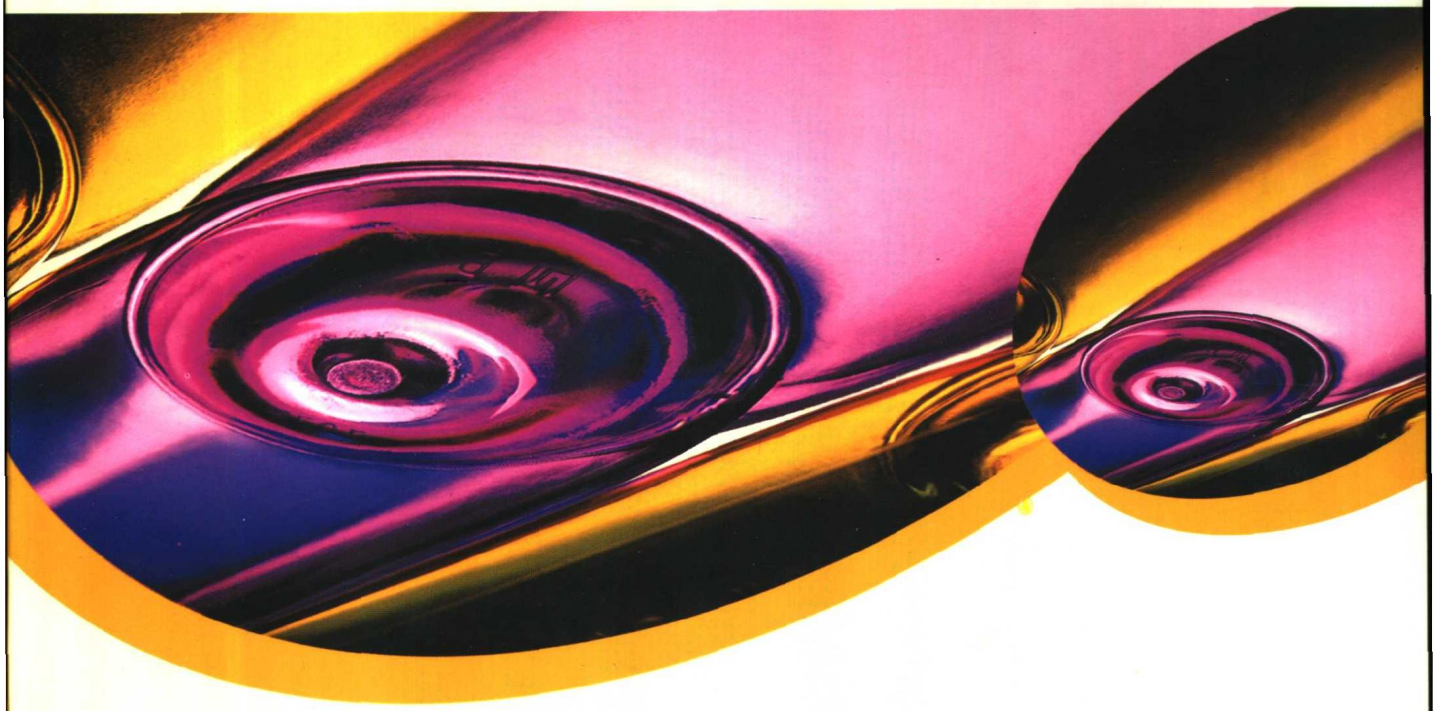
车静光 著

# 微机集群组建、优化和管理



机械工业出版社
CHINA MACHINE PRESS

献给复旦大学100周年校庆

# 微机集群组建、优化和管理

车静光 著

机械工业出版社

微机集群是把微机用网络连接起来，用 Linux 系统软件控制的并行计算机。本书详细讲解了微机集群所需的网络功能，并给出组建微机集群的完整过程和操作步骤，以及调试和测试技术。即使连 Linux 也没有学过的人，也可以通过本书学会组建微机集群所需的知识和技术；甚至只需拥有两台带网卡的微机和一根网线，就可以一步步地跟着本书进行微机集群实践。此外，本书还介绍了微机集群的管理、网络唤醒、网络启动、网卡捆绑以及任务排队等较深层的内容，供准备或已经组建大型微机集群的读者参考。

