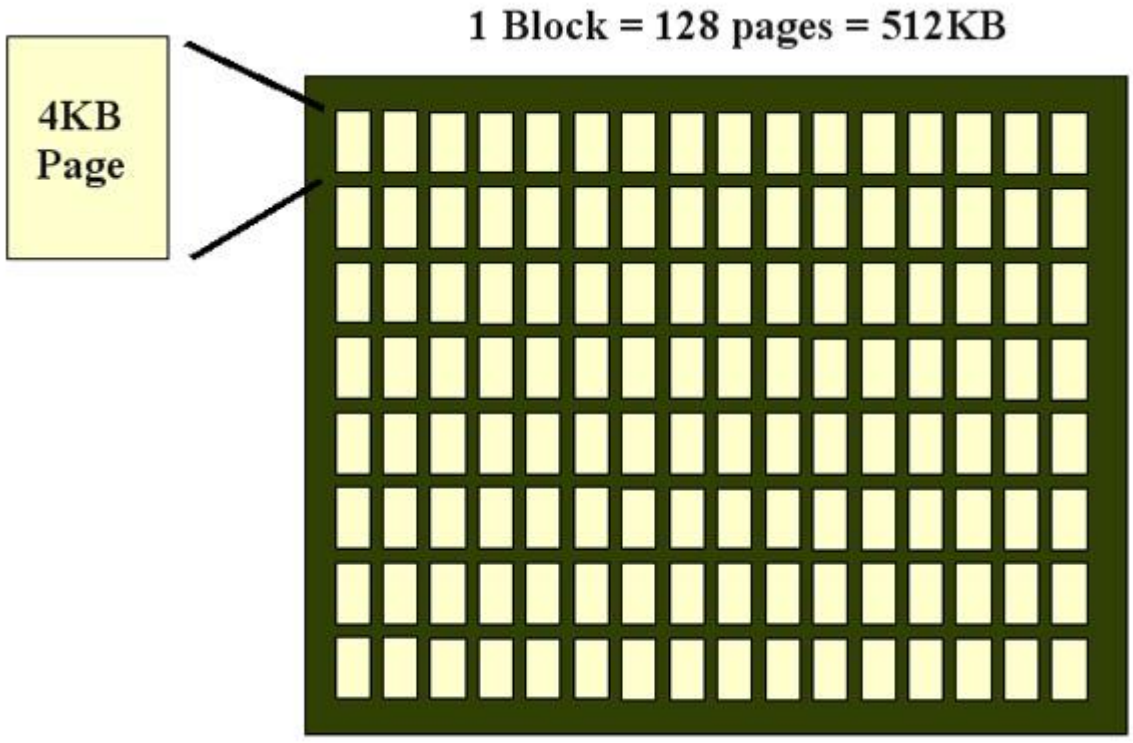
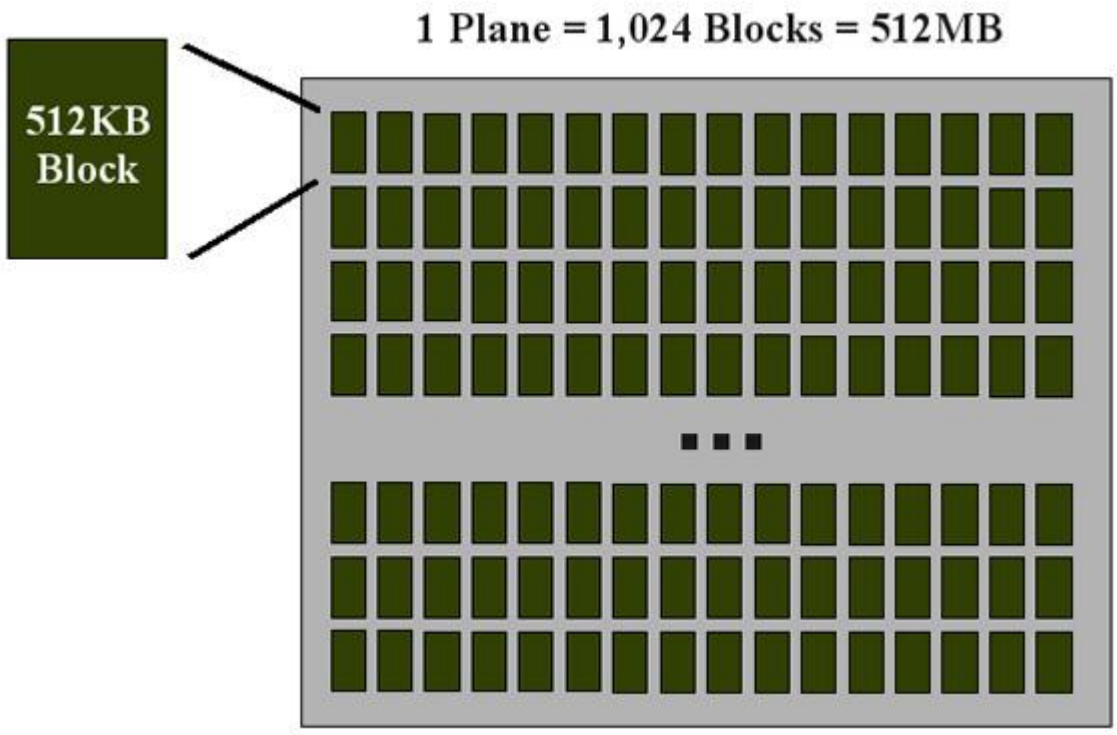
【SSD 科普】为啥越用越慢？揭秘 NAND 闪存的读写过程

SSD 与传统 HDD 有着不同的架构和原理，读写数据的过程也是不一样的，这种原理上的不同带给 SSD 优异的性能，但是也决定了 SSD 固有的一些缺点。



HDD 磁盘有扇区、柱面之分，SSD 的基本组成也有 Page（页面）、Block（区块）、Plane（平面）之分，page 是最基本的组成，大小一般是 4KB，，每个 block 通常包含 64 个 page，容量是 256KB，也有 128 个 page 的，容量就是 512KB，不过目前主流的 25nm 工艺闪存普遍都是 8KB page 容量，128 个 page 配置。

多个 block 再组成 plane，而 plane 就是就是闪存中的一颗核心（die）了，而我们看到的闪存片其实是多颗 die 封装在一起的，一般是 2-8 颗，而整个 SSD 上则会由多片闪存组成。



实际上，如果 SSD 内部是以 die 颗粒的 RAID 0 模式组建的，那么 block 层级之上还有一个 band 之分，它是 RAID 0 模式中所有芯片的同一块 block 区块的总和。SSD 横向评测的基础知识部分大家一定要细读，这样可以为我们看更多 SSD 数据提供补充。

