Elastic Queue: A Universal SSD Lifetime Extension Plug-in for Cache Replacement Algorithms

**1. Introduction**

【基本要重写，篇幅会较长】

P1. SSD Cache的重要性，已经在使用。历史上Cache算法有很多，很难说哪个算法一定最好，不同应用情景下，表现优异的算法很可能不同。给出一个我们的实验结果加以证实。如果经过测试知道某种算法在某个环境下表现特别出色，就可以设为这种策略来提高性能。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | LRU | LFU | LIRS | ARC |
| 云数据库 | 56.6% | **59.5%** | 58.4% | 58.4% |
| 随机文件访问 | 1.9% | 4.6% | 18.7% | **20.3%** |
| 网络文件系统 | 13.4% | **18.8%** | 15.4% | 15.8% |
| VOD | 9.24% | 4.0% | **17.3%** | 14.4% |

注：5%的Cache空间

P2. 但是SSD Cache上面临耐久性问题，传统的Cache无法解决，因此有一些改进的Cache算法提出[引用SieveStore, LARC, L2ARC, WEC, ETD-Cache等]，但是已有算法选择缓存对象的方式确定(see Section 2.X和表XX)，不能充分利用各种Cache算法的优势，在各种应用环境中选择最合适的算法。所以需要一种新的plug-in，能够与各种Cache策略配合，在其基础上显著减少SSD写入，延长SSD寿命。

(解释为什么plug-in就可以解决寿命问题)

P3. 缓存算法都是用某种规则形成一个数据的排序（按照evict的优先级从低到高），如下图所示，其中实线框表示SSD缓存空间，虚线框表示假设将未缓存的其他所有数据也按照缓存算法进行排序。//那么理想情况是缓存数据不断命中，始终停留在实线框中。然而在实际应用中，由于缓存空间有限和好数据的热度不稳定这两个原因，好数据不加保护的话会不断通过边界。如图a所示，造成很多写入量；通过将他们pin住，和动态调整边界的位置，可以减少数据在边界来往的次数，减少额外的写入。

增加在边界进出的问题，好数据不加保护的话会不断通过边界，造成很多写入量；通过将他们pin住和动态调整边界位置，可以减少很多额外的淘汰和进入。要画一个对比图，本来进出的很多，现在很少



P4.我们的一个统计证实了这一点，以LRU为例，被缓存算法evict的数据中，有相当一部分是热门数据，只是暂时冷，就被踢出了。

//将原来3.2的内容和图3拿来，不谈no-hit rate了。（原表1省略，只在正文中用一句话提一下各个Trace代表什么应用）

P5. 所以，已经缓存算法虽然多样，但是在SSD写入量过多的这个问题上，他们的问题是一样的，都是对这样的好数据没有保护好，让其evict，后面又会由于表现优异再次进入SSD Cache，循环往复，产生大量没必要的写入量。因此我们提出算法的思路就是Elastic Queue，在其他算法排序队列的基础上，判别好坏数据，将好数据Pin住，即使进入虚线区域也不evict。实验结果blablabla。

P6. 下文组织

**2. Related Work**

2.1 Write Endurance of SSDs

不用变

2.2 Cache Replacement Algorithms

P1 不变

P2 加入我们自己的工作引用，WEC和ETD-Cache

增加P3，说明以上经典算法，以及面向SSD的改进方法都是用固定的策略选择SSD缓存对象，不能很好适应不同应用特征。经典算法（LRU等）是倾向于recency，或frequency，或二者的混合，这里就不总结了。主要总结一下面向SSD的改进算法选择缓存对象的固定策略：

|  |  |
| --- | --- |
| 算法 | 选择缓存对象的固定策略 |
| SieveStore | Frequency，访问数达到阈值作为准入条件 |
| LARC | 两个LRU队列（再确认一下） |
| L2ARC | 内存队列ARC算法，SSD队列FIFO |
| WEC | LRU队列 |
| ETD-Cache | 两个LRU队列，加采样 |

**3. Elastic Queue**

【基本要重写，有些素材可以使用，或改动后使用】

**3.1 Principle of Elastic Queue**

尽管cache算法有各种各样的形式，但从整体的角度看，都可以看做是将数据按照evict优先级排列在队列中，然后淘汰高优先级数据，补充新数据的过程。所以不论算法的实现方式如何，都可以用一个缓存队列来描述，而不同算法之间的核心区别在于如何给数据划分优先级。