本书可供使用并行计算机从事科学和工程计算的科研、工程技术人员以及 Linux 发烧友和玩家自建和管理微机集群时参考，本书亦可作为大专院校、高职、高专及相关培训班的“Linux 网络应用”课程的教材。

**图书在版编目（CIP）数据**

微机集群组建、优化和管理/车静光著. —北京：机械工业出版社，2004.1
ISBN 7-111-13188-6

Ⅰ. 微… Ⅱ. 车… Ⅲ. 微型计算机—集群—基本知识 Ⅳ. TP36

中国版本图书馆 CIP 数据核字（2003）第 091771 号

机械工业出版社（北京市百万庄大街 22 号 邮政编码 100037）
策　　划：胡毓坚
责任编辑：孙　业
责任印制：路　琳

北京机工印刷厂印刷 · 新华书店北京发行所发行

2004 年 1 月第 1 版第 1 次印刷
787mm×1092mm　1/16 · 13.5 印张 · 329 千字
0001－3000 册
定价：24.00 元

凡购本图书，如有缺页、倒页、脱页，由本社发行部调换
本社购书热线电话（010）68993821、88379646
封面无防伪标均为盗版

# 前　言

我的微机集群实践始于 1998 年，为了调试并行计算程序，将两台微机设置成了微机集群。当时所用的网络带宽只有 10Mb/s，自己怎么也没有意识到，这种微机集群有朝一日竟会发展到足以与传统超级计算机竞争的地步；更没想到，100Mb/s 网络成为主流后，自己也会组建一台速度可以进入世界超级计算机 500 强的微机集群。2001 年 5 月，我在复旦大学计算凝聚态物理 985 计划 100 万元人民币的资助下，开始实践大型微机集群。经过一年多时间的试制，于 2002 年 5 月底建成了一台采用网卡捆绑、网络启动、OpenPBS 管理的有 96 个 P4 计算节点的大型微机集群。用世界超级计算机 500 强排序所采用的标准测试程序测试，这台微机集群的最大速度可达每秒 1417 亿次浮点运算，可排当时（2002 年 6 月）公布的世界超级计算机 500 强第 468 位。受此鼓舞，我就想在国内大力推广这种符合我国国情的超级计算机。将这些经验写出来，让更多的人分享。本书是我在给学生开课讲解自己动手组建超级计算机之后，在学生的要求下，和王迅院士的鼓励和推动下写成的。

书中尽可能手把手地教读者如何组建、优化和管理微机集群。从未接触过 Linux 的读者，可先阅读附录 A“Linux 系统基础”；第 2、3、4 章的内容足以帮助读者组建、调试和测试一台微机集群；虽然附录 B“微机集群的硬件选择”的内容，由于硬件市场变化很快而可能有点过时，但一些要领对组建一台可靠的微机集群还是有益的；第 5 章的内容是优化微机集群性能的一些有用方法；而第 6 章则介绍了一个常用管理工具的基本用法。

我从事物理研究，只是在研究工作需要、研究经费又捉襟见肘的情况下，才走上了自己组建微机集群的道路的，本书是这个过程的经验总结。由于教学、研究工作繁忙，成书时间仓促，书中错误在所难免，真诚欢迎计算机专家、学者和广大 DIY 发烧友批评指正。

附录 A 由洪峰收集整理，附录 B 由赵坚收集整理，在此谨表谢意。

在本书出版之际，我要感谢我的硕士导师谢希德院士和张开明教授，是她们引导我走上了表面物理研究道路；感谢我的博士导师德国明斯特大学的 J. Pollmann 教授，我的计算物理学和计算机编程方面的严格训练是在明斯特大学完成的；感谢王迅院士，他的鼓励和鞭策是本书写作启动的第一推动力；感谢清华大学李家明院士，他的关心和指导是我不断尝试微机集群新技术的源泉；感谢好友香港科大冯永嘉博士的无私帮助，否则我还需摸索更长的时间；感谢同事资剑教授的信任和支持，使我得以完成大型微机集群的实践；最后，还要感谢我的妻子顾青和女儿车逸文，她们的理解、支持和牺牲是如期完成本书的保证。

