大数据
分析统计应用丛书

全国应用统计专业学位研究生教育指导委员会推荐用书

大数据分布式计算与案例

DISTRIBUTED STATISTICAL COMPUTING FOR BIG DATA AND CASE STUDIES

主编　李丰

中国人民大学出版社

全国应用统计专业学位研究生教育指导委员会推荐用书

大数据分布式计算与案例

DISTRIBUTED STATISTICAL COMPUTING FOR BIG DATA AND CASE STUDIES

主编　李丰

中国人民大学出版社
·北　京·

图书在版编目（CIP）数据

大数据分布式计算与案例／李丰主编. — 北京：中国人民大学出版社，2016.7
（大数据分析统计应用丛书）
ISBN 978-7-300-23027-6

Ⅰ.①大… Ⅱ.①李… Ⅲ.①统计数据-分布式数据-处理-研究生-教材 Ⅳ.①O212

中国版本图书馆 CIP 数据核字（2016）第 140162 号

大数据分析统计应用丛书
大数据分布式计算与案例
主编　李丰
Dashuju Fenbushi Jisuan yu Anli

出版发行	中国人民大学出版社		
社　　址	北京中关村大街 31 号	**邮政编码**	100080
电　　话	010－62511242（总编室）		010－62511770（质管部）
	010－82501766（邮购部）		010－62514148（门市部）
	010－62515195（发行公司）		010－62515275（盗版举报）
网　　址	http://www.crup.com.cn		
	http://www.ttrnet.com（人大教研网）		
经　　销	新华书店		
印　　刷	北京七色印务有限公司		
规　　格	185 mm×260 mm　16 开本	**版　　次**	2016 年 7 月第 1 版
印　　张	9.25 插页 1	**印　　次**	2016 年 7 月第 1 次印刷
字　　数	220 000	**定　　价**	29.00 元

大数据分析统计应用丛书编委会

总 序

一

统计学是收集、分析、展示和解释数据的方法性质的一门科学。信息技术的蓬勃发展，使统计在经济、社会、管理、医学、生物、农业、工程等领域有了越来越多、越来越深入的应用。2011 年 2 月，国务院学位委员会第 28 次会议通过了新的《学位授予和人才培养学科目录 (2011) 》，将统计学上升为一级学科，这为统计学科建设与发展提供了难得的机遇。

一般认为，麦肯锡公司的研究部门——麦肯锡全球研究院 (MGI)，在 2011 年首先提出了大数据时代 (age of big data) 的概念，并引起了全球广泛的反响。大数据是指随着现代社会的进步和信息通信技术的发展，在政治、经济、社会、文化等各个领域形成的规模巨大、增长与传递迅速、形式复杂多样、非结构化程度高的数据或者数据集。它的来源包括传感器、移动设备、在线交易、社交网络等，其形式可以是各种空间数据、报表统计数据、文字、声音、图像、超文本等各种环境和文化数据信息等。大数据时代是一个海量数据开始广泛出现、海量数据的运用逐渐普遍的新的历史时期，也是我们需要认真研究与应对的一个新的社会环境与社会形式。

大数据时代对统计专业的学生提出了更高的要求。他们不仅需要具有扎实的统计理论基础，并且要熟练掌握各种处理大数据和统计模型分析的计算机技能，还要懂得如何提出研究问题、如何判断数据质量、如何评价模型和方法，以及如何准确清晰地呈现分析结果。这对统计教育和人才培养提出了新的目标和方向。

二

顺应时势，在教育部全国应用统计专业学位研究生教育指导委员会推动下，由中国人民大学、北京大学、中国科学院大学、中央财经大学、首都经济贸易大学五所高校发起，集中统

计学科、计算机学科、经济与管理学科的相关学院优势, 依托应用统计专业硕士项目, 组建了北京大数据分析硕士培养协同创新平台。2014 年 9 月首届实验班正式招生并开始授课。

实验班每年招收约 50 ~ 60 名学生, 分别来自中国人民大学、北京大学、中国科学院大学、中央财经大学、首都经济贸易大学等院校。他们均是以优异成绩进入上述高校应用统计硕士项目的本科毕业生, 对大数据分析有浓厚的兴趣, 励志为大数据分析领域的发展做出贡献。

