

RAID5 小写更新的冗余管理机制

刘秀菊

(菏泽学院 计算机与信息工程系, 山东 菏泽 274000)

摘 要: 为改进传统 RAID5 机制小写性能低下问题, 提出了一种高性能的冗余数据读写机制 RMB(Redundancy Management for RAID5). RMB 采用异步冗余校验方法 APC(Asynchronous Parity Calculating), 对现有 RAID5 更新算法的关键步骤进行了优化, 将统一计算校验块的同步操作拆分成同步镜像写更新和异步冗余计算更新. 该操作分解了数据与其校验数据更新的紧耦合现状, 提升了写操作并发度; 通过合并多个版本镜像块写操作减少异或计算操作的次数. 通过实验验证表, RMB 机制, 可将小写性能大幅提高, 峰值可达 4 倍以上.

关键词: 磁盘冗余阵列; 冗余数据; 小写更新; 异步冗余计算

中图分类号: TP751

文献标识码: A

文章编号: 1000-7180(2012)07-0112-04

A Redundancy Management Policy For Raind5

LIU Xiu-ju

(Department of Computer and Information Engineering, Heze University, Heze 274000, China)

Abstract: To solve the problem that RAID5 performance ineffective for small writes, a new high performance read and write mechanism for redundancy data is brought up, which is called RMB. By Asynchronous Parity Calculating, the mechanism RMR optimized the write-update mechanism of tradition al RAID5 and decomposed the concentrated operations of calculating parity blocks into writing mirror and asynchronous calculating. The new algorithm accomplish the decomposition of the tight data, improve the concurrency state and decrease the count of calculate by the way of merge ring several versions. It shows from experiments that RMR will substantially improve the performance of small writes, the peak value will be over four times as before under the out-of-band mode.

Key words: RAID; redundancy data; update small writes; asynchronous parity calculating

1 引言

随着信息技术的迅猛发展, 人类进入了数据爆炸的时代, 大规模存储系统正在面对大容量、高性能、高可靠和高可用保障以及高空间利用率多方面挑战^[1]. RAID 技术很好地解决了设备级数据的可靠性和可用性, 其中 RAID5 是一种兼顾了存储性能、数据安全和存储成本的解决方案, 得到了广泛的研究和应用^[2]. 但 RAID5 存在着小写更新性能低下问题的严重缺陷^[3]. 解决 RAID5 小写性能低下的问题, 文献^[4]提出了 DGP(Dynamic Parity Grouping)机制, 文献^[5]以数据版本模块管

理新数据块和冗余校验数据块, 以数据存储模块管理多个镜像块, 通过缓存替换操作推动异步计算冗余校验数据块, 从而实现 RAID5 的异步冗余计算方法. 在总结前人研究成果的基础上, 设计了一种高性能的冗余管理机制 RMR(Redundancy Management for RAID5). RMR 通过异步冗余计算方法 APC(Asynchronous Parity Calculating), 对传统 RAID5 写更新机制的关键技术进行了优化, 将集中计算校验块的同步操作分解为同步的并行镜像写操作和异步冗余计算操作. 该操作分解了数据与其校验数据更新的紧耦合特性, 提高了写操作并发度; 通过合并多个版本镜像块写操作减少

收稿日期: 2011-11-25; 修回日期: 2011-12-28

基金项目: 山东省自然科学基金项目(Y2011A03)

异或计算操作的次数. 试验结果表明, 可将小写性能大幅提高, 峰值可达 4 倍以上.

2 RAID5 小写性能低下问题分析

造成 RAID5 小写性能低下的原因主要有两点, 下面分别给出分析.

2.1 RAID5 的小写性能低下

RAID5 写更新可分为如下三个模式^[6]: 全条带写(Full-stripe Write) 重构写(Reconstruct Write) 读修改写(Read-Modify Write).

其中, 读修改写(Read-Modify Write)是导致小写性能低下的主要原因: 如果需要写更新的数据块数目不足冗余校验组宽度的一半, 这类小请求写操作只能采取读修改写方式. 在读修改写模式中, 先从冗余校验组中读取需要修改的旧版本数据块 D_i^j ; 再读取旧的冗余校验数据块 P^k ; 根据旧版本数据块 D_i^j 、旧版本冗余校验数据块 P^k 和新版本数据块 D_i^{j+1} , 计算得到新版本冗余校验数据块 P^{k+1} , 满足式(1):

$$P^{k+1} = D_i^j \oplus D_i^{j+1} \oplus P^k \quad (1)$$

之后将新版本数据块 D_i^{j+1} 和新版本冗余校验数据块 P^{k+1} 写回. 小写的读修改写操作比重构写操作效率更低, 因为一次读修改写需要两次读操作(读出旧版本数据块和旧版本冗余校验数据块)、两次冗余校验计算和两次写操作(写新版本数据块和新版本冗余校验数据块)才可以完成.