作　者

# 目　录





此为试读，需要完整PDF请访问：www.ertongbook.com

# 第1章　超级计算机的发展方向

微机集群最早的倡导者和实践者在2001年合作出版了一本题为《基于Linux的Beowulf集群计算》的书，书的开头就是一句如下的话："在经历了20多年错误和死路一条的高性能计算机架构的研究、探索之后，现在的道路已经很清楚：Beowulf集群[㊀]"。Beowulf集群是指用市场上可以买到的标准硬件（计算节点如微机或工作站、网络设备如交换机）组建的、主要用免费软件操控的并行计算机[㊁]。如果计算节点用微机，就是微机集群。

当然，作为一本由微机集群的倡导者和实践者出版的书，他们对传统超级计算机所下的断言难免倾向于Beowulf集群，也许失之偏颇，我们不必很在意他们的观点。然而，从这句断言中所透出的信息还是可以看出：超级计算机的发展应该具备Beowulf集群的主要特征：开放、通用、兼容。

该书认为，所有在传统高性能计算机上的努力都失败了。据估计，传统超级计算机硬件研制已经投入的费用仅在美国就可能超过六十亿美元，而应用软件开发上的投资至少加倍。乍听之下有点不可思议——动辄上百亿美元的传统超级计算机的研制、开发已经是死路一条？今后，像Cray这种类型的传统超级计算机，这种几年之前还高不可攀，即使在西方发达国家也只有少数几个国家级超级计算中心才能配置的、不可一世的庞然大物，除了因为已经花费巨资为这类超级计算机开发的应用代码无法移植、重要的数据库已经建立而不得不继续使用外，这类超级计算机的命运就此彻底结束？

## 1.1　市场是决定性的因素

任何技术的发展，离不开市场。或者说，市场是技术发展的主要推动力。某些技术，由于特殊的需要，会得到有关方面特殊的扶持。比如涉及国防军事、国家安全的关键技术，国家有关部门在其发展的起步阶段会给予大力扶持。传统超级计算机的出现，就是出于冷战需要、由美国政府予以大力扶持的结果。1991年以"美国总统倡议"形式开始的"高性能计算与通信"（HPCC）计划，对超级计算机的研制、开发起了很大的促进作用。但是，超级计算机的发展终究要由市场说了算。何况，现在超级计算机不仅在军事、国防、科技领域，同时也更多地渗透到其他的工业、商业、服务业等领域，能否持续发展终究要靠其自身的市场竞争力！

软件奇才比尔·盖茨在他的《未来之路》一书中说过：在一个发展市场上，两种相互竞争的产品，一种只要比另一种稍有优势，就会形成正反馈，从而形成事实标准，占据市场[㊂]。这在高科技市场上尤其如此。

请看一个例子：比苹果（Apple）电脑晚五年推向市场的IBM个人电脑虽然失去先机，

---

㊀ Thomas Sterling. Beowulf Cluster Computing with Linux.The MIT Press, 2001。

㊁ http://www.beowulf.org。

㊂ 比尔盖茨等. 未来之路. 辜正坤主译. 北京大学出版社。



此为试读，需要完整PDF请访问：www.ertongbook.com

但由于 IBM 开放技术标准，形成所谓的 IBM 兼容机事实标准，稳稳地占据着个人电脑几乎全部的市场份额。当年在技术上，特别是在图象技术方面领先，在市场上占得先机的苹果电脑，由于没有开放标准，颓势早现，现在几乎已经销声匿迹。这样的例子比比皆是。这说明，当产品技术本身相差不大时，兼容、通用、开放的标准是决定市场的关键因素。

由于用途的不同，作为个人用途发展起来的微机在技术方面的侧重有所不同。因此，当它被用作并行计算机的计算节点时，在某些方面的性能现在可能还不如传统超级计算机中的计算节点。但是，由于个人微机市场非常庞大，微机的发展速度远远快于传统超级计算机：各种新技术会被不断地开发、应用到微机上，这是传统超级计算机不能相比的。这样，微机在处理器速度方面的优势可以逐渐弥补微机在其他方面的不足。这一点可从下面两个例子来说明。

大家都知道，在 133MHz 主频的 Pentium 微机以前，由于处理器速度不够，在微机上看 VCD 还要用解压卡，否则影像、声音不流畅。这可归于微机某些功能的不足。没过多久，胜任解压任务的 166MHz 主频的 Pentium 芯片发布，解压卡很快就被逐出市场。

