闪存基础

结构

NAND 闪存是分层组织的。图 1 说明了 3D NAND 闪存芯片的组织结构。多个（例如 24 至 176 个）闪存单元（图 1(a)）垂直堆叠并形成连接至位线 (BL) 的 NAND 串（图 1(b)）。不同BL的NAND串组成一个子块。子块中同一垂直位置的每个单元的控制栅极连接到同一字线(WL)，这使得同一WL上的所有单元同时操作。一个块由几个（例如，4到8个）子块组成，数千个（例如，3,776[37]）块构成一个平面。 NAND 闪存芯片包含多个芯片（图 1(c)），每个芯片包含多个平面（例如，每个芯片有两个或四个平面 [32]）。 NAND 闪存芯片中的芯片可以彼此独立运行，而芯片中的平面可以在有限的条件下同时运行，因为它们通常共享相同的行解码器。

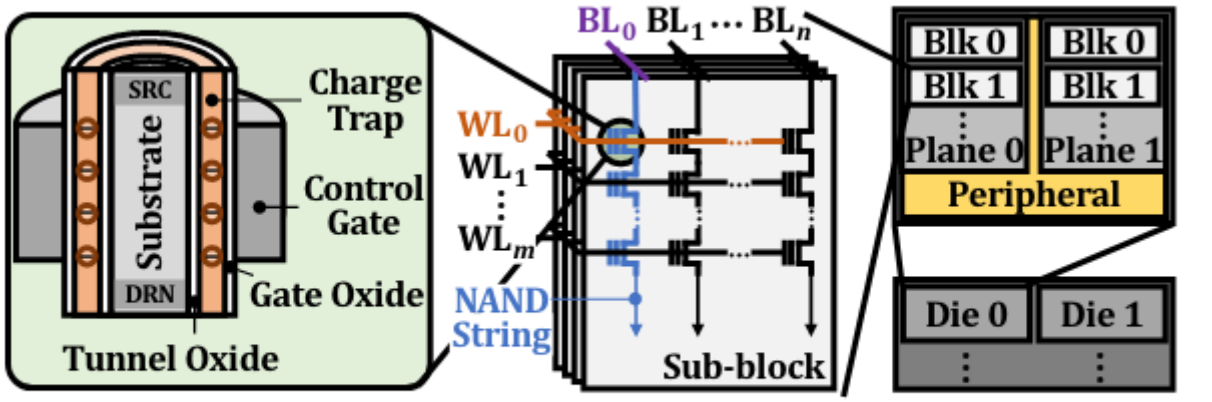


Figure 1: Organization of 3D NAND flash memory.

闪存单元介绍

传统的SLC或单级单元存储通过在闪存单元的浮栅（FG，Floating Gate）上不带电或带电来区分“1”和“0”。通过增加电荷数或电压阈值(Vt)水平，每个单元可以存储超过1位信息。通过将Vt电平数量增加到代表11、10、01和00的4位，可以启用每单元2位（MLC）存储。同样，通过将电压阈值水平数量增加到8位和16位，可以启用每个单元3位（TLC）和4位（QLC）的存储。多层单元存储的好处是可以在不增加工艺复杂性的情况下增加存储容量。用于生产SLC产品的硅片的相同fab设备也可以用于生产MLC，TLC和QLC的器件。然而，多级存储单元需要准确定位电压阈值，确保电荷分布不重叠，并准确感知不同电荷水平。随着Vt分级数量的增加，精确编程和传感所需的时间也会增加。我们需要额外的电路和编程算法以改善设备的性能和耐久性下降的问题。从SLC过渡到MLC，TLC乃至QLC，相当于在没有额外资本投资的情况下将设备内存进行扩展。

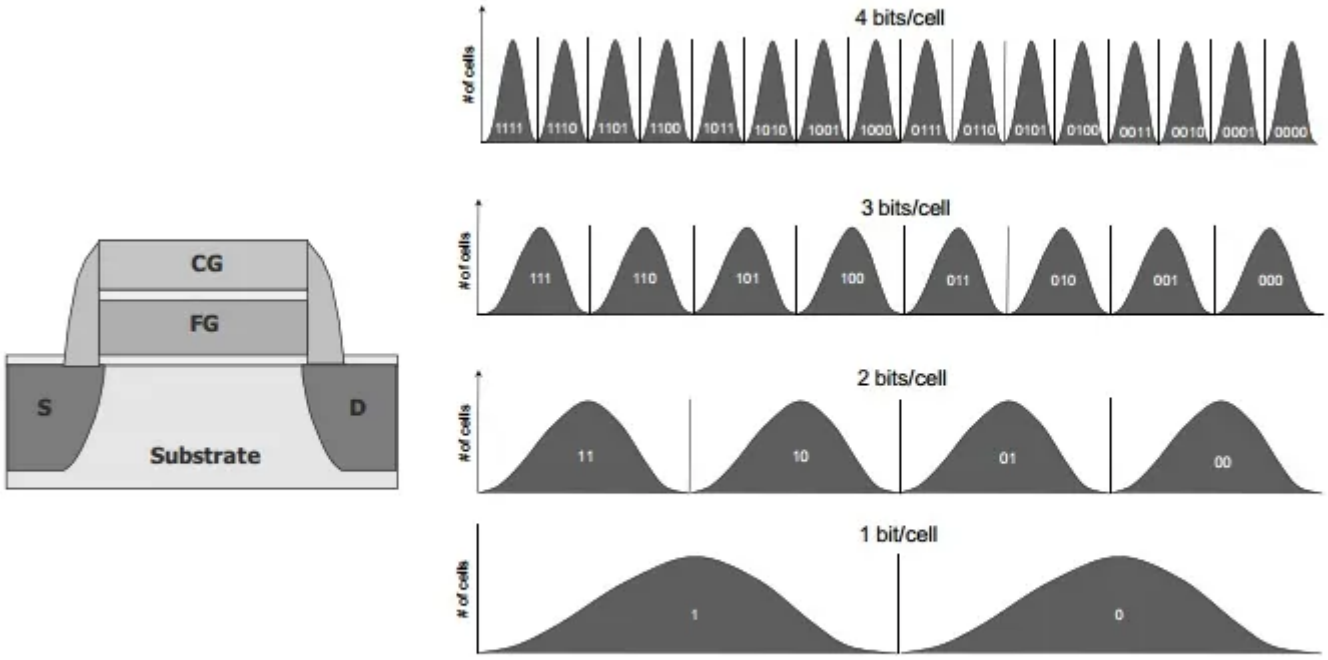


Fig. 1.7. Multi-level storage in floating gate NAND Flash memory

阵列（Array）

为了最大化提高硅片的使用率，存储单元被堆积成一个矩阵。根据存储单元在矩阵中的排列方式，我们可以区分NAND闪存和NOR闪存。在NAND串行中，存储单元以32个或64个为一组进行串联，如图2.2所示。两个选择晶体管被放置在行边缘，以确保与源线(通过Msl)和位线(通过Mdl)的连接。每个NAND行与另一个行共享位行联系。控制门通过字线(wordlines, WLs)连接。

