

內部碎片與小文件優化



上篇「系統中的大多數文件有多大？」提到，文件系統中大部分文件其實都很小，中位數一直穩定在 4K 左右，而且這個數字並沒有隨着存儲設備容量的增加而增大。但是存儲設備的總體容量實際是在逐年增長的，，總容量增加而文件大小中位數不變的原因，可能是以下兩種情況：

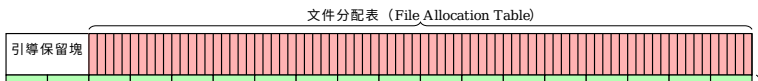
1. 文件數量在增加
2. 大文件的大小在增加

實際上可能是這兩者綜合的結果。這種趨勢給文件系統設計帶來了越來越多的挑戰，因為我們不能單純根據平均文件大小來增加塊大小（block size）優化文件讀寫。微軟的文件系統（FAT系和 NTFS）使用「簇（cluster）」這個概念管理文件系統的可用空間分配，在 Unix 系文件系統中有類似的塊（block）的概念，只不過稱呼不一樣。現代文件系統都有這個塊大小或者簇大小的概念，從而基本的文件空間分配可以獨立於硬件設備本身的扇區大小。塊大小越大，單次分配空間越大，文件系統所需維護的元數據越小，複雜度越低，實現起來也越容易。而塊大小越小，越能節約可用空間，避免內部碎片造成的浪費，但是跟蹤空間所需的元數據也越複雜。

具體塊/簇大小對文件系統設計帶來什麼樣的挑戰？我們先來看一下（目前還在用的）最簡單的文件系統怎麼存文件的吧：

FAT系文件系統與簇大小

在 FAT 系文件系統(FAT12/16/32/exFAT)中，整個存儲空間除了一些保留扇區之外，被分為兩大塊區域，看起來類似這樣：



前一部分區域放文件分配表（File Allocation Table），後一部分是實際存儲文件和目錄的數據區。數據區被劃分成「簇（cluster）」，每個簇是一到多個連續扇區，然後文件分配表中表項的數量 決定了後面可用空間的簇的數量。文件分配表（FAT）在 FAT 系文件系統中這裏充當了兩個重要作用：

1. 管理簇空間分配。空間分配器可以掃描 FAT 判斷哪些簇處於空閒狀態，那些簇已經被佔用，從而分配空間。
2. 對現有文件，FAT 表中的記錄形成一個單鏈表結構，用來尋找文件的所有已分配簇地址。

目錄結構		
文件名. 擴展名	文件屬性	起始簇
文件名. 擴展名	文件屬性	起始簇
文件名. 擴展名	文件屬性	起始簇
文件名. 擴展名	文件屬性	起始簇

直觀上理解，FAT表像是數據區域的縮略圖，數據區域有多少簇，FAT表就有多少表項。FAT系文件系統中每個簇有多大，由文件系統總容量，以及FAT表項的數量限制。我們來看一下微軟文件系統默認格式化的簇大小（數據來源）：

Volume Size	FAT16	FAT32	exFAT	NTFS
< 8 MiB	N/A	N/A	4KiB	4KiB
8 MiB – 16 MiB	512B	N/A	4KiB	4KiB
16 MiB – 32 MiB	512B	512B	4KiB	4KiB
32 MiB – 64 MiB	1KiB	512B	4KiB	4KiB
64 MiB – 128 MiB	2KiB	1KiB	4KiB	4KiB
128 MiB – 256 MiB	4KiB	2KiB	4KiB	4KiB
256 MiB – 512 MiB	8KiB	4KiB	32KiB	4KiB
512 MiB – 1 GiB	16KiB	4KiB	32KiB	4KiB
1 GiB – 2 GiB	32KiB	4KiB	32KiB	4KiB
2 GiB – 4 GiB	64KiB	4KiB	32KiB	4KiB
4 GiB – 8 GiB	N/A	4KiB	32KiB	4KiB
8 GiB – 16 GiB	N/A	8KiB	32KiB	4KiB
16 GiB – 32 GiB	N/A	16KiB	32KiB	4KiB
32 GiB – 16TiB	N/A	N/A	128KiB	4KiB
16 TiB – 32 TiB	N/A	N/A	128KiB	8KiB
32 TiB – 64 TiB	N/A	N/A	128KiB	16KiB
64 TiB – 128 TiB	N/A	N/A	128KiB	32KiB
128 TiB – 256 TiB	N/A	N/A	128KiB	64KiB

> 256 TiB N/A N/A N/A N/A

用於軟盤的時候 FAT12 的簇大小直接等於扇區大小 512B，在容量較小的 FAT16 上也是如此。FAT12 和 FAT16 都被 FAT 表項的最大數量限制（分別是 4068 和 65460），FAT 表本身不會太大。所以上表中可見，隨着設備容量增加，FAT16 需要增加每簇大小，保持同樣數量的 FAT 表項。

到 FAT32 和 exFAT 的年代，FAT 表項存儲 32bit 的簇指針，最多能有接近 4G 的 FAT 表項，從而表項數量理應不再限制 FAT 表大小，從而理應使用同樣的簇大小。不過事實上，簇大小仍然根據設備容量增長而增大。FAT32 上 256MiB 到 8GiB 的範圍內使用 4KiB 簇大小，隨後簇大小開始增加；在 exFAT 上 256MiB 到 32GiB 使用 32KiB 簇大小，隨後增加到 128KiB。增加簇大小的原因是爲了限制 FAT 表整體的大小，因爲在使用 FAT 表的文件系統中，需要將 FAT 表整體裝入內存才能滿足文件訪問和簇分配時的性能，如果讀寫 FAT 表的範圍需要訪問磁盤，那麼整個文件系統的讀寫性能將暴跌到幾近不可用。到針對閃存優化的 exFAT 上，雖然在 FAT 表外還有額外簇分配位圖，但是也同樣要限制 FAT 表整體大小，減少對 FAT 表區域的隨機讀寫。

使用如此大的簇大小，導致的劣勢在於極度浪費存儲空間。

