计算机科学2002Vol.29№.12

VoD系统中磁盘存放策略的研究*)

Data Placement Policy in Distributed VoD Systems

乔 浩 朱晴波 张 潇 陈道蓄 谢 立

(南京大学软件新技术国家重点试验室 南京大学计算机科学与技术系 南京210093)

Abstract One of the significant tasks when designing a VoD(Video-on-Demand) system is to support as many simultaneous streams as possible. Obviously, it is closely related to the storage subsystem. In this paper, we compare the performance of the storage subsystems based on two different designs by simulation, one data placement policy is using the technique of striping, and the other is simply placing a whole video only in one server. We find that, when data is CBR (Constant-Bit-Rate) coded, the numbers of simultaneous streams respectively supported by the two schemes are equal, but the latter can present better service.

Keywords Constant-bit-rate, Video-on-demand, Striping

1. 引言

过去十年里,网络和多媒体技术的飞速发展,使VoD成为可能。VoD是一个一般的实例,它由客户发起接收远地存储的视频,要求在可接受的短时间内收到(有一定的实时要求),并在客户端回放,同时服务方提供的视频量具有相当大的规模[1]。它改变了传统的用户被动接受的方式(如收看电视),而代之以交互的方式,变革了人们传统的娱乐方式,具有极为广阔的前景。因此,引起人们广泛而深入的研究[1~4]。

VoD系统设计中,最终是为了提供尽可能多的并发数据流,因此设计系统时主要考虑以下几个问题:1、CPU的处理能力,2、网络的出口带宽,3、磁盘的出口带宽,4、内存的大小。网络的出口带宽和磁盘的出口带宽显然是与系统能支持的并发流数目直接相关的,就目前的状况而言,相对于网络的出口带宽和磁盘的出口带宽,CPU处理能力是足够的,因为服务器的CPU几乎没有大的计算任务,而主要用于处理发送数据。内存之所以与系统支持的并发流数目有关,是因为内存除了要存储即将转发给用户的数据,而且可以提高内存的命中率,使用户请求的数据直接在内存中得到满足,避免从磁盘中读取,从而减缓磁盘的压力,提高磁盘的吞吐率。但同时,VoD数据的大容量决定了内存只能存放极小部分的数据,大部分用户的服务需要磁盘来支撑,所以磁盘存放策略的选取至关重要。本文就是研究在多服务器的环境下,分片(stripping)与不分片两种磁盘存放策略对于系统性能的影响。

2. 相关工作

Striping是一种较为流行的存放方式[5~8],即将每个影片均匀地分成N块,再将这N块以轮转的方式存放到磁盘上。Striping最初是用来解决磁盘的带宽甚至连一个流都不能支持的情况,即用户的播放速率大于磁盘的带宽。同时发现striping也有利于服务器之间的负载平衡。流媒体文件有两种编码方式:CBR和VBR(Variable-Bit-Rate),stripping被认为是适合于存放CBR编码的流媒体文件[6]。后来,又有学者对Striping进行了改进[7],其策略如图1所示,将各影片的首块分散到各服务器上,以避免将所有影片的首块放到同一个服务器上,因为处于用户点播的高峰时段时,存放所有影片首块的服务器可能会成为瓶颈。而针对在VBR情形下,Striping,有些学者提出所谓的RIO,Santos比较了RIO与stripping的性能[9],VCR情形下,两者区别不大,而VBR情形下,RIO优于Striping。

