

基于分组的 NandFlash 块管理方法

胡 宁, 杨 琼, 王 冬

(中航工业西安航空计算技术研究所, 陕西 西安 710065)

摘 要: 为有效管理 NandFlash 存储空间, 提出了单替换块组相联映射方法. 将 NandFlash 擦除块的集合划分成若干虚拟组, 逻辑空间和虚拟空间映射时, 组内的页任意映射, 组间一一映射. 一个虚拟组只设置一个替换块, 用于组内无效页的回收和重用. 该方法能够保持较高的空间利用率, 垃圾回收时间确定. 利用存储设备访问的局部性, 仅把需要的组映射表装入内存, 内存开销小. 分析表明, 虚拟组大小为 16 个擦除块时工程实践是合适的.

关键词: 闪存; 块映射; 垃圾回收; 文件系统; 嵌入式系统

中图分类号: TN402

文献标识码: A

文章编号: 1000-7180(2015)03-0019-04

NandFlash Memory Management Based on Blocks Set

HU Ning, YANG Qiong, WANG Dong

(Aeronautics Computing Technique Research Institute, Xi'an 710065, China)

Abstract: In order to manage the NandFlash memory space effectively, a set associative mapping method with single replacing block is presented. The NandFlash erasable blocks is divided to several virtual block sets. When the logical space is mapped to the virtual space, the logical pages in one block set are mapped to any virtual pages in the same one block set, and the logical sets are mapped to the virtual sets one by one. Only one replacing block is set in one virtual set, which is used to withdraw and reuse the invalid pages. This method can raise the storage space utilization ratio and assure the time consumptions of the garbage withdrawing. Only the required mapping tables have to load into the memory due to the locality of the data accessing, which decreases the memory usage. The analysis results show that 16 blocks is appropriate in one virtual set for engineering practice.

Key words: flash; block mapping; garbage recycle; file system; embedded system

1 引言

与传统的机械式存储介质相比, NandFlash 闪存介质具有抗震动、低能耗、随机访问时间短等优势^[1]. 目前, NandFlash 存储介质已广泛使用于从消费类电子产品到安全关键电子设备的各型电子产品, 特别在航空航天领域, 大容量存储设备中一般都使用 NandFlash 介质.

目前, NandFlash 芯片可以做到很大的容量, 但是由于生产工艺和使用环境的关系, NandFlash 芯片在出厂时会存在一定量的坏块. 一般 NandFlash 芯片出厂时, 厂商会将已确认的坏块标记出来. 而

且, NandFlash 在正常使用的过程中也会产生坏块, 一般在某存储块擦除或写入失败时, 管理软件也将这些块标识为坏块, 避免因存储特性不稳定而导致数据损坏.

NandFlash 存储介质还有一个重要特性是数据写入时必须“先擦后写”, 不能“原地改写”, 而且每个存储块擦写次数是有限的(通常在 10 万次左右)^[2], 当接近该擦写极限时, 很容易产生坏块, 此时整个芯片的寿命也即将终结.

因此, 使用 NandFlash 时必须综合考虑上述特性, 进行均衡设计. 其中一个重要的设计考虑就是 NandFlash 的块管理机制, 目标是屏蔽 NandFlash

的专有特性,让用户感知不到坏块存在,可原地改写数据,并且在内存开销、空间利用率、写入延迟等参数间取得一种让用户可接受的平衡.

2 问题的提出

NandFlash 存储介质的访问管理主要需要解决好两个问题:一是存储块擦写寿命有限的问题;二是数据不能“原地改写”的问题.

对前一个问题,文献[1,3-4]中从数据写入平衡角度分别提出了磨损平衡算法,该问题不在本文讨论.本文重点研究“原地改写”和坏块管理问题.

由于坏块的存在,无法在 NandFlash 存储介质上做到逻辑地址到物理地址的一对一映射,必须使用一定的方法建立逻辑地址与物理地址的映射关系.

基于上述需求,Intel 在 1998 年提出了闪存翻译层 FTL(Flash Translation Layer)的概念^[5],以“页”为单位在逻辑页和物理页之间建立直接的映射关系.当时 NandFlash 芯片容量普遍很小,该方法确实很有效.但随着微电子技术的发展,现在的 NandFlash 芯片容量已经非常大,与 1998 年相比容量增大了数千倍,这种基于“页”映射的方式已不适用,否则会占用巨大的内存空间去维护页映射表.

