Beowulf-T 机群系统高可扩展性的研究*)

祝永志　李丙锋　魏榕晖
（曲阜师范大学计算机科学学院　日照 276826）

摘　要　并行计算技术是衡量一个国家科技水平的重要标志之一，PC 机群计算机是最廉价的高性能计算机。本文构建了一个 Beowulf-T 机群系统，提出了在该系统上的加速比和效率计算公式。通过实例的测试说明该机群系统具有高可扩展性。

关键词　MPICH，Beowulf，并行计算，并行算法，可扩展性

Research on the High Scalability of the Beowule-Tcluster System

ZHU Yong-Zhi　LI Bing-Feng　WEI Rong-Hui
（College of Computer Science，Qufu Normal University，Qufu 273165）

Abstract　Parallel computing technique is one of the important signs of measuring a country's science and technical level，PC cluster system is the cheapest High-Performance. This paper develops a Beowulf-T cluster system，proposed the computing formula in the speed and efficiency of this system. The test shows that the cluster of examples with high scalability.

Keywords　MPICH，Beowulf，Parallel computing，Parallel algorithm，Scalability

1　引言

由于在大型科学与工程计算问题中对运算数据的安全性、稳定性、高速性和准确性等方面的要求的不断提高，高性能并行计算越来越受到人们的重视和青睐。大型并行计算机系统大致分为单指令多数据流 SIMD、并行向量处理机 PVP、对称多处理机 SMP、大规模并行处理机 MPP、分布共享存储多处理机 DSM 和工作站机群 COW 六类机器[1]。前五类由于受系统结构的限制或研制费用及成本较高等诸多因素，其市场受到一定的限制。机群计算机由于具有投资风险小、可扩展性好、可继承现有软硬件资源和开发周期短、容易编程等突出特点，目前已成为并行计算的热点和主流，在中小企业和大专院校受到普遍重视。

本文以我校研究生教学实验室为平台构建了一个 Beowulf-T 机群系统。实验性地论证了该系统具有高可扩展性。

2　机群和 Beowulf 机群

机群是由一组独立的计算机系统构建的一个松耦合的多处理机系统，系统中各进程借助网络实现通信、共享内存传递信息，从而实现分布式并行计算。一组廉价的微机协同工作可以达到超级计算机的性能。

近年来，机群系统之所以发展如此迅速，主要是因为是机群结点为工作站的系统的处理性能越来越强，更快的处理器和更高效的多 CPU 机器已大量进入市场；局域网新技术和新协议的引入，机群结点间的通信能获得更高的带宽和较小的延迟；机群系统比传统的并行计算机更易于融合到已有的网络系统中去；机群上的开发工具日臻成熟，而传统的并行计算机缺乏统一的标准；机群价格便宜并且易于构建；机群的可扩放性良好，结点的性能也很容易通过增加内存或改善处理器性能获得提高。

随着微机性价比的提高和以太网等局域网技术的成熟和硬件成本的降低，以及消息传递标准和相应软件的发展，为用一组微机建立并行计算机群（常称为 Beowulf 系统）铺平了道路。Beowulf 系统具有高可用性、高扩展性、高性价比等优点，其关键技术是可用性支持、单一系统映像、作业管理、高效的通信[2]。

Bowulf 系统还有如下优点：首先，系统依赖于成熟的、容易获得的计算机技术和通信技术以及硬件设备。其次，构建系统所用的软件，如 MPICH 编程环境、测试工具软件等，可以从网上免费下载。最后，系统具有良好的可移植性，可伸缩性，对系统的构建、维护以及资源的充分利用十分有利。

3　Beowulf-T 系统的硬/软件构建

3.1　硬件环境构建

24 台联想微机的硬件配置为：CPU：Intel(R)(R)4 3.20GHz AT. AT Compatible；内存：DDR 512M；硬盘：7200R 40G。网络设备配置为：交换机：D-Link DES-1024R＋ 24 个交换端口；网卡：100Mbps 自适应网卡；超五类双绞网络线缆、接头：适量。

3.2　软件环境构建

系统和软件配置为：每台微机上安装 Windows2000 server 操作系统并采用 MPICH. NT 1. 2. 5. 2 作为并行计算的支撑环境，编程语言 VC6. 0。

