

基于多阈值的闪存磨损均衡算法

拓 晶, 唐 磊

(西安微电子技术研究所, 陕西 西安 710054)

摘 要: NAND Flash 专用文件系统 Yaffs2 在磨损均衡算法方面仍存在缺陷, 通过详细分析 Yaffs2 中磨损均衡算法的实现原理, 综合考虑磨损均衡与垃圾回收的矛盾关系, 提出了一种基于多阈值的磨损均衡机制(属于静态磨损均衡算法), 利用三个关键技术: (1) 改进回收块选择机制, 选择公式(1)的计算结果落入目标区间中的所有块中擦除次数最大的块进行回收; (2) 利用有序双向链表管理空闲块, 分配当前链表中擦除次数最小的块给新写入的数据, 而将擦除擦除次数最大的块分配给静态数据; (3) 针对磨损均衡的两级分化引入冷热数据分离策略. 实验结果表明, 在仅付出 4.05% 的额外擦除代价下, 将 NAND Flash 的寿命延长了近 4 倍.

关键词: Yaffs2 文件系统; 磨损均衡; 回收块选择机制; 冷热数据分离

中图分类号: TP302

文献标识码: A

文章编号: 1000—7180(2018)01—0006—04

MAND Flash Wear Leveling Based on Mutil-Threshold

TUO Jing, TANG Lei

(Xi'an Microelectronics Technology Institute, Xi'an 710054, China)

Abstract: Yaffs2, special to NAND Flash, still exists defect aiming at wear leveling. A static wear-leveling algorithm based on Yaffs2 is designed in this paper, considering the contradiction between wear-leveling and garbage collection. Firstly, Improve retrieve block selection mechanism; choose the block for recycle whose the value of benefit/cost falling to the target range according to the formula (1), and its erasure number is biggest; Secondly, allocation free page whose erasure number is smallest to new data and whose erasure number is biggest to cold data in line of orderly two-way linked list; Last, introduce hot and cold data separation strategy orienting to hot and cold data coexists in a block. The experimental results show that the algorithm prolongs NAND lifetime nearly four times at the expense of 4.05% additional erasure number.

Key words: yaffs2 file system; wear leveling; retrieve block selection; separation of cold and hot data

1 引言

NAND Flash 每个物理块的擦除次数存在上限, 对于 SLC (Single-Level Cell) 型 NAND 其块擦除上限为 100 k, 而 MLC (Multi-Level Cell) 型 NAND 仅为 10 k^[1-2], 一旦某块磨穿 (wear out), 则不能继续用于数据存储, 还会引起相邻块迅速坏掉. 为了保障数据的可靠性, 当 NAND 中出现磨穿块时就会停止继续使用^[2-3]. 磨损均衡旨在使 NAND 中各个块的擦除次数均匀分布, 在几乎不增加额外擦除的代价下, 最大限度的延长闪存的使用寿命^[4]. 磨

损均衡技术大致分为两类-静态磨损均衡和动态磨损均衡. 动态磨损均衡^[3-6]主要针对更新频繁的热数据块, 不考虑冷数据块, 显然很难在整个闪存颗粒中实现磨损均衡, 优点是算法简单, 占用主控的资源少. 静态均衡^[7-9], 从全局的角度出发, 定期将冷数据搬移到擦除次数较大的块上, 利用冷数据^[10-11]很少更新的特点减小热块继续磨损, 但是搬移静态数据时会引起写放大, 尤其是块上数据均为有效数据时, 强行回收冷块需要进行整块数据的搬移, 有时为了得到系统所需的空闲空间必须进行多次回收. 显然, 这样的回收块择机制会占用主控的大量资源, 实

现复杂度也相对较高. 通过深入细致分析上述磨损均衡算法,归纳出以下三个影响磨损均衡效果的因素:(1)回收块选择机制,磨损均衡要求擦除当前时刻具有最小擦除次数的块,而垃圾回收则偏向于选择有效数据最少的块进行回收,显然二者很难统一^[6]; (2)空间分配策略,如何处理冷数据的搬移以及新写入的数据如何存放,直接由块分配策略决定; (3)冷热数据分离机制,好的算法应能够清晰地区分冷热数据,并且能够解决冷数据突然变热以及热数据变冷的块反转现象^[10-11]. 总之,嵌入式系统中仍然没有很好的实现闪存磨损均衡,事实上好的均衡算法仅需保证在闪存颗粒寿命即将耗尽时,所有块的擦除次数趋于平均,没有必要要求任意时刻闪存的磨损都呈现均匀分布.

