# 3D NAND Flash存储器的系统级应用

## 固态硬盘背景

图1‑1显示了计算机系统的内存层次结构。顶层的存储器速度更快，但容量更小（比特成本更高）。在合同中，底层的内存速度较慢，但容量更大（比特成本较低）。scm、NAND闪存和硬盘驱动器(HDD)是非易失性的。在存储器层次结构，NAND闪存位于SCM和硬盘之间。由于NAND闪存通过微缩和MLC技术不断降低位成本，SSD作为HDD的替代方案变得具有成本效益。

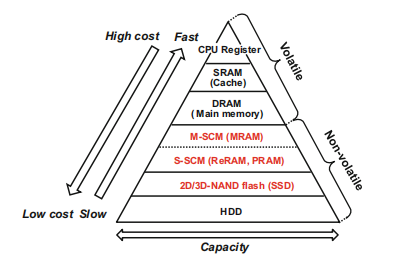


图1‑1 分级存储器体系

图1‑2说明了NAND闪存的组织结构。与同一字行连接的存储器单元由一个页面组成，该页面是NAND闪存的读写单元。一个块是擦除单位。MLC NAND闪存一个块中有典型的128-256页。

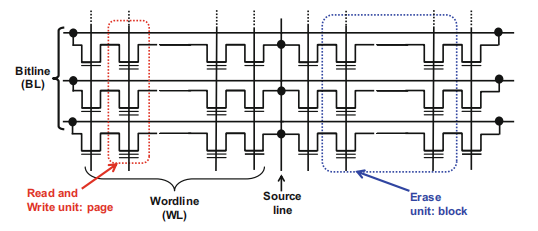


图1‑2 NAND闪存组织

所有3D-NAND闪存SSD、混合M-SCM/3D-NAND闪存SSD和所有S-SCMSSD的体系结构如图1‑3所示。SSD的关键组件是SSD控制器，它集成了Flash转换层(FTL)，使SSD可以作为一个块设备工作。如图1‑4所示，FTL中基本但非常关键的功能是逻辑到物理的地址转换，这是由于NAND闪存的会出现坏块而需要映射到其他区块。根据映射的粒度，可以将地址转换划分为页面级映射、块级映射和混合映射。当一个页面数据被覆盖时，从旧页面读取旧数据，与新数据合并，并写入新页面。在此之后，旧的页面将失效。因此，在SSD中有三个页面状态：可用的页面、包含有效数据的页面和包含无效数据的页面。频繁访问的数据（热数据）会产生大量的无效页面。

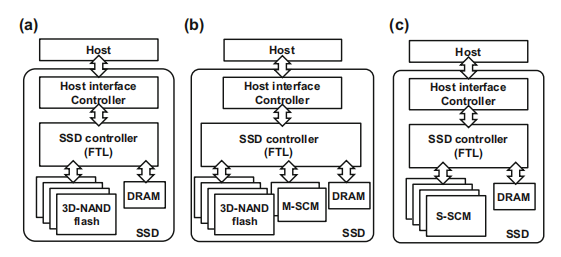


图1‑3全SSD架构，a) 3D-NAND闪存SSD b)混合M-SCM/3D-NAND闪存SSD和c)所有S-SCMSSD

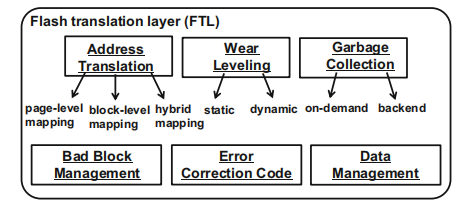


图1‑4 flash转换层中的基本功能(FTL)

当SSD可用空间减少到低于阈值（平面中的几个空闲块）时，FTL将按需或后台触发操作、垃圾回收（GC)，以回收一个或多个旧块。在删除旧块之前，必须将块中的所有有效页面复制到另一个块中的空闲空间，如图1‑5所示。因此，GC的延迟随着回收块中有效页数的增加而增加。当这种页面复制开销很大时，GC将成为SSD写入性能的瓶颈。此外，FTL中的磨损水平保证了NAND闪存块的均衡磨损，以最大限度地提高SSD的使用寿命。根据NAND闪存块中的静态数据是否周期性地移动，磨损水平可分为静态和动态磨损水平。其他功能，如纠错代码(ECC)和坏块管理(BBM)也是必不可少的。

