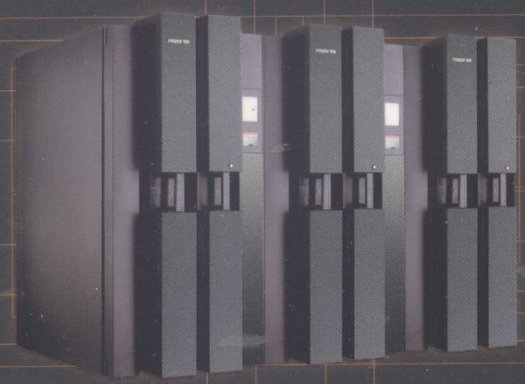
inspur 浪潮

高效能计算机系统设计与应用

王恩东 等 著



- 国内**第一本**系统论述**高效能计算机**系统的专业图书，**国家重点实验室**科研团队集体智慧的结晶
- 紧密围绕**工程实例**展开论述，深入剖析**异构并行计算**原理和特征，全面总结面向应用特征分析的**方法论**

科学出版社

高效能计算机系统设计与应用

王恩东 等 著

科学出版社

北京

内 容 简 介

高效能计算机是解决大规模复杂计算任务的重要科研工具。本书系统地介绍高效能计算机系统的设计原理与应用案例。全书共 9 章。第 1、2 章分别介绍高效能计算机的基本概念和发展现状；第 3 章阐述高效能计算系统构建的一般原则；第 4、5 章结合实例详细论述十万亿次级别和百万亿次级别高效能计算系统的设计方法；第 6 章介绍高效能计算应用运行环境的构建与配置；第 7 章介绍高效能计算系统并行应用软件的开发策略、方法和优化等；第 8 章介绍高效能计算的综合评测体系；第 9 章展望高效能计算机的发展未来，分析了百亿亿次级别高效能计算机的实现前景。

本书由高效能服务器和存储技术国家重点实验室组织编写，是我国第一本系统论述高效能计算机系统的专著。本书主要面向从事高效能计算的程序员、工程师等科研技术人员，也可作为高等院校计算机科学与技术等专业开设相关课程的教材。

图书在版编目(CIP)数据

高效能计算机系统设计与应用 / 王恩东等著. —北京：科学出版社，2014.4
ISBN 978-7-03-040281-3

Ⅰ. ①高… Ⅱ. ①王… Ⅲ. ①计算机系统－系统设计 Ⅳ. ①TP302.1

中国版本图书馆 CIP 数据核字(2014)第 055270 号

策划编辑：张 濮 / 责任编辑：张 濮 / 责任校对：刘小梅
责任印制：张 倩 / 封面设计：迷底书装

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com

骏主印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2014 年 4 月第 一 版 开本：787×1 092 1/16
2014 年 4 月第一次印刷 印张：11 3/4
字数：306 000

定价：52.00 元

(如有印装质量问题，我社负责调换)

序　言

20 世纪 40 年代问世的数字计算机已经渗透到人类社会的每个角落，使人们的工作和生活方式发生了翻天覆地的变化。大到上天入地，小到娱乐购物，到处都有计算的印记。科学研究是最早应用计算机的领域之一。计算已经和传统的理论分析与实验观察一道成为人类从事科学研究的基本方法。基于计算的模拟有助于人类正确认识客观世界，它可以加快原本漫长的过程，探索宇宙的起源和演变；可以把转瞬即逝的事件放慢，研究核裂变和核聚变的过程；可以在原子水平探索物质的微观结构，依据第一原理进行新材料的设计；也可以揭示宏观世界的发展规律，预测千年气候的变化。复杂系统的模拟需要巨大的计算能力，因此高性能计算机、高效的计算方法、成熟的科学计算与模拟仿真软件已经成为现代科学研究的必备利器。

随着计算技术的发展与普及，高性能计算机的应用已不局限于基础科学研究等少数领域，它正在经济建设、社会发展、人民生活的方方面面发挥着重要作用。在油气勘探、新能源、飞机设计优化、大型流体机械节能、汽车碰撞模拟、高速列车设计、大型工程项目的模拟仿真等领域，都有高性能计算的用武之地。高性能计算也和人们的生活息息相关，精准数值天气预报、基于计算机模拟的新药发现、减灾防灾、环境污染治理、大规模信息服务、数据挖掘与知识发现、数字媒体与文化创意等，都需要高性能计算机的支持。

