**Software Engineering and Applications** 软件工程与应用, 2017, 6(2), 21-27
Published Online April 2017 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2017.62003

Hans汉斯

# A Survey of Distributed File System

**Zhennan Du, Chongjun Zhu**

College of Computer Science, National University of Defense Technology, Changsha Hunan
Email: 121422193@qq.com

Received: Mar. 27th, 2017; accepted: Apr. 10th, 2017; published: Apr. 14th, 2017

## Abstract

**The file system is an important part of the computer system. With the development of personal computer and network technology, the generation of distributed file system has effectively solved the problem of infinite growth of massive information storage. This paper summarizes the origin of the distributed file system and comprehensively analyzes and organizes the architecture of several typical distributed file systems by consulting a large number of documents.**

## Keywords

**Distributed File System, Storage Technology**

# 分布式文件系统综述

**杜振南，朱崇军**

国防科学技术大学计算机学院，湖南 长沙
Email: 121422193@qq.com

收稿日期：2017年3月27日；录用日期：2017年4月10日；发布日期：2017年4月14日

## 摘　要

**文件系统是计算机系统的重要组成部分，随着个人计算机和网络技术的发展，分布式文件系统的产生有效解决了无限增长的海量信息存储问题。本文通过查阅大量文献，概述了分布式文件系统的起源并综合分析整理了几种典型分布式文件系统的体系架构。**

## 关键词

**分布式文件系统，存储技术**

文章引用: 杜振南, 朱崇军. 分布式文件系统综述[J]. 软件工程与应用, 2017, 6(2): 21-27.
https://doi.org/10.12677/sea.2017.62003

## 1. 引言

随着科学技术的发展以及云计算、P2P 等技术的普及，全球数据量呈现爆炸式的增长，尤其是大数据时代的到来，通过互联网，用户制造了海量的数据。2017 年 1 月 22 日中国互联网络信息中心发布的《第 39 次中国互联网络发展状况统计报告》中指出：我国 2016 年全年共计新增网民 4299 万人，增长率为 6.2%，其中，手机网民规模达 6.95 亿，占比达 95.1%，增长率连续 3 年超过 10%。手机网民最常使用即时通信 APP：2016 年，网民在手机端最经常使用的 APP 应用前三位分别是微信、QQ、淘宝，无论是微信、QQ、微博等社交通信软件还是淘宝、京东等电商软件，其图片的上传、分享与展示所产生的数据量都达到了指数级。传统的存储系统无法满足呈爆炸性增长的海量数据存储需求，为解决信息存储容量、数据备份、数据安全等问题，分布式文件系统应运而生，如今已得到广泛应用。使用分布式文件系统时，用户不必考虑底层的存储设备以及实现细节，系统会将用户的数据进行存储、归档、备份，实现对数据的使用、共享以及保护的目的。

本文阐述了分布式文件系统的概念和发展历程，结合近年来分布式文件系统的应用情况，对几种典型分布式文件系统的概念、特点、体系架构进行研究，旨在帮助学习研究人员进一步了解分布式文件系统。

## 2. 分布式文件系统概述

本地文件系统只能访问与主机通过 I/O 总线直接相连的磁盘上的数据。当局域网出现后，各台主机间通过网络互连起来。如果每台主机上都保存一份大家都需要的文件，既浪费存储资源，又不容易保持文件的一致性。于是就提出文件共享的需求，即一台主机需要访问其它主机的磁盘。这直接导致了分布式文件系统的诞生。

### 2.1. 分布式文件系统的概念

分布式文件系统(Distributed File System，DFS)是指文件系统管理的物理存储资源不一定直接连接在本地节点上，而是通过计算机网络与节点相连文件系统管理的物理存储资源。分布式文件系统基于客户机/服务器(C/S)模式而设计，通常一个网络内可能包括多个可供用户访问存储资源的服务器。同时，分布式文件系统的对等特性也允许一些系统在扮演客户端的同时扮演服务端。例如，用户可以发布一个允许其他客户机访问的目录，一旦被访问，这个目录对于其他客户机来说就像使用本地驱动器一样。

