软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(8):1999−2009 [doi: 10.13328/j.cnki.jos.005203]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向海量高清视频数据的高性能分布式存储系统*

操顺德[1,2], 华 宇[1,2], 冯 丹[1,2], 孙园园[1,2], 左鹏飞[1,2]

[1](武汉光电国家实验室(华中科技大学),湖北 武汉 430074)

[2](华中科技大学 计算机科学与技术学院,湖北 武汉 430074)

通讯作者: 华宇, E-mail: csyhua@hust.edu.cn

摘 要: 通过对视频监控数据的特点和传统存储方案进行分析,提出一种高性能分布式存储系统解决方案.不同于传统的基于文件存储的方式,设计了一种逻辑卷结构,将非结构化的视频流数据以此结构进行组织并直接写入RAW 磁盘设备,解决了传统存储方案中随机磁盘读写和磁盘碎片导致存储性能下降的问题.该方案将元数据组织为两级索引结构,分别由状态管理器和存储服务器管理,极大地减少了状态管理器需要管理元数据的数量,消除了性能瓶颈,并提供了精确到秒级的检索精度.此外,该方案灵活的存储服务器分组策略和组内互备关系使得存储系统具备容错能力和线性扩展能力.系统测试结果表明,该方案在成本低廉的 PC 服务器上实现了单台服务器可同时记录400 路 1080P 视频流,写入速度是本地文件系统的 2.5 倍.

关键词: 海量存储;视频存储;流媒体存储;分布式系统;逻辑卷

中图法分类号: TP316

中文引用格式: 操顺德,华宇,冯丹,孙园园,左鹏飞.面向海量高清视频数据的高性能分布式存储系统.软件学报,2017,28(8):1999−2009. http://www.jos.org.cn/1000-9825/5203.htm

英文引用格式: Cao SD, Hua Y, Feng D, Sun YY, Zuo PF. High-Performance distributed storage system for large-scale high-definition video data. Ruan Jian Xue Bao/Journal of Software, 2017,28(8):1999−2009 (in Chinese). http://www.jos.org.cn/1000-9825/5203.htm

High-Performance Distributed Storage System for Large-Scale High-Definition Video Data

CAO Shun-De[1,2], HUA Yu[1,2], FENG Dan[1,2], SUN Yuan-Yuan[1,2], ZUO Peng-Fei[1,2]

[1](Wuhan National Laboratory for Optoelectronics (Huazhong University of Science and Technology), Wuhan 430074, China)

[2](School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China)

Abstract: This paper presents a high-performance distributed storage solution for video surveillance data via analyzing the characteristics of video surveillance data and conventional file storage solutions. This design proposes a logical volume structure rather than file-based storage to efficiently organize unstructured video stream data. The scheme hence directly writes these stream data into raw disk devices, to address the problem of storage performance decrease caused by the random access and disk fragmentation in traditional storage systems. A two-stage index strategy is also implemented to manage metadata by the state manager and storage servers, which significantly reduces the amount of metadata managed by the state manager, eliminates the performance bottlenecks, and provides the second-level video retrieval accuracy. Moreover, the design has the salient features of fault tolerance and linear scaling abilities with the help of the flexible storage server grouping policy and the mutual backup relationship in a storage group. Experimental results show that the solution can simultaneously record 400 ways of 1080P video streams with a single low-cost PC server, and the system's average write speed is 2.5 times faster compared with the local file systems.

Key words: massive storage; video storage; stream media storage; distributed system; logical volume

* 基金项目: 国家重点研发计划(2016YFB1000202)
Foundation item: National Key Research and Development Program of China (2016YFB1000202)
收稿时间: 2016-08-11; 修改时间: 2016-09-21, 2016-11-11; 采用时间: 2016-12-01; jos 在线出版时间: 2017-01-12
CNKI 网络优先出版: 2017-01-12 10:36:01, http://www.cnki.net/kcms/detail/11.2560.TP.20170112.1036.005.html

随着城市项目的推进,城市视频监控系统的规模不断扩大,视频监控系统朝着“数字化、网络化、高清化、智能化”的方向不断前进.不断增长的监控点设备数量、更高清的摄像头和更长的视频留存时间,对视频监控数据存储系统提出了更高的要求.

传统的基于 IP-SAN(Internet protocol storage area network(IP 存储网络))存储方案通过 iSCSI(Internet small computer system interface(Internet 小型计算机系统接口))协议将磁盘阵列挂载给流媒体服务器做标准磁盘,采用文件方式进行视频数据的集中存储.在这种模式下,数据流要经过流媒体转存服务器才能写入磁盘阵列,存在着单点故障和性能瓶颈等问题.视频监控系统通常要求 7×24 小时持续运行,面对多路视频流的并发持续写入,基于通用文件系统构建的存储系统将会产生大量的文件碎片,导致随着时间的推移,存储效率不断下降.因此,传统模式不能满足视频监控系统规模上升时对存储系统的高可靠性、高性能、可扩展性和易管理等方面的需求.随着技术的发展,分布式云存储[1]变得越来越重要,一些中等规模的视频监控系统采用了分布式云存储方案.

视频监控的数据和传统数据,如邮件、文件、图片、数据表等相比,有以下特点.

