多媒体存储系统的两级粒度方式的设计*)

刘晓光[1] 王 刚[2] 吴 英[2] 刘 璟

(广发证券博士后站 广州510075)[1] (南开大学计算机系 天津300071)[2]

A Research on TWO-Level Granularity in Multimedia on Demand System

LIU Xiao-Guang[1] WANG Gang[2] WU Ying[2] LIU Jing[2]

(GF Securities Research, Guangzhou 510075)[1] (Department of CS, Nankai University, Tianjin 300071)[2]

Abstract Multimedia on demand, such as VOD and AOD, have popularized with the development of network technology recently. The Storage systems of multimedia on demand have some special requirements. The two-level granularity is a new method to stripe unites in storage devices. It integrates the virtues of coarse granularity and fine granularity.

Keywords Granularity, Stripe, Storage, Multimedia

1. 问题的描述

随着多媒体技术的发展,多媒体的应用日益广泛。同普通文件相比,多媒体文件一般都具有实时性的要求,即文件的传输必须能够保证一定的传输速率。例如,MPEG-1标准的视频文件的传输率要不低于1.5Mbps。因此,这类文件又被称为连续媒体文件(CM,Continuous Media),主要包括 video,audio 等。相应地,多媒体存储服务器也需要具有一些不同于传统存储服务器的特点:第一,实时性要求。第二,要进行大量的定期和顺序性的读写操作。第三,大容量的要求。本文的内容就是分析在点播服务中,存储系统条纹单元的不同粒度方式对系统性能的影响。

一个点播服务器自身拥有的资源(包括存储系统、内存和网络带宽等)都是有限的。因此,它必须设定一个 Admission Control 机制。这一机制的基本流程如下:在一个新的合法 Client 提交服务请求后,首先,服务器根据这一请求计算出所需要提供的系统资源。然后判断现有的可用资源是否可以满足这一请求。如果可以满足,则接受该请求,否则拒绝这一请求。点播服务器的基本工作原理如下:因为需要对 CM 文件定期进行 I/O 操作,所以点播服务器要设定一个访问周期 T。在 T 时间内,必须完成对合法 Client 请求的响应。假设第 i 个 Client 要求的媒体播放率为 r_i,则系统内存中预先读取的相应数据量 $D \geqslant r_i \times T$。实际上,系统为每个 Client 设定两个 Buffer,大小一般都设定为 D。其中一个用于发送而另一个则用于接收。发送 Buffer 负责向 Client 发送数据,接收 Buffer 负责从存储系统读取数据。正常状态下,当发送 Buffer 的数据发送完毕后,接收 Buffer 中数据也应读取完成。这时发送 Buffer 和接收 Buffer 的角色互换,原有的接收 Buffer 成为发送 Buffer,从而保证了 Client 端 CM 文件的连续播放。

本文主要讨论存储系统条纹单元粒度方式的影响,因此设定下列前提假设:第一,所有工作都是针对 CBR(Constant Bit Rate)情况,即假设所有 Client 的播放率都相同为 r_{disp}。第二,假设点播服务器拥有足够的网络带宽和内存等资源为 Client 提供访问服务,即只考虑存储系统的影响。第三,为了简化表述,除非特别说明。否则都是以读操作和 RAID 5条纹为例进行分析。

2. 相关的工作

在不同环境下,如何确定最优的存储系统条纹单元粒度,是多媒体存储系统领域的一个重要研究课题。Chen 和 Katz 研究了以 non-CM 文件为主的负载情况下,条纹单元大小和条纹长度的确定问题[1,2]。UCLA 的 Rio 是为虚拟现实研究服务的存储系统[3]。它研究的重点是 VBR(Variable Bit Rate)情况下(如虚拟现实和交互式游戏等),条纹单元的设定问题。Rio 将数据随机地分配到系统所有磁盘的物理空间中,将所有的负载模式在磁盘块的层次上转化为一种均衡的随机访问模式。它的目的是平衡虚拟现实中不同媒体格式对数据传输率的不同要求。Bell Labs 的 Fellini 多媒体存储服务系统[4]侧重于提高有线电视网络环境下不同格式的 CM 文件的综合服务质量。具体到点播服务器的存储系统,现有的系统主要使用两种粒度方式:粗粒度方式和细粒度方式[5]。Ozden 分析了 CBR 情况下条纹单元不同粒度的影响,他的结论是粗粒度条纹比细粒度条纹更适合应用于点播服务器的存储系统。