### 2.2. 分布式文件系统发展历程

分布式文件系统的发展主要经历了四个阶段[1]：

第一代分布式文件系统(1980~1990)

早期的分布式文件系统一般以提供标准接口的远程文件访问为目的，更多地关注访问的性能和数据的可靠性。早期的文件系统以 NFS (Network File System)和 AFS (Andrew File System)最具代表性，它们对以后的文件系统设计也具有十分重要的影响。

第二代分布式文件系统(1990~1995)

这一时期的分布式文件系统主要需求是广域网和大容量。XFS (Extended File System)、Tiger Shark 并

行文件系统及 Frangipani 等分布式文件系统应运而生。其中，XFS 借鉴了当时对称多处理器的设计思想，解决了广域网上缓存和减少网络流量的难题。后来出现的 Tiger Shark 文件系统则是专门针对规模比较大的多媒体应用。它做的创新主要集中在预留资源和针对资源优化的调度策略，保证了访问的高性能。

第三代分布式文件系统(1995~2000)

这一阶段，网络技术的发展和普及极大地推动了分布式文件系统的研究与应用，出现了许多优秀的分布式文件系统，如 General Parallel File System (GPFS)等。数据容量、性能和共享的需求使得这一时期的分布式文件系统管理的系统规模更庞大、系统更复杂，更多的先进技术也得以应用到系统中实现，如分布式锁、缓存管理技术、Soft Updates 技术、文件级的负载平衡等。

第四代分布式文件系统(2000 年后)

随着 SAN (Storage Area Network and SAN Protocols)和 NAS (Network Attached Storage)两种体系结构逐渐成熟，研究人员开始考虑如何将两种体系结构结合起来，以充分利用两者的优势。另一方面，基于多种分布式文件系统的研究成果，人们对体系结构的认识不断深入，网格的研究成果等也推动了分布式文件系统体系结构的发展。各类应用对于分布式文件系统的要求也越来越高：如大容量、高性能、可扩展性、高可用性、可管理性等。

### 2.3. 分布式文件系统的优势

相对于传统存储方式，分布式文件系统具备如下优势：

一是节约成本。分布式文件系统使用大量廉价的设备存储数据，对于企业，减少了购买昂贵存储服务器的成本，分布式文件存储技术的应用，使得企业对设备的维护以及管理成本大幅度降低。

二是方便管理。分布式文件系统在设计时就考虑了数据的管理，特别是海量数据的管理，通过使用虚拟化技术，可以方便的完成数据的备份以及迁移等操作。

三是扩展性好。支持线性扩容，当存储空间不足时，可以采用热插拔的方式增加存储设备，扩展方便。

四是可靠性强。分布式文件系统包含冗余机制，自动对数据实行备份，在数据发生损坏或丢失的情况下，可以迅速恢复。

五是可用性好。用户只需要拥有网络就可以随时随地的访问数据，不受设备、地点的限制。

## 3. 典型分布式文件系统介绍

### 3.1. Google 文件系统

Google 文件系统(Google File System，GFS)是 Google 公司为了存储海量搜索数据而设计开发的面向搜索引擎的分布式文件系统，为大量用户提供高可靠、高性能、良好扩展的数据存储服务[2]。主要用于大型的、分布式的、需要对大量数据进行访问的应用。它对硬件要求不高，只需运行在廉价的普通硬件上即可为大量用户提供总体性能较高的服务。在 Google 搜索引擎中，最大的分布式存储系统集群中已经广泛的在 Google 内部进行部署，能够同时支持数百个客户端的访问，是处理整个 WEB 范围内难题的一个重要工具。

