城市轨道交通视频云存储系统架构及功能模块设计

蔡京军[1,2] 刘晓宇[3] 王珊珊[4] 沈 强[1,2] 潘 皓[1,2]

(1. 北京市轨道交通建设管理有限公司,100068,北京;
2. 城市轨道交通全自动运行系统与安全监控北京市重点实验室,100068,北京;
3.北京市轨道交通运营管理有限公司,100068,北京;
4.北京全路通信信号研究设计院集团有限公司,100160,北京//第一作者,高级工程师)

摘 要 以北京大兴国际机场线视频监视系统的视频云存储系统为例,介绍了视频云存储系统的架构选型和功能模块。建议视频云存储系统采用非对称式架构,从统一管理、动态负载均衡、集群性能、数据可靠性、智能运维 5 个方面设计管理模块,并介绍了各个功能模块的设计思路。视频云存储系统的应用提高了数据可靠性,为城市轨道交通视频监控系统的视频深化应用提供了高可用及在线扩容等保证。

关键词 北京大兴国际机场线; 视频监视系统; 云存储系统; 功能模块设计

中图分类号 U29-39

DOI:10.16037/j.1007-869x.2019.12.042

Urban Rail Transit Video Cloud Storage System Structure and Functionality Module Design

CAI Jingjun, LIU Xiaoyu, WANG Shanshan, SHEN Qiang, PAN Hao

Abstract Taking the video cloud storage system adopted by Beijing Daxing International Airport Express as an example, the structure selection and the functionality module of video cloud storage system are introduced. It is suggested to adopt the asymmetrical structure for video cloud storage system, and design the management modules from 5 aspects: centralized management, dynamic load balance, cluster performance, data reliability and smart operation, the design idea of each module is introduced. The application of video supervision system will improve the data reliability, guarantee the further deepening of video supervision system application in rail transit and the online memory expansion.

Key words Beijing Daxing International Airport Express; video supervision system; cloud storage system; functionality module design

First-author's address Beijing MTR Construction Administration Corporation, 100068, Beijing, China

1 云存储系统介绍

分布式存储系统组成集群资源池,资源共享,管理统一,扩展灵活,这类存储系统被业界称为云存储[1]。北京大兴国际机场线(以下简称"大兴机场线")视频监视系统存储系统应用的是视频云存储系统,该系统有别于传统存储技术,在数据可靠性、空间扩展性、统一管理能力、动态负载均衡、集群性能、智能运维等方面均可达到传统存储无法比拟的水平。

云存储系统采用了基于云架构的分布式集群设计和虚拟化设计,在系统内部实现了多设备协同工作、性能和资源的虚拟整合,可最大限度利用硬件资源和存储空间。整个系统从逻辑上由设备接入层、第三方接入、流媒体服务、图片服务、分布式文件系统组成。提供从前端数据采集、存储、转发于一体的数据层解决方案。同时,通过开放透明的应用接口和简单易用的管理界面,为整个视频监视系统提供高效、可靠的数据服务。本文介绍视频云存储系统的构架选型及功能模块设计,其系统构架及功能模块如图 1 所示。

2 视频云存储系统架构选型

按照元数据的管理方式,云存储集群文件系统可分为对称式和非对称式。对称式云存储集群文件系统中每个节点的角色均等,共同管理和维护元数据,节点间通过高速网络进行信息同步和互斥锁等操作。非对称式云存储集群文件系统中有专门的一个或者多个节点负责管理元数据,其他节点需要频繁与元数据节点通信以获取最新的元数据,如目录列表、文件属性等。下文分别从系统可用性和扩展性方面对两种架构进行比较分析。

1）系统可用性。对称式架构中，节点间的耦合性非常紧密，一旦某个节点出现问题，比如响应延迟，向其加锁就会迟迟得不到应答，会影响整个集群的性能。因此，如果某个节点把文件系统元数据破坏，整个集群系统都会受到影响，轻则丢失数据或元数据不一致，重则系统整体瘫痪。非对称式架构中，节点间采用松耦合机制，数据节点故障可以得到很好的隔离，系统的健壮性更强。