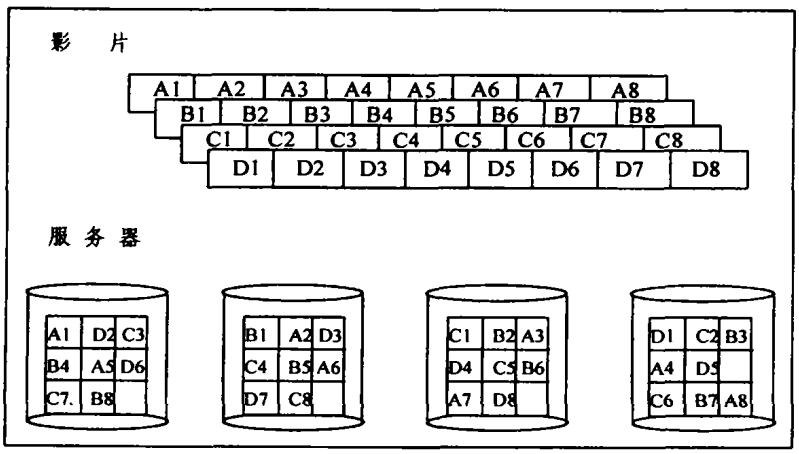

图1

*)基金项目:国家863项目——应用服务器的运行、管理及调度技术(编号2001AA113050)。

相对于分片存放的方式，不分片的存放方式是指，每一个影片作为一个整体存放在某一个服务器上，这样一个要考虑的问题就是，由于影片之间的热门程度不同，对于各部影片的用户请求率（到达率）就会有不同，导致服务器之间的负载不平衡，因此我们在分配影片时，应尽量使得各服务器之间保持平衡。

本文通过理论分析，比较了不分片与分片的两种策略对于系统支持的并发流的数目的影响，并给出模拟的结果。

这两种方式各有优劣，下面这个例子很好地说明了这一点：假设每个服务器的磁盘带宽只能支撑一个流，不考虑内存对磁盘的减负作用，即数据都需要从磁盘读取，而不能在内存中得到满足。假设图1描述的系统中的数据是 VCR 编码的，每个数据块对应的时间长度是 T，对于下面的用户到达过程：首先观看影片 A 的用户到达，T 时间后观看 B 的用户到达，显然第二个用户的请求将得不到满足。而对于另一种到达过程：首先观看影片 A 的用户到达，T 时间后又有一个观看影片 A 的用户到达，第二个用户的请求能得到满足。但如果系统是以不分片的方式组织的，且将 A、B 分配到不同的服务器上，则我们得到的结论恰恰相反：对于第一种到达过程两个用户的请求都能得到满足，而对于第二种到达第二个用户的请求将得不到满足。这提示了我们很可能对于系统能接入的用户数而言，两者是近乎相等的，下面的内容主要也是试图说明这一点。

3. 理论分析

为了讨论方便我们假设：1）所有影片有相同的时间长度；2）每个服务器所能支撑的用户数相等；3）影片是用 VCR 编码的，且影片的大小一致；4）用户到达是一个 Poisson 过程且没有两个用户同时到达（即用户到达的时间间隔服从指数分布）；5）服务器的存储容量相等。假设有 N 个服务器，用0到（N－1）加以编号，每个服务器所能支撑的用户数设为 C。每个服务器上的用户数用随机变量 X_i 表示，其中 i 为 N 个服务器的编号，用户到达率为 λ。这样出现服务器过载的概率 P 可以用右式表示：$P=Pr\{Max(X_i)>C\}$，其中 $Max(X_i)$ 表示诸 X_i 中最大的一个。

为了便于计算，在分片的情况下，我们假设每个影片都被均匀地分割到各个服务器上，即每部影片在每个服务器上的长度都相等。在不分片的情况下，由于观看各个影片的用户到达率是不相同的，我们假设将诸影片分配到各个服务器上之后，每个服务器上影片的到达率之和相等，这在一个影片数量较多的大系统中是可以近似做到的。我们分别计算了一下分片情况下的概率 P1和部分片情况下的概率 P2，发现两者相等（见附录略），说明在上述条件下，虽然分片能使得在一个影片长度的时间内，各服务器的平均负载相等，但就接入用户数而言，两者是没有区别的。

4. 模拟结果

图2中，横坐标表示服务器数目，纵坐标表示一个影片长度时间内服务不能被满足的用户数，影片长度为100分钟，每个服务器的可以支撑15个用户，平均每个服务器的到达率为 $\lambda=\lambda/N=1$人/（8分钟）。

从图2可以看到在用户到达率相同的条件下，请求不能被满足的用户的数目在两种情况下是相差无几的。这说明了两种情况下服务器过载的概率是近乎相等的。而一个系统所能支持的流的数目，是指保证一定服务质量的情况下的流的数目，而服务质量与服务器过载的概率是直接相关的，由于不考虑两个用户是同时到达的情况，而且，每个用户都有一个独立的流为他服务，图2的数据说明，两种策略在系统支撑的流的数目相同的情况下，系统中的服务器出现过载的概率是相同的，这也说明了在保证相同的服务质量的前提下，根据两种不同策略设计的系统可以提供的并发流的数目是相同的。

