BW-netRAID:一种后端集中冗余管理的网络RAID系统

那文武[1),2)] 柯 剑[1),2)] 朱旭东[1),2)] 孟晓烜[1),2)] 卜庆忠[1)] 许 鲁[1)]

[1)](中国科学院计算技术研究所 北京 100190)

[2)](中国科学院研究生院 北京 100039)

摘 要 在大型网络存储系统中,在设备间采用冗余策略是提高数据可靠性的重要方法.针对目前网络存储系统前端集中冗余管理中存在的性能瓶颈问题,结合带外虚拟化存储管理架构,文中提出了一种前端并行数据传输和后端集中冗余管理的网络 RAID 存储系统.应用服务器从元数据服务器获得地址映射信息后可直接并行访问存储设备,充分利用了所有存储节点的聚合 I/O 性能;冗余管理服务器在磁盘上以日志方式缓存镜像块数据,然后在后台异步计算校验块,并将新校验块数据更新到对应的存储设备节点,从而避免了前端集中冗余管理的单点性能瓶颈和可靠性问题.对不同访问活跃度的数据采用 RAID1/RAID5 异构分布的管理方法,取得了系统性能、可靠性和价格的平衡.

关键词 网络 RAID 系统;后端集中冗余管理;冗余管理协议;活跃数据缓存;数据可靠性分析

中图法分类号 TP302 **DOI 号**: 10.3724/SP.J.1016.2011.00912

A Network RAID System with Backend Centralized Redundancy Management

NA Wen-Wu[1),2)] KE Jian[1),2)] ZHU Xu-Dong[1),2)] MENG Xiao-Xuan[1),2)] BU Qing-Zhong[1)] XU Lu[1)]

[1)](*Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

[2)](*Graduate University of Chinese Academy of Sciences, Beijing* 100039)

Abstract It is an important scheme that data redundancy among the storage devices is used to enhance the reliability in the large storage system. But the network storage systems with frontend centralized redundancy management have the performance bottleneck problem. The paper presents a novel network RAID storage system based on the out-of-band storage virtualization. It adopts the backend centralized redundancy management and the frontend parallel data transfer architecture. The application servers can acquire the mapping information of the virtual disks from the metadata server of storage system and directly access the storage device nodes. It maximizes the aggregated I/O performance of all storage device nodes. And the redundancy management server can cache the mirrored data blocks in the local disks with log-structured mode, then asynchronously calculate the parity blocks in the background process. Therefore it can relieve the performance bottleneck and enhances the reliability. The heterogeneous layout of RAID1/RAID5 based on the different access model of data blocks has acquired the trade-off between performance, reliability and cost of the whole storage system.

Keywords network RAID system; backend centralized redundancy management; redundancy management protocol; active data cache; data availability analysis

---

收稿日期:2008-09-24;最终修改稿收到日期:2011-03-16.本课题得到国家"九七三"重点基础研究发展规划项目基金(2004CB318205)、国家"八六三"高技术研究发展计划项目基金(2007AA01Z402,2009AA01A403)资助.**那文武**,男,1977 年生,博士研究生,研究方向为网络虚拟存储和数据冗余. E-mail: naww77@gmail.com. **柯 剑**,男,1971 年生,博士研究生,研究方向为虚拟存储和自适应性能管理. **朱旭东**,男,1979 年生,博士研究生,研究方向为虚拟存储和 I/O 访问模式识别. **孟晓烜**,男,1980 年生,博士研究生,研究方向为网络存储和 Cache 管理. **卜庆忠**,男,1972 年生,博士,助理研究员,研究方向为网络存储和存储管理模型. **许 鲁**,男,1962 年生,博士,研究员,博士生导师,研究领域为网络存储和文件系统.

1 引　言

网络存储是传统的存储技术借助于网络技术发展形成的一个新兴的交叉技术，研制大容量、高性能、高可用、可扩展、高性价比、高效能和易管理的存储系统是目前网络存储领域的重要发展方向. 数据冗余技术是以存储空间代价换取数据可靠性的有效策略，能够在一定的可靠性指标范围内，容忍部分存储设备故障，保证任何数据不丢失，同时还能保证数据可以连续访问.

在网络存储系统中保障用户数据的可靠性非常重要. 因此如何实现存储节点间的数据冗余功能成为网络存储研究中相当重要的问题. 当前的网络存储系统主要有以下两种冗余管理架构：

(1) 前端集中式主从管理

在应用服务器和存储节点之间有一个性能强劲的控制服务器，能够快速存储转发数据到存储设备节点，同时也负责管理节点间的数据冗余. 其优点是系统实现和存储管理简单，数据冗余一致性语义易于保证，而且可以采用高性能硬件或 Cache 等技术对读写进行优化. 缺点是控制服务器存储转发所有的数据导致 I/O 吞吐率受限于其网络接口带宽和处理能力，而且 I/O 路径增长带来的转发延迟导致 I/O 响应时间增加，难以充分发挥分布式网络存储系统并发通信和并行存储的能力，系统扩展性比较差；另外 RAID5 冗余即时计算模式以及数据小写更新问题将导致用户数据写性能严重下降.

(2) 分布式对等管理

分布式存储没有集中的冗余管理服务器，多个存储设备节点协调管理，优点是能充分发挥分布式系统并行读写能力，缺点是每个节点既是 Client 又是 Server，节点间冗余管理较复杂. 目前系统多采用镜像冗余的方式来降低管理复杂度，但镜像方式的存储系统的空间利用率只有 50%. 如采用 RAID5 冗余方式，冗余计算将会抢占存储服务器的数据传输带宽和 CPU 资源，冗余请求竞争磁盘资源而且干扰正常用户读写请求，降低了存储服务质量.

在带外虚拟化存储架构中，数据传输的显著特点是应用服务器直接访问存储设备节点. 针对上述前端集中冗余管理的问题，本文提出了一种新的网络 RAID 系统：BW-netRAID. 主要创新点是虚拟存储元数据管理和冗余管理功能分离，在数据传输通道后端的冗余管理服务器上后台异步执行计算，从而减少数据冗余对用户数据读写性能的影响.

BW-netRAID 采用后端集中冗余管理方式，读请求直接从存储设备读取数据，写请求数据既要存入存储设备，同时还镜像到冗余管理服务器；然后当系统空闲或者资源不足时才在后台异步计算 RAID5 校验块，并更新到另外的存储设备上. 优点是：带外虚拟化并行访问存储设备能够有效聚合所有存储设备的 I/O 性能；而后端冗余管理能够避免前端集中冗余管理的单点性能瓶颈问题. 对不同访问活跃度的数据采用 RAID1/RAID5 异构分布的管理方法，取得了系统的性能、可靠性和价格的平衡.

本文第 2 节概述系统结构和主要的软件模块功能以及关键技术特点；第 3 节讨论系统正常、冗余计算和重构状态的数据读写协议；第 4 节描述系统实现方法，包括元数据结构、设备创建和初始化、节点故障监测和恢复以及活跃数据缓存机制；第 5 节评价系统性能的扩展性和数据恢复速度；第 6 节分析系统如何保证数据冗余的完整性以及具体的数据布局方式，并给出系统的数据可靠性分析结论；第 7 节介绍相关研究工作；第 8 节总结并指出下一步的研究方向.

