Software Engineering and Applications 软件工程与应用, 2016, 5(3), 198-203
Published Online June 2016 in Hans. http://www.hanspub.org/journal/sea
http://dx.doi.org/10.12677/sea.2016.53022

Hans汉斯

Research and Implementation of Distributed File Storage System Based on TFS

Hairu Zhang

Tianjin Tonglian Electric Co. Ltd, Tianjin
Email: 2401079597@qq.com

Received: May 28th, 2016; accepted: Jun. 12th, 2016; published: Jun. 15th, 2016

Abstract

With the rapid development of Internet, the explosive growth of unstructured data such as text, audio, pictures and other data has caused the technology of distributed file storage to develop rapidly. The efficient distributed storage system is designed. And the distributed file storage scheme based on TFS (Taobao File System) is put forward, including the web platform, core data cache cluster, file management server cluster, database and TFS data storage cluster.

Keywords

Distributed System, File Storage, TFS

基于TFS的分布式文件存储平台研究与实现

张海茹

天津通联电气有限公司，天津
Email: 2401079597@qq.com

收稿日期：2016年5月28日；录用日期：2016年6月12日；发布日期：2016年6月15日

摘　要

随着互联网的迅猛发展，以文本、音频、图片等数据为代表的非结构化数据呈现爆炸式增长，分布式文

文章引用: 张海茹. 基于 TFS 的分布式文件存储平台研究与实现[J]. 软件工程与应用, 2016, 5(3): 198-203.
http://dx.doi.org/10.12677/sea.2016.53022

件存储技术应运而生。论文设计高效的分布式存储体系，提出基于TFS (Taobao File System)的分布式文件存储方案，包括平台Web端、核心数据缓存集群、文件管理服务集群、数据库和TFS数据存储集群。

关键词

分布式系统，文件存储，TFS

1. 引言

随着互联网迅速发展，用户创造了大量的小文件，一般文件大小小于 64 MB，这些文件如何存储，如何分析，成为当前一个重要问题。

Google 设计了 GFS (GoogleFile System)用于存储庞大数据，HDFS [1] (Hadoop File System)是 GFS 的一个简化版实现，二者都采用单一主控机 + 多台工作机的模式，通过数据分块和复制来提供更高的可靠性和更高的性能。但是根据文献[2] [3]可知，访问大量小文件的速度远小于访问少量大文件的速度，效率较低。

淘宝自行研发文件的存储框架(TFS) [4]，其高效率、高容错性、高并发性使得它能够处理海量的小文件存储。但是对于中小企业来所，TFS 搭建，存在浪费时间、金钱和精力的问题，不能适应现在高速发展的互联网行业。

为了更快的搭建分布式存储体系，提高小文件存储效率，本文研究并实现了基于 TFS 的文件存储平台，实现对文件的管理，主要包括：文件上传，文件下载，查看文件细节等，解决负载均衡的问题，同时保证了用户文件的安全性、保密性和隐私性。

2. TFS 文件存储技术

TFS 的设计初衷是为了解决淘宝网站海量小文件的存储问题。TFS 文件存储系统是一个分布式的文件系统，分为两部分 NameSever 节点和 DataServer 节点，可以运行在同一台服务器上，但是 TFS 一般搭建集群，由两个 Nameserver 节点和若干个 DataServer 节点组成，两个 NameServer 节点分别为一主一备，提高系统的安全性。TFS 文件存储结构也采用了块的概念，TFS 文件存储系统中的块的大小默认 64 M，但是可以根据需求更改配置项更改块的大小。NameServer 负责 DataServer 的状态管理，并维护块与 DataServer 的映射关系。NameServer 不负责实际数据的读写，实际数据的读写由 DataServer 完成。

TFS 能够完成用户自定义文件名存取小文件，但是仍然存在一些问题，TFS 在执行读写操作时，读写文件低效；TFS 的 NameServer 记录了所有块的信息，如果失效，则必须重建块信息和映射关系[5]。为了克服这些问题，本文在 TFS 基础上，设计文件存储集群负责文件的基本操作，实现均衡负载，保证数据安全与一致性，并设置核心数据缓存集群，提高文件读写效率。