大数据分析硕士的培养是为了满足政府部门和企业等用人单位利用大数据决策的需求, 其核心竞争力是快速部署从大数据到知识发现和价值的能力, 培养方案与国际接轨, 核心内容是面向大数据的统计分析和挖掘技术。经过前期的充分论证, 大数据分析硕士培养方案确定了核心必修课与分方向的选修课。必修课的重点内容为统计学和计算机科学的交叉部分, 侧重于培养从大数据到价值的实践能力, 包括大数据分析必备的计算机基础技能、面向大数据分析的计算机编程能力、大数据统计建模和挖掘能力。每门必修课均配备了 5 人以上的教学团队, 由包括国家 “千人计划” 入选者、长江学者、国家杰出青年基金获得者在内的在相关领域有较高造诣的中青年学者组成。

大数据分析硕士培养协同创新平台是一个面向政府部门和企业等大数据分析人才需求单位开放的平台, 目标是建成一个政产学研有机融和的协同创新平台。2014 年 5 月 19 日平台成立大会就汇集了《人民日报》、新华社、中央电视台、中国移动、中国联通、中国电信、全国手机媒体专业委员会、SAS (北京) 有限公司、华闻传媒产业创新研究院、北京华通人商用信息有限公司、龙信数据 (北京) 有限公司等, 成为该平台的第一批实践培养和研发基地。在 2014 年 9 月开学典礼上又有中国科学院计算机网络信息中心、中国中医科学院、商务部国际贸易学会、国家食品安全风险评估中心、北京商智通信息技术有限公司、史丹索特 (北京) 信息技术有限公司、北京太阳金税软件技术有限公司、北京京东叁佰陆拾度电子商务有限公司、北京知行慧科教育科技有限公司、中关村大数据产业联盟、艾瑞咨询集团 11 家单位加入平台建设的联盟协作单位。实际部门的踊跃参与说明大数据分析人才培养的巨大发展空间。为了加强大学与实际部门专家的双导师制度, 开学典礼上为第一届实验班专门聘任了 26 名实际部门专家担任硕士研究生指导教师。

2015 年 1 月 15 日, 大数据分析硕士培养协同创新平台联合京东、奇虎 360、艾瑞咨询集团、华通人等多家公司举办了针对学生实习的宣讲会。会后组织学生到各相关部门进行有关数据挖掘、大数据分析的实习工作, 学生们得到了锻炼。

为活跃学术氛围, 拓展学生视野, 大数据分析硕士培养协同创新平台组织了大数据分析学术系列讲座, 邀请学界、业界相关人士交流分享学术、行业前沿的经验, 共同推进大数据人才培养以及学术成果的转化。

三

迄今为止, 五校联合大数据分析硕士实验班已经成功开展两届。在此基础上, 课程组全体教师及时收集学生反馈意见, 积极组织讨论, 联合中国人民大学出版社, 启动了《大数据分析统计应用丛书》的编写工作。

本套丛书第一期出版四本。《大数据分析计算机基础》着重介绍数据分析必备的计算机

技能, 包括 Linux 操作系统与 Shell 编程, 数据库操作与管理; 面向大数据分析的计算机编程能力, 我们重点推荐了 Python 语言。《大数据探索性分析》的内容包括大数据抽样、预处理、探索性分析、可视化以及时空大数据案例。《大数据分布式计算与案例》介绍了单机并行计算以及 Hadoop 分布式计算集群, 在此基础上介绍了 HDFS 文件管理系统以及 MapReduce 框架、各种统计模型的 MapReduce 实现, 此外还介绍了处理大数据最常使用的 Hive, HBase, Mahout 以及 Spark 等工具。《大数据挖掘与统计机器学习》介绍了常用的统计学习的回归和分类模型、模型评价与选择的方法、聚类和推荐系统等算法, 所有方法均配有 R 语言实现案例, 支持向量机和深度学习方法给出了 Python 实现案例, 最后是两个数据量在 10G 以上的大数据案例分析, 所有的数据和程序均可下载。相信读者在学习本套丛书的过程中, 数据处理与分析能力会得到锻炼和提高。