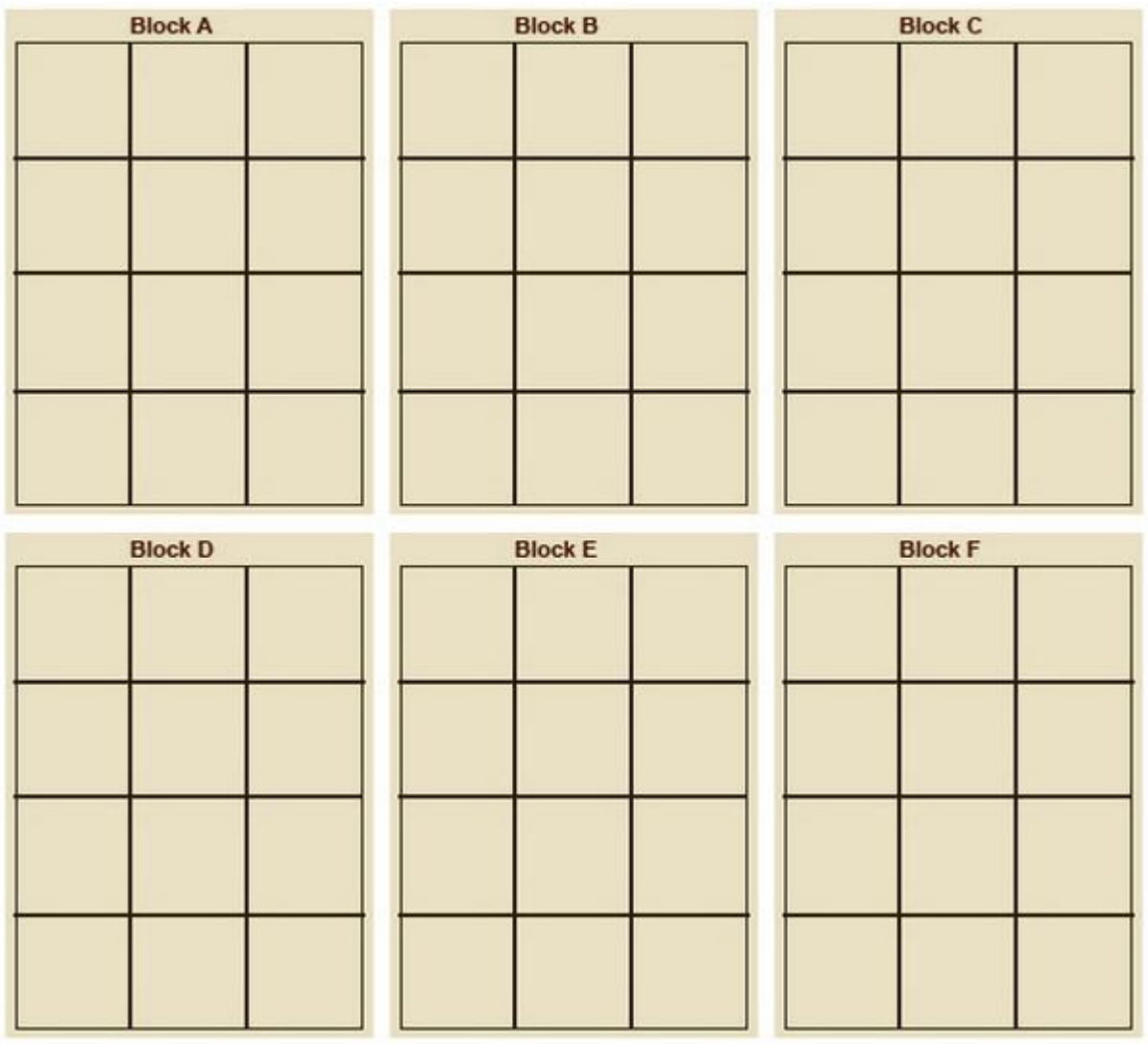
简单的描述就是这样：page→block→band→plane-die→闪存片→SSD。

数据读写的主要过程就在 page、block 以及 band 三个层面上。

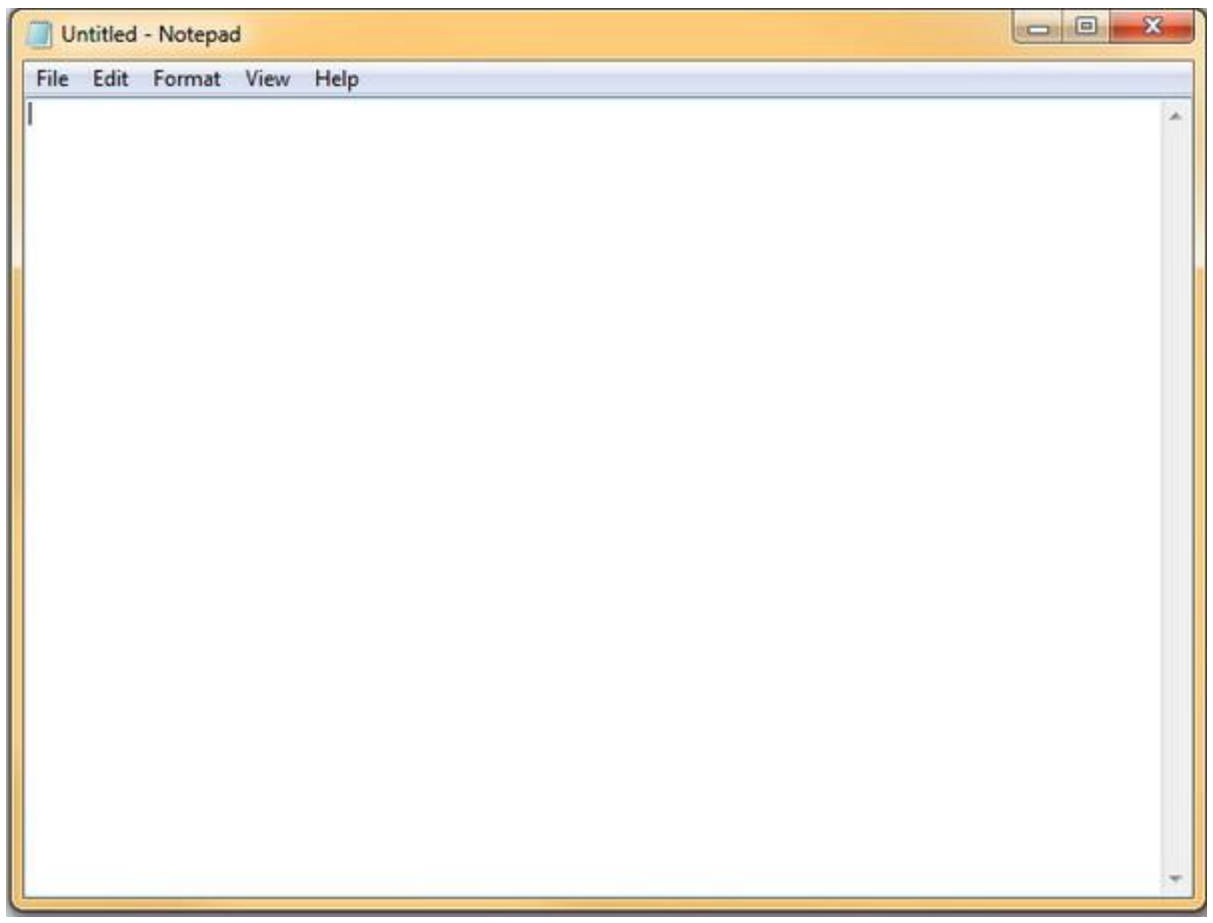
在系统中，数据写入是以 page 为单位的，SSD 写入新数据擦除原有的数据，但是擦除过程只能以 block 为单位，要清除就得擦除整个 block 单元，哪怕只写入了一个 page 的文件。

