FFS：一种基于网络的PB级云存储系统

吴海佳，陈卫卫，胡谷雨，董继光
（解放军理工大学 指挥自动化学院，江苏 南京 210007）

摘 要：针对传统海量存储系统成本高、管理复杂、升级困难等问题，借鉴云计算中效用计算理念与存储虚拟化技术，设计实现了一种PB级云存储系统（FFS，formicary file system），该系统基于通用PC集群构建，具有海量并行扩容、自动故障切换和数据恢复、动态负载均衡等关键能力，且通过虚拟化技术，将云存储资源以磁盘的方式提供给客户端使用。详细阐述了FFS的设计思路与核心机制，并对FFS进行了性能测试。测试结果表明，在由17台普通PC构建的云存储环境下，FFS最大聚合读带宽可达105M bit/s，最大聚合写带宽可达49.4M bit/s，且系统具有良好的负载均衡表现。

关键词：蚁穴文件系统；海量存储；云存储；效用存储；存储虚拟化
中图分类号：TP302 **文献标识码**：A **文章编号**：1000-436X(2011)9A-0024-10

FFS: a PB-level cloud-storage system based on network

WU Hai-jia, CHEN Wei-wei, HU Gu-yu, DONG Ji-guang
(Institute of Command Automate, University of Science and Technology of PLA, Nanjing 210007, China)

Abstract: Most of traditional mass-storage systems are high-cost, hard to manage, and difficult to upgrade. So a new PB-level cloud-storage system named formicary file system (FFS) was prevailed, which draw on the idea of utility-storage and storage-virtualization. FFS is built from many inexpensive commodity components that often fail, it has parallel expansibility, it can switch failure and renew data automatically, and it can balance work-load and data-load dynamically. Through the technology of virtualization, it can provide the aggregate storage resource to the client in the form of a common disk. The forth the architecture and key-technic of FFS, and at last, a test towards FFS is made. The test results show that at the experiment condition of a 17-PC-built cluster, the maximum aggregate read-rate can reach to 105M bit/s, as well as the maximum aggregate write-rate can reach to 49.4M bit/s, and during the test, FFS behaves well at load balance.

Key words: FFS; mass-storage; cloud-storage; utility-storage; storage virtualization

1 引言

图灵奖获得者Jim Gray提出了一个经验定律：网络环境下每18个月产生的数据量等于有史以来数据量之和[1]。信息资源的爆炸性增长，对存储系统的容量、可扩展性、数据可用性以及I/O性能等方面提出了越来越高的要求。目前，企业在存储超过1TB以上数据时，一般采用DAS/NAS/SAN架构[2]。用这类架构存储几十TB以下数据的成本，企业通常还可以承受，若要扩展到PB级存储，成本异常高。除此之外，这类架构还存在管理复杂、扩展困难，易出现存储孤岛等问题[3]。为了低成本、无存

收稿日期：2011-02-17
基金项目：国家自然科学基金资助项目（60603029）；国家高技术研究发展计划（“863”计划）基金资助项目（2008AA01A309）
Foundation Items: The National Natural Science Foundation of China (60603029); The National High Technology Research and Development of China (863 Program) (2008AA01A309)

储孤岛地集中存储海量数据，需要改变原有思想，采用新的云存储架构。云存储是服务器虚拟化的一种途径，各类服务器通过云存储平台进行整合，使各服务器的应用程序融入同一个存储平台，从而减少硬件采购成本，降低管理成本，并能提高资源的利用率。

云存储这个概念一经提出，就得到了众多厂商的支持和学者的关注。Amazon 推出的 Elastic Compute Cloud（EC2）中包含 Simple Storage Services（S3）[4]云存储服务，旨在为客户提供可靠、易用以及低成本的网络存储服务。Google 设计的大型分布式文件系统 GFS[5]能为其云计算提供海量存储。HDFS 是 Hadoop[6]的底层分布式文件系统，是 GFS 的开源实现。国内一些大型企业也开始部署和使用云存储系统，如淘宝网和阿里巴巴网站分别利用 Taobao File System (TFS)[7]和 Alibaba Distributed File System (ADFS)[8]构建了其底层云存储系统。

然而，以上各类云存储系统都是针对各自的应用而设计的专用存储系统。S3 适合于网络备份文件；GFS 和 HDFS 为满足上层 Map/Reduce 分布式计算，专门对存储大文件做了优化；TFS 和 ADFS 为了满足淘宝网和阿里巴巴网站存储海量图片、邮件文件等的需求，专门对小文件存储做了优化。本文则是在借鉴各类云存储系统、分布式文件系统等的基础上，设计并实现了一种用于构建通用云存储系统的文件系统 Formicary File System (FFS)。

2　FFS 的设计与实现

2.1　系统架构

FFS 云存储系统的部署如图 1 所示。FFS 云存储系统由 3 大模块组成：主控服务器模块、存储服务器模块和 FFS 客户端模块。

主控服务器模块对云存储客户端提供目录服务和元数据服务，并对存储服务器集群进行监控，部署于一台性能较好的服务器中；存储服务器模块负责文件数据的具体存放，部署于存储服务器集群中的每个节点中；客户端代理模块负责对云存储客户机提供虚拟磁盘服务，将云存储客户机对虚拟磁盘的操作请求提交给主控服务器，并从存储服务器读取/写入数据，部署于有海量数据存储需求的应用服务器中（例如，图 1 所示的数据库服务器、FTP 服务器、邮件服务器、Web 服务器、流媒体服务器等）。FFS 云存储系统模块间通信过程如图 2 所示。