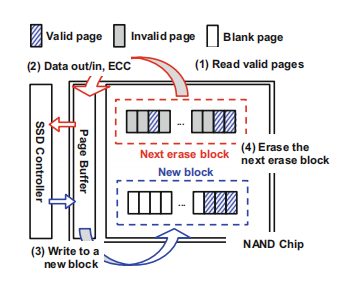


图1‑5 垃圾回收操作

## SSD性能改进技术

本章介绍了提高基于3D-NAND闪存SSD写入性能的三种技术：存储引擎辅助SSD（SEA-SSD）、逻辑块地址(LBA)）加密SSD和混合M-SCM/3D-NAND闪存SSD。前两种技术都是基于SSD控制器和中间件的协同设计。最后一种技术是将M-SCM引入SSD系统。最后，设计了所有S-SCMSSD作为长期解决方案。

### 存储引擎辅助SSD（SEA-SSD）

数据库是企业服务器中应用最广泛的应用程序之一。数据库的存储引擎，控制着数据应该存储在存储器中的时间和位置。因此，第一种技术，存储引擎辅助SSD(SEA-SSD)，与SSD控制器共同设计了存储引擎，以提高数据库的SSD写入性能。它是基于存储堆栈的上层比底层拥有更丰富的信息的思想。对于当前的SSD，它只接收来自操作系统(OS)的块设备层的信息，其中包括数据、数据大小和数据地址。这些信息相当有限。

图1‑6显示了传统计算机系统与建议采用SEA-SSD的计算机系统之间的比较。由于原因(i)存储层，如文件系统、块层等。针对传统的硬盘进行了优化，但不针对SSD，传统的OS对于SSD存储是低效的，(ii)如果提示信息必须通过操作系统中的所有层，则需要大量的工程努力，那么在提议的SEA-SSD中只是简单地绕过操作系统。图1‑6所示，从SE（存储引擎）传递到SSD控制器，以便更有效地存储数据。

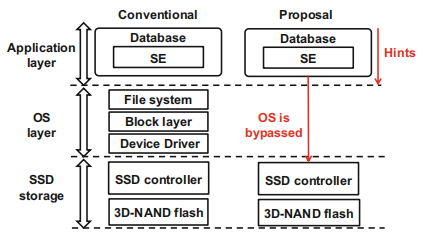


图1‑6 SEA-SSD概念

图1‑7显示了SEA-SSD的体系结构。每个3D-NAND闪存芯片被分为两个逻辑段。热数据（经常访问数据）和冷数据（很少访问数据）。通过在同一块中聚合具有类似活动的数据，可以减少GC开销。为了确定每个段的大小，第一种提示被发送到SSD控制器，这是基于SE设置和热数据大小之间的强相关性。对于Innodb存储引擎，这些设置包括缓冲区池和重做日志大小。缓冲区池缓存频繁访问的数据（热数据），而重做日志用于崩溃恢复，以保证Innodb存储引擎的持久性。第二个提示是用动态数据模型进行数据初步分类。如果数据在刷新时被存储引擎判定为热，则将逻辑“1”发送到SSD控制器，表示数据为热，因此存储在Seg-hot中，如图1‑8所示。否则，它只是存储在Seg-Cold中。当存储在3D-NAND闪存中的数据的活动随时间变化时，当在SSD中触发GC时，再次用第三个提示预测数据。第三个提示是第一次进入刷新列表的页面数据的逻辑地址，因为这些数据将很快被刷新到SSD。如图1‑8所示，这些数据应存储在Seg-Hot中，而其他数据在GC后存储在Seg-Cold中。

为了评估SEA-SSD，基于同步系统平台架构师开发了一个数据库和SSD耦合模拟器，其速度比虚拟平台快20多倍。从评价结果来看，写入性能最大值提高了24%。此外，最大16%的能耗和19%的寿命提高。