3. 不同粒度方式的对比

细粒度方式的条纹单元很小,在一个访问周期内一个数据请求会跨越整条条纹,因此组成条纹的所有硬盘都参与了数据操作。假设一个条纹中有 c 个数据单元,则对 RAID 5方式而言,条纹长度为 $c+1$。在一个访问周期内,一个数据请求由 c 个硬盘并行完成。由于 CM 文件具有顺序性的特点,数据在硬盘上是顺序存储的,所以磁头的移动也基本上是单向的。在最差情况下,硬盘的寻道时间是平均寻道时间的两倍。根据 CM 文件的实时性要求,可以得到:

$$\frac{q \times T \times r_{disp}}{c \times r_{disk}} + q \times t_{rot} + 2t_{seek} \leqslant T \qquad (1)$$

其中,q 是存储系统支持的并发数据流的数目,r_{disp} 是媒体播放率,r_{disk} 是硬盘内部平均传输率,t_{seek} 是寻道时间,t_{rot} 是最差旋转延迟时间。存储系统的最大并发数据流数目为:

*)课题研究得到国家"863"计划资助,编号:863-306-ZD01-02-6.刘晓光 博士,主要研究方向为并行与分布式系统等。王 刚 博士,主要研究方向为并行与分布式系统等。吴 英 博士研究生,主要研究方向为并行与分布式系统等。刘 璟 教授,博士生导师,主要研究方向为并行与分布式系统、算法分析、VLSI等。

$$Q_1=\max(q)=\left\lfloor\frac{T-2t_{seek}}{T\times r_{disp}/c\times r_{disk}+t_{rot}}\right\rfloor$$
$$=\left\lfloor\frac{c\times r_{disk}\times(T-2t_{seek})}{T\times r_{disp}+c\times r_{disk}\times t_{rot}}\right\rfloor \quad (2)$$

粗粒度方式的条纹单元很大，在一个访问周期内一个数据请求只在一个硬盘上完成。同一时刻，存储系统中的不同硬盘可以处理不同的 Client 请求。则对于一个硬盘而言有：

$$\frac{q\times T\times r_{disp}}{r_{disk}}+q\times t_{rot}+2t_{seek}\leqslant T \quad (3)$$

在式(3)中，q 是一个硬盘支持的并发数据流的数目，则存储系统支持的并发数据流数目是 $c\times q$。存储系统的最大并发数据流数目为：

$$Q_2=\max(c\times q)=\left\lfloor\frac{c\times(T-2t_{seek})}{T\times r_{disp}/r_{disk}+t_{rot}}\right\rfloor$$
$$=\left\lfloor\frac{c\times r_{disk}\times(T-2t_{seek})}{T\times r_{disp}+r_{disk}\times t_{rot}}\right\rfloor \quad (4)$$

因为 c 一般是不小于1的正整数，对比式(2)和式(4)，可以知道：$Q_2\geqslant Q_1$(其中 $c=1$时相等)。这也是在点播服务器中，粗粒度条纹存储系统支持的最大并发 Client 数目多于细粒度存储系统的原因。因此，粗粒度方式更适合于对并发性要求比较高的应用环境。粗粒度方式的缺点是容易形成热点：因为热门的 Video 或 Audio 文件的点播率非常高，使得存储这些文件的区域成为热点，造成系统服务的堵塞。

细粒度方式和粗粒度方式各有其优缺点：细粒度方式并发性较差，但是它实现了最大的负载均衡，不会出现热点现象。细粒度方式适合应用于对并发性要求不高，而对服务质量要求很高的点播服务。粗粒度方式的并发性能好，但是容易出现热点现象，降低了点播服务器的服务质量，它适合于对并发性能要求很高的应用环境。由此可见，粗粒度和细粒度方式两者是互相补充的，细粒度的优点正好可以弥补粗粒度的不足，反之亦然。但是，大多数现有存储系统使用的单一数据布局方式决定了在一个存储系统中只能存在一种粒度方式，两者不可能混合使用。

本文给出的两级粒度方式正是结合了粗粒度和细粒度的优点而提出的一种新的混合粒度方式。两级粒度方式是在两极数据布局的基础上实现的。所谓两极数据布局就是存储系统的硬盘分为两个层次，分别使用不同的数据布局形式，以提高存储系统的性能和效率。例如，常用的 RAID 10就是在第二级(下层)使用 RAID 1数据布局，而在第一级(上层)使用 RAID 0数据布局。在两级粒度方式中，第一级使用的是粗粒度方式，而第二级使用的是细粒度方式。一个外部的数据请求被发送给存储系统以后，它首先在第一级上以粗粒度方式被处理。这一数据请求被分解为大小为粗粒度条纹单元的多个数据请求。然后这些请求被转向第二级。在第二级上，这些粗粒度条纹单元请求被分解为多个细粒度条纹单元请求，平均分布到各个物理硬盘上。由它们并行完成具体的 I/O 操作，则有：

$$\frac{q_g\times T\times r_{disp}}{r_{array}}+q_g\times t_{array-rot}+2t_{array-seek}\leqslant T \quad (5)$$