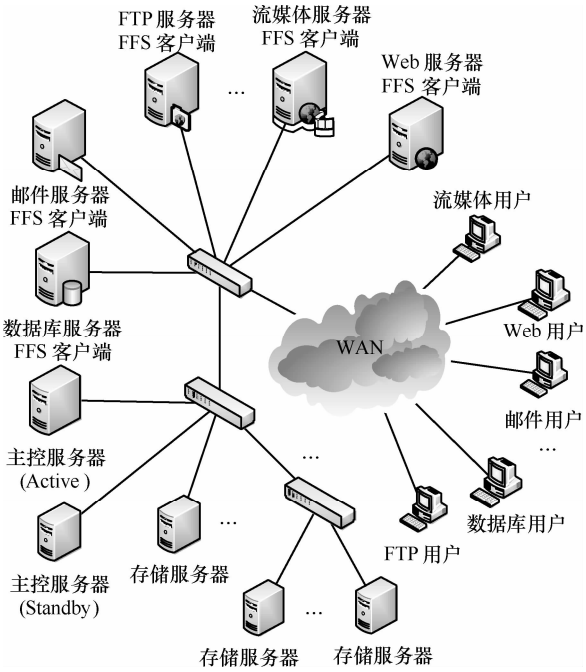


图 1　FFS 云存储系统部署

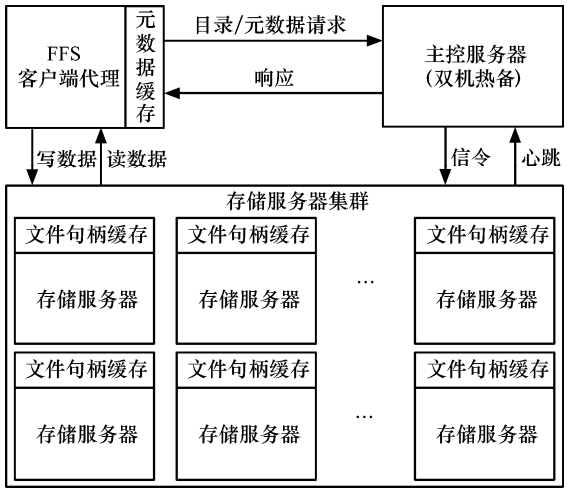


图 2　模块间通信

2.2　核心机制

为了在通用低性能的 PC 集群上构建高可靠高性能的通用云存储系统，在 FFS 中设计了集群监控机制、自动扩展机制、动态负载均衡机制、在线备份与自动恢复机制、AC-RU (adjust-controlled recently-used)缓存机制，客户端虚拟磁盘机制以及主控服务器双机热备机制。

2.2.1　集群监控

集群监控机制是集群自动扩展机制、动态负载均衡机制，以及数据自动恢复机制的基础。

主控服务器在内存中维持一个在线存储服务器列表，列表中的每一个逻辑节点保存其对应存储

服务器最近一次心跳所汇报的状态。如图3所示，逻辑节点中还设有一个监视变量state，主控服务器被动接收来自存储服务器的心跳信息，一旦收到某存储服务器的心跳，则更新其对应逻辑节点中的状态信息，并将该节点的state变量值重新设置为2。主控服务器按照逻辑节点列表周期性检查存储服务器，检查周期等于存储服务器的心跳周期，每检查一轮将被检节点的state值减1，若state值为0，则做出“该存储服务器已经掉线”的判断，并启动数据自动恢复过程。

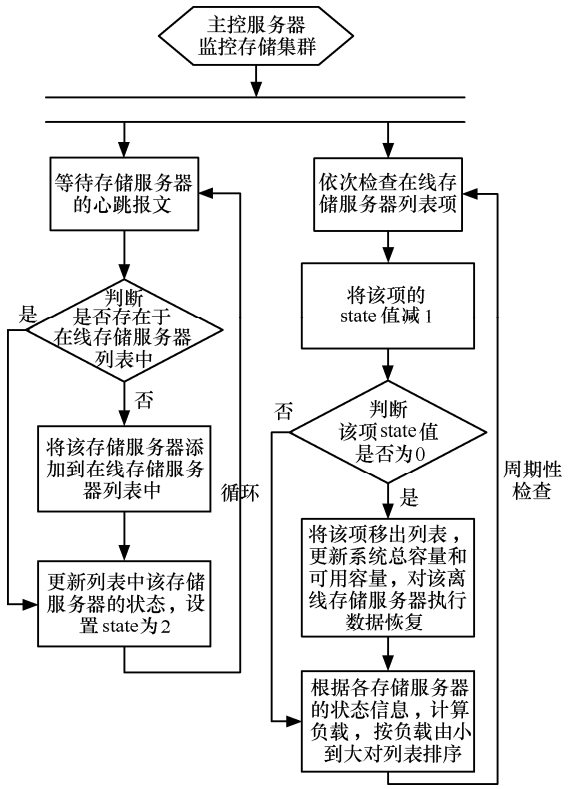


图3 主控服务器监控存储服务器集群

2.2.2 自动扩展

存储集群自动扩展机制解决传统海量存储系统扩展困难的问题。

当需要扩展FFS云存储系统总容量时，管理员只需将新的存储服务器接入存储服务器集群，该存储服务器启动后将周期性地向主控服务器发送心跳报文，心跳报文中包含该存储服务器当前的状态（包括CPU占用率、内存占用率、网卡上/下行速率、磁盘耗费）。如图3左半部分所示，主控服务器收到该心跳报文后将在内存中新建一个对应于该存储服务器的逻辑节点，并将该逻辑节点添加到在线存储服务器列表中。主控服务器会周期性地对在线存储服务器列表进行遍历检查，当下一次检查完毕，主控服务器将更新当前FFS云存储系统的总容量。

2.2.3 动态负载均衡

云存储集群在长期运行后会出现数据分布不平衡的问题。FFS云存储系统采用动态负载均衡机制解决了这一问题。FFS从以下两方面进行系统负载均衡。

一方面，当云存储客户端提出创建文件的请求时，主控服务器从在线存储服务器列表中挑选出负载最轻的若干台存储服务器，并从这若干台中随机挑选两台作为该文件的主存储服务器和备份存储服务器。这样做的目的是避免当同一时刻有大量并发任务时，把这些任务都交给负载最轻的存储服务器可能导致这台存储服务器过载或宕机。