2 系统概述

2.1 BW-netRAID 系统结构

图 1 描述了 BW-netRAID 的系统组件图，系统主要有 4 类组件：

(1) 存储设备节点(Storage device Node，SN)

它有独立的 CPU、内存、网络接口和大容量的磁盘阵列，用来存储应用服务器的数据. 当存储设备节点出现软硬件故障或者内部 RAID 损坏时，可能导致其上的数据不能被访问甚至丢失.

(2) 应用服务器(Application Server，AS)

它上面运行应用服务程序，如文件服务器、Web 服务器、数据库服务器和流媒体服务器等. 应用服务器通过网络虚拟磁盘读写存储设备节点上的数据. 存储设备间数据冗余关系对应用服务器透明.

(3) 虚拟化元数据服务器(Virtualization Metadata Server，VMS)

它以带外虚拟化方式管理应用服务器的数据请求到存储设备节点上数据之间的映射. 应用服务器先从元数据管理服务器获取并缓存网络虚拟磁盘的地址映射信息，在数据读写时直接访问存储设备节点，执行读写请求操作.

(4) 冗余管理服务器(Redundancy Management Server，RMS)

它维护和管理存储设备节点之间数据的冗余关系,保证不会因为单个存储设备节点的故障而导致用户数据不能访问或丢失,并提供数据重构机制和恢复策略.

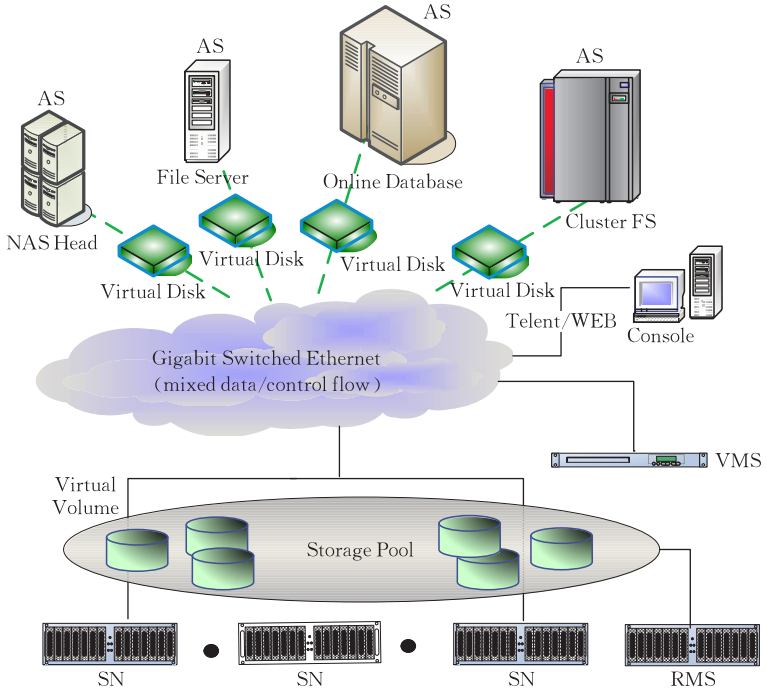


图 1 BW-netRAID 系统组件图

应用服务器、元数据管理服务器和存储设备节点是网络存储系统的带外虚拟化存储管理的基本结构①.传统的前端集中管理系统中控制服务器既负责资源地址映射和数据转发,又管理多个存储设备间的冗余计算.BW-netRAID 系统中新增加了冗余管理服务器,它专职负责存储设备节点间的数据冗余管理,而资源地址映射由元数据服务器管理.

元数据管理和冗余管理功能分离是本系统的主要特点.元数据管理服务器在数据通道的前端以带外方式管理资源地址映射,冗余管理服务器在数据通道的后端以 RAID1/RAID5 分级存储方式管理存储设备节点上数据块的冗余关系.

2.2 软件模块

图 2 给出 BW-netRAID 系统中主要的软件模块和数据请求传输示意图.应用服务器上的网络磁盘从元数据服务器获取地址映射关系后,缓存到地址映射表(mapping table),再将请求通过虚拟磁盘(Virtual Disk,VD)直接发送到对应的存储设备节点上的数据收发模块(data-server),根据相应状态的数据读写协议完成读写操作.存储设备节点的数据卷(data-lv)存储应用服务器读写的原始数据,而冗余卷(parity-lv)存储不同设备间的 RAID5 校验块数据.冗余管理服务器上的数据收发模块接收处理存储设备节点发起的数据读写请求,当后台异步计算 RAID5 时进程(RAID-update)也向冗余数据收发模块(parity-server)发起校验块的读写请求.对于镜像写请求,冗余管理服务器先以日志写方式缓存到本地的磁盘(LogDisk)上,当 RAID5 计算时再读回.如果冗余管理服务器要读取或者恢复存储设备节点上的数据,则直接访问存储设备节点的数据收发模块获取数据.

2.3 关键技术

本节从系统冗余级别、读写性能和冗余计算技术等 3 个方面分析 BW-netRAID 的关键技术特点.

(1) 系统冗余级别

从应用服务器上的用户数据读写角度看,BW-netRAID 是一个 Stripping+Mirroring 的系统,并发读写多个存储设备节点能更充分发挥分布式网络存储系统的性能优势.而从存储设备节点上的数据存储关系看,是一个 RAID5 系统,系统有较优的空间利用率和可靠性/空间利用率比的优势.

(2) 读写性能

整个系统的数据读性能不再受限于前端的集中管理服务器,充分发挥多个存储设备节点的并发读性能,系统扩展性高,解决了前端集中式管理的读性

① Rob Peglar, Virtualization I - What, Why, Where and How? http://www.snia.org/education/tutorials/2008/spring/virtualization/Peglar-R_Virtualization_I_Why_What_Where_How.pdf, 21-23

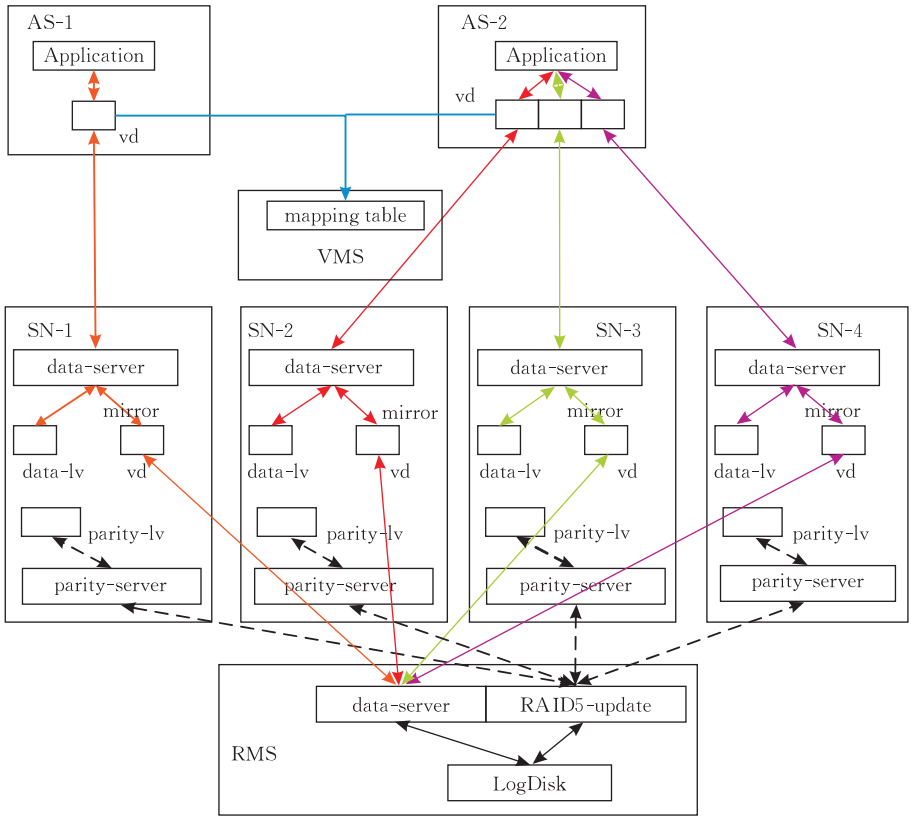


图 2　BW-netRAID 的软件结构图

能瓶颈问题.