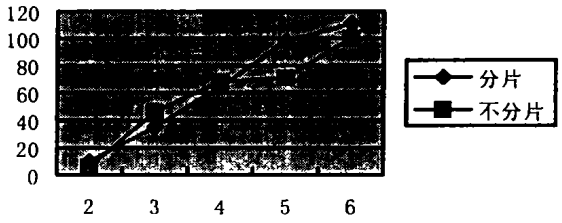

图2

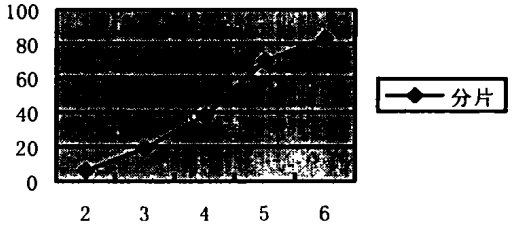

图3

此外，在分片的情况下，由于用户请求会在服务器之间的跳转，造成有可能跳转之后原本没有出现过载的服务器，发生了过载，因此不能保证开始时用户的请求能被满足，以后用户的请求也能被满足，一个设计良好的 VoD 系统应该有基本的接入控制机制，所谓接入控制是指如果系统认为自己不能很好地满足某个用户的请求，就拒绝该用户接入[10]。由于接入控制只是在用户接入时判断系统是否能支撑该用户的请求，无疑分片给接入控制增加了难度。而无分片的情况下，用户不会在各服务器之间跳转，所以接入控制会相对简单。图3中我们统计了分片情况下，如果只是将用户当前请求的数据块能否被满足作为接入与否的标准，用户在观看过程中会发生由于磁盘带宽不足而导致用户请求不能被满足的情况的数目。可以看出这种情况出现的数量还是很多的，决不是个别现象，一个好的接入控制策略应该考虑到这种情况的发生。

结论 我们在数据是以 VCR 编码的情况下，比较了采取分片策略设计的存储系统与采取不分片策略的系统之间的性能，发现相同条件下，两者能支撑的用户数是相同的，但前者会给接入控制策略的设计带来困难，而后者的接入控制策略相对简单。

根据引理1，利用归纳法，不难得到对于某部影片 j，观看影片 j 的人数服从参数为（λ_j，L/N）的 poisson 分布，再根据引理2，用归纳法，不难得到各服务器上的用户数服从参数为（$\sum_j \lambda_j$，L/N）的 Poisson 分布。

再来考虑不分片的情况，我们假定我们能够将影片按各影片到达率的不同分开，使得每个服务器有相同的用户到达率，设总的用户到达率为 λ，由于影片按到达率均匀地分配到各服务器上，这样每个服务器的到达率为 λ/N，则各服务器上

（下转第94页）

用的角色，但我们认为它们对本体库的使用机制是很类似的，所以把它们在“基于 Web 的分布式本体”环境下归于一类。

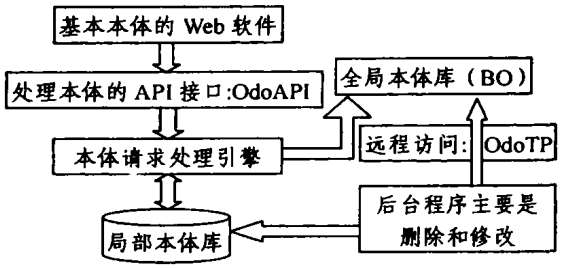

图3 SUO 上本体系统结构

在图3中，我们的 OdoAPI（Open distributed Ontology API）参考了 DAML API。使用这个 API，第三方开发的基于本体的 Web 软件可以在不用关心分布式本体的处理机制的情况下，简单而透明地使用一切符合规划的本体。图3中的“局部本体库”完成了对本地本体的存储与管理。

图3中的“本体请求处理引擎”是关键，本体的分布式策略及相关算法在这里实现。上图中的“后台程序”是按一定算法对局部本体库的本体进行删除和修改的操作。而这两者的操作都需要通过一定的协议来和远端的 GOB 交互。我们将开发 OdoTP（“Open Distributed Ontology Transport Protocol”）来完成这种交互的的信息格式（建立在 DAML 基础之上）和交互粒度的定义。