为新建文件分配存储服务器时，不仅需要考虑磁盘负载，还要考虑存储服务器的工作负载，它们之间是乘性关系。存储服务器的磁盘负载可用磁盘可用空间百分比来表示，存储服务器的工作负载可用网卡上/下行速率与网卡的最大速率来表示，上/下行速率之间是加性关系。式(1)为存储服务器的负载计算方法。

$$l_i = a_i \quad (u_i + d_i)/n_i \tag{1}$$

其中，l_i为存储集群中第i台存储服务器的负载值，a_i为此存储服务器的磁盘可用空间百分比，u_i为此存储服务器的网卡上行速率，d_i为其下行速率，n_i为该网卡的最大速率，u_i、d_i和n_i的单位都是kbit/s。如图3右半部分所示，主控服务器会根据各节点的负载值周期性地对存储服务器列表进行排序。

另一方面，主控服务器会周期性地计算当前存储集群的总工作负载和数据分布不平衡度。当总工作负载低于某设定值$L_{Threshold}$，且数据分布不平衡度高于某设定值$U_{Threshold}$时，则对存储负载最重的存储服务器实行数据迁移，将数据迁移到存储负载最轻的存储服务器。

存储集群总工作负载的计算公式如下：

$$L = \sum_{i=1}^{N}[(u_i + d_i)/n_i] \tag{2}$$

其中，N表示存储集群中存储服务器的总数，u_i、d_i、n_i的意义同式(1)。

数据分布不平衡度可用存储集群中各存储服务器磁盘可用空间百分比的方差表示，计算公式如下：

$$U = \frac{\sum_{i=1}^{N}(a_i - \bar{a})^2}{N} \tag{3}$$

其中，a_i的意义同式(1)，N 的意义同式(2)，$\bar{a}$表示存储集群中各存储服务器磁盘可用空间百分比的数学期望。

$L_{Threshold}$ 通过采样评估的方法确定，具体方法是：主控服务器维护一个变量 L_{max}，该变量的初值为 0，保存主控服务器自启动以来存储集群总工作负载 L 的最大采样值。$L_{Threshold}=L_{Max}\times 20\%$ 。

$U_{Threshold}$ 通过样本统计的方法确定，具体方法是：首先选定某大型服务器上所有文件作为模拟样本，若以轮询（round-robin）的方式存入 N 台存储服务器，可计算出该情况下 N 台存储服务器的数据分布不平衡度 U_{RR}。$U_{Threshold}=U_{RR}\times 80\%$ 。

实验中，样本选择的是本校校园网 FTP 服务器上的所有文件，根据统计可知，该 FTP 服务器上共有 25 169 个文件夹，261 242 个文件，文件大小总和为 4.75TB。图 4 所示的为样本中文件大小的分布情况，从图中可看出，该组样本中最小文件为 0 字节，最大文件为 34.29GB（对应横坐标为 24.3），其中绝大部分文件的大小分布在e^5~e^{20}(e≈2.718)字节范围内。将这些文件以轮询的方式依次保存到 50 台存储服务器，每台存储服务器总容量为 250GB，各存储服务器的存储耗费情况如图 5 所示。经过计算可知，$U_{RR}=6.88$，$U_{Threshold}=5.504$。

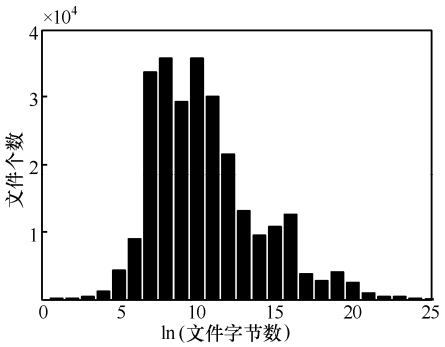


图 4　某 FTP 服务器文件大小分布

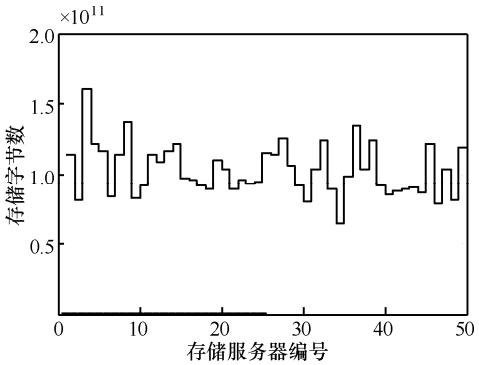


图 5　50 台存储服务器容量耗费态势

2.2.4　在线备份与自动恢复

在线备份与自动恢复机制解决了采用通用低性能 PC 集群提供云存储服务时的可靠性问题。

FFS 的数据自动恢复基于在线备份机制。在线备份是指 FFS 客户端在创建文件时，会同时在 2 个存储服务器上创建文件副本，其中一个为主文件，另一个为备份文件。为了保持 2 个副本的一致性，FFS 采用了 Read-one Write-all 策略[9]：由于文件读操作是安全性操作，只需要从主文件读取数据即可；对于文件写操作，客户端需要将数据同时写入所有副本，并获得所有副本的写成功回复，才判断为文件写操作成功。这种策略实现简单，而且尽可能保证了每一个副本在任何时刻都有最新数据，同时也减小了服务器负载，因为数据传播负担由服务器转移到了客户端，这反过来又提高了可扩展性。

在线备份机制保证了常态下每份保存在FFS中的文件都有 2 个完全一致的副本。当存储集群中某台存储服务器失效时，主控服务器通过心跳机制可及时发现该失效节点，并启动数据自动恢复过程。如图 6 所示，在主控服务器中维持着一个目录结构和一个存储服务器逻辑节点列表。

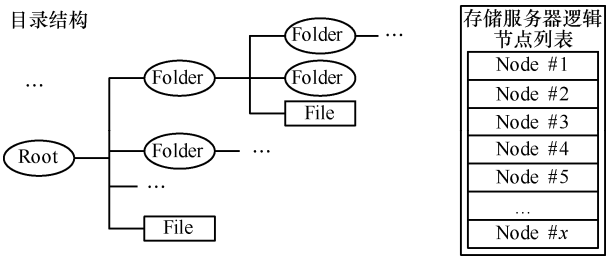


图 6　目录结构和存储服务器逻辑节点列表

目录结构是一种树形结构。在目录树中，内节点为子目录节点，保存该目录的路径信息、目录属性信息、创建日期信息；叶节点为空目录节点或文件节点，文件节点中保存着对应文件的路径信息、文件属性信息、创建/修改日期，以及该文件所对应的 2 台存储服务器的逻辑节点指针和其在存储服务器上对应的主文件和备份文件的 GUID。GUID 是文件被创建时由主控服务器分配的一个全局唯一编号，使用该编号在存储服务器上保存文件可避免同名文件冲突的问题。存储服务器逻辑节点中保存了对应存储服务器的 IP 地址与端口号、state 变量、CPU/

