具有分布式并行 I/O 接口的分布式并行服务器系统的性能研究

刘心松

(电子科技大学 8010 研究室,四川成都 610054)

摘 要: 分布式并行是多年来的一个研发热点,但其研发仅限于分布式并行系统的内部.服务器和外界(如客户机)的交互仅通过一条通道(如一台计算机)完成,这一瓶颈限制了系统可靠性/可用性,加速比与效率,系统频带和 I/O 响应速度等.本文提出的具有分布式并行 I/O 接口的分布式并行服务器使其中的各节点机均可直接与众多客户机交互作用,设计保证 80%以上的客户请求均一次直接发往能提供服务的服务器中的节点机,仅有少量请求需要在节点间进行动态负载再调度.这是一种面向宽带的宽带网络服务器.本文的重点是将这种服务器与单机、镜像/双机热备份、群集服务器进行性能比较.结果表明,具有这种 I/O 接口的分布式并行服务器性能最佳.这种服务器的第一个版本已经研发成功,投入使用,并将进入市场.

关键词: 分布式并行 I/O 接口; 分布式并行服务器; 宽带网络服务器

中图分类号: TP301; TP302.7 **文献标识码:** A **文章编号:** 0372-2112 (2002) 12-1808-03

The Performance Research of the Distributed Parallel Server System with Distributed Parallel I/O Interface

LIU Xin-song

(8010 *Division, University of Electronic Science & Technology of China, Chengdu, Sichuan* 610054, *China*)

Abstract: Distributed parallel is an important research and development field in many years, but the researches and developments are only limited to the inside of the distributed parallel system. The interactions between the server and its outside (such as clients) are performed by only one channel (for instance, a computer). The channel is a bottleneck, limiting the reliability/availability, speed up ratio, efficiency, frequency band and I/O response speed of the system. The nodes of the distributed parallel server with the distributed parallel I/O interface presented in this paper can directly interact on the clients. The client demands over 80 percent are directly sent to the nodes that they can provide required services, while only few demands need dynamical schedule again. The is a broad band network server. The emphasis of the paper is to compare the performances among the distributed paraallel server and single computer server, mirro-image server, two-redundancy computer server and cluster server. The results show that this new server is the best one among these servers. The first version of this server is already researched and developed successfully, it is used, and it will go to market.

Key words: distributed parallel I/O interface; distriibuted parallel server; broad band network server

1 引言

随着 Internet 2 等宽带技术、产品和应用的迅猛发展,例如信息社会中信息系统的一个薄弱环节将突出的显现出来,这就是服务器.

我们可以把一个信息系统简单地描绘成如图 1 所示.图中,交换机的总体交换速度已达 Tb/s,采用 DWDM 的单膜光纤的传输速率已达 Tb/s,属于宽带范畴的 Internet 2 将比现有 Internet 的速度高出 100 倍以上,现有高档个人计算机的处理速度和存储容量已足以满足单个用户的需求.然而,现有市场上的单机服务器、双机热备份或镜像服务器却成为了制约其发展的瓶颈.主要原因在于其有限的可靠性、处理能力和存储容量等.

采用分布式服务器可使负载均衡,但例如偶遇大型计算任务则无能为力,为此引入并行处理机制,将分布式动态任务调度机制和并行处理机制融于一体形成分布式并行服务器,则可大幅度提高系统可用性(Availability)、吞吐力、I/O 响应速度和存储容量等[1-2].这种服务器有着广泛的应用前景,如 WWW、视频和教育服务等[3~7].

国内外已对分布式并行服务器进行了大量的研究[8~12],但他们只局限于服务器内部的研究,而没有很好地解决用户和服务器之间的接口问题.请求和结果均从其中的一个计算机进出,这个计算机就成为了其系统的瓶颈,影响了系统的可靠性(或可用性)、吞吐力、*I/O* 响应速度等.

在分布式并行服务器中,客户端和服务器端有相应分布式并行 *I/O* 接口,如图 1-2 所示,其功能是使客户端请求按当时服务器中可用节点机情况基本均衡而又随机地发往服务器中的各相应节点机,这一过程对用户是透明的[13].在完成每一用户请求的这一步之后,若服务器中各节点机之间的负

收稿日期:2002-01-23;修回日期:2002-06-14
基金项目:四川省科学研究项目

载之差超过一定限值,则在服务器内进行智能化动态负载再调度.