在丛书第一期的基础上, 我们也在积极策划第二期, 内容包括非结构化大数据分析、大数据统计模型、统计计算与统计优化方法等, 希望可以涵盖更多的数据类型与统计方法。

该丛书面向的读者主要是应用统计专业硕士, 也可以作为统计专业高年级本科生、其他专业的本科生、研究生以及对大数据分析有兴趣的从业人员的参考书, 希望这套丛书可以为我国大数据分析人才的培养奉献我们的绵薄之力。

丛书编委会

前　言

本书的编写受益于中央财经大学联合中国人民大学、北京大学、中国科学院大学和首都经济贸易大学五所高校与政府部门和产业界联合共建的大数据分析硕士培养协同创新平台。我有幸作为该平台主要课程设计和讲授的教师之一，负责大数据分析方向研究生课程“大数据分布式计算”的建设和教学。本教材是以该课程 2014—2015 年的教学内容和讲义辅以教学案例为蓝本编写的。

目前市面上与大数据相关的计算类书籍有很多，但是均面向计算机相关专业人员。有的侧重于大数据分布式平台 Hadoop 或者 Spark 的架构，有的侧重于大数据计算相关计算机语言介绍，有的侧重于大数据平台的系统开发，但是针对大数据分析最为重要和骨髓部分之一的统计模型，相关实践类书籍还相对较少。

本书侧重于统计和机器学习模型在大数据分布式平台的应用，从案例入手，介绍常见统计模型的大数据分布式计算原理。基于单机共享内存背景开发的统计软件很难直接应用于分布式存储的海量数据。对于初学者而言，在大数据平台下，即便是开发简单的回归模型或者逻辑斯蒂模型都非常困难，更不用说复杂的统计、机器学习算法，这直接阻碍了高效的统计模型在大数据中的开发和部署。

考虑到数据相关工作者在企业实际策略开发和建模中 R 语言与 Python 语言是基础语言，为了方便相关读者快速入门，本书的主要语言采用 R 语言和 Python 语言，但是本书中提到的大数据建模思想是不受语言限制的，读者可以根据自己擅长的语言实现相关模型的大数据开发。

与传统的大数据计算类书籍不同，本书的侧重点是统计模型的实际案例解决，因此本书每章均附有较完整的统计案例。考虑到市面上对于大数据平台的搭建和配置书籍已经很多，而且对于企业而言，这样的平台往往已经很完善，本书淡化了该部分，感兴趣的读者可以参考相关书籍或者本书的附录。

本书按照如下结构组织：第 1 章介绍大数据分布式计算的背景和基于 R 语言和 Python 语言的单机并行原理，让读者熟悉分布式的基本概念。第 2 章介绍目前流行的大数据分布式计算框架 Hadoop 的历史、文件存储系统以及大数据分布式计算的各个击破原理，即 MapReduce。与 Hadoop 相关的安装配置参见附录 A。第 3 章介绍现有大数据分布式平台中常见的

统计模型的原理以及案例分析。与之相关的 Mahout 安装和配置参见附录 B。第 4 章以多个案例的形式介绍如何在大数据平台开发常见统计模型。第 5 章介绍分布式文件系统的访问和操作。与此相关的 Hive, HBase 的安装参见附录 C 和附录 D。第 6 章对学有余力的读者介绍 Spark 平台下统计分析的基础, 并配有 Pyspark 使用基础和基于 Scala 语言的案例。附录 E 介绍 Spark 和 Scala 的安装和配置。

在此要特别感谢中国人民大学统计学院吕晓玲老师以及李天博、王小宁、丁维悦、曹昕、李荣庆、王张浩、王高斌同学在本书的编写过程中对文字和内容的大力贡献。感谢参加五校大数据分析方向研究生课程的同学对本书案例的贡献, 他们是成慧敏、陈思聪、陈晞、刘利恒、刘智彬、魏诗韵、吴雅雯、辛思、张楚妍、张诗玉、赵哲汇、郑巧筠、朱述政。没有吕老师和几位同学的协助, 就没有《大数据分布式计算与案例》一书的最终及时成稿。感谢百度大数据部高级工程师康雁飞博士、中央财经大学统计与数学学院方剑和刘静同学对本书的认真校对。