(1) 数据量大.视频监控朝着高清化方向发展,随着摄像头数量的不断增加,一个中等规模的城市一天就能产生 PB 级数据量.监控视频的保存周期为 30 天~90 天,视频监控的存储系统必须满足视频数据的长时间大容量存储需求,且要具有线性扩展的能力.

(2) 写密集.传统数据的读写符合二八定律,即 20%的时间写数据,80%的时间读数据.而视频监控数据完全相反,写操作几乎占到了 100%,只有在回放和检索视频时才读取视频数据.视频监控的存储系统在设计和实现上需要优先考虑如何提高写入带宽.

(3) 码率恒定.传统数据的码率比较随机,而视频监控数据的码率比较恒定,因为监控视频的分辨率和格式一般不会发生变化.

(4) 7×24 小时持续服务.传统数据一般遵循访问周期,可在访问量很少的情况下将系统停机下线进行维护和升级而不影响正常的商业活动.视频监控数据不存在这样的周期,视频数据流持续不间断地涌向存储系统,任何升级维护操作都不能影响输入视频流的存储,必须动态地进行.

本文针对上述问题和视频监控数据的特点,提出一种面向海量高清视频监控数据的高性能分布式存储系统——DVSS(distributed video surveillance storage).本文的主要贡献如下.

(1) 首先,设计了一种基于 RAW 磁盘设备的逻辑卷结构来组织并存储非结构化视频数据,将大量并发数据流随机写转化成大段的连续写,极大地提升了磁盘写入带宽,支撑大规模高清视频流的并发写入.

(2) 其次,设计了两级索引结构来管理视频元数据.两级索引结构设计极大地减少了状态管理器需要管理的元数据数量,消除了性能瓶颈,且可实现录像段的秒级检索.

(3) 最后,实现了 DVSS 系统原型,并对磁盘写入速度和支持的并发视频路数等关键指标进行测试,证明了 DVSS 系统的高效的存储性能.测试结果表明,DVSS 系统在单台廉价 PC 服务器上能够同时记录 400 路 1080P 视频流数据,写入速度达本地文件系统的 2.5 倍.

1 相关工作

针对传统文件系统存储方案的问题,研究人员根据视频监控数据的特点,提出了直写裸磁盘设备的流式存储策略,即设计一种磁盘逻辑结构,将录像段的数据和索引信息以此结构组织并直接写入裸磁盘设备中[2−4].文献[2]采用固定大小的数据区域存放每一视频帧,提供帧级的检索精度.每一视频帧数据大小不一,采用固定大小区域存储会造成内部碎片,且只采用帧内压缩而没有考虑相邻帧之间的冗余信息,存储空间利用率低下.文献[3]以图像群组(group of pictures,简称 GOP)为单位组织存储,消除了相邻帧之间的冗余信息,在文献[2]的基础上提升了一定的空间利用率.文献[5]提出一种面向连续数据存储的高效能盘阵 Ripple-RAID,该方案采用新的局部并行数据布局,运用地址转换和异地更新等技术带来了写性能的提升和节能效率.文献[6]提出一种分布式大规模监控视频存储系统 THNVR.THNVR 系统采用了将结构化和非结构化数据分别存储和检索的设计思想,用定长文件存储非结构化视频数据,并将定长视频文件的元数据保存在 SQLite 中,从根本上避免了磁盘碎片,提高了

存储的性能.THNVR 虽然避免了外部碎片,但仍存在内部碎片问题.

文献[7]提出一种基于 Hadoop 分布式文件系统(HDFS)的云视频记录系统框架,使用 HDFS 来存储监控视频数据,并提出将来可以用 Map/Reduce 机制执行视频分析任务.然而,用 HDFS 存储视频无法提供细粒度的视频检索.为了解决大规模视频监控系统存在的高并发、高比特率写瓶颈问题,文献[8]提出了一种面向大规模并发高清视频流的高性能分布式文件系统 DSFS.DSFS 设计了一种连续存储模型(continuous storage model,简称 CSM)将对磁盘的随机高并发写转化成顺序写,提高了磁盘的写速度.但 DSFS 元数据全部由 MetaServer 管理,存在性能瓶颈和单点故障,且同一路视频流数据存储分散,读操作需要频繁移动磁头,视频回放性能不佳.

2 DVSS 设计与实现

视频监控数据具有高并发、大容量、有序的特点.通用文件系统如 NTFS 和 Ext4 等并非针对视频监控数据而设计,传统存储方案以文件来组织并存储视频流数据,存在以下两个问题.

(1) 以文件方式存储视频流数据不能保证数据在磁盘上的存放是连续的,同时,在记录多路高清视频流时,系统需要维护大量打开的文件描述符,并在这些文件描述符之间来回切换.随着系统长时间运行,文件被频繁地创建和删除,势必产生大量的磁盘碎片,使得在写连续的视频流时磁头需要频繁地移动,造成磁盘访问性能下降.

(2) 文件分片的长度决定了视频检索的精度:分片太小会产生海量的小文件,造成 inode 号不够用;分片太大,造成检索精度不高.

2.1 DVSS系统结构

DVSS 系统由一个状态管理器(state manager)、多个存储服务器(storage server)和多个客户端(client)组成.系统结构如图 1 所示.

