杨 利 昌月楼 等编著

并行数据库技术

国防科技大学出版社

并行数据库技术
杨利 昌月楼 等编著
国
社

并行数据库技术

杨利　昌月楼　等　编著

国防科技大学出版社
·长沙·

TP

内 容 简 介

本书全面介绍了并行数据库研究领域的现状和各种并行数据库处理技术。全书共分六章，重点介绍数据库并行处理中用到的各种实用算法。第六章中介绍的几种国内外并行数据库系统可为研制人员提供有益的参考。

本书可作为高等院校和科研机构的教学和科研参考书。

图书在版编目(CIP)数据

并行数据库技术/杨利等编著.—长沙:国防科技大学出版社,2000.5
ISBN 7-81024-625-9

Ⅰ.并… Ⅱ.杨… Ⅲ.并行数据库 Ⅳ.TP311.133.2

中国版本图书馆 CIP 数据核字(2000)第 25554 号

国防科技大学出版社出版发行
电话:(0731)4555681 邮政编码:410073
E-mail: gfkdcbs@ public.cs.hn.cn
责任编辑:黄 煌 责任校对:文 慧
新华书店总店北京发行所经销
长沙交通学院印刷厂印装
*
787×1092 1/16 印张:12 字数:277 千
2000 年 6 月第 1 版第 1 次印刷 印数:1-2000 册
*
定价:18.00 元

前　言

数据库技术是一项十分重要的计算机技术，它在很多应用领域发挥着关键作用。尤其是关系数据库技术，它在商业应用领域取得了巨大的成功，并以惊人的速度向其他领域拓展。除了传统的 OLTP(联机事务处理)领域之外，新的数据库应用领域不断增加，例如：CAD/CAM、CASE、GIS、DSS(决策支持系统)和数据挖掘、VOD、知识库系统、实时系统等等。同时，随着应用的积累和 Internet 技术的日益普及，众多一直在发挥作用的数据库的数据量和工作负荷日益加重，并呈现出以下特点：

1．数据库规模十分庞大

Meta Group 的 Aaron Zornes 预测，21 世纪建立 1 万亿字节数据仓库的企业将从 7%增至 17%。据不完全统计，目前采用 IBM 数据库超过 1TB 的客户就达 20 多家。数据库界称这类数据库为海量信息数据库。

2．数据库查询越来越复杂

人们期望数据库具有支持智能化应用的特性，要求在数据库处理技术中糅合人工智能和专家系统式的处理手段，比如对规则系统的支持等等。数据仓库和数据挖掘应用中要求支持用户对数据进行多维分析和获取决策信息，这绝不是简单的查询检索能完成的。

3．实时性要求高

指挥与决策支持系统、高速目标识别、大型股票市场分析等应用对查询响应时间的要求极为严格。

越来越多的应用表明，运行于串行计算机上的传统数据库管理系统缺乏支持高性能联机事务分析处理(OLAP)和复杂查询操作的能力，日益加重的负载使其达到了性能的极限，使它很难适应迅速增长的应用需求。

另一方面，20 世纪 90 年代中期以来，由于微处理机技术高速发展以及高带宽低延迟互联网技术的成熟，促使计算机体系结构快速向基于通用微处理机芯片的对称多处理(SMP)和大规模并行处理机(MPP)转变。各种高性能应用需求的不断涌现正在使并行计算机结构受到越来越普遍的重视。但是，并行计算机的这种发展形势并没有得到软件系统的相应支持，并行计算机结构的性能还没有得到充分发挥。这种现象在数据库技术中体现得更明显。

近十年来，在上述应用需求和并行处理技术发展的双重驱动下，并行数据库技术一直是研究界和工业界的瞩目焦点。不仅出现了像 Gamma、Bubba、Volcano、DBS3 等并行数据库系统原型系统，更有商品化的并行数据库系统问世，如 Oracle 并行数据库服务器、Informix并行查询服务器、IBM DB2 UDB EEE 等。并行数据库技术正在各种大规模应用中发挥作用，并将不断面对新的应用挑战，成为未来的高性能数据库系统的必由之路。

并行数据库系统能支持并行处理体系结构，获得比串行系统下高得多的性能，关键在于其中的并行数据库算法技术、并行查询优化技术以及先进的数据存储技术的支持。因此，本书全面介绍了并行数据库技术的产生和发展，以及并行数据库技术的基本知识，阐