2）系统扩展性。对称式架构中，节点数量不能太多，否则节点间相互的通信量将迅速激增，达到瓶颈。比如：系统中有 10 个节点，每个节点可能同时在与其他 9 个节点通信，此时系统连接总数近似为 10×10；如果 100 个节点，则连接总数为 99×99。随着节点数量增加，信息同步复杂度呈几何级数增长，节点性能奇差。非对称式构架中，由专用的服务器维护元数据，节点增加带来的元数据复杂度是线性的，因而集群性能能够实现线性增长。

架构选型建议：视频监视系统每路存储码流为 6 Mbit/s（《北京市轨道交通视频监视系统应用规范》[2] 要求）、存储时长为 90 d（《中华人民共和国反恐怖主义法》要求），所需数据空间大、节点多、安全性要求高，因此建议视频云存储系统采用非对称式架构。

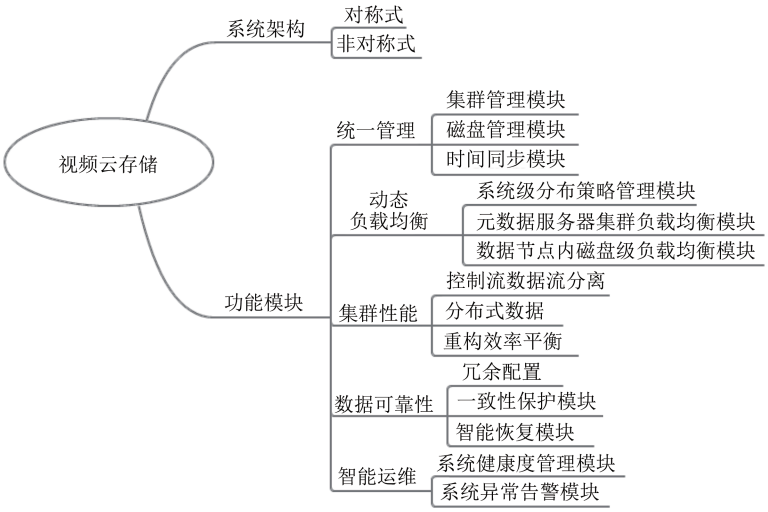

图 1　视频云存储系统的架构及功能模块示意图

3　视频云存储系统功能模块设计

3.1　统一管理功能模块

云存储使用分布式文件系统，将硬盘、服务器等设备进行集群，系统应支持对分布部署的存储设备的统一配置管理，方便接收存储设备告警上报。单站系统由多个节点机、控制设备等构成的空间体现为一个大的存储空间，提供存储服务时空间命名及 IP 地址应唯一。因此，云存储系统中应至少设计统一资源池管理模块，对全局的文件的元数据信息进行统一管理，并提供出 bucket（存储对象的容器），让用户将文件按 bucket 组织。元数据的信息中记录文件的组织信息，即：一个文件存储在哪些节点上，有哪些数据块组成。在系统一致性检测时，就可以通过比较发现元数据记录的信息和节点上记录的数据块信息的差异，用于修正差异。通过统一资源管理，可以方便有效地共享文件。整个系统的文件最终是统一使用所有存储节点所提供的存储空间，让弹性扩展更加简单高效。

统一管理功能模块包括集群管理、智能硬盘管理和时间同步 3 个模块。

1）集群管理模块。管理和识别系统中的所有数据节点，为负载均衡、数据恢复、客户端升级资源提供数据源，并负责节点在集群中的生命周期管理以及处理节点的扩容变更等操作。在节点第一次注册加入到集群，就为该节点分配身份 ID，即使修改节点的网卡的 IP 也仍能识别到该节点，避免由于节点 IP 的修改而导致数据迁移，进而触发大量的数据恢复等问题。集群管理模块识别处理按节点的重复加入、下线、删除等操作。在节点长时间下线，系统触发针对该节点上的数据块进行恢复，避免真正由于节点网络问题或者硬件问题导致数据恢复

延时而丢失。

2）磁盘管理模块。直接管理数据节点内的磁盘。将每个磁盘抽象成一个磁盘对象，并将磁盘对象交由磁盘管理模块统一管理，形成数据节点内部的存储层，为数据块在节点的统一存储和管理提供便利，实现所有磁盘相关的管理操作对外无感知。磁盘管理模块能感知磁盘的热插拔事件、磁盘异常损坏、磁盘变慢盘、触发磁盘自动上下线，对于新盘可以自动感知格式化，对于同一集群磁盘可以自动上线加载磁盘内的数据索引。对于热插拔事件，在节点出现异常时，通过磁盘漂移现数据快速恢复，也可以根据该特性实现新扩容之后，使节点容量快速均衡。对于磁盘故障事件，能感知到异常损坏或者慢盘，可提前触发系统进行恢复，实现免维护。