由于编写时间仓促和本人水平有限, 书中的错误和纰漏一定有很多, 恳请读者不吝指出以便作出修正。

李丰

2016 年于北京

目 录

第 1 章　统计分析与并行计算

1.1　并行计算与并行计算机

1.1.1　并行计算概述

在大数据时代, 全球数字数据每两年增加一倍。海量的数据给了数据科学家们更广阔的应用空间。利用丰富的数据分析工具, 从数据中挖掘出有价值的信息, 让决策更加智慧, 让生活更加便利, 数据科学颠覆了人们传统的生活方式和商业模式。同时, 大数据时代的到来使人们从担心没有数据、不知道从哪里获取数据, 变成了面对如此大规模的数据无从下手, 不知道如何应用。数据量的激增对计算速度提出了更高的要求。并行计算这一概念在当下并不新鲜。如果你的个人计算机的中央处理器 (CPU) 是多核的话, 那么你就已经在使用并行计算技术了。近几年我国超级计算机获得了长足发展。2011—2014 年, 中国国防科技大学研制的 "天河二号" 计算机 (见图 1—1) 以每秒 33.86 千万亿次的浮点运算速度, 连续 4 年夺得全球最快计算机称号。这个超级计算机就是一种并行计算机, 它有 32 000 个 Ivy Bridge 处理器和 48 000 个 Xeon Phi, 总计有 312 万个计算核心。把一个复杂的任务分割成若干小任务, 分配给不同的计算核心同时计算, 以达到省时高效的目的, 这就是并行计算技术。

图 1—1　"天河二号" 计算机

并行计算 (parallel computing) 这一概念是与串行计算 (serial computing) 相对应的, 它是一种可一次同时执行多个指令的算法以达到提高计算速度目的的计算过程, 它可以使需要

大量计算的问题分割成若干子问题同时计算, 从而提高计算速度, 节省时间。一般意义上, 并行计算已经和高性能计算 (high performance computing)、超级计算 (supercomputing) 是同义词了, 因为任何高性能计算机和超级计算机都要使用并行计算技术。并行计算有四个层次, 分别是单核指令并行、多核并行、多处理器并行、集群或者分布式并行。

并行计算经过了几十年的发展 (见表 1—1), 最早可以追溯到 20 世纪 50 年代。20 世纪六七十年代超级计算机的兴起进一步推动了并行计算技术的发展。1963 年世界上第一台巨型机 CDC6600 诞生, 共安装了 35 万个晶体管, 运算速度为 1MFLOPS。此时的超级计算机都是共享内存结构, 各个处理器之间在共享的数据上肩并肩工作。80 年代中期, 一种新的并行计算系统由加州理工大学提出, 在其 Concurrent Computation 计划中, 搭建了一台由 64 个 Intel 8086/8087 处理器构成的超级计算机用于科学计算。这套系统是首个由市场上现成的微处理器搭建而成的超级计算机。由此, 人们不再追求单芯片的高性能计算, 转而把着眼点放在并行计算系统的构造上。1997 年 ASCI Red 超级计算机首次打破了每秒万亿次浮点计算的屏障, 从此大规模并行处理 (massively parallel processing, MPP) 占据了高性能计算的主流, MPP 系统开始持续在体量和功率上不断扩展。并行计算的另外一个 "科技树杈" 是集群系统。集群系统发展的源头可以追溯到 20 世纪 80 年代末期。集群系统是用现成的网络连接起现成的计算机构造的并行计算系统。到 21 世纪, 集群系统因为便利性和超高性价比最终取代了 MPP 的地位, 成为并行运算系统的主流。在承载这个大数据时代的数据中心方面, 世界上几乎所有的计算工作都是由集群系统完成的。在个人计算机方面, 并行计算也是时代的主流。当下大多数的个人计算机都是双核或者四核处理器, 甚至手机都已经使用八核处理器。芯片厂商 (如 Intel, AMD, Qualcomm 等) 似乎更愿意通过增加核心的方式来提高处理器的计算性能。这样做的原因是通过并行计算提高处理器的性能要远比提高单个处理器的时钟频率更加高能效, 在计算机更注重便携和续航的趋势下, 这一点显得尤为重要。