述了并行数据库系统使用的关键技术，突出算法的描述和分析以及技术的实用性。

全书共分为六章。第一章简单介绍了高性能数据库应用和并行处理技术的背景，说明并行数据库技术是如何成为未来高性能数据库系统的必由之路。同时，这一章介绍了并行数据库的研究内容、基本的概念和理论，国内外并行数据库研究的状况，以及发展展望。第二章主要论述并行处理环境和并行数据库系统中的并行性开发、并行处理体系结构、并行处理的粒度、数据偏斜的概念以及并行性的度量标准等问题。这一章可以作为并行数据库技术的预备知识。

并行数据库的技术基础是数据库中的数据必须分散存储，以便得到并行 I/O 的支持。第三章主要论述这方面的技术。这些技术包括并行数据库中为了提高查询操作的性能而采取的各种数据分布技术，分为一维数据划分和多维数据划分技术，介绍了各种技术的理论依据、适用范围和相关研究情况。这一章结合了作者在该领域的最新研究成果，重点讨论了多维数据划分方法在并行数据库系统中的应用。

第四章是本书分量较重的一章，详细介绍了关键的二元数据库操作算法和并行化技术，对并行排序，特别是并行连接算法做了大篇幅的论述。对各种基于 Hash 的并行连接算法、对称 Hash 连接算法、连接算法的加速技术、数据偏斜的处理技术以及基于多维数据划分方法的并行连接算法等做了详细讨论和分析。第五章则介绍并行数据库系统中另一个重要的课题——并行查询及优化技术。其中主要论述了复杂多元连接优化技术，介绍了各种线性和非线性优化模型。在任务调度技术中介绍了处理机分配算法、连接调度算法等问题。作为并行数据库系统实用化的关键，第六章介绍了并行数据库系统的实现问题，给出了国际、国内几个主要研究原型，以及重要的数据库厂商如 Oracle、Informix、DB2 等的并行性解决方案、实现技术、实例和体系结构。这一章对于在传统数据库系统上如何改造，使之成为具有并行处理功能的数据库系统具有参考价值。

在本书的写作过程中，作者参考了大量的文献和技术资料，同时加入了最新研究成果。书中的图表都经过测试或有引用出处。要求阅读本书的读者具备数据库原理知识，并且最好有数据库管理系统或应用实践背景。

本书第二章、第四章、第五章由杨利博士编写，第三章由谭郁松编写，第一章和第六章由昌月楼和叶常春编写，昌月楼主持了全书的审校工作。本书既是一本教学和研究参考书，又是一本研究专著。读者对象为计算机或信息系统专业研究生、计算机技术工程技术人员、大学相关专业教师以及企业研究机构的研究人员。

杨　利

2000 年 4 月于长沙

目　　录

第一章　并行数据库——高性能数据库发展的必由之路

社会需求带动科技发展，科技发展推动社会进步。纵观科技发展史，莫不如此。数据库技术的发展亦不例外。当数据库代替文件系统使计算机应用更上一层楼时，人们当初甚感兴奋——尽管当时的数据库还比较初级。然而，随着社会的进步，人们并不满足已有的成绩，不断地提出新的要求，不断地改进数据库系统，使得数据库的研究和发展取得了长足进步。数据模型由最初的层次型和网络型发展到现在广泛使用的关系型，并在向对象关系型发展；数据类型也由当初简单的数值型和字符型发展到包括大对象在内的复杂数据类型；编程界面也由导航式语言发展到非过程语言，以至今天的可视化编程。

然而，有一种至关重要的需求使得数据库科研工作者在改进数据库技术时上下求索，孜孜不倦，这就是如何提高数据库的处理能力。许多应用是要求高性能数据库才能完成的，例如，预测模型及模拟、工程设计及自动化、能源开发、医药、军事和基础研究等等，它们都需要有大容量和高处理能力的数据库系统。然而，追求这一目标的道路是坎坷不平的。20 世纪 70 年代中期到 80 年代中期，数据库研究者在研制数据库机器方面付出了辛勤的劳动，研究出诸如 IFAM、MOULDER、ACPS、SURE、SHAW、CAFS、DIRECT、RELACS、DIALOG、MBDS、MIRDM、DBC、HYPERTREE 和 SABRE 等系统结构纷纭复杂的数据库机器。然而，由于这类机器硬件系统的昂贵和磁盘的处理瓶颈，这一探索未能如愿以偿。

山重水复疑无路，柳暗花明又一村。并行计算机系统的出现为高性能数据库系统带来了曙光。研究工作转向了以通用的并行计算机为基础的数据库系统。目前，并行数据库系统已成为一个新的数据库研究领域，并取得了一些引人注目的进展。实践将证明，并行数据库是高性能数据库研究发展的必然方向。