数据写操作将新写的数据块镜像到后端的冗余管理服务器上，写性能和镜像冗余方式相当，比 RAID5 方式得到大幅度提高. 而且针对写性能可能受限单个集中点的问题，冗余管理服务器采用延迟写和写聚合，以及在磁盘上用日志方式缓存活跃数据等优化技术来进一步提高写性能.

(3) 冗余计算

由单个服务器集中管理冗余存储的机制和策略，能够简化数据冗余一致性的管理. 降级和重构状态时先从冗余管理服务器的缓存磁盘中读取，没有时才由冗余管理服务器重构丢失的数据，而且还支持优先恢复活跃的数据，缓存的数据不必执行 RAID5 重构，减少了 RAID5 冗余计算时数据块的同步传输和计算负担. 存储设备节点不参与冗余存储管理，即使某个节点有故障也不会直接影响其它节点，非故障节点照常读写，最大程度上减少冗余存储对存储服务器性能的影响，保证了整个系统的存储服务质量.

3　数据读写协议

BW-netRAID 系统主要有 4 种运行状态：正常状态、降级状态、重构状态和失效状态. 当每个存储设备节点都正常工作时，整个系统处于正常状态；而当其中某个存储设备节点故障时，导致其上的数据不能被访问甚至丢失，系统进入降级状态；当故障设备被修复或者用新设备替代后，系统转入重构状态，冗余管理服务器负责重构丢失的数据并恢复到存储设备节点上；当所有丢失的数据都被恢复后，存储设备节点间的数据冗余关系重新回到一致的状态，系统也恢复到正常状态. 如果多个设备同时故障超过了系统冗余能力，则会导致数据不能被访问或者部分数据丢失，系统变为失效状态.

下面通过分析每种状态下的读写操作过程中的内存、磁盘和网络上的 I/O 操作，描述数据读写协议流程. 限于篇幅，这里只给出正常状态下数据写协议、后台 RAID5 计算协议和重构状态下丢失数据恢复协议.

3.1　正常状态下数据写协议

正常状态下数据写协议包括存储设备节点的写请求处理和冗余管理服务器的镜像数据缓存处理. 具体步骤如图 3 所示：

(1) 应用服务器上的用户应用程序向网络虚拟磁盘发起写请求；

(2) 网络虚拟磁盘根据地址映射信息将写请求

通过网络数据传输协议传给对应的存储设备节点；

(3)存储设备节点将写请求数据分别存储到冗余管理服务器和本地卷设备上，组成 RAID1 镜像冗余；如果冗余管理服务器上没有这个数据块的旧数据块，存储设备节点在写新数据块之前还需要读取存储设备节点上的旧数据块并传输到冗余管理服务器上；

(4)冗余管理服务器先在内存中缓存接收的镜像数据块，返回镜像写完成信息给存储设备节点；当缓存到一定数目的数据块后，再将多个数据块以一次连续写方式写到本地日志磁盘上；

(5)存储设备节点的底层卷设备和冗余管理服务器的写请求都完成后，存储设备节点返回写请求完成信息给应用服务器的网络虚拟磁盘；

(6)网络虚拟磁盘返回写请求完成信息给上层存储应用.

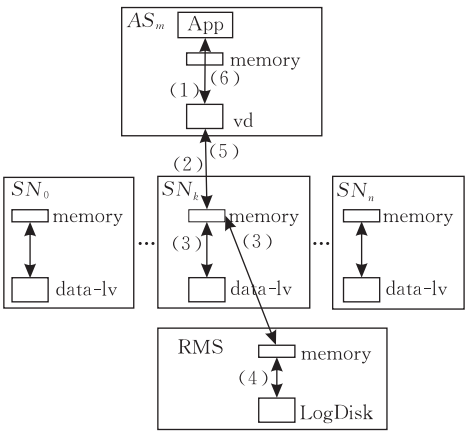


图 3 正常状态写协议

3.2 正常状态下后台计算 RAID5 协议

在冗余管理服务器上，当系统空闲或者磁盘剩余空间超过指定阈值时，后台进程异步计算 RAID5 校验块. 具体步骤如图 4 所示：

1. 从冗余管理服务器的磁盘日志缓存中读取数据块，同时也读取其对应的旧数据块和校验块；如果缓存了校验块，转到步 3，如果没有转到步 2；

2. 根据 RAID5 布局关系从对应存储设备节点上读回校验块：

2.1. 冗余管理服务器向存储设备节点发起校验块的读请求；

2.2. 存储设备节点将读请求转发给底层卷设备；

2.3. 从磁盘阵列设备上获取所需的读请求数据后，返回读请求完成信息；

2.4. 将读请求数据返回给冗余管理服务器；

3. 根据 RAID5 公式计算得到新的校验块并缓存到磁盘日志中；

4. 冗余管理服务器再发起新的校验块写请求，将新校验块写到这个 RAID5 条带对应的存储设备节点上；

5. 存储设备节点将接到的写请求转发给底层卷设备；

6. 存储设备节点的磁盘阵列设备返回写请求完成信息；

7. 冗余管理服务器收到存储设备节点返回的新校验块的写请求完成信息.

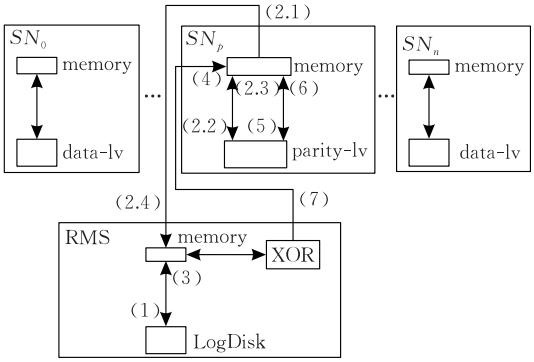


图 4 后台计算 RAID5 协议

冗余管理服务器根据数据活跃程度和存储空间占用情况决定是否释放磁盘上缓存的 RAID1 镜像块数据和 RAID5 校验块数据. 对于不再有用的旧数据块和校验块，由资源回收进程删除并释放其占用的空间.

