



# 布袋文件系统

Budai Filesystem (BFS)

[h@ibudai.com](mailto:h@ibudai.com) 2009

# 主题

- **什么是布袋文件系统**
- 布袋文件系统 vs. Hadoop File System(HDFS)
- 布袋文件系统 vs. Google File System(GFS)
- Q & A

—

# 布袋文件系统是为在线存储服务设计的

在线存储服务对存储系统的要求：

- **海量数据** - 全球信息正在以爆炸性的方式增长，想象一下您的个人照片在过去一年内的增长速度
- **数据的可靠性** - 没有用户能够接受任何借口导致的数据丢失
- **成本** - 个人用户可以承受的起的存储价格
- **安全** - 保护用户数据的隐密性为用户接受云计算的关键

# 传统文件系统无法满足在线存储的需要

在线存储系统的要求	传统的本地文件系统
海量数据	通常都有最大容量限制， Reiserfs/ext3 最大为 16TB 。并且在文件数目增加到某个极限后，文件访问的性能会显著下降。 大部分文件系统都不支持动态扩展文件系统容量。
数据的可靠性	文件系统是脆弱的。文件系统本身不保障数据可靠性，一般依赖于卷管理软件或者磁盘阵列。文件系统一般不提供高可用支持。
成本	为了实现高可用和高性能，通常需要采购昂贵的磁盘阵列和额外的卷管理软件。
安全	企业级文件系统一般提供文件加密功能。

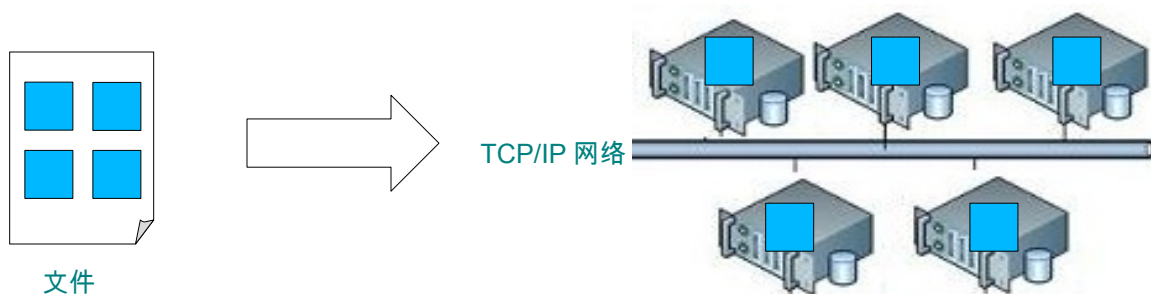
# 布袋文件系统是

- 高度可扩展的分布式文件系统
- 通用硬件平台 (Commodity hardware) 上提供的容错
- 大量并发访问下的高性能
- 内建的块级别的重复数据识别 (Dedup) 技术

在线存储系统的要求	布袋文件系统
海量数据	分布式文件系统在容量几乎是没有限制的。
数据的可靠性	特有的容错和高可用设计
成本	基于通用硬件平台大幅降低存储成本。 重复数据识别技术减少重复数据有助于减少对磁盘存储的需求，从而降低存储成本。
安全	文件级别和块级别 256 位 AES 加密

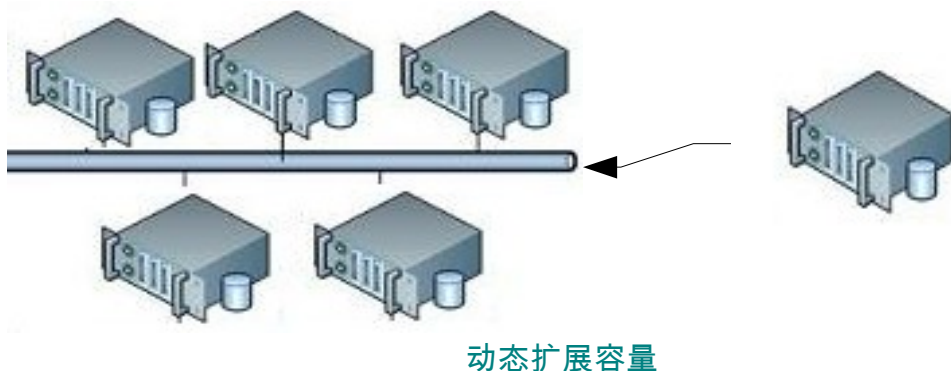
# 布袋文件系统：分布式文件系统

- 数据存储分布于多台计算机
- 提供“海量”的存储容量
  - 每个文件系统实例最大可存储 1000TB 数据
  - 远远超过传统本地文件系统 (NTFS/EXT3..) 的最大容量
- 更高的 I/O 吞吐量
  - 解决本地文件系统受到的单个服务器的磁盘和网络瓶颈
- 轻量级的用户态实现



