

99,26(10)

1-5

元计算

分布计算

并行计算

计算机

计算机科学 1999 Vol 26 No. 10

面向 21 世纪的广域高性能元计算技术^{*}

Wide-Area High Performance Metacomputing for 21 Century

肖 依 卢宇彤 王意洁 卢锡城

(国防科技大学并行与分布处理国家重点实验室 长沙 410073)

TP301.6

86-
97-97

Abstract In order to solve the questions of scientific applications and commercial applications, many high-performance applications require the ability to exploit diverse, geographically distributed computing and data resources. These applications use high-speed networks to integrate supercomputers, large databases, archival storage devices, advanced visualization devices, and/or scientific instruments to form networked virtual supercomputers or metacomputers. The metacomputing, in fact, is trend of high performance parallel and distributed computing in the future.

Keywords Parallel and distributed computing, Wide-area computing, Metacomputing

一、引言

高性能并行与分布计算是关系到国家战略利益的核心高技术,也是国民经济发展的基础,是一个重大的基础研究问题。由于计算机微处理器的速度越来越高,网络通信技术的迅猛发展,网络规模的扩大和速度的提高,特别是 Internet 技术的兴起和广泛应用,网络已经交叉纵横整个世界。与此同时,人类的应用需求朝着高性能、多样性、多功能发展,许多大规模科学计算应用和复杂系统的应用不仅需要一台超高性能的计算机,更需要由多种机器组成、多个系统合作、多个科学仪器设备相连的网络虚拟超级计算机。这些应用要求将地理上分布、异构的多种计算资源通过高速网络连接起来,共同完成计算问题。这就决定了并行与分布计算技术将朝着基于高速网络的、面向大区域的高性能并行分布计算技术(简称元计算技术)发展,它将突然并行计算技术局限于局部范围的使用,将高性能计算融入到分布计算系统当中,提高分布式计算的性能,结合并行与分布处理技术,通过高速网络连接并集成地理上分布、异构的各种高性能计算机、大型数据存储系统、各种软件系统和科学仪器等,实现跨地域的、分布的高性能联合计算,共同完成一些重大的应用研究问题。通过利用全球网络的庞大计算资源,丰富的开发工具和友好的人机界面,以及各种不同性能和功能的

计算机和软件系统,进行高性能的并行与分布式计算,以解决宇宙学、生物学、生物化学、工程、流体力学、材料、神经和环境、核模拟等重大科学应用领域的计算问题,它不仅对国防建设具有重要的意义,也对国民经济产生重大影响,并从目前由 Internet 推动的面向信息计算应用发展到面向二十一世纪的计算应用。因此,研究基于高速网络的广域高性能元计算技术具有重大的理论意义和应用价值,是综合国力的重要体现,必将带动其它相关技术的大力发展,是下一个世纪初计算机的重大主流技术之一。

二、高性能并行与分布元计算技术的发展

高性能元计算技术将是下一个世纪高性能并行与分布技术的主流技术,其主要原因有以下几个方面。

1. 技术进步。一方面微处理器的主频已经达到 600MHz 左右,32 位跃至 64 位,而且仍旧以摩尔定律在飞速发展,桌面机、工作站、服务器等都具有相当高的性能;另一方面,网络通信技术迅猛发展,特别是 Internet 技术宣布了以网络为中心的信息时代已经到来,网络的速度也越来越高,特别是 ATM 和光纤技术的进步,建立高速和通信高质量的网络势在必行,这些也已经是各国目前正积极实施的事情。例如 Internet2、美国 DREN 网、NSF 的 vBNS 网络等都是以高速网络为目标,在下个世纪初我们可以看到高速网络

^{*} 本文得到国家高技术研究发展计划项目 863-306-ZT01-06-1 的资助。肖依 博士,副研究员,研究方向为并行与分布式处理技术;卢宇彤 讲师,研究方向为并行与分布式处理技术;王意洁 博士,助理研究员,研究领域为并行分布处理技术和数据库;卢锡城 教授,博士生导师,研究方向为 MPP 实现技术、分布处理技术和网络技术。