3. 系统分析与设计

3.1. 系统分析

基于 TFS 的文件存储系统是针对中小文件的安全和一致性的读写，设计的轻量级的文件存储平台，主要实现对文件的管理，主要包括：文件上传，文件下载，查看文件细节等，解决负载均衡的问题，同时保证了用户文件的安全性、保密性和隐私性。

系统运行在普通计算机组成的集群上，而不是运行在高性能的专业服务器上，达到减低存储成本，

同时便于系统扩展。该平台为程序开发者提供一个存储文件的服务，减少开发者工作量，同时减少开发者对文件安全性的担忧，减少开发者对文件的空间的担忧。

3.2. 系统架构设计

基于 TFS 文件存储系统平台架构采用的是对等式结构，主要是文件存储服务器集群和 TFS 文件存储集群。文件存储集群负责文件的基本操作，并实现均衡负载，TFS 文件存储集群主要是负责文件存储，TFS 文件存储集群有多个 TFS 文件存储节点组成，一个节点可以由单台服务器也可以由多台服务器共同组成，其中 TFS 文件存储节点均有备份，同时各个节点互相独立，增强了系统的容错性和抗压能力。其整体结构如图 1。

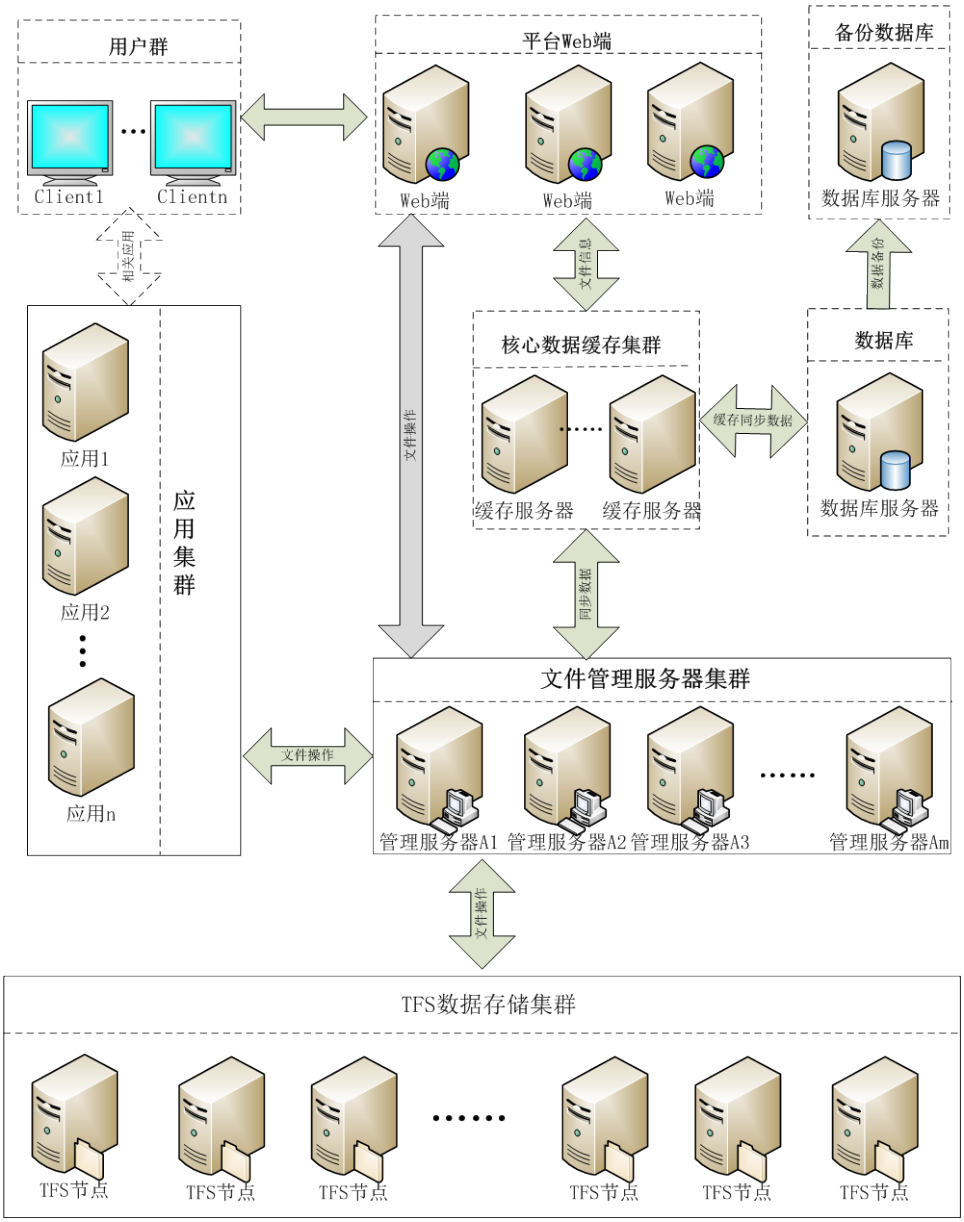

Figure 1. System architecture
图 1. 系统架构

系统组件主要有五个：平台 Web 端，核心数据缓存集群，文件管理服务集群，数据库，TFS 数据存储集群。

1) 平台 Web 端

平台 Web 端采用 HTML + CSS + JAVASCRIPT 实现，主要用到 java 语言进行开发，采用 SSH 框架。平台 Web 端服务器主要是给平台注册的用户提供文件存储服务。平台 Web 的功能主要包括用户信息的管理，用户创建的应用的管理，各应用的文件管理，其中文件管理包括上传、下载、移动位置、删除等功能。