下面将叙述数据库应用的新特点，并行处理为数据库新应用要求带来的生机，以及研制并行数据库要解决的问题。

1.1　数据库应用的新特点和新要求

数据库是用来存储数据和提供信息的。在当今信息作为重要的社会资源和信息“爆炸”的年代，数据库作为信息产业的基石，自然出现了许多不同于 20 世纪 70～80 年代数据库的新特点；同时，人们对它也提出了许多新要求。

1.1.1　数据类型的多样化和复杂化

由于数据库应用已从早期简单的商业事务处理发展到今天的工程数据库应用，例如 CIMS、地理信息系统(GIS)、多媒体信息系统(MDIS)和科学与统计数据应用，其应用领域出现了五彩缤纷的局面，因而，早期简单的数据类型，如整数、浮点、字符、日期、逻辑和货

币，就不能适应上述新的应用领域的要求了。这些新的数据类型表现如下：

1. 支持声音、图像、视频和超大文本等的二进制大对象类型(BLOB)

这些数据类型主要是为了满足数据库的多媒体应用而出现的。

2. 支持点、线、圆、多边形等几何元素的数据类型

这些数据类型为数据库的 GIS 应用和 CAD、CAI、CAM 以及 CIMS 应用提供了很好的基础。

3. 时间段数据类型

这些数据类型为数据库的在线事务处理(OLTP)、在线事务分析(OLAP)、数据仓库(DW)和数据挖掘(DM)等应用提供了有效的支持，因为这些应用一般都涉及到对过去数据的回归分析和处理。

4. 复合数据类型

按照关系数据库的范式理论，关系数据库中关系的字段必须是具有原子性的，这是它的第一范式。然而，新的应用对这一理论提出了挑战。MDIS、GIS 和 CIMS 等应用要求打破这一约束，于是出现了非原子字段的要求。

一般的商用数据库管理系统都支持 BLOB；开放源码数据库管理系统 PostgreSQL 除支持 BLOB 外，还支持几何数据类型和时间段数据类型[4]。国产数据库管理系统 DM2 支持复合数据类型。

1.1.2 数据模型的演变

数据模型从最初的层次型和网络型发展到后来的关系型。目前，由于关系型数据库具有许多优点，所以它已完全取代了层次型和网络型数据库。然而，数据库新应用的要求发现关系模型尚有不足之处。多媒体应用就是其中的一个例子。在多媒体应用中，由于数据类型的多样性和数据操作的复杂性，面向对象数据库管理系统(OODBMS)显得更为合适。由于 OODBMS 具有抽象数据类型和类的封装性、层次性与多继承性，在管理和操作多媒体数据库方面具有其得天独厚的条件。而在更为广泛的商务应用中，关系型数据库管理系统(RDBMS)却优于 OODBMS。因此，在目前的商品化的 DBMS 中，纯 OODBMS 的产品并不多见，常见的是具有面向对象功能的 RDBMS。它是这两者的“杂交”，因此被称为对象关系数据库(ORDBMS)，它具有这两者的长处。此外，对于 GIS、CAD、CAM、CIMS 和 CASE 等应用，ORDBMS 也优于 RDBMS。

著名的商品化数据库管理系统 DB2 就是 ORDBMS 的典型代表，开放源码数据库管理系统 PostgreSQL 也具有面向对象的特点[4]。

1.1.3 数据库的智能性

信息“爆炸”的现实使得人们在浩瀚的数据海洋中显得茫然，从而希望管理数据的数据库管理系统能够提供某些智能，根据人们不同的需要，提供有用的数据。于是，20 世纪 80 年代出现了主动数据库、专家数据库、演绎数据库等方面的研究。这些数据库的特点是，它们能按照人们的要求自动进行某些逻辑推理，提供有用的信息。目前，在信息工程领域中，人们比较关注的数据挖掘(DM，Data Mining)就是基于上述要求而出现的，一些这