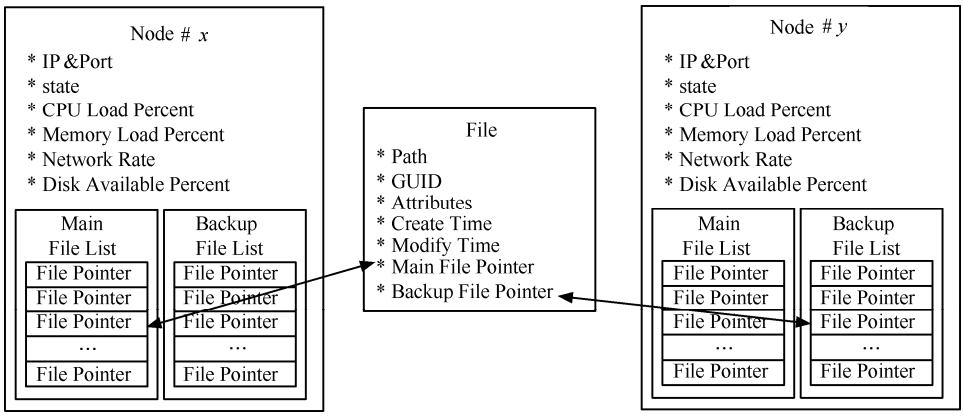


图7 文件节点与存储服务器逻辑节点

内存/网卡/磁盘状态，以及主文件指针列表与备份文件指针列表。图7为文件节点与存储服务器逻辑节点的数据视图。在文件节点中，通过主/备份文件指针可找到存放该文件的主/备份存储服务器逻辑节点；在存储服务器逻辑节点中，通过主/备份文件列表可找到目录结构中对应的文件节点。

当主控服务器发现某存储服务器掉线后，将启动数据恢复过程。数据恢复过程通过遍历该掉线存储服务器逻辑节点中保存的文件指针列表，找到目录结构中对应的文件节点。若掉线存储服务器为该文件的主存储服务器，则从文件叶节点中找其备份存储服务器，将其备份存储服务器升级为主存储服务器，并选择一台负载较轻的存储服务器作为该文件新的备份存储服务器，然后通知该新备份存储服务器创建该文件，并从该文件的当前主存储服务器读取文件重构备份；若掉线存储服务器为该文件的备份存储服务器，则选择一台负载较轻的存储服务器作为该文件新的备份存储服务器，通知该新的备份存储服务器创建该文件，并从该文件的主存储服务器读取文件重构备份。

2.2.5 AC-RU 缓存

AC-RU（adjust-controlled recently-used）缓存机制解决了云存储系统中使用传统缓存调度策略时出现的缓存序列频繁调整的问题，从而进一步提高系统性能。

如图2所示，AC-RU 缓存被应用于存储服务器模块和客户端代理模块。存储服务器的 AC-RU 缓存保存最近打开的若干个文件句柄，通过该缓存可避免文件句柄被频繁的打开和关闭，从而提高存储服务器响应数据请求的速度。客户端代理的 AC-RU 缓存保存最近访问的目录信息和文件的元数据信息，通过该缓存可避免客户端频繁地向主控服务器请求目录信息和文件元数据信息。

AC-RU 缓存调度算法如图8所示。若设定 AC-RU 缓存的最大容量为 N，则该缓存的前 N/2 使用 FIFO (first in first out)[10]调度策略，后 N/2 使用 LRU (least recently used)[10]调度策略。新对象加入缓存时，添加到缓存的头部，缓存中原有对象依次后移。若缓存已被占满时，加入新对象，则缓存尾部的对象被移出缓存。当缓存中已有对象被命中时，若该对象处于 FIFO 调度策略段，则不进行缓存序列的调整，若该对象已经处于 LRU 调度策略段，则将该对象调整到缓存的头部。由此可见，当命中的缓存对象处于前 N/2 缓存时，并不会调整缓存序列；当加入新对象或命中的缓存对象处于后 N/2 缓存时，调整缓存序列。通过 AC-RU 缓存技术，既保证了缓存具有 LRU 调度策略下较高的缓存命中率，又克服了单纯使用 LRU 调度策略时，频繁