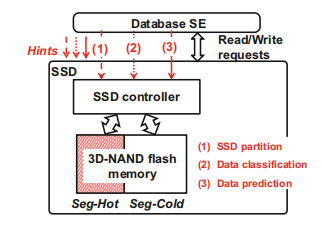


图1‑7 SEA-SSD架构

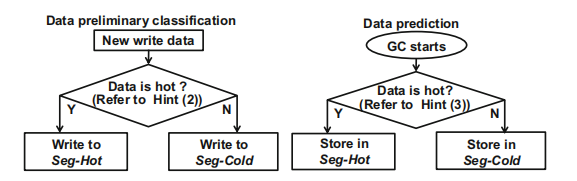


图1‑8 SEA-SSD数据管理算法

### 存储引擎辅助SSD（SEA-SSD）

SEA-SSD是一个专门为数据库应用程序而优化的设计。作为提高所有3D-NAND闪存SSD写性能的通用解决方案，提出了逻辑块地址(LBA)置码SSD。在现有的SSD系统中添加了一个基于地址重映射技术的中间件LBA scrambler。

LBA scrambler的概念是积极地减少GC的页面拷贝开销，如图1‑9所示。SSD中有三种页面：有效页面、可用页面和无效页面。在有效页面中，仍然拥有可用空间的页面称为碎片页面。下一个擦除块中的所有有效页面必须复制到另一个块的可用空间，这将导致SSD写性能下降。因此，提出了LBA scrambler将小数据主动写入下一个擦除块中的碎片页面的剩余空闲空间。由于覆盖，下一个擦除块中的所有这些碎片化页面都变得无效，SSD中的数据也变得不那么碎片化。

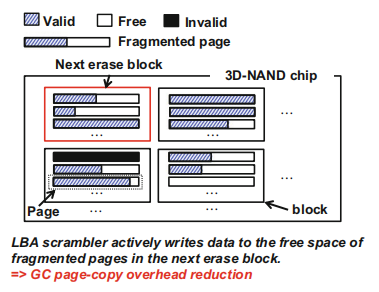


图1‑9 LBA scrambler的概念

图1‑10说明了基于LBA scrambler SSD的计算机系统。为了实现地址重映射，LBA scrambler引入了另一个逻辑地址，称为scrambler LBA(SLBA)。在LBA置乱后，数据地址SLBA被发送到SSD控制器。SSD控制器通过FTL(SSD控制器)中的逻辑到物理表将数据写入到物理3D-NAND闪存页。

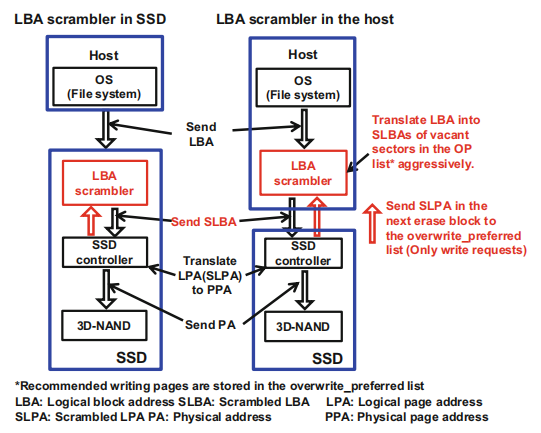


图1‑10 LBA scrambler SSD

为了通知LBA scrambler碎片化的页面地址，SSD控制器将下一个删除块中碎片化页面的打乱逻辑页面地址(SLPA)发送到LBA scrambler。要记录LBA和SLBA之间的地址重映射，LBA\_to\_SLBA表和unused\_SLBA表维护在DRAM中。如图1‑10所示，LBA scrambler 可以在SSD或主机中进行定位。当它位于SSD中时，SSD将需要一个很大的DRAM容量，但不需要进行接口修改。相比之下，如果LBA scrambler位于主机中，则需要一个较小的DRAM容量。但是，由于LBA scrambler 和SSD控制器之间的通信，SSD的接口必须进行升级。LBA scrambler 的算法流程图如图1‑11所示。