2 Yaffs2 FS 层次结构

与通用 Flash 文件系统 JFFS/JFFS2 不同, Yaffs2^[7] 文件系统专为 NAND Flash 设计,带有 NAND 驱动程序,同时通过 Yaffs direct 向上层应用提供 API,用户可以不使用 Linux 系统而直接操作 Yaffs2. NAND Flash 大多采用 MTD+Yaffs2 的模式. MTD(Memory Technology Device,内存技术设备)^[1] 提供操作 Flash 的统一接口,从而将硬件驱动设计与文件系统设计分离,最终,在 NAND 器件上以怎样的格式组织和存取数据由 Yaff2 文件系统决定. 图 1 列出了嵌入式系统的文件系统结构. NAND 器件的数据组织存储单位是 page(Yaffs2 使用 chunk,实质相同),一页又根据状态不同,可以分为干净页、脏页、空闲页. 干净页中存放有效数据,脏页中存放无效数据,空闲页是经过擦除后可以直接用于写入数据的页. NAND Flash 需遵循 Erase-Before-Write,当 block 中某个 page 更新时,须先将其他干净页中数据拷贝到其他空闲页上,并将原数据页标记为无效,对应的块随之变为脏块. 随着上述操作的不断进行,大量无效页累积在存储器中使得空闲空间告急. 当 Flash 中的空闲块达到预制阈值时,启动垃圾回收操作,擦除脏块. NAND Flash 中块的擦除次数存在上限,通常为 10K 次~100K 次^[1-2]. 当块擦除次数超过该值时,继续使用则无法保证其上数据的可靠性. 因此,设计合理的回收块选择机制,从待回收块中找到回收后能产生良好磨损均衡效果且付出较少额外代价的块来擦除,从而最大限度的发挥闪存的性能.

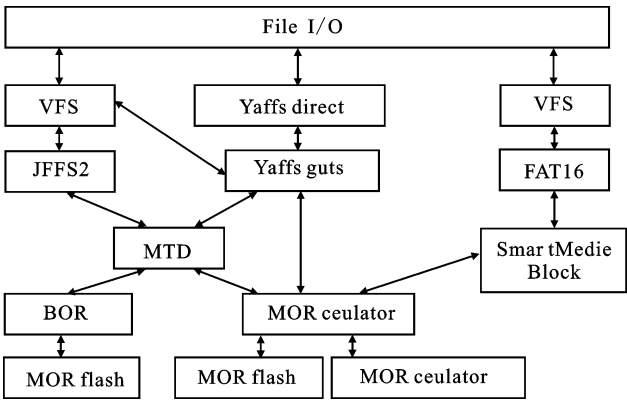


图 1 嵌入式系统中文件系统结构

3 Yaffs2 在磨损均衡方面的缺陷

Yaffs2 中采用贪心算法选择回收块,擦除有效数据最少的块,这样做可以最大限度减少有效数据搬移的代价,同时具有最小的时间开销. 当数据更新趋于一致的时,Yaffs2 的回收块选择机制具有最佳的执行效率. 但是,通常情况下数据页的更新频率并不相同,部分数据页更新频繁,而大量数据几乎不更新^[8]. 频繁更新的数据称为热数据,很少更新的数据称为冷数据^[10-11]. 如果某个块上的数据不断更新,该块上有效数据就会逐渐减少,根据最少有效数据选择策略,Yaffs2 容易选中热数据块来回收. 而存储大量有效冷数据的块却难以被 Yaffs2 选中擦除. 回收的热块再次分配存放待写入的新数据,而新数据一般又是热数据,进而使得热数据所在块被反复擦除. 由此可知,热数据所在的块擦除次数较大,冷数据所在的块擦除次数较少,物理块的擦除次数趋于两级分化,即最大块擦除次数与最小块擦除次数差值不断增大.

Yaffs2 回收脏块时,需将该块上的剩余有效数据拷贝到其他空闲块上. 而剩余有效数据不能占满整个块,系统会将新写入的数据继续存放到该块的剩余空闲页里,直到该块写满为止. 被保留下来的数据通常是热数据,这样会造成冷热数据共存于一个物理块. 该块上的热数据不仅更新频繁而且所占比例较大,容易变为无效数据进而使对应块满足回收条件,系统会再次将保留下来冷数据和待写入热数据存放到下一个空闲块中. 随着上述更新方式进行,会引起局部热数据块的反复更新导致冷数据被迫更新,增加了非必须的数据搬移开销,同时使得数据块擦除次数标准差逐渐增大,即磨损不均衡.