3.3 重构状态下丢失数据恢复协议

当存储设备节点的故障被修复或者新设备加入到 BW-netRAID 系统中，系统进入重构状态. 冗余管理服务器作为存储设备节点间数据冗余关系的集中管理点，启动数据重构进程，重构丢失的数据并同步到对应的存储设备节点上. 重构进程首先恢复缓存的活跃数据，然后再重构其它的数据. 丢失的活跃数据只需从冗余管理服务器的磁盘缓存日志中就可以找回，而其它非活跃数据需要利用 RAID5 算法进行重构恢复. 恢复非活跃数据块的具体步骤如图 5 所示：

1. 根据要重构数据块的 RAID5 冗余关系，查询磁盘缓存日志上是否缓存了相同条带的数据块和校验块，如果已经缓存则读回到内存中，如果没有则从存储设备节点读回；

2. 向组成 RAID5 的多个存储设备节点发起读请求；

3. 每个存储设备节点将读请求转发给底层卷设备；

4. 从卷设备上获取读请求数据后，返回读请求完成信息；

5. 冗余管理服务器接从存储设备节点收到读请求的数据；

6. 当相同条带的所有数据块和校验块都被读回到冗余管理服务器后，根据 RAID5 恢复算法计算得到故障的存储设备节点上丢失的数据；

7. 冗余管理服务器将计算得到的数据以写请求方式传送给刚修复的存储设备节点；

8. 刚修复的存储设备节点将接收的数据写到底层卷设备上；

9. 存储设备节点的底层卷设备完成写操作；

10. 存储设备节点将写完成信息返回给冗余管理服务器.

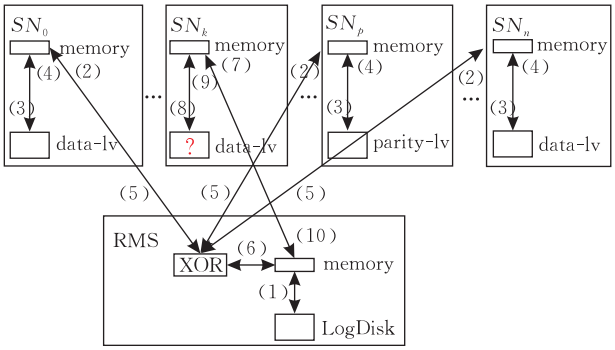


图 5　重构丢失数据协议

当丢失的非活跃数据被正确恢复后，后续的用户读写请求直接访问相应的存储设备节点.

4　系统实现

本节介绍 BW-netRAID 系统实现中关键的数据结构和性能优化方法. 首先概述系统的元数据结构，然后描述系统的启动运行、故障监测和恢复机制. 最后分析使用磁盘缓存镜像数据的优点和缓存数据结构，以及怎样计算 RAID5 数据和回收失效数据占用的存储空间.

4.1　元数据

BW-netRAID 的元数据包括记录网络 RAID 设备状态的系统元数据，以及系统运行过程中标识数据块状态的数据块元数据.

系统元数据记录了组成网络存储系统的存储设备节点的名称、逻辑卷和校验卷名字和大小、网络接口标识等. 这些元数据信息存储在存储设备节点的逻辑卷和校验卷的保留区域内以及冗余管理服务器上. 当系统启动时，各个存储设备节点分别从各自的逻辑卷和校验卷上读取与系统状态信息相关的元数据. 在运行过程中，设备状态发生变化时(如节点故障)，要同步更新每个节点的系统元数据到磁盘上，保持系统元数据的一致性.

数据块元数据包括存储设备节点上的 RAID5 初始化位图表、RAID5 重构位图表和数据块更新位图表以及冗余管理服务器上的缓存数据块索引表. 存储设备节点用数据块更新位图表. 冗余管理服务器用缓存数据块索引表同时记录数据块的状态，保证数据块状态的元数据不会因为单个节点异常宕机而丢失；然后定时将修改的元数据刷新到磁盘上，避免每次改变都更新到磁盘的同步写操作产生过大的 I/O 负担的问题.

4.2　创建和数据同步

存储设备节点启动后根据配置访问冗余管理服务器，协商网络 RAID 的组成. 目前系统支持 N 个存储设备上的相同容量(C)的逻辑卷组成 RAID5，另外每个存储设备节点还要保留容量为 $C/(N-1)$ 的冗余卷，用来存储节点间的 RAID5 校验块. 然后利用初始化位图表同步 RAID5 条带数据，分别在多个存储设备节点对数据块和校验块置零，与读取数据块到一个集中点再计算 RAID5 校验块的同步计算方法相比，能够更大地发挥多个存储设备节点并发读写能力，数据初始化同步时间短.

4.3　节点失效监测和恢复

冗余管理服务器通过“心跳”机制来监控和管理所有的存储设备节点，存储设备节点自己监控内部的磁盘状态. 当存储设备节点主动上报内部磁盘故障或者冗余管理服务器监测到某个节点失效时，系统进入降级状态，执行降级状态读写协议. 应用服务器的存储代理发现不能访问故障节点，则会从元数据管理服务器得到指向冗余管理服务器的新映射地址. 存储设备节点的内部故障修复或者新的存储设备节点加入到系统中后，系统进入重构状态，执行重构状态读写协议. 直到所有的数据重新恢复一致的冗余关系后，网络 RAID 系统再恢复到正常状态.

4.4　在磁盘上缓存活跃数据

多个存储设备节点的写请求都要镜像到冗余管理服务器，随着 RAID 规模增大和 I/O 写负载增加，冗余管理服务器可能成为性能的瓶颈. 缓存写请求并立即向存储设备节点返回写确认能够提高 I/O 响应时间和写性能，缓存数据越多则提高写性能越多.

用内存缓存速度快但容量有限，异常宕机还会丢失内存中的数据，而 SSD 价格还比较昂贵. 磁盘容量大价格便宜，但磁盘连续写速度快而随机写速度慢. 因此采用日志方式缓存写请求数据，能改变随机写为连续写操作，从而提高磁盘上缓存数据的性能.

4.4.1　缓存数据块

根据数据冗余完整性的需要(参见第 6 节)，缓存的数据既要包括新数据块也要包括旧数据块. 日志技术能够记录写请求的多个数据版本，当后台 RAID5 计算时，只需利用最新的版本和最老的版本来计算. 资源回收程序需要保留最新和最老的数据块，优先回收中间的失效的数据版本，并释放其占用的存储空间.

4.4.2 缓存校验块

冗余管理服务器既缓存活跃数据块，当计算RAID5校验块后，也尽可能地缓存校验块.如果不缓存这些校验块，那么在小写更新情况下每次计算RAID5校验块时，都需要从相应的存储设备节点上读回旧校验块，会增加网络数据传输和存储设备节点上读数据的负担.即使将前后数据块的变化差异传给存储设备节点，由存储设备节点接着完成RAID5计算，也仍然会增加存储设备节点的计算和读校验块的负担.

4.4.3 RAID5更新

尽管磁盘容量较大，但是相对于网络存储系统的总存储容量还是较小.因此为避免冗余管理服务器上的磁盘缓存区最终被写满，需要设置剩余容量下限，启动后台进程及时做异步RAID5计算，并释放部分存储空间.在正常状态下只有数据写请求到达冗余管理服务器，读请求时冗余管理服务器空闲，另外利用内存暂时缓存部分新数据块使得不必每次写操作都访问磁盘.上述方法使得冗余管理服务器能够得到一定的空闲时间用于RAID5计算和校验块更新.

4.4.4 资源回收

日志方式缓存数据需要有资源回收进程不断删除部分数据以便重新利用它们占用的空间.系统优先回收写请求中的中间版本，再根据数据的活跃度回收最近很少访问的数据块占用的空间.如果剩余空间仍然低于阈值，则先执行RAID5计算后，再回收已经完成RAID5计算的数据块，直到剩余空间大于资源回收下限的阈值.