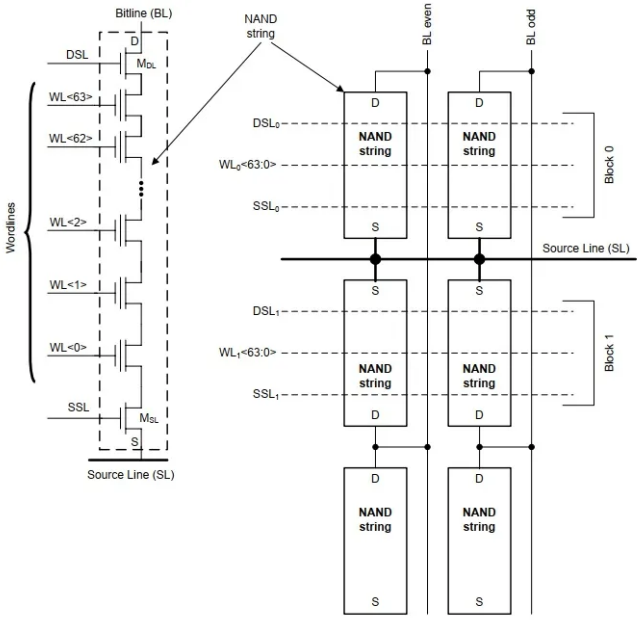


Fig. 2.2. NAND string (left) and NAND array (right)

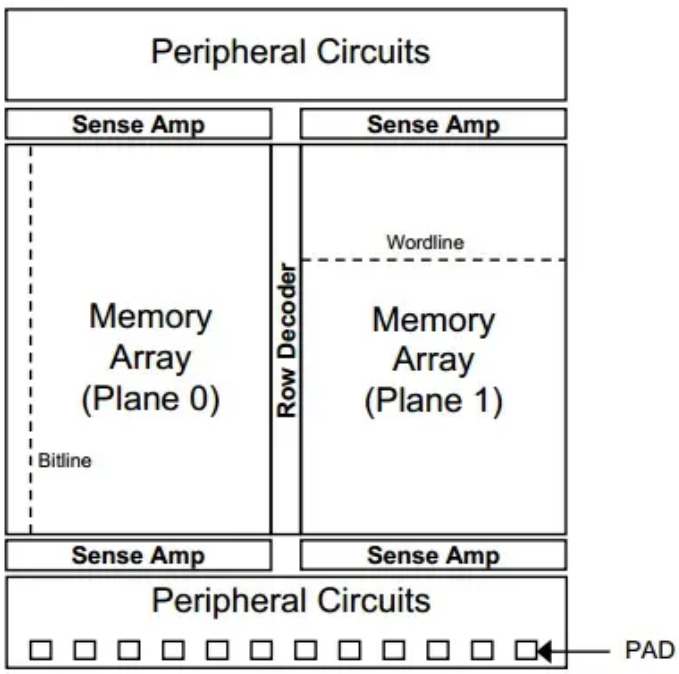


Fig. 2.3. NAND Flash memory floorplan

闪存设备主要由存储阵列组成。但是为了执行读取、编程和擦除操作，我们需要额外的电路。由于封装NAND芯片需要一个明确的尺寸，因此在早期设计阶段组织好所有电路和阵列十分重要，也就是说， 确定一个平面图是很重要的。图2.3给出了一个平面布置图的例子。存储阵列可以被分割成不同的平面(图2.3中的两个平面)。在水平方向上突出显示Wordline，而在垂直方向上显示Bitline。

行译码器（Row Decoder）位于两个平面之间:该电路的任务是使所有属于所选NAND字符串(Sect.2.2.2)的字行适当偏置。所有位线都连接到检测放大器(Sense Amp.)。每个感知放大器可以有一个或多个位行;感测放大器的目的是将存储器切利所沉没的电流转换成数字值。