3.3　Beowult-T 机群结构体系

*)项目基金：山东省教育厅高等学校实验技术改革项目（编号：2005-400）。**祝永志**　教授，硕士生导师，主要研究方向为网络与分布式计算。**李丙峰**　硕士生，主要研究方向为分布式计算。**魏榕晖**　硕士生，主要研究方向为分布式计算。

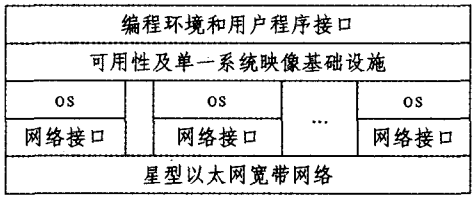

图 1　Beowulf-T 机群系统的体系结构

4　可扩展性与扩展性理论

可扩展性是设计并行算法和高性能计算机所追求的一个重要目标。如果能通过增加或减少系统资源而达到对计算性能、功能的提高或成本的降低，则称该计算机系统是可扩展的。

4.1　可扩展并行计算机体系结构

(1)无共享(Shared-nothing)系统结构：

所有节点之间通过 I/O 总线互联。多数机群采用此结构。

(2)共享磁盘体系结构：

磁盘模块全从节点机中移出。较小规模的商用性机群常采用此结构，其优点是当某个节点出现故障时，其他节点可接替其工作。

(3)共享存储器体系结构：

互联系统与每个节点中的存储总线相连，而在其他两种结构中，互联系统则是与节点的 I/O 总线相连[3]。

4.2　可扩展性范围

可扩展性范围一般包括资源可扩展性、应用可扩展性和技术可扩展性。

·资源可扩展性：

规模可伸缩性：增加处理器数。有时包括互联网络、接口以及通讯软件在内的子系统也可能要增加。

资源扩展：保持 CPU 数不变，通过增加多存储容量、更大的芯片外高速缓存以及大容量磁盘来扩展系统。

软件可扩展性：包括操作系统、友好的编程环境等。

·应用可扩展性：

要充分挖掘并行机的可扩展能力，应用程序也必须是可扩展的。即，同一程序在一个可扩展系统上运行时，其性能可随规模扩大而成比例加以改进，尤其在并行应用问题对机器和问题规模已有限制的情况下，对程序修改可使可扩展性提高。

·技术可扩展性：

可扩展系统的技术可扩展性是指该系统能适应技术的改变。它可进一步分为代可扩展性、空间可扩展性以及异构可扩展性[4]。

5　机群可扩展性测试与分析

加速比是并行计算的一个重要评测性能的概念。它指的是并行程序相比相同算法的串行程序所获得的性能提高的倍数。目前主要有三种加速比性能定律：适合于固定计算负载的 Amdahl 定律(Amdahl's Law)、适合于可扩放问题的 Gustafson 定律(Gustafson's Law)和受限于存储器的 Sun 和 Ni 定律(Sun and Ni's Law)。

定义可扩展机群系统 Beowulf-T 的加速比公式：

$$S(p)=\frac{T1+T_p}{T1+T_p/P} \tag{1}$$

其中 $T1$ 是单处理器运行时间，T_p 是 P 台处理器运行时间，P 是并行系统中处理器数。

衡量并行系统性能另一个技术指标——并行效率。P 个结点的并行效率为：

$$Ep=\frac{S(p)}{P} \tag{2}$$

理想状态下，加速比接近于 P，并行效率接近于 100%。以计算 Π 为例，N=间隔数。

实验步骤：①将 MPICH 安装到每台主机上；②在所有主机上建立一个相同的帐户，用户名 ZYZMPI，密码 3245(也可在每台机器上使用不同的用户名和帐户，然后建立一个配置文件，使用命令的方式运行程序)；③将 MPICH 安装到每台主机上；④运行"mpich\mpd\bin\MPIRegister. exe"，将在每台计算机上申请的帐号与密码注册到 MPICH 中，信息写入硬盘，重启动仍存在，这样 MPICH 才能在网络环境中访问每台主机；⑤在主控机上(IP：169. 254. 6. 41)，启动图形界面的程序 MPICH Configuration，以便主控机能获知各主机的信息；⑥将计算 Π 的程序，在 VC 下编译成可执行文件 MyPi. exe，放入各主机相同文件夹 Mytemp 下；⑦运行 MPIRun. exe 启动图形方式的 MPI 环境，运行程序。见图 2。