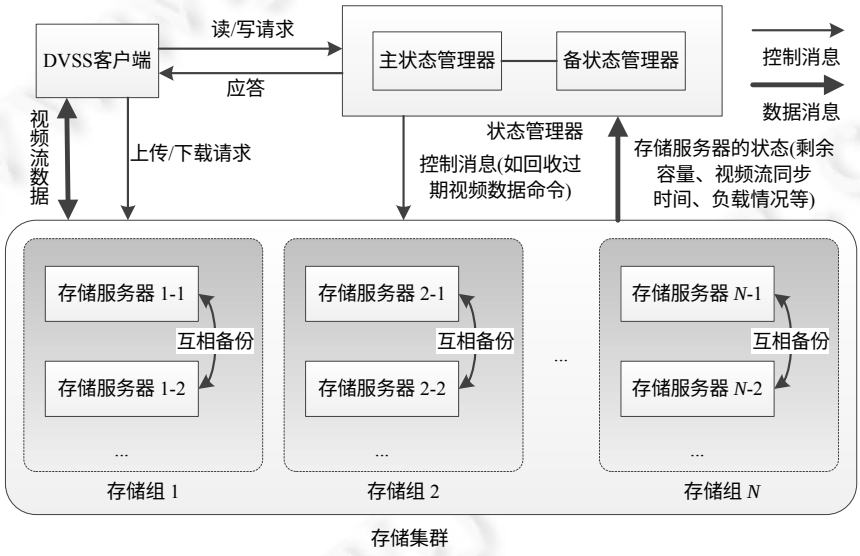

Fig.1 Architecture of DVSS

图 1 DVSS 系统结构

状态管理器主要做调度工作,起负载均衡的作用.状态管理器除了需要在内存中记录集群中所有存储组和存储服务器的状态信息以外,还需要记录录像段的起止时间和对应的存储组.存储服务器采用分组的组织方式,一个存储组由 1 台或多台存储服务器组成,同组内的存储服务器互相备份,一个组的存储容量由该组内存储容量最小的存储服务器决定,存储集群总容量为集群中所有组的存储容量之和.HDFS 等主流的副本实现方式通

常采用动态分配的方式,一个文件实际存储节点位置是不确定的,通常是 3 个备份.DVSS 系统采用的分组存储方式更加灵活,可操作性更强.例如,对于重点监控点的视频数据,可以由管理员指定存储分组.当一个分组的存储服务器访问压力较大时,可以在该组添加存储服务器来提升服务能力;当视频监控规模扩大时,可以增加存储组来扩充存储容量,实现系统线性扩展.存储服务器保存视频流数据和其索引信息.客户端拥有全局唯一识别号(命名为 SID),通常为网络摄像头的 IP 地址.

2.2 视频数据组织

为了提高存储性能,方便对视频进行检索和管理,DVSS 采用了将视频流和其元数据分开存储的设计思想.元数据和视频数据存储在不同的磁盘驱动器上,保证元数据的读写请求不会影响对视频流数据的读写,同时方便在视频录像段过期后对其进行空间回收.

2.2.1 视频流数据管理

考虑到视频流编码格式的特点,DVSS 以 GOP 为单位组织数据并建立索引,提供精确到秒级的高效检索.为了便于视频流数据的管理,定义如图 2 所示的磁盘逻辑存储结构.

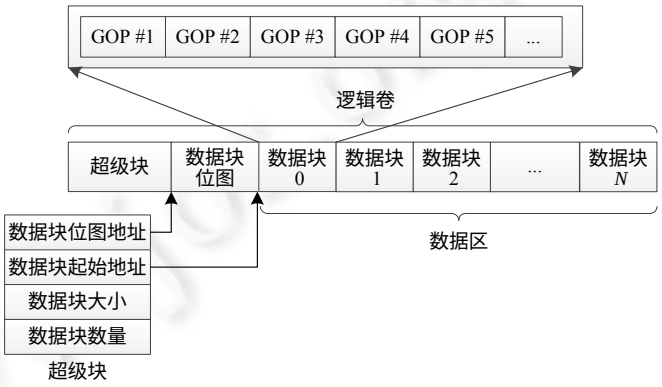

Fig.2 Logical volume architecture design

图 2 逻辑卷结构设计

(1) 逻辑卷

逻辑卷将多个磁盘或磁盘分区在逻辑上聚合,提供大小一致的存储空间,无需担心不同磁盘分区大小不一致的情况,方便实现存储空间动态扩容.

用 LVM(logical volume manager(逻辑卷管理器))创建逻辑卷之后,按如图 2 所示的结构格式化逻辑卷.在逻辑卷的头部设计有“超级块(super block)”结构,包含数据块位图的起始地址、数据块大小、数据块起始地址等信息,用来描述逻辑卷各项参数信息.数据块位图描述数据块的使用情况,0 表示未分配,1 表示已分配.逻辑卷剩下的空间都为数据区,由固定长度的数据块组成,用于存储视频流数据.

(2) 数据块

数据块是存储空间分配和回收管理的基本单位.系统每次分配一个数据块给某一视频流,并将对应的位图 bit 表示置 1.一个数据块上存储的是同一路视频流的数据,方便进行顺序读写,提高读写带宽.空间回收时,因为数据块里的数据是按时有序的,只要此数据块存放的最后一个图像数据过期就回收此数据块.数据块的大小一般设置为 512MB,具体的讨论见第 3.2 节.

