分布式视频会议服务器媒体管理系统①

崔梦华[1,2], 廉东本[2]

[1](中国科学院大学, 北京 100049)
[2](中国科学院沈阳计算技术研究所, 沈阳 110168)

摘 要: 由于其方便易用及可靠性, 又得益于网络的普及, 基于 TCP/IP 的视频会议已经成为目前市场主流, 因此搭建可靠的视频会议服务器也成为了需要解决的主要需求. 视频会议服务器不同于普通的流媒体服务器, 视频会议服务器需要较高的实时性, 而且其接受的上传/下载速度比要比普通的流媒体下载/在线观看服务器高得多. 视频会议服务器的媒体服务器主要负责流媒体的接收保存及同步转发. 本文借鉴 HDFS 的最基本原则, 提出了一种可靠的分布式结构用于解决视频会议服务器的媒体存储问题.

关键词: 视频会议; 流媒体存储; 分布式

Distributed Media Management System of Video Conference Server

CUI Meng-Hua[1,2], LIAN Dong-Ben[2]

[1](University of Chinese Academy of Sciences, Beijing 100049, China)
[2](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)

Abstract: For the TCP/IP based video conference has been the main requirement because of its convenience and thanks for the well implemented networks now, building a stable video conference server has been required. Differed from normal video-on-demand servers, video conference server requires more on real-time demands, and it faces higher rate of download and upload. Using the basic principles of HDFS for reference, this essay comes up with a reliable distributed file system for the video conference server.

Key words: video conference; media management; distributed file system

1 引言

基于 TPC/IP 网络的视频会议服务已经成为市场主流, 过去的专线专设备视频会议系统由于其成本和应用场景限制已经逐渐被基于IP网络的视频会议系统所取代. 基于 IP 网络的视频会议大大降低了用户进行视频会议的成本. 在现有的 IP 网络上进行视频会议, 用户只需要为每次会议付出少量的服务费即可. 而且只要有网络的地方, 只要有音视频设备, 随时随地都可以进行视频会议, 再不受场景限制.

视频会议服务器的核心是流媒体服务器. 为了应对随时可能突发的媒体数据流, 为用户提供稳定可靠的服务, 服务器必须具有较大的数据吞吐量, 在这方面, 分布式系统是一个目前比较成熟且应用效果较好解决方案. 分布式让企业拥有相当于整片云的存储能力, 成功地解决了视频存储的难题[1].

2 系统概述

视频会议服务器从功能上划分成几大部分. 为了方便用户完成一场完整的视频会议, 服务器端至少需要以下功能模块: 业务逻辑、网络模块(RTP、RTCP)、sip 服务器、媒体管理、会议管理等. 图 1 所示为视频会议服务器的基本结构模块及其基本数据流方向.

在视频会议过程中, 除了创建、发起、加入会议、用户管理等功能外, 主要的工作模块是网络模块和媒体管理模块, 服务器的主要工作量在于媒体管理.

媒体管理模块主要负责接收网络模块传输来的数

① 收稿时间:2014-07-23;收到修改稿时间:2014-08-19

据流并进行组播分发以及实时存储、下载回放等. 其中, 实时存储、下载回放的功能, 由于视频会议所产生的实时数据量非常大, 需要高速的、吞吐量和容量大、扩展性好、数据安全性高的文件系统. 分布式文件存储系统是目前针对这个需求的最佳选择.

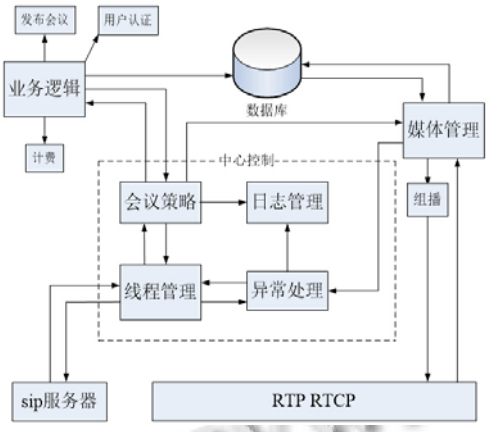

图 1　视频会议服务器结构示意图

针对以上需求, 系统的总体设计包括: 调度节点、存储节点、对外接口(类).