基本操作

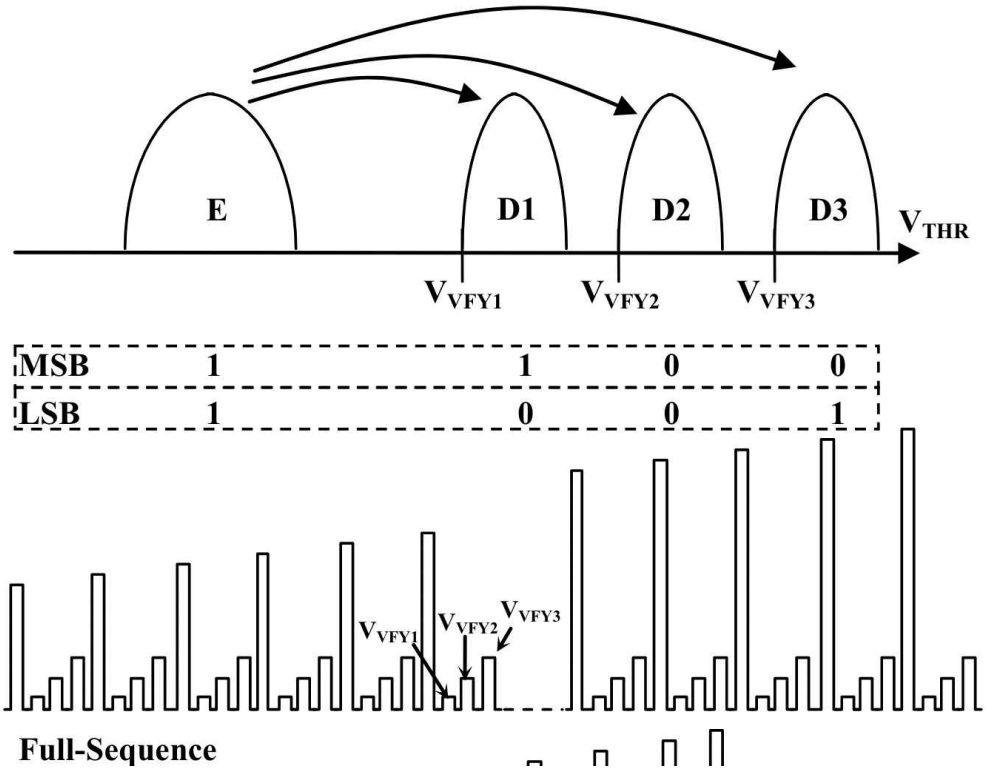
写

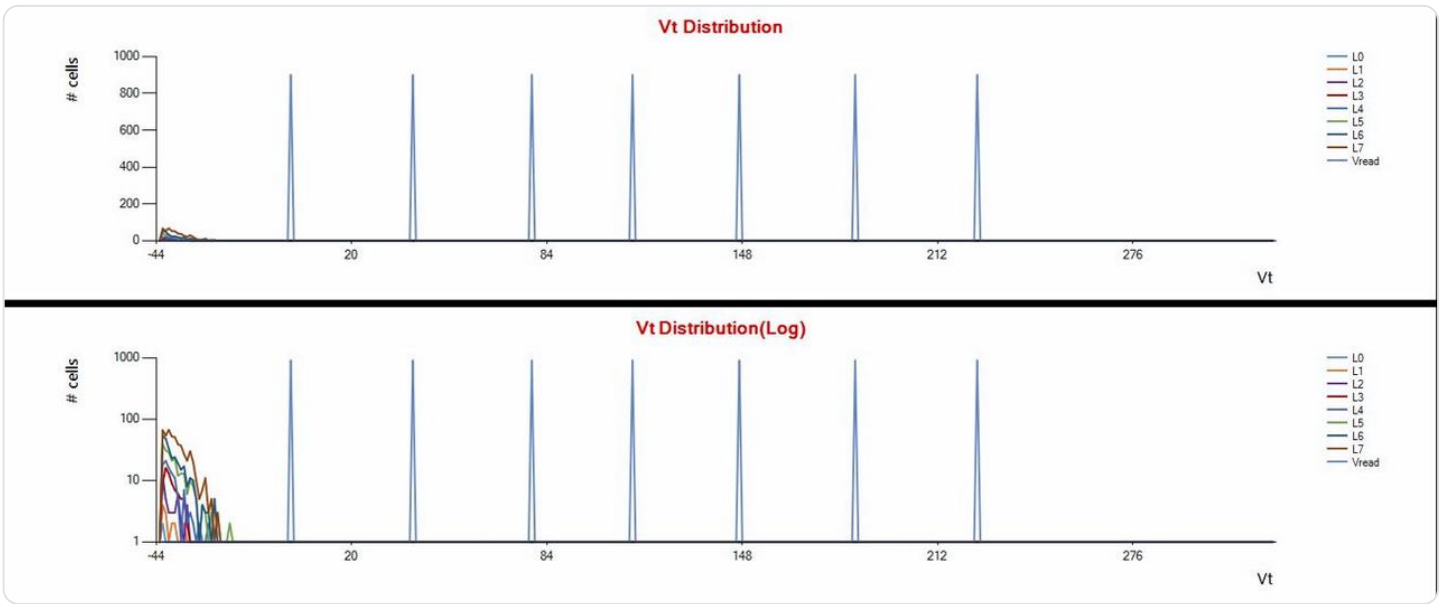
编程操作通过向 WL 施加高电压 (> 20 V) 将电子从衬底注入单元的电荷陷阱，这会增加单元的 VTH 电平（即，编程操作只能将单元的数据从“1”更改为“ 0' 假设采用上述 SLC 编码）。由于一组闪存单元连接到 NAND 闪存中的单个 WL（即，将相同的电压施加到同一 WL 中每个单元的控制栅极），数据以页粒度（例如 16 KiB）写入，这样同一 WL 中的每个单元都存储该页的一位。

TLC（一次编程）

https://www.flashmemorysummit.com/English/Collaterals/Proceedings/2018/20180808_FTEC-202-1_Ye.pdf

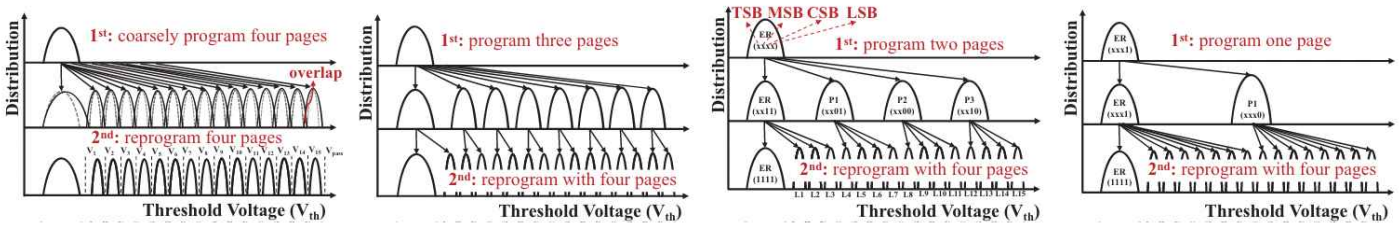
闪存编程是通过向 WL 施加高电压 (> 20 V)将 WL 上的闪存单元充电至与 BL 的输入bit value匹配的电压电平的过程。匹配由编码位和电压电平之间的预定义编码表确定。编程过程通常采用交互式编程验证策略，该策略在每个器件编程脉冲之后验证单元电压，如果单元达到其期望的电压电平，则将其排除在随后的编程脉冲之外。





QLC（两步编程）

由于编程效率高，3D闪存传统上对大型技术节点的低密度单元采用一步编程（OSP）。随着技术的快速进步，**OSP 无法确保 QLC 和每单元更多位闪存的编程可靠性**。具体地，当对WLn+1的存储单元进行编程时，WLn的存储单元的电压很可能受到干扰。为了应对这一挑战，业界恢复了最初为 2D MLC 闪存提出并广泛采用的两步编程 (TSP) 算法来对 3D 闪存单元进行编程。



(a) TSP(16,16) (b) TSP(8,16) (c) TSP(4,16) (d) TSP(2,16)

Fig. 2. Two-step programming (§II-B).