为此,文献[6]中提出了以“块”粒度为单位的映射机制 NFTL(Nand Flash Translation Layer),为每个逻辑块分别分配一个主块和一个替换块,且替换块与逻辑块一一对应.该方法降低了内存开销,但是空间利用率较低.

为提高空间利用率,文献[7]中提出了 FAST(Fully Associative Sector Translation)方法,该方法改进了逻辑块与替换块绑定的方式,允许逻辑块与替换块之间实现任意映射——全相联映射,显著提高了空间利用率.但这种块级的多对多映射方式在进行垃圾回收时效率较低,对一个块的回收往往牵连其他多个块,从而引发连锁效应,垃圾回收的时间也不确定.

因此,需要一种既能管理坏块,又能高效进行页面映射,占用内存空间合理,垃圾回收时间确定的 NandFlash 块管理方法.

3 块管理方案

3.1 块映射管理

针对前述问题,本文提出一种新的单替换块组相联映射方法.该方法与 Cache 的组相联管理思路

类似,即组内任意映射,同时一个组内只设置一个替换块.这样,一能够保持较高的空间利用率;二可以有效缓解垃圾回收的时间确定性问题;三可以利用存储设备访问的局部性,仅把需要的组映射表装入内存,有效降低了内存开销.单替换块组相联映射方法方案要点如下:

(1)将整个 NandFlash 存储空间划分成 n 个逻辑组,每个组包含 m 个逻辑块(这里的块等同于 NandFlash 芯片的块,由固定数量的页组成).这里的逻辑块从用户视角是可见的(见图 1 左边,假设 $m=8$),在用户看来,一个组内的逻辑块是相邻的连续块,各组逻辑块连接在一起组成连续的逻辑块空间.

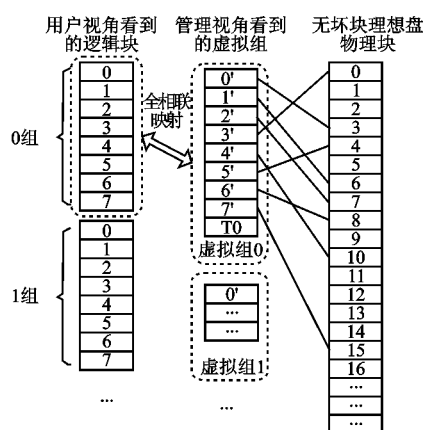


图1 单替换块组相联映射机制

(2)提出虚拟组概念,每个虚拟组与前述用户可见的逻辑组一一对应,进行全相联映射(见图1中间).该虚拟组内的逻辑块数量为 $m+1$ 块,比逻辑组多出1块,但是任意时刻参与全相联映射的块数仍然为 m 块.全相联映射时,逻辑组中的任何逻辑页可以映射到虚拟组中参与映射的 m 块中任意块的任意页上.

(3)每个虚拟组内设置且仅设置一个替换块(就是虚拟组比逻辑组多出的那一个块),并在组内共享(见图1中间虚拟组0中的 T_0 块).该替换块从用户视角是不可见的,仅仅用于组内数据的维护管理.该替换块也并非固定的,也就是说虚拟组内任意一个块都可能成为替换块.替换块的产生算法见3.2节.

(4)虚拟组中的块在物理上不一定是连续的,即使是在无坏块理想盘中也是如此(见图1右边).出于“写均衡”考虑,NandFlash 存储管理软件中必须引入写均衡算法.本文虽不讨论写均衡算法,但不论采用何种算法,必然会导致同一虚拟组中的块物理上不连续的现象.