在一篇国外博文中找到一个非常简单形象的 SSD 写数据的描述，我们来看看 SSD 到底是怎么写入数据的。

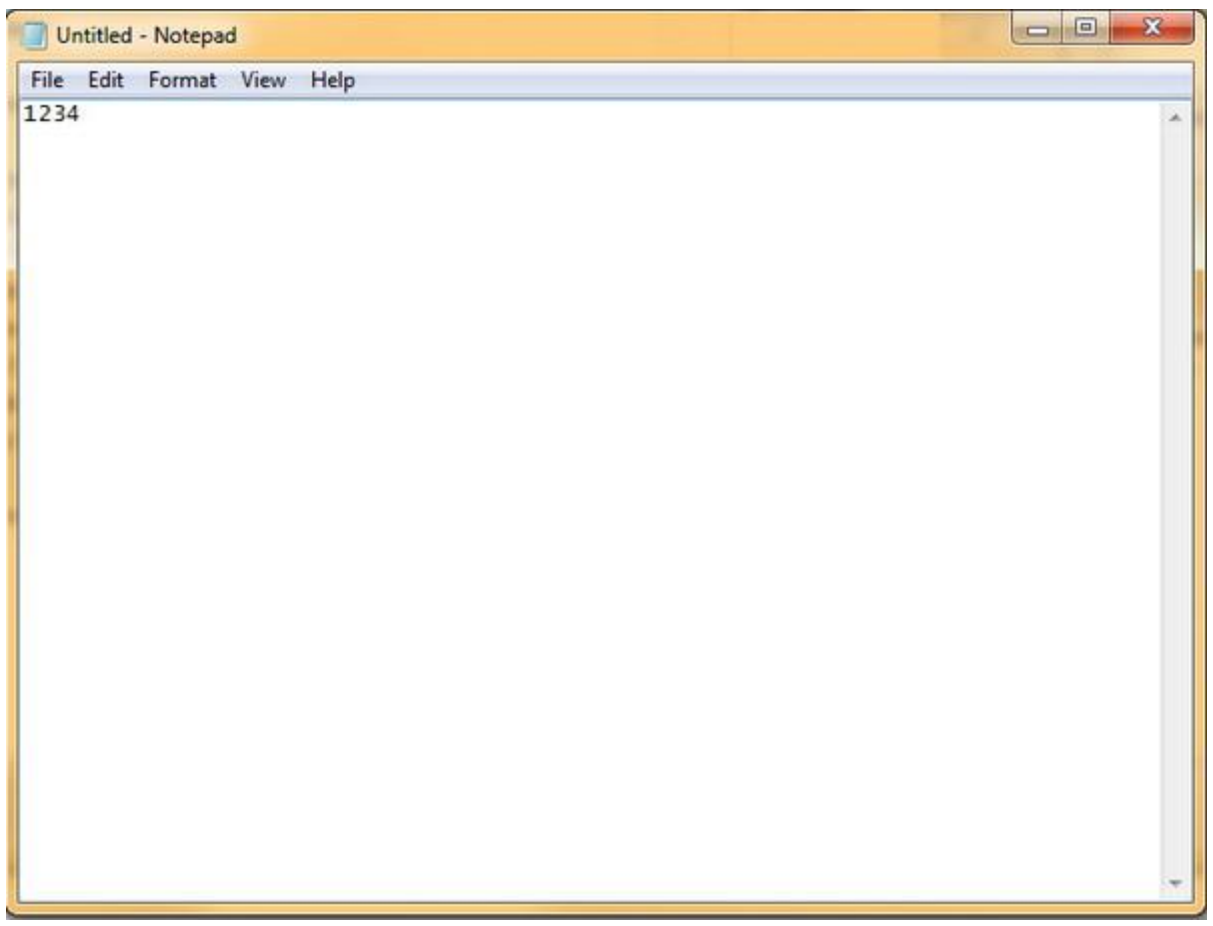
这里为了简化说明，假设每个 block 只有 12 个 page，每个 page 大小 1Byte。