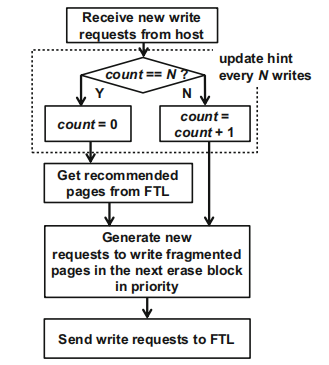


图1‑11 LBA scrambler 的算法流程图

对于每N个写入请求，提示(overwrite\_preferred列表)将通过从FTL传输到LBA scrambler的信息进行更新。通过引用overwrite\_preferred列表，将生成新的写入请求，用于在下一个擦除块中主动写入碎片页面。此外，未对齐的写入将创建碎片化的页面和额外的覆盖，如图1‑12所示。通过LBA scrambler的地址重映射，可以消除NAND闪存未对齐写的问题。

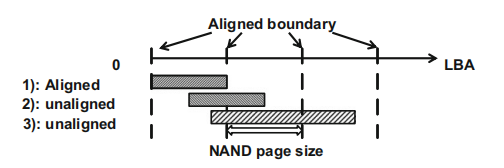


图1‑12 针对NAND闪存的对齐和非对齐写入

从评估结果来看，与没有LBA scrambler的SSD系统相比，使用LBA scrambler实现了最大的394%的写入性能改善，降低了56%的能耗和55%的耐力增强。

### 混合M-SCM/3D-NAND闪存SSD

SEA-SSD和LBA置码SSD均采用中间件和SSD控制器协同设计方法，通过减少SSDGC开销来提高所有3D-Nand闪存SSD的写入性能。然而，SSD的写性能的提高受到3D-NAND闪存的读写性能的限制。

另一方面，SCM比3D-NAND闪存更快、更节能、更耐用。它是非易失性，支持就地覆盖。因此，由于单片机的存在，内存和存储系统都处在一次旋转之中，如图1‑13所示。M-SCM同时用于内存和存储系统。

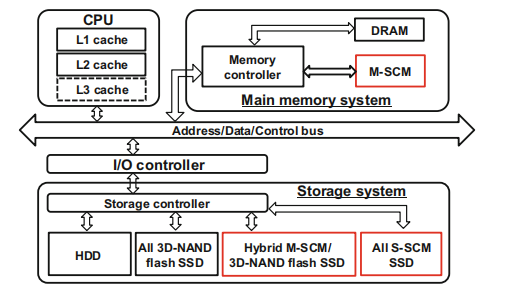


图1‑13 主内存和存储系统的革命，由于单片机

另外，S-SCM仅用于存储系统。通过将SCM引入SSD系统，提出了混合M-SCM/3D-NAND闪存SSD来提高所有3D-NAND闪存的写入性能。从SCM(ReRAM)设备的测量结果来看，写入成功的验证周期随写入/删除（W/E）周期而变化。因此，单片机采用就绪/忙状态的nand接口。如图1‑13所示，M-SCM被用作SSD的存储设备，而不是一个简单的高速缓存。针对混合M-SCM/3D-NAND闪光SSD的数据活动度和数据大小，开发了数据碎片抑制算法和冷数据驱逐算法。在SSD控制器中，使用最少的(LRU)表来记录页面数据访问历史记录。

如图1‑14所示，当页面数据的逻辑页面地址(LPA)到达LRU时，认为页面数据为热数据（频繁访问）。否则，页面数据将被判定为冷数据（很少被访问）。此外，根据页面利用率（数据大小除以页面大小），页面数据分为两种：随机（碎片）和顺序（未分段）。当页面数据大小超过一个阈值（θ）时，将将其视为顺序数据。混合SSD的数据存储策略是在M-SCM中存储热数据或随机数据，而在3D-NAND闪存中存储冷数据和顺序数据。热数据可以就地更新，随机数据可以在M-SCM中积累和顺序。混合M-SCM/3D-NAND闪存SSD的数据管理算法的算法流程图如图1‑15所示。当M-SCM几乎变满时，冷的和较少碎片化的M-SCM数据被驱逐到3D-NAND闪存。对于驱逐过程，判断数据的阈值是顺序的或随机的，是一个动态值，根据M-SCM中的数据存储状态递增/递减。当很难找到具有当前阈值的驱逐候选者时，会降低阈值以放宽限制。