4 Yaffs2 的改进方法

4.1 基于阈值的回收块选择策略

针对 Yaffs2 回收块选择机制的缺陷,可以通过加入块的年龄 age 参数^[4]解决. 增加 age 参数后的回收块选择策略遵循式(1),这可以综合考虑数据更新速率、擦除时的写放大效应对磨损均衡的影响. 再者,为了尽可能的延长闪存的使用寿命,应该延迟各个回收块的擦除时间^[2].

$$\frac{\text{Benefit}}{\text{Cost}} = \text{age} \times \frac{1-u}{2u} \quad (1)$$

式中, age 指从本次分配开始到当前擦除时刻闪存块的使用时间,擦除后将 age 清零. u 表示单个块中有效数据的占有比率,由于有效数据的搬移会引起不必要的时间开销. 新的回收块选择机制选取当前时刻年龄最大且有效数据最少(即 $\frac{\text{benefit}}{\text{cost}}$ 比值最大)

的块作为回收对象,但是一般情况下, age 最大的块其上的有效数据的比例也会较大,因为 age 最大的块上存储的数据可认为是 cool data,而 cool data 一般又是有效数据,所以上式的分母也会较大,导致比值较小^[2]. 基于此,我们设定一个阈值 Th_1 ,将上式的计算结果落入区间 $\left[\frac{\text{benefit}}{\text{cost}}_{\max} - \text{Th}_1 \sim \frac{\text{benefit}}{\text{cost}}_{\max}\right]$ 所对应的块归入待回收块池,再根据擦除次数由大到小的顺序依次擦除回收块,直到满足系统的空闲区间需求.

4.2 基于双向列表的空闲块管理机制

将回收后的干净块插入 clean_list 链表,本文使用有序双向链表管理空闲块,空闲块按照擦除次数由小到大排列,采用二分查找算搜索当前回收块的插入位置. 显然,每次写入新数据时选择表头元素标记的物理块进行分配,而进行冷数据搬移时则选择表尾元素标记的物理块,这样可以将热数据写入擦除次数较小的块,同时将冷数据搬移到磨损最大的块,而且插入回收块的算法时间复杂度为 $O(\log_2 N)$ (N 表示列表的长度)^[11].

4.3 冷热数据分离策略

对于 Yaffs2 中冷热数据混存的问题,可以引入冷热数据分离策略^[5]解决. 冷热数据分离策略不仅能够收集冷数据,有效地避免热数据更新时冷数据的被迫更新现象,而且可以有效解决冷数据突然变热或者热数据突然变冷的块反转情况. 在系统运行过程中,一旦检测到块的最大、最小擦除次数差值达到阈值 Th_2 ,将擦除次数最小的块上的有效冷数据

搬移到擦除次数多的块上. 当冷池中的块被热数据占据或者热池中块被冷数据占据时,冷热数据分离策略也能处理这类块反转问题. 数据的冷热程度对块的擦除次数会产生直接影响,因此冷热数据分离策略通过块擦除次数来判断数据温度. 操作步骤如下:

(1)初始化:剔除初始坏块后,将 NAND Flash 中剩余块按 3:1 分为两个部分,前者放到冷池中,后者放入热池中.

(2)数据交换:找出热池中擦除次数最多的块 A(用 max_wear 表示擦除次数),冷池中擦除次数最少的块 B(用 min_wear 表示擦除次数). 当 $\text{max_wear} - \text{min_wear} > \text{Th}_2$ 时,立即进行数据搬移. 首先将块 A 上的有效数据转存到空闲块链表 clean-list 的表头块,擦除该块后将块 B 中收集的冷数据拷贝到块 A 中,最后擦除块 B 并根据擦除次数将块 B 插入到 clean-list 中. 利用冷数据不易更新的特点降低热块的磨损程度,同时选择 clean-list 中擦除次数最小的块存储新写入的数据,从而达到磨损均衡的目的.

$$\text{Th}_2 = 0.1 \times (10k - \text{max_wear}) \quad (2)$$

块反转:当热池中的某个块突然被冷冻时,由于冷数据很少更新导致该块被滞留在热池中且逐渐成为热池中最冷的块. 针对上述情况,引入第三个阈值 Th_3 . 当热池中擦除次数最多的块 $H_{\max}$ (用 max_wear_1 表示其擦除次数)和热池中擦除最少的块 $H_{\min}$ (用 min_wear_1 表示其擦除次数)擦除次数满足 $\text{max_wear}_1 - \text{min_wear}_1 > \text{Th}_3$ 时,进行热块反转操作,将块 $H_{\min}$ 放到冷池中去. 另外当冷池中的某个块突然变热逐渐成为冷池中最热的块时,同样如果冷池中擦除次数最大的块 $C_{\max}$ (用 max_wear_2 表示其擦除次数)与擦除次数最小的块 $C_{\min}$ (用 min_wear_2 表示其擦除次数)间的擦除次数差值满足 $\text{max_wear}_2 - \text{min_wear}_2 > \text{Th}_4$ 时,将块 $C_{\max}$ 归入热池.

