·计算机工程与应用·

分布式并行BOD系统资源管理算法研究

刘 丹，刘心松，杨曙锋

(电子科技大学计算机科学与工程学院 成都 610054)

【摘要】研究分布式并行宽带点播服务系统中资源的优化存储，以充分发挥系统性能；提出了最大化访问成功率的资源放置算法，以均衡分布式并行服务器系统负荷；提出了动态阈值预测机制的副本自适应放置算法，以提高系统吞吐量。通过实用系统，验证了它们在宽带点播服务系统中提高访问成功率及系统性能方面优于其他算法的良好特性。

关 键 词 访问频度； BOD系统； 预测机制； 放置算法

中图分类号 TP393.02 **文献标识码** A

Resource Management Algorithm Research of Distributed and Parallel BOD Systems

LIU Dan，LIU Xin-song，YANG Shu-feng

(School of Computer Science and Engineering, UEST of China Chengdu 610054)

Abstract The resource data storage in distributed Broad-band-service on Demand (BOD) systems is studied quantitatively . Firstly, a suitability resource placement algorithm is presented to balance the system load and to improve the system throughput. Secondly, an algorithm of replication placement policy is proposed depending on prognosticate mechanism. Applications in BOD systems approve that the algorithm can make the better access succession rate and system performance than using some other algorithms.

Key words access frequency; BOD system; prognosticate mechanism; placement algorithm

近年来，以基于微机的分布并行机群作为宽带点播服务(Broad-band-service On Demand，BOD)系统成为一种趋势[1]。在网络传输带宽及磁盘I/O带宽确定时，优化资源管理能进一步提高这类系统的服务性能。

对分布并行服务系统资源存储的研究工作已有很多[1-7]，一般是以最优负荷平衡原则来选择一个资源并为其创建副本，很少考虑创建副本的时延及带宽消耗对系统性能的实时影响。但在实际系统中应综合考虑多媒体文件尺寸大、创建时延长，以及创建时对系统传输带宽消耗大这些因素的影响。本文以层次模型为基础研究多媒体资源管理，主要针对资源在线决策及资源副本管理等问题进行研究，并给出优化管理算法，提出采用合理预测机制，通过动态阈值来选择适当资源创建副本的思想。对于资源的合理存储层次，文中给出了一个基于资源访问频度，最大化系统访问成功率的资源在线决策算法。

1 分布式并行BOD系统存储模型及算法逻辑

分布并行BOD系统主要包括集中式结构、分层式结构[1]、分布式结构[2]、基于代理的分布式结构[3]。为简化问题，本文以两层模型来描述，如图1所示。图中分布式并行宽带服务器组由高速交换机连接，由于各服务器磁盘的存储空间有限，其中存放访问率相对高的资源作为在线资源，而其余资源则作为非在线资源存放在二级存储器中。

定义用户访问成功为能得到系统的满足服务质量(Quality of Service，QoS)要求的及时响应。从资源放置角度来看，影响用户访问成功率的因素有两个：(1)在线资源有限，当用户请求非在线资源时，由于资源装载延时用户无法接受，此情况定义为访问失败；(2)对于在线资源，请求是否成功则要看系统当时的负荷情

收稿日期：2004－06－24

作者简介：刘 丹(1969－)，男，博士，讲师，主要从事分布式技术及网络技术方面的研究.

况，若有空闲带宽能够服务则访问成功，否则访问失败。由于各服务器的磁盘空间B_{disk}、磁盘I/O带宽$B_{\text{I/O}}$以及网络带宽B_{network}的限制，服务器服务能力是限定的。定义单服务器的同时服务数为S，每路服务需要的带宽为B_u，有：