基于这种调度技术的分布式并行服务器消除了任何瓶颈,从根本上提高了系统的可用性,吞吐力和 I/O 响应速度均因此而提高.因

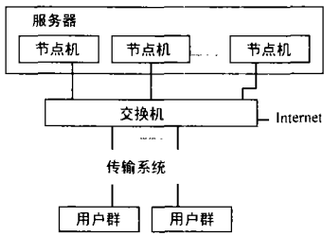

图 1　一个信息系统的粗略构架

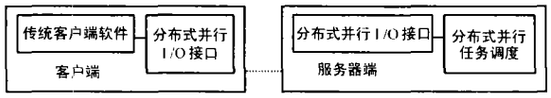

图 2　分布式并行服务器任务调度示意图

为服务器中各节点机均接入交换机,而交换机的总交换速度已达 Tb/s,而且还在不断提高,所以基于分布式并行 I/O 接口和有效的调度算法,服务器的处理能力和存储容量可根据需求扩缩节点机数而得到增减.

研究了这种分布式并行服务器的可用性、加速比、效率、系统使用频带和 I/O 响应速度,并与单机服务器、双机热备份/镜像服务器和群集系统进行比较.结果表明,这种服务器具有明显的性能优势.目前已投入实际应用,极受用户欢迎.

假设服务请求间彼此独立,请求的到达和服务服从指数分布,服务器中各节点之结构相同.下面各节将分别阐述.

2　可靠性和可用性

定义 1　可靠性[14]　在时间 0 至 t 期间服务器一直正常工作的概率即称之为服务器在时间 0 至 t 期间的可靠性.设单一节点机在时间 0 至 t 期间能正常工作的概率为 P_n,服务器中共有节点机数为 N,假设服务器中有一个以上(含一个)的节点机能正常工作,即认为服务器工作正常.则对镜像服务器或双机热备份服务器的可靠性 R_d 为

$$R_d = P_n(2 - P_n) \tag{1}$$

分布式并行服务器的可用性 R_{dp} 为

$$R_{dp} = \sum_{i=1}^{N} \frac{N!}{i!\ (N-i)!} P_n^i (1 - P_n)^{N-i} \tag{2}$$

群集系统因存在瓶颈节点,故其可靠性与单一节点机服务器无异,其可靠性为 P_n.

这样,分布式并行服务器可靠性与单机服务器/群集系统服务器和镜像服务器/双机热备份服务器可靠性之比分别为

$$R_1 = \frac{R_{dp}}{P_n} = \sum_{n=1}^{N-2} \frac{N!}{i!\ (N-i)!} P_n^{i-1} (1 - P_n)^{N-i} \tag{3}$$

$$R_2 = \frac{R_{dp}}{R_d} = \sum_{i=1}^{N-2} \frac{N!\ P_n^{i-1} (1 - P_n)^{N-i}}{i!\ (N-1)!\ (2 - P_n)} \tag{4}$$

当 $P_n = 0.9$, $N = 5$,则有 $R_d = 0.99$, $R_{dp} = 0.99999$, $R_1 = 1.1111$, $R_2 = 1.0100909$

可见,这种分布式并行服务器的可靠性相对于单机/群集系统服务器和镜像/双机热备份服务器都有极大的提高.

当节点机失效时,经修复后可重新回入系统正常工作时,则引入可用性概念.

定义 2　可用性[14]　服务器在时间 t 能正常工作的概率即称之为服务器在时间 t 的可用性 A.通常, $A \geqslant R$.

由于这种分布式并行服务器解决了传统分布式服务器/并行服务器/分布式并行服务器与客户机之间的界面瓶颈,使其可靠性和可用性获得了极大的提高.

结论 1　具有分布式并行 I/O 接口的分布式并行服务器的可靠性和可用性大于传统带有瓶颈性质的群集系统的可靠性和可用性.