表 1—1　　并行计算机演化

年代	结构	代表
第一代, 20 世纪六七十年代	单芯片系统	CDC 系列, IBM 360
第二代, 20 世纪八九十年代	向量处理系统	Cray XMP, Cray YNP, 银河 -I, 银河 -II
第三代, 20 世纪 90 年代	大规模并行处理 (MPP) 体系	IBM SP2, Cray T3D, Intel Paragon, 曙光 2000
第四代, 20 世纪 90 年代	分布共享内存系统	SGI Origin 2000, 银河三号
第五代, 21 世纪初	集群系统	Berkeley NOW, Princeton SHPIMP, HKU Pearl Cluster

在并行计算的进化浪潮中, 并行软件扮演着非常积极的角色。并行程序要比串行程序难写得多, 因为要把程序分割成几个子程序, 而且需要子程序之间的同步与通信, 这是非常困难的事情。现在一些并行程序的标准已经成型了。对于 MPP 和集群系统来说, 在 20 世纪 90 年代中期, 一些应用程序接口就整合为一个统一标准, 称之为 MPI (Multi Point Interface)。对于共享内存式的并行计算系统, 在 90 年代后期也形成了两个标准: pthreads 和 OpenMP。在数据挖掘中常用的程序设计语言和一些应用, 比如 C, Python 以及 R 都不是并行语言, 但是利用一些已有的程序接口且使用并行算法可以大大提高计算效率, 基于 MPI 的编程请参考

罗秋明 (2012)。当然, 并行计算并不是解决计算速度的万能药, 并行计算机发展目前还存在很多问题。首先, 计算机的存储器访问速度与计算速度不平衡。处理器速度每年提高约 60%, 而存储器的访问速度每年只能提高约 9%, 这使得处理器经常 "空转", 数据传输跟不上。利用高速缓存等方式虽然在一定程度上可以缓解数据传输问题, 但是由于其容量有限, 加上高昂的费用, 难以得到广泛应用。其次, 并行计算机的规模越来越大, 结构也越来越复杂, 不同并行计算机结构各异, 处理器数目数以万计, 数据存储方式、通信模式迥异, 同构异构并存, 这使得为超级计算机编写程序更加困难。另外, 并行计算机耗能极大, "天河二号" 每年的电费高达上亿元人民币! 超级计算机占地面积大、功耗高, 不具备广泛使用的条件。

1.1.2　并行计算机

并行计算机是指由一系列处理器协作解决计算问题的计算机。通俗地说, 能执行并行计算的计算机就是并行计算机。1966 年 Michael J. Flynn 首先提出计算机结构按照指令 (程序) 和数据分类可以分为四类: 单指令单数据 (single instruction single data, SISD)、多指令单数据 (multiple instruction single data, MISD)、单指令多数据 (single instruction multiple data, SIMD)、多指令多数据 (multiple instruction multiple data, MIMD)。其中后两种是并行计算机 (见图 1—2)。