(3) 图像组 GOP

当前广泛采用的视频编解码标准 H.264/AVC 将 I 帧、P 帧、B 帧组成图像组来编码图像序列,一个 GOP 通常存放 1s 的视频数据.

DVSS 利用视频数据流的编码特征,以 GOP 为单位读写数据,同一个视频流的 GOP 在数据块中连续存放,直到数据块剩下的空间不足以存放一个 GOP 为止.

2.2.2 元数据管理

为了方便实现负载均衡和精确到秒级的检索,DVSS 设计了两级索引结构,并使用基于内存的键值对数据库 Redis 作为元数据管理引擎.

(1) 一级索引

一级索引存放在状态管理器上,用来存储录像段索引.如图 3 所示,录像段记录为一级索引的基本单元,表示一段连续视频的信息,包括该录像段的起始时间、时长以及对应的存储服务器组 ID 等.视频流标识(SID)为 *Key* 值,按起始时间排序的录像段记录列表作为对应的 *Value* 值.由于监控视频流本身按时有序,所以录像段记录的插入也是有序的,可以实现 *O*(1)的插入时间复杂度.

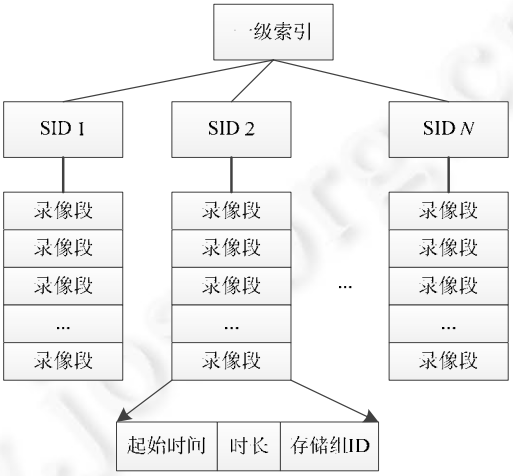

Fig.3 First-Level index structure

图 3 一级索引结构

录像段是从客户端发起视频写请求开始,到客户端主动停止写入或者因为存储服务器空间不足拒绝继续写入为止的一段连续视频.

(2) 二级索引

二级索引存放在存储服务器上,结构上类似于一级索引,其基本索引单元为 GOP,用于对视频流的 GOP 进行描述.一个 GOP 的描述信息包括该 GOP 对应的起始时间戳、存放的逻辑卷号、对应的数据块、块内偏移和长度等信息.视频流标识为 *Key* 值,按时间戳有序的 GOP 描述信息列表作为对应的 *Value* 值.同样,GOP 描述信息的插入也是有序的,可以实现 *O*(1)的插入时间复杂度.

2.3 视频流上传和检索

2.3.1 视频流上传

视频流上传操作由客户端主动发起,客户端在上传固定时长的视频流后结束写操作,再发起新的上传操作.通常,一个视频段的时长为 1 个小时.视频段时长太短,会增大状态管理器管理的元数据的负担;时长太长,则不利于系统的负载均衡和动态扩容.

上传步骤如下.

(1) 客户端向状态管理器发送上传请求,包括视频流标识、视频段起始时间、预计时长和码流大小.

(2) 状态管理器根据负载均衡策略查询可用的存储服务器,在一级索引里增加一条视频段记录,将分配的存储服务器的 IP 地址和端口号返回给客户端.

(3) 客户端向分配的存储服务器发起写操作,包括视频流标识、起始时间和视频流数据.

(4) 存储服务器的视频流编解码模块解析出 GOP,按需给此视频流分配数据块,将 GOP 连续地写入数据

块,并在二级索引里增加对应的 GOP 描述信息记录.

(5) 在客户端完成此视频段的上传后,存储服务器返回状态信息.

2.3.2 视频流检索

监控视频回放以视频流检索为基础,当用户需要查看某一段视频时,客户端发起视频流检索操作.精确到秒级的检索精度,减少了不必要数据的传输,极大地提高了视频回放的效率.

检索步骤如下.

(1) 客户端向状态管理器发送检索请求,包括视频流标识、视频段起始时间和回放时长.

(2) 状态管理器以视频流标识为 *Key* 在键值对数据库中查询对应的视频段信息列表,在有序的视频段信息列表中定位对应的起始时间和时长的视频段,得到此视频段的存储服务器组编号.状态管理器再根据存储服务器组内的负载情况和视频数据同步情况选择一个可用的存储服务器,并返回其 IP 地址和端口.

(3) 客户端向定位的存储服务器发起数据读请求,包括视频流标识,起始时间和时长.

(4) 存储服务器根据客户端的读请求,通过二级索引找到视频段的视频数据存储所在的数据块区间,以 GOP 为单位连续读出视频数据.

(5) 存储服务器读取的视频段数据返回给客户端.

2.4 高并发写优化

DVSS 支持大规模视频流的高并发写操作.状态管理器会根据存储服务器的负载情况,将写请求分配给合适的存储服务器.对于单个存储服务器,在同时写入的视频路数较高的情况下,因不同的视频流写的数据块不同,多视频流高并发的写入会导致磁头的频繁移动,使得磁盘存取性能降低.