应用是高性能计算发展的根本动力，正是各行各业对计算不断增长的需求推动着高性能计算技术的持续进步。同时，高性能计算应用的开发、实施、推广和普及需要千千万万的高素质人才，人才是促进高性能计算应用进步的第一要素。审视我国目前高性能计算发展现状，缺乏自主研发的大型高性能计算应用软件，缺少具有跨学科知识和视野的高水平人才，是制约我国高性能计算事业发展的两大瓶颈因素。

这本论述高性能计算的专著正是在这样的大背景下应运而生的，它由参加了世界最快计算机“天河 2 号”研发的浪潮技术团队撰写。内容涵盖高效能计算机系统设计和应用两大部分，包括高效能计算机的系统架构、高效能计算系统构建的基本原则、十万亿次至百万亿次高效能计算系统的设计方法、高效能计算机应用环境构建和大型并行应用软件开发、高效能计算系统的综合评测体系，以及对百亿亿次级高效能计算机的展望等多个方面。特别针对目前流行的并行异构体系结构，探讨了 CPU+GPU 和 CPU+MIC 两种异构并行软件的开发技术。此外，该书还介绍了计算化学、计算生物学、计算流体力学、分子动力学、计算材料、大气科学、石油地质勘探等领域的典型应用软件，以及这些软件与计算机结构和性能之间的关系。这些内容丰富而实用，是作者多年实践经验的结晶。该书不仅可以作为从事高性能计算研发的科技人员的参考书，也可以作为相关专业的本科生、研究生的教材。

1993 年，浪潮集团推出其第一台基于 Intel 486 处理器的并行小型机，迈出了浪潮服务器产业的第一步。现在，浪潮集团已经在高性能服务器、容错计算机、海量存储设备、云计算、大型应用软件等多个领域取得了令人瞩目的成就，成为有重要影响力的大企业。特别是与国防科学技术大学合作，成功研制了世界上最快的高效能计算机“天河 2 号”，这标志着浪潮集

团的技术水平上了一个新台阶。当然，与 IBM、英特尔、微软等国际信息技术产业巨头相比，浪潮集团在引领技术发展方面还有很长的路要走。可喜的是，浪潮集团近年来依托其国家重点实验室不断推出学术型、技术型的著作，这正反映了浪潮集团在引领技术发展上的努力，该书就是这种努力的具体体现。衷心希望浪潮集团能不断推出新的高水平学术著作，为我国的科学技术进步和新兴产业发展作出应有的贡献。

钱德沛

2013 年 12 月 16 日

于北京航空航天大学

前　言

高效能计算是信息领域的前沿高新技术。随着信息化社会的飞速发展，人们对信息处理能力的要求越来越高，高效能计算已在科学研究、石油勘探、气象预报、航天国防、生命基因等领域得到广泛应用，金融、政府、企业、互联网等新兴应用对高效能计算的需求也在迅猛增长。目前，百亿亿次级别的高效能计算系统已经处于研制过程中，我们期待采用百亿亿次计算机解决更大规模和更复杂的问题。

从“天河 1 号”超级计算机系统问鼎全球超级计算 Top500 榜首，到全国产化的千万亿次“神威蓝光”问世，再到天河 2 号两次问鼎 Top500 榜首，中国在超级计算机和高效能计算技术方面不断取得突破，计算系统的峰值计算能力已经处于国际领先水平。但由于在处理器等核心技术方面的落后，以及应用、专业人才的欠缺，我国还不是高效能计算的强国。应用的落后与相关人才的缺乏已成为中国高效能计算产业发展的瓶颈，我国在高效能计算知识的普及和传播、计算人才的培养和储备等方面还有很多工作要做。

为了推进高效能计算在中国得到更好的普及和发展，特别是在对高效能计算感兴趣的科研人员以及开设了相关课程的高校中间广泛传播高效能计算知识，培养高效能计算人才，浪潮集团与高效能服务器和存储技术国家重点实验室先后举办了中国大学生超级计算机竞赛和亚洲大学生超级计算机竞赛等活动，已初步在全国乃至亚洲范围形成了大学生学习和关注高效能计算的良好氛围。参赛的师生们普遍反映目前还缺少一本深入浅出的书籍来系统地介绍高效能计算知识。