的广泛建立。以美国 NSF 的 vBNS 网络为例(见下表),1999 年局域网将达到 6.4Gb/秒,广域网为 2.4Gb/秒,2002 年广域网为 38.4Gb/秒。一旦如此高速的网络建成,人们便能够快速访问远程计算资源,能够更加方便地通讯和交互。

表 1 美国科学计算联盟局域网和广域网带宽表

vBNS	1998	1999	2000	2001	2002
局域网	0.8	6.4	12.8	15.6	51.6
广域网	0.644	2.4	9.6	19.2	38.4

2. 高性能需求的应用增加。技术的进步和人类需求的发展推动了许多应用要求得到高性能计算资源的支持,例如核模拟与仿真、经济模型的构造、全球天气预报、医学、生物工程学、宇宙学等都需要高性能计算的支持。这些应用不仅希望能够得到 MPP 技术的支持,还要求能在桌面机上访问到远程多种类别和功能的高性能计算资源,并能够操纵计算过程。

3. 空闲资源的利用。网络上的大量高性能 PC、工作站和服务器的资源和利用率据统计不超过 30%,剩下的 70%空闲资源未得到充分利用,这种巨大的计算资源应当采用各种技术加以利用。应当注意的是,不仅仅是硬件资源,而且大量的软件与信息资源也应得到共享。

4. 计算结果的大范围共享。大量的计算结果和大量的信息数据都存在大范围、异地共享的需求,这是一些大型科学应用的要求,也是大型复杂系统的合作计算要求。重要的是能够使各类研究学者能够跨地域相互交流,访问异地仪器设备、共享数据和计算资源,能够访问信息资源,共同在地域不同的条件下合作共同进行科学研究,共同解决科学问题。

5. 解决问题的新技术和新工具。科学计算已经成为人们解决科学问题的重要手段,可视化技术、虚拟现实技术、计算机仿真模拟等都为解决科学问题提供了新的、良好的方法,而这些都需要能够更易、更快地访问各种计算资源和共享信息和数据。

正是以上各个方面的发展促进了并行与分布计算技术朝着基于高速网络的大区域高性能并行与分布计算技术发展。它将是高性能网络、计算和高级软件技术的综合,是一种面向局域网和广域网范围的并行与分布计算技术。

三、元计算系统的特性

元计算系统不同于传统的分布式系统和并行系统,高性能的要求导致元计算的程序设计模型和接口不同于分布式系统,并行计算而非分布式系统往往希望精确地安排通讯以满足高性能需求,元计算的异构

性、动态性和区域性限制了并行计算的工具和技术在其中的应用。尽管,元计算系统与分布式系统和并行系统有所不同,但是它们有许多相似之处。元计算系统和分布式系统都强调能够集成处于不同管理域的、通过网络连接的各种计算资源,元计算系统的高性能计算要求和并行计算是一致的。因此,元计算系统结合了分布式系统和并行系统的特性,可以吸取分布式软件和并行软件技术,并通过新的技术和工具在高速网络上将各区域的计算资源连接起来,形成一个网络虚拟超级计算机。下面我们介绍一下元计算的主要特点。

1. 可扩充性和可选择性 元计算系统首先必须保证系统的可扩充性。广域网范围内的并行与分布式计算拥有庞大的计算机资源,可扩充性是一个基本的特点,任何应用都可以根据自己在连接特性、开销、安全和可靠性等方面的要求选择计算资源完成计算。

2. 多层次上的异构性 构成虚拟网络超级计算机的计算资源和网络连接常常是高度异构的,这种异构特性表现在各个层次上,从硬件设备到系统软件、调度策略、安全策略、使用策略等等,这是由元计算系统所面临的现实环境所决定的。