#### GFS 架构

一个 GFS 由一个 master 和多个 chunk servers 组成，并且能够被多个客户访问。master 在 GFS 中处于中心位置，存储整个文件系统的元数据，包括命名空间、访问控制信息、文件到 chunk 的映射信息以及任意 chunk 的位置信息。文件被 master 分成了固定大小的 chunk，chunk servers 在本地的磁盘上存储

chunk。为了保证数据的可靠性，GFS 中实行冗余数据机制，对于每个 chunk，存储 3 份冗余数据，这样即使其中某一个 chunk servers 死机后，仍可通过访问其他 chunk servers 来获得数据。master 可通过心跳消息与各个 chunk servers 保持同步，侦听各个 chunk server 的状态，并获取 chunk 位置信息。图 1 是 GFS 的基本架构。

### 3.2. Lustre

Lustre 是一个大规模的、安全可靠的，具备高可用性的集群文件系统，Lustre 名字是由 Linux 和 Clusters 演化而来，它是由 SUN 公司开发和维护。该项目主要的目的就是开发下一代的集群文件系统，可以支持超过 10000 个节点，数以 PB 的数量存储系统。因为其超强的计算能力和开源，所有通常用于超级计算机，Top 100 的 60%超级计算机都使用该文件系统[3]。

#### Lustre 架构

Lustre 是一个面向对象的文件系统。主要由三个部分构成：元数据服务器(Metadata Servers，MDS)，对象存储服务器(Object-Based Storage Servers，OSSs)和客户端，Lustre 使用块设备来作为文件数据和元数据的存储介质，每个块设备只能由一个 Lustre 服务管理。Lustre 文件系统的容量是所有单个 OST 的容量之和。客户端通过 POSIX I/O 系统调用来并行访问和使用数据。图 2 是 Lustre 的基本架构。

### 3.3. Hadoop 分布式文件系统

Hadoop 是一个分布式系统基础架构，由 Apache 基金会所开发[4]。用户可以在不了解分布式底层细节的情况下，开发分布式程序。充分利用集群的威力高速运算和存储。Hadoop 的一个核心子项目就是 Hadoop 分布式文件系统(Hadoop Distributed File System，HDFS)是一个基于 Java 的文件系统。HDFS 有高容错性的特点，并且设计用来部署在低廉的硬件上；而且它提供高传输率来访问应用程序的数据，适合那些有着超大数据集的应用程序。HDFS 放宽了 POSIX 的要求，可以流的形式访问文件系统中的数据。

#### HDFS 架构

HDFS 文件系统采用主从系统结构[5]，一个提供元数据服务的 Namenode 节点和若干个提供数据存储的 Datanode 节点构成一个 HDFS 集群，分布式文件系统将元数据和应用数据分别存储在 Namenode 及 Datanode 中，各个服务器之间通过 RPC 协议进行通信。HDFS 命令空间是分层次的文件及目录结构，Namenode 维护命名空间树及文件块与 Datanode 之间的映射文件，Datanode 负责存储实际的数据。Datanode 与 Namenode 之间通过心跳信息进行通信，在通信中报告自己的运行状况。图 3 是 HDFS 的基本架构。

### 3.4. Fast 分布式文件系统

Fast 分布式文件系统(Fast Distributed File System, FastDFS)，是由余庆根据 MogileFS 的设计思想改进而来的类 GFS“键值对”的轻量级开源分布式文件系统，目前只支持 LinuX 等 UNIX 系统。FastDFS 是专门针对互联网应用(存储海量小文件)量身定做的系统，其主要功能包括：文件的存储、文件的同步、文件的访问等。它通过 C 语言实现，为文件的存取访问设计了专用的应用程序接口，因此可以为其定制客户端，目前客户端已经有了 PHP、Python、JAVA 等版本，已经有多家公司使用 FastDFS 来搭建存储平台及应用，如 UC、支付宝、赶集网等大容量存储应用。