2.2 RAID5 的并发写操作性能低下

由于传统 RAID5 机制采用冗余校验组内一个冗余校验数据块对应多个数据块的方式, RAID5 机制的存储资源空间利用率比 RAID1 高, 因此存储成本相对较低^[7]. 但是为了保证冗余校验组内数据和冗余校验数据的冗余校验关系一致性, 数据进行写更新时, RAID5 需要同时对冗余校验数据进行更新. 也就是说, 数据块和冗余校验数据块必须保证同步修改才能保证冗余校验组内冗余校验关系一致性, 这种紧耦合关系影响了 RAID5 写更新的并发操作. 数据块更新和冗余校验数据块更新的这种紧密耦合特性, 极大地影响了 RAID5 并发写操作的性能, 并发写请求较多时, 会导致系统性能严重下降. 在大规模网络分布是存储系统中, 由于存在网络转发延迟大等特性, 这种影响表现更为明显.

3 冗余管理机制 RMR

为了解决 RAID5 小写性能低下问题, 提高系统写性能, 设计了一种高性能的冗余管理机制 RMR (Redundancy Management for RAID5). RMR 通过异步冗余计算方法 APC (Asynchronous Parity Calculating) 对冗余校验组进行更新, APC 方法将集中计算分解为并行写操作和异步冗余计算两个部分, 提高了写操作并发度; 而且通过合并多个写操作, 减少了异或计算操作的次数, 也就减少了对底层设备的 I/O 访问次数, 提高了系统的写性能.

RMR 在对 RAID5 冗余校验组进行写操作时, 可分为三步操作, 下面分别进行描述.

(1) 对冗余校验组内的数据块写更新

当需要对冗余校验组内的数据块第一次进行写更新时, 按照如下步骤操作:

① APC 将新版本数据块 D_i^1 写入旧版本数据块 D_i^0 (D_i^0 通常表示为 D_i) 同一个存储设备 SN_i 中;

② 同时将新数据块的镜像 D_i^1 写到另一个存储设备 SN_i' 中, 形成 RAID1 镜像, 为数据提供冗余保护;

③ 同时, APC 需要保留和旧冗余校验数据块 P 呈紧耦合关系的旧版本数据块 D_i . 数据块与其耦合的校验块必须同时更改, 旧校验块被覆盖写后才能删除旧版本数据块.

更新前后如图 1 中(a)(b)所示.

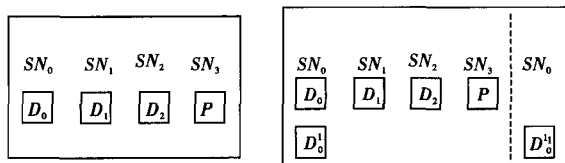


图 1 冗余校验组状态之一

异步冗余计算方式(APC)将计算校验块的同步操作分解为并行镜像写操作和延迟进行的异步冗余计算操作. 并行镜像写操作只需同步写数据块和镜像块, 共有 2 次 I/O 写操作, 低于同步冗余计算方式的 I/O 操作次数.

(2) 对经过写更新的冗余校验组内数据块进行写更新

对冗余校验组内数据重复进行写操作时, 分为两种情况: 对冗余校验组内尚未写更新的数据块进行修改时, 方式同上; 对冗余校验组内已经更新过的数据块 D_i^j 进行修改时, 需要如下所示三个步骤:

① 用新版本数据块 D_i^{j+1} 更新一版本数据块 D_i^j ;

② 同时用新版本数据块 D_i^{j+1} 的镜像数据块 D_i^{j+1} 更新 D_i^j , 形成 RAID1 镜像模式, 为数据提供冗余保护;

③ 同时, APC 需要保留和旧冗余校验数据块 P 呈紧耦合关系的旧版本数据块 D_i . 数据块与其耦合的校验块必须同时更改, 旧校验块被覆盖写后才能删除旧版本数据块.

更新前后如图 2 中(a)(b)所示.