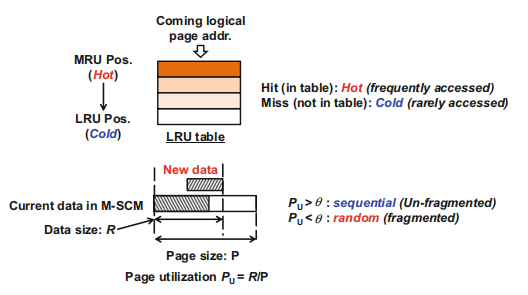


图1‑14 数据分类标准

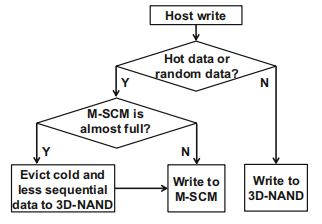


图1‑15混合M-SCM/NAND闪存SSD的数据管理算法

对于混合的M-SCM/3D-NAND闪存SSD，有两个重要的设计考虑因素。首先，了解M-SCM对代表性应用程序的容量和延迟要求。其次，了解3D-NAND组织对SSD写入性能的影响。在分析之前，SSD工作负载分为热和随机、热和顺序、冷、冷和随机、冷和顺序四类，如图1‑16所示。平均覆盖的一个大值，定义为总写数据大小除以用户数据大小，表明工作负载是热的，因为它包含许多热数据。此外，随机写请求的百分比决定了工作负载是随机的还是顺序的。在这里，一半的NAND闪存页面大小被用作判断页面是随机的或顺序的阈值。

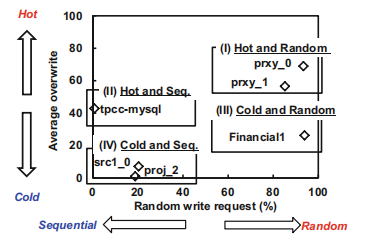


图1‑16 根据数据活动和大小进行SSD工作负载分类

从评估结果来看，增加M-SCM容量比增加3D-NAND闪存过度配置（额外容量）更有效地提高SSD的写性能。增加M-SCM容量和3D-NAND闪存过度配置都能够提高SSD的写入性能。然而，两者都没有增加M-SCM容量或3D-NAND闪存过度配置对于提高冷载和顺序工作负载的写性能是非常有效的。因此，在SSD中引入M-SCM最适合热和随机工作负载，但对于冷和连续工作负载成本不高。一般来说，对于M-SCM延迟固定为100ns/扇区的代表性工作负载，小于10%的M-SCM/3D-NAND闪存容量比就足够了。

另一方面，通过增加内存芯片设计的芯片面积，可以实现更快的速度。例如，可以通过增大内部写入单元来提高写入速度，并且可以通过添加选择设备来降低位线电容来提高读取速度。当M-SCM的速度提高时，混合的M-SCM/3D-NAND闪存SSD的最大吞吐量得到了提高。如图1‑17所示，当M-SCM的容量超过代理服务器应用程序(prxy\_0)的阈值时，混合SSD的写入性能达到饱和，这是一个热的、随机密集的。对于其他工作负载，在增加M-SCM/3D-NAND闪存容量比时没有饱和的趋势。从各种工作负载的评估结果来看，混合的M-SCM/3D-NAND闪存SSD的写入性能依赖于工作负载/应用程序。此外，更少的M-SCM容量需要更快的M-SCM速度来达到目标应用程序吞吐量。从系统的角度来看，在M-SCM的容量要求和M-SCM的速度之间存在一个权衡。