$$S=\min(B_{\text{I/O}}, B_{\text{network}})/B_{\text{u}} \tag{1}$$

设整个分布式并行宽带服务器组的服务器数为m，则整个系统的最大服务能力为：

$$S_{\max}=\sum_{i-1}^{m} S_i \tag{2}$$

式中 S_i代表服务器i的可服务流数。

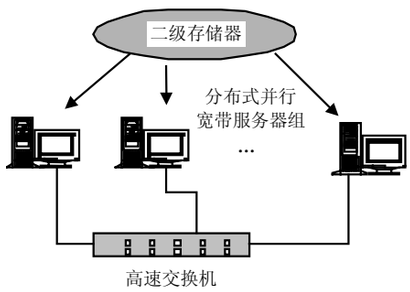

图1 层次存储模型

要提高系统的服务能力，一方面可以扩大系统规模，即增加m；另一方面应使系统内各服务器负荷均衡。对于访问率很高的资源，其宿主服务器因负荷可能很重而不能满足用户要求，但系统中的其他服务器此时可能尚未达到最大服务能力，应通过对访问率高的资源建立副本来实现负荷均衡。

因此，在系统规模一定的情况下，为提高访问成功率，资源的优化存储分为两个方面，一是选择哪些资源作为在线资源；二是被选中的资源在各个服务器中如何存放，以及是否建立副本、建立多少副本以达到系统的负荷均衡，发挥系统的最大效率。

1.1 决策资源在线策略

研究表明[4-5]，用户对资源的访问具有很强的局部性，客户的访问集中在一小部分热门资源上。当资源库的容量增大时，局部性表现得更加明显。

设系统要求访问成功率为$\eta(0\leqslant\eta\leqslant1)$，资源库中有$x$种资源，资源$R_i$的访问概率为$P_{R_i}(1\leqslant i\leqslant x)$，$P_{R_1}+P_{R_2}+\cdots+P_{R_x}=1$，将各资源按访问概率降序标号，即$P_{R_1}\geqslant P_{R_2}\geqslant\cdots\geqslant P_{R_x}$，必须保证资源的在线率$P_{\text{online}}$满足：

$$P_{\text{online}}=(P_{R_1}+P_{R_2}+\cdots+P_{R_n})\geqslant\eta \tag{3}$$

式中 资源$R_i(1\leqslant i\leqslant n,\ n\leqslant x)$为在线资源。定义统计周期为$T$，$T$可以由用户根据实际系统需要自行定义；定义访问频度为资源在本周期内被访问的次数。按以下式(4)计算第k个周期的$P_{R_i}(T_k)$，即：

$$P_{R_i}(T_k)=f_{R_i}(T_k)/C(T_k) \tag{4}$$

式中 $C(T_k)$为第k个周期统计的资源总访问频度，$f_{R_i}(T_k)$为第k个周期内资源R_i的访问频度。

根据式(3)可知哪些资源应作为在线资源存放。若η=1，则所有资源都要在线存放。当某资源需要在线存放时，选择哪台服务器存放资源是重要的。统计各服务器在线资源的访问频度，设服务器i的在线资源访问频度为X_i，有$X_i(T_k)=\sum_{R_j\in i} f_{R_j}(T_k)$，式中$R_j$为在服务器$i$上的资源。为均衡负荷，选择$X_i$最小的服务器存放资源。

1.2 决策在线资源增删副本的策略

1.2.1 增加副本

定义资源R_i当前周期k的访问频度阀值为$\delta_{R_i}(T_k)$，当$f_{R_i}(T_k)>\delta_{R_i}(T_k)$时，即使系统有负荷能力，访问$R_i$也会失败。这时应为资源$R_i$建立副本以均衡负荷，防止访问失败的情况发生。设集合x是拥有资源R_i的服务器节点集合，则$\delta_{R_i}(T_k)$取值为$\delta_{R_i}(T_k)=\sum_{j\in x}(S_j-L_j(T_k))$，式中$S_i$为机器$j$的可服务流数目，$L_j(T_k)$为机器$j$第$k$周期的已服务流数目。

取$f_{R_i}(T_k)>\delta_{R_i}(T_k)$作为建立副本的标准，当条件成立时，则$k$周期拥有资源$R_i$的节点集合是满负荷，再消耗其带宽建立副本会引起该节点集合用户服务质量下降或建立副本不成功。本文提出预测机制的资源访问频度计算方法，并以此建立副本。