5 系统性能评估

本节对BW-netRAID系统的性能进行测试和评估.测试系统包括1个元数据管理服务器、4个应用服务器、6个存储设备节点和1个冗余管理服务器；节点间通过千兆以太网连接.每个存储设备上的数据卷和校验卷组成一个网络RAID设备，然后在网络RAID设备上划分逻辑卷，每个逻辑卷大小为20G，逻辑卷以stripping方式映射到多个存储设备节点上.每次测试都重新创建网络RAID设备和逻辑卷，然后从应用服务器端读写网络虚拟磁盘，统计多个应用服务器聚合的读写性能.使用IOMeter[①]作为I/O性能评测工具，负载类型为4KB块的连续读和连续写.测试机的软硬件配置如表1所示.

表1 测试系统软硬件配置

	CPU	内存/G	磁盘	系统	程序
AS	Xeon 2.4G	1	网络虚拟磁盘	Linux 2.6.11	vd-agent
SN	Xeon 2.4G	1	1个120G的磁盘	Linux 2.6.11	sn-server
VMS	Xeon 2.4G	1	1个120G的磁盘	Linux 2.6.11	meta-server
RMS	Pentium D 3.0G	2	4个200G的磁盘	Linux 2.6.11	raid-server

5.1 扩展性测试

扩展性测试用来比较后端集中冗余管理方式(BW-netRAID)和前端集中冗余管理方式(RAID0/RAID5 controller)随应用服务器个数和存储设备节点个数变化时的读写性能.在前端集中管理方式中，聚合的读写性能受限于集中管理服务器转发性能和存储设备节点聚合读写性能.在BW-netRAID中，读性能只受限于存储设备节点聚合读性能，而写性能受限于存储设备节点聚合写性能和后端冗余管理服务器缓存写性能.因此测试结果显示，在同样配置下BW-netRAID系统比前端集中冗余管理方式性能高，而且随应用服务器和存储设备节点个数增加，系统性能扩展性更好.

5.1.1 应用服务器个数变化

由4个存储设备节点组成网络RAID，然后分别测试1～5个应用服务器时的顺序读性能，测试中所有应用服务器同时并发读同一个卷.由于磁盘读请求的搜索距离小以及数据预取和缓存效果，使得读操作多数在内存中完成.图6中前端集中方式中应用服务器个数为4个时就到达最高(千兆网络最大传输带宽)，聚合读性能受限于集中的控制点的最大网络传输带宽.BW-netRAID读性能随应用服务器个数线性增加，预期最高可到达4个存储设备节

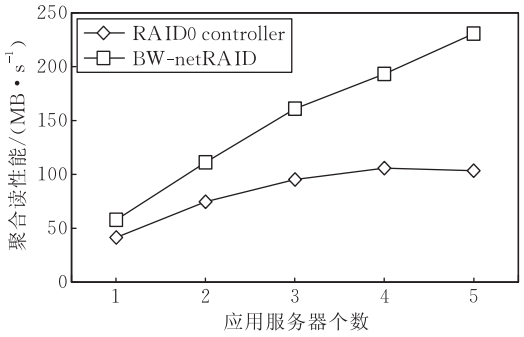


图6 应用服务器个数变化的读性能(多个应用服务器顺序读同一个卷)

① IOMeter Disk and FileSystem benchmark, http://www.iometer.org

点的聚合网络带宽，扩展性更好. 另外由于前端集中方式比后端集中方式要多一次存储转发，所以即使没到达集中控制点最大性能时，前端集中方式系统的聚合读性能也一直低于 BW-netRAID 系统.

由 4 个存储设备节点组成网络 RAID，然后分别测试 1～5 个应用服务器时顺序写性能，测试的应用服务器同时分别写不同的卷. 多个应用服务器共享相同的网络存储设备，同时写会导致不同卷的磁盘写请求的搜索距离大，系统性能受限于存储设备节点的磁盘写性能. 图 7 的测试结果表明后端集中冗余管理的方式比前端集中冗余管理方式写性能高. RAID5 前端集中方式由于还要同时进行冗余计算，其性能比 RAID0 前端集中方式更低.

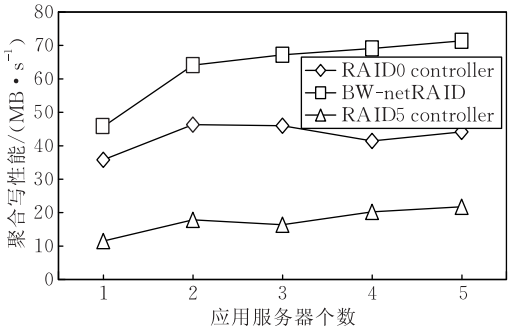


图 7　应用服务器个数变化的写性能(多个应用服务器分别顺序写不同的卷)

5.1.2　存储设备节点个数变化

测试由 2～6 个存储设备节点组成网络 RAID，然后 4 个应用服务器分别顺序读不同的卷. 图 8 表明存储设备节点个数相同时后端冗余管理比前端冗余管理读性能高一些. 两种结构的读性能都随存储设备节点个数线性增加. 但前端冗余管理预期最大只能达到前端集中控制服务器的最大网络带宽，而后端冗余管理预期最大可达到所有存储设备节点的最大网络带宽.

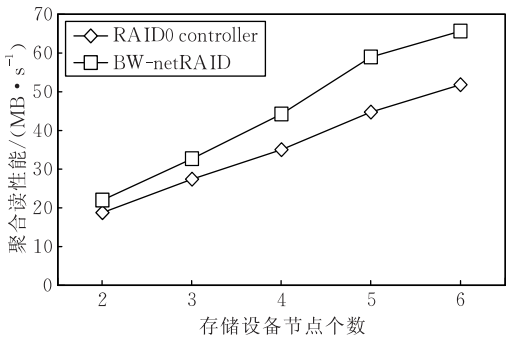


图 8　存储设备节点个数变化的读性能(4 个应用服务器分别顺序读不同的卷)

测试由 2～6 个存储设备节点组成网络 RAID，然后 4 个应用服务器分别顺序写不同的卷. 图 9 表明后端冗余管理比前端冗余管理写性能高一些. 后端冗余管理的写性能当存储设备节点个数较少时随节点个数增加，当节点个数超过 4 个时不再增加，说明此时写性能受限于后端冗余管理服务器的缓存写性能. RAID0 方式前端冗余管理写性能当存储设备节点个数较少时随节点个数增加，但节点个数超过 5 个时不再增加，说明此时写性能受限于前端冗余管理服务器的转发写性能. RAID5 方式前端集中管理由于冗余计算的负担导致其性能最低.

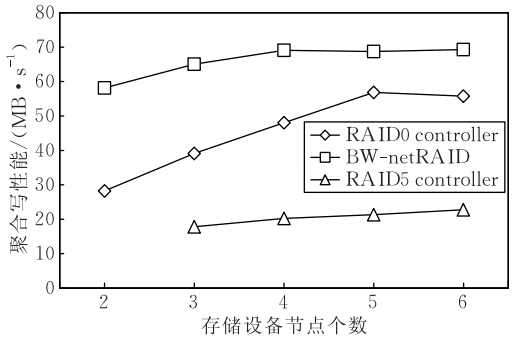


图 9　存储设备节点个数变化的写性能(4 个应用服务器分别顺序写不同的卷)

5.2　磁盘缓存读和重构读速度比较

当系统降级和重构状态时要读取丢失的数据块或者重构进程主动恢复丢失数据时，根据数据块状态有两种执行方式：(1) 从冗余管理服务器上的磁盘缓存读数据，记作 logging-read；(2) 先从多个存储设备节点上读数据再做 RAID5 计算后得到读请求数据，记作 RAID5-degrade-read. 测试由 3～6 个存储设备节点组成网络 RAID，当一个存储设备节点故障时，系统进入降级状态，再测试从一个应用服务器上顺序读一个卷的速度. 图 10 表明从磁盘缓存读性能是 RAID5 计算恢复读性能的 1.25～1.5 倍，RAID5 计算恢复读性能随存储设备节点个数增加而减少，因此说明缓存数据能缩短重构时间. 上面测