方面的文章和原型具有研究价值。

1.1.4 数据操作的复杂性

数据操作的复杂性源于数据库应用的多样性。在多媒体数据库中，对数据库的检索只提供传统意义上的基于表示的检索是不够的，还要有基于内容的检索[2]。例如，“找出含有‘多媒体’单词的句子”；或者根据图像上的某些面部特征，在图像数据库中找出具有此面部特征的人的面部图片。这些都属于基于内容的检索。前者属于基于超文本内容的检索，后者属于基于图像内容的检索。对于前者，它的实现并不困难，努力的方向在于加快检索速度；对于后者，实现起来是很困难的，它牵涉到查询要求的表示、图像识别和模糊数学等多领域的学科内容。在 GIS 应用中，常常涉及到检索对象之间的位置关系问题，例如，查询要求为“找出离火车站距离少于 100 米的所有三星级宾馆”。这里要求找出以火车站为圆心、以 100 米为半径的区域内所有的三星级宾馆，体现在数据库底层的操作表现为“面包含点”的问题。此外，还有点和线、线和线、线和面、面和面等等之间的距离、包含、相交、左边、右边等各种错综复杂的位置关系问题。这些特殊操作的复杂性来源于检索要求的多样性。当然，在图形数据库中此类问题同样会出现。

1.1.5 高性能数据操作的紧迫性

高性能数据库是指处理速度特别快和容量特别大的数据库。而“处理速度”通常是指查询速度，因为“查询”是出现最频繁的操作，其反应速度对业务工作影响很大，也关系到数据库在用户心目中的形象。

高性能数据库系统的需求反映在诸如银行系统、证券系统、实时点播、电信实时计费系统等覆盖面广并对系统的实时反应要求很苛刻的应用中。这类系统覆盖全国甚至全球，它除了要求极可靠的安全性外，还需要有很快速的响应时间和很大的吞吐量。一般说来，增大数据库容量是不难的，现在硬件技术的发展已能解决这一问题，TB 级的数据库系统已不鲜见。难度在于数据库系统的速度上。数据库容量的增大和数据处理的复杂化都会降低数据查询的反应时间和数据处理的吞吐量。长期以来，数据库工作者努力奋斗的目标之一是不断追求高速访问速度。在 20 世纪 70～80 年代，国外不少学者潜心研究数据库机器，希望采用专用数据库机器的方法来获得数据库的高速访问，但由于其造价昂贵和磁盘I/O瓶颈问题而未能商品化，因而其研究工作到 80 年代就基本无人问津了。

于是，数据库研究人员另辟蹊径，力求在造价相对低的通用机上，利用今天迅速发展的硬件技术和并行技术，实现高性能数据库。

1.2 并行处理为满足新的数据库应用带来生机

1.1.1 节至 1.1.4 节中提到的数据库发展的新特点和新要求，由于面向对象技术、多媒体技术和人工智能技术等交叉学科技术的发展和成熟，有的得到了较好的解决，有的得到了部分解决。而高性能数据库的研究和实现，还有很多工作要做，还有许多困难要克服。高性能数据库的基础是并行处理技术，包括并行硬件机构、并行软件系统和并行算

法。而并行处理技术正是现在不少科技工作者都在啃的“硬骨头”,这就决定了并行数据库的难度。

1.2.1 一个并行数据处理的例子

下面介绍一个国外 20 世纪 80 年代末至 90 年代初的并行处理的例子。该例子是 Intel 公司的并行处理系统 iPSC/860,是当时美国有名的几大并行处理系统之一。

1. iPSC/860 超级计算机介绍

iPSC/860 是 Intel 公司的 iPSC(Intel Powerful Supercomputing)超级计算家族的第三代产品,它出现于 20 世纪 80 ~ 90 年代,其微处理器是 RISC 结构的 i860。它最多可有 128 个 i860 微处理器,系统提供的峰值速度是 10GFLOPS(32 位)或 7.6GFLOPS(64 位)。它是一个分布式存储的 MIMD(Multiple Instruction Multiple Data)系统,每个处理机有 8MB 或 16MB 局部内存。各处理器之间用直连(Direct-Connect)结构进行通信。此种通信方式使通信开销少到可以忽略,同时通信路径上的中间结点也不受其他结点通信的影响。直连网络的带宽是可变的,在 128 个结点上的综合消息传递能力为 700MB/s。这种方式的最大好处是使得程序员在将应用程序映射到系统时不必考虑其拓扑结构。系统结构的灵活性是它最重要的特性之一,当需要的处理机数增减变化时,原来的程序代码可以不必改写。