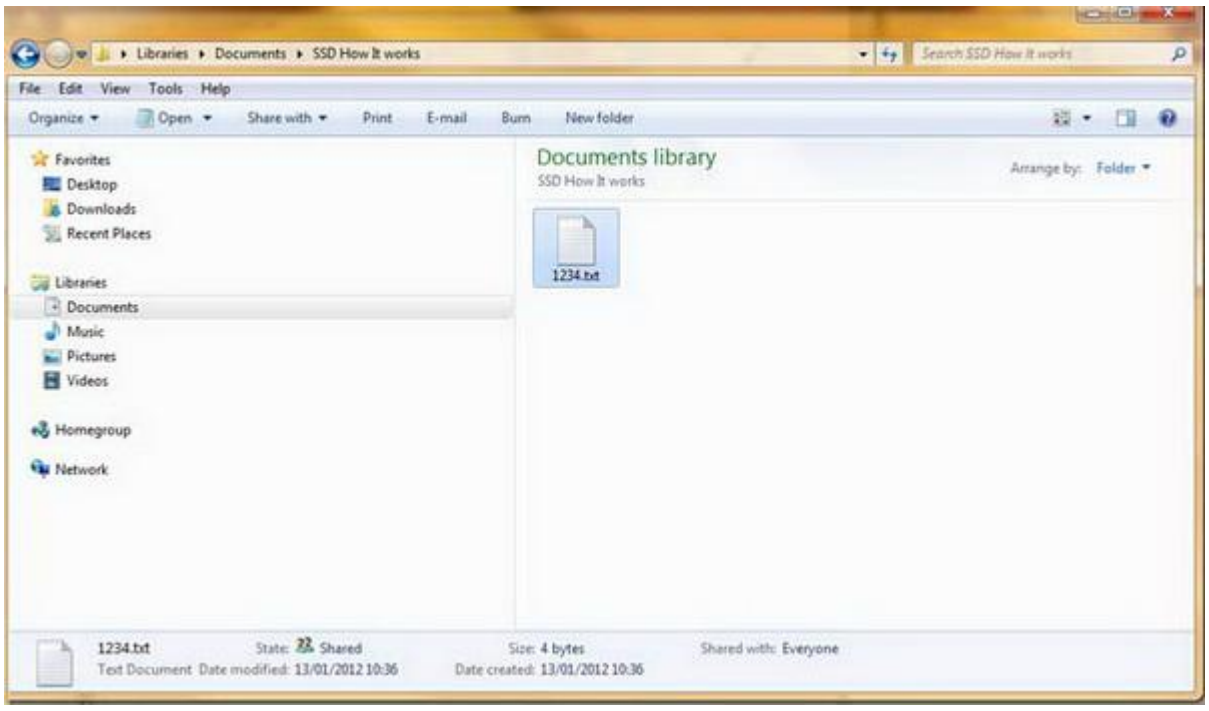
无数据的情况就是这样的，SSD 性能最好的状态



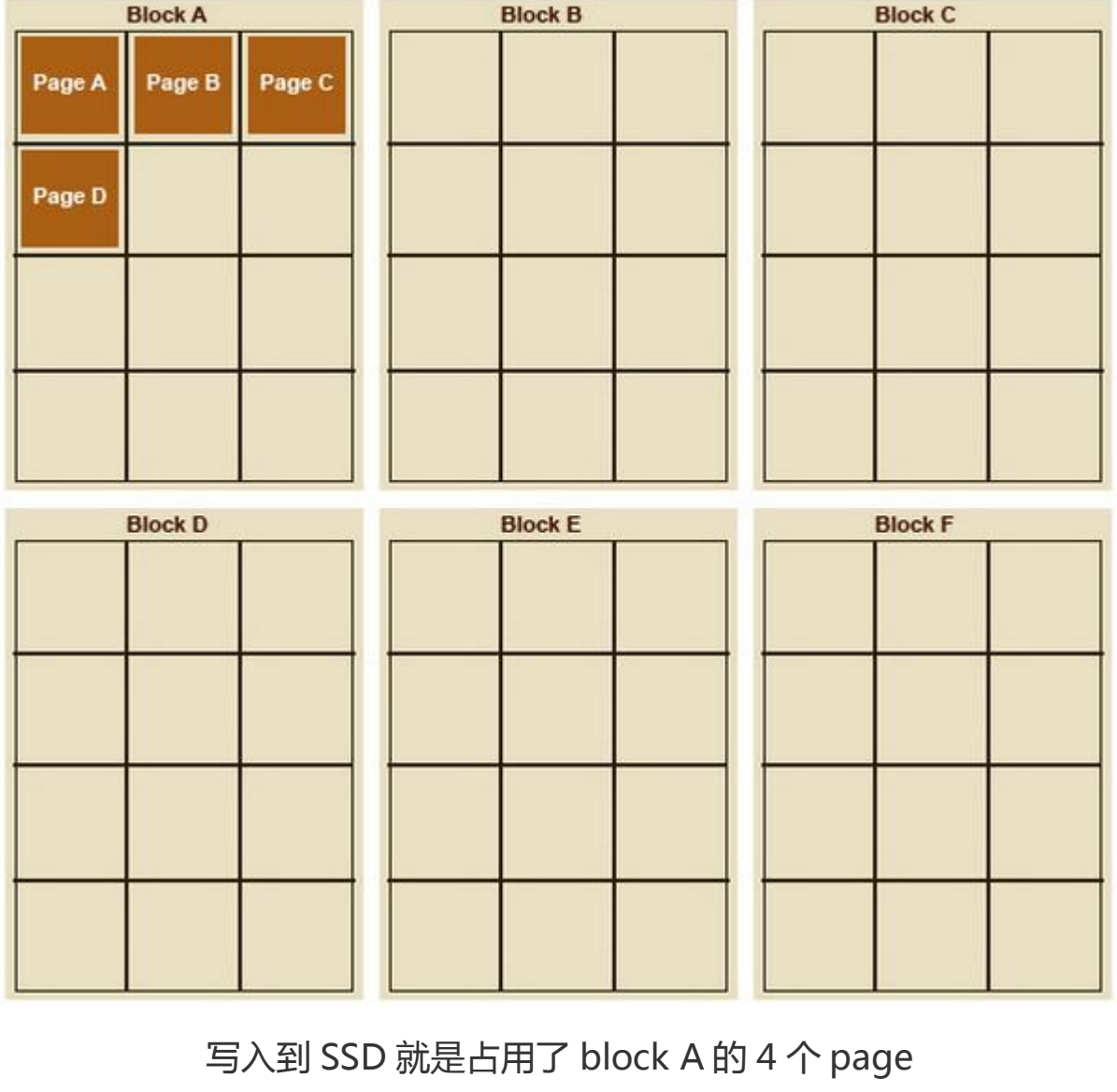
打开记事本程序



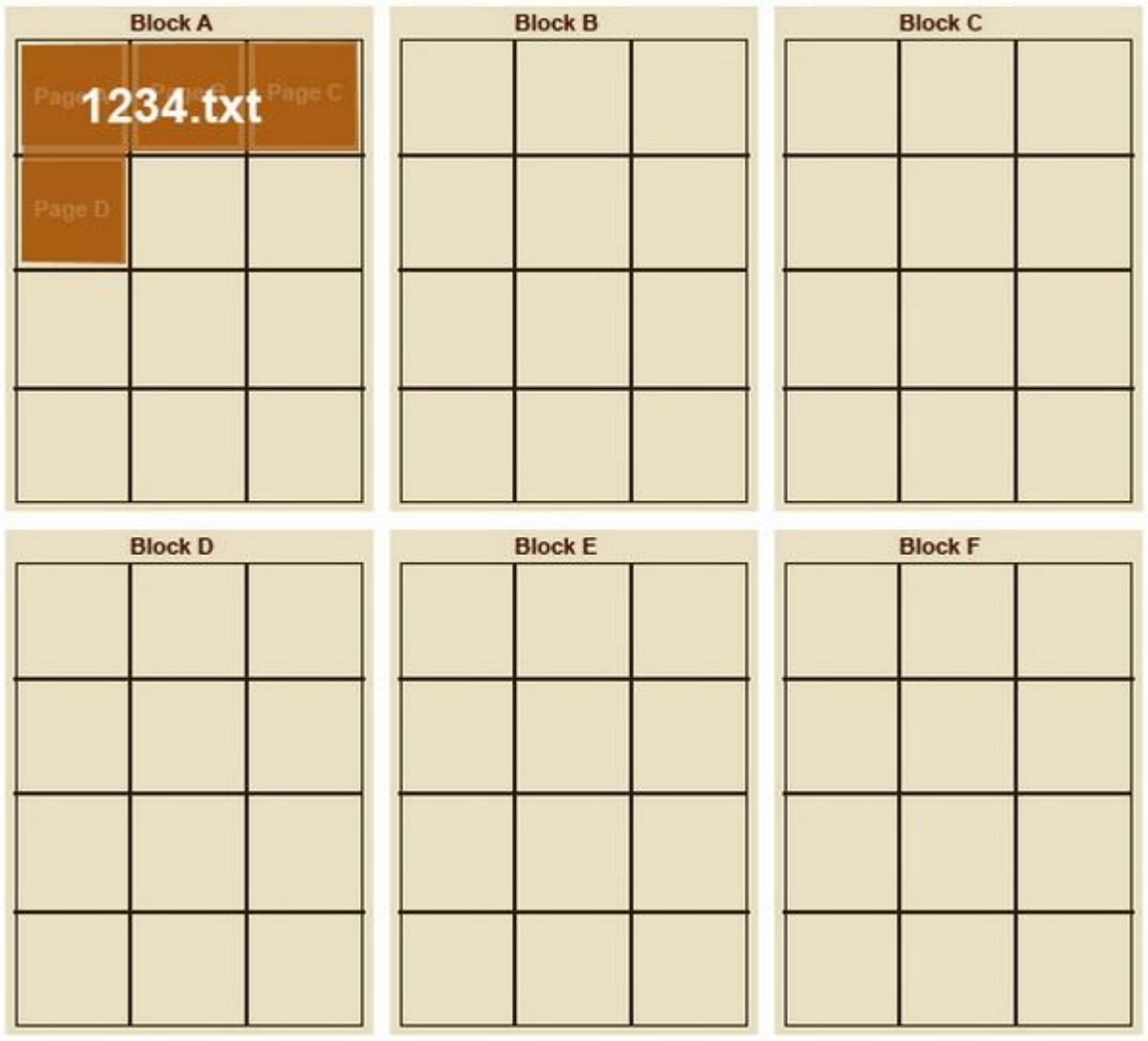
输入 1234 并保存

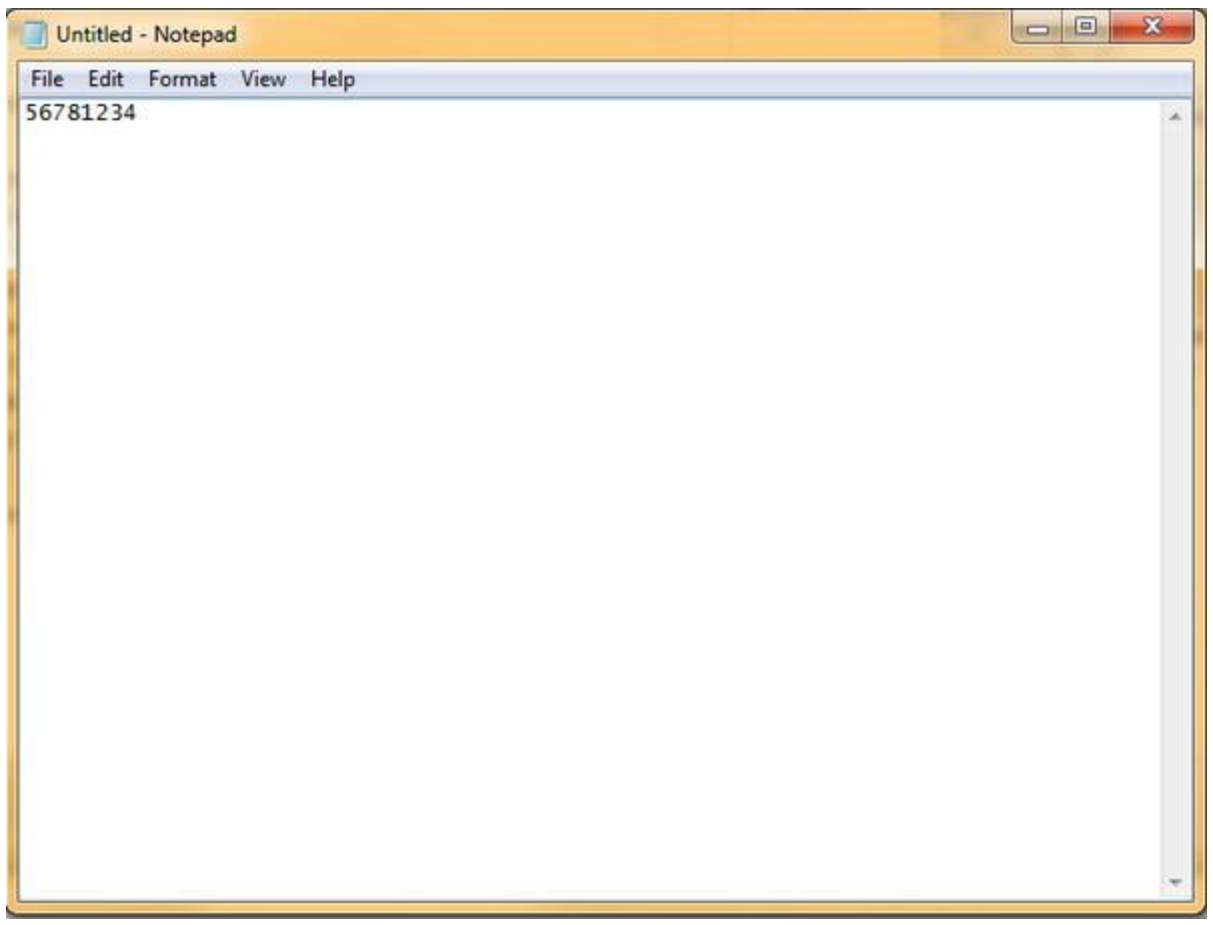


这个文件的大小正好是 4Byte

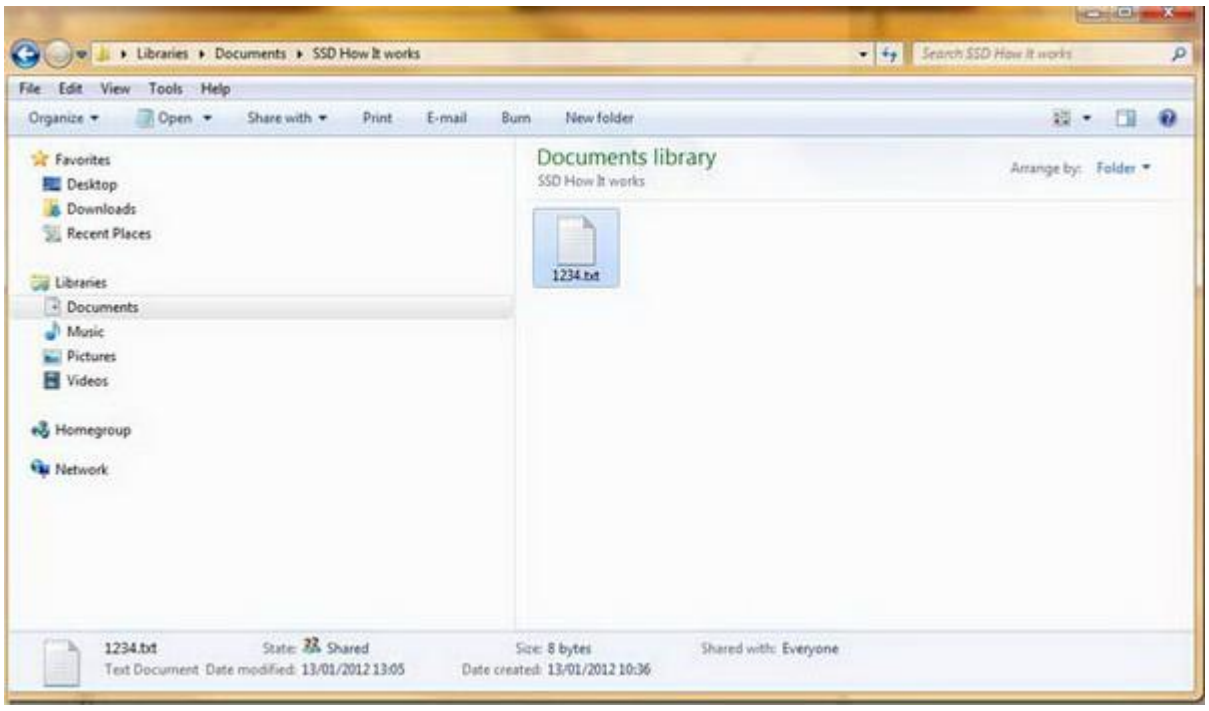


写入到 SSD 就是占用了 block A 的 4 个 page



