


# 闪存类存储的转换层



---

上篇「柱面-磁头-扇区寻址的一些旧事」整理了一下我对磁盘类存储设备（包括软盘、硬盘，不包括光盘、磁带）的一些理解，算是为以后讨论文件系统作铺垫；这篇整理一下我对闪存类存储设备的理解。

这里想要讨论的闪存类存储是指 SSD、SD卡、U盘等基于 NAND 又有转换层的存储设备（下文简称闪存盘），但不包括裸 NAND 设备、3D Xpoint（Intel Optane）等相近物理结构但是没有类似转换层的存储设备。闪存类存储设备这几年发展迅猛，SD卡和U盘早就

替代软盘成为数据交换的主流，SSD 大有替代硬盘的趋势。因为发展迅速，所以其底层技术变革很快，不同于磁盘类存储技术有很多公开资料可以获取，闪存类存储的技术细节通常是厂商们的秘密，互联网上能找到很多外围资料，但是关于其如何运作的细节却很少提到。所以我想先整理一篇笔记，记下我搜集到的资料，加上我自己的理解。本文大部分信息来源是 [Optimizing Linux with cheap flash drives](#) 和 [A Summary on SSD & FTL](#)，加上我的理解，文中一些配图也来自这两篇文章。

## 封装结构

从外部来看，一个闪存盘可能有这样的结构：

从上往下，我们买到的一个闪存盘可能一层层分级：

1. 整个闪存盘有个控制器，其中含有一部分 RAM 。  
然后是一组 NAND Flash 封装芯片（chip）。
2. 每个封装芯片可能还分多个 Device ，每个 Device 分多个 Die ，这中间有很多术语我无法跟上，大概和本文想讨论的事情关系不大。
3. 每个 Die 分多个平面（Plane），平面之间可以并行控制，每个平面相互独立。从而比如在一个平面内做某个块的擦除操作的时候，别的平面可以继续读写而不受影响。
4. 每个平面分成多个段（Segment），段是擦除操作的基本单位，一次擦除一整个段。
5. 每个段分成多个页面（Page），页面是读写操作的基本单位，一次可以读写一整页。

6. 页面内存有多个单元格（Cell），单元格是存储二进制位的基本单元，对应 SLC/MLC/TLC/QLC 这些，每个单元格可以存储多个二进制位。

以上这些名字可能不同厂商不同文档的称法都各有不同，随着容量不断增大，厂商们又新造出很多抽象层次，不过这些可能和本文关系不大，如果看别的文档注意区别术语，本文中我想统一成以上术语。重要的是有并行访问单元的平面（Plane）、擦除单元的段（Segment）、读写单元的页（Page）这些概念。抽象地列举概念可能没有实感，顺便说一下这些概念的数量级：

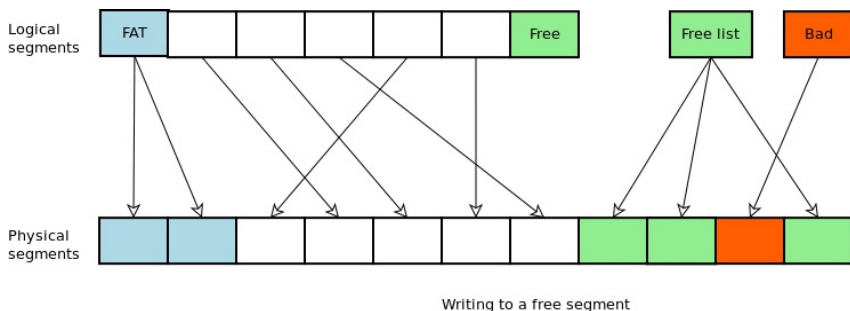
1. 每个 SSD 可以有数个封装芯片。
2. 每个芯片有多个 Die。
3. 每个 Die 有多个平面。
4. 每个平面有几千个段。比如 2048 个。
5. 每个段有数百个页，比如 128 个，外加一些元数据。
6. 每个页面是 4KiB、8KiB 这样的容量，外加几百字节的元数据。

和硬盘相比，一个闪存页面大概对应一个到数个物理扇区大小，现代硬盘也逐渐普及 4KiB 物理扇区。每次读写都可以通过地址映射直接对应到某个闪存页面，这方面没有硬盘那样的寻址开销。不过闪存有写入的限制，每次写入只能写在「空」的页面上，不能覆盖写入已有数据的页面。要重复利用已经写过的页面，需要对页面所在段整个做擦除操作，每个段是大概 128KiB 到

8MiB 这样的数量级。每个擦除段需要单独跟踪和统计自己经历的擦除次数，以进行擦写均衡（wear leveling）。

# 擦写均衡（wear leveling）和段映射

Animation: wear leveling on SSD drives



擦除段的容量大小是个折衷，更小的擦除段比如 128KiB 更适合随机读写，因为每随机修改一部分数据时需要垃圾回收的粒度更小；而使用更大的擦除段可以减少元数据和地址映射的开销。从擦除段的大小这里，已经开始有高端闪存和低端闪存的差异，比如商用 SSD 可能比 U 盘和 SD 卡使用更小的擦除段大小。

闪存盘中维护一个逻辑段地址到物理段地址的隐射