系统由计算结点、I/O 结点和硬盘组成。计算结点的结构如图 1.1 所示。I/O 结点的结构如图 1.2 所示。

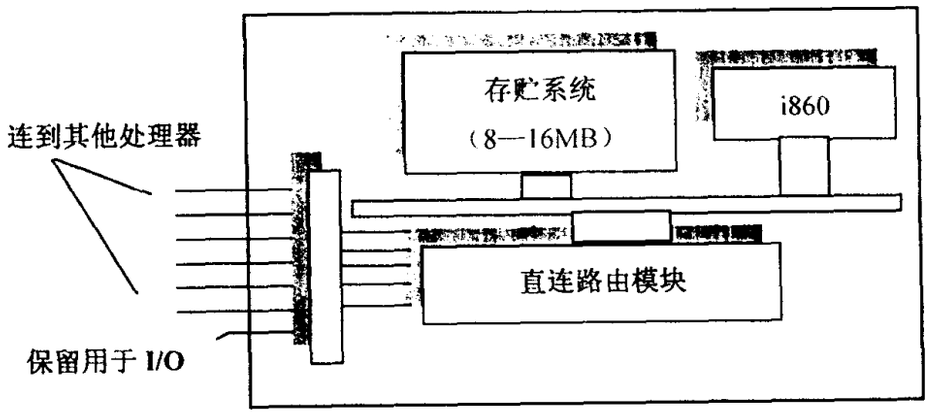

图 1.1 计算结点简图

虽然图 1.1 中一个 I/O 结点只连接到一个计算结点,但是直连网络的特点允许所有计算结点访问每一个 I/O 结点。一个具有 8 个计算结点、4 个 I/O 结点和 8GB 磁盘的例子的系统结构图如图 1.3 所示。

系统的 0 号结点连接到一个 80386 微型机上(当时最高档的微机),该微机被称之为系统资源管理器(SRM),SRM 上运行的是 UNIX V.3,它提供软件开发平台和从它访问 iPSC/860系统的通道。标准开发语言是 Fortran 和 C,它们可以使用系统调用在 iPSC/860 内部各结点之间或者在 iPSC/860 和 SRM 之间传递数据;可以将进程广播到所有结点进行全局算术运算,例如加、减、求最大值,等等,然后通过 SRM 直接进行 I/O 操作。

