◎研发、设计、测试◎

分布式大规模监控视频存储系统 THNVR

邬建元[1,2],顾 瑜[1,2],鞠大鹏[1,2],汪东升[1,2]
WU Jian-yuan[1,2],GU Yu[1,2],JU Da-peng[1,2],WANG Dong-sheng[1,2]

1.清华大学 计算机科学与技术系,北京 100084
2.清华信息科学技术国家实验室,北京 100084
1.Department of Computer Science and Technology,Tsinghua University,Beijing 100084,China
2.Tsinghua Information Science and Tech. National Lab,Beijing 100084,China
E-mail:jy-wu02@mails.tsinghua.edu.cn

WU Jian-yuan,GU Yu,JU Da-peng,et al.THNVR:Distributed large-scale surveillance video storage system.Computer Engineering and Applications,2009,45(31):56-59.

Abstract: A novel approach to achieve high surveillance video storage performance via commodity SATA disks is introduced to meet the storage requirements of network video surveillance applications.The solution uses fixed-size files and manages structured and unstructured data separately to resolve the disk fragment problem thoroughly.It employs an adaptive buffering strategy,which regularizes and schedules the accesses to disks efficiently.It also implements features such as fault-tolerance and scalability. Performance analysis shows that the solution can carry out 256 ways of D1 or 512 ways of CIF video streams with one commodity computer setup.

Key words: video surveillance;video storage;disk fragmentation;fixed-size files

摘 要:通过分析网络视频监控应用中的存储需求特点,提出了一种采用低成本 SATA 硬盘的高性能存储解决方案。该方案使用定长文件,并将结构化与非结构化数据分别存储和检索,彻底克服了传统解决方案中磁盘碎片导致存储性能下降的问题。该方案还采用了自适应的缓冲技术,对磁盘访问进行了有效的规整和调度,充分利用了磁盘的写入带宽。此外,该方案还实现了容错、扩展等重要功能。性能分析显示,该方案在普通计算机和 SATA 磁盘上能够同时记录 256 路 D1 或 512 路 CIF 视频数据流。

关键词:视频监控;视频存储;磁盘碎片;定长文件
DOI:10.3778/j.issn.1002-8331.2009.31.018 **文章编号**:1002-8331(2009)31-0056-04 **文献标识码**:A **中图分类号**:TP3

1 引言

随着视频监控系统不断向数字化和网络化方向发展以及视频监控技术的逐渐普及,使用普通计算机及磁盘进行视频监控数据的存储已成为一种趋势,逐步取代传统的磁带等存储介质。

随着网络视频监控的发展,新的应用场景带来了新的挑战。网络视频监控应用本身的特点决定了其对比传统视频监控有更高视频质量的要求。视频监控的数字化为实现更高的清晰度提供了可能性。但随着清晰度的提高,需要存储的数据量相对于传统的监控视频将显著上升,加之摄像头数量的快速增长,对运行视频监控的计算机存储系统提出了更高的要求[1]。

当前已有的一些解决方案采用 SAN 作为视频监控应用的存储系统[2-3],但其成本较高,在系统普及和规模上升时遇到成本问题。

针对上述问题提出了一种采用普通 SATA 硬盘实现的高码率监控视频存储系统——THNVR(Tsinghua Network Video Recorder)。

2 THNVR 设计与实现

通过对视频监控应用特点和传统存储方案的分析可以发现,制约存储性能的关键原因有以下两点:

(1)在视频监控的应用场景中,同时会有多路并发的监控视频需要记录。而根据视频监控应用本身的特点,各路视频的码率和记录时间不同,由此而产生文件的大小也将各不相同。如果采用传统的技术,依靠非定长文件系统来组织这些随机长度的文件,势必在经历几轮视频文件的创建和删除等操作之后,产生大量由无序操作导致的磁盘碎片,造成磁盘在读写文件的过程中需要频繁地移动磁头。这样磁盘的访问性能会随着

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60273006)。
作者简介:邬建元(1984-),硕士研究生,主要研究方向:存储系统;顾瑜(1981-),博士研究生,主要研究方向:海量存储、存储安全;鞠大鹏(1967-),副研究员,主要研究方向:并行处理与分布计算机系统、存储系统等;汪东升(1966-),教授,博士生导师,主要研究方向:计算机体系结构、多核与片上系统等。
收稿日期:2009-05-05 **修回日期**:2009-10-08