关于本书

第 1 章从高效能计算的诞生历程回顾开始，介绍高效能计算的定义、发展、应用，以及高效能计算的架构分类和目前发展遭遇的瓶颈，并重点介绍异构并行技术的兴起和新型众核芯片的发展。

第 2 章介绍当前国际上领先的千万亿次高效能计算机，以及中国的千万亿次计算机天河 1 号、天河 2 号与神威蓝光，给读者世界顶尖高效能计算系统的直观印象。

第 3 章阐述高效能计算系统构建的一般规则，包括系统的组成结构，如何构建性能均衡的高效能计算系统，以及如何对它进行管理和调度。

第 4 章和第 5 章阐述高效能计算系统的设计方法，从较为简单的十万亿次级别的高效能计算系统到比较复杂的百万亿次级别，涵盖系统设计的绝大部分内容，具有较强的实用性。

第 6 章介绍如何构建高效能计算机的应用环境，包括硬件环境、软件环境和开发环境，同时也介绍了异构系统的应用环境配置方法。

第 7 章介绍高效能计算系统并行应用软件开发的知识，包括同构系统和异构系统的计算模式、开发策略、开发方法和优化方法等，并有实际案例供读者参考。

第 8 章介绍高效能计算的综合评测体系，对高效能计算系统硬件性能的评测方法、应用性能的评测方法和预测方法作了结构性说明，同时以 NAMD、VASP 等应用加以实证。

第 9 章展望了高效能计算的未来与发展。

本书力求以通俗的语言科普高效能计算的知识，本着简洁实用的原则，基于大量的实践经验并结合具体案例进行论述，期望能够对高效能计算机的建设者、使用者和管理者有所裨益。

致谢

对于本书的编写工作，多位在高效能计算领域有丰富的理论与实践经验、见地深刻的专家给予了无私的支持。北京航空航天大学钱德沛教授百忙之中为本书作序，西安交通大学董小社教授、伍卫国教授尽心审阅书稿并提出中肯意见。本书的编写小组成员通力合作，将从业多年积累的知识和经验进行整理归纳，从而完成了这本书，在此一并感谢！

本书编写组成员有王恩东、刘军、魏健、王渭巍、刘羽、张清、王娅娟、张广勇、沈铂、吕文静、金莲、陈博文等。

王恩东

高效能服务器和存储技术国家重点实验室主任

目　录

第 1 章　高效能计算机的系统架构

1.1　高效能计算机发展与系统分类

1.1.1　HPC 的定义

高效能计算机也称为高性能计算机(High-Performance Computer，HPC)，指信息处理能力尤其是计算速度超过普通计算机的机器，广义上是指解决大型复杂任务的计算工具。现代高性能计算机由成千上万个处理器组成，包括定制的高性能处理器和加速部件等，共同完成普通单台个人计算机(PC)和服务器不能完成的大型复杂运算任务。高性能计算机的概念来源于超级计算机(super computer)。

超级计算机诞生的初衷是解决大规模计算求解问题。计算科学与传统的两大科学——理论科学和实验科学，被并列认为是人类认识自然的三大支柱。在许多情况下，当理论模型复杂甚至理论尚未建立，或者实验费用昂贵甚至无法进行时，计算就成为求解问题的唯一或主要手段。超级计算机具备超强的处理能力，因此人们可以通过超级计算机进行数值模拟来预测和解释无法通过实验检验的原理和现象，如天气预报、核爆模拟和基因比对等。

以天气预报为例，在超级计算机没有诞生之前，要想知道天气变化，需要专人去布置卫星和探测器等，捕捉大气变化，再根据自身经验来判断。而使用超级计算机后，结合其他监测仪器发回的数据就可以模拟大气变化，自动分析出天气变化情况，既能提高气象预报的准确性，又能提升时效性。这样，人们经常听到的“局部地区有雨”等字眼在超级计算机的帮助下也会逐步消失，因为超级计算机能精准地计算出哪个地区什么时间会下雨。