结束语 几乎从 Internet 开始普及推广的时候起，人们就一直在描绘智能搜索引擎、个性化智能信息推荐、智能代理、智能电子商务、个性化电子服务、基于 Web 的 KOD（Knowledge on Command）等的美好前景。但这些从来就没有应用推广过，它们就象“科幻小说”一般遥不可及。究其原因，Web 智能服务所必需的面向 Web 的基于本体的信息及知识表达相关标准、理论和技术还有待发展。

另一方面，本体论在人工智能中的运用也由来已久。但在实际应用中并没有广泛的效果。Web 的出现为本体论提出了新的、更广泛的需求，也提供了一个建立真正全球化的、统一的、标准化的本体环境的途径。

因此，基于本体的智能 Web 服务 IWSBO 是必然的发展方向，相关的研究课题如 Semantic Web，DAML 等已成为研究热点。本文提出了自己对 IWSBO 的理解，同时提出它必须要有一个重要基础——面向 Web 的分布式本体系统 WODOS。本文提出了一个 WODOS 的模型，并对其进行了初步设计。

参 考 文 献

1 Hendler J. Agents and the Semantic Web. IEEE Intelligent Systems, March/April 2001

2 LuRu-qian, Jin Zhi, Wang Rong-lin, et al. An approach of acquiring requirement information based on domain knowledge. Journal of Software, 1996, 7(3)

3 McGuinness D L, Fikes R, Rice J, Wilder S. The Chimaera Ontology Environment. In: Proc. of the Seventeenth National Conf. on Artificial Intelligence (AAAI 2000). Austin, Texas, 2000

4 McGuinness D L. Conceptual Modeling for Distributed Ontology Environments. In: Proc. of the Eighth Intl. Conf. Conceptual Structure Logical, Linguistic, and Computational Issues (ICCS 2000), Darmstadt, Germany, 2000

5 Heflin J, Hendler J. A Portrait of the Semantic Web in Action. IEEE Intelligent Systems, March/April 2001

6 Neches R, Fikes R E, Finin T, et al. Enabling technology for knowledge sharing. AIMagazine, 1991, 12(3)

7 McIlraith S A, Son T C, Zeng Honglei. Semantic Web Service. IEEE Intelligent Systems March/April 2001

（上接第113页）

的用户服务参数为($\lambda/N, L$)的 poisson 分布，由于 $\lambda=\sum_j \lambda_j$，因此($\sum_j \lambda_j$, L/N)的 poisson 分布=(λ, R/N)的 poisson 分布=(λ/N, L)，所以 $P=Pr\{Max(X_i)>C\}=\prod_i (Pr(Xi>C))$对于两种存放方式是一致的。

参 考 文 献

1 Cleary K. Video on demand-competing technologies and services, Broadcasting Convetion, 1995. IBC 95., International, 1995. 432～437

2 Anderson D P, Osawa Y, Govindan R. A File System for Continuous Media. ACM Transaction on Computer System, 1993, 11(2)

3 Barnett S A, Anido G J. A cost comparison of distributed and centralized approaches to video-on-demand, Selected Areas in Communications. IEEE Journal on 1996, 14(6): 1173～1183

4 吴敏强，周刚，陈晓林，陆桑璐，陈道蓄，谢立. 分布式视频点播服务器关键技术分析. 计算机科学（录用待发表）

5 Berson S, Muntz R, Wong W. Randomized data allocation for real-time disk I/O. In compcon 96, 1996. 286～290

6 Berson S, et al. Staggered striping in multimedia information systems In ACM SIGMOD 94, ACM, 1994. 79～90

7 Ozden B, Rastogi R, silberschatz A. Disk striping in video server environments. In: Proc. of the Intl. Conf. on Multimedia Computing and Systems, Hiroshima, Japan, June 1996. 172～180

8 Brinkmann A, Salzwedel K, Scheideler C. Efficient, Distributed Data Placement Strategies for Storage Area Networks. In: Proc. 12 ACM Symposium on Parallel Algorithms and Architectures (SPAA), 2000

9 Santos J R, Muntz R R, Ribeiro-Neto B. Comparing Random Data Allocation and Data Striping in Multimedia Servers SIGMETRICS 2000, 6/00, Santa Clara, California, USA

10 Vin H M, Goyal P. A statistical admission control algorithm for multimedia servers. In: Proc. of ACM Multimedia, San Francisco, Oct. 1994. 33～40