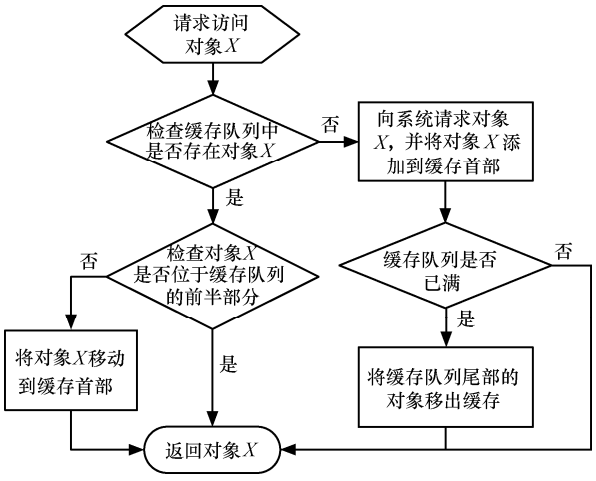


图8 AC-RU 缓存调度算法

调整缓存序列而带来的延迟问题。

2.2.6　客户端虚拟化磁盘

客户端虚拟磁盘机制屏蔽了客户端应用对 FFS 云存储系统的访问细节，使客户端应用可采用与访问传统磁盘一致的方法访问 FFS 云存储系统。

如图 9 所示，FFS 设计并支持从 Windows 客户端和 Linux 客户端以磁盘方式进行访问。

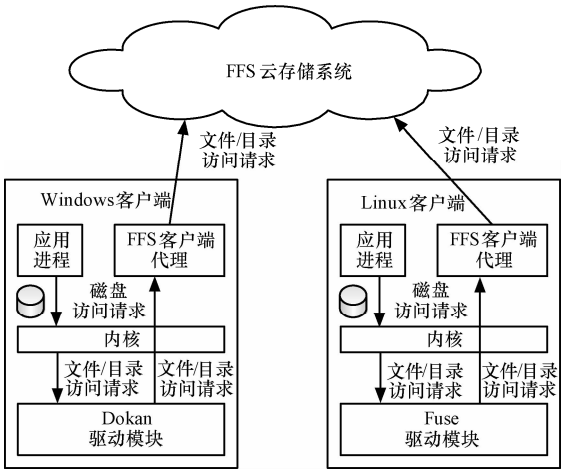


图 9　客户端虚拟磁盘机制

在 Linux 下，应用进程对云存储系统中文件的访问请求转化为对 Fuse 驱动模块的访问请求；在 Windows 下，转化为对 Dokan 驱动模块的访问请求。Fuse 是 Linux 下开发应用层文件系统的开源开发分组，提供实现 Fuse 的用户态目录与文件操作接口。Dokan 是 Windows 下类似于 Fuse 的一个开源开发分组。

2.2.7　主控服务器双机热备机制

FFS 云存储系统中文件数据的可靠性由在线备份与自动恢复机制保证，对于主控服务器的可靠性，则采用基于共享磁盘阵列的双机热备机制来保证，其部署方法如图 10 所示。

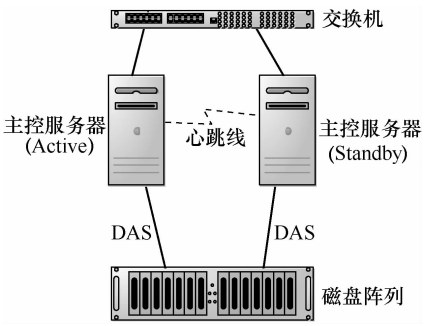


图 10　主控服务器双机热备机制

主控服务器模块部署于两台主控服务器上，其中一台处于活动态（Active），另一台处于待命态（Standby），它们之间使用一条心跳线（双机互联线）连接。活动态主控服务器将目录结构、各逻辑节点的文件列表以 XML 的方式保存在共享磁盘阵列中。待命态主控服务器通过心跳线侦测活动态主控服务器的工作状况。当活动态主控服务器出现故障时，待命态主控服务器立即启动主控服务器模块，从共享磁盘阵列中读取 XML 文件，重构目录结构与各逻辑节点，并接管原活动态主控服务器工作。

2.3　协议设计

FFS 通过网络传输的信息有两类，一类是信令，一类是数据。信令通过 UDP 报文传送，数据则通过 UDT（UDP-based data transfer）协议进行传输。这里所指的协议格式是指信令报文的格式。

如图 11 所示，FFS 传送的信令报文由前导码、信令类型和信令内容组成。

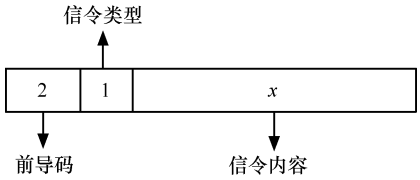


图 11　协议格式

前导码为一个 UInt16 类型数值，用来标识该报文为 FFS 信令报文。所有 FFS 信令报文具有相同的前导码。信令类型为一个 Byte 类型数值，FFS 各类信令依次编号，所有请求信令为奇数，对应的应答信令为下一相邻偶数。信令内容的长度并不固定，其与信令类型有关。

表 1 所示为 FFS 中主要的 19 种信令类型，表中描述了各信令以及应答报文的功能。

2.4　主要工作流程

2.4.1　主控服务器目录服务流程

目录服务包括存储容量查询、目录查询、文件查询、目录创建/删除、文件创建/删除等。目录删除需要递归地删除其子目录。图 12 为主控服务器的目录服务流程。

2.4.2　存储服务器数据交互流程

在 FFS 中，客户端采用被动读/写数据的方式。如图 13 所示，客户端读数据时，首先打开 UDT 接收端口，并通知存储服务器向该端口写入数据。

表1 信令类型

描述符	功能	应答
DEV_JOIN	存储服务器向主控服务器发送的注册报文	返回是否成功
HEARTBEAT	存储服务器向主控服务器发送心跳报文	无
GET_CHUNKSV	云存储客户端向主控服务器查询当前负载较轻的存储服务器	返回当前负载最轻的存储服务器地址
DISKSPACE	云存储客户端向主控服务器查询所磁盘空间	返回云存储系统总容量、已用容量
FOLD_EXIST	云存储客户端向主控服务器查询目录是否存在	返回是否存在
FILE_EXIST	云存储客户端向主控服务器查询文件是否存在	返回是否存在
UPDATE_SIZE	云存储客户端向主控服务器提交文件大小更新命令	返回是否成功
FOLDINFO	云存储客户端向主控服务器请求获取目录信息	返回所请求目录的目录信息
FILEINFO	云存储客户端向主控服务器请求文件信息	返回所请求文件的元数据信息
CREAT_FOLD	云存储客户端向主控服务器提交创建目录命令	返回是否创建成功
CREAT_FILE_M	云存储客户端向主控服务器提交创建文件命令	返回是否创建成功
CREAT_FILE_S	主控服务器向存储服务器提交创建文件命令	返回是否创建成功
DEL_FOLD	云存储客户端向主控服务器提交删除目录命令	返回是否删除成功
DEL_FILE_M	云存储客户端向主控服务器提交删除文件命令	返回是否删除成功
DEL_FILE_S	主控服务器向存储服务器提交删除文件命令	返回是否删除成功
MOV_FOLD	云存储客户端向主控服务器提交移动目录命令	返回是否移动成功
MOV_FILE	云存储客户端向主控服务器提交移动文件命令	返回是否移动成功
READ	云存储客户端向存储服务器请求读取数据	存储服务器通过UDT协议向客户端发送数据
WRITE	云存储客户端向存储服务器请求写数据	存储服务器通过UDT协议从客户端读取数据