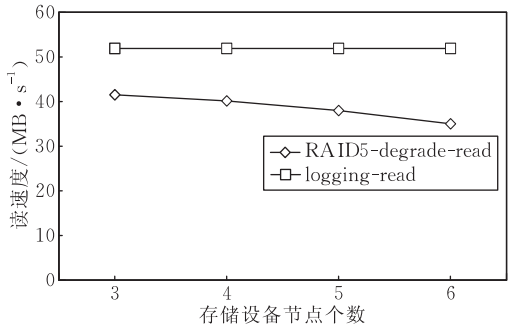


图 10　从磁盘缓存读和 RAID5 计算恢复读的对比

试是无负载的理想情况下，如果存储设备节点上有应用服务器的读写请求，彼此竞争资源会导致重构读性能降低.

6 数据布局和可靠性分析

本节首先讨论在数据更新时保证冗余关系完整性的必要性，给出冗余组中数据块所属的 RAID1 和 RAID5 冗余级别的变换过程，最后针对上述数据块异构分布，分析了不同类别的数据块可靠性.

6.1 数据布局

系统必须保证正常状态下每个数据块都有冗余保护，才能在某个存储设备节点故障或者其上的磁盘损坏导致数据丢失后，通过剩余节点上的数据恢复丢失的数据. 存储设备节点上数据在网络 RAID 初始化操作后，建立 RAID5 冗余关系；当存储设备节点上某个数据块被更新后，它和冗余管理服务器上的镜像块建立 RAID1 冗余关系. 但是其它节点上相同的 RAID5 条带中的数据块缺少了冗余保护，也就是说此时新写块所在的节点故障，丢失的数据可以用冗余管理服务器上的镜像块数据恢复，但是如果其它节点故障丢失数据，就不能利用剩余节点上的数据恢复.

因此 BW-netRAID 在冗余管理服务器上保留旧数据块，在冗余管理服务器和其它存储设备节点间维持原来的 RAID5 冗余关系. 只有当后台 RAID5 计算新校验块并将其写到对应存储设备节点上后，才重新在多个存储设备节点间形成新的 RAID5 冗余关系. 图 11 描述了一个数据块的 RAID1/RAID5 冗余级别变换的过程. 如果一个 RAID5 条带的多个块都被更新，那么对应的旧数据块也都被缓存到冗余管理服务器上. 当其它存储设备节点故障丢失数据时，这些旧数据块将作为一个整体参与 RAID5 恢复计算. 如果冗余管理服务器故障丢失了这些旧数据块，则只需重新计算同步的这个 RAID5 条带，并不会导致数据丢失.

6.2 数据可靠性分析

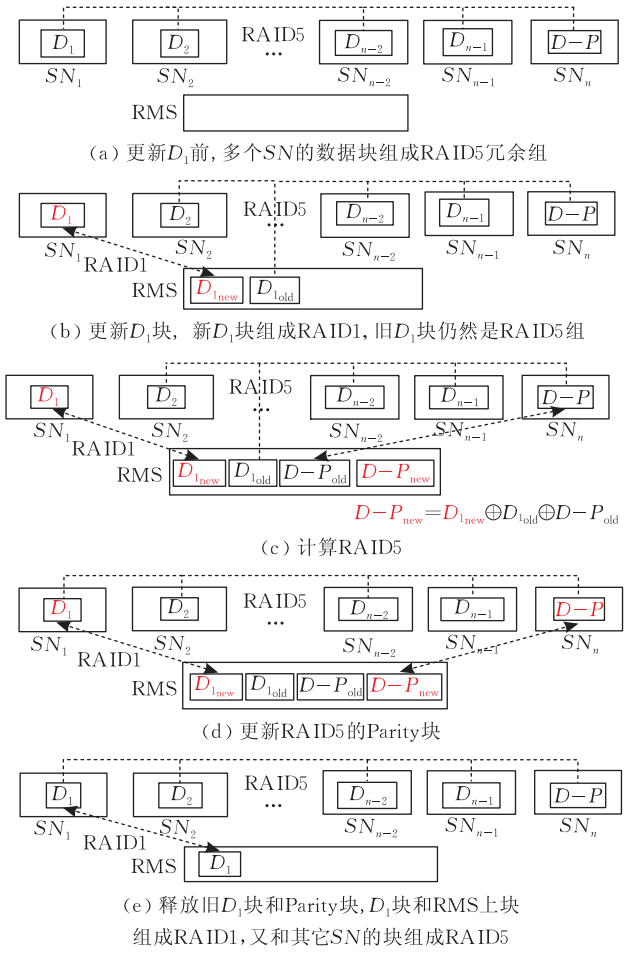
将数据块按照 RAID1/RAID5 的不同布局分类讨论数据块的平均丢失时间. $MTTF_{node}$ 是节点平均失效时间，$MTTR_{mirror}$ 是从冗余管理服务器上磁盘日志缓存中恢复镜像数据块的时间，$MTTR_{RAID5}$ 是冗余管理服务器从其它存储节点读取数据然后 RAID5 重构恢复数据的时间.

(1) 如图 11(a)所示，N 个存储设备节点上的数据块建立 RAID5 冗余关系. 此状态下数据块 D_1 属于由 N 个存储节点上的数据组成的 RAID5，数据可靠性为

$$MTTDL=\frac{MTTF_{node}^2}{N\cdot(N-1)\cdot MTTR_{RAID5}}.$$

(2) 如图 11(b)所示，刚做完写更新的数据块 D_1 在存储设备节点和冗余管理服务器上各有一份数据拷贝，并且在冗余管理服务器上也保留了其对应的旧块 $D_{1_{old}}$. 此状态下块 D_1 属于由所在存储设备节点和冗余管理服务器组成的 RAID1，数据可靠性为

$$MTTDL=\frac{MTTF_{node}^2}{2\cdot MTTR_{mirror}}.$$



(a) 更新D_1前，多个SN的数据块组成RAID5冗余组

(b) 更新D_1块，新D_1块组成RAID1，旧D_1块仍然是RAID5组

(c) 计算RAID5

(d) 更新RAID5的Parity块

(e) 释放旧D_1块和Parity块，D_1块和RMS上块组成RAID1，又和其它SN的块组成RAID5

图 11 RAID1/RAID5 冗余级别变换过程

(3) 如图 11(e)所示，RAID5 计算后，新校验块拷贝到对应的存储设备节点，并且在冗余管理服务器上保留数据块 D_1. 此状态下 D_1 块既属于由 N 个存储设备节点上的数据组成的 RAID5，而且也属于由 D_1 所在的存储设备节点和冗余管理服务器组成的 RAID1. 图 12 给出了活跃数据的 Markov 状态变化图，状态 0 表示无故障状态，状态 1 表示冗余管理服务器或 D_1 所在存储设备节点之一故障，状态 2 表