TSP 分两步对闪存字线进行编程。图2(a)显示了基本的TSP(16,16)。首先，对字线（例如WLn）进行粗略编程，使得每个单元处于中间状态。两个相邻状态可能具有电压重叠，这表明读取这样的字线往往会产生错误。在下一个字线WLn+1的第一编程步骤之后，控制器进一步编程WLn，每个单元都被编程为其最终状态。使用TSP编程的字线比使用OSP编程的字线具有更大的噪声容限，从而保证更好的数据可靠性。然而，TSP(16,16)的一个主要问题是它的编程延迟，平均是OSP的1.53倍。**为了减少TSP 造成的编程延迟**，提出了几种优化方法来提高第一步的编程速度。例如，2019 年，Shibata 等人。东芝提出的 TSP(8,16)（图 2(b)）在第一步将三个页面编程为 8 级状态 [60]。 TSP(8,16) 比 TSP(16,16) 减少了 18% 的编程延迟。 2021 年，Khakifirooz 等人。 Intel 提出的 TSP(4,16)（图 2(c)）[26] 在第一步对两个页面进行编程。与TSP(8,16)相比，TSP(4,16)可以将第一步的编程延迟减少60%，从而将整体写入延迟减少约22%[26][25]。图2(d)示出了在第一步对一页进行编程的TSP(2,16)，这可能是最快的TSP。然而，TSP(2,16)是不现实的，因为将单元从2级状态编程到16级状态将引入大的字线间干扰，从而无法保证可靠性。研究表明，即使采用 TSP，字线间干扰仍然很严重，并且当实现更多堆叠层时，闪存制造商正在开发不同的程序干扰减少技术 [11]。研究表明，部分编程的闪存单元比完全编程的单元更容易受到单元间干扰和读取干扰的影响 [4]。因此，本文其余部分不考虑 TSP(2,16)。由于TSP(16,16)第一步已经完成了4页16级状态的编程，因此第二步只需要微调每个状态的电压范围即可。综上所述，在 TSP(4,16)、TSP(8,16) 和 TSP(16,16) 三种 TSP 方案中，我们发现 **TSP(4,16) 具有最佳编程性能，而 TSP(16,16)) 具有最佳的数据可靠性**。

读

NAND闪存通过识别对应的BL是否有电流流过来确定单元的数据（即单元的VTH电平）。读取操作被设计为寻址存储器单元并推断其中存储的信息。在 NAND 型闪存的情况下，存储单元以 2^k 单元为一组（串）串联连接。在图 8.1 中，说明了 NAND 闪存的串：MBLS 和 MSLS NMOS选择器晶体管分别将串连接到位线BL和源极线SL。与其他类型的闪存一样，存储的信息与单元的阈值电压 VTH 相关：图 8.2 中显示了包含一个逻辑位的单元的阈值电压分布。如果单元具有属于擦除分布的VTH，则其包含逻辑“1”，否则，如果其属于写入分布，则其包含逻辑“0”。包含 n 位信息的单元有 2^n 个不同级别的 VTH。

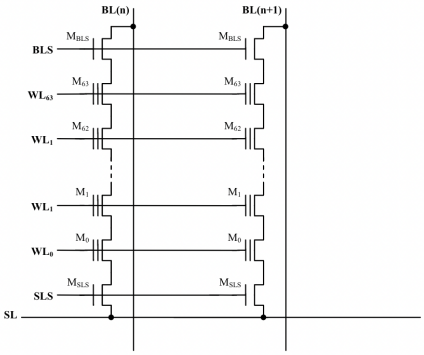


Fig. 8.1. Two strings in a NAND architecture

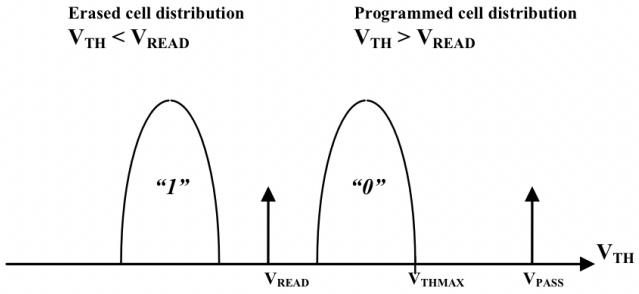


Fig. 8.2. Threshold voltage distributions for erased and programmed cells

闪存单元的工作方式与普通 MOS 晶体管类似。给定固定栅极电压，单元电流是其阈值电压的函数。因此，**通过电流测量，可以了解存储单元属于哪个VTH分布**。存储单元属于由其他单元组成的串这一事实存在一些问题。首先，未选择的存储器单元必须以一定方式偏置，使得未选择的单元必须充当传输晶体管。因此，它们的栅极必须驱动至高于最大可能 VTH 的电压（通常称为 VPASS）。在图 8.2 中，VPASS 必须高于 VTHMAX。

根据编码策略，从 字线读取闪存页面需要施加多个参考电压，以便页面读取延迟 (tR) 与所施加的参考电压的数量呈线性关系。如公式 1 所示。这里，tSENSE 表示单个电压感测延迟，nSEN SE 表示施加的参考电压的数量。由于需要区分的级别越多，因此需要更多的参考电压，因此具有更多级别的读取性能往往会降低。这种情况下，请求的读延迟会明显放大。tPRE、tEVAL 和 tDISCH 是分别定义预充电、评估和放电阶段延迟的时序参数。制造商仔细决定三个时序参数以确保正确运行。例如，如果 tPRE 太短而无法对 BL 和 CSO 充满电，则即使目标单元已编程，VSO 在评估阶段也可能低于 VSR。tDISCH 太短还可能导致一些 BL 部分充电，从而导致原始位错误。由于与所有 BL 完全放电相比，当有一些 BL 部分充电时需要更多时间来稳定所有 BL，因此下一个预充电阶段可能无法在 tPRE 延迟内将所有 BL 正确设置为 VPRE。