例如，要研发飞机，飞机发动机的涡轮叶片是一种耐高温的材料，这种材料一般都是由四五种金属材料还有一些其他材料合成的。从已知的 100 多种材料里选出这 5 种材料的组合，至少有 9×10^9 种可能；这还不够，每种元素还有不同的比例组合，这就是个无穷尽的问题；再加上不同温度和压强的外界条件，如果没有超级计算机，而是通过实验来尝试，那么周期和成本将是不可想象的。怎样实验才是高效的呢？只有虚拟实验，即在计算机上建立能够反映该材料状态的准确模型，通过计算机软件模拟，这里的硬件载体就是超级计算机。

之所以选择超级计算机作为这些复杂运算的载体，是因为超级计算机的计算速度是相当惊人的。例如，我国研发的亿亿次超级计算机天河 2 号的峰值运算速度为每秒 33.86 千万亿次浮点运算(PFLOPS)。天河 2 号运算 1 小时相当于 13 亿人同时用计算器计算 1000 年，其存储总容量为 12.4 PB，相当于存储 600 亿册每册 10 万字的图书。

当今，超级计算机的效能比越来越被重视，于是高效能计算机的概念随之兴起。

1.1.2　HPC 的诞生

超级计算机的历史可追溯到 20 世纪 60 年代，当时一系列由西摩·克雷(Seymour Cray)设计和实现的具有卓越计算峰值性能的计算机在控制数据公司(CDC)诞生。CDC 6600 诞生

于 1964 年，通常被认为是第一台超级计算机。1972 年，西摩·克雷离开 CDC，创办了克雷公司(Cray)；1976 年，他设计了 80 MHz 的Cray-1超级计算机，如图 1-1 所示，这是当时最成功的一台超级计算机。

1985 年Cray-2诞生，如图 1-2 所示，当时包含了 8 个处理器，执行速度达到 1.9 GFLOPS，直到 1990 年它都是世界上运行速度最快的计算机。

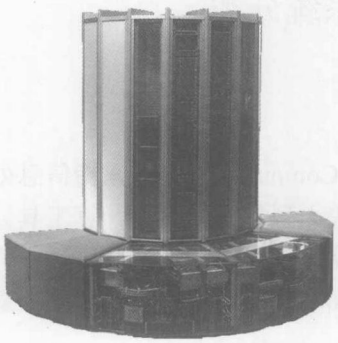


图 1-1　Cray-1 超级计算机

图 1-2　Cray-2 超级计算机

20 世纪 80 年代的超级计算机只包含几个处理器，而 20 世纪 90 年代在美国和日本就出现了包含上千个处理器的计算机，创下当时计算机处理能力的新世界纪录。

1.1.3　HPC 发展与应用

超级计算机的发展经历了从向量机、对称多处理机、大规模并行处理机到集群的体系架构的变化；在超级计算机互连网络方面，早期向量机和大规模并行处理机都是定制的专用互连网络，采用商品化部件的集群超级计算机系统出现后，集群互连网络从快速以太网发展到千兆位以太网，再到目前流行的 InfiniBand 网络等。

1. 向量机

向量机(PVP)主要集中在 20 世纪 70 年代，主要代表产品是克雷公司 CDC 系列和 Cray 系列，如 CDC 6000、CDC 7000、Cray-1 和 Cray-2 等。当时最快的计算机都是向量机。向量机是由专门设计定制的向量处理器(VP)经专门设计的高带宽交叉开关网络和共享存储模块互连组成的计算机系统。

2. 对称多处理机

对称多处理机(SMP)系统使用商用的微处理器经高速总线连接共享存储器组成。该系统是对称的，每个处理器都可以等同地访问共享存储器、I/O 设备和操作系统服务，典型代表机型有 IBM R50、SGI POWER Challenge 等。

3. 大规模并行处理机

大规模并行处理机(MPP)和对称多处理机一样，也采用商用的微处理器，不同之处是系统有物理上的分布式存储器，程序由多个进程组成，每个进程都有其私有的地址空间，进程间通过消息传递进行交互。