示冗余管理服务器和 D_1 所在存储设备节点都故障，状态 3 表示另外一个存储设备节点故障，状态 4 表示另外一个存储设备节点和冗余管理服务器或 D_1 所在存储设备节点之一故障，状态 F 表示冗余管理服务器和 D_1 所在存储设备节点都故障而且又有另外一个存储设备节点故障或者两个存储设备节点同时故障. 状态 F 时数据发生丢失. 因此根据活跃数据的 Markov 状态转换图，计算得到其数据可靠性为

$$MTTDL=MTTF_{\text{node}}^{3}/(2(N-1)\cdot MTTR_{\text{mirror}}\cdot [N\cdot MTTR_{\text{RAID5}}+(N-1)\cdot MTTR_{\text{mirror}}]).$$

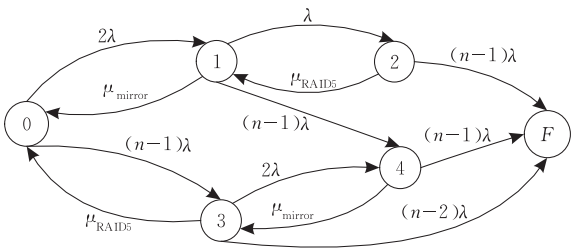


图 12　活跃数据的 Markov 状态转换图

假设存储设备节点个数为 $N=10$；每个存储设备节点的数据块个数为 C（总数据量为 4TB）；一个存储设备节点在冗余管理服务器上的活跃数据块占其总数据块的比例为 $r=0.1$；从磁盘日志恢复和 RAID5 计算恢复的速度分别为 $v_1=50$MB，$v_2=v_1/1.5$；平均恢复时间为数据块数量除于恢复速度. 根据上面的数据块分类和可靠性公式，图 13 出示一组随存储设备节点平均无故障时间变化的数据可靠性数值，缓存到冗余管理服务器上活跃数据块比普通数据块的数据可靠性高 3～5 个数量级.

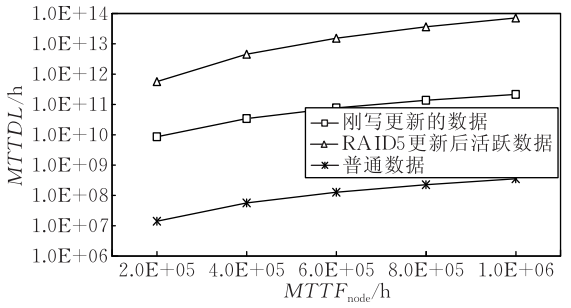


图 13　不同类别数据的可靠性数值

7　相关研究

本节首先讨论存储设备节点间采用数据冗余的必要性，然后对比分析典型的网络存储系统中的冗余管理方法. 最后比较 BW-netRAID 的采用 RAID1/RAID5 异构分布和日志方式缓存数据的性能优化方法.

7.1　存储节点间数据冗余必要性

文献[1]详细分析了单个冗余难以保证大型存储系统足够的数据可靠性，进而讨论了采用多级冗余（如 mirror2，mirror3，RAID5＋1）后的存储系统的数据可靠性. 文献[2]则分析了 RAID0、RAID1 和 RAID5 算法组合得到的双层 RAID 的可靠性模型和性能. 文献[3]也是通过组合不同 RAID 算法的方式，分析比较了由通用部件“IBM brick”组成存储系统的可靠性，并且通过仿真方式分析了不同参数对可靠性的影响. 综上所述，在大规模存储系统中存储节点间采用数据冗余技术能显著地提高整个系统的数据可靠性.

7.2　冗余管理

在存储系统使用镜像冗余的方式，可以避免校验块计算的开销，但用户数据占用存储系统资源的总量变为原来的 2 倍，系统的性价比较低. 因此多采用 RAID5 方式提高系统整体性价比，但冗余管理的问题也变得更为复杂. 冗余管理重点解决在哪个节点如何计算 RAID 冗余信息和为计算冗余信息如何传输数据的问题.

在传统的前端集中主从冗余管理方式中，系统需要有一个集中的控制点（如 AutoRAID[4]），负责冗余计算和数据传输，这种方式不需要复杂的管理协议，但问题是为避免成为性能瓶颈点而必须配置一个性能远高于存储设备节点的控制服务器，系统成本变高. 如果采用分布式对等存储管理方式（如 RAID-x[5]），不需要集中管理服务器，任意两个存储设备节点之间都能相互并发通信，但是需要采用严格的锁机制提供单一的地址空间（Single I/O Space Image），来保证用户数据读写的一致性. 如果在应用服务器上计算 RAID5 校验块（如 NetRAID[6]），再将数据块和校验块传输到存储节点，不但对应用服务器性能影响大，而且也需要采用锁机制保证多个服务器并发写的数据一致性. 如果在存储设备节点上计算 RAID5 校验块（如 ClusterRAID[7]），数据经存储节点转发，用户请求的延迟较大，而且节点间数据传输占用存储系统的网络带宽多.

冗余计算和数据传输过多占用系统资源而影响应用性能，现有的冗余存储管理架构难以满足系统的需求. 表 2 给出了几种典型存储系统的冗余管理方式对比.

表 2 冗余管理方式对比

系统名称	管理方式	冗余计算方式
AutoRAID	前端集中管理	控制器集中计算 RAID5
RAID-x	无管理服务器，存储设备节点通信采用锁方式管理	镜像，无冗余计算
BW-netRAID	带外存储管理，后端集中冗余管理	冗余管理服务器集中计算 RAID5
NetRAID	没有管理服务器	应用服务器集中计算
ClusterRAID	没有管理服务器	存储设备节点集中计算
NSSM[8]	元数据集中管理	存储设备节点上的 RAID 模块独立计算
VNS-II[9]	2 个控制服务器在前端集中管理	控制服务器集中计算 RAID5

BW-netRAID 是在带外虚拟化存储管理系统基础上提出的新的后端集中冗余管理的存储结构：

(1) 对于冗余计算问题，BW-netRAID 采用独立的后端管理服务器在后台集中计算，避免了在数据通道上集中计算引起的性能瓶颈问题，也避免了每次读写过程中存储设备节点间的锁协议开销问题.

(2) 对于数据传输负荷问题，BW-netRAID 采用在后端冗余管理服务器上缓存活跃数据，只在冗余计算缺少所需数据时才从存储设备节点读数据，减少了存储设备节点的过多输入和输出数据问题，以少量空间代价换取到更高的性能和可靠性.

(3) 优化读写请求处理，使得读请求只需直接访问存储设备节点，读性能不再受限于前端集中控制器性能，具有分布式存储的高扩展性. 而且通过日志方式在磁盘上缓存写请求数据，再异步计算校验块的方法，有效地提高了写请求的性能.

7.3 RAID1 和 RAID5 异构分布

表 3 对比了不同存储系统的 RAID1 和 RAID5 方式. 上述系统的数据变化的共同特征是新修改的活跃数据先写到 RAID1，不活跃数据再后台迁移或者重新计算得到 RAID5，数据块关系都先是 RAID1 再变为 RAID5. 它们之间的区别在于镜像块和校验块的存储方式. AutoRAID 采用分级存储，Hot mirroring[10] 采用交叉分布存储在磁盘不同的区域，DPGADR[11] 和 BW-netRAID 的镜像块是单独存储到另外的存储设备上，但 DPGADR 中非活跃数据是降级 RAID5，而 BW-netRAID 中是标准 RAID5.