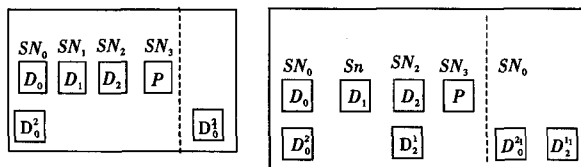


图 2 冗余校验组状态之二

(3) 对经过写更新的冗余校验组进行冗余校验计算并更新冗余校验数据块时, 步骤如下:

① 读取最新的数据块(包括冗余校验组内尚未写更新的数据块和冗余校验组内已经更新过的新版本数据块);

② 读取旧冗余校验块 P^k ;

③ 计算新的冗余校验数据块 P^{k+1} , 并用新冗余校验数据块更新原有冗余校验数据片 P^k ;

④ 当新的冗余校验组重新建立后, 方可删除被更新过的旧版本数据块, 系统重新得到一个新的 RAID5 冗余校验组, 保持数据冗余一致性;

⑤ 删除 RAID1 镜像数据块 $D_0^{2'}$ 和 $D_0^{1'}$, 释放存储资源.

以下图 3 所示的一个冗余校验组为例说明: 在需要进行冗余计算操作时, 读取最新版本的数据块 D_0^2 、旧版本数据块 D_0 和旧冗余校验块 P ; 计算新的冗余校验数据块 p^1 ; 用新冗余校验数据块 p^1 更新原有冗余校验数据片 P ; 当新的冗余校验组重新建立后, 方可删除旧版本数据块 D_0 和 D_2 , 系统重新得到一个新的 RAID5 冗余校验组, 保持数据冗余一致性; 删除 RAID1 镜像数据块 $D_0^{2'}$ 和 $D_2^{1'}$, 释放存储资源.

之所以旧冗余校验数据块被更新后才能删除旧版本数据块, 是由于冗余校验组内数据块和冗余校验数据块更新的紧耦合特性: 冗余组内部数据块和冗余校验数据块必须保证数据冗余校验的一致性, 因此数据块和耦合的校验块必须同时更改. 异步冗余计算(APC)方式将冗余校验数据块的计算和更新

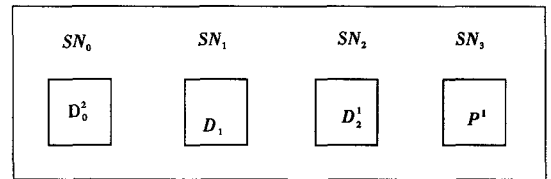


图 3 冗余校验组状态之三

延迟, 如果更新数据块的同时没有保留旧版本数据块, 则冗余组内数据块和冗余校验数据块的冗余关系一致性遭到破坏.

4 RMR 与 RAID5 性能对比测试

为了验证冗余管理机制 RMR 解决 RAID5 小写性能低下问题的效果, 用 mdadm 创建了软 RAID5, 然后在 RAID5 上建立 EXT3 文件系统, 进行性能对比测试. 通过对小写性能进行比较, 可以客观地观察到采用高性能冗余管理机制 RMR 相对于传统 RAID5 机制在小写性能上的优势.

4.1 测试 Benchmark

采用 IOzone [Norcott 2003] 基准测试程序作为测试工具. IOzone 是一个文件系统的基准 Benchmark 工具, 可以测试不同的操作系统中文件系统的读写性能. IOzone 的配置非常丰富, 可以高效地定义负载属性如文件大小、I/O 大小和访问模式. 所以可以比较准确的模拟特定负载. 目前 IOzone 可以测试 Read, write, re-read, re-write, read backwards, read strided, fread, fwrite, random read, pread, mmap, aio_read, aio_write 等等不同的模式下的硬盘的性能.

4.2 测试环境

(1) 文件系统配置

条带宽度为 3(对应 3 块硬盘, 创建 3 个数据卷); EXT3 所采用的资源分配粒度 256K;

(2) 软 RAID5 配置

通过 mdadm 实现, 版本 3.2.1; mdadm 是一个单一的程序, 创建、管理 RAID 方便, 运行稳定. mdadm 的条带粒度默认为 64K; 支持 RAID0、RAID1、RAID4、RAID5、RAID6; 3 块 146GB SCSI 接口的磁盘, 构成一个软 RAID5 的虚拟设备, 对应条带宽度为 4.

4.3 测试设计与结果分析

为了保证测试的公平性, 将 EXT3 的条带宽度设置为与实际使用的数据卷一致, 从而保证和 RMR 的条带宽度一致. 用 3 块数据盘创建软 RAID5, 并在 RAID5 的基础上实现了 RMR 机制,

进行写读操作,利用 IOzone 以不同的请求粒度作随机写读测试,收集测试数据;再在 3 块数据盘上建立数据卷,创建 EXT3 文件系统,利用 IOzone 以不同的请求粒度作随机写读测试,收集测试数据.为了消除缓存对读写操作的性能影响,测试文件大小设为内存容量的 2 倍,即 4GB.随机读写部分测试结果如图 4 所示.