3. 不可预测的结构 传统上,高性能应用一般在单独一类系统上开发,很多特性是固定的、可知的。然而,元计算的应用要求在较广泛的环境中执行,这种环境由可适合的资源动态构成,导致了执行系统的结构和特性很难确定,动态变化,无法进行预测,同时地理上的分布和网络等的复杂性更加剧了这一点。

4. 动态的、不可预测的行为 传统的高性能应用系统经常使用空间共享或者组调度的调度策略,可以预测其对处理机和网络的访问,而在元计算系统中,资源特别是网络更可能是共享的,其结果是系统行为和性能随时间而变化,而且大规模的元计算系统也可能遇到资源和网络的失败,甚至无法保证基本的服务质量要求,这就导致无法预测应用系统的行为。

5. 多个管理区域 元计算应用所使用的资源不可能只被一个实体所拥有和管理,必然是由多个管理实体管理自身的计算资源。不同实体可能采用不同的管理机制,不同的验证机制、授权机制和访问机制使得本已令人头疼的网络安全问题更加复杂,这也要求元计算系统能解决在不同地点上执行用户提供的代码等问题。

四、高性能并行分布计算的应用

应用的驱动是技术发展的动力,以下我们介绍几种需要大区域高性能并行分布计算技术支持的主要应用。

1. 分布式超级计算应用 这种应用无法由单一

系统来解决,需要由多种机器,多个系统合作,甚至多个科学仪器设备相连的网络虚拟超级计算机来共同完成。它是一种计算密集型应用,网络带宽、延迟、调度、容错等是其关键技术,例如大型分布式可交互仿真系统以及复杂物理过程的精确仿真。这种计算系统也可以称为元计算系统,即将地理上分布、异构的各种高性能计算机、数据服务器、大型检索存储系统和可视化、虚拟现实系统等,通过高速互连网络连接并集成起来,共同完成一些缺乏有效研究办法的重大应用研究问题。例如,美国 SF-EXPRESS 工程基于 Globus 系统提供的服务,使用了横跨七个时区的九个地方的先进计算机,进行了十万辆独立机载物的军事模拟。

2. 高吞吐率计算的应用 这种应用通常由大量松散耦合的、独立任务组成,主要目标是调度大量的松散耦合的独立任务,利用网络上空闲、可用的计算资源完成多个任务的计算,从而达到完成整个应用。这种应用常常利用空闲的工作站系统,例如 AMD 公司的 K6、K7 芯片设计就采用了几千台计算机利用空闲时间进行计算,一些加密算法问题也采用了此类技术。美国威斯康辛大学的 Conder 系统就是支持此类应用的一个典型代表。

3. 按需求计算的应用 是指一些局部无法解决计算资源的应用往往要求提供短时间的服务,这些资源可以是计算、软件、数据和特殊传感器等等,它们主要以性能价格比作为最主要的因素,例如:远程医学仪器的使用和远程教育等。这类应用根据需要在适当的时间要求对自己无法解决的局部计算资源提供支持。NEOS 和 NET Solve 就是这种基于网络的、集成远程资源到局部计算中来的系统。

4. 数据密集计算的应用 这种应用从地理上分布的数字库、数据库和数据仓库中的数据中综合出信息,这种综合过程往往是计算密集和通讯密集的。大区域高性能并行与分布计算底层基础技术、基于信息的计算和数字库技术等的综合发展将是此类应用的有效解决途径。例如:未来的高能物理实验每天将产生太(10^{12})字节量数据,每年将达拍它(10^{15})字节量数据,需要使用复杂的查询方法访问巨大的数据,从中抽取、组织有效的信息。访问这些数据的科学工作者往往处在不同的地理位置,需要对数据进行共享和不同的计算处理,因此,希望这种庞大数据上的数据系统能够支持分布存储和管理,并提供良好的并行查询、提取等操作。在 SC'98 会议上 NPACI 和 IBM 公司等针对数据密集型计算及应用进行了展示,显示出良好的应用前景。