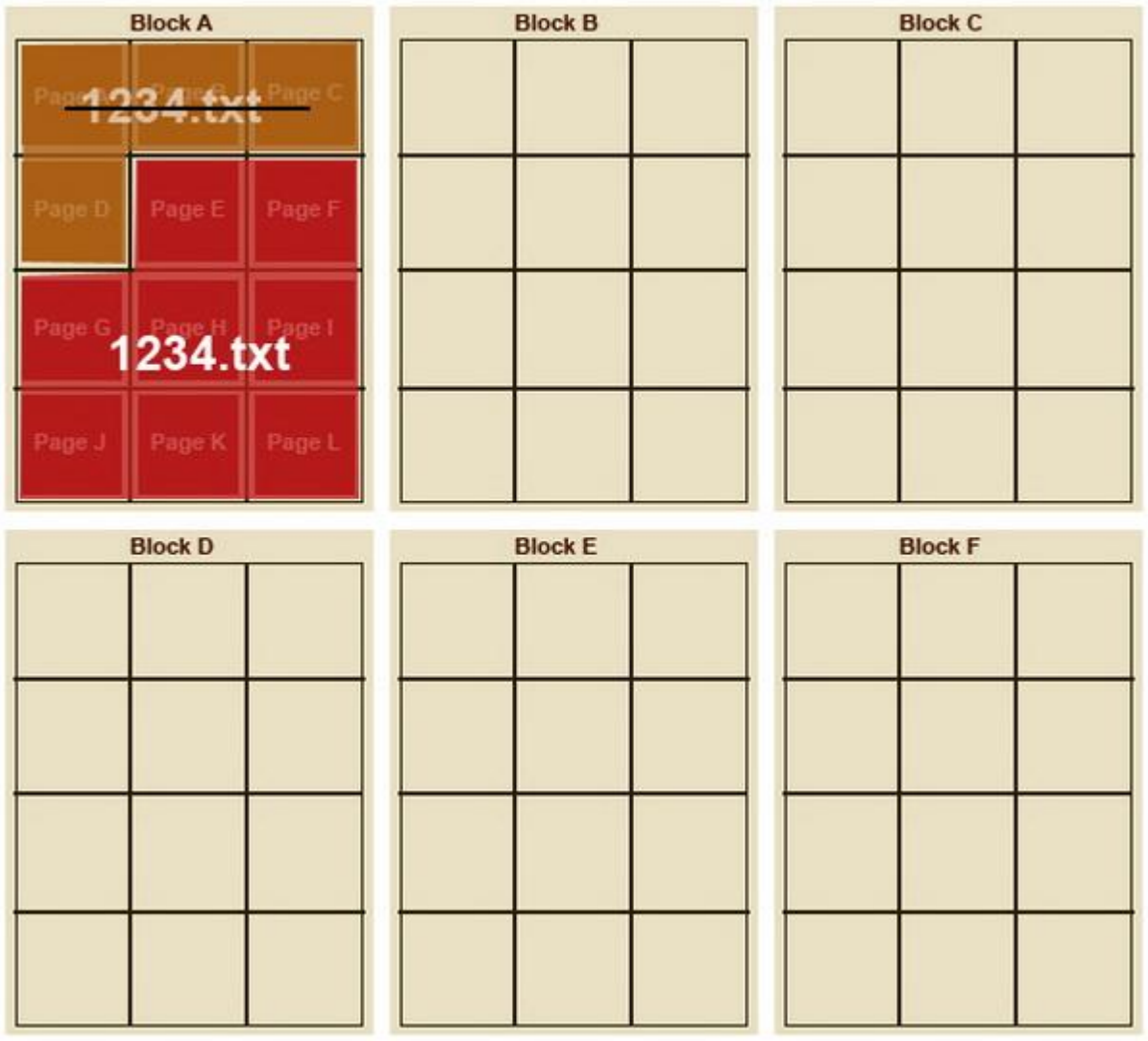


改变原来的文档内容，变成 56781234



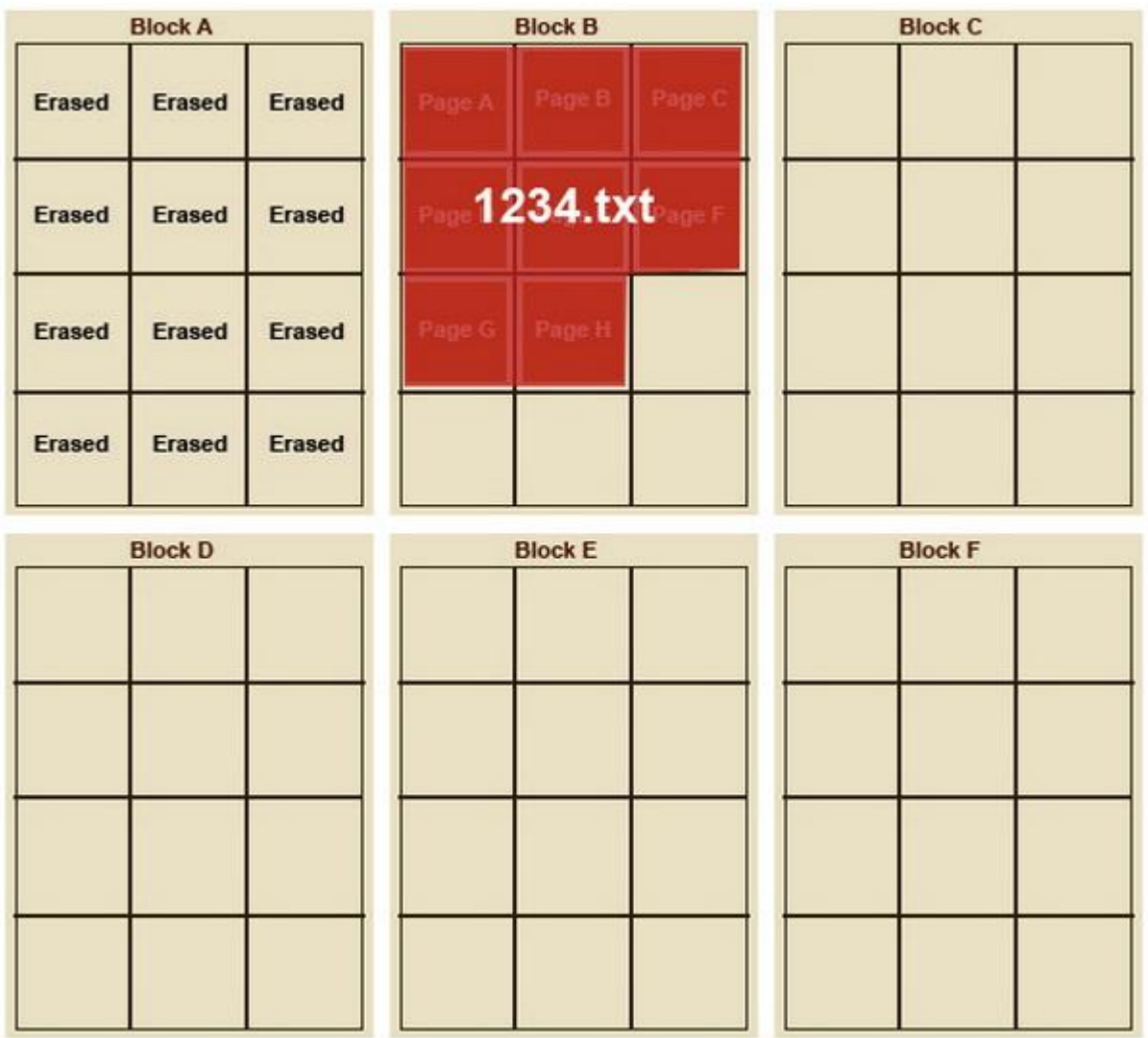
现在文档大小变成了 8B

保存时 SSD 不能直接覆盖原有文件，需要重新占用 8 个 page 文件

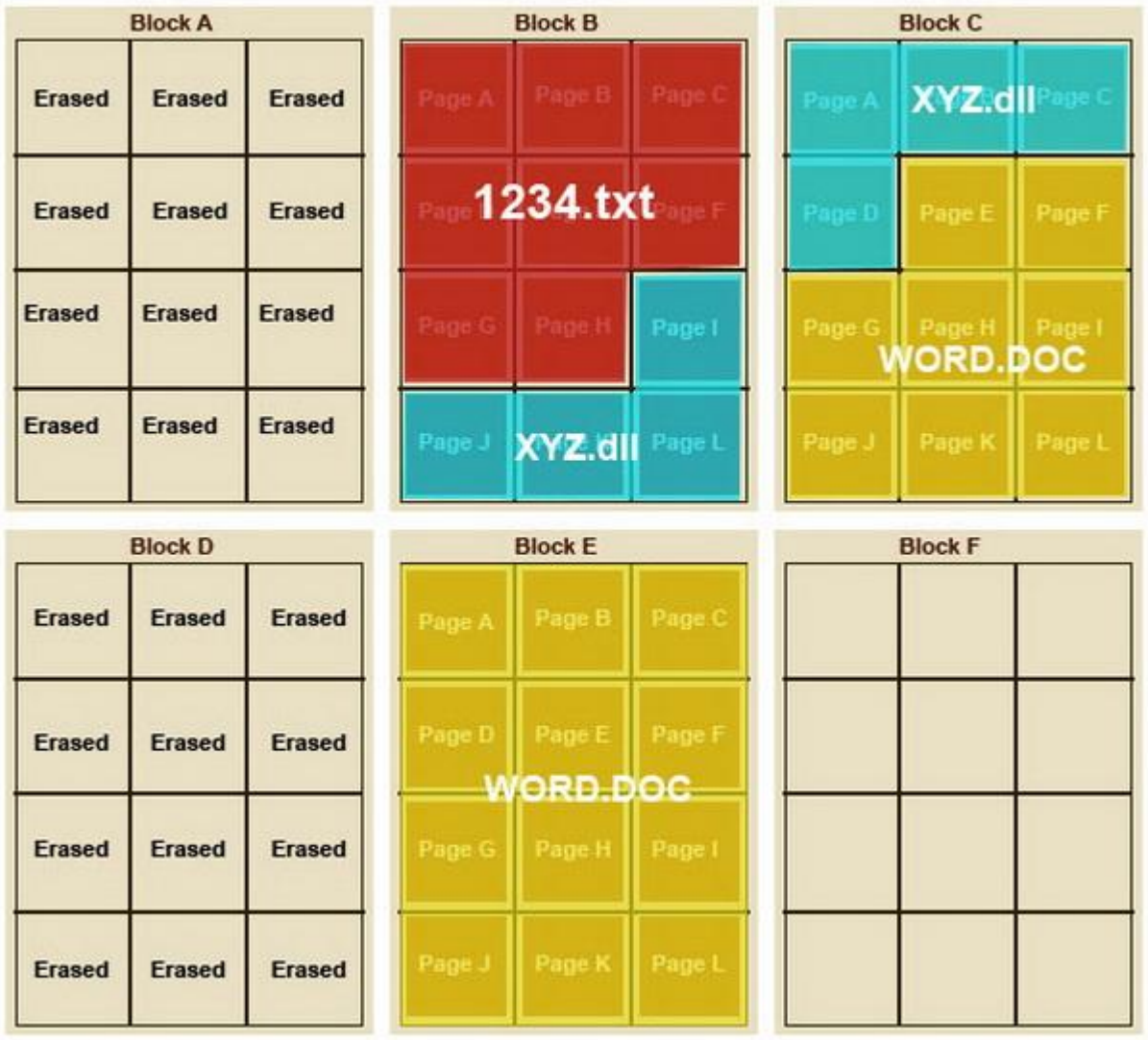


现在就是这个样子了

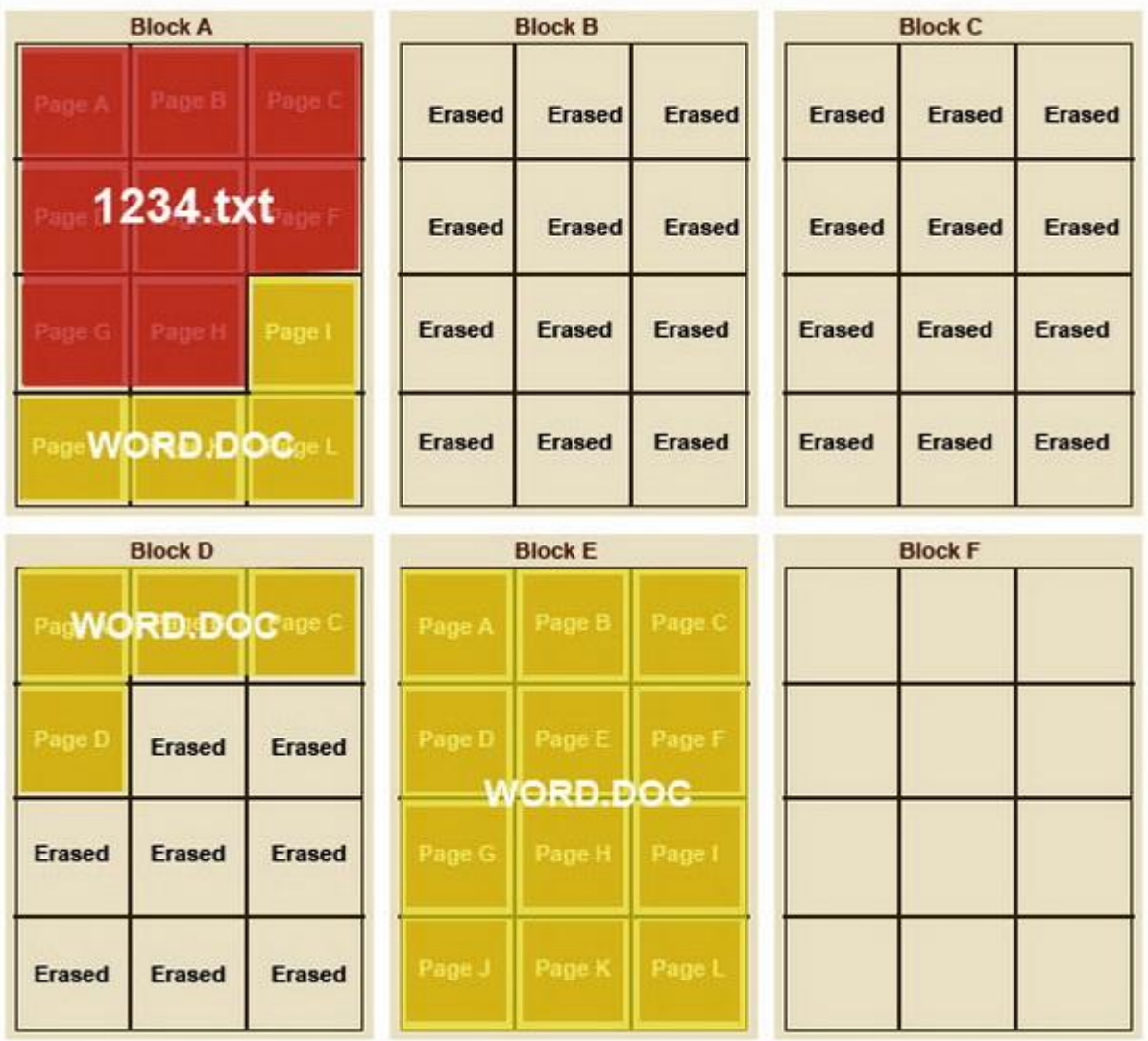
Block A 实际上已经写“脏”了，要恢复性能就需要删除整个 block 区块，此时需要把有用的数据拷贝到另一个空白 blokc 中然后再清除 Block A。