20 世纪 80 年代末 90 年代初，许多公司对 MPP 进行了实践，比较著名的有思维机器公

司(Thinking Machines Corporation)、nCUBE、KSR(Kendall Square Research)、英特尔和 IBM。

思维机器公司成立于 1982 年，其生产的 CM(Connection Machine)系列高性能计算机为 MPP 的互连进行了一系列的探索。1986 年思维机器公司发布的 CM-1 采用 SIMD 方式，包含了 65536 个 1 位的处理器，每个处理器有 4 Kbit 的存储器，连接成超立方体(hypercube)结构。在 CM-1 推出一年后，该公司又推出了 CM-1 的升级版本——CM-2，CM-2 被认为是第一台成功的大规模并行处理计算机。

英特尔公司在 1971 年推出世界上第一款微处理器 4004，如图 1-3 所示，在微处理器市场上获得了成功。1984 又创建了 Intel Scientific Computers(后改名为 Intel Supercomputer Systems Division)，研究基于超立方体结构的大规模并行处理计算机。1985 年推出了 iPSC/1，由以太网控制器连接的 80286 微处理器组成。两年后推出了 iPSC/2，微处理器升级为 80386/7，如图 1-4 所示。1990 年，英特尔公司推出了第二代基于超立方体结构的高性能计算机 iPSC/860，它采用了 128 个 i860 微处理器，以及电路交换(circuit-switched)的连接方式。由于其具有出色的性价比(800 MFLOPS 每百万美元)，获得了戈登·贝尔性价比奖(Gordon Bell prize in price/performance category)。

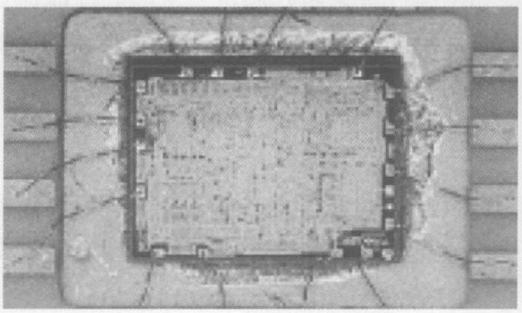


图 1-3　Intel 4004 微处理器

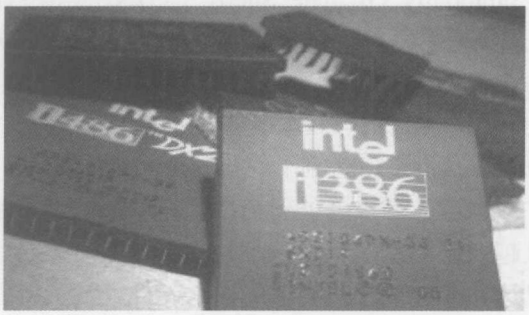


图 1-4　Intel 386 & 486 微处理器

针对大规模并行处理系统的发展趋势，IBM 也于 1991 年启动了 SP(Scalable Power parallel)项目，并于 1993 年公布了第一个产品 9076 SP1。SP1 使用 RS/6000 处理器，通过一个 8×8 的交换芯片进行连接，因此系统最大包含 64 个处理器。由于连接方式限制了系统的规模，SP1 系统的浮点计算能力小于 10 GFLOPS。在设计下一代产品时，IBM 综合考虑了性价比、通用性、可用性等各方面因素，于 1994 年推出了采用集群(cluster)体系结构构建的大规模并行处理计算机 9076 SP2。

4. 集群计算机

集群计算机是指利用商品化高速通信网络将一组高档工作站或服务器甚至个人计算机按某种结构连接起来，在并行程序设计与可视化人机交互集成开发环境的支持下，统一调度，协调处理，实现高效并行工作的系统。

20 世纪 90 年代中期，随着微处理器和动态随机存储器(Dynamic Random Access Memory，DRAM)速度的提升以及 PCI(Peripheral Component Interconnect)总线的出现，个人计算机市场日趋成熟。随着互联网的拓展，局域网(Local Area Network，LAN)技术飞快发展，1982 年出现的 10 Mbit/s 以太网(Ethernet)已经渐渐被 100 Mbit/s 快速以太网(fast Ethernet)所代替，

同时一些专用系统网络(System Area Network，SAN)如 Myrinet 也发展迅速，在带宽和延迟上与传统高性能计算机所采用的专有网络的差距日渐减小。