5. 合作计算的应用 这种应用主要关心如何允许和加强人们之间的交互行为和合作共同完成任务。

这种应用通常是根据虚共享空间而构造的,允许共享使用计算资源,能够进行远地的实时交互,完成一个共同的任务。这类系统目前有 CAVE5D 和 NICE 等。

从以上各种应用来看,高性能广域并行分布计算技术将是这些应用的基础,也正是这些应用反过来推动着这些技术的发展。在各种应用中我们可以得到这样的结论:首先,尽管目前技术还不成熟,但是许多基于高速网络、跨地域联合计算应用的成功,已显示出这种技术旺盛的生命力;第二,几乎所有以上应用对计算资源都具有极大的需求,共享计算、共享数据访问和共享通讯介质是非常重要的;第三,许多应用是可交互的,或者说依赖于和计算构件的紧密同步,也就依靠于并行与分布计算技术提供鲁棒性的性能保证。

五、元计算的关键技术

广域高性能并行分布计算将通过高速网络连接地理上分布的异构计算资源完成各种应用。分布式计算的低性能无法满足高性能计算的要求,而且跨地域计算的异构性、动态特性和区域性限制了并行计算的工具和技术在其中的应用。广域高性能元计算技术提出了一些新的挑战性问题,实际上也是并行计算和分布式计算发展将要遇到和解决的问题。分布式系统已经讨论了不同机器、不同网络 and 不同管理策略下的各种计算资源的连接技术,并行计算强调了高性能计算。问题是广域高性能元计算将如何结合并行计算技术和分布计算技术。我们认为其中的关键技术有以下几个方面:

1. 网络和网络技术

高速网络技术为高性能的并行分布计算提供底层基础性支持,大规模高速网络的建立是未来应用发展所必需的。高级通讯协议应当支持易用性、高性能、可靠性、容错性,通信服务质量保证 QoS 应当在未来网络得到有力的支持,网络接口和操作系统应协调,以减少网络通讯软件的开销,应保证应用通讯高带宽和低延迟特性,以及开发多播、可变优先级、自适应管理、带宽预定等技术。

2. 高性能并行分布计算的基础设施

大量的应用和工具需要基础设施的支持,这种底层结构不仅仅能够连接资源集合,隐蔽网络和机器的异构性,还要能够提供计算所需的基本服务,支持和简化包括计算密集、数据密集和通讯密集等各种计算类型的应用和工具的开发,支持各种高级服务策略在其基本服务上的有效实现,这是元计算技术非常关键和重要的技术。这种基础设施的基本服务至少有以下几种:

• 资源管理与分配 为应用提供一种表达资源需

求的方式、资源定位机制,以及资源的动态分配、进程管理和调度机制,应当支持各种不同管理策略和高性能调度技术的开发和实现。

- 通讯 它将提供基本的通讯机制,允许各种通讯方式在其上的有效实现,如消息传递、远程过程调用、分布共享主存和广扩技术等。

- 安全机制 在广域网络上进行计算,安全保证至关重要。安全机制将提供基本的安全保护验证机制,以验证合法的用户和资源,并为其它安全服务如授权和数据安全等提供接口,允许用户选择不同的安全策略、安全级别和加密方法等。

- 统一的资源信息服务 为访问动态、分布和多样性的并行与分布系统中的资源信息提供一个统一的、有效和可扩展的访问机制,以获取有关系统资源结构和状态的实时信息。

- 数据访问 通过对 I/O 的抽象,提供一个远程 I/O 接口,以提供对数据的高速远程访问,支持通过串行和并行接口进行远程访问和高性能的数据访问。

3. 程序设计环境

高性能的并行与分布计算环境应当给用户提供一个良好的程序设计环境,这一点类似于传统计算应用所面临的使用何种程序设计语言和环境问题。但是它不同于传统计算应用仅限于局域环境,面对的是大区域的异构环境,这是它的独特性所在,以下是主要的几种方法。