其中，r_{array}、$t_{array-rot}$和 $t_{array-seek}$是使用第二级布局方式磁盘阵列系统的稳定数据传输率、旋转延迟和寻道时间。如果不考虑其它因素的干扰，则有 $r_{array}=N\times r_{disk}$，而旋转延迟和寻道时间与硬盘的相应指标相同。在这里，假设存储系统规模是$(G+1)\times(N+1)$，$G+1$是第一级布局的粗粒度条纹长度，$N+1$是第二级布局的细粒度条纹长度，q_g 是第一级中单个粗粒度条纹单元支持的并发数据流的数目，则有：

$$\frac{q_g\times T\times r_{disp}}{N\times r_{disk}}+q_g\times t_{rot}+2t_{seek}\leqslant T \quad (6)$$

两级数据布局存储系统支持的并发数据流数目 $G\times q_g$，则最大并发数据流数目

$$Q_3=\max(\lfloor G\times q_g\rfloor)=\left\lfloor\frac{G\times(T-2t_{seek})}{T\times r_{disp}/N\times r_{disk}+t_{rot}}\right\rfloor$$
$$=\left\lfloor\frac{G\times(T-2t_{seek})\times N\times r_{disk}}{T\times r_{disp}+N\times r_{disk}\times t_{rot}}\right\rfloor \quad (7)$$

从公式7中可以看出，G 和 N 都与 Q_3成正比关系。同时 N 的增加，也意味着分担数据请求负载的硬盘数目的增加，则出现“热点”的概率也将降低。但是考虑到成本等因素，G 和 N 的大小不可能无限扩大，需要根据实际应用情况来确定。

表1对比分析了不同粒度方式的差异。其中，在对具体数值的计算中，依据下列假设条件：$c+1=(G+1)\times(N+1)=36$，$t_{seek}=4.9$ms，$t_{seek}=2.99$ms，$T=1$s，$r_{disp}=1.5$Mbit/s，$r_{disk}=36$Mbytes/s。此外，在两级粒度方式的计算中，分为 $G=11$、$N=2$和 $G=8$、$N=3$两种情况分别计算。

从表1中可以看出，两级粒度方式的每个数据盘允许的最大并发请求数目可以达到粗粒度方式的60%左右，同时它的负载情况是粗粒度方式的 N 倍。综合考虑，两级粒度方式更适合应用于大规模的点播服务器系统。

表1 不同粒度方式的对比

粒度方式	特点	负载情况	最大并发数目	每数据盘并发数目
细粒度(RAID 5)	负载均衡，但并发性差	c	315	9
粗粒度(RAID 5)	并发性好，但容易出现热点	1	4270	122
两级粒度(两级布局)	并发性能和负载的均衡都比较好	N	1947 (G=11, N=2) 1676 (G=8, N=3)	88.5 (G=11, N=2) 69.8 (G=8, N=3)

总结 在多媒体的点播服务中，存储系统的性能是影响点播服务器服务质量一个重要因素。与现有的单纯的粗粒度方式和细粒度方式相比，两级粒度方式在结合了两者的优点的同时，还避免了它们各自的缺点。表现在：

· 并发性能好。两级粒度方式通过第一级的粗粒度提高了存储系统的并发性能。在这一方面，它与粗粒度方式相同。

· 实现了负载的均衡。两级粒度方式通过第二级的细粒度把外部的数据请求平均分散到多个硬盘上，实现负载的均衡，避免了热点的出现。在这一方面，它与细粒度方式相同。

事实上，在我们所实现的两级网络 RAID 存储系统中[6]，已经应用两级粒度方式作为存储系统条纹单元的组织方式，并且在实际应用中表现出了良好的性能表现。

参考文献

1 Chen P M, et al. RAID: High Performance, Reliable Secondary Storage. ACM Computing Surveys, 1994, 26(2): 145～185

2 Chen P M, et al. An Evaluation of Redundant Arrays of Disks Using an Amdahl 5890. In: Proc. of the 1990 ACM Conf. on Measurement and Modeling of Computer Systems (SIGMETRICS), Boulder CO, May 1990

3 Muntz R, Santos J R, Berson S. RIO: A Real-time Multimedia Object Server. Performance Evaluation Review, ACM Press, 1997, 25(2): 29～35

4 Martin C, et al. The Fellini Multimedia Storage System. Journal of Digital Libraries, 1997

5 Ozden B, Rastogi R, Silberschatz A. Disk Striping in Video Sever environments. In: Proc. of the IEEE Intl. Conf. on Multimedia Computing and Systems, Hiroshima, Japan, June 1996

6 刘晓光，王刚，曾昭智，刘璟. 多节点分级网络 RAID 存储结构研究. 计算机科学，2000，28(11)