英国剑桥大学做分子动力学计算的一些同行曾经开发过快速傅立叶变换专用模块，以提高分子动力学计算中大量出现的快速傅立叶变换，其功能与解压卡类似，可以称为傅立叶变换卡，现在也早已销声匿迹。

这就是说，微机在某些能力上的不足，比如解压缩图象功能的不足，比如傅立叶变化计算能力的欠缺，是可以用处理器速度优势来弥补的，但前提是市场。一种技术只要有市场，就会形成正反馈，就会形成事实标准，其本身的某些不足可以借助于市场的力量得到改善。而任何一项技术，如果高高在上，产品失去市场，就不可能得到持续发展。即使在某些方面一时还可以占有一定的优势，这些优势随着时间的推移也会丧失殆尽，最终被逐出市场。市场的巨大需求量使得微机发展必定远远超过传统超级计算机的发展。因此，微机在处理器速度上的飞速提高很快就可以弥补其他方面的劣势。

构成微机集群另一个重要硬件的网络，则依托互联网技术的发展，同样是一个快速增长的领域。从 1982 年出台 10M 以太网，到 1994 年 100M 的快速以太网的开发成功，用了整整 12 年时间。但是由于互联网需要，网络技术发展的速度明显加快。仅仅过了三年，千兆以太网就已经出现，现在万兆（10G）以太网的标准和产品也已面世。此外，已经有近十年发展历史的一些集群专用的系统域网（SAN）技术如 Myrinet 等现在也已日趋成熟，在带宽和延迟性能上与传统超级计算机网络技术的差距也日益缩小。

因此，尽管目前微机集群与传统超级计算机相比，还有这样那样的不足，我们仍然可以大胆预言，随着微机市场的不断扩大，随着互联网技术的不断发展，市场决定未来的超级计算机就是微机集群，因为微机集群符合兼容、通用、开放标准的特征。

## 1.2 微机集群的历史

自从世界上第一台电子计算机 ENIAC 在 Mauchly 和 Eckert 的主持下，于 1945 年在美国宾夕法尼亚大学的莫尔电子工程学院诞生起，人类对计算机计算速度的追求就一直没有停止过。但是，更快、更强的处理器终究会受到物理极限的限制。而另一种提高计算机速度的可能途径就是利用多个处理器的协同能力——这就是所谓的并行计算机系统。

并行计算的概念早在第一台电子计算机诞生后不久就由现代电子计算机之父 von

Neuman 提出。让众多计算机一起工作这样的计算机集群的实践可以追溯到 20 世纪 50 年代末、60 年代初。当时，美国空军用真空管电子数字计算机建造了被命名为 SAGE 的第一个计算机集群，作为一个早期预警系统，用以监视苏联的核攻击。

1980 年，DEC 公司将集群概念应用到 VAX 微处理器集成的系统。虽然这种 VAX 集群和当时的 Cray 传统超级计算机在性能上相差 100 倍，但性能价格比正好倒过来。很快，VAX 群系统成了大学和科研机构占统治地位的计算机系统，成为研究人员能自己拥有和管理的属于自己的计算机系统。

1991 年以美国总统倡议形式进行的为期五年的“高性能计算与通信”HPCC 计划，加速了大规模并行计算机的研制。该计划要求到 1996 年以前实现计算速度、存储容量和网络带宽都达到所谓的 3T（万亿）的目标。而在 1995 年开始实施的“加速战略计算创新”（ASCI）计划，更是首次将计算科学提高到战略的高度，强调这是关系到国家安全、经济发展和科技进步的关键环节，是事关国家命脉的大事。该计划要求计算机能力从 1994 年的每秒 10 亿次，提高到 1996 年的每秒万亿次、2000 年的每秒 10 万亿次、2004 年的每秒 100 万亿次浮点运算的目标。10 年期间要求提高 5 个量级，而且制造成本要基本相近。

可见，科学计算的硬件基础——超级计算机的研制受到了很大的关注。但是，靠国家专项资金扶持的传统超级计算机，却是高成本、低收益的。它的组成部分——专用的处理器、特制的高速网络、复杂的操作系统和昂贵的应用软件使得传统超级计算机成本高昂，令市场望而却步。这无疑也制约了传统超级计算机的发展。