- 面向应用工具开发 针对特殊应用领域开发各种面向应用问题的方法和工具,这些方法和工具要广泛使用,必须能够适应在广域范围的环境下不同能力的网络和计算资源。

- 基于对象的方法 面向对象技术将成为软件实现的主要技术之一,以 Java、Web、和 CORBA、HPC++ 等技术为代表的对象程序设计环境和工具将适合构造复杂的计算对象和移动计算对象,实现软件重用和模块可交互性,这些方法将在高性能并行与分布计算中起着重要的作用,而高性能和实时性将是基于对象技术所需解决的问题。

- 高性能商业计算 目前,商业计算技术主要以三层结构的 Client/Server 技术为主,主要使用 CORBA、COM、数据库、Java、Web、网络技术等。这些技术都能够以不同方法用于高性能的计算当中,也能够得到高性能计算。目前,从传统的三层结构模型中发展出了分布商业计算和信息系统(Dccis),这种系统是一个面向企业的分布计算环境,需要高性能的计算,高性能和 Dccis 结合有相当大的困难,然而使用高性能的 Dccis 作为一个基础,巨大的商业投资将能够保存,利润也将随之而来。

• 4 •

- 编译器、语言和库 异构高性能分布计算对程序设计语言、编译器和运行系统提出新的问题,这些问题在传统的同构并行计算中并未得到广泛的讨论,并行编译技术、运行支撑系统技术、各种应用并行库、运行编译、程序管理、程序之间编译技术等都需要进一步研究。

4. 可视化、合作、交互、沉浸式虚拟工作环境技术

研究交互、沉浸式虚拟现实系统,并将网络上各种可视化虚拟现实系统和超级计算机连接起来,开发共同的工具和环境,以达到对多模式、交互、沉浸式虚拟环境的使用,更加逼真地模拟应用问题。

5. 大型分布式资源存储的建立和集成化的管理

在广域高速网络中,数据资源将达到拍它字节规模,如何有效地将它们分布存储、有效地检索、提取、集成优化和发掘是摆在我们面前必须逾越的鸿沟。

六、元计算的实际应用领域

广域元计算系统建立一个具有太(10^{12})—拍(flop)性能的、拍它(10^{15})字节大小规模的高性能并行与分布计算环境,为大型科学应用计算问题提供一个国家级范围内的计算平台。值得一提的是广域元计算技术的研究并不能够替代和否认百万和千万亿次超级机的研究,因为这些超级计算机技术是大型科学应用研究的关键技术。然而大型科学应用研究不仅需要超级计算机,还要依靠大量的其它计算资源和仪器设备,这也正是广域元计算系统所能提供的。下面,我们简单介绍在广域元计算系统上的一些大型科学应用研究。

- 核模拟和核库存维护 核模拟和核库存维护在核禁试条件下要求具有超高性能的计算机,希望能充分利用现有的 Internet/Intranet、软件技术、多媒体及信息压缩技术,虚拟现实技术的支持,构建国家级范围内的合作开发环境,同时要求有一整套的高级计算测试和仿真模拟工具。

- 宇宙学 针对几十个太字节的多光谱数据的收集、分布存储、检索等,并且对大量图像进行形态学、统计学研究,并从中抽取进行分析,这种研究不仅需要高性能的计算,还需要异地合作收集、存储、发掘、分析和研究,同时要求超高性能的计算机能够和大量仪器设备以及各种不同类型的计算机共同组合成一个系统。

- 环境水文学 需要创造数字模型、数据处理分析工具和可视化系统,以预测和评价环境状态,进行有效的洪水管理、生态系统的管理和后勤保障,并能够将各种独立的环境水文中的模型连接起来,共同发挥效应。

- 流体力学 流体力学中的湍流特性研究需要建模,并构造仿真模拟环境,进行验证,这其中需要单个

的高性能计算机,以及可视化、虚拟现实系统和设备等。例如,对于 1024^3 阶的湍流模型求解和模拟仿真必须在太-拍(Tera-flops)的大规模多计算系统组成的系统下运行,才能完成。在化学非平衡、真实气体效应的计算研究中同样如此。