表 3 RAID1/RAID5 异构分布对比

系统名称	存储方式	RAID1/RAID5 异构分布
AutoRAID	分级存储	上层逻辑空间是 RAID1，下层逻辑空间是 RAID5
Hot mirroring	交叉分布	磁盘设备的上半部用于 RAID1，下半部用于 RAID5
DPGADR	镜像块存储在复制节点	活跃数据是 RAID1，非活跃数据是降级 RAID5
BW-netRAID	镜像块缓存在冗余管理服务器	冗余管理服务器缓存的活跃数据和存储设备节点相关数据是 RAID1，存储设备节点间数据是 RAID5

7.4 数据块和校验块日志

表 4 对比了几种数据块和校验块日志技术. 通过日志方法缓存数据，然后后台再计算 RAID5 是上述方法的共同特征. 区别在于缓存哪类数据和存储的方式. Logging RAID[12] 仅仅缓存了新写的数据到日志区，当重新计算 RAID5 检验块时仍然需要从数据区读回旧数据块. Parity logging[13] 仅仅缓存了校验块的更新数据，Data logging[14] 缓存校验块更新数据或者新数据，当重新计算 RAID5 检验块时仍需读回旧校验块. BW-netRAID 用磁盘缓存日志来增大日志区，不但缓存新数据，而且还尽可能地缓存旧数据和旧检验块，因此以少量空间代价减少存储设备节点间数据重读的性能开销，而且也提高了数据可靠性.

表 4 数据块和校验块日志技术对比

	数据区	日志区
Logging RAID	旧数据块	新数据块
Parity logging	新数据块	新数据块与旧数据块的异或数据块
Data logging	新数据块	旧数据块，新数据块
BW-netRAID	新数据块	旧数据块，新数据块，旧校验块

8 结 论

本文提出了一种新的后端集中冗余管理的网络 RAID 存储系统. 通过带外冗余管理的方式解决了前端冗余管理的性能瓶颈问题，并且系统根据数据更新的活跃度采用 RAID1 和 RAID5 异构分布以及活跃数据日志缓存的方法，以数据缓存的空间代价换取系统的整体性能和可靠性提升.

在目前工作的基础上我们还将在以下 3 个方面继续深入研究：

(1) 数据读写和冗余管理分离的分布式网络 RAID 系统；

(2) 借助工作负载特征分析，改进活跃数据缓存和资源回收算法，根据访问模式调整 RAID1/RAID5 的数据块分布；

(3) 支持多重数据冗余功能，能够容忍多个存储设备节点故障.

参 考 文 献

[1] Xin Q, Miller E L, Schwarz T, Long D D E, Brandt S A, Litwin W. Reliability mechanisms for very large storage systems//Proceedings of the 20th IEEE/11th NASA Goddard Conference on Mass Storage Systems and Technologies. San Diego, California, USA, 2003: 146-156

[2] Baek S H, Kim B W, Joung E J, Park C W. Reliability and performance of hierarchical RAID with multiple controllers//

Proceedings of the 20th Annual ACM Symposium on Principles of Distributed Computing. Newport, Rhode Island, USA, 2001: 246-254

[3] Rao K K, Hafner J L, Golding R A. Reliability for networked storage nodes//Proceedings of the International Conference on Dependable Systems and Networks. Philadelphia, Pennsylvania, 2006: 237-248

[4] Wilkes J, Golding R, Staelin C, Sullivan T. The HP AutoRAID hierarchical storage system. ACM Transactions on Computer Systems, 1996, 14(1): 108-136

[5] Hwang Kai, Jin Hai, Hoy R S C. Orthogonal striping and mirroring in distributed RAID for I/O centric cluster computing. IEEE Transactions on Parallel and Distributed Systems, 2002, 13(1): 26-44

[6] Sobe P. Reconfiguration of RAID-like data layouts in distributed storage systems//Proceedings of the 18th International Parallel and Distributed Processing Symposium, Workshop on Fault-Tolerant Parallel and Distributed Systems. Santa Fe, New Mexico, USA, 2004: 212-219

[7] Wiebalck A. ClusterRAID: Architecture and prototype of a distributed fault-tolerant mass storage system for clusters [Ph. D. dissertation]. Ruprecht-Karls-University of Heidelberg, Germany, 2005

[8] Zhang Hong-Can, Xue Wei, Shu Jiwu. An expandable distributed RAID storage cluster system. Journal of Computer Research and Development, 2008, 45(4): 741-746(in Chinese)
(章龙灿，舒继武，薛巍. 一种可扩展分布式 RAID 存储集群系统. 计算机研究与发展，2008，45(4)：741-746)

[9] Wu Ying, Wang Gang, Liu Jing. Design and implementation of network softRAID system based on dual center-node. Computer Engineering, 2006, 32(8): 73-75(in Chinese)
(吴英，王刚，刘璟. 双中心节点的网络软 RAID 系统的设计与实现. 计算机工程，2006，32(8)：73-75)

[10] Mogi K, Kitsuregawa M. Hot mirroring: A method of hiding parity update penalty and degradation during rebuilds for RAID5//Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data. Montreal, Quebec, Canada, 1996: 183-194

[11] Chandy J A. Parity redundancy strategies in a large scale distributed storage system//Proceedings of the 21st IEEE/12th NASA Goddard Conference on Mass Storage Systems and Technologies. Greenbelt, Maryland, USA, 2004: 185-191

[12] Chen Ying, Hsu W W, Young H C. Logging RAID- an approach to fast, reliable, and low-cost disk arrays//Proceedings of the 6th International Euro-Par Conference. Munich, Germany, 2000: 1302-1312

[13] Stodolsky D, Gibson G A, Holland M. Parity logging: Overcoming the small write problem in redundant disk arrays//Proceedings of the 20th Annual International Symposium on Computer Architecture. San Diego, California, USA, 1993: 64-75

[14] Gabber E, Korth H F. Data logging: A method for efficient data updates in constantly active RAIDs//Proceedings of the 14th International Conference on Data Engineering. Orlando, Florida, USA, 1998: 144-153

NA Wen-Wu, born in 1977, Ph. D. candidate. His research interests include network virtual storage and data redundancy.

KE Jian, born in 1971, Ph. D. candidate. His research interests include virtual storage and adaptive performance management.

ZHU Xu-Dong, born in 1979, Ph. D. candidate. His research interests include virtual storage and I/O access pattern recognition.

MENG Xiao-Xuan, born in 1980, Ph. D. candidate. His research interests include network storage and cache management.

BU Qing-Zhong, born in 1972, Ph. D. , assistant professor. His research interests include network storage and storage management model.

XU Lu, born in 1962, Ph. D. , professor, Ph. D. supervisor. His research interests include network storage and file system.

Background

All customers of the storage systems wish that their data can be accessed correctly and quickly. The storage system must provide data redundancy features and solve the data loss problems because of the unavoidable failures in the large computer system. But the network storage systems with frontend centralized redundancy management have the serious performance bottleneck problem. This paper presents a novel network RAID storage system with backend centralized redundancy management based on the out-of-band storage virtualization architecture. It can aggregate I/O performance of all storage device nodes and relieve the performance bottleneck. And it can also balance the performance, reliability and cost of the whole storage system.

The research proposed in this paper is a part of work supported by the National Basic Research Program (973 Program) of China (2004CB318205) and the National High Technology Research and Development Program (863 Program) of China (2007AA01Z402 and 2009AA01A403). The research team has been focusing on network storage technology and already implemented the network storage system: BW-VSDS. It also achieved many advanced functions: in-of-band metadata management and out-band data transferring, dynamic resource allocation, virtual snapshot framework, network RAID, performance insulation of virtual device, I/O access pattern recognition and block reorganization, cache algorithms for storage service.