$$tR = nSENSE * tSENSE \quad (1)$$

$$tSENSE = tPRE + tEVAL + tDISCH$$

公式 1 中的延迟适用于一页读取。为了纠正错误以增强可靠性，基于低密度奇偶校验码（LDPC）的 ECC 需要发出额外的读取操作（读取重试），这会应用几个额外的感测电压以获得更准确的信息。读取重试操作会显著延长读取延迟，如公式 2 所示。nRETRY 是重试次数，tDMA 是数据传输成本，tLDPC 是解码成本。在实际应用中，如果读重试次数达到阈值，就应该重写所访问的数据，以保证数据的完整性。

$$tREAD =(1+nRETRY) * (tR + tDMA + tLDPC) \quad (2)$$

综上所述，读取闪存页面的延迟取决于读取次数和读取重试次数以及施加的参考电压次数在每次读取或读取重试中。下面我们详细阐述信息编码和读取延迟之间的关系。

3D NAND闪存的格雷码

TABLE II
THE 16 VOLTAGE LEVELS OF GRAY-CODE FOR QLC FLASH.

Gray-code		ER	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13	L14	L15	$nSENSE$
GC(1,2,4,8)	LSB	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1
	CSB	1	1	1	1	0	0	0	0	0	0	0	0	1	1	1	1	2
	MSB	1	1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	4
	TSB	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	8
GC(1,2,6,6)	LSB	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1
	CSB	1	1	1	1	0	0	0	0	0	0	0	0	1	1	1	1	2
	MSB	1	1	0	0	0	1	1	0	0	0	1	1	1	0	0	1	6
	TSB	1	0	0	1	1	1	0	0	0	1	1	0	0	0	1	1	6
GC(1,4,5,5)	LSB	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1
	CSB	1	1	1	0	0	0	0	1	1	0	0	0	0	1	1	1	4
	MSB	1	1	0	0	1	1	0	0	0	0	0	1	1	1	1	0	5
	TSB	1	0	0	0	0	1	1	1	1	1	0	0	1	1	0	0	5
GC(3,4,4,4)	LSB	1	1	0	0	0	1	1	1	1	1	1	0	0	0	0	0	3
	CSB	1	0	0	0	0	0	0	1	1	1	0	0	1	1	1	1	4
	MSB	1	1	1	0	0	0	0	0	0	1	1	1	1	0	0	1	4
	TSB	1	1	1	1	0	0	1	1	0	0	0	0	0	0	1	1	4

高密度闪存通常采用格雷码对一个单元中的多个位进行编码，将多个位值分配给不同的电压电平[17]。表 II 列出了 QLC 闪存的四种常用格雷码 — GC(1,2,4,8)、GC(1,2,6,6)、GC(1,4,5,5) 和 GC(3,4,4,4) 。这些代码以不同方式将 4 位值映射到 16 个电压电平。正如我们在第 1 节中列出的那样，一个单元中的四位分别称为 LSB、CSB、MSB 和 TSB。特定 TSP 的格雷码选择存在限制。特别地，给定 TSP(2^m,2^n) 和 GC(A,B,C,D)，如果 A+B<4(m=2) 或 A+B+C<8(m=3)，我们可以采用相应的 TSP 对数据进行编码并对闪存单元进行编程。显然，所有格雷码都与 TSP(16,16) 兼容。 GC(1,2,4,8)是唯一与TSP(8,16)兼容的，而GC(1,2,6,6)和GC(1,2,4,8)都与TSP(4,16)兼容。2）格雷码的性能和可靠性：采用不同的格雷码表现出不同的读取和编程性能特征。例如，要从 GC(1,2,6,6) 编码的单元中读取 LSB 页，我们只需施加一个参考电压来区分低于或高于 REFL7/L8 的单元电压（L7 和 L8 之间的参考电压）。类似地，我们应用两个参考电压REFL3/L4和REFL11/L12来读取CSB页。读取 TSB 页要慢得多，因为它需要 6 个参考电压电平。因此，GC(1,2,4,8)、GC(1,2,6,6) 和 GC(1,4,5,5) 都称为 UGC（不平衡格雷码）。相反，GC(3,4,4,4) 被称为 BGC（平衡格雷码），因为读取不同的页面具有大致相同的性能。由于读取重试的可靠性会降低，总体读取性能也取决于读取重试的次数，如公式 2 所示。

TABLE III
COMPARISON AMONG VARIOUS GRAY-CODES.

Gray-code	TSP algorithm	Read.perf	Write.perf	Reliability
GC(1,2,4,8)	TSP(16,16), TSP(8,16), **TSP(4,16)**	low	**high**	low
GC(1,2,6,6)	TSP(16,16), **TSP(4,16)**	middle	**high**	low
GC(1,4,5,5)	**TSP(16,16)**	middle	low	**high**
GC(3,4,4,4)	**TSP(16,16)**	**high**	low	**high**

格雷码的性能和可靠性：采用不同的格雷码表现出不同的读取和编程性能特征。例如，要从 GC(1,2,6,6) 编码的单元中读取 LSB 页，我们只需施加一个参考电压来区分低于或高于 REFL7/L8 的单元电压（L7 和 L8 之间的参考电压）。类似地，我们应用两个参考电压REFL3/L4和REFL11/L12来读取CSB页。读取 TSB 页要慢得多，因为它需要 6 个参考电压电平。因此，GC(1,2,4,8)、GC(1,2,6,6) 和 GC(1,4,5,5) 都称为 UGC（不平衡格雷码）。相反，GC(3,4,4,4) 被称为 BGC（平衡格雷码），因为读

取不同的页面具有大致相同的性能。由于读取重试的可靠性会降低，总体读取性能也取决于读取重试的次数，如公式 2 所示。

table III 比较了使用最佳兼容 TSP 时不同格雷码及其支持的 TSP 的读写性能。每个格雷码的最佳 TSP 在表中以粗体显示。研究表明，BGC 在可靠性方面更加稳健，因为 QLC 闪存单元的四位具有相似的参考电压，并且在可靠性下降的情况下，它比 UGC 需要更少的读取重试。