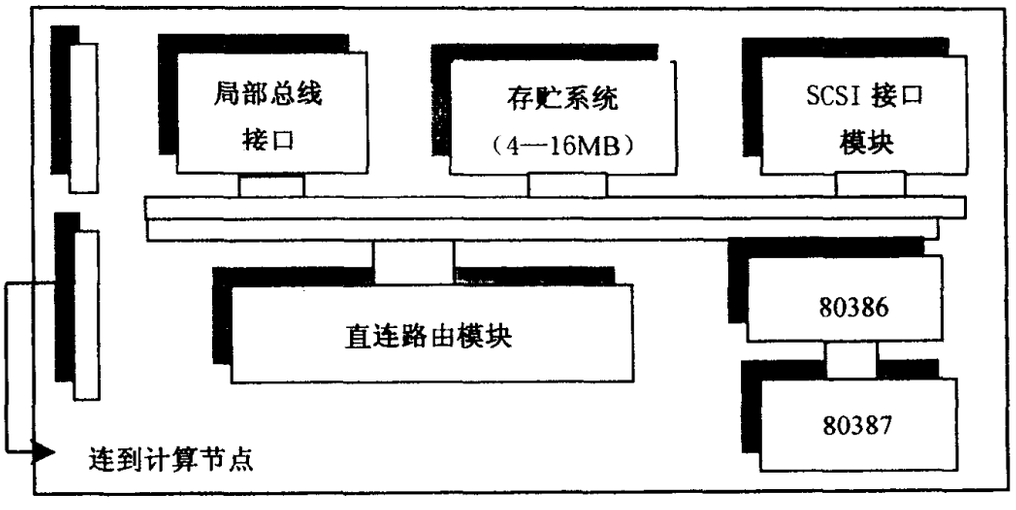

图 1.2　I/O 结点模块简图

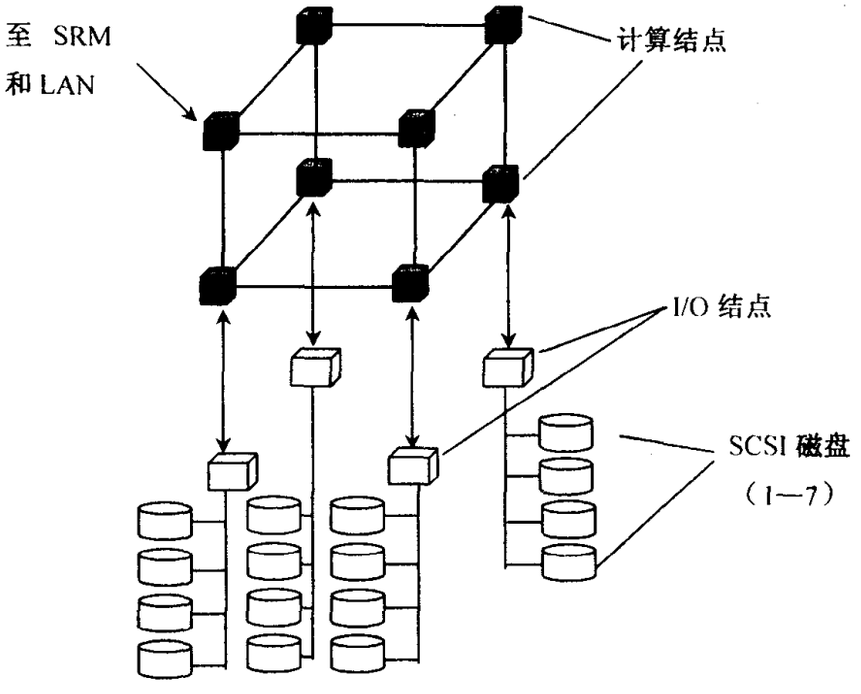

图 1.3　iPSC/860 简例

该并行工作平台的另一个特点是它的被远程访问的能力，即在 SRM 和 SUN 或 VAX 之间可以通过局域网（TCP/IP 以太网）进行通信。虽然局域网的传输速度不如 iPSC/860 内部消息传递的速度高，但是，与人机接口有关的代码可以保留在传统的工作站上，而大量的计算可以在 iPSC/860 上以超级计算速度进行，从而实现真正的交互式超级计算。最终用户只要关心传统的工作站上的计算，而不必关心并行系统的超级计算，也不必关心它在什么地方。

iPSC/860 是一个多用户系统，各用户可以动态瓜分系统中的结点。这种动态分配结点的特点加上它被远程访问的能力，使得它成为工作站网络上一个极其灵活的计算服务器。

iPSC/860 的文件系统是并行文件系统，而且其并行性对最终用户或应用程序是透明的。过去，人们对将多个磁盘挂接到多个计算实体的问题进行了许多研究和调查，最后形成了由分布式文件系统组成的分布计算模型。SUN 公司的 NFS 就是一个例子。它的缺点是一个文件只能存放在一个单一的磁盘上，在并行访问同一个文件时会带来严重的I/O瓶颈问题。而 iPSC/860 的并行文件系统将挂接到系统中的所有磁盘作为单一的逻辑磁盘来管理(通过 127 个 I/O 结点可以挂 889 个容量为 500MB 的磁盘)。它把一个文件分散保存在各个磁盘上，这一技术被称为分布(Declustering)技术。由于文件的不同部分可以同时被访问，从而克服了并行处理中的 I/O 瓶颈问题。用户创建的文件可以大到整个逻辑盘，也可以将数百个文件分散到许多磁盘上，而不必关心任何文件、块和目录的物理位置。文件块的大小是 4KB，每个磁盘有个卷号，每个文件块的块号由卷号和该卷中的逻辑块号惟一确定。进行写操作时，文件的各块被并行文件系统自动地分配到所用的I/O结点和它们的磁盘上。I/O 结点使用的缓存(Cache Memory)大小是 2MB。

2. 一个在 iPSC/860 上的并行数据处理的例子

下面简单介绍一个 iPSC/860 应用在金融系统上的例子。在国外，金融系统(银行、证券、保险)的运作是很活跃的，其业务量是很大的。在高风险的金融服务行业中，使决策变得十分困难的主要因素是：全球性竞争、全天候(24 小时)交易和大的市场波动。因此，能否提供及时观察市场活动和证券价值的有效决策工具变得至关重要，从而使超级计算机扮演了一个关键性的角色。超级计算使得过去无法想象的金融业务计算成为现实，使得费时的批处理操作成为及时的交互操作(过去传统主机上要几个小时才能完成的运算变成站在电话机旁的委托人只要等几分钟就能完成)。iPSC/860 并行超级计算代表了一种低成本高回报的投资，它已广泛应用于资产评价、数量经济、外汇交易、投资组合评估、投资组合优化和风险分析等方面。下面是抵押证券的一个例子。