时间而明显降低[4-6]，无法继续满足大数据量视频监控应用的需求。

(2)随着系统采集摄像头数量的不断增加，数据量显著上升，传统的视频记录技术受到性能的制约无法满足应用规模的增长，无法支持更多的视频数据路数。

对于这两个关键问题，提出了一种新的解决方案：分布式大规模监控视频存储系统 THNVR。

2.1　THNVR 系统结构

THNVR 系统采用分布式架构的设计思想，由一个控制节点带动很多个 THNVR 节点。控制中心通过网络向每个 THNVR 节点发送命令，控制它们的运行；每个 THNVR 节点通过网络连接相应的多个摄像头获取视频数据并在该节点存储。系统结构设计如图 1 所示。

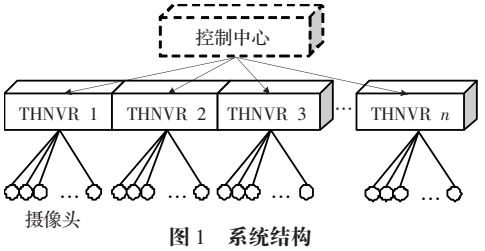

图 1　系统结构

THNVR 系统的模块化设计如图 2 所示。

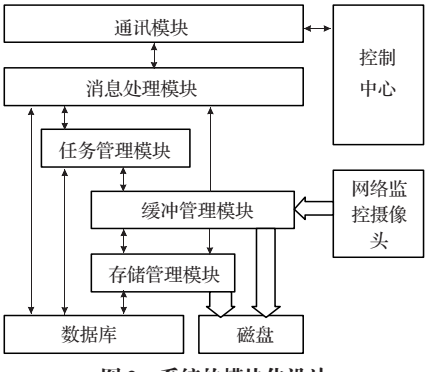

图 2　系统的模块化设计

THNVR 依照高聚合、低耦合的面向对象设计思想，采用分层化的程序结构，控制中心或其他 THNVR 节点通过通讯和消息处理模块向该 THNVR 节点下达录像任务；任务管理模块管理和调度各路录像任务的运行；缓冲管理模块从网络摄像头获取视频数据并对数据进行缓存和排队写入；存储管理模块管理系统存储空间，负责初始化、分配和回收录像文件；数据库模块存储录像描述信息等元数据，以及记录当前运行状态以实现错误现场恢复，保障系统可靠性。

2.2　关键设计

2.2.1　数据存储及检索

为了彻底克服传统视频存储方式产生的磁盘碎片问题，提高存储的性能，便于管理，THNVR 采用了将结构化和非结构化数据分别存储和检索的思想设计存储模块。

在非结构化视频数据存储的实现上，THNVR 引入了一种在数据库实现中被采用的方法[6]：为磁盘进行应用层的格式化操作，创建定长文件。

与此同时，为了便于管理定长文件系统中存储的数据以及描述视频文件的元数据，THNVR 采用了轻量级数据库 SQLite[7] 来维护一张存储这些结构化数据的表。在数据表中记录每个定长文件的分配信息、优先级信息、起止时间戳、视频信息等，如图 3 所示。

文件 1	已分配	优先级	时间戳	信息
文件 2	未分配			
文件 3	已分配	优先级	时间戳	信息
…	…	…	…	…
文件	未分配			

图 3　元数据表

使用结构化数据与非结构化数据的分别存储及检索的实现方式，可以带来如下的优势：结构化数据、非结构化数据均易于扩展；当进行多字段的元数据检索时，数据库可以很高效而便捷地实现。

2.2.2　磁盘空间管理

(1)磁盘初始化

在一个空的磁盘分区被挂载到系统中准备服务于视频监控存储应用的时候，系统在这个空磁盘分区中预先静态地分配等长度的连续大文件，充满整个磁盘分区，从而在根源上避免了磁盘碎片。

(2)空间分配

在系统运行的过程中每次有一路视频监控需要进行记录的时候，系统从数据表中为其分配一个处于空闲状态的文件；每当写满一个文件后，若录像记录仍需要继续，则再为其分配另一个空闲状态的文件，直至该路录像记录写入完毕。

在系统记录视频数据时，因为不再进行文件的动态分配与删除操作，所以不会产生任何文件碎片，磁盘的性能不会随着使用时间变长而降低。

系统每次记录视频数据都是在对磁盘分区上的已有文件进行“顺序写”操作，克服了磁头的频繁移动，所以磁盘访问的效率会大大提高[6]。每次视频记录所占用的磁盘空间都是文件定长的整数倍，每个视频任务至多浪费不超过一个定长文件的储存空间。

由于 THNVR 系统采用了将结构化的元数据与非结构化的视频数据分别存储和检索的技术，在数据的存储方面，系统可以快速地从数据库中获取一个对应空闲定长文件的记录，分配给上层使用。在数据的检索方面，系统可以快速地从数据库中获取一个或多个符合条件的定长文件的记录，提供给上层使用。