HPCA23的工作中闪存读取需要增强以适应多种格雷码。控制器利用每个格雷码的编码表来确定参考电压电平的数量以及每个电平的幅度。为了降低硬件复杂性，MGC 在块级别启用，即一个块对该块中的所有页面采用相同的格雷码及其 TSP。一旦块被擦除，它可以采用不同的格雷码及其TSP。请注意，我们在页面级别跟踪应用程序使用特征，正如我们在下一节中详细说明的那样。鉴于采用了两种格雷码，我们使用位图（称为格雷码位图（GCB））来跟踪为每个块选择的格雷码。位图足够小，可以缓存在 DRAM 中，例如，对于块大小为 16MB 的 1TB SSD，位图为 8KB。如果有两个以上的格雷码，则每个块可能需要更多位。 MGC 还可以在不同级别（例如平面和芯片）启用。我们把它留给我们未来的工作。

读过程

基于3D NAND闪存的SSD的读取与基于平面闪存的SSD类似，包括以下四个步骤：（1）向闪存芯片发出读取命令； (2)施加一个或多个参考电压来感测所访问的字线； (3)将感测到的信息传送至闪存控制器； (4)通过控制器内部的ECC模块对信息进行纠正。

在预充电步骤(图2左侧部分中的P)中，NAND闪存芯片通过使能预充电晶体管MPRE 1 将所有目标BL及其读出电容器(CSO)充电至预充电电压VPRE。同时，芯片将读取参考电压VREF施加到目标WL，同时将大得多的通过电压VPASS施加到同一块中的其他WL。如果 VTH≤VREF（图 2 左侧的 a），则目标单元将作为电阻器运行；如果 VTH>VREF (b)，则目标单元将作为开路开关运行；同一 NAND 串中的所有非目标单元将始终作为电阻器运行，因为 VPASS 足够高（> 6 V）。

然后，芯片通过将BL与VPRE 3断开并启用锁存电路来开始对目标单元进行评估。如果目标单元的 VTH 电平低于 VREF，CSO 中的电荷会快速流过 NAND 串 (c)，被检测为“1”。如果 VTH>VREF，则 CSO 的电容几乎不会变化 (d)，因为目标单元阻止 BL 放电电流，该电流被感测为“0”。

最后，芯片对 BL进行放电，使 NAND 串返回到其初始稳定状态（即可以进行预充电之前的状态）以供将来操作。

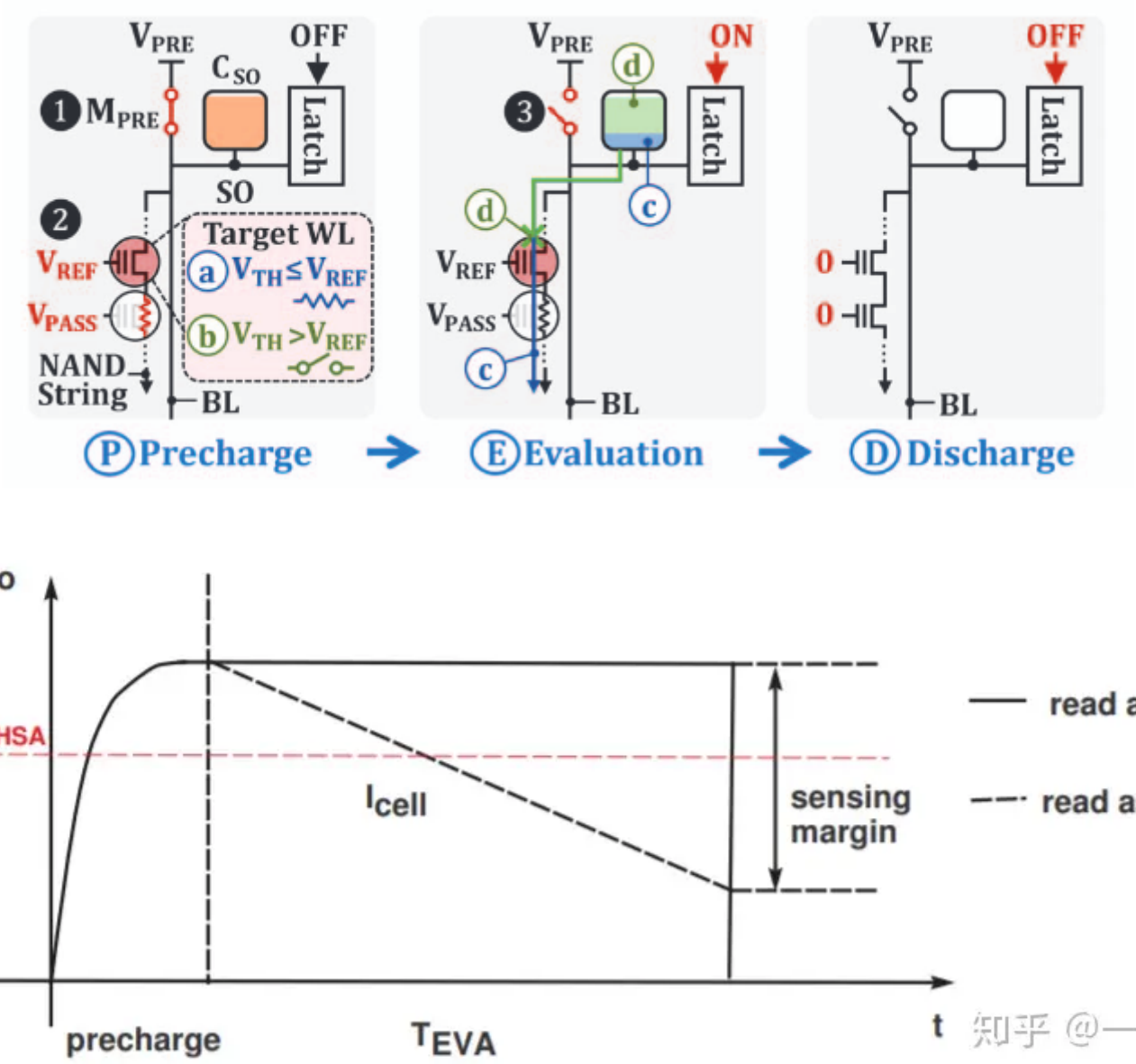
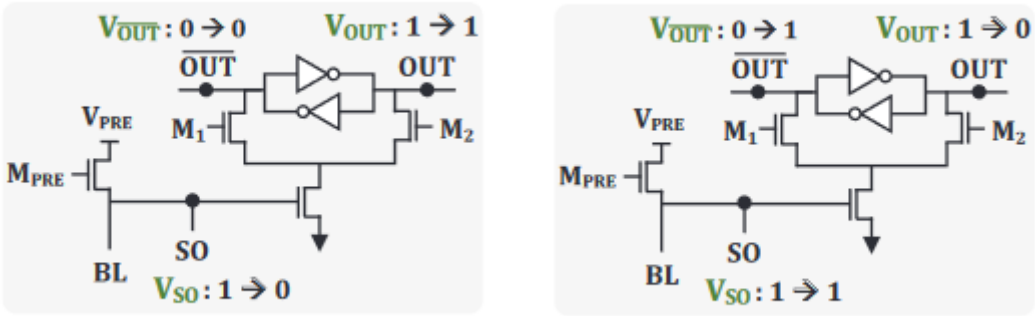


Fig. 3: ABL read operation timing diagram.