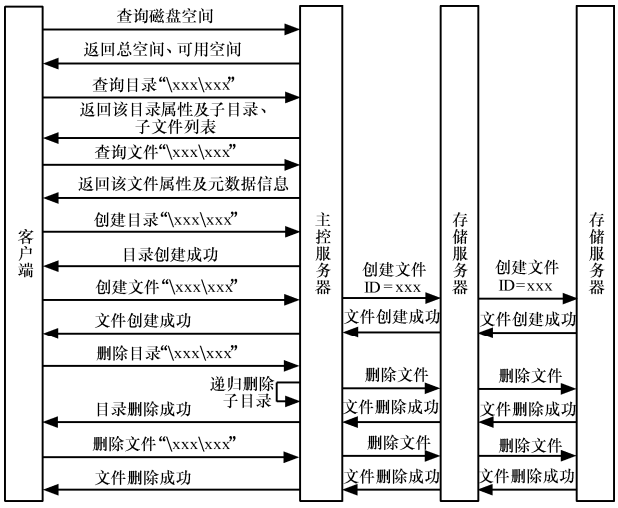


图12 主控服务器目录服务流程

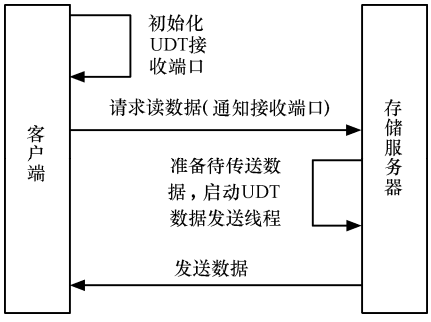


图13 读数据流程

如图14所示，客户端写数据时，将数据置于一个打开的UDT发送端口处，通知存储服务器从该端口取走数据。

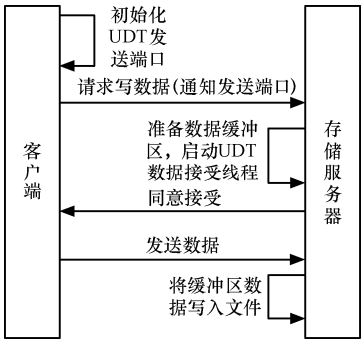


图14 写数据流程

3 性能测试与分析

对FFS云存储系统进行了扩展性测试和工作负载测试，测试环境包括一台主控服务器、16台存储服务器和16台客户机。主控服务器与存储服务器通过100M bit/s全双工网卡连接到HUAWEI Quidway S3000交换机，所有客户机连接到另一台同样配置的交换机上，两台交换机使用1G bit/s链路相连。它们的配置如表2所示。测试工具采用Intel

公司开发的通用文件系统基准测试程序：IOMeter。

表 2　　测试环境

角色	CPU	内存	硬盘	数量
主控服务器	Core2 2.6GHz	4GB	250GB	1
存储服务器	Core2 2.6GHz	2GB	2TB	16
客户端	Core2 2.6GHz	2GB	250GB	16

3.1　扩展性

图 15~图 17 是用 IOMeter 测得的不同数目存储服务器情况下聚合读写带宽的变化情况，存储服务器的数目分别为 2、9、16，客户端最多有 16 个。100M bit/s 链路最大带宽为 12.5M bit/s，1G bit/s 链路最大带宽为 125 M bit/s。从图中可看出，当存储服务器数为 2 时，系统理论最大读写带宽为 25M bit/s；当存储服务器数为 9 时，系统理论最大读写带宽为 112.5M bit/s；当存储服务器数为 16 时，系统理论最大读写带宽仅为 125M bit/s(因为受限于 1G bit/s 链路的最大带宽)。

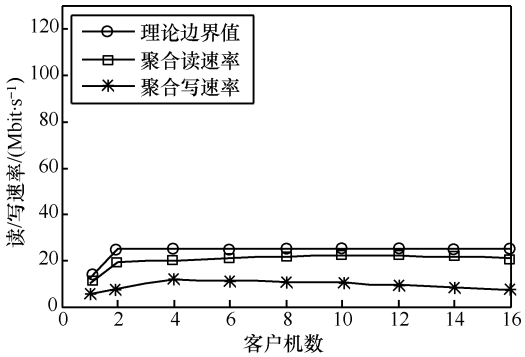


图 15　存储服务器数为 2 时的读/写性能

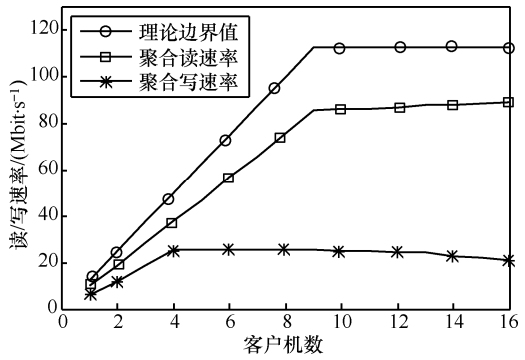


图 16　存储服务器数为 9 时的读/写性能

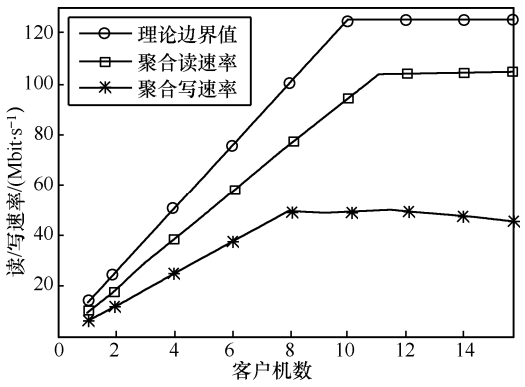


图 17　存储服务器数为 16 时的读/写性能

单台客户机对系统的读速率为 9.4M bit/s，随着客户机的增加，系统聚合读带宽趋近理论最大带宽的 75% 。单台客户机对系统的写速率为 6.1M bit/s，由于对系统的写操作包括写主文件和写备份文件，因此单台客户机的写速率约为理论最大带宽的 48% ；当客户机增加到存储服务器数目的一半时，系统聚合写带宽达到最大值；随着客户机数目的进一步增加，系统聚合写带宽反而有所下降，这是因为存在多台客户机对同一个存储服务器并行写数据，导致该存储服务器磁盘频繁调度，从而影响系统整体性能。