平台 Web 端是本系统的推广的关键所在，为了增强用户的体验效果，提高网站的速度，将文件管理服务分离，并添加缓存服务增强访问效果，同时 web 端采用了两台服务器共同搭建，负载均衡，从而提高效率。

2) 核心数据缓存集群

核心数据缓存集群服务[6]是本系统的核心服务，采用 Memcache 缓存服务技术进行搭建。

Memcache 缓存在业界内应用很广泛，是一个分布式的内存对象缓存系统，该系统在内存中维护一张 Hash 表，该 Hash 表中可以存储图片，视频，还有从数据库中检索的数据等。由上可见，其提高读取的速度，主要是将数据放到内存中，然后各个应用从内存中读取数据。Memcached 采用的是守护程序的方式，并若干个服务器上运行，实时监测连接等操作。现在许多动态的网站常常为了缓解数据的压力，常常采用 Memcached 缓存技术，将常用的数据预先加载到内存，通过唯一标识符在缓存中获得数据，降低读取数据库的次数，从而提高网站的速度。

Memcache 缓存服务集群主要是解决文件上传下载速度慢的问题，通过减少数据库的访问，提高获得文件上传下载的地址的速度。

3) 数据库服务

数据库服务由两台服务器搭建，一主一副，主数据库面向应用，副数据库与主服务器数据同步，作为主数据库的备份，同时数据库中的数据定期做冷备份。

数据服务主要采用 Mysql 数据库，主要是因为 Mysql 数据库是免费的数据库，没有 Oracle 和 SqlServer 数据库的版权问题；其次，Mysql 数据库是一个快速的，多线程的、多用户和健壮的 SQL 数据库服务器。MySql 服务器支持关键任务、重负载生产系统的使用；再次，Mysql 是开源的，意味着任何人都可以使用和修改该软件，方便开发者根据项目需求更改 Mysql 底层代码的实现，从而更加高效的获取数据。