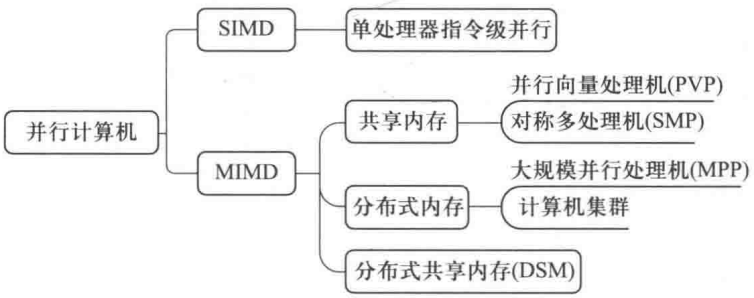

图 1—2　并行计算机分类

SIMD 并行方式是指指令级并行 (instruction-level parallelism, ILP), 是最简单、最基础的并行方式, 它是单处理器并行层级上的并行, 在多核多处理器的并行中, SIMD 存在于每一个核中。MIMD 并行方式是现代主流的、更高级的并行方式。MIMD 按照数据的存储方式可以进一步划分为共享内存 (shared memory) 、分布式内存 (distributed memory) 和分布式共享内存 (distributed shared memory)。所谓共享内存, 是指多个处理器共用同一个内存, 而分布式内存则指每个处理器都有自己的内存, 需要高速网络或者 LAN 进行不同节点之间的通信。分布式共享内存是综合了共享内存和分布式内存的特点, 每个节点都有自己的内存, 同时不同节点共享一个虚拟的地址空间, 这个地址空间映射到每个节点的物理地址。并行向量处理机 (parallel vector processor, PVP)、对称多处理机 (symmetric multi processor, SMP) 是典型的共享内存存储方式的并行计算机系统。大规模并行处理机 (MPP) 与计算机集群 (Cluster) 则是分布式内存的存储方式。

单处理器指令级并行

单处理器指令级并行一般有两种技术: 向量处理机 (vector processor) 与超标量处理机 (superscalar processor) 。向量处理机指一条指令可以对一组数据 (称为向量) 实行相同的操

作的并行计算机。图 1—3 描绘了这种结构。向量处理机只有一个指令执行单元 (instruction execution unit, IEU), 如果当前指令没有执行结束, 下一条指令是不会被执行的。与串行标量处理机不同的是向量处理机一次处理一组数据, 向量处理机使用的寄存器是向量寄存器。向量运算很适合流水线计算机的结构特点。向量型并行计算与流水线结构相结合, 能在很大程度上克服通常流水线计算机中指令处理量太大、存储访问不均匀、相关等待严重、流水不畅等缺点, 并可充分发挥并行处理结构的潜力, 显著提高运算速度。

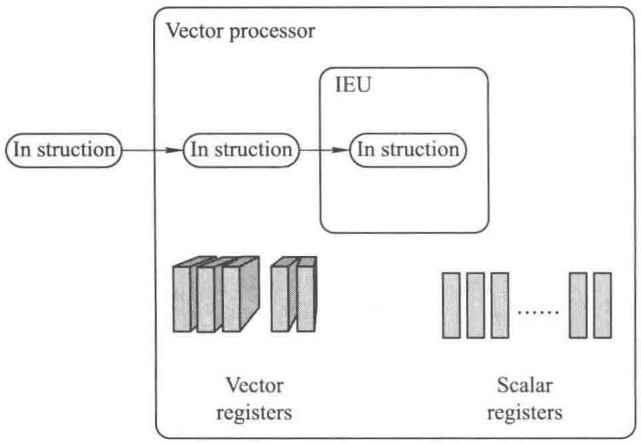

图 1—3　向量处理机结构

资料来源: http://flylib.com/books/en/2.601.1.9/1.

超标量处理机通常指一个时钟周期内能够同时发射多条指令的处理机。为了能够在一个时钟周期内同时发射多条指令, 超标量处理机必须拥有两条或者两条以上能够同时工作的指令流水线和指令执行单元。值得注意的是, 超标量处理机是在单个处理器 (核) 内的指令级的并行, 并不属于 MIMD 的范畴。超标量处理机使用的是标量寄存器。图 1—4 展示了超