#### FastDFS 架构

FastDFS 中只有两个角色，跟踪服务器 Tracker Server 和存储服务器 Storage Server [6]。

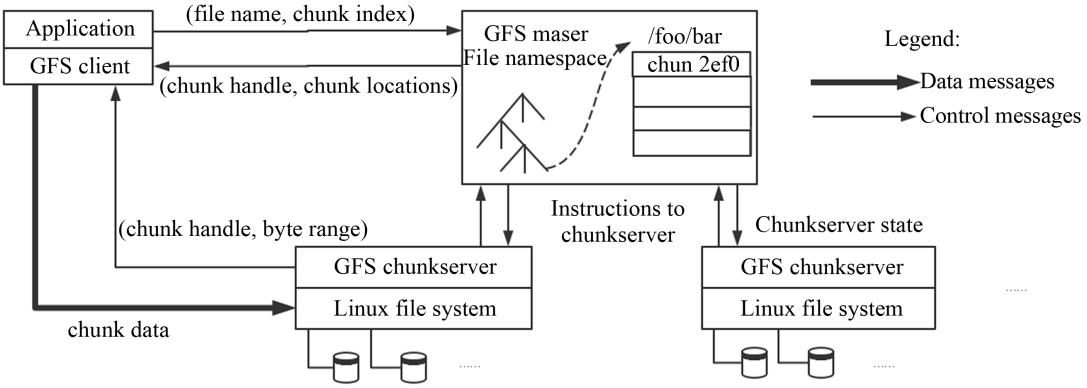


**Figure 1.** The architecture of GFS
**图 1.** GFS 基本架构

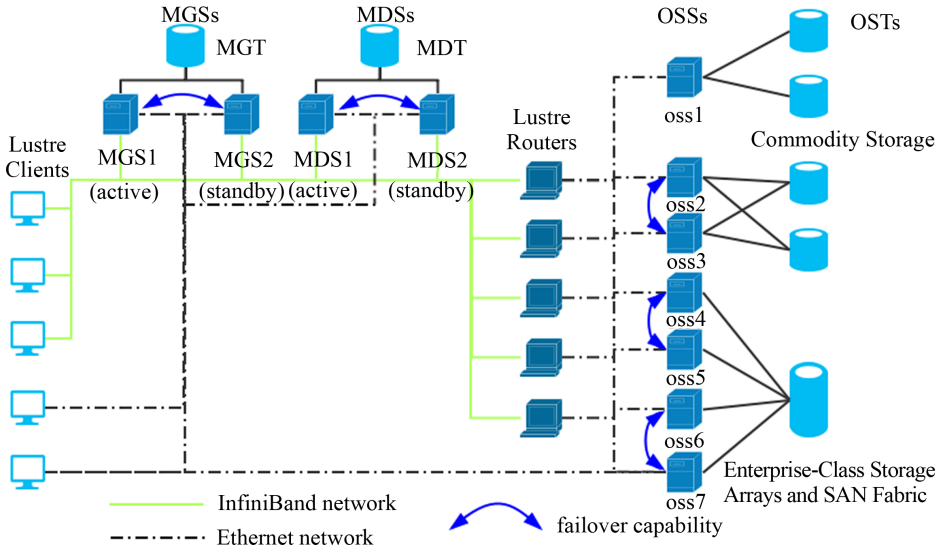


**Figure 2.** The architecture of Lustre
**图 2.** Lustre 基本架构

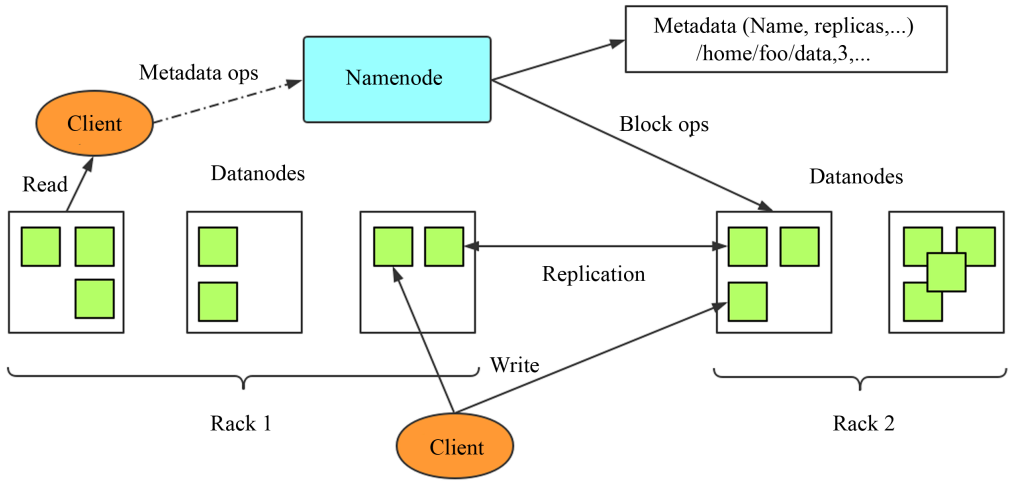


**Figure 3.** The architecture of HDFS
**图 3.** HDFS 基本架构

Tracker Server 是 FastDFS 中心节点，它的任务是收集 Storage Server 的状态信息、调度客户端请求和负载均衡。Tracker Server 将收集的存储服务器状态信息存储在内存中，并且记录分组的信息，但是没有记录文件的任何索引信息，所以需要占用很少的内存。当收到 Client 和 Storage Server 发来的请求时，Tracker Server 在记录的分组和 Storage Server 信息中寻找需要的信息，然后做出应答。Storage Server 直接通过本地操作系统中的文件系统来存储文件，存储文件时直接存储整个文件，而不会对文件进行分块存储，客户端上传的文件和 Storage Server 上的文件一一对应，图 4 是 FastDFS 基本架构。