1.1 调度节点

本系统不是完全分布式的系统, 为了保证较高的工作效率, 系统以半分布式的形式组织形成. 调度节点在系统中负责任务分配、负载均衡、系统状态维护、数据备份、数据查询等重要功能.

1.2 存储节点

系统接收请求后, 首先由调度节点进行任务分配, 然后被分配到存储任务的节点就进入工作状态, 开始接受用户发送的数据并存储到本地. 但存储节点的功能不止如此.

为了保证数据不易丢失, 数据节点之间的数据需要相互备份, 即每个节点上的数据需要在其他节点上备份至少一份, 同时当一个节点失效时, 需要将其保存的数据重新备份到其他节点. 此外, 为了防止调度节点的失效, 系统同时需要运行一个备选节点与调度节点同时运行, 在调度节点失效时即刻成为调度节点进行调度任务并选出下一个备选节点.

1.3 对外接口

分布式文件系统是以一个单独实体的形式存在, 用户对系统的使用应该像使用普通的文件系统一样, 不需要知道实现细节、通信协议, 而是直接调用系统所提供的接口, 这里以类的形式提供给用户.

2 系统设计

分布式系统从分布的程度上可以分为集中式、半分布式、完全分布式等几种类型. 分布的程度越高, 节点的自主性就越高, 系统应对节点失效的可靠性就越高, 扩展性越好, 但资源查询效率也越差, 整体性能不容易得到发挥. 作为一个特定应用的分布式文件系统, 采用集中式的或者半分布式的结构更有利于系统性能的发挥.

采用这种结构的, 比较著名的是开源的 HDFS (Hadoop File System). 与以往不同的是, HDFS 认为硬件失效是常态, 而非异常[2], 所以要格外应对硬件失效问题. HDFS 是一个集中分布式系统, 系统由一个称为 NameNode 的节点进行集中调度(Hadoop 中不只是文件系统调度, 还负责分布式计算任务调度), 多个 DataNode 进行存储及计算任务. HDFS 对于文件丢失问题, 采用的是文件完全备份, 默认情况下所有文件备份三份[1]. HDFS 还对文件进行了分割存储, 通常是将文件分为以 64M 为基础大小的块分别存储在不同的 DataNode 中, 这样可以并行从不同的节点获得数据, 从而获得很高的文件读取速度. 但作为视频会议服务器的存储系统, 本系统与 HDFS 的需求稍有不同, 具体内容会在下面章节中详细说明.

2.1 系统结构

对视频会议内容进行存储是为了方便用户日后随时读取调用, 所以视频文件的存储是实时的, 必须在视频会议进行的同时, 将视频会议服务器收到的数据流实时存储. 所以本系统是作为视频会议系统的应用或服务的形式存在的.

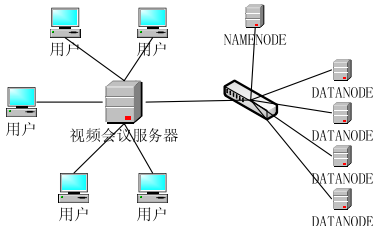

图 2　基于分布式文件系统的视频会议系统架构图

系统架构如图 2 所示. 整个分布式文件系统通过交换机连接成为一个集群, 对外提供存储服务, 视频会议服务器直接与交换机相连, 与文件系统进行交互并调用系统提供的服务、传输并存取数据.

2.2 任务调度

NameNode 负责任务调度.

当用户(此系统中是视频会议服务器)有文件存取需求时, 首先与 NameNode 进行通信, 请求获得服务. NameNode 获得请求或对请求进行分析, 获得请求的具体内容, 然后综合当前系统中各个节点的工作状态进行任务分配, 最后将执行任务的节点返回给用户, 此后用户直接与工作节点进行通信、数据交换, 直到任务结束, 再将任务结束请求发送给 NameNode, 完成此次文件存取请求. 系统对外部请求的处理流程如图 3 所示.

