

閃存類存儲的轉換層



上篇「柱面-磁頭-扇區尋址的一些舊事」整理了一下我對磁盤類存儲設備（包括軟盤、硬盤，不包括光盤、磁帶）的一些理解，算是爲以後討論文件系統作鋪墊；這篇整理一下我對閃存類存儲設備的理解。

這裏想要討論的閃存類存儲是指 SSD、SD卡、U盤等基於 NAND 又有轉換層的存儲設備（下文簡稱閃存盤），但不包括裸 NAND 設備、3D Xpoint（Intel Optane）等相近物理結構但是沒有類似轉換層的存儲設備。閃存類存儲設備這幾年發展迅猛，SD卡和U盤早就

替代軟盤成爲數據交換的主流，SSD 大有替代硬盤的趨勢。因爲發展迅速，所以其底層技術變革很快，不同於磁盤類存儲技術有很多公開資料可以獲取，閃存類存儲的技術細節通常是廠商們的祕密，互聯網上能找到很多外圍資料，但是關於其如何運作的細節卻很少提到。所以我想先整理一篇筆記，記下我蒐集到的資料，加上我自己的理解。本文大部分信息來源是 [Optimizing Linux with cheap flash drives](#) 和 [A Summary on SSD & FTL](#)，加上我的理解，文中一些配圖也來自這兩篇文章。

封裝結構

從外部來看，一個閃存盤可能有這樣的結構：

從上往下，我們買到的一個閃存盤可能一層層分級：

1. 整個閃存盤有個控制器，其中含有一部分 RAM 。然後是一組 NAND Flash 封裝芯片（chip）。
2. 每個封裝芯片可能還分多個 Device ，每個 Device 分多個 Die ，這中間有很多術語我無法跟上，大概和本文想討論的事情關係不大。
3. 每個 Die 分多個平面（Plane），平面之間可以並行控制，每個平面相互獨立。從而比如了一個平面內做某個塊的擦除操作的時候，別的平面可以繼續讀寫而不受影響。
4. 每個平面分成多個段（Segment），段是擦除操作的基本單位，一次擦除一整個段。
5. 每個段分成多個頁面（Page），頁面是讀寫操作的基本單位，一次可以讀寫一整頁。

6. 頁面內存有多個單元格（Cell），單元格是存儲二進制位的基本單元，對應 SLC/MLC/TLC/QLC 這些，每個單元格可以存儲多個二進制位。

以上這些名字可能不同廠商不同文檔的稱法都各有不同，隨着容量不斷增大，廠商們又新造出很多抽象層次，不過這些可能和本文關係不大，如果看別的文檔注意區別術語，本文中我想統一成以上術語。重要的是有並行訪問單元的平面（Plane）、擦除單元的段（Segment）、讀寫單元的頁（Page）這些概念。抽象地列舉概念可能沒有實感，順便說一下這些概念的數量級：

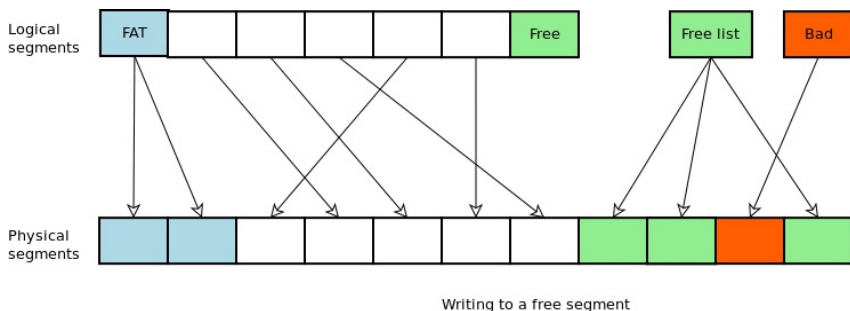
1. 每個 SSD 可以有數個封裝芯片。
2. 每個芯片有多個 Die。
3. 每個 Die 有多個平面。
4. 每個平面有幾千個段。比如 2048 個。
5. 每個段有數百個頁，比如 128 個，外加一些元數據。
6. 每個頁面是 4KiB、8KiB 這樣的容量，外加幾百字節的元數據。

和硬盤相比，一個閃存頁面大概對應一個到數個物理扇區大小，現代硬盤也逐漸普及 4KiB 物理扇區。每次讀寫都可以通過地址映射直接對應到某個閃存頁面，這方面沒有硬盤那樣的尋址開銷。不過閃存有寫入的限制，每次寫入只能寫在「空」的頁面上，不能覆蓋寫入已有數據的頁面。要重複利用已經寫過的頁面，需要對頁面所在段整個做擦除操作，每個段是大概 128KiB 到

8MiB 這樣的數量級。每個擦除段需要單獨跟蹤和統計自己經歷的擦除次數，以進行擦寫均衡（wear leveling）。

擦寫均衡（wear leveling）和段映射

Animation: wear leveling on SSD drives



擦除段的容量大小是個折衷，更小的擦除段比如 128KiB 更適合隨機讀寫，因為每隨機修改一部分數據時需要垃圾回收的粒度更小；而使用更大的擦除段可以減少元數據和地址映射的開銷。從擦除段的大小這裏，已經開始有高端閃存和低端閃存的差異，比如商用 SSD 可能比 U 盤和 SD 卡使用更小的擦除段大小。

閃存盤中維護一個邏輯段地址到物理段地址的隱射