(3)空间回收

针对目前的视频监控应用，由于大量的视频监控数据的时效性较强，旧有的数据经过一段时间(如 30 天或 90 天)后，其重要程度会逐渐降低。为了更有效地利用存储资源，进一步降低系统的成本，THNVR 提出了一种高效的回收过期视频监控数据技术，实现在系统正常运行视频监控任务的同时回收过期数据的功能。

得益于 THNVR 采用的结构化数据与非结构化数据的分别存取设计，过期文件的回收操作仅需要对储存在数据库中的元数据进行，不需要实际删除视频数据文件。这样的实现方式大大简化了空间回收操作，减小了对系统性能的影响。

在具体实现上，THNVR 对视频监控产生的文件按优先级分级处理，并由用户配置视频文件的生存周期参数。当空闲的

定长文件少于配置的下限时，系统对数据库中过期的文件记录，按优先级由低向高的次序进行回收。

2.2.3 磁盘访问优化

仅依靠上述避免磁盘碎片的方法仍然不足以解决大量视频路数的需求问题。在同时运行的监控视频路数较高的情形下，造成磁盘存取性能降低的另一个原因是不同的视频存储任务需要在磁盘上写的文件位置是不同的。如果多个任务并发地写磁盘，就会使得磁头的移动变得频繁，从而表现为磁盘效率降低。另外，由于多路视频流是并发执行的，若不对它们进行有效的调度，而允许它们无序地并发写入的话，前面所提出的各种方法都将无法有效而充分地发挥它们的优化效果。

为了解决这个问题，THNVR 提出了一种缓冲存储的方案，即为每一路视频的数据分配较大的缓冲区，当缓冲满时一并写入磁盘，即“大写”。

由于定长文件系统中，文件的逻辑编号与其在磁盘上的物理位置呈单调递增关系，所以通过对缓冲区中等待写入磁盘的数据按照文件编号进行排序，就可以使它们单线程、有序地访问磁盘。这样做就保证了对磁盘的访问是顺序的，并且是连续大量的，从而更有效地利用了磁盘写入带宽，获得了磁盘的高效访问特性，能够比传统的设计方案提供更高的磁盘访问性能，支持更高的视频码率及更多的并发视频数量。同时，缓冲管理模块根据系统内存资源自适应地调整分配给每路视频的缓冲区数量，从而控制内存使用量，在内存资源不足时仍可以保障高优先级视频数据的正常记录。

2.2.4 可扩展性

(1)节点扩容

由于视频监控本身的应用特点，产生的数据量十分巨大。虽然系统提供了空间回收策略以便于有效利用现有存储空间，但当磁盘空间不能满足新的需求时，会遇到追加磁盘空间的需求。THNVR 在系统的设计中考虑到了这一点，提供了扩充储存空间的功能。

在实现上，追加磁盘空间与初始化文件系统类似。用户通过 LVM 等方式为录像分区追加容量后，系统在扩充后的剩余空间中创建新的定长文件，并向数据库中追加相应的记录，在此之后系统即可继续正常工作。

(2)系统扩容

完整的系统中，由一个控制中心来组织和管理各个 THNVR 录像服务器。控制中心与 THNVR 节点之间仅传输控制命令等消息，实际的视频数据由 THNVR 节点直接从网络摄像头获取。在这样的结构下，扩展系统的视频路数可以通过相对简便地增加新的 THNVR 节点并向控制中心报告的方式来完成。一个控制节点可以带动大量的 THNVR 节点，从而达到超强的系统扩展性，支撑超多数量的摄像头。

2.2.5 容错与恢复技术

作为一套服务于实际应用的系统，可靠性是设计中非常重要的一个方面。THNVR 系统的可靠性主要体现在以下几个方面：

(1)容错

系统的某些模块工作不正常的情况下，整个系统仍能够稳定运行。

在某一路正在记录的视频编码器出现故障时，系统将停止该路数据的存储，保证其他路视频继续正常记录；在 CPU 占用率、内存占用率、磁盘性能由于其他进程或意外原因而持续恶化，以至于不足以保障系统对所有数据的记录带宽时，系统将停止一部分优先级较低的视频监控任务，释放其占用的系统资源，从而缓解系统资源的紧张状况，保障剩余的高优先级视频监控任务的继续运行，同时向控制单元发出警报消息；在依照空间回收策略的前提下，若空闲的空间仍出现不足以接受新的视频数据的情况，系统将按照优先级由低向高的顺序停止正在运行的视频监控任务，并向控制单元发出警报消息。