为了解决这个问题,存储服务器采用单线程处理多路视频流的并发写入,并为每个视频流分配一个缓冲区,在缓冲区满时将数据写入到对应的数据块.同时,根据数据块的分配特性,多个视频流的数据块编号呈递增关系,对应地,在磁盘上的物理位置也呈递增关系,故可对待写入的数据按数据块编号排队.这种机制将无序的小段数据高并发写入转变成了有序的大段连续数据写入,在增加磁头有效连续写入时间的同时,减少了磁头的频繁移动,提高了磁盘存取性能,可以支持更多的并发视频路数.

2.5 存储空间管理

2.5.1 空间回收

监控视频的保存时间从 1 个月到 3 个月不等,除了需要永久保存的录像段外,其他视频数据随着时间的推移,其重要性逐步降低.由于磁盘存储空间有限,为了充分利用存储资源,需要对达到存储周期的视频数据进行空间回收.

DVSS 采用循环覆盖的回收策略.空间回收由状态管理器控制.状态管理器根据用户设定的存储周期,周期性地删除一级索引里的过期的录像段记录信息,并通知相应的存储服务器删除对应的二级索引.存储服务器删除过期的二级索引信息,并将对应的数据块的位图表示置 0,无需擦除数据块里的实际视频数据,回收效率较高.

得益于数据块的分配策略,连续数据块存放的视频流数据在时间上有序,回收的空间也是连续的.当回收的空间再次分配给数据流时,并发写入的数据流分得的数据块也是连续的,可以减少磁头的移动.相邻的数据块总是被连续分配,又被连续回收,使存储系统不会因为长时间运转而导致存取性能下降.

2.5.2 扩 容

为了应对摄像头路数增加以及摄像头分辨率提高的挑战,需要及时地对 DVSS 进行扩容.

(1) 组内扩容

同组内的多台存储服务器互为备份,类似于木桶短板效应,一个组的存储容量为该组服务器中最小的存储容量.扩充组内容量时,需要给组内的每台服务器增加相同的磁盘空间.

追加磁盘空间后,用 LVM 创建新的逻辑卷,并按逻辑卷结构格式化,分配逻辑卷号,与裸设备(/dev/raw/

raw[N])绑定后,即可提供服务.

(2) 系统扩容

当系统容量不足时,可以增加存储组组来扩展系统存储集群的总容量.新增的存储服务器启动后主动向状态管理器汇报自己的状态,状态管理器对比自己内存里记录的所有存储服务器状态后就会发现此存储组是新增的,将此存储器组加入活动存储服务器队列,负载均衡机制会将新到来的上传视频请求导向给新增的存储服务器组,此新增服务器组就开始提供服务,正常工作,实现了系统的动态扩容.

2.6 高可用

DVSS 由成本低廉的 PC 服务器组成,在降低视频监控存储系统成本的同时,需要提高可靠性,保证系统 7×24 小时持续服务的能力.DVSS 的高可用主要体现在如下几个方面.

(1) 状态管理器高可用

类似于 HDFS NameNode,DVSS 的状态管理器存在单点故障.在 Hadoop 2.0 中,HDFS NameNode 单点故障问题已经解决.DVSS 参照了 HDFS HA[9]的实现,状态管理器由主状态管理器和备状态管理器构成,主备状态管理器共享存储.正常情况下,只有主状态管理器对外提供服务.当主备切换控制器检测到主状态管理器故障时,进行主备切换,由备状态管理器对外提供服务.当主状态管理器恢复后,再切换回来.状态管理器高可用架构消除了单点故障,可持续对外提供服务.

(2) 组内互备

一个视频录像段的备份个数为所在存储服务器组内的服务器个数.DVSS 分组存储的方式可以根据视频流的重要程度灵活控制视频段的副本个数.将存储服务器组按重要程度,配置不同数量的组内服务器.状态管理器根据视频流标识将重要的视频流导向到互备服务器较多的存储组.

若一台存储服务器失效,状态管理器通过和存储服务器之间周期性的心跳机制检测出存储服务器断线,状态管理器会将此存储服务器剔除服务队列直到其重新上线.由于同一组内的服务器是对等关系,同组的其他服务器可以继续提供服务,对此存储服务器的读数据请求,会被分派给同组的其他存储服务器.

(3) 断点续传

针对摄像头可能出现的网络故障问题,前端设备需配有小容量的本地存储,如 SD 卡等.当网络出现故障时,视频录像会被存储在本地存储上;在网络恢复之后,再将暂存在本地存储的未上传的视频录像补录到 DVSS 中,以保证不会因为网络故障而发生数据丢失.

同时,当 DVSS 系统通过周期性的心跳机制发现摄像头断开连接时,需通过发送报警信息等方式通知维护人员及时处理.

3 实现与评估

3.1 测试环境

DVSS 原型系统由一台状态管理器和 40 台存储服务器组成.存储服务器采用 Intel Xeon 2.4GHz×2 CPU,4GB 内存,6 块 1TB 15000RPM SATA Disks 和 Mellanox InfiniBand QDR 40Gb/s 网卡.操作系统为 Ubuntu Server 14.04LTS,键值对数据库 Redis 采用 V3.2 版本.