SSD 评测数据写入到空白的 block B 中

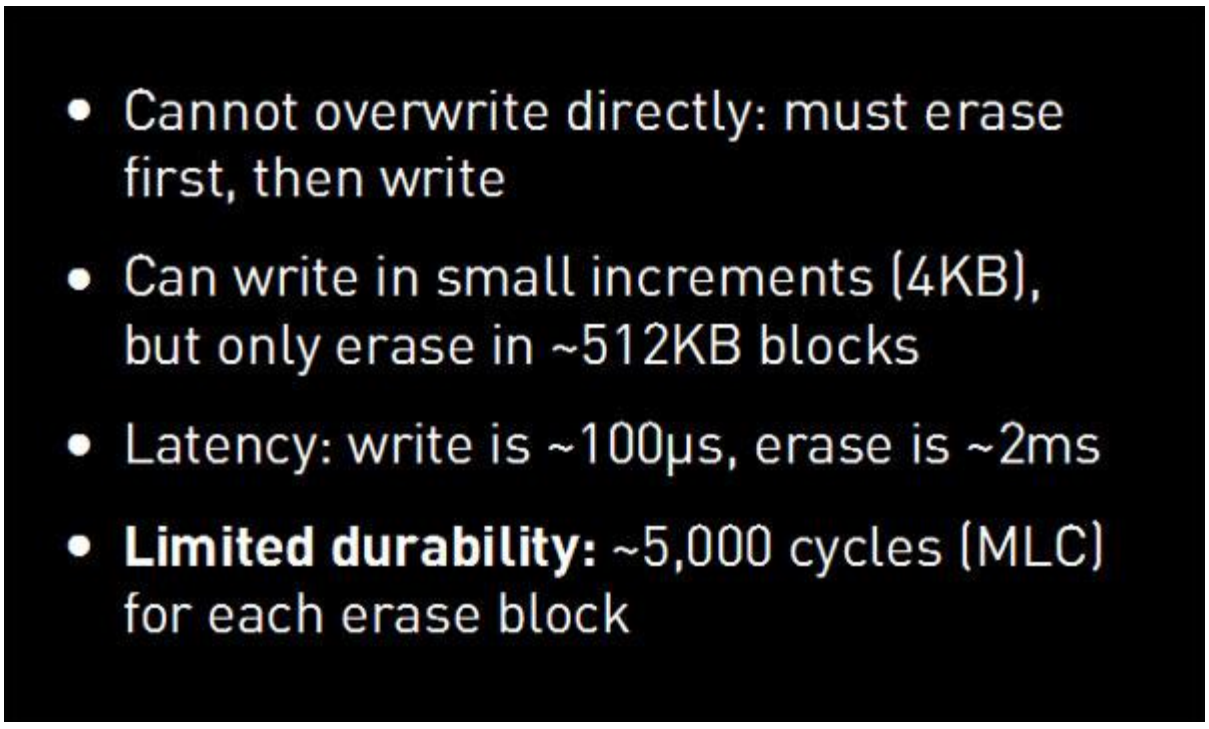


实际过程中数据显然不是只有这么简单，略微复杂一点的情况就如上图所示，1234.txt 文档占用了 8 个 page，xyz.dll 也是占用 8 个 page，但是分别在两个 block 区块中，word.doc 文件也是占用了两个 block，其中占满一个，另一个占用了 8 个 page。



此时如果用户删除了 xyz.dll 文件，那么数据就要重新洗牌，Block B 中的 1234.txt 重新拷贝到 block A 中，doc 文件中的 4B 也要写入 