学校编码：10384　　分类号_____密级_____

学号：23020111153074　　UDC_____

厦门大学

硕士学位论文

高清视频服务器磁盘 I/O 调度算法的研究

The Study of Disk I/O Scheduling Algorithm for High Definition Video Servers

刘硕

指导教师姓名：卢 伟 副教授

专 业 名 称：计算机软件与理论

论文提交日期：2014 年　　月

论文答辩日期：2014 年　　月

学位授予日期：2014 年　　月

答辩委员会主席：_____

评　　阅　　人：_____

2014 年　　月

2014 年 5 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

2014 年 5 月 20 日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（ √ ）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

2014年 5月 20日

摘要

服务器前置方案是否具有可行性的关键之一，就是如何尽可能地发挥硬件潜能，提高单台视频服务器的性价比。目前视频服务器的主要性能瓶颈在于资源磁盘的带宽，尤其是在多线程环境下，使用主流操作系统的磁盘 I/O 将导致磁盘带宽大大下降，严重制约服务器的性能。本文的主要工作是，详细研究了 Linux 内核块设备的 I/O 子系统，对 Linux 提供的四种 I/O 调度算法做了深入的剖析；在此基础上，针对分条式高清视频点播服务器读取资源磁盘的特点，设计了一个专用的 I/O 调度算法——HVOD，并且在 Linux 2.6.32 内核中实现了该算法。HVOD 算法通过精确的读预测提升磁盘顺序读的性能，使用超时队列避免进程的 I/O 饥饿，从而确保尽可能多的视频流能够流畅播放。在真实 VOD 系统和模拟 VOD 系统上进行的测试表明，在分条式高清视频点播服务器中，HVOD 算法与 Linux 现有的 I/O 调度算法相比，可以使视频服务器的服务能力提升 40～50%左右。我们的研究结果证明，服务器前置方案即使是对于大规模开展 4K 高清视频的点播服务也具有现实可行性。

关键词：HVOD；服务器前置；磁盘性能；I/O 调度算法

Abstract

It is one of the keys to the feasibility for the SDFC (servers are deployed along the frontier of cilents) to make good use of the hardware capacity so as to promote the cost performance as much as possible for a single video server. At present, the bandwidth of resource hard disks is the main bottleneck for video servers, because, especially in a concurrent environment, simultaneous access to a hard disk will lead the disk bandwidth to a serious decline to severely restrict the capacity of video servers. The followings are the major works of this paper. We investigate the I/O block devices subsystem in Linux and give in-depth analysis and evaluation for all of the four I/O scheduling algorithms provided in Linux. On the basis of the above, we design a dedicated I/O scheduling algorithm, called HVOD, for a video server in which the video resources are striped and stored in every resource hard disk, and implement the algorithm in Linux 2.6.32 kernel. The HVOD algorithm makes precise read prediction to improve the performance for sequential disk reads and employs timeout queues to prevent I/O starvation, so as to make sure as many videos are replayed smoothly as possible. The tests, which are made in a real VOD system and in a VOD simulator, show that, in a video server that video resources are striped, the performance of HVOD is 40~50% higher than the present I/O scheduling algorithms provided in Linux. Our research confirms that the SDFC scheme have realistic feasibility even for the large-scale VOD services that provide 4K high definition videos.

Keywords: HVOD; SDFC; Disk Performance; I/O scheduler

目录

第一章 绪论 1

1.1 高清视频点播系统 1

1.2 需要解决的关键问题 2

1.2.1 机械磁盘的并发访问 2

1.2.2 固态硬盘的局限性 4

1.3 研究内容及其意义 5

1.4 结构组织 6

第二章 Linux 块设备及其 I/O 子系统 7

2.1 块设备 7

2.2 I/O 操作 7

2.2.1 虚拟文件系统 8

2.2.2 磁盘高速缓冲 9

2.2.3 文件系统 9

2.2.4 通用块层 10

2.2.5 I/O 调度层 12

2.2.6 块设备驱动 13

第三章 Linux I/O 调度算法的剖析和评价 15

3.1 算法简介 16

3.2 算法接口 17

3.3 调度过程 19

3.4 Noop 20

3.5 Deadline 22

3.6 Anticipatory 24

3.7 CFQ 26

第四章 HVOD 算法设计 33

4.1 分条式存储策略 33

4.2 调度策略 34
4.3 读预测 34
4.4 超时处理 36
4.5 算法描述 37
第五章 HVOD 算法实现 41
5.1 用户层与内核通信 41
5.2 数据结构 42
5.3 算法接口 46
第六章 HVOD 性能测试 51
6.1 实验环境 51
6.2 并发环境下的性能 52
6.2.1 测试方法 53
6.2.2 测试结果与分析 53
6.3 真实系统测试 55
6.3.1 真实系统简介 55
6.3.2 测试方法 56
6.3.3 测试结果与分析 57
6.4 模拟系统测试 58
6.4.1 模拟系统简介 58
6.4.2 测试方法 58
6.4.3 测试结果与分析 60
6.5 通用性测试 61
第七章 总结与展望 65
7.1 总结 65
7.2 未来工作 66
参考文献 67
附录 69