以往其他的分布式并行服务器是用一台或两台计算机作为与外界的接口(I/O);但在该系统中,每个服务器节点都为外界提供 I/O 接口,从根本上解决了与外界间的瓶颈问题,此时系统的 R 即为 R_{dp}.当 $P_n = 0.9$, $N = 5$ 时, $R_d = 0.99$,但 R_{dp} 已为 0.99999.若 N 为数十甚至数百(设计值为 255,且 G 或 T 级线速交换机已为此提供了可能)时, R_{dp} 则无限逼近于 1;且因为 $A \geqslant R$,所以在引入可修复的概念之后,设计开发出一个永不崩溃的服务器系统完全成为了可能.这就是本系统的关键意义之所在.

3　加速比和效率

3.1　加速比

定义 3　加速比　对一给定任务或任务组,单节点机完成时间与多节点机系统完成时间之比.

设系统中单一节点机接收一请求、调度、处理相应任务和结果返回的时间分别为 t_r, t_s, t_p 和 t_d,这些时间长度间的比例为 $t_s : t_r + t_d : t_p = l : n : m$,而且 $l + n + m = 1$.

为简化讨论,我们假设有 N 个任务几乎同时到来则

单机服务器、镜像或双机热备份服务器(忽略镜像和转接时间)的完成时间为

$$t_{si} = N(n + m) \tag{5}$$

群集系统的完成时间为

$$t_c = N(1 + n) + m \tag{6}$$

若忽略再调度时间,则新型分布式并行服务器系统的完成时间为 $t_0 = 1$

群集系统的加速比为 $S_c = (N(n + m)/(N(l + n) + m))$, $S_c > 1$,新型分布式并行服务器的加速比为 $S_d = N(n + m)$, $1 < S_d < N$,并且 $S_d > S_c$.

结论 2　这种新型分布式并行服务器的加速比大于传统群集系统的加速比.

随着 G 或 T 级交换机的使用,网络带宽的不断增加,并不断研发设计新的系统结构,服务器系统中的节点机数可以根据需求不断增加,系统对一给定任务或任务组的完成时间可以足够短,以满足用户要求.需要指出的是,这种系统中的节点机可以是 PC 服务器,也可以是高档机型.这种系统完全有可能取代大型机,而且性价比更高.

3.2　效率

定义 4　效率　多节点机系统加速比与节点数之比.

根据 3.1 节的结果和定义 4,群集系统的效率为 $E_c = \dfrac{n + m}{N(l + n) + m}$, $E_c < 1$;

新型分布式并行服务器的效率为 $E_d = n + m$, $E_d > E_c$.

结论 3 这种新型分布式并行服务器的效率大于传统群集系统的效率,并且其效率不随节点机数的增加而降低.这又是以往其它服务器系统所无法比拟的效率优势.

4 系统频带

定义 5 周期 从一个客户服务请求的产生开始直至完成该请求之任务所经历的时间.

定义 4.2 系统频带 每个周期系统接收请求的平均数.

设周期间彼此独立,一节点机接收服务请求的概率为 P_r,则对分布式并行服务器而言,在一个周期内有 i 个请求的概率为

$$q_n(i)=\binom{N}{i}P_r^i(1-P_r)^{N-i} \tag{7}$$

在一个周期中有 i 个请求的期望值为

$$E_n(i)=\left(\frac{A_N^i-A_{N-1}^i}{A_N^i}\right)N \tag{8}$$

若交换机的交换速度和节点机的处理速度足够高,期望的频带则为

$$B_n(N)=\sum_{0\leqslant i\leqslant N}E_n(i)q_n(i) \tag{9}$$

$B_n(N)$即为系统的请求到达率.

顺便指出,节点利用率为

$$U_n=\sum_{0\leqslant i\leqslant N}q_n(i)\left(\frac{A_N^i-A_{N-1}^i}{A_N^i}\right) \tag{10}$$

单机、镜像和双机热备份的服务器的系统频带为 P_r;对传统带有瓶颈性质的群集系统,若在一个周期中的请求数 $\geqslant m/(l+n)+1$,则系统频带不再随节点数的增加而增加.

结论 4 若在一个周期中的请求数 $\geqslant m/(l+n)+1$ 且 $N\geqslant m/(l+n)+1$,则这种新型分布式并行服务器的系统频带大于传统带瓶颈性质的群集系统的频带.而且可以通过增加服务器节点机的数目来满足应用的高服务请求率要求,避免至少缓解爆发性(burst)请求到来时被阻塞甚至被拒绝服务的情况.