1. Precharge, 给SO(sense out)电容充电至VDD(没有SO capacitor的时候可能是将整个BL看成一个电容来充电，i.e., BL precharge)
2. 关闭MPCH，导通的BL中流过一个由SO电容放电产生的恒定电流Icell(evaluation phase).
3. 经过Teva秒后, 读取SO电容电压并与参考电压(Vthsa)比较，Vsensing>Vthsa即为"1"，Vsensing<Vthsa即为"0"，读的结果通过latch输出。

下图描述了 NAND 闪存如何通过其锁存电路感测 BL 的电导。图(a)和(b)分别描述了当闪存单元的阈值电压VTH低于和高于读取参考电压VREF时锁存电路的操作。我们展示了从预充电步骤到评估步骤时三个节点 (SO/OUT/OUT’) 中每个节点的电压状态转变。预充电时，NAND闪存芯片对BL充电，使VSO=1。在评估步骤之前，芯片通过仅激活晶体管 M1 来初始化锁存电路，导致 VOUT ‘=0，从而 VOUT=1。评估步骤禁用 MPRE 和 M1，同时启用 M2。在图 (a) 中，评估步骤使 VSO=0，因为 CSO 中的电荷快速流过 NAND 串（因为 VTH≤VREF），这导致 VOUT=1。由于 CSO 的电荷保持率较低，闪存单元的位值立即存储在锁存电路中。在图 (b) 中，评估步骤导致 VSO=1 且 VOUT=0，因为当 VTH>VREF 时闪存单元作为开路开关运行。



(a) $V_{TH} \le V_{REF}$ (b) $V_{TH} > V_{REF}$

MLC 外围电路

下图描述了 NAND 闪存如何通过其锁存电路感测 BL 的电导。读出放大器SA从 SRO 获取数据，并提供与输出上的下部页或上部页相关的位。图 10.9 显示了可用于读取图 10.1 中描绘的分布的电路。仅描述与读操作相关的元素。在该示意图中，两个静态锁存器L1和L2能够通过启用NMOS M1和M2来存储节点SO的值。通常，单个锁存器提供输出数据：在图10.9所示的情况下，它是L2。

Upper page 1	1	0	0
Lower page 1	0	0	1

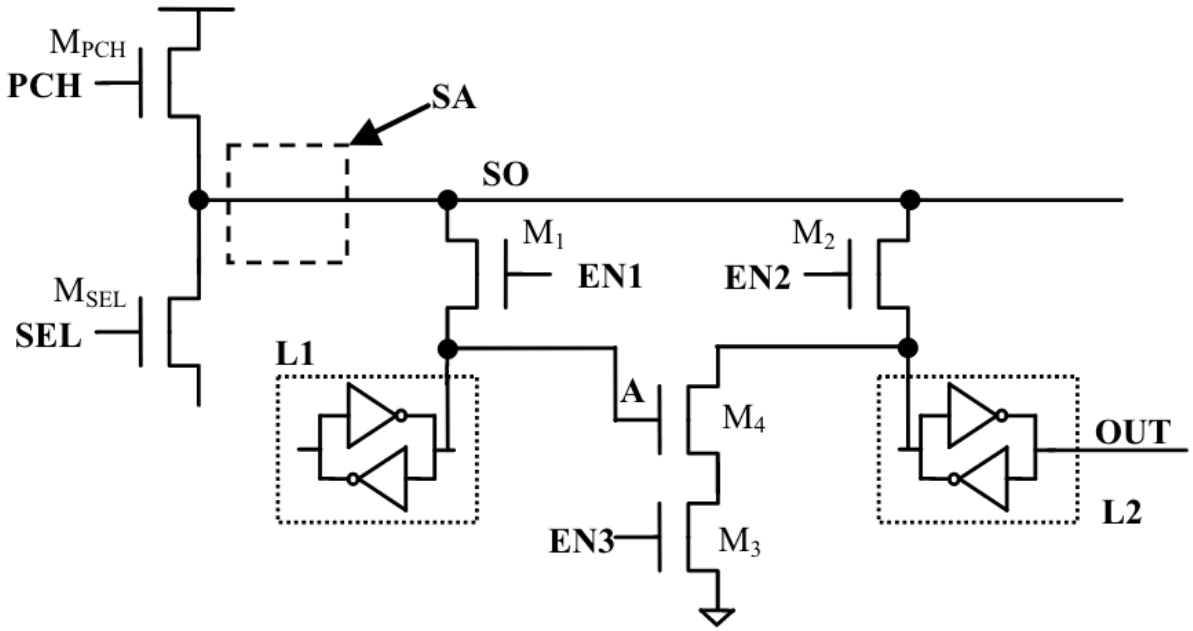


Fig. 10.9. Basic MLC sensing circuit

在读Upper page时施加一次Vread2即可。若Vthr属于E或者D1分布，SO逻辑值为0。若Vthr属于D2或者D3分布，SO逻辑值为1。通过使能M2，SO逻辑数据被锁存到L2中，并且经过反相器，当输入SO为0011（分别对应于E、D1、D2、D3）时，OUT分别为1100。