3）时间同步模块。在大规模部署的时候，为了保证数据的一致性需要保证使每个节点统一时间，通过 NTP（网络时间协议）时间同步功能，避免运行过程中因调整时间导致数据丢失的风险。运维获取到 NTP 时钟源 IP 地址后，各个节点通过 NTP 协议获取到 NTP 时钟源时间，并与本地时间做对比，进行时间同步。

3.2 动态负载均衡功能模块

视频业务正常工作情况下，存储系统应根据各节点承担的业务压力动态调整各节点的业务负荷，使系统始终处于均衡稳定状态中，避免单节点过载。在故障及数据恢复工作情况下，要求系统自动将流量重新分配到其他的节点机上，且各个节点机分配流量均衡一致，同时不能影响正常的视频存储业务。

一个存储集群内部，众多存储节点组建形成的一个统一空间，从整体性能、避免单点故障、数据热点瓶颈等方面都需要一个良好的动态负载均衡功能。动态负载均衡指集群内部自动根据各存储节点的 IO（输入输出）负载、空间容量、CPU、内存负载等因素，调度数据流向，实现 IO 读写的负载均衡。

视频云存储系统设计采用两级负载均衡调度。首先由元数据服务器选择一个负载轻的数据节点作为当前请求的读写节点，同时节点内部还会根据每个硬盘的负载选择最合适的硬盘参与数据写入。因此动态负载均衡功能模块包括高可靠的系统级数据分布策略管理模块、元数据服务集群负载均衡模块和数据节点内磁盘级负载均衡模块。

1）系统级数据分布策略管理模块。根据数据分布算法将数据块分布存储，以满足节点级容错以及硬盘级容错，即支持 N+M:B（N 为原数据模；M 为校验模块；B 为备用模块）。通过在集群负载均衡模块之上接入数据分布策略管理模块，能使负载均衡模块选择出分布更加合理的数据节点。

数据分布策略管理模块让系统可以支持多种 N+M:B 的策略。另外，当系统规模小，不满足节点容错的时候也可以通过 N+M:0 让数据分布降级为支持磁盘级容错，使系统逐步扩容而不用修改任何配置，后续所有新写入的数据可以自动提升为最优的容错数据分布，使数据分布从磁盘级容错提升节点级容错（节点级容错可以自动提升直至到最高的冗余数所相对应的节点数）。

2）元数据服务集群负载均衡模块。元数据服务器针对集群中所有节点汇报的实时负载压力（如 CPU 占用率、内存使用情况、网络流量大小、磁盘 IO 数据）进行汇聚，收集到集群负载均衡模块内，做统一的调度，优化节点间的负载，让所有节点均衡均摊系统压力，提升系统的整体读写性能，实现各个节点的容量均衡，使得系统能够支持异构容量和性能的数据节点。集群负载均衡模块为文件写入分配随机节点，满足 N+M 的节点级容错。

3）数据节点内磁盘级负载均衡模块。通过实时收集磁盘的负载、磁盘空间使用情况，调度写入到该节点内的数据流，均衡地分布到各个低负载、高可用容量的磁盘，使写入更加平滑，最大粒度发挥磁盘的顺序写入能力，并在长期负载下使得各个磁盘的容量能最终达到均衡，实现系统容忍异构的磁盘。

3.3 集群性能功能模块

视频云存储系统采用数据离散技术，使得客户端可以有效利用众多存储节点提供的聚合网络带宽，实现高速并发访问。客户端在访问云存储时，首先访问元数据服务器，获取将要与之进行交互的数据节点信息，然后直接访问这些数据节点完成数据存取。

客户端与元数据服务器之间只有控制流，而无数据流，这样就极大地降低了元数据服务器的负载，使之不成为系统性能的一个瓶颈。客户端与数据节点之间直接传输数据流，同时由于文件被分散到多个节点进行分布式存储，客户端可以同时访问多个节点服务器，从而使得整个系统的 IO 高度并行，系统整体性能得到提高。