从之前的分析中可以看到，在如此多的应用场景中，针对某一个或某一类应用，一般都可以找到一种缓存算法，可以比其他算法更准确地鉴别出其中的热门数据。但是由于缓存空间有限，一旦这些数据离开了缓存队列，就会触发缓存替换，增加SSD写入。虽然这些数据离开了缓存空间，但我们仍然可以将它们看作延伸队列的一部分。如果一个数据被驱逐出了缓存空间，但在延伸队列中多次命中，不久后它会再次进入缓存空间。如此一来，这一次替换就带来了额外的写入，这在内存替换中并不是太严重的问题，但是在SSD这样的不耐久设备上就会白白损耗宝贵的寿命。

基于这个考虑，EQ的思想是将好数据pin住，保护起来，不让其由于短暂的冷却而被evict。但核心问题是如何判定好数据？主要挑战在于数据的访问间隔不稳定，即使是热门数据[引用WEC结论]。

我们提出一个低负载的简单方法，统计每个数据块在下图虚拟队列中各个段的命中数，这样可以求出平均的access interval，平均interval小，说明热度高，就是好数据。应该长期缓存，不要暂时靠后就退出，即不要因为噪音干扰“这是个好块”的判断。

// 给出一好、一坏两个例子，可以清楚看出这种方法思路。



下面可以再给一组图，LRU、MQ、LIRS、ARC、LARC的队列管理示意图，用来说明上图Cache队列是一个抽象的概念，不一定是一个单队列，形式可能每个算法都不同，但其中各个块按照被淘汰的优先级会形成一个优先级队列，我们的Cache队列指的是这个虚拟的优先级队列。

**3.2 Architecture of Elastic Queue**

但上述方法维护要维护所有数据块的元数据，负载太大。因此我们做了简化，即将原有Cache队列（实线框）扩展为Elastic Queue。

在系统中有两类数据：General block, Pinned block(always on SSD).

PB是EQ由Pinning Block module负责将Cache size内的数据块pin住（如果考察其为好数据）；Pin的状态不会一直保持，由Unpinning Block module负责判断pinned block何时不好了，应该退出，变为unpin状态。

General blocks：除了Pinned block之外，其他的block也可能被原始的算法认为是值得缓存的数据，但由于EQ的存在，SSD空间被PB占据了，所以只记录GB的元数据。当有Pinned block退出时，这些GB将会被Pinning module考察和挑选成为Pinned block，真正进入SSD并停留一段时间。而如果未能成为Pinned block，而且超过了cache size，就不再被Pinning module考虑，而是直接evict，此时的evict本质上只是删除其在SSD queue中的元数据。

Elastic Queue就是在原有的优先级队列上延长了队列长度，容纳原本只能直接被Evict而现在被Pin住的数据。EQ的长度是弹性变化的，两种极端情况：当原有的优先级队列都是Pinned block时，EQ的长度等于原有优先级队列长度；当所有Pinned block都在延长队列中时，原有队列只有GB的元数据，此时队列长度为2倍。

// 介绍EQ和上述完全队列的对应关系：每次有数据退出，会加1，这样能看出来每个pinned block在上述虚拟队列中的位置。

只有cache size内才可能有general block，cache size外只可能有pinned block。

我们会统计每个数据块（包括general block和pinned block）的命中数，包括在cache size内的，和cache size外的，用于判定其热度。

// 介绍General Block和Pinned Block的处理规则区别

Pinned block真正存在于SSD中，general block只是元数据。

Pinned Block：由Pinning Block module负责将Cache size内的数据块pin住（如果考察其为好数据）；由Unpinning Block module负责判断pinned block何时不好了，应该退出，变为unpin状态。

General Block：过了cache size的线就evict（实际只有元数据）。



给出一个具体的例子，将原Fig. 5的图改造一下，符合EQ就行。

3.2部分要强调一下EQ名字的来历，我们把原来的优先级队列的长度可能延伸出去，但是长度是弹性变化的，可长可短，最长是2倍的元队列长度，最短是1倍。

**3.3 Pinning Block**

3.3.1 触发Pinning block

记录当前pinned block数量，由此知道SSD Cache是否还有空间。如果有空间就会触发pinned block，从EQ首部开始找第一个general block pin住。

// 原4.5 Location Transition内容改一改。

3.3.2 Black List

简要介绍，unpinned block要记录一下，这里的数据不能pin，会skip。

Black list是一个长度有限的FIFO队列，时间长了有可能再次进入。

// 还有原4.3 EB Entrance内容考虑拿过来一些。

**3.4 UnPinning Block**

3.4.1 数据分类

根据在不同段的命中数，将数据分为三类，赋予不同的最远距离（退出条件）

**4. Evaluation**

先保留原内容，具体措辞上修改。

**5. Conclusion**

最后修改。