·其它领域 如复杂系统、材料、工程、神经科学、化学等重大应用领域都需要一个高性能的国家级虚拟网络合作工作环境。

七、国内外研究现状

国际上已经确定了高速网络广泛建立的策略,从美国政府提出的下一代 Internet 网 NGI 到美国国防部(DoD)的国防研究和工程网络 DREN,以及能源部(DOE)的能源科学网 ESnet 和美国国家科学基金会(NSF)的 vBNS 等都是高速网络为目标在进行实施,这奠定了未来大区域高性能并行分布计算的基础。在高性能的网络计算方面,研究包括 SMP、SMP 的机群系统到分布的同构机群系统和异构机群系统的大区域计算 Grid。Grid 是 1997 年 10 月美国国家科学基金会(NSF)宣布面向 21 世纪的“计算科学联盟”重大基金项目,该项目计划以美国全国范围内的元计算为目标,历时五年时间,联合六十余所大学、科研机构和公司,集中力量在全国范围内建立一个高级计算的基础设施,以解决大区域的高性能并行与分布式计算,在大范围内将由巨型网络结点把分布在各地的计算资源和数据资源连接在一起,达到太-拍规模的系统,形成新的科学计算和信息应用结构,并在上面集成各种系统和开展各种应用研究,以解决宇宙学、生物化学、生物化学工程、流体力学、材料、神经等重大科学应用领域的计算问题。在未来五到十年时间使美国在全国范围内的元计算达到实用水平。

目前,其它国家包括日本、加拿大、新加坡、西欧各国和俄罗斯也开始加入美国的科学计算联盟。同时,美国的 DoD、NASA、NIH 和 DOE、DARPA 除了支持 MPP 技术的研究,也强力支持高性能并行与分布计算技术。核模拟技术确定将采用此技术,ASCI 计划采用 DCE 技术连接美 Lawrence Livermore, Los Alamos 和 Sandia 三大实验室构成合作计算环境,以使任何一个实验室研究人员可以通过桌面机访问其它实验室的计

算资源,尽管这种方式的缺陷是明显的,基于 RPC 机制,而非高性能计算,然而实质上它也是一种跨地域的联合计算,重要的是它说明了需求的存在和强烈要求。美 DOE 的 ESnet 也作为计算 Grid 系统初始网络的一部分。总而言之,美国在此方面的研究处于世界领先水平,研究工作已经开展起来了,并取得了一些实验性研究成果,例如 Globus 系统,但是整个研究处于起步阶段。

高速网络的建立也将是我国网络发展的必然趋势。现行国内的研究基本集中在局部网的并行计算,对大区域高性能的并行与分布计算技术的研究较少,且无实验环境,其它一些院校和国防科学技术大学并行与分布处理国家重点实验室在此方面开展了初步的研究工作。但是,我国在该技术的相关领域 MPP 技术和实现、网络、通信领域、中间件、虚拟现实和并行算法、分布计算等都有大量而深入的研究和技术储备、如果能够在国家的指导下,集中一批有实力的计算机科学家、应用领域专家、大学、研究机构组成团体共同攻关,是能够尽快缩短与美国的差距的。

八、建议

基于高速网络的广域高性能元计算技术是下个世纪初计算机技术的重大技术,是一个国家重要的战略技术资源,对增强国家的国力和军力具有重要的意义。因此,我们建议国内应当集中一定的财力和人力,依托有实力的单位建立若干个高性能计算中心,用高速网络将它们连接起来,研究高性能广域并行与分布计算技术,提供高性能计算服务,以使我国在该重要技术领域研究赶上世界先进水平,为国民经济和国防建设作出实际的贡献,也为银河、曙光和其它国内巨型机产业的应用提供一种有效的途径。

参考文献

- 1 Smarr L. et al. Computational Infrastructure: Toward the 21st Century. CACM, Nov. 1997. 28~94.
- 2 Grimshaw A S. Legion-A View Form 50,000 Feet. [Computer Science TR96-23]. University of Virginia, 1996.