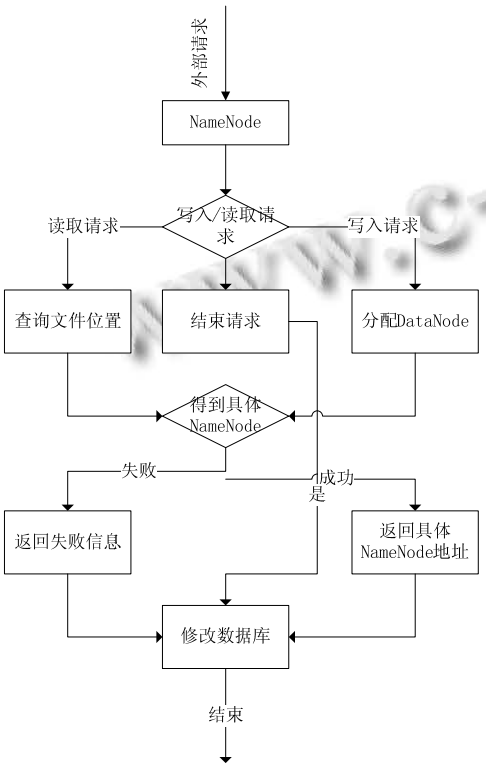

图 3 分布式文件系统对外部请求的处理过程

在响应外部请求成功后, NameNode 对系统的元数据进行更新, 即将系统节点的工作状态进行更新, 将文件存储目录(即文件存储在哪些 DataNode)记录在数据库中.

2.3 文件存储、备份

用户得到分配的 DataNode 的地址之后就开始直接与其进行通信, 进行文件的写入或读取.

当用户进行文件写入时, 系统对写入的数据行进实时备份. NameNode 在响应写入请求时, 同时根据节点的负载状态分配 N 个节点, N 的大小可在系统初始化时确定, 默认为 2, 并将其中的一个节点的地址返回给用户, 将其他的 N－1 个节点的地址返回给直接与用户通信的节点, 作为实时备份节点－与用户通信的节点将接收到的数据即时传输到其它节点备份保存.

与 HDFS 不同的是, 此处没有将文件分块存储, 而是直接将一场视频会议的数据保存为一个文件, 备份亦是如此. HDFS 将文件分块存储是为了获得较高的读取和计算性能, 而视频会议系统的需求与此不同, 用户对已保存的视频文件的读取不需要很高的速度, 达到实时播放即可, 而且现实中受网络带宽限制, 读取速度不会很高, 而最关键的因素是: 同时对一个视频文件读取的用户数量极其有限, 因为只有参加此会议的用户才有读取权限. 而如果将文件进行分块存储的话无疑将增大大加系统需要保存的元数据的量. 这将加重 NameNode 的负担, 使 NameNode 的瓶颈效应更加明显. 使用纠删码[3] [4]可以节省很多存储空间, 但同样需要较多计算和元数据, 而且对于大小未知的文件也不太适合.

2.4 节点失效

节点失效是每个分布式系统需要解决的基本问题之一, 节点失效会造成数据丢失、任务出错. 在 HDFS 中更是将失效当成了硬件的常态. 为了保证提供的文件存取服务安全可靠, 本系统主要需要解决 NameNode 和 DataNode 的失效问题.

(1) NameNode 节点失效

NameNode 是系统负责调度的节点, 其突然失效不会对当前正在进行的存取任务造成影响, 但将无法响应新的请求, 更严重的是系统的元数据将全部丢失, 虽然理论上元数据可以从 DataNode 的数据中重新生成, 但那将耗费太多时间和带宽.

与 DataNode 节点的数据一样, NameNode 中的元数据也进行实时的备份, 可以看成是一台备用机与 NameNode 同时运行, 当 NameNode 失效时, 备用机立刻接替 NameNode 成为新的 NameNode.