分布式系统由于数据分散存放在不同的节点，因而出现磁盘故障或者节点故障时不可避免地会进行跨节点的数据重构。当追求重构速度时，节点间的数据交互压力很大。为了避免网络拥塞，拖慢整个系统，需要将业务网络和存储网络分离。业务网络和存储网络分别使用不同的物理网卡以达到从网络上相互隔离的目的，可以根据现有网络状况选择千兆和万兆连接。前端 IPC（进程间通信）接入的数据流走单独的业务网络，后端数据离散流和控制信令流走单独的存储网络，以满足不同场景的组网需求。无论哪种组网，系统中所有节点网络都是冗余的，任何单一网口故障或者单一交换机故障均不影响系统使用。

3.4 数据可靠性功能模块

数据是业务系统核心应用的最终保障，其可靠性至关重要。云存储系统的核心是一个分布式文件系统，设计时假设任意机框、任意节点、任意硬盘都可能出现故障，通过分布式的数据冗余、数据操作日志、元数据主备冗余、数据自动恢复等多种机制来处理这些故障。

云存储系统针对视频数据主要采用 Erasure Code（纠删码或者叫擦除码）算法，以较小的数据冗余实现较高的可靠性，而没有采用互联网采用的多副本和监控领域常用的数据备份方式，存储空间利用率高。数据可靠性功能模块包括一致性保护模块和智能恢复模块，对数据进行可靠性保障。

1）一致性保护模块。系统长期运行过程中，由于断电、人为破坏、写入异常、程序 bug 等原因，都可能导致写入的文件出现损坏。针对可能出现的数据和写入时不一致，通过读写加入校验值记录和判断内部周期性检测数据块是否和记录的校验值不一致，当发现不一致时，汇报给元数据服务器，由智能恢复模块进行数据恢复，从而保证数据一致性。在分布式系统中，由于各种各样的原因，会有小概率的集群内多台节点数据块和元数据服务内管理数据块信息出现的差异。为了解决该问题，通过周期性触发数据节点进行全量汇报，报告自身拥有数据块索引信息，使元数据可以据此不断修正自身记录的信息，从而使外部读取文件正常或者感知到文件缺失时能触发自动恢复功能，以保障云存储系统的可靠性。

2）智能恢复模块。实时感知文件在云存储中出现的异常块，对于出现异常块的文件按照调优适合视频监视系统特点的恢复策略进行恢复，尽量让时间最近、文件损坏更严重的优先恢复。并针对冗余度高、可靠性高的文件，在出现可容忍的少量数据块损坏时，可以减少恢复（比如：如果 12 个存储器中坏了 3 个，由于本身已经很可靠，则可以不进行恢复）。同时为了支持更加紧急的数据文件，在自动恢复策略之上引入优先恢复队列，用于在某些特殊情况下人工判断某天的数据需要优先恢复。优先恢复会打断自动恢复，优先完成指定的某一天内的数据文件。智能恢复模块能根据当前系统的负载压力实时调整恢复速度，在保证读写不受影响的情况下，高速完成异常文件的快速恢复。

3.5 智能运维功能模块

云存储系统运行过程中出现比较严重的问题时，如果不能及时感知，可能会造成系统停止服务，甚至数据丢失的严重问题。为了使管理人员能够及时发现云存储故障，需要设计智能化运维管理模块，使系统轮询各个节点的监控项，使运维人员能第一时间介入恢复系统。同时综合应用多种技术延长系统生命期，比如硬盘休眠技术，支持硬盘分时上电，达到节能、延长硬盘寿命目的。智能运维功能模块包括系统健康度管理模块和系统异常告警模块。

1）系统健康度管理模块。针对系统的各个节点，一起参与系统多维度的数据收集，并汇总分析系统整体运行状况。系统健康度集成在运维管理模块中，提供实时流量、设备在线率、系统服务状态、硬件信息检测、历史流量波动、平均 IOPS（磁盘性能评价指标）、IO 平均延时、系统容量变化等信息，并通过图表展示给运维人员。

2）系统异常告警模块。对系统的流量波动、磁盘容量达到使用阈值、节点上下线、磁盘告警、各个节点的 CPU 温度、内存使用量、风扇转速、系统盘使用量、ssd（存储介质）磨损程度、网卡速率波动、网线异常等进行实时告警，可以使技术支持人员快速感知到当前云存储系统存在的风险，快速响应，减少因为系统错误累计超过系统能容忍的阈值而引发的异常事件，从而保障系统高可靠地持续运行。