3.2 参数测试

数据块和缓冲区的大小将直接影响磁盘的写性能.通过改变数据块大小和为每路数据流分配的缓冲区大小来测试和分析这两个参数对磁盘写性能的影响.对单台存储服务器发起 100 路数据流并发写入,为了测试最大吞吐率,不限定数据流的比特率大小.测试结果如图 4 所示.

测试数据显示:64MB~512MB 大小的数据块,在缓冲区大小为 64KB 时,相对于 32MB 大小的数据块,写入速度有显著的提升.在数据块大小为 64MB~600MB 时,100 路数据流并发写入速度先随着缓冲区大小的增大而提升,随后下降.对测试结果的分析如下:记数据块大小为 *blocksize*,缓冲区大小为 *bufsize*.在将 *bufsize* 大小的一路

数据流写入磁盘后,需要调用 *lseek*()将写位置指向下一路数据流对应的数据块位置,偏移量 *offset* 为 *blocksize* *bufsize*.在 *offset* 比较小的情况下,上一次 *write*()调用结束到下一次 *write*()调用时,磁盘已转过要写的位置,需要多转一圈才能接着写,使得写性能下降.这解释了数据块增大时写性能提高的原因,同时,随着 *bufsize* 的继续增大,使得 *offset* 减小,写性能也会下降.数据块也不是越大越好,当数据块大小增大到 600MB 时,由此带来的寻道时间开销使得写性能显著下降.

根据以上的测试结果,DVSS 原型系统将数据块和缓冲区大小分别设置为 512MB 和 1MB.

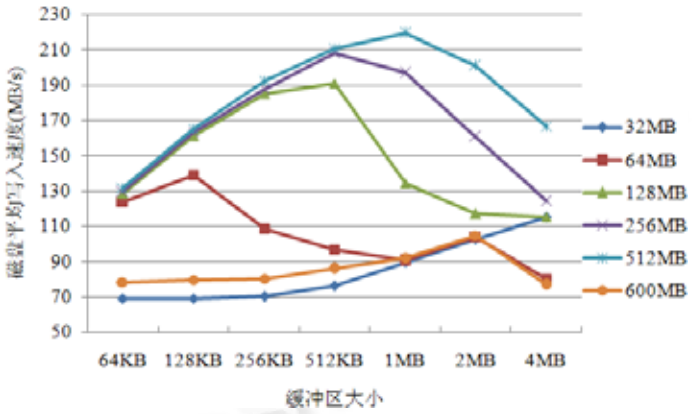

Fig.4 Effect of data block and buffer size on disk write performance

图 4 数据块和缓冲区大小对磁盘写性能的影响

3.3 单机性能测试

为了证明 DVSS 系统的视频数据组织结构在多路视频流高并发写入时的性能优势,将 50 路~500 路分辨率为 1 080P 的视频流分别并发写入 DVSS 单台存储服务器和 Ext4 文件系统中,每一路视频流的码率恒定为 4Mbit/s,测得的平均写入速度如图 5 所示.

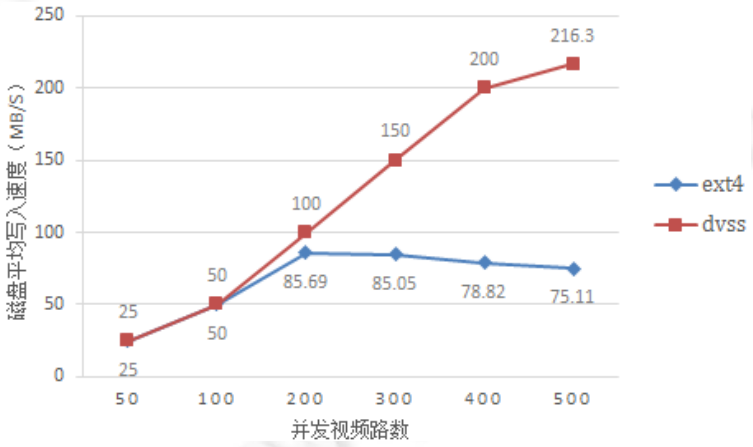

Fig.5 DVSS single storage server and Ext4 performance comparison evaluation

图 5 DVSS 单台存储服务器与 Ext4 性能对比测试

当视频路数为 50 和 100 时,总的输入带宽分别为 25MB/s 和 50MB/s,Ext4 和 DVSS 的写入速度和输入总带宽一致.但是当视频路数为 200 时,输入带宽为 100MB/s,DVSS 能够顺利地实时写入,而 Ext4 的写入速度只有 85.69MB/s,低于输入带宽,造成较大的写入延迟.视频路数上升到 300 时,输入总带宽达到 150MB/s,DVSS 依然

能够实时地将视频数据写入磁盘,而 Ext4 的写入速度稍微下降到了 85.05MB/s.随着视频路数的继续上升,DVSS 的写入速度持续稳定增长,Ext4 的写入速度却显著下降.视频路数上升,用 Ext4 存储大量并发的视频流数据时,系统需要维护打开的文件表项增多,内存开销变大,且需要来回切换要写入的文件描述符,磁头移动更加频繁,导致其写入性能显著下降.