抵押证券带来的复杂性来源于资金流的不确定性。在任何时候都可以发生的抵押预付款是不确定性的主要原因。此外，可调利率抵押的抵押利息的资金流取决于抵押的特殊结构。处理抵押证券的常用方法是采用蒙特卡罗(Monte Carlo)模型。它使用计算机产生的随机数作为模型的输入，该模型输出随时间变化的利率走向。然后，与某证券有关的预付款模型被用来计算利息和本金的资金流，并将它们贴现。计算时需要几百个这样的资金流模型，最后从所有这些模型的结果中取平均值。

蒙特卡罗方法最令人烦恼的是计算需要消耗大量的时间，但它很适合 iPSC/860 这种系统的超级计算。图 1.4 给出了不同系统对同一个应用表现出的性能比较。

1.2.2 并行数据库作为高性能数据库的理由

上一节的例子只从一个侧面——金融业务需要——说明了并行数据处理带来的好处，实际上，还有很多部门都需要有超级速度和超大容量的超级计算机。例如，物探部门的地震数据处理、气象部门的气象卫星数据处理、工程设计部门的有限元分析、计算型流体学，以及军事部门的核爆炸模拟数据处理、经济部门的全国乃至全球经济模型的数据处理，等等；甚至高科技本身也需要超级计算，例如高分子合成、遗传工程(基因研究)等，不一而足。虽然上一节的例子未用到并行数据库——因为这个例子是出现在并行数据库的

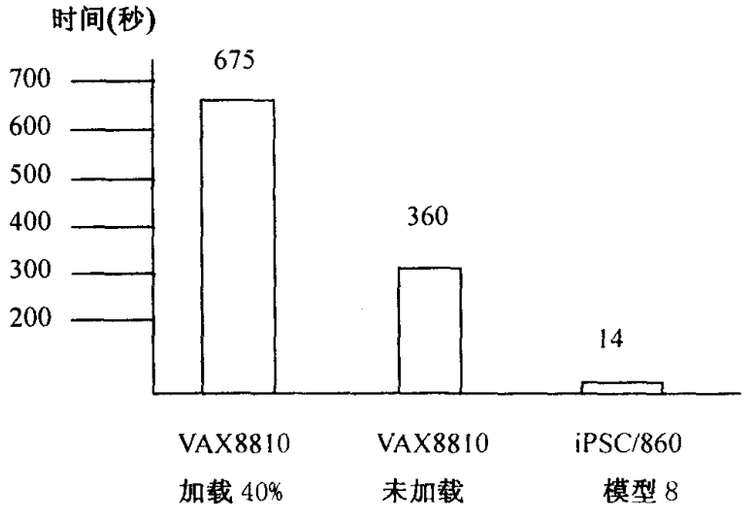

图 1.4 性能比较[5]

研究刚刚起步的时候,但并行数据处理中的并行数据库的需要是十分显然的。并不是不需要并行数据库,而是当时的水平开发不出并行数据库。

超级计算的要求必须由并行处理来满足,这一点是不难理解的。尽管人们使尽浑身解数来提高计算机硬件的计算速度,例如,不断地开发出主频越来越高的 CPU,采取多级缓冲及指令预测,提高内存及磁盘的访问速度,以及改进工艺以便提高集成度和减小功耗,等等。然而硬件速度提高的幅度却会越来越小,而且最终会走到它的极限,因为硬件的改进是有极限的。当然,也可以另辟蹊径来寻找新的硬件,例如,研制光计算机、生物计算机和化学计算机,等等,而且目前有人在跃跃欲试并有所收获,但这些对目前来说至少是远水不解近渴——如果它能成功的话。所以,比较现实可行的办法是在目前的软硬件条件下研究并行处理技术,将一个任务分配给尽可能多的处理机去同时完成。在理想的情况下,由 n 个处理机完成一个任务所需的时间是由单机完成同一任务的 1/n。

基于上述理由,自 20 世纪 80 年代以来,并行处理的研究成为计算技术研究中一个重要的越来越受人关注的分支,而且研究成果越来越成熟。在国外,类似于上一节中的例子是很多的,美国的 Cray、SGI、IBM 和日本的 NEC、富士通等都是著名的超级计算机(并行机)研究和生产公司。据报道,现在国外超级计算机的处理速度已达到万亿次每秒的数量级;而共享主存(SMP,Shared Memory Processing)类型的并行处理系统几乎到处可见,就连高档微机也是具有 4~8 个 CPU 的 SMP 并行处理机。在国内,"银河"和"曙光"是早已投入使用的大规模并行处理机(MPP,Massive Parallel Processing),清华大学和黑龙江大学都已研制出基于局域网的并行处理机。