资源访问规律的时间函数用正态分布来描述，采用正态分布预测机制，以$F_{R_i}(t)$表示资源R_i的访问规律函数，定义为$F_{R_i}(t)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{\frac{-(t-\mu)^2}{2\sigma^2}}$。取$\mu$=0，则有$\frac{F_{R_i}(t)}{F_{R_i}(t-1)}=\mathrm{e}^{\frac{1-2t}{2\sigma^2}}$，$\frac{F_{R_i}(t-1)}{F_{R_i}(t-2)}=\mathrm{e}^{\frac{3-2t}{2\sigma^2}}$和$\frac{F_{R_i}(t)}{F_{R_i}(t-2)}=\mathrm{e}^{\frac{-1}{\sigma^2}}$。已知

$F_{R_i}(t)$ 和 $F_{R_i}(t-2)$，可以预测 $F_{R_i}(t+2)$ 为：

$$F_{R_i}(t+2)=\frac{F_{R_i}^2(t)}{F_{R_i}(t-2)} \tag{6}$$

由式(6)，统计k周期及k-2周期的资源访问频度可预测k+2周期的资源访问频度 $F_{R_i}(T_{k+2})=\frac{F_{R_i}^2(T_k)}{F_{R_i}(T_{k-2})}$，由此，在$k$周期预测有 $F_{R_i}(T_{k+2})\geqslant\delta_{R_i}(T_k)$，则应该为资源$i$建立副本。

关于统计周期的选取，因节点的负荷变化较快，以常规资源的平均服务时间为其周期是适当的。但资源访问规律的变化周期一般较长，也可根据实际资源类型定制单位，通常以天为单位。上述预测机制分两种周期来实现。以天作为大周期，每一大周期中，再以常规资源的服务时间作为小周期统计访问频度及节点负荷，各大周期中的各小周期统计指标取平均值作为大周期的统计指标，预测以大周期为单位来实现。

1.2.2　副本删除

和创建不同，在没有足够空闲存储空间时，才选择适当的副本删除。按最近最久不用置换 (Least Recently Used，LRU)算法选择资源删除，定义资源R_i在周期k的加权访问频度 $B_{R_i}(T_k)$ 为最近n个周期的访问频度的指数加权平均，即 $B_{R_i}(T_k)=\frac{1}{n}\sum_{j=0}^{n-1}f_{R_i}(T_{k-j})\times e^{-j}$，当需要删除资源时，定义资源的副本访问频度 $A_{R_i}(T_k)$ 为 $A_{R_i}(T_k)=\frac{B_{R_i}(T_k)}{n_{R_i}(T_k)}$，式中 $n_{R_i}(T_k)$ 表示k周期时资源R_i的副本总数，选择 $A_{R_i}(T_k)$ 最小的资源删除其副本。

2　性能测试

可以用实验对上述的正态预测副本创建算法进行性能测试，并与文献[6]中的创建副本算法，以及线性预测、二次曲线预测算法对比，如图2～4所示。

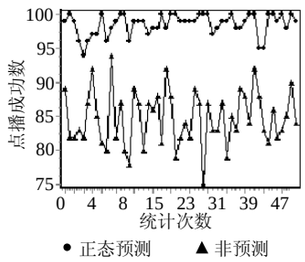

图2　正态预测与非预测的比较

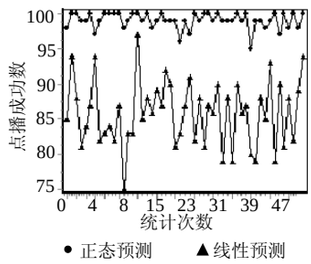

图3　正态预测与线性预测的比较

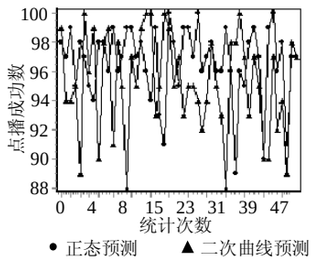

图4　正态预测与二次曲线正态预测的比较

测试时，服务器数为4，各服务器性能相同，服务能力为30路mpeg1视频流，系统总服务能力为4×30=120路；资源总数为500，每台服务器能容纳的在线资源数为50，系统总容纳资源数为4×50=200。