block A 中，还有多余的 4 个 page 要再占用 block D 的 4 个 page 空间，而 block E 中的数据是满的，不需要移动，此时的排列就如上图所示，腾出来的 block B 和 block C 也就可以清除数据以恢复性能了。

上述过程还只是非常简单的例子，如果是真实的应用环境情况会更复杂，SSD 需要不断地在各个 block 之间进行写入-转移-清空操作 ,而且 SSD 的写入速度与擦除速度相差很大 ,这也会影响 SSD 的性能发挥 ,SSD 评测中就会体现出相应的不可避免的性能下降。



上图文字的翻译：

1. 不能直接覆盖写入，必须先擦除，然后再写入
2. 写入的时候是按照 PAGE 的，为 4KB，但擦除的时候是按照 BLOCK 的，为 512KB
3. 操作时间：写入 100us，擦除 2ms=2000us
4. 寿命限制：每个 BLOCK 的擦除次数为 5000 次（对也 MLC 颗粒来说）

总之，**SSD 的特性决定了它的写入方式，不能直接覆写数据使得 SSD 多了擦除的操作，而写入单位与擦除单位的不统一又让 SSD 不停地在各个 Block 区块之间折腾，而写入数据的延迟约为 0.2ms，但擦除操作需要 2ms 左右，SSD 用久了需要擦除的区块就会越多，性能自然也会变慢。**

这些问题都是 SSD 必须处理的，影响可大可小，也让很多人开始对 SSD 的可靠性不放心，下一个关注点自然就是 SSD 的使用寿命了。