具体的实现方法是, 在系统初始化时 NameNode 从所有的节点中选取一个节点作为候选节点 Candidate. 在系统运行中, NameNode 在元数据发生任何变化时就立即将所发生的变化同步到 Candidate. NameNode 节点失效时, Candidate 节点立即成为新的 NameNode, 实现无缝切换, 同时从所有的节点中选出一个新的 Candidate, 将所有的元数据同步到新 Candidate 节点, 此后并行运行. 如果 Candidate 节点先于 NameNode 失效, 则由 NameNode 重新选出一个

Candidate, 并将所有元数据同步到新 Candidate 上, 此后并行运行.

(2) DataNode 节点失效

对于已完成的视频会议, DataNode 节点的失效将造成其某一备份数据的丢失; 对于正在进行的视频会议, 节点的失效将造成通信的失败. 所以为了应对以上两点, 当一个 DataNode 节点失效时:

NameNode 节点首先从元数据中查询出正在此节点上进行的视频会议, 找到其实时备份节点, 将服务请求重定向到其中一个点, 实现无缝切换,并分配一个新的节点进行实时备份. 同时, 查找失效节点的上的文件的备份节点, 将这些文件重新备份到其他节点上, 保证数据备份数为 N.

3 对外接口

为了完成与分布式文件系统的通信, 必须提供一个封装好的接口提供给用户, 使用户对本系统的使用像对普通文件系统的使用一样, 用户只需要声明一个接口类的对象, 调用类中提供的类似 Open、Read、Write、Close 的方法, 便可以使用分布式文件系统所提供的服务.

对外接口的实现, 实质是通信协议的实现[5].

系统初始化后, NameNode 和 DataNode 分别监听不同端口(默认 13333, 14444), 等待用户发送来的命令:

3.1 **Open**

用户在接口对象初始化时指定 NameNode 地址, 调用 Open 方法时发送“open”命令到 NameNode 的 13333 端口, 然后等待 NameNode 完成节点分配或文件查找, 接收返回的 DataNode 地址, 再与 DataNode 的 14444 端口建立连接, 如果失败, 则重新请求 NameNode 分配或发送 Close 终止此次调用.

3.2 **Read/Write**

直接与 Open 阶段建立好的 DataNode 连接通信, 中间如果节点失效, 自动向 NameNode 重新请求新 DataNode 节点. 如果 NameNode 节点失效, 则发送 UDP广播, 触发Candidate节点使其成为新NameNode.

3.3 **Close**

发送“end”命令, NameNode 更新元数据, 断开端口连接.

4 系统应用及结语

分布式文件系统具有很好的安全性、可靠性、可扩展性, 而且具有很高的数据吞吐量, 完全可以满足视频会议的实时写入请求, 对于视频会议系统本身就不大的读取请求, 也可以完全满足.

对于视频会议服务器本身, 分布式存储系统还可以分担更多的任务, 将流媒体服务器的功能完全应用于分布式系统中, 在充分利用系统的存储性能的同时, 充分利用系统的计算性能.

此外, 在将数据的存储改为固定块大小(比如 100M、200MB)[6], 并且增加相应的元数据后, 系统还可以应用在视频监控等需要较多存储空间的应用中.

参考文献

1 雷玉堂.云技术及其在视频存储中的应用.中国公共安全,2013,20:178–187.

2 刘统阁,刘波,杨志文.Hadoop 在 VOD 系统中的应用研究.计算机与现代化,2012,6:195–199.

3 黄震.大规模分布式存储系统中数据冗余技术研究[学位论文].长沙:国防科学技术大学,2012.

4 张凯.分布式存储系统中节点修复问题研究[学位论文].成都:西南交通大学,2012.

5 Toharia P, Sánchez A, Bosque JL, Robles OD. Efficient grid-based video storage and retrieval. Proc. of On the Move to Meaningful Internet Systems: OTM 2008, OTM 2008 Confederated International Conferences, CoopIS, DOA, GADA, IS, and ODBASE 2008, Monterrey, Mexico, November 9-14, 2008, Part I, 2008.

6 罗丽丽.视频存储优化技术研究与应用[学位论文].长沙:国防科学技术大学,2009.