量与多点测量;根据测量内容,分拓扑测量与性能测量。它们采用分布式测量结构,在网络内部的相关节点上通过测量代理采集有关测量数据(如丢包率、延迟等)并汇集到一个中心节点上,通过建立的适当数学模型对测量数据进行分析,从而实现网络系统的性能评估。这些测量技术均属于网络内部测量,与网络体系结构和网络协议密切相关,并且需要网络内部相关节点的密切协作。由于GN不仅在技术上而且在语义上不同于其他的分布式环境,传统测量技术不能直接用于网格的性能分析[3]。另外,这些测量技术基于传统网络而存在自身的缺点:首先,它们依赖于特定的网络协议,无法实现与网络结构和协议无关的测量;其次,它们依赖于自治系统内部节点之间的协作,出于网络安全和商业利益等原因,有些自治系统并不对外开放,难以实现内部节点的协作和信息交流;最后,它们是建立在网络拓扑结构比较稳定的网络基础上,难以实现对动态GN的测量。

4 基于网络断层扫描的GN性能测量分析

4.1 网络断层扫描技术

近年来,国际上提出网络断层扫描(Network Tomography,NT)[8,9]网络测量技术,将医学断层扫描(CT)思想引入到网络测量中,根据网络外部(网络端点或边界)的测量来分析和推断网络的内部性能和拓扑结构。NT技术使用路径和链路概念来描述网络内部节点之间的连接关系。路径是指两个节点之间的间接连接,它们之间包含一个或多个间接节点。链路是指两个节点之间的直接连接,它们之间不包含间接节点。在NT测量中,首先将根节点和多个叶子节点之间的一对多通信关系抽象成一个逻辑树,然后在网络边界节点上观测它们之间的通信行为并采集有关测量数据,如报文是否成功到达、到达报文数量和时延时间等。由于报文丢失和时延具有一定的随机性,这些测量数据并不能直接作为链路性能参数,需用统计学相关方法对测量数据进行处理,最后推断出网格内部链路延时和报文丢失率等网络性能。

因此,NT测量过程分两个阶段:一是数据测量阶段,建立测量模型,在网络端点或边界上观测和采集测量数据,它使用逻辑树形图来描述介于根节点(源节点)和多个叶节点(目的节点)之间经网络不同内部节点所进行的一对多通信;二是数据分析阶段,建立统计分析模型,运用统计学理论对大量的测量数据进行分析和评估,推断出网络性能和拓扑。NT测量技术的优点在于:它通过测量端到端的通信行为来推断网络内部性能,无需内部网络的任何协作,这不但降低了由于测量所带来的网络负载,并可实现与被测网络内部结构和协议无关的测量。由此可见,NT测量技术代表着网络测量技术发展的先进思想。

4.2 基于NT的GN测量架构

为了对GN的性能进行测量,可以将GN及其资源节点看作是一个逻辑的树型结构,各资源节点通过逻辑链路或路径来连接。GN的源节点发送探测数据包经共享路径上若干个节点的传送而到达目的节点。根据GN叶节点得到的测量值及节点间的关联关系,来推测GN内部每个资源节点的性能值。我们将GN测量问题描述为线性模型$Y=A\theta+\varepsilon$。其中,Y是个度量向量(如GN端到端延迟),A是GN路由矩阵,而θ是待估的参数向量(如平均延迟),ε是误差向量。为了简化问题,忽略ε并用θ的向量函数X代替上式中的θ,可得:

$$Y=AX \tag{4.1}$$

为此,我们用$X_{*,j}$表示GN节点j处的性能测量值,V表示所有节点集合,测量所得到测量值集合为$X_d=(X_{*,j})_{j\in V}$,$\theta=(\theta_j)_{j\in V}$为网格网络中每个节点需要估测的性能参数。定义$p(x,\theta)=P\{X_d=x\}$,则有:

$$p(x^1,x^2,\cdots,x^n;\theta)=\prod_{m=1}^{n}p(x^m;\theta)=\prod_{x\in\Omega}p(x;\theta)^{n(x)} \tag{4.2}$$

