内部碎片与小文件优化

上篇「系统中的大多数文件有多大？」提到，文件系统中大部分文件其实都很小，中位数一直稳定在 4K 左右，而且这个数字并没有随着存储设备容量的增加而增大。但是存储设备的总体容量实际是在逐年增长的，，总容量增加而文件大小中位数不变的原因，可能是以下两种情况：

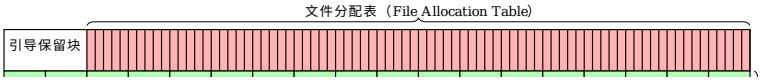
1. 文件数量在增加
2. 大文件的大小在增加

实际上可能是这两者综合的结果。这种趋势给文件系统设计带来了越来越多的挑战， 因为我们不能单纯根据平均文件大小来增加块大小（block size）优化文件读写。 微软的文件系统（FAT系和 NTFS）使用「簇（cluster）」这个概念管理文件系统的可用空间分配，在 Unix 系文件系统中有类似的块（block）的概念，只不过称呼不一样。 现代文件系统都有这个块大小或者簇大小的概念，从而基本的文件空间分配可以独立于硬件设备本身的扇区大小。 块大小越大，单次分配空间越大，文件系统所需维护的元数据越小，复杂度越低，实现起来也越容易。 而块大小越小，越能节约可用空间，避免内部碎片造成的浪费，但是跟踪空间所需的元数据也越复杂。

具体块/簇大小对文件系统设计带来什么样的挑战？我们先来看一下（目前还在用的）最简单的文件系统怎么存文件的吧：

FAT系文件系统与簇大小

在 FAT 系文件系统(FAT12/16/32/exFAT)中，整个存储空间除了一些保留扇区之外，被分为两大块区域，看起来类似这样：



前一部分区域放文件分配表（File Allocation Table），后一部分是实际存储文件和目录的数据区。 数据区被划分成「簇（cluster）」，每个簇是一到多个连续扇区，然后文件分配表中表项的数量 决定了后面可用空间的簇的数量。文件分配表（FAT）在 FAT 系文件系统中这里充当了两个重要作用：

1. 管理簇空间分配。空间分配器可以扫描 FAT 判断哪些簇处于空闲状态，那些簇已经被占用，从而分配空间。
2. 对现有文件，FAT 表中的记录形成一个单链表结构，用来寻找文件的所有已分配簇地址。

目录结构		
文件名. 扩展名	文件属性	起始簇
文件名. 扩展名	文件属性	起始簇
文件名. 扩展名	文件属性	起始簇
文件名. 扩展名	文件属性	起始簇

直观上理解，FAT表像是数据区域的缩略图，数据区域有多少簇，FAT表就有多少表项。 FAT系文件系统中每个簇有多大，由文件系统总容量，以及 FAT 表项的数量限制。 我们来看一下微软文件系统默认格式化的簇大小（数据来源）：

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

用于软盘的时候 FAT12 的簇大小直接等于扇区大小 512B ，在容量较小的 FAT16 上也是如此。 FAT12 和 FAT16 都被 FAT 表项的最大数量限制（分别是 4068 和 65460 ），FAT 表本身不会太大。 所以上表中可见，随着设备容量增加， FAT16 需要增加每簇大小，保持同样数量的 FAT 表项。

到 FAT32 和 exFAT 的年代，FAT 表项存储 32bit 的簇指针，最多能有接近 4G 的 FAT 表项，从而表项数量理应不再限制 FAT 表大小，从而理应使用同样的簇大小。不过事实上， 簇大小仍然根据设备容量增长而增大。 FAT32 上 256MiB 到 8GiB 的范围内使用 4KiB 簇大小，随后簇大小开始增加；在 exFAT 上 256MiB 到 32GiB 使用 32KiB 簇大小，随后增加到 128KiB 。增加簇大小的原因是为了限制 FAT 表整体的大小，因为在使用 FAT 表的文件系统中，需要将 FAT 表整体装入内存才能满足文件访问和簇分配时的性能，如果读写 FAT 表的范围需要访问磁盘，那么整个文件系统的读写性能将暴跌到几近不可用。到针对闪存优化的 exFAT 上，虽然在 FAT 表外还有额外簇分配位图，但是也同样要限制 FAT 表整体大小，减少对 FAT 表区域的随机读写。

使用如此大的簇大小，导致的劣势在于极度浪费存储空间。