在软件方面，1991 年出现的 Linux 操作系统借助互联网获得成千上万开发者的支持，到 1994 年已经相当稳定。1989 年，PVM(Parallel Virtual Machine)发布，这是第一个通用的基于消息传递方式的并行环境，PVM 可以使用互连的计算机构造一个虚拟的并行机。但由于 PVM 是免费的，只有少数厂家进行维护，而 1993 年发布的 MPI(Message Passing Interface)功能与 PVM 相似，但 MPI 是由国际组织维护的国际标准，许多厂商都为其提供具体的实现版本。因此，最终 PVM 停止开发，MPI 成为名副其实的消息传递接口标准。

处理器、网络、操作系统和软件的高速发展推动了集群技术的发展，第一台集群系统源自美国航空航天局(National Aeronautics and Space Administration，NASA)Goddard 航天中心的 Beowulf 项目。Goddard 航天中心的地球与空间科学(Earth and Space Science，ESS)项目需要一台能够处理大数据的高性能计算机，要求其具备 1 GFLOPS 的峰值和 10 GB 的存储能力，而价格却不能高于用于高端科学计算的工作站的价格。无奈之下，NASA 的唐纳德·贝克(Donald Becker)和托马斯·斯特林(Thomas Sterling)只能选择市场上可以买到的商用现货(Commodity Off The Shelf，COTS)的微型计算机和网络硬件来组建满足要求的高性能计算机，于是 Beowulf 项目产生了。1994 年名为 Wiglaf 的第一台 Beowulf 集群在 Goddard 航天中心诞生，它由 16 台 100 MHz 的 486DX4 微机通过 10 Mbit/s 以太网集线器连接而成，在实际的应用中每个节点(一般指集群中具有独立 I/O 资源的计算硬件单位)的计算速度达到 4.6 MFLOPS，整体达到 74 MFLOPS。

集群计算机是处理器技术和网络技术不断提高的产物。商业处理器运算速度飞速提高而且越来越便宜，网络技术的进步使得商业网络的带宽已经很高。目前，很多商业网络的带宽已达 Gbit/s 量级。高速的网络硬件再加上特殊设计的网络协议，其传输速度已能达到甚至超过某些 MPP 专门定制的网络，这就为并行计算的通信提供了有力的保障。集群系统在被提出之后发展十分迅猛，已成为目前研究的热点。在 2013 年 6 月发布的 Top500 排行榜中，以集群方式构建的机器为 417 台，占总数的 83.4%，MPP 架构为 83 台，占总数的 16.6%。

由于集群计算机具有投资风险小、可扩展性好、可继承现有软硬件资源、开发周期短、容易编程等突出特点，目前已成为并行处理的热点和主流。集群中的节点数越多，系统的整体处理能力就越强，但节点数的增多受限于消息传递的通信速度和带宽。对于集群系统，可用的节点有工作站、个人计算机、SMP 服务器甚至超级计算机。节点的操作系统是多用户多任务多线程系统，如 Linux 等。节点可以是同构的，也可以是异构的。

集群计算机系统作为当前世界上并行处理的热点和主流，具有许多其他系统不可替代的优势：性价比高、可扩展性好、可用性高和能用性高。尤其是个人计算机并行集群系统有系统开发周期短、用户投资风险小、节约系统资源、用户编程方便等优点，它将给各行各业提供性价比较高的高性能并行计算资源。

随着集群技术的出现，高性能计算机从 20 世纪 90 年代开始得到迅猛发展，计算能力在 1998 年首次突破万亿次，到 2008 年达到千万亿次，10 年间提高 3 个数量级。在 2013 年 6 月 Top500 排名中，中国的天河 2 号以 33862 TFLOPS 的运算速度夺冠，而 Top500 排行榜中前 26 位均已进入千万亿次时代。业界预计 2018—2020 年高性能计算机将会进入百亿亿次(ExaScale)时代。

5. 高性能计算机的应用

随着高性能计算机的发展，其应用也越来越广泛，目前主要运用在许多工业领域，如汽车和航空航天器的设计制造、石油勘探、地震资料处理与核爆炸模拟等。在教育、科研领域，高性能计算机有着更广泛的应用空间，在生命科学、基因比对、材料设计、气象气候研究、宇宙演变、量子物理学、计算化学、分子模型和天体物理模拟等学科中已成为科学研究的必备工具。