随着高性能微机、工作站和互联网技术飞速发展以及它们的商品化、标准化，随着免费开放的 Linux 操作系统内核的成熟稳定和性能完善，集群式并行计算机系统应运而生。它以传统超级计算机无法匹敌的性价比，成为并行计算的理想工具，导致了超级计算机的“平民化”。从此，速度超过每秒千亿甚至十万亿次浮点运算，几年前在美、日等发达国家也只有少数几个国家级超级计算中心才能配置的超级计算机，开始进入普通的大学、研究所、甚至实验室。

这一切令人瞠目结舌的变化都开始于美国航天航空署（NASA）的 Goddard 航天中心的那个无奈之举、那个被称为 Beowulf 计划的项目。

1994 年，Goddard 航天中心需要一台能被用来计算、采集、控制、显示、解决地球和空间问题大量数据的计算机。要求其峰值速度达每秒 10 亿次，存储数据大于 10G。当时具备这样能力的计算机系统的价格大约为一百万美元。可是，NASA 只能提供 5 万美元。因此，Goddard 航天中心的科学家们，迫于经费短缺，不得不选择用市场上可以购买到的微机和网络硬件自己组建满足这样要求的并行计算机系统。这就是 Beowulf 项目的由来。

当然，当时 Beowulf 出现的软、硬件条件都已经基本具备。1994 年，主流微机的配置是 intel 80486 处理器，10Mb/s 以太网，10MB 数量级的内存，100MB 数量级的硬盘。此外，从 1991 年开始流传的 Linux 免费操作系统，到 1994 年已经趋于成熟，内核已经相当的稳定。而且，PVM[⊖]（Parallel Virtural Machine），一种以消息传递并行编程和可连接通信函数库为基础的并行计算平台已经开发成功，1993 年推出的 V3.0 版本已经相当稳定。根据 Beowulf 项目研究人员估计，他们惟一需要做的就是在 Linux 系统下开发这样的系统的网络软件技术。

---

⊖ 由 Oak Ridge National Laboratory 于 1989 年开始开发，1990 年发布 v1.0 版，1994 年推出 v3.3 版后基本停止继续开发。

1994 年第一台 Beowulf 式的微机集群 Wiglaf 在美国航天航空署的 Goddard 航天中心诞生。它由 16 个 66MHz 的 486 处理器的微机组成（很快被换成 100MHz 的 486 DX4 的处理器），采用了 10Mb/s 以太网集线器。它的速度已经达到每秒 10 亿次操作。只是当时 intel 486 处理器的浮点运算能力还很差。即使如此，Wiglaf 的速度也达到了 7200 万次浮点运算，能与当时 Intel 公司的 Paragon 和 Thinking 公司的 CM-5 等大规模并行系统超级计算机使用相同计算节点数时的速度一争高下。以后，这样的微机集群就被开发者之一的 Thomas Sterling 统称为 Beowulf。从此，并行计算开始了一个新纪元。

微机集群发展的原始动力是市场需求。Beowulf 这一产物之所以能在市场上立足是因为它的三大技术基础——标准的、商品化的、廉价的高性能微处理器和高速网络技术以及免费、开放的系统及并行软件。

- 由于多媒体应用的需要，带动了微机处理器浮点运算能力的快速发展。从 Pentium 处理器开始集成的所谓多媒体扩展指令集（MMX），大幅改进了微机的浮点计算性能。同时，出于与其他微处理器如 alpha 和 PowerPC 等芯片竞争的需要，从 Pentium Pro 开始增加缓存专用总线。从此，微机处理器在浮点运算能力上也开始逼近甚至超过工作站用微处理器。
- 1994 年出现的快速以太网带宽达到了 100Mb/s；仅过了三年，1997 年发布的千兆以太网，带宽达 1Gb/s；现在 10Gb/s 以太网标准也已出现并有产品上市。此外，1993 年开发的第一个标准、开放、通用的集群用系统域网络技术——Myrinet，到 2002 年性能已经达到了 2Gb/s 带宽和 11 微秒延迟的高水平。
- 微机集群主要采用消息传递模式并行计算。1990 年发布的 PVM（Parallel Virtural Machine）是第一个广泛使用的消息传递并行标准和函数库，在此基础上可以相对比较容易地并行编程。1994 年，Massive Passing Interface（MPI）标准问世，成了后来占统治地位的消息传递并行通信标准。