## 3.5. 淘宝文件系统

淘宝文件系统(Taobao File System，TFS)是淘宝内部使用的分布式文件系统，承载着淘宝主站上所有的图片、商品描述等数据存储。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，在淘宝开发人员的努力下，TFS 文件存储系统完全解决了淘宝海量小文件存储的问题。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。

### FastDFS 架构

TFS 文件存储系统分为两部分 NameServer 节点和 DataServer 节点，采用主从式架构，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[7]。

NameServer 节点负责管理和维护 block 和 DataServer 的相关信息，包括 DataServer 加入、退出、心跳信息，block 和 DataServer 对应关系的建立和解除。DataServer 节点负责实际数据的存储和读写，正常情况下，每一个 block 都会在多个 DataServer 节点上存在，也就是说每一个文件都有多个备份，确保了数据的可靠性。在 DataServer 节点上，block 都是以主块+扩展块的形式存在的，一个 block 对应一个主块和多个扩展块。扩展块的应用是为了在文件大小发生变化时，如果主块的存储空间不够的话可以将数据放到扩展块里面。DataServer 内部为每一个 block 保存了一个与该 block 对应的索引文件(index)，在 DataServer 启动时会把自身所拥有的 block 和对应的 index 加载到内存。图 5 是 TFS 基本架构。