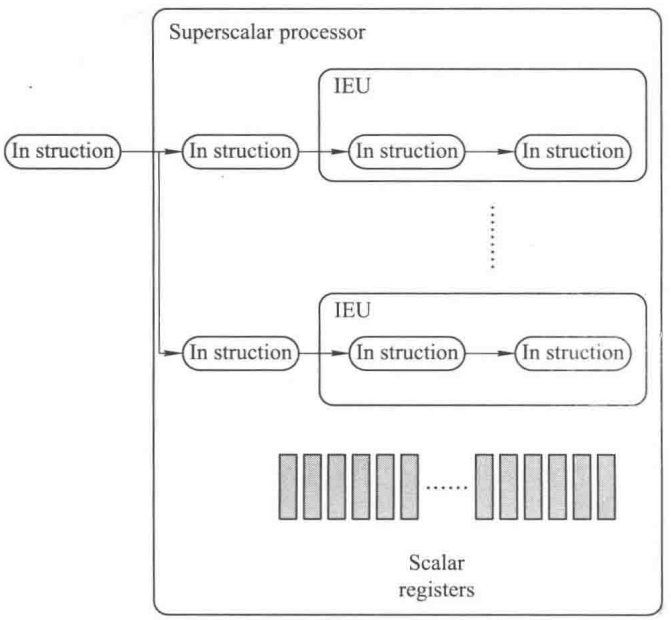

图 1—4　超标量处理机结构

资料来源: http://flylib.com/books/en/2.601.1.9/1/.

标量处理机的结构模型。 1964 年超标量技术首先在 CDC6600 处理机上使用, 而 1969 年 CDC7600 则首次对每条指令使用了流水线作业。当今几乎所有工业级的微处理器都使用了流水线超标量处理技术。

并行向量处理机

并行向量处理机如图 1—5 所示。并行向量处理机系统就是将若干向量处理机并行起来, 使用专门设计的高宽带的交叉开关网络将 VP 连向共享存储模块, 存储器以每秒兆字节的速度向处理器传递数据。并行向量处理机并不使用高速缓存而是使用大量的向量寄存器和指令缓冲器。

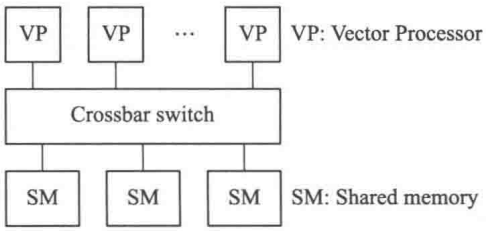

图 1—5　并行向量处理机结构

资料来源: http://slideplayer.com/slide/5677116/.

对称多处理机

对称多处理机使用多个处理器分享内存和系统总线, Cray CS6400 Enterprise Server (up to 64 CPUs), SGI Power Challenge (up to 18 CPUs), HP T-Class Servers (up to 14 CPUs), HP K-Class Servers (up to 4 CPUs), HP D-Class Servers (up to 2 CPUs), HP J-Class Workstations (up to 2 CPUs) 等并行计算机都是基于对称多处理机的结构。对称多处理机的优势在于结构简单, 相关应用的编程比较容易。由于所有的 CPU 共享单一的内存, 使得操作系统很容易在不同的处理器之间分配任务。对称多处理机的弱点在于拓展性较差, 增加新的处理器会增加系统总线的流量负担, 一旦流量饱和, 除非增加系统总线带宽, 增加处理器将不能提高计算速度。主流的多核处理器大多是 SMP 结构的。

大规模并行处理机

MPP 与 SMP 的差别在于: MPP 的处理器之间并不共享内存, 而是分布式内存, 也就是说每一节点都是由一个或者多个处理器、高速缓存、内存组成, 不同节点之间通过系统总线或者高速网络连接。当今大规模并行处理机的示意图如图 1—6 所示。IBM SP 和曙光 2000 使用的都是 MPP 架构。MPP 的优点在于拓展性较好, 不存在内存—总线瓶颈问题。随着处理器的增加, 内存、I/O 容量都可以成比例增加。MPP 对系统总线的流量要求较低, 因为系统总线不需要承担 CPU 与内存之间的数据交换, 但是由于对比 SMP, MPP 各个节点的独立性更强, 所以编程起来更加复杂, 在源代码中, 内存与处理器之间需要消息传递指令, 程序的源代码对硬件更加依赖, 程序的可移植性变差。并行虚拟机 (parallel virtual machine, PVM) 以及标准信息传递接口 (MPI) 的使用可以解决这个问题。还有一种与大规模并行处理机类似的并行计算机系统叫做工作站集群 (cluster of workstations, COW), 二者之间的区别在于节点的通信。COW 使用的是标准的 LAN 互连, 而 MPP 使用的是高速、专用高宽带、低延迟的互连网络。在现代并行计算机中广泛使用的计算机集群中, 主要使用的就是 MPP 和