(5)通过坏块管理机制实现无坏块理想盘中的线性块到真实NandFlash存储块的映射。鉴于NandFlash芯片无法避免出现坏块,且坏块率较低的特点,设计了坏块管理机制见3.3节。

3.2 替换块产生算法

NandFlash存储介质的特性是“先擦后写”,不能“原地改写”,并且擦除的单位是块,而分配的单位是页。该特性要求介质在使用中尽量减少擦除次数,设计中一般采取将改写的的数据写入空闲页,而原废弃数据(无效页)待到不得不擦除时才擦除的策略。

前述块映射管理方法中,在虚拟组内仅设置一个替换块,该替换块的作用就是辅助虚拟组内块管理时进行垃圾块回收。垃圾块回收过程实质就是替换块的产生过程,其算法如下:

时机一:系统初始时,选择虚拟组内任意一个块作为替换块;

时机二:系统运行过程中,仅当某逻辑组(m 个块)没有可写入页时,启动该逻辑组的垃圾块回收活动,借助当前替换块工作并产生新的替换块。最终效果是腾出了可用的空闲页继续进行写操作。该活动步骤如下:

(1)选取虚拟组内所有块中包含无效页最多的块A;

(2)复制块A的所有有效页到替换块;

(3)擦除块A,并将块A标记为替换块。

若考虑写均衡问题,根据某种写均衡算法,可以设计跨虚拟组选择替换块的算法,本文不作讨论。

3.3 坏块管理机制

NandFlash这种存储介质的特性决定了芯片出现坏块是不可避免的,不仅芯片出厂时就可能已经存在坏块,而且在使用过程中也随时可能出现坏块。坏块一旦出现,就必须标识并避免使用。业内认为NandFlash芯片的坏块率一般不超过2%,如果超过,就认为芯片使用寿命已到,需要进行更换。因此,本文利用NandFlash芯片坏块率较低的特点,设计了一种简单的坏块管理机制。

把整个NandFlash芯片的存储区域划分两部分,前一部分是数据区,后一部分是预留区(见图2)。数据区占据了整个存储空间绝大部分,供存放无坏块影响的数据使用;预留区占据剩余存储空间,专门供因坏块而转存的数据使用。不论是数据区中的坏块,还是预留区中的坏块,其替换块都在预留区。

预留区占据NandFlash存储空间尾部的多个

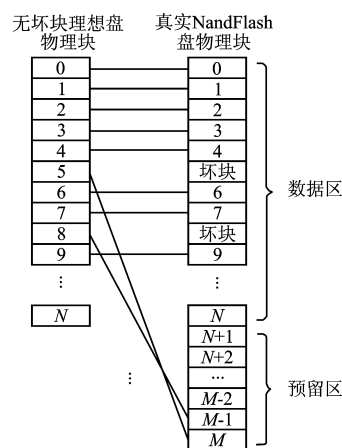


图2 坏块管理示意图

连续物理块,其使用的方式是“从后往前”分配空闲的替换块。该坏块管理机制不会导致较大系统开销,由于NandFlash芯片坏块率较低,预留区不会很大,即使进行一致性维护时需要逐块扫描,仍然可以保持较高的效率。

系统需要维护坏块管理表,记录坏块号和替换块号。该表存放在预留区的固定位置,例如尾部。从容错角度考虑,进行一备一存储。

4 方案参数分析

以大容量存储模块中常用的三星公司的K9NCG08U5M芯片为例分析,该芯片有 32×1024 个块,每块有64个页,每个页有4kB,整个芯片容量为8GB。

如果以页为单位进行全相联管理K9NCG08U5M芯片中的存储空间,则页映射表的全部表项个数为 $32 \times 1024 \times 64$ 个。以每个表项占用4个字节的空间计算,整个页映射表就要占用 $32 \times 1024 \times 64 \times 4 \text{ B} = 8 \text{ MB}$ 的内存空间。目前典型机载大容量模块的存储空间已超过128GB,即使按128GB计算,也需要128MB的内存空间存放全部映射表,显然占据这样大的内存空间是不允许的。

如果按照本文的方案进行组相联管理,假设虚拟组包含 m 个块($m \in [1, 32 \times 1024]$,且 $m = 2^n$),随着 m 的增大,每个虚拟组页映射表占据的内存空间 S 也越大(见表1)。

表1 虚拟组规模与映射表大小的关系

$m/\text{块}$	8	16	32	512	4096
S/kB	2	4	8	128	1024

单从页映射表占用内存空间大小角度, m 越小越好。根据存储访问局部性原理,只需要把少数虚拟

组空间的映射表维护在内存里.但是, m 越小,虚拟组个数越多,所需的替换块 T 越多,空间利用率越低.图 3 中给出了虚拟组大小与空间利用率的关系.可以看出,当 $m=8$ 时,空间利用率低于 90%,不可接受;而 $m=16$ 时,空间利用率已达 94%.因此,从空间利用率方面,选择 $m \geq 16$ 都是可接受的.