在(4.2)式中,$x^1,x^2,\cdots,x^n$为n次测量结果,Ω是所有可能测量结果的集合,$n(x)$是测量结果中测量值为x的次数。这样,需要推测的GN性能参数值就可表示为求(4.2)极值:

$$\hat{\theta}=\arg\max_{\theta}\prod_{x\in\Omega}p(x;\theta)^{n(x)} \tag{4.3}$$

利用上述的思想并结合代理机制,建立图1所示的网格网络测量架构。

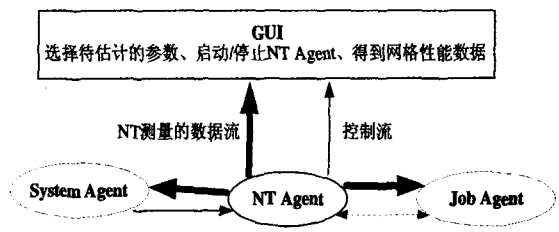

图1 基于NT的网格性能测量架构

在图1中,GUI是用户就某项Job对网格网络性能测量的界面;System Agent用来对GN环境的计算机资源进行管理;NT Agent启动Job Agent,得到GN各资源节点采集的性能数据(样本)并采用合适的推断分析理论对样本进行分析;Job Agent启动应用程序任务并为NT Agent提供必要的Job网格网络环境信息。

4.3 GN测量方式的选择

一般，被动测量不会带来额外的网络负载，但它难于部署，测量时要采集大量的数据，数据分析的计算量很大，尤其是它会威胁到数据的安全性，因而难以发挥 NT 的优越性。而主动测量能够根据不同的场景控制探测包（如流量特征、采样频率等），其灵活性更能体现 NT 的优势。主动测量可分为多播测量和单播测量。多播测量虽然在 GN 中传输的报文数量相对较少，给被测量 GN 增加较轻的网络负担，但它要求 GN 必须支持多播。由于网格资源的异构性，对网格网络的性能测量适合采用单播测量，其探测模式可采用包对（Packet Pair）以克服单播测量本身的局限性。针对 GN 环境的特点，GN 的测量不能独立于网格的基础架构（如网格数据的获取）[10]，利用网格传感器（一种服务 daemon）来进行数据的采集，对各节点的测量依靠 Globus 的统一部署[11]。

4.4 GN 性能推断分析方法

在 GN 的端节点上观测并收集测量数据后，就可进行数据分析，采用统计推断分析理论来推测 GN 的内部性能。推断分析理论直接关系到 GN 性能测量的效率和准确性，这也是 GN 测量问题的关键所在。

基于 NT 的 GN 性能测量，可以采用最大概似估计（Maximum-Pseudo Likelihood Estimation，MPLE）[12] 来求解（4.2）式中的极值。针对 GN 性能测量的复杂性，在（4.1）式中所有的 X 分量相互独立的前提下，我们采取如下的措施：把 GN 性能测量问题分解为若干简单的子问题并忽略其相关性，然后将这些子问题的边缘概率相乘得到概似函数（Pseudo Likelihood Function，PLF），从而得到（4.2）式。

上述的子问题是从 GN 通信矩阵 A 中的若干对行（Pair Rows）来选择（如图 2 所示）。如果用 S 表示 A 中所有可能的子问题的集合，对每个子问题 $s\in S$ 而言，由（4.1）式可得 $Y^s=A^sX^s$，X^s 是含 s 的待估计量，A^s 表示子路由矩阵（如（$i1$，$i2$），$i1$ 和 $i2$ 分别是 A 的第一、二行）。而对于每个 s，用 $l^s(Y^s;\theta^s)$ 表示边缘似然函数（含有 s 的参数 θ^s），就可得到 PLF 函数：

$$L^p(y_1,y_2,\cdots,y_N)=\prod_{n=1}^{N}\prod_{s\in S}f^s(y_n^s;\theta^s) \qquad (4.4)$$

（4.4）式的最大化表示参数 θ 的 MPLE 估计。因此，PLF 把对全局参数的分析转化为一些边缘概率的分析，它所处理的是许多简单的子问题计算，这样就解决了 GN 测量的复杂性。

显然，MPLE 可产生许多子期望值。利用这种迭代算法，通过选择合适的起点将会对此类问题进行快速的估计。设 $\theta^{(k)}$ 是在第 k 步所获得的 θ 估计值，迭代函数（目标函数）$Q(\theta,\theta^{(k)})$ 在第 $k+1$ 步进行最大化，如此反复，直至它达到 θ 的一个稳定值。这样，利用 Pseudo-EM Algorithm[8] 方法可使 PLF 取得最大值，于是：

$$Q(\theta,\theta^{(k)})=\sum_{s\in S}\sum_{n=1}^{N}E_{\theta^{(k)}}(l^s(x_n^s;\theta^s)\mid y_n^s) \qquad (4.5)$$