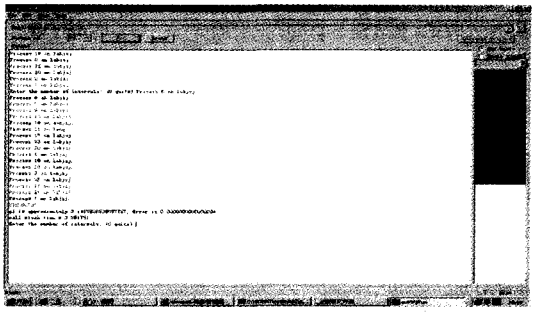

图 2　Beowulf-T 机群 24 节点计算结果示意图

表 1　固定问题规模、变换系统规模所得的时间与效率对照表

P	N=123154216		N=214341726		N=423567210		N=812392731	
	Tp(S)	E(%)	Tp(S)	E(%)	Tp(S)	E(%)	Tp(S)	E(%)
1	12. 064003		21. 106305		42. 086649		80. 605132	
2	6. 020065	59. 98	10. 502084	59. 96	20. 927106	59. 95	40. 079362	59. 95
4	2. 919539	29. 27	5. 255424	29. 39	10. 394251	33. 28	20. 119538	33. 36
8	1. 407608	13. 75	2. 583233	13. 81	5. 268984	17. 71	9. 996991	17. 61
16	0. 710061	6. 59	1. 293776	6. 61	2. 545488	8. 99	4. 988147	9. 08
24	0. 455611	4. 26	0. 850897	4. 32	1. 711825	6. 77	3. 281751	6. 72

表 1 实验数据表明：如果问题规模 W 保持不变，只增加系统的规模 P，则并行开销将会增加，使得系统效率降低。而

当系统规模 P 不变，只增加问题规模 W，由于并行效率增长得比 W 慢，从而使得效率增加。因此，可以让 W 和 P 同步增加，以保持效率不变。结论符合适合于可扩放问题的 Gustafson 定律，说明本文给出的计算加速比、效率新方法是简捷有效的。

结束语 本文构建的 Beowulf-T 机群系统是由若干独立的微机和可扩展的星型结构以太网组成，上述实验从资源可扩展性方面表明该系统具有高可用性和高可扩展性。另一方面，机群环境下通讯技术问题、并行程序设计环境问题、负载平衡问题以及全局资源管理与使用等，将是我们下一步研究的内容。

参考文献

1 陈国良. 并行计算-结构. 算法. 编程 [M]. 北京：高等教育出版社，2004
2 李继民，马力，王凤先. PC 机群系统的关键技术[J]. 河北大学学报(自然科学版)，2002，22(1)
3 熊盛武，王鲁，杨婕. 构建高性能集群计算机系统的关键技术[J]. 微计算机信息，2006，22(1-3)
4 黄淑玲. 可扩展并行计算的应用与研究 [J]. 电脑知识与技术，2005. 12
5 黄铠，徐志伟. 可扩展并行计算-技术、结构与编程[M]. 北京：机械工业出版社，2000

(上接第 255 页)

转到(6)；否则，用该客体的属主证书对变更后的客体进行签名，然后进入下一步。

(9)访问结束。

4 安全和性能性分析

(1)自身的安全性。用户信息、密钥或特征码正这些在认证过程中需要的重要信息存储在 TPM 中，防止了非法用户读取敏感信息。

(2)可信验证系统既可以保证主体在访问可信静态客体的时候需要验证客体的身份和完整性，又可以保证客体内容的真实性，解决了安全操作系统对静态客体的处理存在的缺陷。

(3)系统的工作效率。鉴于丰富的计算资源和较高的传输速率，TASSOBT 的主要开销在于内部的计算时间以及与 TPM 的通信时间。选用 SLE66 型号的嵌入式安全模块作为 TPM(通信速率为 38.4k/s)实现 TASSOBT 的计算量和时间花销(包含通信和计算时间)。由于鉴别和签名过程在用 SLE66 型号的嵌入式安全芯片内部，具有较高的安全性，因此可选用速度较快的 HSAH 算法实现鉴别和签名。在 SLE66 型号的嵌入式安全芯片中，鉴别和签名产生的随机数长度均为 160B，HASH 算法选用 SHA. 1。另外，在 TASSOBT 过程中，由于传输的数据量少，通信开销小，通信时间可以忽略；而此过程中，SLE66 型号需要进行 4 次 HASH 运算。对于 SLE66 型号的嵌入式安全模块，一次 SHA. 1 运算化的时间大概是 200ms，所以 TASSOBT 过程所需要的时间开销约为 0.8s。