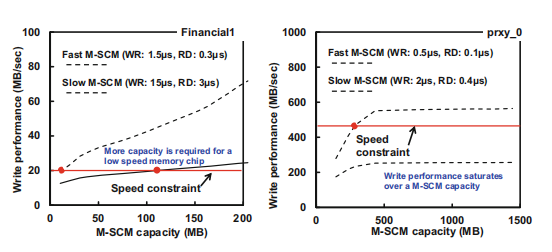


图1‑17 SSD的写入性能依赖于M-SCM的容量、延迟和应用程序

通过分析SSD的工作负载，混合SSD的最小M-SCM容量如图1‑18所示。它说明了累积的扇区写入频率与写入用户数据的地址范围之间的关系。例如，在20%的用户数据地址范围内的25%的访问频率意味着25%的访问发生在用户数据的前20%的地址上。曲线的转折点表示频繁访问的数据的结束，通常是随机数据，需要每秒高输入输出(IOPS)。曲线的高斜率值显示了最关键的数据，确定了混合SSD的最小M-SCM容量。对于某些工作负载，M-SCM的容量应该超过用户数据大小的40%，以覆盖75%的部门访问。然而，由于时间和空间位置，作为写缓存缓冲区的实际所需的SCM容量远远小于40%。该曲线的上升趋势与图1‑17中的结果一致。提高M-SCM容量可以有效地提高工作负载的混合SSD吞吐量。此外，由于曲线的斜率值小于prxy\_0和prxy\_1，因此增加SCM容量可以更有效地提高代理服务器应用程序(prxy\_0和prxy\_1)的性能。从图1‑18中可知，M-SCM的容量小于用户数据大小的20%，对于代理服务器应用程序就足够了。

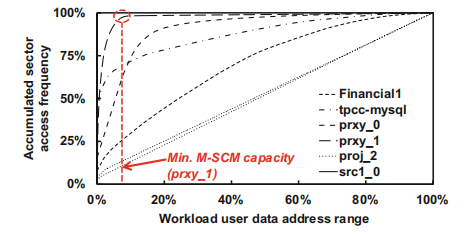


图1‑18 M-SCM的最低容量要求

由于传统的NAND闪存平面尺度面临着各种限制，使得降低制造成本和保证内存可靠性的难度越来越大，3D技术成为延续NAND闪存比特成本降低趋势的可行方法。 以下列举几个3D-NAND flash体系结构。例如，terabit单元阵列晶体管(TCAT)、管形比特成本可伸缩（P-BiCS）、垂直堆叠阵列晶体管(VSAT)，以及围绕浮动栅（DC-SF）的双控制栅。3D阵列有两种类型：垂直沟道和垂直栅。垂直沟道型阵列和垂直栅阵列的电流分别为垂直和水平流动。除了xy维度外，三维与非d还增加了垂直方向（z维度）上的位密度。在PiBCS3D-NAND闪存中，3D-NAND闪存的容量通过叠加更多的层来增加，这也弥补了由于不可扩展的BiCS孔的直径而导致的xy维细胞密度降低的问题。

对于3D-NAND闪存的设计，NAND的架构对电路的性能和成本至关重要。一组NAND闪存单元串联成一个NAND闪存string。共享同一衬底的多个NAND闪存string由一个NAND闪存块作为擦除单元组成。通过在衬底上施加高电压以将电子从块中的存储单元的浮栅极吸出，从而执行擦除操作。在该块中，与同一个字行连接的存储单元是一个页面。它是一个读/写单元。如图1‑19，随着微缩，页面和块大小呈增加趋势。NAND flash页面大小为8kb，块大小为2MB（一个块中256页）。随着3D-NAND闪存的出现，采用大的页面和块大小。以P-BiCS 3D-NAND为例，通过双片叠封技术，可以使NAND闪存的大小加倍。另一方面，如图1‑20所示，与采用小页面大小的设计相比，采用大页面大小的设计减少了NAND闪存芯片的字线解码器面积开销。然而，更大的页面或块大小并不会更适合实际应用程序的性能。

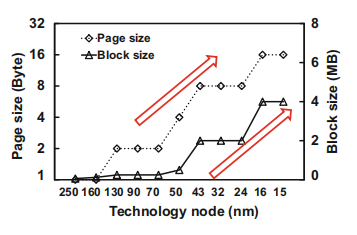


图1‑19 NAND闪存块和页面大小的变化趋势

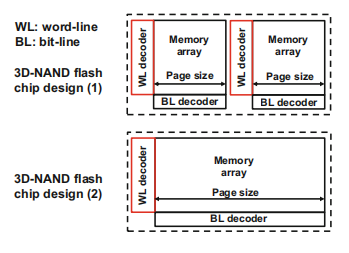


图1‑20 3D-NAND闪存的芯片设计

图1‑21为所有3D-NAND闪存的块大小和性能分析的评价结果。页面大小固定为16kb。具有一定容量的所有3D-NAND闪存SSD的写入性能具有最大值。太小或太大的块大小会降低写入性能。大的块可能会导致长的GC延迟，而小的块会降低擦除吞吐量。假设所有3D-NAND闪存SSD都可以容忍10%的写性能，那么对于具有25、50和100%过度配置的所有3D-NAND闪存，可接受的块大小分别为2、4和8MB。更高的所有3D-NAND闪存SSD容量可以接受更大的块大小。此外，当块大小超过一个阈值时，写性能达到饱和。即使块的大小高达16MB，也不会导致写性能下降，这对于3D-NAND闪存的应用是很好的。