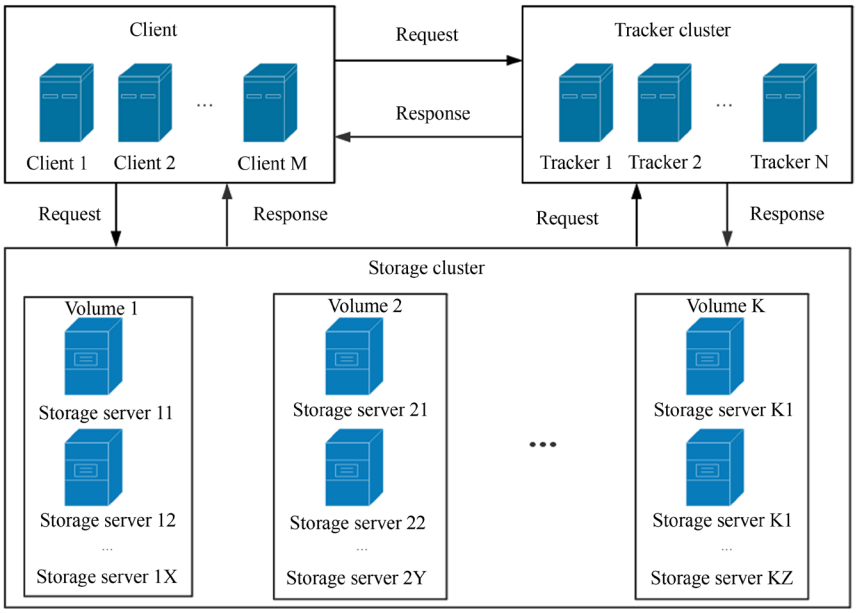


**Figure 4.** The architecture of FastDFS
**图 4.** FastDFS 基本架构

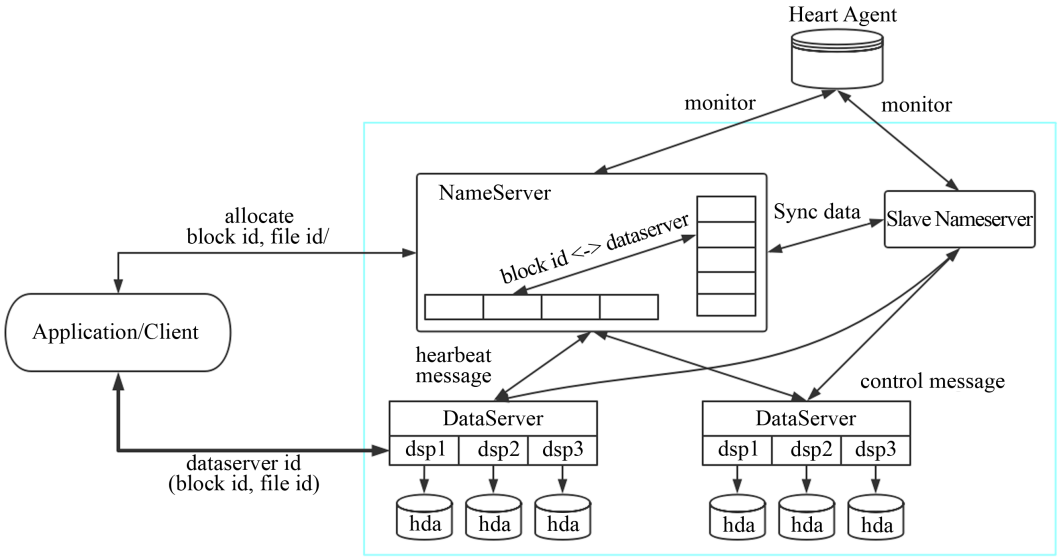


**Figure 5.** The architecture of TFS
**图 5.** TFS 基本架构

## 4. 结束语

随着计算机和网络技术的进一步发展，越来越多的应用需要磁盘容量更大、处理速度更快、扩展性更高的分布式文件系统，在这种需求拉动下，越来越多的优秀的分布式文件系统出现在大众面前，根据各类应用的不同需求提供不同的服务，它们在数据存储、数据管理、数据共享、安全性、稳定性、可用性和扩展性方面也各有所长、各具特色，深入探析各类分布式文件系统的优势和特性，整合资源取长补短，在互联网+、人工智能等新的领域中发现更大价值。

## 参考文献 (References)

[1] 韩增曦. 分布式文件系统 FastDFS 的研究与应用[D]: [硕士学位论文]. 大连: 大连理工大学, 2014.

[2] 周子涵. 基于 FastDFS 的目录文件系统的研究与实现[D]: [硕士学位论文]. 成都: 电子科技大学, 2015.

[3] 白铖. 一种分布式文件系统的设计与实现[D]: [硕士学位论文]. 成都: 电子科技大学, 2015.

[4] 周小玉. HDFS 分布式文件系统存储策略研究[D]: [硕士学位论文]. 成都: 电子科技大学, 2015.

[5] 胡军杰. 云平台下分布式文件系统评测技术研究[D]: [硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2014.

[6] 王波. 基于 FastDFS 的轻量级分布式文件系统的设计与实现[D]: [硕士学位论文]. 沈阳: 东北大学信息科学工程学院, 2013.

[7] 翟猛. 基于 TFS 的分布式文件存储平台研究与实现[D]: [硕士学位论文]. 天津: 河北工业大学, 2014.