在此基础上，越来越多的研究人员开始自己动手组建 Beowulf 集群。各项软、硬件应用技术也随之开发。此外，众多的传统超级计算机制造商如 IBM、Sun、HP 和 SGI 等也开始介入这一市场。从此，Beowulf 集群就有了更加迅速的发展。

- 1996 年采用 16 个 Pentiun Pro 处理器和 100Mb/s 以太网交换机的 Hyglac 和 Loki 分别在加州理工学院和美国能源部的 Los Alamos 国家实验室建成，这是微机集群发展历史的另一个重要的里程碑。它耗资 5 万多美元，速度首次超过了每秒 20 亿次浮点运算。1996 年出版的美国自然科学的权威刊物《科学》对这一成就以《自己动手装超级计算机》为题作了详细的报道[⊖]。Hyglac 后来在 1997 年赢得了 Gordon Bell 性能价格比大奖。
- 1997 年，用 100~200 个 Pentium Pro 处理器的微机集群分别在加州理工学院、Los Alamos 国家实验室和美国航天航空署等处建成，性能首次突破每秒 100 亿次浮点运算。
- 1998 年采用 UltraSparc 芯片的工作站和 Myrinet 交换机建成的工作站集群——NOW-2，是于 1993 年在加州大学伯克力分校启动的工作站集群项目（NOW）的延续。NOW-2 的速度达每秒 480 亿次浮点运算，是第一个挤身当时的世界超级计算机 500 强（第

⊖ Gary Taubes. Do it yourself surpercomputer, Sience **274**, 1840 (1996)。

113 位）的集群式计算机。

- 1999 年建成的集群式计算机 Linpacj，速度首次达到每秒万亿次浮点运算。
- 现在，规模超过 1000 台微机的集群也已出现，速度最高的微机集群已达每秒 11 万亿次浮点运算[㊀]。

Beowulf 所揭示的经济和社会现象，已经消除了高性能计算机架构之争，对超级计算机的研制和开发产生了深刻的影响。Beowulf 计划负责人之一的 Thomas Sterling 在 2001 年 7 月和 8 月出版的《科学美国人》上接连发表两篇专题文章——《如何建造超级计算机》和《自己建造的超级计算机》，对 Beowulf 现象作了详细的介绍和深刻的剖析[㊁]。

## 1.3 微机集群的现状

图 1-1 是从 1993 年起至 2002 年底、用 Linpack 速度测试排序的世界超级计算机 500 强内各种不同架构的超级计算机的数量按年份分布的情况。图中没有标注出的类型的是单指令多数据（SIMD）处理机，数量很少，并且已在 1998 年后完全退出世界超级计算机 500 强。

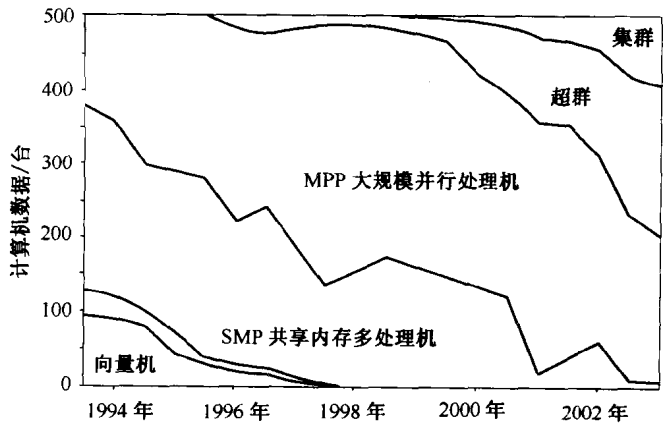


图 1-1

用线性代数程序包 Linpack[㊂]测试工具软件 xhpl 对超级计算机的最大浮点运算速度进行测试，进而排出世界超级计算机 500 强顺序并在网上公布的活动开始于 1993 年。Linpack 测试虽是线性代数计算，给出的是每秒能达到的最大浮点运算次数，但并非只对线性代数计算有效。并行计算实际上包含数值计算和数据交换通信两部分。任何形式的并行计算不管是不是线性代数计算，如果数值计算和数据交换通信量与 Linpack 测试程序的相当，那么 Linpack 速度指标就可以真实地反映该计算机的速度。如果数值计算和数据交换通信量比 Linpack 测试程序低，Linpack 速度指标相同时，但通信带宽、延迟好的计算机，性能有冗余；带宽、延迟差的计算机则还可有更佳的表现。