THNVR 的设计考虑了由误操作等原因而导致的定长文件损坏问题，在为视频数据分配定长文件时会检查该文件的可用性，避免使用无效文件。

(2)错误恢复

系统遇到意外断网的情况时，会造成两类问题，一是无法与控制单元进行命令和状态等通讯；二是无法继续获取来自网络摄像头的视频数据。

当第一类故障发生时，系统将工作在自治模式下，继续运行已经启动的各路视频监控记录任务。与此同时，将需要向控制单元发送的状态、警报等消息在数据库中进行持久化，在系统与控制单元的网络连接恢复时重新发送。当第二类故障发生时，系统无法接收到一路或多路视频的新数据，等待通讯恢复正常，或控制单元发送来新的命令。

对于意外断电的情况，在 THNVR 的设计中采用了将正在执行的任务信息在数据库中进行持久化的方式来解决。当断电造成内存中的任务信息丢失时，系统重新启动后，将从数据库中读取并恢复断电之前的所有任务信息，快速恢复现场，并继续运行那些仍处在执行周期内的任务，将由电源故障而导致的视频监控服务的中断时间降到最低。同时 THNVR 在启动时对维护视频源数据的数据库进行一致性检查，回收因意外断电而产生的无效数据文件，保障数据的正确性。

(3)状态监控和日志管理

为了进一步提供更好的系统可靠性，THNVR 的设计中提供了向控制单元报告系统的 CPU 占用率、内存占用率、剩余存储空间等数据的功能，便于系统管理员更好地观测系统运行状态和及时采取维护措施。另一方面，系统在运行过程中，分级别地记录日志，方便管理员根据紧迫程度查看系统运行历史。

3 系统测试

为了分析 THNVR 系统设计中采用的各项关键技术对性能产生的实际效果，采用了如下设备环境进行系统测试：

CPU：Intel PIV Xeon 2.8 GHz；内存：2GB DDR；磁盘：Seagate SATA II 500GB；Ubuntu Linux Server 8.04，Sqlite3。

3.1 缓冲大小测试

通过改变为每路视频数据分配的缓冲区大小的方法来测试和分析缓冲技术对磁盘写入性能的影响，结果数据如图 4 所示，图中包含了系统在同时存储 8 路视频时，采用不同的定长文件大小以及不同的缓冲区大小得到的磁盘写入带宽。

测试数据反映出，当缓冲区较小(4~32 kB)时，磁盘的写入带宽远远低于其最大带宽。随着每路视频数据的缓冲区大小的增加，磁盘的实际写入带宽显著上升，直到逼近其最大写入带宽。

为每路视频数据分配较大的写缓冲区，可以减少磁盘访问请求数量。通过缓冲、排序、大写技术，可以平滑磁盘写操作，减少磁头移动，从而带来更高的磁盘写性能。

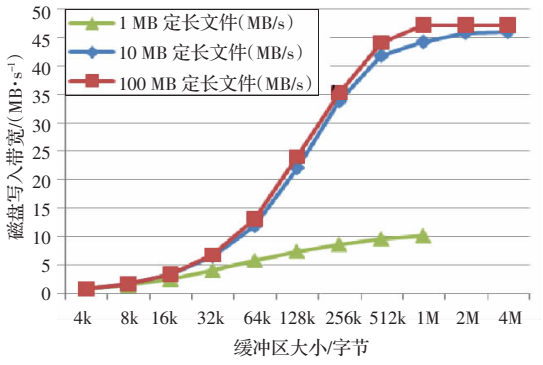

图 4　定长文件大小及缓冲区大小对磁盘写入带宽的作用(8 路并发视频)

当然,增大缓冲需要消耗更多系统内存,同时会增大写入延迟,这在一定程度上影响了数据写入的实时性。

3.2　文件大小测试

对于定长文件系统的参数对磁盘写入性能的影响,使用不同的定长文件大小来测试,观测到的系统实际写入性能数据如图 4 所示。

通过实验数据可以清晰地看出,采用较小的定长文件时,系统的写入带宽比较明显地低于其他两组数据。以较大的定长文件(大于或等于 10 MB),可以在系统运行的过程中较少地切换文件,减小额外开销,从而得到更高的性能。但采用较大的定长文件会带来较多的空间损失。在每次录像的时间较短,产生的视频数据较少的应用场合,可以选择适中的定长文件大小(10 MB 量级),以获得性能和磁盘空间利用率的折中。

3.3　视频并发数测试

为了进一步观察系统在大负载下的磁盘写入性能,将视频数据路数增加到 256 路,测试数据如图 5 所示,图中包含了系统在采用不同的定长文件大小以及不同的缓冲区大小得到的磁盘写入带宽。

