分类号 TP3 _____

UDC _____ 编号 _____

中国科学院研究生院
博士学位论文

分布式文件系统可扩展元数据服务关键问题研究

杨德志

指导教师 许 鲁 研究员

中国科学院计算技术研究所

申请学位级别 工学博士 学科专业名称 计算机系统结构

论文提交日期 2007 年 11 月 论文答辩日期 2008 年 2 月

培养单位 中国科学院计算技术研究所

学位授予单位 中国科学院研究生院

答辩委员会主席 刘志勇 研究员

声　　明

我声明本论文是我本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，本论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

作者签名：　　日期：

论文版权使用授权书

本人授权中国科学院计算技术研究所可以保留并向国家有关部门或机构送交本论文的复印件和电子文档，允许本论文被查阅和借阅，可以将本论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编本论文。

（保密论文在解密后适用本授权书。）

作者签名：　　导师签名：　　日期：

摘　要

文件系统元数据描述文件系统及其管理的文件，其访问效率是文件数据访问性能的关键因素。在海量网络存储环境中，随着系统应用的多样化、应用需求规模的不断扩大，如何充分利用系统资源，提供高扩展能力的文件系统元数据服务，成为大规模分布式文件系统研究的热点问题。

多种典型应用的访问统计结果表明，文件系统元数据具有活跃性、局部性、动态性、需要更改文件系统的多元数据请求的比例很少等特点。基于此特性，本文的主要创新点是，提出动态灵活的分布式文件系统元数据服务机制和策略，支持文件系统元数据服务的动态扩展。它包括元数据存储和访问两个方面：

1) 集中共享的元数据虚拟存储模型。以虚拟存储技术支持的存储资源透明扩展、分层的存储资源管理和动态的元数据资源分配、元数据对所有请求服务器可见等为基础，元数据存储服务有效分离元数据的存储和访问，为动态灵活的元数据请求服务提供支持。
2) 动态灵活的元数据请求服务机制和策略。文件系统元数据表现出活跃性、局部性和动态性等特征。动态分布决策将文件系统名字空间结构与元数据类型相结合，由用户访问动态驱动活跃元数据的请求分布。初步的对比评估结果表明，相对于目录子树分区法，其请求处理能力将提高 30%左右。

动态灵活的元数据服务机制为服务扩展能力提供基础。为解决两阶段提交等传统协议的不足，本文提出通过动态迁移协议，集中处理跨服务器请求，有效保证请求的原子性。在最坏情况下，动态迁移协议也能够减少 25%的处理时延，且其错误恢复的影响也小得多。

针对传统和新兴应用的实验验证了元数据服务扩展能力的有效性。针对生物信息计算 BLAST 的评估表明，元数据服务器的增加将带来 20%左右的元数据服务时间降低。通过对系统实现的优化，将可能获得近线性的元数据服务扩展能力。同时，实验结果还为未来的工作方向提供了参考：1）原型系统的实现优化；2）自适应的元数据请求分布决策模型；3）系统结构的优化，以更好支持超大规模的系统扩展需求。

关键词：网络存储系统；可扩展分布式文件系统元数据服务；蓝鲸集群文件系统；蓝鲸元数据服务器集群系统

Research on Key Issues of Scalable Distributed File System Metadata Service in Large-scale Networked Storage Systems

Yang Dezhi (Computer Architecture)

Directed By Xu Lu

Efficiency of metadata access is a key factor of that of file access. In large-scale networked storage systems, with increase of file system metadata requests, doing research on key issues of scalable file system metadata service becomes a key problem of distributed file system researches.

Based on characteristics of locality, dynamicity and rare cross-server requests of active metadata, a dynamically scalable distributed file system metadata service is provided which includes:

a) A symmetric architecture of metadata servers which bases on fully shared virtual storage model and hierarchical storage resource management which supports scalable storage resource management.

b) Flexible management of file system metadata requests distribution. Supported by hierarchical distribution management, flexible metadata requests distribution which keeps metadata locality in mind decides requests distribution dynamically, which supports dynamic changes in system.

And more, there is a lightweight protocol for atomicity of cross-server metadata requests. Based on the symmetric architecture and fully shared storage model, requests which involve metadata distributed on multiple servers will be processed on single server by migrating active metadata between servers.

Evaluations on the prototype verify its scalability. And it also points out future works on architecture adjustment, adaptive metadata requests distribution, performance tuning and high-available metadata service, and so on.

Keywords: networked storage system; scalable distributed file system metadata service; BlueWhale file system; BlueWhale Multiple Metadata servers System

目　录