由于没有更合理的标准，Linpack 测试仍是目前可以被普遍接受的超级计算机的排序标准。世界超级计算机 500 排序是要自己登记，然后经该网站核实认可后公布的。从 1993 年开始到目前为止，500 强中数量最多的仍然是传统超级计算机。可见，这些制造传统超级计算机的大公司也是认可和接受这个排序标准的，是有一定权威性的。

从图 1-1 中可以看出，从 1998 年第一台集群式超级计算机进入世界超级计算机 500 强起，

---

㊀ http://www.top500.org。

㊁ Thomas Sterling. How to Build a Hypercomputer. Scientific American, July, 2001; William W. Hargrove, Forrest M. Hoffman and Thomas Sterling。 "The Do-It-Yourself Supercomputes", Scientific American, August, 2001。

㊂ http://www.netlib.org。

进入 500 强的集群式超级计算机的数量逐年增加。实际上，世界各国的大学、研究机构还有数量众多的自建微机集群，由于种种原因没有到世界超级计算机 500 强网站登记排序，因而不在该网站公布的 500 强清单上。在 2002 年 11 月公布的 500 强清单中，已有 93 台属于集群式超级计算机，其中 32 台是采用 Intel 处理器、8 台是采用 AMD 处理器的微机集群。而超群[⊖]和大规模并行处理机 MPP 分别占 206 和 195 台。单一共享内存多处理机 SMP 在 500 强中占有数量下降的趋势不可避免，现在只占 6 台，最终退出 500 强只是时间问题。向量机和 SIMD 单指令多数据处理机已分别在 1997 和 1998 年先后退出 500 强。从发展趋势看，这几年有较大发展的是集群和超群。

## 1.4 微机集群的技术

微机集群技术分硬件和软件两部分：市场上可以买到的标准硬件（计算节点、网络设备）和免费软件（Linux 操作系统、各种并行、管理平台等）。

计算节点的性能主要由处理器决定。

最早的微机集群使用的处理器是 Intel 486，浮点运算能力非常有限，因此那时微机集群的性能瓶颈在处理器速度。由于市场的作用，现在 32 位的微机处理器不仅在主频上早就超过了工作站微处理器，而且浮点运算能力上也开始逼近甚至超过 64 位的工作站微处理器。目前速度最快的 32 位微机处理器 Pentium 4 的主频已经超过 3GHz，64 位架构的 Itanium3 处理器也已经发布上市。

此外，从 Pentium Pro 处理器开始，Intel 在处理器中增加了支持多处理器设计的线路，PentiumII 和 PentiumIII 也都延续了这一做法。因此，出现了 Pentium Pro、Pentium II、Pentium III 和同级 Xeon（至强）的多处理器服务器，这可以说是一种廉价的对称多处理机 SMP。由于 Intel 的销售策略，Pentium 4 不再包含支持多处理器的线路，只有同级的 Xeon 处理器才支持。现在，最新（3.06GHz 主频以上）的 P4 也采用了超线程技术。因此，除了在处理器芯片中增加了用以支持多处理器线路之外，同主频、同缓存的 Pentium4 与 Xeon 是基本相同的[⊜]。可能由于销售策略，同样主频的 Xeon 总是比 Pentium 4 晚半年左右的时间才推出。

但是，由于总线结构，基于微机处理器的 SMP 服务器不能解决多处理器同时访问存储器的问题。即在一个时钟周期内，一个存储器单元只能由一个处理器进行读或写操作。这多少影响了基于微机处理器的 SMP 的性能。据报导，使用多处理器时并行计算的速度在某些情况下甚至还不如在同一服务器上仅使用单个处理器的速度。如果要改进这样的局面，其芯片设计的难度和成本都不小。目前尚未见到这方面的努力。所以，目前看来这种 SMP 还不适合用作微机集群的计算节点。用这种 SMP 作为微机集群计算节点的惟一考虑是昂贵的网络硬件开销。对于系统域网如 Myrinet 等，采用 SMP 在处理器数量相同的情况下可以降低网络硬件的成本，因此，还是有相当的微机集群采用这种 SMP 作计算节点。