图 18 显示存储服务器数为 2、9、16 时所测得的系统最大聚合读写带宽。2 个存储服务器的最大读带宽为 21.9M bit/s，最大写带宽为 11.7M bit/s；9 个存储服务器的最大读带宽为 88.9M bit/s，最大写带宽为 25.6M bit/s；16 个存储服务器的最大读带宽为 105M bit/s，最大写带宽为 49.4M bit/s。通过提高链路带宽或增加存储服务器数目可进一步增加系统聚合读写带宽。

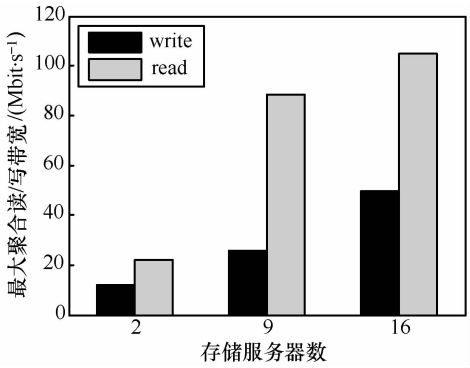


图 18　最大读/写带宽的变化

3.2　工作负载状况

可通过加入新存储服务器和已存在存储服务器失效时系统中各节点工作负载的变化来观测 FFS 的工作负载均衡能力。使用 4 台客户机进行负载生成，各客户机每秒向系统写入一个 2.5M B 的文件。各存储服务器上启动一个收集网卡负载的程序，以周期为 1s 的速度记录网卡负载。

图19为加入新存储服务器过程中存储集群内各节点的负载状况。横坐标为时间，单位为s；纵坐标为网卡负载百分比（单位为%）。初始时，存储集群中只有节点A和B；在60s时节点C加入，原来由A和B承担的工作负载被平分至3个节点；在120s时节点D接入，4个节点比较均衡地分担了系统总负载。

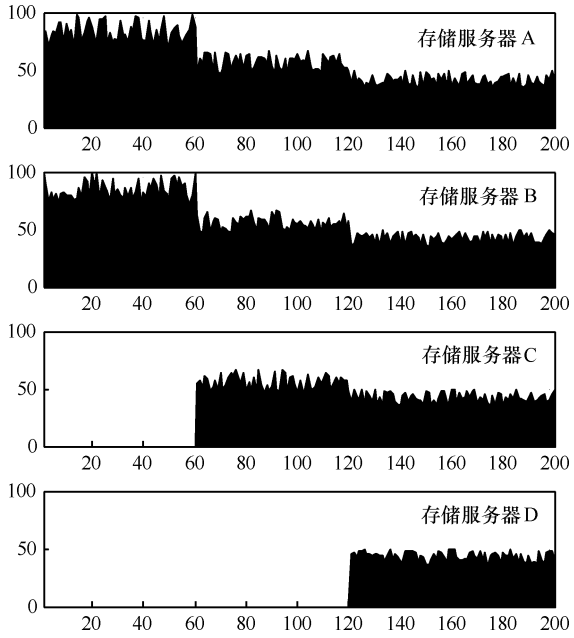


图19　加入存储服务器过程中各节点工作负载状况

图20为存储服务器离线过程中存储集群内各节点的负载状况。初始时，存储集群中有4个节点A、B、C和D，它们比较均衡地分担了系统总负载；

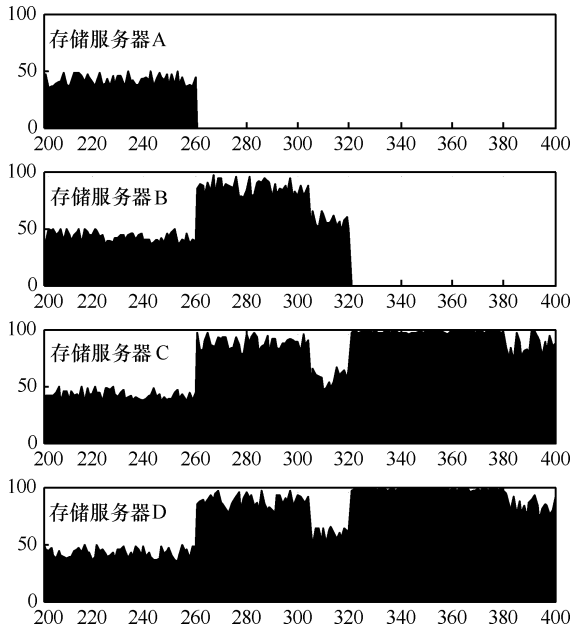


图20　存储服务器离开过程中各节点工作负载状况

在260s时节点A离线，系统工作负载将由剩余3个节点分担；同时，由于A的离线将导致系统启动数据恢复工作，这进一步增大了系统的工作负载；在300s左右，数据恢复工作完毕，系统负载有所下降；在320s，节点B离线，系统工作负载将由C、D分摊。

图19和20反映了FFS具有良好的动态负载均衡性能，同时图19也反映了在存储服务器掉线后，数据自动恢复过程会占用系统一部分资源。

4　FFS的特征和优势

传统海量存储通常采用DAS/NAS/SAN架构，这类架构依赖于昂贵的专用存储设备，且存在扩展困难、管理复杂等问题。本文所设计的云存储系统通过FFS将集群中大量廉价通用的存储设备协同起来，共同对上层应用提供海量存储服务，并通过虚拟化机制在FFS客户端将云存储服务虚拟成符合标准文件访问协议的存储设备。使用该云存储系统的客户端通常是有海量存储需求的应用服务器，包括邮箱服务器、FTP服务器、Web服务器或数据库服务器等。相对于传统海量存储架构，FFS云存储系统具备如下几个特征。