图2～4表明，本文提出的正态预测算法比非预测算法、线性预测算法和二次曲线预测算法都获得了较高的访问成功率，如表1所示。

表1　正态预测算法与其他几种算法访问成功率的比较

算　　法	访问成功率
非预测	0.850
线性预测	0.870
二次曲线预测	0.940
正态预测	0.975

(下转第231页)

参考文献

[1] Rosen E, Viswanathan A, Callon R. Multiprotocol label switching architecture[S]. IETF RFC3031, 2001.
[2] Ooms D, Livens W. IP multicast in MPLS networks[C]// Proceedings of the IEEE Conference on High Performance Switching and Routing, Heidelberg Germany, 2000.
[3] Ooms D, Sales B, Livens W, et al. Overview of IP multicast in a multi-protocol label switching environment[S]. IETF RFC3353, 2002.
[4] Boudani A, Cousin B. Simple explicit multicast [R]. IETF Internet Draft, 2002.
[5] Boudani A, Cousin B. A new approach to construct multicast trees in MPLS networks[C]// Seventh International Symposium on Computers and Communications, Proceedings ISCC, Taormina Italy, 2002.
[6] Boudani A, Cousin B. An effective solution for multicast scalability: the MPLS multicast tree [R]. IETF Internet Draft, 2003.
[7] Cui J H, Faloutsos M, Gerla M. An architecture for scalable, efficient, and fast fault-tolerant multicast provisioning[J]. Network IEEE, 2004, 18(2): 26 - 34.
[8] Mohammadi B M, Barzoki S S, Nikoopour M, et al. A case for dense-mode multicast support in MPLS[C]// Ninth International Symposium on Computers and Communications, Proceedings ISCC 2004, Gaithersburg, 2004.
[9] Fei A, Cui J, Gerla M, et al. Aggregated multicast: an approach to reduce multicast state[C]// Global Telecommunications Conference, San Antonio US., 2001.
[10] Yang B J, Mohapatra P. Edge router multicasting with MPLS traffic engineering[C]// 10th IEEE International Conference on Networks, Mumbai India, 2002.

编　辑　熊思亮

(上接第227页)

3　结 束 语

本文提出了一个基于动态阈值预测的副本自适应放置算法，同时给出了资源在线及在线位置的决策算法，实际系统上的性能测试结果说明运用这两个算法能进一步提高分布式并行BOD系统的服务性能，两个算法都可广泛应用于分布式并行BOD服务器，以提高其系统性能。

参 考 文 献

[1] Brubeck D W, Rowe L A. Hierarchical storage management in a distributed video on demand system.[J]. IEEE Multimedia, 1996, 3 (3): 37-47.
[2]Bisdikian C C, Patel B V. Cost-based program allocation for distributed multimedia on demand systems[J]. IEEE Multimedia, 1996, 3 (3): 62-72.
[3] Kyung A, Hoon C. Architecture of a VOD system with proxy servers[J]. IEICE IEICE Transactions on communications, 2000, E83-B(4):850-857.
[4] 李　勇, 吴　非, 陈福接. 大规模层次化视频点播存储系统的设计与管理[J]. 软件学报,1999, 10(4): 355-358.
[5] Zhang Z L, Wang Y W. Video staging: a proxy-server-based approach to end-to-end video delivery over wide-area networks networking[J]. IEEE/ACM Transactions on Networking, 2000, 8(4): 429-442.
[6] Hongtao Y, Chor P L, Atif Y. Design issues on video-on-demand resource management[C]// Proceedings of the 8th IEEE International Conference on Networks (ICON 2000), Washington DC USA, 2000.
[7] Kit S T, King T K, Chan S. Optimal file placement in BOD system using genetic algorithm[J]. IEEE Transactions on Industrial Electronics, 2001, 48(5): 891-897.

编　辑　熊思亮