如前所述,在数据库领域中,人们在追求高性能数据库方面曾走过弯路——从早期的专用数据库机转到研究并行数据库。既然并行处理机技术的研究还算顺利,并且至今可谓硕果累累,在现有并行系统的基础上研制并行 DBMS(PDBMS),何乐而不为呢? 数据库机器是从硬件入手研制专用机,既昂贵又费时,而且速度受限于 I/O 能力;而在成功运行的并行处理机上研制 PDBMS,主要牵涉到软件的研制,它不需要昂贵的硬件,可以在通用机上进行,好处很多。

尽管困难很大,但研制 PDBMS 是完全可能的。首先,十几年来关于并行处理和并行数据库方面的研究为成功研制 PDBMS 奠定了基础。这些研究在解决并行系统结构、并行处理机调度、处理机通信、并行数据划分、并行操作算法和并行查询优化等方面做出了有益的探索。其次,一些 PDBMS 原型系统的研制为进一步开发商品化的 PDBMS 提供了丰富的经验。20 世纪 80 年代末的 Bubba、Gamma、Valcano 和 Tandem 等都是著名的 PDBMS 原型;在国内,黑龙江大学的 Hcluster、中国人民大学的 Cobase 和国防科技大学的 ParaBase 以及 GDB/P 都是 PDBMS 的原型系统。20 世纪 90 年代后期,国内 PDBMS 的研究有很大进展。上述原型系统从不同的侧面对 PDBMS 的设计与实现进行了有益的探索。更有说服力的是,一些著名的商品化数据库管理系统都已成功地增加了并行处理能力。Oracle 公司的 OPS(Oracle Parallel Server)、Informix 公司的 Online Dynamic Server 和 Sybase 公司的 VSA (Virtual Server Architecture),甚至微机上的 MS SQL Server 7.0 都提供对多 CPU 的支持。虽然还不能认为它们就是 PDBMS,但能执行并行数据处理,提高查询速度,能有限地满足当前高速数据处理的需要,这也是数据库技术发展的一个里程碑式的进步。因为它表明并行数据库是完全有可能走向商品化应用的,并行数据库是高性能数据处理的必由之路。

1.3 要解决的问题

如前所述,现有的商品化 DBMS 中实现并行数据处理的途径是在原有的 DBMS 的基础上增补并行操作功能,以尽可能满足超级数据处理的需要。但是,它们都是基于 SMP 结构的并行机的。在这种结构中,系统中多个 CPU 共享内存和磁盘,并行的方式是将一个客户的请求通过专用的并行算法拆成若干个可并行的片段,由各个处理机来并行执行,最后再进行汇总。事实上,由多个处理机共享磁盘或内存并不能有效地提高系统性能,因为各处理机要争用有限的资源封锁和进行频繁切换,这将增大系统开销。这种结构在 CPU 增多的情况下,加速性能会不断下降,甚至 CPU 多到一定程度时不但不能加速,反而会减速,其性能曲线对 CPU 个数成抛物线形状。所以,SMP 结构的并行系统的扩充性很差。

要克服这一缺点,就要使每一个处理机都有自己的内存和硬盘,这就是目前研究得比较多的无共享(SN,Shared Nothing)并行机构。通常讲的大规模并行处理(MPP)就是指这种结构。这种结构的共享资源少,系统开销少,有很好的加速比(Speedup);更重要的是它有很好的扩充性(Scaleup),因为它的加速比受规模扩充(处理机增加)的影响不大,至少不会出现负效应。但是,研制 MPP 并行机的难度要远远大于研制 SMP 结构机器的难度。国防科技大学研制的银河-Ⅱ是 SMP 结构的并行机,银河-Ⅲ是 MPP 结构的并行机。

要在 MPP 结构的并行机上研制 PDBMS 是一件很艰巨的工作。Oracle 公司在 nCUBE 并行机上实现的并行数据库和 Sybase 公司与 NCR 合作研制的 Navigation Server 属于大规模并行机上的 PDBMS。但是,MPP 上的 PDBMS 的开发进展一直很慢,这有两个方面的原因。从商业角度出发,这类 DBMS 的需求面不大,而研制它需要的投资却很大,所以,公司不愿意出资开发。据有关统计报导,95%的用户使用单 CPU 的机器,使用含 2~8 个 CPU 的机器的用户只有 4%,而使用含 8 个以上的 CPU 的机器的用户只占 1%(1994 年数据);