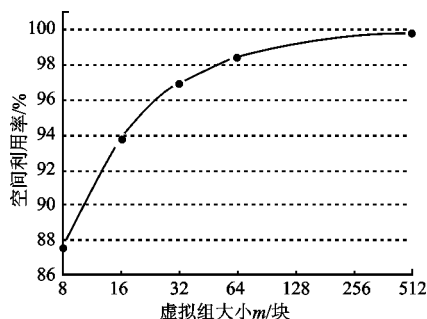


图 3 虚拟组规模与空间利用率的关系

根据图 3 所示的趋势,当 $m \geq 16$ 时,后续随着 m 的增大,空间利用率提高不显著,而映射表的大小却是随着 m 的增大线性增加的.因此,工程实践中选择 $m=16$ 较为合适.此时,映射表的大小仅为 4 kB,可管理 1024 个页的全相联映射.

5 结束语

本文设计了一种单替换块组相联映射方法,对 NandFlash 存储空间进行有效管理.通过参数调节,在 NandFlash 存储空间利用率和内存开销间进行均衡,分析得出了工程实践可行的参数方案.目前,该方法已在某机载大容量存储模块的 NandFlash 管理软件中得到应用.

当前设计的每个虚拟组都有一个替换块,虽然较其他文献提出的一比一替换方法显著提升了空间利用率,但在虚拟组内部块数较少时仍然有一定的

空间浪费,仍存在改进空间.

参考文献:

- [1] Chang Lipin, Huang Luchun. A low-cost wear-leveling algorithm for block-mapping solid-state disks[C]// Proc of Conference on Languages, Compilers and Tools for Embedded Systems. New York: ACM Press, 2011: 31-40.
- [2] 郑文静,李明强,舒继武. Flash 存储技术[J]. 计算机研究与发展, 2010, 47(4): 716-726.
- [3] 邢春波,杨良怀,龚卫华,等. 一种有效的混合式闪存磨损均衡算法[J]. 小型微型计算机系统, 2009, 30(9): 1903-1906.
- [4] 王浩,邵高平,胡泽明. 基于分组—循环擦写的闪存磨损均衡算法[J]. 计算机应用研究, 2013, 30(5): 1353-1356.
- [5] Intel Corporation. Understanding the flash translation layer (FTL) specification[R]. California: Intel Press, 1998.
- [6] Kim J, Kim J M, Noh S H, et al. A space-efficient flash translation layer for compact-flash systems[J]. IEEE Trans. on Consumer Electronics, 2002, 48(2): 366-375.
- [7] Lee S W, Park D J, Chung T S, et al. A log buffer-based flash translation layer using fully-associative sector translation [J]. ACM Trans. on Embedded Computing Systems, 2007, 6(3): 18.

作者简介:

胡 宁 男, (1978-), 博士, 高级工程师. 研究方向为嵌入式操作系统和软件工程. E-mail: hnhawk@163.com.

杨 琼 女, (1987-), 硕士, 工程师. 研究方向为嵌入式操作系统.

王 冬 男, (1987-), 硕士, 工程师. 研究方向为嵌入式操作系统.

(上接第 18 页)

- [4] 何先波,唐宁九,吕方,等. ELF 文件格式及应用[J]. 计算机应用研究, 2001, 18(11): 144-145.
- [5] 齐宁,丁松阳,孙维新,等. 基于加载机制分析的 ELF 文件 main 函数定位技术[J]. 计算机工程, 2007, 33(6): 89-91.
- [6] 王红梅,张铁军,王东辉. 面向 VLIW 处理器的 GAS 汇编器实现[J]. 微计算机应用, 2010, 31(5): 73-76.

作者简介:

鲍丽丹 女, (1987-), 博士研究生. 研究方向为编译器与编译优化技术. E-mail: badidan 1987@163.com.

张铁军 男, (1974-), 博士, 研究员, 硕士生导师. 研究方向为 VLSI 设计、DSP 设计.

王东辉 男, (1973-), 博士, 研究员, 博士生导师. 研究方向为 VLSI 设计、DSP 设计.