一个子问题由一颗仅包含两个叶节点的子树所组成，而一颗子树是通过选择矩阵 A 的两行来构成。如果用 I 表示叶节点的数量，那么子树的总数就是 $I(I-1)/2$。假设网络拓扑的多播树如图 2 左边所示，右边是子问题及其边缘似然函数的形成过程。

利用这种性能推断方法对 GN 的性能进行推断时，MPLEA 不但能够克服极大似然估计的计算复杂度，而且还保持了良好的统计效率[5]。其关键是构建 PLF：（1）每次也可选择 A 中的 3 行或更多行构建一个 PLF，但需保持计算复杂度和估计效率之间平衡，一般可以选择两行来进行。（2）理论上，可选择当前所有的对（Pairs）来构建 PLF，但为了克服计算难度，可以考虑选择一个子集，但需要对推断的准确性加以验证。（3）将分治法（Divide-and-Conquer）的思想和概似方法相结合适合于 GN 环境下大规模的 NT 推断。

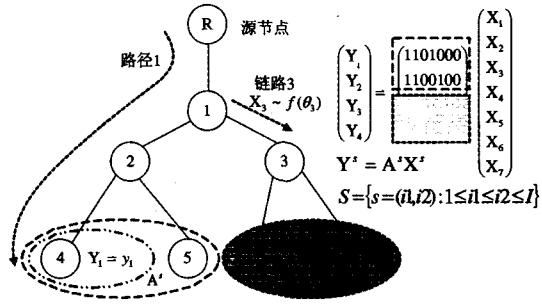

图 2 GN 性能的子问题是通过选择矩阵 A 的两行来确定，这对应于选择两个目的节点

总结和展望 网格计算将分布在不同地点的超级计算机用高速网络连接，并用网格中间件软件“粘合”起来，形成强大的计算平台。尽管国内外对并行和分布式系统的性能评估技术进行了许多研究，但它们并不能直接应用于 GN 性能的测量，大多数网格中间件只在某种程度上（缺乏全面性、准确性）对网格性能进行了监测，而 NT 技术的出现为 GN 性能的测量带来了希望。本文通过对 GN 性能测量的难点、现有的相关测量技术进行分析，提出了一种基于 NT 的 GN 性能测量方法，它有效地克服了网格环境的异构性和动态性，为 GN 性能的测量带来了新途径。然而，由于网格和网络断层扫描技术还处于发展之中，基于 NT 的 GN 性能测量尚需应用于实际的网格计算环境中做进一步的研究。

参 考 文 献

1 Baker M, Apon A, Ferner C, et al. Emerging Grid Standards. Computer. Published by the IEEE Computer Society, 2005, 38 (4): 43～50

2 Zhang X, Schopf J. Performance analysis of the globus toolkit monitoring and discovery service, mds2. In: Proceedings of the International Workshop on Middleware Performance (MP 2004), April 2004

3 N'emeth Zs, Gomb'as G, Balaton Z. Performance evaluation on grids: Directions, issues, and open problems. In: Proceedings of the Euromicro PDP 2004, A Coruna, Spain, IEEE Computer Society Press, 2004. 6

4 GLOPERF – Globus Network Performance Measurement Tool. http://www-fp.globus.org/details/gloperf.html, 2004

5 Peng Liang, See S, Jiang Yueqin, et al. Performance Evaluation in Computational Grid Environments, hpcasia, High Performance Computing and Grid in Asia Pacific Region. In: Seventh International Conference on (HPCAsia'04), 2004. 54～62

6 张宏莉，方滨兴，胡铭曾，等. Internet 测量与分析综述. 软件学报，2003，14(1)：110～116

7 林宇，程时端，邬海涛，等. IP 网端到端性能测量技术研究的进展. 电子学报，2003，31(8)：1227～1233

8 Castro R, Coates M, Liang G, et al. Network Tomography: Recent Developments. Statistical Science, 2004, 19(3): 499～517

9 Duffield N G, Presti F L. Network Tomography From Measured End-to-End Delay Covariance. IEEE/ACM TRANSACTIONS ON NETWORKING, 2004, 12(6): 978～992

10 Gerndt M, Wismüller R, Balaton Z, et al. Performance Tools for the Grid: State of the Art and Future. White paper. Shaker Verlag, 2004

11 Lee D, Dongarra J, Ramakrishna R. VisPerf: Monitoring Tool for Grid Computing. Lecture Notes in Computer Science, Springer Verlag, Heidelberg, 2003, 2659(1): 233～243

12 Liang G, Yu B. Maximum Pseudo Likelihood in Network Tomography. IEEE Trans. Signal Processing, 2003, 51(8): 243～253