总结 随着可信计算技术的兴起，可信操作系统逐渐成为人们关注的焦点。在可信操作系统中如何保证静态客体可信成为了一个紧迫的重要课题。随着网络技术的不断发展和 Internet 的日益普及，人们对 Internet 的依赖也越来越强。人们在 Internet 浩瀚的信息海洋中，获取无穷无尽的知识的同时，也会获得一些无用或虚假的信息。随着内容服务的发展，静态客体内容可信的需要越来越强烈。本文首先分析了操作系统中客体的类型，总结了安全操作系统中对静态客体的处理存在的问题，提出可信静态客体的概念并分析其特点。为了保证可信静态客体内容的真实性，提出了基于 TPM 的静态客体可信验证系统。该系统将生成可信静态客体的映像文件，映像文件记录某可信静态客体的来源、各次处理行为和内容变化的签名并存于 TPM 中。为保证静态客体的可信提供了一种解决方案。我们的下一步工作是将 TASSOBT 在 Linux 系统上实现。

参考文献

1 Jajodia S, Samarati P, Subrahmanian V, et al. A unified framework for anforcing multiple access control policies. In: S IGMOD '97, Tucson, AZ, May 1997. 474～485
2 Galiasso P, Bremen O, Hale J, et al. Policy Mediation for Multienterprise Environments. ACSAC, 2000. 100～106
3 Abrams M, LaPadula L, Eggers K, et al. A Generalized Framework for Access Control: an Informal Description. In: Proceedings of the 13th National Computer Security Conference, Gct. 1990. 134～143
4 Bertino E, Jajodia S, Samarati P. Supporting Multiple Access Control Policies in Database Systems. In: IEEE Symposium on Security and Privacy, Oakland, 1996
5 Osborn S, Sandhu R, Munawer Q. Configuring Role-based Access Control to Enforce Mandatory and Discretionary Access Control Policies. ACM Transactions on Information and System Security, 2000, 3(2): 85～105
6 Secure Computing Corporation. DTOS Lessons Learned Report: [Technical Report]. DTOS CDRL A008. Secure Computing Corporation, Secure Computing Corporation, 2675 Long Lake Road, Roseville, Minnesota, June 1997. 55113～2536
7 Bell D E, La Padula L J. Secure Computer Systems: A Mathematical Model. MTR 2547-II (AD 771 543). The MITRE Corporation, Bedford, Massachusetts, May 1973
8 Harrison M H, Ruzzo W L, Unman J D. Protection in operating systems. Communications of the ACM, 1976, 19(8): 461～471
9 Biba K J. Integrity considerations for secure computer systems: [Technical Report]. MTR 3153. The Mitre Corporation, April 1977
10 Denning D E. A lattice model of secure information flow. Commun ACM, 1976, 19(5): 236～242
11 Brewer D, Nash M. The Chinese Wall security policy. In: Proceedings of the 1989 IEEE Symposium on Security and Privacy, IEEE Computer Society Press, May 1989. 206～214
12 Boebert W E, Kain R Y. A Pratical Alternative to Hierarchical Integrity Policies. In: Proceedings of 8th National Computing Security Conference, Gaithersburg, October 1985
13 侯方勇，王志英. 可信计算研究[J]. 计算机应用研究，2004(12)：1～4
14 刘鹏，刘欣. 可信计算概论[J]. 信息安全与通信保密，2003(7)：17～19
15 周明辉，梅宏. 可信计算初探[J]. 计算机科学，2004，31(7)：5～8
16 屈延文. 软件行为学[M]. 北京：电子工业出版社，2005
17 林闯，彭雪梅. 可信网络研究[J]. 计算机学报，2005，28(25)：751～758
18 谭良，周明天. CRL 分段-过量发布新模型[J]. 电子学报，2005，33(2)：227～230
19 谭良，周明天. CRL 增量-过量发布新模型[J]. 计算机科学，2005，32(4)：133～136
20 SHAN Zhiyony. Research on the Framework for Multi-Policies and Practice in Secure Operating System [D]. Beijing, P R China: Institute of Software Chinese Academy of Sciences Beijing, 2002
21 屈延文. 软件行为学[M]. 北京：电子工业出版社，2005
22 胡建伟，汤建龙，杨绍全. 网络对抗原理[M]. 西安电子科技大学出版社，2004