# 布袋文件系统：高度扩展性

- 动态扩展容量
  - 加入一个新的存储结点既可自动识别扩展文件系统容量
  - 完全没有传统文件系统在扩容方面的种种不便
- 串联多个文件系统实例可以扩展到更大容量 (Scale Out)



# 布袋文件系统：通用硬件架构上的高可靠

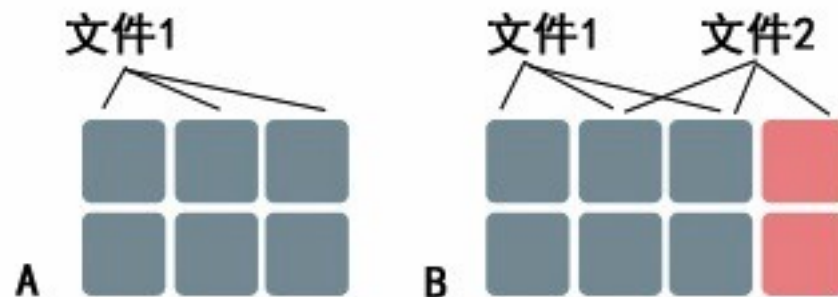
- 采购通用硬件平台 (Commodity Hardware) 搭建
  - AMD X86\_64 CPU/SATA 硬盘 /Ubuntu Linux
  - 大幅度降低每字节存储成本
- 容错
  - 每个数据块最少保留 3 个副本在不同结点上 - 避免单点故障
  - 保证同一数据块的副本至少分布于两个数据中心 - 避免数据中心故障导致数据丢失
  - 自动平衡算法 - 自动检测由于结点故障导致数据块副本数减少，动态复制数据块到健康节点上
  - 块级别的数据校验 - 检测数据块的完整性
- 高可用
  - 文件系统的元数据被实时复制到备份节点上
  - 任何结点的故障都不会影响整个文件系统的访问

# 布袋文件系统：性能

- 传统的文件系统的性能瓶颈
  - 元数据（文件名，大小，修改日期，数据块的位置等）：数据量小，要求响应速度快，可靠性要求高
  - 数据块（文件内容） - 数据量大
  - 元数据和数据块存储于单一文件服务器
  - 文件服务器磁盘和网络接口成为文件访问性能的瓶颈
- 布袋文件系统采用了类 pNFS 的并行架构设计
  - 并行架构将文件元数据和数据块根据数据访问的不同要求分布在不同的服务器上
  - Catalog Node - 文件系统元数据
  - Data Node - 专门存储文件的数据块
  - 并行架构已经被证明具有更高的 I/O 吞吐量

# 布袋文件系统：重复数据识别

- 内建块级别的重复数据识别
  - 文件内容被分割成 1M 大小的数据块
  - 写入新的数据块时，和系统中已有的数据块比对。已存在的数据块只需增加引用
  - 比文件级别的重复数据识别技术有更高的适用性
  - 布袋网的使用情况表明至少可以节约 20% 左右的磁盘空间
- 重复数据识别技术 (dedup) 已经被广泛应用于数据备份领域
  - Data Domain



当新的文件2被加入到储存系统时 (B)  
只有差异的部分需要占用新的磁盘空间

# 主题

- 什么是布袋文件系统
- **布袋文件系统 vs. Hadoop File System(HDFS)**
- 布袋文件系统 vs. Google File System(GFS)
- Q & A

—

# 布袋文件系统 vs. Hadoop File System

- Hadoop File System 设计用来支持那些需要特大数据集的应用程序
  - Apache Project - <http://hadoop.apache.org/core/>
  - 高的数据吞吐量
  - 文件只需写一次，读无数次，无需修改
  - Hadoop 云计算框架的一部分

	布袋文件系统	Hadoop File System
用户态实现	是	是
并行架构	是	是
容错 / 可靠性	高	高
加密支持	有	无。依赖于应用程序
重复数据识别	有	无
追加文件内容	支持	不支持

# 布袋文件系统 vs. Google File System

- Google File System 专门为 Google 的搜索引擎提供存储服务
  - 特大的数据量 / 高的数据吞吐量
  - 文件只需写一次，读无数次，无需修改
  - 对性能的要求远大于对数据可靠性的要求

	布袋文件系统	Google File System
用户态实现	是	是
并行架构	是	是
容错 / 可靠性	高	高
加密支持	有	未知 - 理论上无此要求
重复数据识别	有	未知 - 理论上无此要求
追加文件内容	有	未知 - 理论上无需支持

马上访问 <http://www.ibudai.com> 开始使用布袋文件系统  
提供的在线备份服务！