由此可见,DVSS 单台存储服务器的写入速度比基于文件系统的存储方案要稳定,能够支持 400 路分辨率为 1 080P 的视频流的并发写入.

3.4 整体性能测试

上述的实验只测试了 DVSS 单台存储服务器在多路数据流同时写入时的性能,为了进一步观察 DVSS 系统作为一个整体对外提供服务的能力,需要对其进行大规模数据流并发写入测试.将 40 台存储服务器分成 20 个存储组,每组 2 台存储服务器.我们用恒定码流生成器(constant-bit-rate traffic generator)模拟 2 000 路、4 000 路、6 000 路、8 000 路、9 000 路、10 000 路、11 000 路、12 000 路数据流,每路数据流的比特率均为 4Mbit/s.测试结果如图 6 所示,DVSS 系统的吞吐量随着视频路数的增加而线性增长.当并发写入的数据流路数为 11 000 时,平均每个存储服务器组处理 550 路数据流,组内单台存储服务器平均处理 275 路数据流.随着视频路数的继续上升,DVSS 吞吐量提升缓慢,慢慢达到了当前系统最大的吞吐量.因为在完整的 DVSS 系统中,为了提高数据的可靠性,同一个存储组内的服务器需要相互备份,带来了额外开销,限制了单台服务器最大支持的并发数据流路数.若要进一步支持更多的并发数据流路数,只需增加存储服务器组的数量就可以使系统线性扩展.

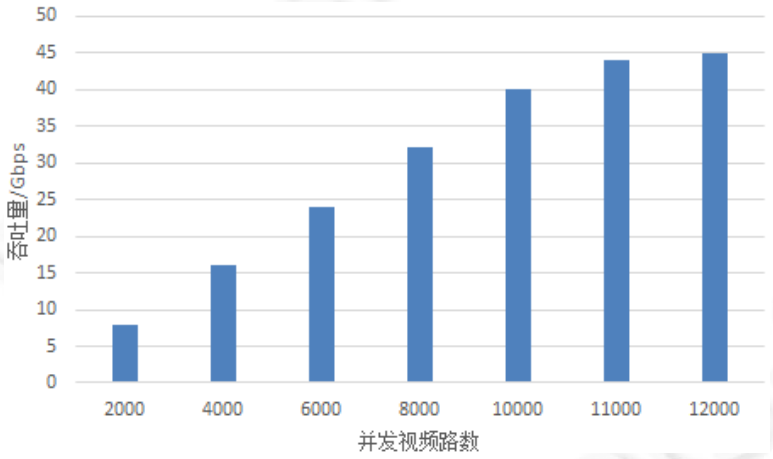

Fig.6 DVSS overall performance evaluation

图 6 DVSS 整体性能测试

3.5 相关工作比较

为了说明 DVSS 系统的优势,将 DVSS 系统与已有工作进行对比分析,分析结果见表 1.

Table 1 Comparison of DVSS with exisiting schemes

表 1 DVSS 与已有工作的比较

现有/本文工作	存储模型	容错能力	性能瓶颈	检索精度	写性能	读性能	性能描述
THNVR[6]	SQLite+定长文件	较弱	无	低	高	高	普通 SATA 硬盘上可同时记录 256 路码率为 1~2Mbit/s 的视频数据
DSFS[8]	连续存储模型 CSM	单点故障	存在	低	高	低	单台存储服务器可同时记录 300 路码码率为 4Mbit/s 的视频数据
DVSS	Redis+逻辑卷结构	高可用	无	高,秒级	高	高	单台存储服务器可同时记录 400 路码率为 4Mbit/s 的视频数据

THNVR 采用定长文件存储视频数据,使用 SQLite 维护视频文件的元数据.THNVR 检索精度与文件大小和视频码率有关,精度较低.THNVR 只有 1 份存储的视频数据,当 THNVR 节点发生故障时,与该节点连接的摄像头的视频数据将面临丢失的风险,容错能力较弱.DSFS 设计了一种连续存储模型(CSM),将对磁盘的随机高并发写转化为顺序写,有较高的写性能.但 CSM 模型中,同一路视频数据存储较为分散,读操作需要频繁移动磁头,导致 DSFS 的读性能较低.DSFS 的元数据全部由 MetaServer 管理,元数据量较大,存在着性能瓶颈和单点故障问题.DVSS 以 GOP 为单位组织视频数据,检索精度能达到秒级.DVSS 采用根据视频监控数据特征设计的逻辑卷结构,并结合多数据流并发写入优化方法,将无序随机的小写转化成有序的大段连续写,提供较高的写性能,并且同一路视频流的数据存储在同一数据块中,能够支持较高的读性能.同时,DVSS 针对分布式存储的特点,设计了元数据服务和视频数据的可用机制,提高了系统的容错能力和可用性.