例如，在天气预报和台风预报等大量预报工作中，超级计算机将大大提高预报的准确率和及时性。目前，国家超级计算济南中心正和气象部门共同研发新的气象预报系统，原来的计算系统只能预测 2h 内的气象变化，应用超级计算后，半小时甚至十几分钟就能预报一次。此外，超级计算可将气象预报的范围精确到 2 km 以内，这意味着可为爬泰山的人提供山上和山下不同区域的温度预报。

6. 高性能计算机的评估组织

随着高性能计算机的发展，越来越多的国家和组织参与到高性能计算领域，并且出现像 Top500、Green500 和 Graph 500 等专门针对高性能计算机进行评估的组织。一年一度的国际超级计算机大会(ISC)和超级计算机大会(SC)则为 HPC 研究人员、技术领导者、科学家和信息技术决策者提供了一起研究 HPC 解决方案的平台，同时在大会期间举办的国际大学生超级计算竞赛有助于超级计算人才的国际交流与互动，激发年轻一代大学生对超级计算的热情，加速超级计算应用人才储备，推动超级计算更快更好地发展。

全球 Top500 超级计算机排行榜诞生于 1993 年，其主要目的是以此为基础来对高性能计算领域的发展趋势进行跟踪和监测。该排行榜对全球范围内性能最强劲的 500 套计算机系统进行排名，每年更新和发布两次，6 月在德国的 ISC 上发布，11 月在美国的 SC 上发布。

Green500 排行榜由美国弗吉尼亚科技大学(Virginia Tech)的 Wu Chunfeng 在 2007 年创立，主要关注超级计算机不断增长的能耗，对当前的 Top500 世界超级计算机以能效标准进行重新排名，该榜单半年更新一次。它与 Top500 一样，会在每年 6 月和 12 月各发布一次。

Graph 500 是美国桑迪亚国家实验室与英特尔、IBM、AMD、NVIDIA 和 Oracle 合作定义并发布的一个新的基准测试排名，它是利用图论分析超级计算机在模拟生物、安全、社会以及类似复杂问题时的吞吐量，并进行排名。Graph 500 排行榜旨在测试关于这些应用类型的系统能力，而不是只看其计算能力。希望通过这个测试促使计算机厂商构建可处理复杂问题的架构。

7. 中国高性能计算机现状

高性能计算机技术日渐成为衡量一个国家科技水平与综合国力的重要标志之一，目前世界上一些发达国家都在争相投入巨额资金对它进行开发和研究。我国从“天河 1 号”超级计算机系统问鼎全球超级计算机 Top500 榜首，到全国产化的千万亿次“神威蓝光”问世，直到 2013 年 6 月 17 日天河 2 号再次问鼎 Top500 榜首，中国在超级计算机研制方面不断取得突破，超级计算机峰值能力已经处于国际领先水平。2010 年 11 月，天河 1 号凭借 2.57 PFLOPS

的运算速度荣登 Top500 排名第一的时候，这一消息当时震惊了世界，美国总统奥巴马在讲话中把它当成中国重要的科技成就之一，并用它来激励美国人奋发向上。我国国家超级计算济南中心装配的神威蓝光系统是国内首台全部采用国产自主中央处理器和系统软件构建的千万亿次计算机系统，标志着我国成为继美国、日本之后能够采用自主中央处理器构建千万亿次计算机的国家。

1.1.4　HPC 架构分类

1. SMP 小型机

小型机是指运行原理类似于个人计算机和服务器，但性能和用途又与它们截然不同的一种高性能计算机，它是 20 世纪 70 年代由数字设备公司(DEC)首先开发的一种高性能计算产品。

SMP 小型机的主要特点是所有的 CPU 共享全部资源，如总线、内存和 I/O 系统等，但只有一个操作系统，如图 1-5 所示。SMP 小型机具有高运算处理能力(high performance)、高可靠性(reliability)、高服务性(serviceability)、高可用性(availability)，以商务计算为主，易编程，可扩展性差。

典型的 SMP 小型机有 IBM System p5 510Q、HP 9000 rp3440-4(PA-8900)、曙光天演 EP850-G20、浪潮 SMP2000。