5 结束语

从三方面对 Yaffs2 文件系统作改进:一是对回收块选择机制作改进,选择块擦除次数落入区间 $\left[\frac{\text{benefit}}{\text{cost}}_{\max} - \text{Th}_1 \sim \frac{\text{benefit}}{\text{cost}}_{\max}\right]$ 中擦除次数最大的块进行回收;二是利用有序双向链表管理空闲块,分配当前链表中擦除次数最小的块给新写入的数据,而将擦除擦除次数最大的块分配给静态数据;

三是针对磨损均衡的两级分化引入冷热数据分离策略. 将上述程序做成 module,更改 Yaffs2 对应的磨损均衡算法,并将上述文件系统 Yaffs2-Updata 移植到 mini2440 开发板中验证. 实验结果表明;仅付出 4.05%的额外擦除代价,将 NAND Flash 的寿命由 2.19 年延长到 8.23 年.

参考文献:

[1] Agrawal N, Prabhakaran V, Wobber T, et al. Design tradeoffs for SSD performance[C]// Usenix Technical Conference. Boston, Ma, Usa, DBLP, 2008:57-70.

[2] 温朝建, 严华. Yaffs2 文件系统中对 NAND Flash 磨损均衡的改进[J]. 电子技术应用, 2015, 41(6): 16-19.

[3] 杨习伟, 贺云健. 车载娱乐系统的 NandFlash 管理模式研究[J]. 单片机与嵌入式系统应用, 2008, 7(1): 19-22.

[4] Gal E, Toledo S. Algorithms and data structures for flash memories[J]. Acm Computing Surveys, 2005, 37(2):138-163.

[5] Wu M, Zwaenepoel W. eNVy: a non-volatile, main memory storage system[J]. Acm Sigops Operating Systems Review, 1994, 28(5):86-97.

[6] Lin M, Chen S. Efficient and intelligent garbage collection policy for NAND flash-based consumer electronics[J]. IEEE Transactions on Consumer Electronics, 2013, 59(3):538-543.

[7] Xu G, Wang M, Liu Y. Swap-aware garbage collection algorithm for NAND flash-based consumer electronics[J]. Consumer Electronics IEEE Transactions on, 2014, 60(1):60-65.

[8] Murugan M, Du D H C. Rejuvenator: A static wear leveling algorithm for NAND flash memory with minimized overhead[C]// IEEE, Symposium on MASS Storage Systems and Technologies. IEEE Computer Society, Washington, DC, USA, 2011:1-12.

[9] Chang Y H, Chang L P. Efficient wear leveling in NAND flash memory[M]. Inside Solid State Drives (SSDs), Springer Netherlands, 2013:233-257.

[10] Hsieh J W. Efficient on-line identification of hot data for flash-memory management[C]// Acm Symposium on Applied Computing. Proceedings of the ACM Symposium on Applied Computing. New York, NY, USA 2005:838-842.

[11] 刘柳, 黄德才. 一种逻辑块温度和物理块年龄的磨损均衡算法[J]. 计算机系统应用, 2011, 20(12): 109-113.

作者简介:

拓晶女,(1989-),硕士研究生. 研究方向为嵌入式存储系统. E-mail:13609117403@163.com.

唐磊男,(1973-),硕士,博士生导师. 研究方向为嵌入式计算机、微系统.

(上接第 5 页)

[11] Rajendran J, Zhang H, Zhang C, et al. Fault analysis-based logic encryption[J]. Computers IEEE Transactions on, 2015, 64(2):410-424.

[12] Subramanyan P, Ray S, Malik S. Evaluating the security of logic encryption algorithms[C]// IEEE International Symposium on Hardware Oriented Security and Trust. Athens, Greece, IEEE, 2015:137-143.

作者简介:

黄一鸣男,(1993-),硕士研究生. 研究方向为硬件木马及可信性设计. E-mail:ymhuang@pku.edu.cn.

冯建华男,(1964-),博士,副教授. 研究方向为数字、模拟和混合信号电路测试和可测试性设计.

张小飞男,(1991-),硕士研究生. 研究方向为硬件木马.