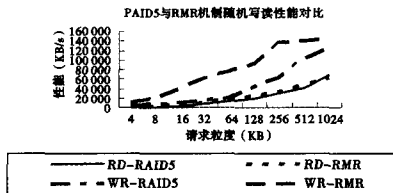


图 4 随机写读性能对比

(1) 随机写性能分析

在条带宽度同样为 3 的情况下,比较 RMR 和 RAID5 的随机小写性能,RMR 最高情况下达到普通 RAID5 性能的 4.25 倍(如表 1 所示).随着块粒度的增大,普通 RAID5 开始以全条带写方式进行写更新,小写问题得到回避普通 RAID5 的随机写性能逐渐与 RMR 缩小差距.

表 1 不同粒度下随机小写性能增速比

条带宽度	4K	8K	16K	32K	64K	128K	256K	512K	1024K
优化增速比(倍)	2.84	2.93	3.61	4.25	3.23	2.1	2.16	1.40	1.19

利用 iostat 监测各磁盘 I/O 状况,RAID5 在较小粒度的随机写操作过程中,各磁盘始终伴随读流量发生,各磁盘聚合写带宽高于 IOzone 测得的实际写带宽(目测估计基本符合 2:1 的关系),这是由于普通 RAID5 小写操作时,每写一次数据,需要两次读操作和两次写操作.而 RMR 写操作没有伴随出现读流量.

观察表明普通 RAID5 小写更新机制对其性能影响确实明显,符合此前的分析.测试数据表明,随机小写情况下,RMR 优势明显.

(2) 随机读性能分析

在条带宽度同样为 3 的情况下,比较 RMR 和 RAID5 的随机小读性能,RMR 略高但优势不明显.经过分析判断,这是因为文件系统在读操作过程中采用了预读机制,预读可以有效的减少磁盘的寻道次数和应用程序的 I/O 等待时间,但随机小写使预读机制无法发挥其功能.RMR 的 RAID1 模式写更

新和普通 RAID5 的写更新都对磁盘的连续性有所破坏,也对预读机制造成影响.RMR 性能稍高的原因可能是 RAID5 系统为了提高性能自身实现了条带缓存,占用内存资源,使得其系统缓存可利用资源减少.因此 RMR 的随机小读性能也具有优势.

5 结束语

为了解决传统 RAID5 机制小写性能低下问题,提高系统性能,设计了一种高性能的冗余管理机制 RMR(Redundancy Management for RAID5).RMR 通过异步冗余计算方法 APC(Asynchronous Parity Calculating),对传统 RAID5 写更新机制的关键技术进行了优化,将集中计算校验块的同步操作分解为同步的并行镜像写操作和异步冗余计算操作.该操作分解了数据与其校验数据更新的紧耦合特性,提高了写操作并发度;通过合并多个版本镜像块写操作减少异或计算操作的次数.RMR 的冗余管理机制解决了 RAID5 小写性能低下问题,从而提高了系统性能.

参考文献:

- [1] Sung Hoon Baek, Kyu Ho Park. Matrix-stripe-cache-based contiguity transform for fragmented writes in RAID-5[J]. IEEE Transaction Computers, 2009, 56(8): 1040-1054.
- [2] Sung Hoon Baek, Kyu Ho Park. Prefetching with adaptive cache culling for striped disk arrays [C]// USENIX. Boston, 2008: 363-376.
- [3] John F Gantz, David Reinsel. Extracting value from chaos [M]. Michelle Baiky; Microsoft, IDC white paper-sponsored by EM, 2011.
- [4] Andrew W Leung, Shankar Pasupathy, Garth Goodson, et al. Measurement and analysis of large-scale network file system workloads [C]// USENIX 2008 Annual Technical Conference on Annual Technical Conference. Boston, 2008: 213-226.
- [5] John L Hennessy, David A Patterson. Computer architecture [M]. 4th ed. A Quantitative Approach, Morgan Kaufmann Publishers, 2009.
- [6] 刘军. 容许多个磁盘故障的 RAID 编码方法研究[J]. 微电子学与计算机, 2011, 28(12): 133-134.
- [7] 任建奇, 蔡皖东. 一种虚拟存储系统的设计与实现 [J]. 微电子学与计算机, 2009, 26(5): 82-84.

作者简介:

刘秀菊 女, (1973—), 博士, 教授. 研究方向为文件系统、计算机网络及应用.