1）海量并行扩容

传统的存储采用串行扩容，不管其接多少扩展箱，总有个极限。FFS采用的架构是并行扩容，几乎没有扩展限制。这是效用计算理念的体现，是云存储区别于传统存储最重要的特征。投资者不必担心对存储的前期过度投资，当需要进行容量扩充时，再向存储集群中添加新的存储设备即可。

2）自动故障切换与数据恢复

硬件损坏会导致服务的停止和数据的丢失。在传统的存储管理中，管理人员通过建立一个全冗余的环境（电源、网络、盘阵等）应对硬件失效，但是这样的成本太高，而且管理工作繁复。在FFS云存储系统中，每个文件都有若干份备份，监控系统通过心跳机制能及时发现意外的硬件损坏，并自动进行故障切换和数据恢复，保证服务的连续性和数据的可靠性。

3）动态负载均衡

传统的存储管理异常复杂，数据中心管理人员常要面对不同的存储产品，而这些存储产品没有一个统一的管理界面，管理员需要了解每个存储设备的使用状况（容量、负载等），并对存储结构进行

人工调整。在 FFS 云存储系统中，系统会自动进行数据迁移与负载均衡工作，管理员只需在整体存储容量快用完时采购新的存储设备即可。

4）支持标准文件访问协议

用户访问 FFS 云存储服务的方式与访问本地硬盘相同，FFS 通过虚拟化技术对客户提供符合标准文件访问协议的客户端虚拟磁盘，从而无需更改现有应用程序访问存储的方式。

5）低成本高可靠

传统存储的可靠性由昂贵的专用存储硬件保证，从而导致存储系统的成本很高。FFS 云存储系统基于通用硬件构建，其可靠性由文件副本机制、自动故障切换和数据恢复机制保证，成本较低。

5　结束语

FFS 是一个基于网络的 PB 级通用云存储系统，通过海量并行扩容机制解决了传统海量存储系统扩容困难的问题，通过集群监控机制解决了传统海量存储系统管理复杂的问题，通过动态负载均衡机制解决了存储系统面对大量并行访问时的任务分配问题，通过在线备份与自动恢复机制解决了采用通用低性能 PC 集群提供云存储系统时的可靠性问题，通过 AC-RU 缓存机制进一步提高了系统的性能，通过客户端虚拟磁盘技术屏蔽客户端与云存储系统的交互细节，通过主控服务器双机热备机制解决了主控服务器单节点故障问题。

FFS 在如下几个方面还有待进一步完善：①目前 FFS 中的文件副本数是固定的，可将其改进为按统计的热度分配副本数，即为读热度高的文件多做副本，为写热度高的文件少做副本，从而减少为保证数据一致性所付出的代价；②当 FFS 存储集群规模达到一定程度后，其总功耗将成为进一步扩展的制约因素，将来可对 FFS 增加动态休眠功能，在不影响在线数据完整性的情况下动态休眠部分节点，需要访问休眠节点时，再次唤醒它们，从而达到降低功耗的目的。

参考文献：

[1] GRAY J. What Next? a few remaining problems in information technology[EB/OL]. http://research.microsoft.com/~gray/talks/Gray_Turing_FCRC.pdf, 1998.

[2] PRESTON W C. Using SANs and NAS[M]. Sebastopol, CA: O Reilly Media, 2002.

[3] ZITO C. Closing critical it gaps and driving towards converged infrastructure with storage virtualization[EB/OL]. http://h18006.www1.hp.com/storage/4AA0-2062ENW, 2009.

[4] DEELMAN J G, VAHIE, et al. Data sharing options for scientific workflows on amazon EC2[A]. 2010 International Conference for High Performance Computing, Networking, Storage and Analysis(SC)[C]. New Orleans, LA, 2010.1-4.

[5] GHEMAWAT S, GOBIOFF H, LEUNG S T. The Google file system[A]. Proceedings of the 19th ACM Symposium on Operating Systems Principles[C]. Bolton Landing, NY, 2003.29-43.

[6] WHITE T. Hadoop: the Definitive Guide[M]. Sebastopol, CA: O Reilly Media, 2009.

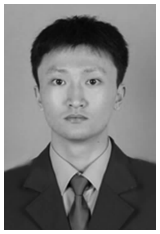
[7] 楚才. TFS 简介[EB/OL]. http://rdc.taobao.com/blog/cs, 2010.
CHU C. TFS Introduction[EB/OL]. http://rdc.taobao.com/blog/cs, 2010.

[8] 叶伟等. 互联网时代的软件革命：SaaS 架构设计[M]. 北京：电子工业出版社, 2010.
YE W, et al. The Software Revolution during Internet Ear: the Design of SaaS Infrastructure[M]. Beijing: Publishing House of Electronics Industry, 2010.

[9] RABINOVICH M, LAZOWSKA E D. An efficient and highly available read-one write-all protocol for replicated data management[A]. Proceedings of the Second International Conference on Parallel and Distributed Information Systems[C]. San Diego, CA, 1993.56-65.

[10] 张艳, 石磊, 卫琳. Web 缓存优化模型研究[J]. 计算机工程, 2009, 35(8): 85-87.
ZHANG Y, SHI L, WEI L. Study on optimal model of Web cache[J]. Computer Engineering, 2009, 35(8): 85-87.

作者简介：



吴海佳（1986-），男，江苏南通人，解放军理工大学博士生，主要研究方向为分布式计算、云计算、网络存储。

陈卫卫（1967-），女，四川隆昌人，解放军理工大学教授、硕士生导师，主要研究方向为计算机算法、软件工程、云计算等。

胡谷雨（1963-），男，浙江东阳人，博士，解放军理工大学教授、博士生导师，主要研究方向为网络管理、云计算等。

董继光（1986-），男，河南周口人，解放军理工大学硕士生，主要研究方向为分布式存储、云计算。