5 请求平均响应时间

定义 7 请求平均响应时间 来自客户端的服务请求,在服务器端的平均排队等待时间与服务时间之和.

设一节点对服务请求的服务率为 μ,则该服务器对一请求的平均响应时间为

$$t_m=c(\rho\cdot\mu)/(\mu N-B)+1/\mu \tag{11}$$

其中,$C(\rho,\mu)=\dfrac{(B/\mu)^N}{\left(\frac{B}{\mu}\right)^N+N!\left(1-\frac{B}{N_\mu}\right)\sum_{n=2}^{N-1}\frac{(B/\mu)^n}{n!}}$ (12)

可以指出,一个请求被接收的概率为

$$P=(B/P_rN) \tag{13}$$

结论 5 若在一个周期中的请求数 $\geqslant m/(l+n)+1$ 且 $N\geqslant m/(l+n)+1$,则这种新型分布式并行服务器对一请求的平均响应时间小于传统的带瓶颈性质的群集系统对一请求的平均响应时间.所以,从系统结构角度讲,这无疑也是一种很好的实时系统结构.

6 结语

设计保证客户端随机地(不是固定地)而又基本均衡地将服务请求发往服务器中的各节点机,使服务器中各节点机接收到的绝大部分请求都不需要再次进行负载平衡调度.这是这种新型分布式并行服务器的核心技术.本文是其系列文章之一,重点在于和其他类型服务器的性能比较,表明理论上的正确性.实际使用的成功使我们坚信了其方向.实践告诉我们,要深入研究和开发的方面还很多.

参考文献:

[1] Serpanos D N, Bouloutas A. IEEE Transactions on Circuits and Systems for Video Technology, 2000, 10(8): 1438 - 1449.

[2] Rumsewicz M, Dwyer M. Proceedings of First IEEE/ACM International Symposium on Cluster Computing and the Grid [C]. Brisbane, Otd, Australia: 2001. 363 - 370.

[3] Mourad A, Huiqun Liu. Scalable Web server architectures [A]. Proceedings of Second IEEE Symposium on Computers and Communications [C]. Alexandria, Egypt: 1997. 12 - 16.

[4] Colajanni M, Yu P S, Dias D M. Scheduling algorithms for distributed Web servers [A]. Proceedings of the 17th International Conference on Distributed Computing Systems [C]. Baltimore, MD: 1997. 169 - 176.

[5] Chan S.-H. G. IEEE Communications Letters, 2001, 5(9): 384 - 386.

[6] Chan S H G, Tobagi F. IEEE/ACM Transactions on Networking, 2001, 9(2): 125 - 136.

[7] Van Reeth F, Raymaekers C, Trekels P, Verkoyen S, Flerackers E. Proceedings of MMM on Multimedia Modeling, 1998: 47 - 48.

[8] Object Management Group. The common object request broker: Architecture and specification [DB/OL]. http://www.omg.org/corba

[9] Felber P, Guerraoui R. IEEE Concurrency, 2000, 8(1): 48 - 58.

[10] Putter P, Roos J D. Proceedings of the IEEE First International Workshop on Systems Management [C]. Los Angeles, USA: 1993. 118 - 124.

[11] Wolff T, Lohr K-P. Proceedings of the IFIP/IEEE International Conference on Distributed Platforms [C]. Dresden, Germany: 1996. 399 - 412.

[12] A Barak, O La'adan. Journal of Future Generation Computer Systems, 1998, 13(4): 361 - 372.

[13] 刘丹,刘心松,刘流.分布式并行服务器中的一种调度方法[J].高技术通讯,2002,8(增刊).

[14] Martin L Shooman. Software Engineering: Design, Reliability, and Management [M]. MCGraw-Hill Book Company, New York: 1983. 587 - 588.

作者简介:

刘心松 男,1940 年 12 月生,重庆市人,教授,博士生导师.已在学术刊物正式发表学术论文 110 余篇.主要研究方向为:分布式并行处理,宽带网络与通信,操作系统与网络软件,数据库系统等.