4 结束语

本文针对传统基于 IP-SAN 视频监控存储系统在规模上升时难以线性扩展的问题,提出了面向视频监控数据的分布式视频流直存系统 DVSS.DVSS 将非结构化视频数据以 GOP 为单位组织直存入裸磁盘设备,摒弃了传统基于文件的存储方式,单台存储服务器在 400 路 1 080P 视频流数据并发写入时的速度达到本地文件系统的 2.5 倍左右.DVSS 采用的两级索引检索结构只需要状态管理器存储少量的元数据,极大地减少了状态管理器的负担,消除了性能瓶颈.同时,两级索引结构提供了精确到秒级的检索精度,能够更好地服务于上层应用.此外,DVSS 灵活的存储服务器分组方式和组内服务器互相备份的关系,让其能够更好地支持视频监控系统的扩展和容错需求.在 DVSS 系统的后续工作中,将考虑用新型存储设备,如 SSD(solid state disk(固态盘))等作为缓存盘,进一步提升系统对高并发高清视频的存储能力.另一方面,系统将会采用纠删码(erasure code)技术对存储时间超过 1 个月的和需要永久保存的录像段进行离线处理,进而用纠删码的技术提高设备的存储效率.

References:

[1] Wang YJ, Sun WD, Zhou S, Pei XQ, Li XY. Key technologies of distributed storage for cloud computing. Ruan Jian Xue Bao/Journal of Software, 2012,23(4):962−986 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4175.htm [doi: 10.3724/SP.J.1001.2012.04175]

[2] He J. Study on key technique in video surveillance storage system based on IP-SAN [MS. Thesis]. Shanghai: Shanghai Jiaotong University, 2011 (in Chinese with English abastract).

[3] Tang JX. The software design of storage subsystem for network video surveillance system [MS. Thesis]. Hangzhou: Zhejiang University, 2013 (in Chinese with English abastract).

[4] Jiang M, Niu ZY, Zhang SP. Design and implementation of video surveillance storage system. Computer Engineering and Design, 2014,35(12):4195−4201 (in Chinese with English abstract). [doi: 10.3969/j.issn.1000-7024.2014.12.027]

[5] Sun ZZ, Zhang QX, Tan YA, Li YZ. Ripple-RAID: A high-performance and energy-efficient RAID for continuous data storage. Ruan Jian Xue Bao/Journal of Software, 2015,26(7):1824−1839 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4606.htm [doi: 10.13328/j.cnki.jos.004606]

[6] Wu JY, Gu Y, Ju DP, Wang DS. THNVR: Distributed large-scale surveillance video storage system. Computer Engineering and Applications, 2009,45(31):56−59 (in Chinese with English abstract). [doi: 10.3778/j.issn.1002-8331.2009.31.018]

[7] Lin CF, Yuan SM, Leu MC, Tsai CT. A framework for scalable cloud video recorder system in surveillance environment. In: Proc. of the Ubiquitous Intelligence & Computing and 9th Int'l Conf. on Autonomic & Trusted Computing (UIC/ATC). Fukuoka: IEEE, 2012. 655−660. [doi: 10.1109/UIC-ATC.2012.72]

[8] Duan HC, Zhan WH, Min GY, Guo H, Luo SM. A high-performance distributed file system for large-scale concurrent HD video streams. Concurrency and Computation: Practice and Experience, 2015,27(13):3510−3522. [doi: 10.1002/cpe.3528]

[9] Deng P, Li MY, He C. Research on namenode single point of fault solution. Computer Engineering, 2012,38(21):40−44 (in Chinese with English abstract). [doi: 10.3969/j.issn.1000-3428.2012.21.011]

附中文参考文献:

[1] 王意洁,孙伟东,周松,裴晓强,李小勇.云计算环境下的分布存储关键技术.软件学报,2012,23(4):962−986. http://www.jos.org.cn/1000-9825/4175.htm [doi: 10.3724/SP.J.1001.2012.04175]

[2] 何炬.基于 IP-SAN 的视频监控存储系统关键技术研究[硕士学位论文].上海:上海交通大学,2011.

[3] 汤家兴.网络视频监控系统存储子系统软件设计[硕士学位论文].杭州:浙江大学,2013.

[4] 江冕,牛中盈,张淑萍.视频监控存储系统的设计与实现.计算机工程与设计,2014,35(12):4195−4201. [doi: 10.3969/j.issn.1000-7024.2014.12.027]

[5] 孙志卓,张全新,谭毓安,李元章.Ripple-RAID:一种面向连续数据存储的高效能盘阵.软件学报,2015,26(7):1824−1839. http://www.jos.org.cn/1000-9825/4606.htm [doi: 10.13328/j.cnki.jos.004606]

[6] 邬建元,顾瑜,鞠大鹏,江东升.分布式大规模监控视频存储系统 THNVR.计算机工程与应用,2009,45(31):56−59. [doi: 10.3778/j.issn.1002-8331.2009.31.018]

[9] 邓鹏,李枚毅,何诚.Namenode 单点故障解决方案研究.计算机工程,2012,38(21):40−44. [doi: 10.3969/j.issn.1000-3428.2012.21.011]

操顺德(1991－),男,湖北蕲春人,硕士,主要研究领域为流媒体存储系统,近似图像检测.

华宇(1978－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为网络存储系统,新型存储器件.

冯丹(1970－),女,博士,教授,博士生导师,CCF 高级会员,主要研究领域为海量存储系统及技术.

孙园园(1992－),女,博士生,CCF 学生会员,主要研究领域为哈希计算,近似查询.

左鹏飞(1992－),男,博士生,CCF 学生会员,主要研究领域为数据去重,非易失内存,存储安全.