Contents

Chapter 1 Introduction 1

1.1 HD Video on Demand 1

1.2 Key Problem 2

1.2.1 Simultanious Process of HDD 2

1.2.2 Limitation of SSD 4

1.3 Contents and Significance 5

1.4 Organization 6

Chapter 2 Block Device and I/O Subsystem in Linux 7

2.1 Block Device 7

2.2 I/O Operations 7

2.2.1 Virtual File System 8

2.2.2 Disk Cache 9

2.2.3 File System 9

2.2.4 Generic Block Layer 10

2.2.5 I/O Scheduler Layer 12

2.2.6 Block Device Driver 13

Chapter 3 Analysis and Evaluation of I/O Schedulers in Linux 15

3.1 Introduction 16

3.2 Scheduler Interface 17

3.3 Scheduling Process 19

3.4 Noop 20

3.5 Deadline 22

3.6 Anticipatory 24

3.7 CFQ 26

Chapter 4 HVOD Algorithim 33

4.1 Striped Storage Arrangement 33

4.2 Scheduling Policy 34
4.3 Read Prediction 34
4.4 Timeout Handling 36
4.5 The Algorithim 37
Chapter 5 Implementation of HVOD 41
5.1 Communication Between User and Kernel 41
5.2 Data Structure 42
5.3 Interface 46
Chapter 6 Performance Tests for HVOD 51
6.1 Experimental Environment 51
6.2 Simultaniously Access Performance 52
6.2.1 Test Methods 53
6.2.2 Test Results and Analysis 53
6.3 Tests on a Real System 55
6.3.1 The Read System 55
6.3.2 Test Methods 56
6.3.3 Test Results and Analysis 57
6.4 Tests on a Simulator 58
6.4.1 The Simulator 58
6.4.2 Test Methods 58
6.4.3 Test Results and Analysis 60
6.5 Generality Test 61
Chapter 7 Summary and Futrue Works 65
7.1 Summary 65
7.2 Futrue Works 66
Reference 67
Appendix 69

第一章 绪论

1.1 高清视频点播系统

视频点播系统（VOD）[1]允许用户根据自身的爱好来实时点播视频内容，同时也可以进行快进、快退、暂停等交互式操作。目前在广域网内进行网络视频点播的技术已经非常成熟，提供 VOD 服务的主要方式有 IPTV，CDN 和 P2P。由于网络限制，以上服务方式仅能提供 VCD 和 DVD 画质的视频，以及分辨率为 1080P 但画质较低的视频。

随着人们生活水平的日益提高，大屏幕数字电视机逐渐普及，1080P 规格的视频已经成为全高清视频的标准。相对于普通的视频点播服务，全高清视频点播需要更高质量的网络传输。若全高清视频码率平均按 15Mbps 计算，则不到 7000 个用户就会消耗掉 100Gbps 的带宽，更何况广域互联网还存在着传输延迟长、带宽不足以及延迟不确定等问题，整个主干网络的带宽也不可能仅仅服务于视频点播。最近几年出现的 4K 视频[2]，可以提供 4 倍于全高清视频的画质，显示分辨率能达到 2160P，可望在不远的将来得到普及。若要提供 4K 的高清视频点播服务，其带宽消耗是全高清的 4 倍，那么 7000 个用户就需要消耗 400Gbps 以上的网络带宽。因此通过广域网大规模地开展真正的高清视频点播服务是不现实的。

对此，可以采用服务器前置方案[3, 4]，即把一台或者多台视频服务器部署在每个小区的交换机房或者 ADSL 局域网内，组成一个分布式服务器集群。如此便可以把绝大部分视频流量限制在局域网内，从而避免广域网的带宽限制。虽然这种解决方案能克服网络的带宽障碍，但是会导致高昂的部署成本。目前一台普通的服务器只能同时提供二、三十个全高清视频流服务，若一个居民小区的平均用户规模按 300 计算，则需要大约 15 台服务器才能满足服务需求；如果要提供 4K 视频的视频点播服务，则需要的服务器数量将超过 60 台。这样每个小区仅仅服务器就需要几十万乃至几百万的部署成本，代价高昂；且服务器数量越多，占用的空间和功耗就越大，故障率也越高，从而导致很高的运行维护成本。因此，

必须大大降低系统成本，尤其是大大提高单台服务器的服务效能，服务器前置方案才具有可行性。

1.2 需要解决的关键问题

高清视频点播系统属于 I/O 密集型应用，整个系统的性价比由它提供的服务带宽所决定，而视频服务器的服务带宽取决于网络带宽和磁盘带宽的最小值。当采用服务器前置方案后，网络带宽不再是问题，磁盘带宽就成了制约服务器服务能力的瓶颈。按最新的 PCI-E 3.0 标准[5]，主板 I/O 总线单向单通道带宽可以达到 1GB/s 左右，因此理论上，单台常规服务器可以对外提供最大 1GB/s 的服务带宽，如果能充分发挥硬件的性能，即便是应付 4K 高清视频的点播，服务器前置方案也具有现实可行性。

1.2.1 机械磁盘的并发访问

受限于传统硬盘的机械式构造，当磁盘处理多个并发访问时，性能会大幅下降。目前服务器使用的主流操作系统有 Windows 和 Linux。图 1-1 显示了在多线程环境下，Windows 和 Linux（Linux 提供了 CFQ、Anticipatory、Deadline、Noop 四种不同的 I/O 调度算法）顺序读取大文件的磁盘带宽，从图中可以看出，单线程读取磁盘时，磁盘带宽可以达到 100MB/s 左右，但是在多线程环境下，Windows 的磁盘带宽下降到了单线程的 10%左右，而 Linux 的 CFQ 和 Anticipatory 算法则下降到了单线程的 70%左右。同时也可以看出，在 Linux 上，不同的 I/O 调度算法的性能是不相同的。