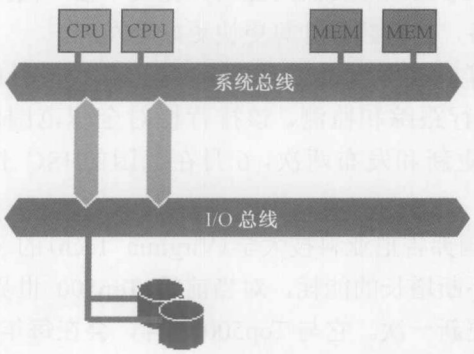


图 1-5　SMP 小型机结构图

2. 大规模并行处理机

大规模并行处理机(MPP)的计算节点采用商用的微处理器，具有物理上的分布式存储器，程序由多个进程组成，每个进程都有其私有的地址空间，进程间通过消息传递进行交互，如图 1-6 所示。MPP 的主要特点是：较 SMP 小型机更易扩展到较大规模，但编程不易，性价比、通用性、可用性相对较低，曾停止发展，每个节点只访问本地内存和存储器，节点之间的信息交互与节点本身的处理是并行进行的，是一种完全不共享的方式。

MPP 的典型机器有 Cray T3E、神威蓝光、曙光-1000、泰坦(2012 年 11 月 Top500 排名第一)、红杉(2012 年 6 月 Top500 排名第一)等。

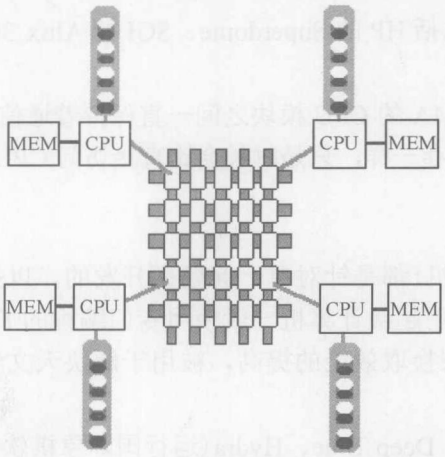


图 1-6　MPP 结构图

3. 向量机

向量机使用超大规模集成电路技术、高密度组装和光电子技术、可扩展技术，共享虚拟存储空间。并行向量机依靠单纯地提高向量处理器的数量来扩展规模，一方面交叉开关网络实现非常庞大，另一方面由于定制的向量处理器价格昂贵，维护费用也比较高，所以并行向量机的性价比迅速下降。然而，超级计算机一味依赖于政府和军方的支持、不计成本、只求性能的时代已经过去，因此向量机走向没落成为必然，目前已不再单独发展。

典型的向量机有克雷公司的 Cray 系列(Cray-1、Cray-2)等。

4. ccNUMA

ccNUMA(缓存一致的非一致性存储访问架构)多处理机模型是将一些 SMP 机器作为一个单节点而彼此连接所形成的一个较大的系统。不同节点内的 CPU 可以访问到其他节点中的内存，利用 NUMA 技术可以很容易地使服务器扩展到 16 路、32 路，甚至 64 路，据悉目前理论上最多可扩展到 512 路。ccNUMA 兼顾可扩展性和可编程性，基本特征是具有多个 CPU 模块，每个 CPU 模块由多个(如 4 个)CPU 组成，并且具有独立的本地内存、I/O 接口等，如图 1-7 所示。访问内存不同的区域具有不一致的延迟。

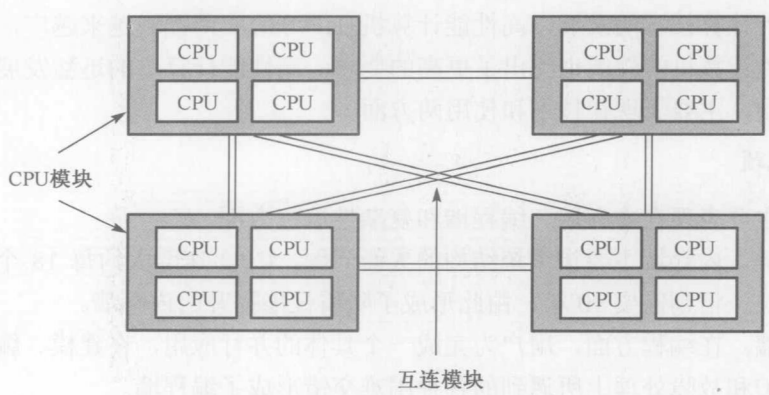


图 1-7　ccNUMA 结构图