所有3D-NAND闪存SSD的页面大小与性能分析如图1‑22所示。块的大小被固定为4Mb。超过大页面大小或小页面大小都会降低3D-NAND闪存SSD写性能。大的页面大小有利于顺序写吞吐量，但是页面覆盖数会很大（更多的页面覆盖开销）。此外，在页面大小越大的情况下，页面就越少。因此，GC的触发频率将更为频繁。相比之下，较小的页面大小会导致更少的页面覆盖开销，但它不利于顺序写。此外，在GC期间可能需要复制更多的页面。

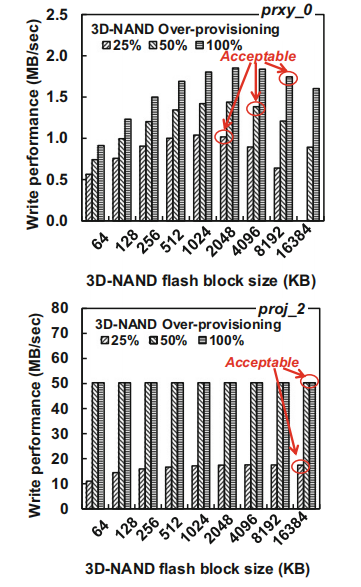


图1‑21所有3D-NAND闪存的块大小评估。2MB是典型的块大小

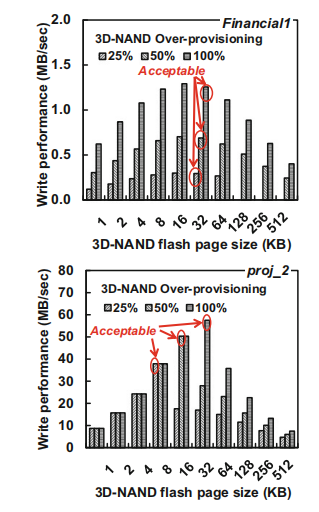


图1‑22 所有3D-NAND闪存的页面大小评估。8kB是典型的页面大小

所有3D-NAND闪存SSD和混合M-SCM/3D-NAND闪存SSD的3D-NAND闪存设计的比较结果见图1‑23。使用M-SCM，可接受的块和页面大小分别放大了4倍和64倍。与所有3D-NAND闪存SSD相比，混合SSD的3D-NAND闪存堆叠层可以增加四倍，而没有任何写入性能下降。

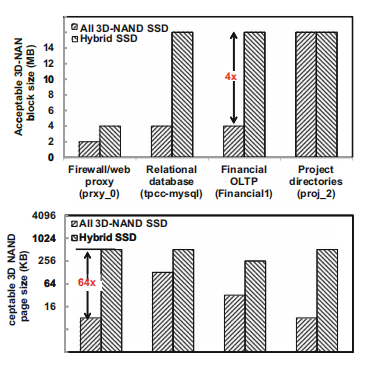


图1‑23 比较所有3D-NAND闪存SSD和混合M-SCM/3D-NANDSSD的可接受的3D-NAND闪存页面和块大小

### All S-SCM SSD

当单片机技术成熟，成本降低，以与NAND闪存竞争，所有S-SCM SSD成为一个可行的解决方案，以取代当前基于NAND闪存的SSD。在本节中，我们介绍了所有S-SCM SSD的磨损水平、S-SCM I/O数据切换率、S-SCM延迟设计。