微机集群实际上是大规模并行处理机 MPP 低成本的变种。现在 MPP 为了降低成本，也

---

⊖ constellation，一种共享内存多处理机 SMP 结合高速超维网络结构的集群。

⊜ 曾有 DIY 发烧友在网站发文论述在 Pentium II 级的赛扬处理器上，如何将 Intel 公司故意断开的支持多处理器线路重新连接以使这种赛扬处理器能被用于双处理器 Pentium II 的主板。

大量使用为工作站或微机开发的微处理器。1997 年建成的大规模并行处理机 Intel/Sandia 的 option red 采用的就是 9216 个 200MHz 主频的 Pentium Pro 处理器。而且，与 MPP 相同，微机集群也是通过消息传递实现计算节点之间的进程互相作用。这样看来，微机集群与大规模并行处理机 MPP 关键的差别在于计算节点间互联技术的优劣。在 MPP 中，节点的存储器总线用高带宽、低延迟的高速专用网络互联（当然，MPP 还有专用的软件以提高通信性能），而微机集群是节点的系统总线用网络互联。现在这样的界限也变得越来越模糊。比如按上述定义基本上属于工作站集群的 IBM 的 SP2 也被认为是 MPP。但实际上，除了用作通信网络的专用高性能交换机外，SP2 是集群结构，即系统总线的网络互联。

比较计算节点规模、性能类似的 MPP 后发现，在大多数情况下，微机集群计算方面的性能相等或优于 MPP，但微机集群的通信能力比 MPP 要低一个量级左右。

微机集群的性能主要取决于网络技术。

如上所述，与传统的超级计算机相比，微机集群主要性能瓶颈是网络带宽和延迟。由于目前已有将 MPP 或其他传统超级计算机的网络技术向集群式并行计算系统移植的倾向，因此，这方面技术的差距正在日益缩小，很难估计 MPP 的网络技术还有多长时间能继续领先集群的网络技术。

下面是微机内部各个主要部分的数据通信带宽：

| 部件 | 延迟 | 带宽 |
|---|---|---|
| L1 缓存[⊖] | 0.5~1.0ns（1~2GHz 时） | |
| L1 缓存 | 4~10ns | 4000~10000Mb/s |
| 内存 | 40~80ns | 1000~4000Mb/s |
| 硬盘 | 5~15μs | 800~1000Mb/s |
| 网络硬盘 | 5~20μs | 100~700Mb/s |
| 网络 | 5~20μs | 100~2000Mb/s |

微机集群要求网络必须具有双向高带宽、直接寻址等能力。现在这些要求已经成为一般网络交换机的标准了。

快速以太网是目前主流的局域互联网技术，带宽为 100Mb/s，如用 TCP/IP 协议，延迟约 90μs。就作者所知，交换机目前最多能被全速堆叠至 384 个端口。随着 internet 互联网技术的发展，快速以太网的地位很有可能在不久被千兆以太网所取代。快速以太网上的应用可以很容易地被移植到千兆以太网。与快速以太网相比，千兆以太网用铜线或光纤连接，虽然千兆以太网的带宽增加到了 1000Mb/s，但是如果仍然采用 TCP/IP 协议的话，延迟基本没有改变。

Myrinet 在 2002 年带宽已经达到 2.0Gb/s，短消息延迟小于 11μs，采用光纤或铜线互联。Myrinet 是从超级计算机网络技术演化而来的，最初开发的用途就是用来构造柜式系统域网互联的计算机集群。它沿用了一些传统超级计算机互联网的成熟技术。一般认为，Myrinet 的潜力在于其硬件标准是开放的，软件是免费的，以及它板卡上的可编程微处理器。目前 Myrinet 网络交换机产品的端口最多可达 128 个，特殊要求还可全速堆叠至 256 个端口。已经有作为微机 PCI 接口的 32 位或 PCI-X 接口的 64 位的网卡，在支持微机集群应用中应该

⊖ L1 缓存的容量通常很小，约 10KB 的量级，一个时钟周期传递所有 L1 数据，它的带宽由主频和 L1 缓存容量决定。



此为试读，需要完整PDF请访问：www.ertongbook.com