4) 文件管理服务集群

文件管理服务是本系统的核心，极有可能成为系统的瓶颈，因此采用多个服务器搭建集群，实现负载均衡，从而解决瓶颈问题。文件管理服务与三方面进行交互,分别是用户的 Applications，核心缓存服务和 TFS 数据存储。文件管理服务给用户 Applications 提供文件操作所需要的接口，这些接口包括文件上传、文件下载、文件更改、文件删除等接口，从而能够很好的为用户 Applications 提供健全的文件存储服务；文件管理服务不与数据库进行直接接触，而是将需要保存的数据同步到缓存服务集群上，然后数据缓存服务定时将数据同步到数据服务上，从而达到异步存储的目的，减轻了文件管理服务的数据处理的压力。文件管理服务与 TFS 文件存储服务交互，将上传的文件存储到 TFS 文件存储服务中，这种设计将用户的 Applications 的文件操作与 TFS 文件存储服务隔离开，极大的保证了 TFS 文件存储的安全性。

5) TFS 文件存储服务集群

TFS 文件存储服务集群由多台 TFS 文件存储节点组成，采用 TFS 文件存储节点，主要是因为 TFS 的容错性高，抗压能力强，同时增大存储空间，搭建集群增加备份点，提高安全性。

Table 1. Platform deployment plan
表 1. 平台部署方案

服务名称	CPU (GHZ)	内存 (G)	硬盘 (G)	网卡 (Mps)	操作系统	服务器数量
Web 端	2.7	4	100	1000	Centos6.4	2
缓存服务	2.7	8	100	100	Centos6.4	2
数据库服务	2.7	4	500	100	Centos6.4	2
文件管理服务	2.7	4	100	1000	Centos6.4	3
TFS 文件存储服务	2.7	4	1000	1000	RHEL 5	2

Table 2. Latency of 100K files
表 2. 10 万个文件延迟对比

存储系统名称	读延迟/μs	写延迟/μs
HDFS 文件存储	9082	92,214
TFS 文件存储	7876	72,865
本文件存储	2642	20,128

4. 平台部署与实验分析

文件存储平台部署比较复杂，需要部署五个服务，分别是 Web 端服务、缓存服务，数据库服务、文件管理服务和 TFS 文件存储系统服务。本平台涉及到的服务比较多，而且再部署的时候还要考虑到负载均衡和文件备份等，需要较多的服务器，为了完成平台的部署，采用某些虚拟服务器。系统的具体部署如表 1。

由表 1 可以看到，各个服务的服务器配置有些许差异，这是根据每个服务的特点进行的配置。例如 Web 端、问价管理服务和 TFS 文件存储服务，对网络要求较高，所以网卡采用的是千兆网卡；缓存服务的承载量与其所在的服务器的内存成正比，所以分配的服务器的内存较高；数据库和 TFS 文件存储服务对硬盘要求比较高，硬盘存储量设定的较高。

平台搭建之后，在本科毕业设计管理系统中进行了实际应用，和之前的文件系统相比，如表 2 所示，能够较快的实现文件的上传、下载，运行效果良好。

5. 结论

由于现在大部分文件存储服务不能够满足中小企业和许多网站的需求，但是自己搭建文件存储服务，对中小企业和许多网站是一个很大的负担。本文从需求和性能上考虑，研究并设计了分布式文件平台，包括了平台 Web 端，核心数据缓存集群，文件管理服务集群，数据库和 TFS 数据存储集群。实践证明，文件存储实现了均衡负载，保证数据安全与一致性，提高文件读写效率。

参考文献 (References)

[1] 马登邑. 基于 Hadoop 存储的文件管理系统的研究与实现[D]: [硕士学位论文]. 武汉: 华中科技大学, 2013.

[2] 王铃惠, 李小勇, 张轶彬. 海量小文件存储文件系统研究综述[J]. 计算机应用与软件, 2012, 29(8): 106-109.

[3] Dong, X.C. (2015) Hadoop HDFS. http://dongxicheng.org

[4] 刘伯睿. 海量数据小文件分布式存储系统的设计与实现[D]: [硕士学位论文]. 长沙: 湖南大学, 2013.

[5] 李洪奇，朱丽萍，孙国玉，等. 面向海量小文件的分布式存储系统设计与实现[J]. 计算机工程与设计，2016(1): 86-92.

[6] 秦秀磊，张文博，魏峻，等. 云计算环境下分布式缓存技术的现状与挑战[J]. 软件学报, 2013, 24(1): 50-66.