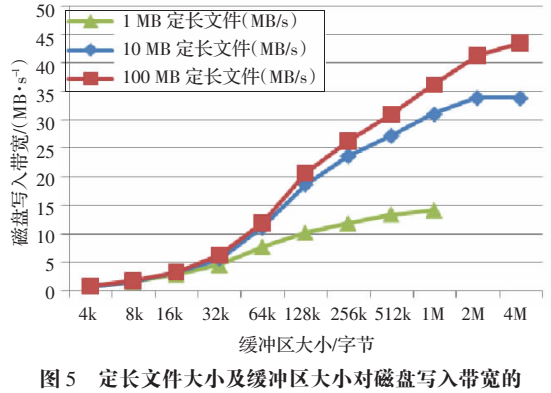

图 5　定长文件大小及缓冲区大小对磁盘写入带宽的作用(256 路并发视频)

测试数据反映出,在提高视频路数的情况下,磁盘的实际写入带宽降低,这是因为系统同时打开的文件数量也随之线性增加,缓冲区中的数据需要交替写入到更多定长文件中,磁头的移动范围和频率相应地增加。同时,测试数据还反映出,这种现象可以通过加大缓冲区大小以及定长文件大小的方式来缓解,这样能够减少磁盘访问请求数量和切换定长文件所带来的开销,逼近磁盘最大写入带宽。

3.4　整体性能测试

上述的测试数据反映了理想情况下 THNVR 的存储模块在不同参数下的满负荷磁盘写入性能。完整的系统由于包含了容错、空间回收等功能模块,它们带来了一些额外的开销。为了获得系统实际服务性能的数据,该文采用 100 MB 定长文件初始化文件系统,以每路码率为 1~2 Mb/s 的 D1 视频数据流对实际系统进行了性能测试。系统在不同的缓冲区大小下实际可以服务的最大视频路数如图 6 所示。当缓冲区大小提高到 2 MB 时,系统可以在普通的 SATA 硬盘上同时记录 256 路 D1 视频数据流。经过测试,这种配置下 THNVR 也能够同时记录 512 路 CIF 视频流。

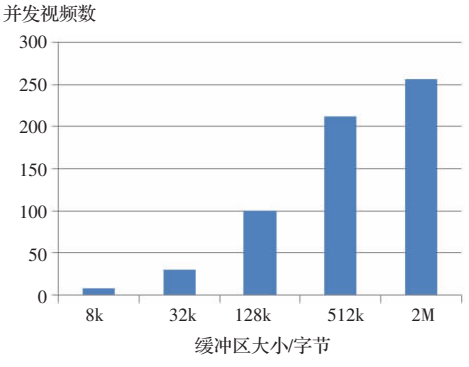

图 6　系统实际服务性能

4　结论与展望

THNVR 系统充分考虑了网络视频监控应用对于存储系统的特殊需求,在普通 SATA 磁盘上实现了同时记录 256 路 D1 或 512 路 CIF 视频数据,让有限的系统资源得到了充分的利用,以较低的成本实现对性能有较高要求的服务。

THNVR 设计中考虑到了空间回收、可靠性、可扩展性等问题,并提出了有针对性的策略和解决方案,可以更好地服务于实际应用,优于现有的其他同类系统。

对于不同的系统参数进行了测试,获得了详细的数据,为实际系统中参数的选择提供了有效的参考和指导。

随着存储技术的快速发展,存储介质的成本将继续下降,性能将继续提高。系统在未来将考虑在成本容许的范围内使用更高效和可靠的存储设备,例如 SSD 等设备,为系统提供更高码率和更多并发视频数量的支持。另一方面,系统将会采用更智能化的管理策略,更有效地利用分布式系统整体资源。

参考文献:

[1] 骆云志,刘治红.视频监控技术发展综述[J].兵工自动化,2009(1):1-11.

[2] 雷凌.视频监控系统中最佳存储方案的实践[J].中国安防,2009(1):122-125.

[3] 中星电子.视频监控核心技术的自主知识产权研究[J].安防科技,2008(12):13-36.

[4] Du David,He Dingshan,Hong Changjin,et al.Experiences in building an object-based storage System based on the OSD T-10 standard[C]//14th NASA Goddard & 23rd IEEE(MSST2006) Conference on Mass Storage Systems and Technologies May 15-18,2006.

[5] Rosenblum M,Ousterhout J K.The design and implementation of a log-structured file system[J].ACM Transactions on Computer Systems (TOCS),1992,10(1).

[6] Vengurlekar N.Oracle disk manager[Z].2002-09.

[7] Owens M.The definitive guide to SQLite[Z].2006-05.