对于像ReRAM这样的S-SCM设备，设备的耐写性是有限的。因此，S-SCM需要进行磨损均衡。一个简单的磨损均衡算法如图1‑24所示，它在页面级操作。因此，对于S-SCM的每一页都保持一个磨损均衡触发阈值δ。此外，为了监视每个扇区（512字节）的耐写性时间，即每个扇区的写入/擦除(W/E)次数。当页面i中的扇区的最大W/E周期小于δpage(i)时，将执行就位页面覆盖。否则，将触发磨损均衡。在磨损均衡过程中，读出旧页面i中的数据，与新数据合并，并写入新页面j。之后，将第i页的磨损水平触发阈值更新为恒定的窗口阈值σ，以提高磨损水平触发的阈值。实际上，在工作负载中有很多热点(一些地址经常被写）。通过磨损均衡， 所有S-SCM SSD的耐力可以大大提高。例如，使用σ=5，没有磨损均衡算法的S-SCM扇区的最大W/E周期比使用有磨损均衡算法的所有S-SCM SSD高3000倍以上。

通过调整σ的值，可以控制磨损均衡触发间隔。一个小的σ很容易触发磨损均衡，使所有的页面磨损均匀。但是，由于额外的页面复制操作，这种配置会降低所有S-SCM SSD的性能。因此，可以调整σ来平衡SSD的性能和耐写性。

图1‑25显示了all S-SCM SSD写入性能都依赖于I/O数据切换速率、S-SCM延迟（假设有一些读写延迟）和应用。M-SCM延迟设置为100ns/扇区。与基于3D-NAND闪存的SSD不同，写性能很少依赖于应用。这个趋势也是一样的。轻微的差异是由于S-SCM磨损量平和总写入数据大小。当S-SCM速度较快时，应采用较高的数据切换速率来充分利用all S-SCM SSD的写入性能。

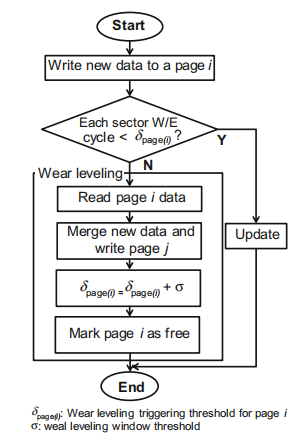


图1‑24 所有S-SCM SSD的磨损均衡算法流程图

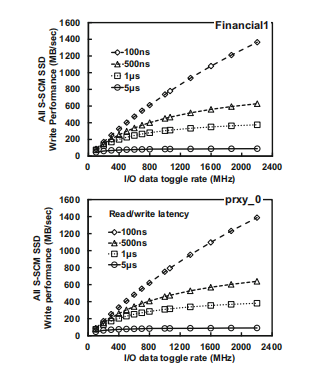


图1‑25 All S-SCM SSD写入性能都依赖于I/O数据切换速率、S-SCM延迟和应用

通过保持S-SCM延迟为1μs， 图1‑26 a比较了all S-SCM SSD和混合M-SCM/3D-NAND闪存SSD的速度，说明了25%固定M-SCM/3D-NAND闪存比下I/O数据切换速率的转折点。以tpcc-mysql工作负载为例，超过500MHz的I/O数据切换速率使所有S-SCMSSD比混合M-SCM/3D-NAND闪存SSD更快。另一方面，在固定的I/O数据切换速率为1066MHz的情况下，可以分析S-SCM延迟的转折点，如图1‑26 b所示。更快的S-SCM设备创建了一个更快的all S-SCM SSD。在某些工作负载中，如果S-SCM延迟超过5μs，混合M-SCM-3D-NAND闪存SSD拥有比all S-SCM SSD具有更快的速度。

此外，通过假设每个存储器设备的比特成本，可以根据SSD内部存储器的容量配置轻松地比较每个SSD的成本.

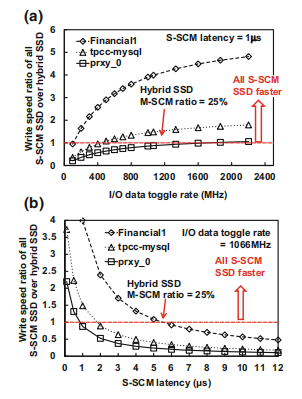


图1‑26 All S-SCM SSD和混合M-SCM/3D-NAND闪存SSD的写入速度比较