3 结语

为了保证云存储系统各个优势功能实现，针对大兴机场线云存储系统，主要设计的系统功能模

（下转第 181 页）

用的无线基带单元 BBU、RRU、合路器及泄漏电缆组成,提供无线接入服务。

3) 车载子系统。由位于车头及车尾的 A、A′网共用的车载接入单元(TAU)及车载天线系统组成。两个 TAU 互为主备冗余配置,当主用 TAU 发生故障时,车上的数据会选择备用 TAU 进行数据传输,保证车地数据传输的高可靠性。

7 方案验证

大兴机场线于 2019 年 6 月 15 日正式试运行。试运行期间 LTE-M 系统不断进行调试和优化,经过多次系统功能和业务数据传输测试,并于 2019 年 9 月 26 日正式开通运营。结果显示:按照上述综合承载业务配置及 LTE-M 系统部署方案,CBTC 列车运行控制业务无丢包现象,PIS 视频业务播放流畅,车载视频业务无卡顿及马赛克现象,列车紧急文本下发及运行状态信息数据能够及时准确传输。北京地铁新机场线 LTE-M 系统能够满足地铁综合承载业务的需求。

8 结语

随着城市轨道交通线路的不断建设,LTE-M 系统应用于城市轨道交通综合承载,可以对列车的车地通信系统进行实时监控,对整个列车系统的状态进行实时观测[7],能够满足综合承载各业务的不同需求,保障列车的安全运行,提高城市轨道交通的管理水平与服务水平,促进整个城市轨道交通行业的良性发展,能够为建设安全、高效、绿色、环保的城市轨道交通线路发挥更多积极的作用。

参考文献

[1] 中华人民共和国工业和信息化部.工业和信息化部关于重新发布 1 785~1 805 MHz 频段无线接入系统频率使用事宜的通知:工信部无[2015]65 号[Z].北京:中华人民共和国工业和信息化部,2015.

[2] 戴克平,张艳兵,朱力,等.基于 LTE 的城市轨道交通车地通信综合承载系统[J].都市快轨交通,2016(7):69.

[3] 中国城市轨道交通协会技术装备专业委员会.LTE-M 系统承载 CBTC 业务及接口规范:T/CAMET 040069—2017[Z].北京.中国城市轨道交通协会技术装备专业委员会,2016.

[4] 李照敬,葛淑云.基于 LTE 技术的城市轨道交通综合承载业务需求分析[J].铁道通信信号,2015(7):74.

[5] 孙寰宇,顾向峰.基于 LTE 技术的车地无线通信组网方案研究[J].铁道标准设计,2014(8):160.

[6] 顾向峰.城市轨道交通车地通信 TD-LTE 综合业务承载测试分析[J].城市轨道交通研究,2016(8):140.

[7] 陈赛印.LTE-M 综合承载和互联互通测试方法的研究[D].北京:北京交通大学,2017.

(收稿日期:2019-09-10)

(上接第 175 页)

块包括元数据服务集群负载均衡模块、数据节点内磁盘负载均衡模块、高可靠数据分布策略管理模块、数据节点的磁盘管理模块、智能恢复模块、一致性保护模块、系统运维管理模块等。在系统实际配置中选用数据可靠性及系统扩展性更好的非对称式架构,采用专门的元数据服务器对元数据进行管理。同时,采用 1+1 冗余配置元数据服务器,进一步保障元数据相关功能高可用性。

通过统一资源池管理模块、集群管理模块、磁盘管理模块,从系统、节点、硬盘三个级别实现统一管理;通过高可靠的数据分布策略,将负载均衡能力深入到系统级、元数据节点级和磁盘级;通过各系统功能模块的设计,保障云存储系统的集群性能和智能管理水平。

视频云存储系统在城市轨道交通领域的应用提高了数据可靠性,为视频深化应用提供了高可用及在线扩容等保证,可将视频监视系统的价值水平带入新的阶段。

参考文献

[1] 许辉.存储技术在铁路综合视频监视系统的应用[J].铁路通信信号工程技术,2016(6):1.

[2] 北京市交通委员会.北京市轨道交通视频监视系统应用规范[Z].北京:北京市交通委员会,2014.

(收稿日期:2019-09-10)