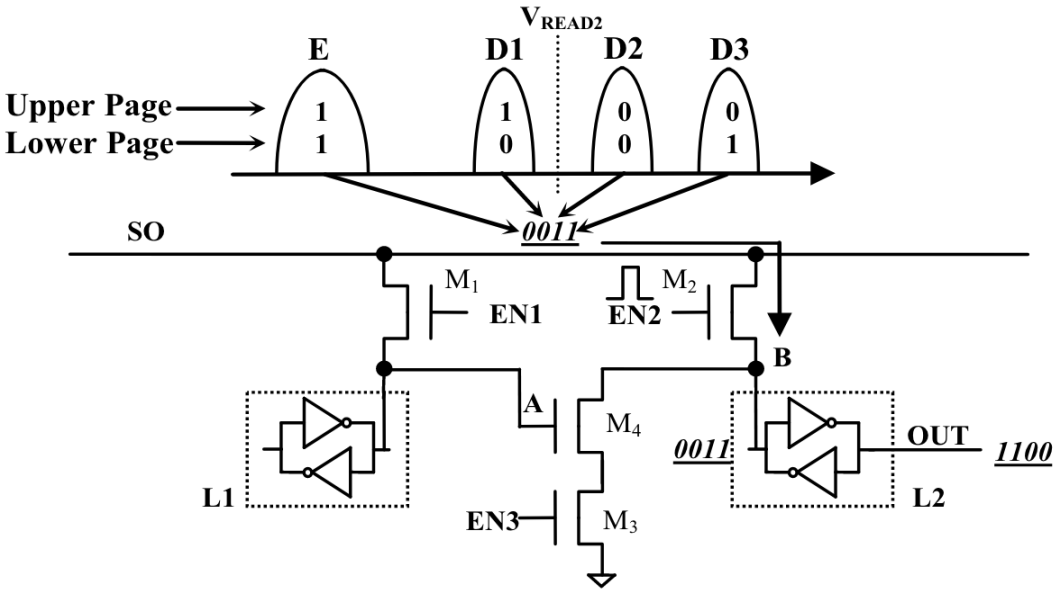
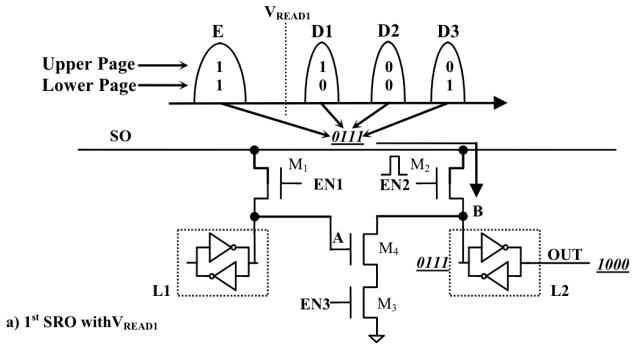
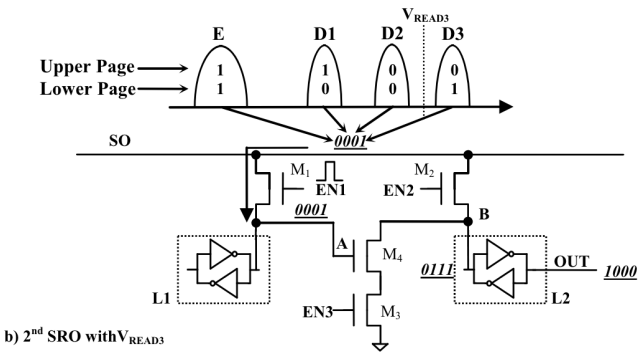


Fig. 10.10. MLC upper page read

在读lower page时，需要施加两次电压，分别是Vread1和Vread3。第一次施加Vread1，电压处于E、D1、D2、D3时，SO分别为0111。通过使能M2，SO逻辑数据被锁存到L2中，输出OUT分别为1000，B=0111。

然后断开M2，施加Vread3，电压处于E、D1、D2、D3时，SO分别为0001。通过使能M1，SO逻辑数据被锁存到L1中，结果A分别为0001。使能M3，只有当A为1时，M4导通，B接地，B=0。所以对应E、D1、D2、D3，B为0110，OUT为1001。





MLC with cache read

但是上述的外围电路不能管理高速缓存管理，L1和L2都用于参与Lower page的读。下面的结构可以。在这种情况下，锁存器 L1 忙于执行读操作，而锁存器 L2 执行数据输出，这意味着即使该结构也属于缓存读类别。

最开始SET=1，M5打开，OUT接地，OUT=0，B=1；SO=1，SO接地，输入为0，EN1=1，M1打开，C=0，A=1；EN3=0，所以A不影响OUT结果。

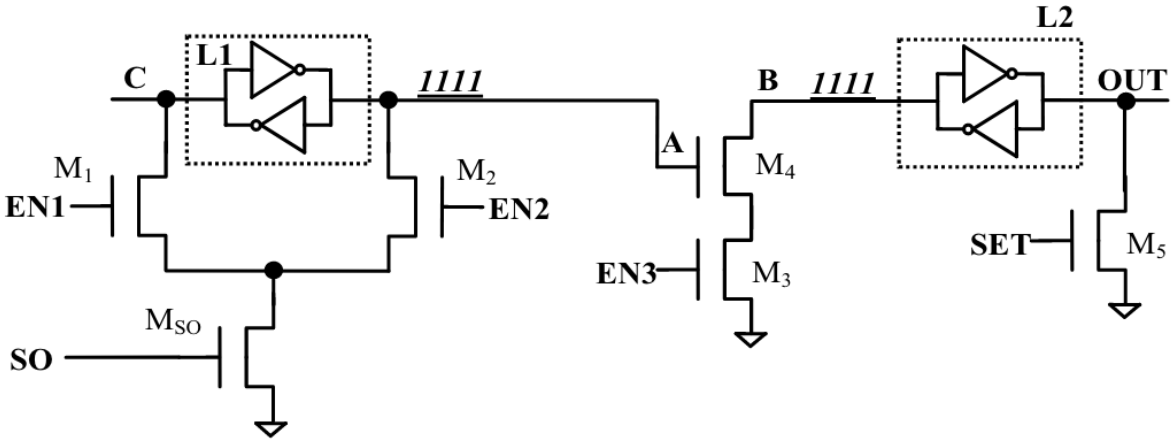


Fig. 10.13. MLC with cache read

施加Vread2，SO=0011，EN2=1，当且仅当S0=1时输入才会接地，此时输入为0，A为0，否则A保持原来的结果为1，所以SO=0011时，A为1100。将L1结果存到L2中，EN3=1，A=1时，B=0，否则维持原来的值1，所以A=1100时，B=0011，OUT为1100。

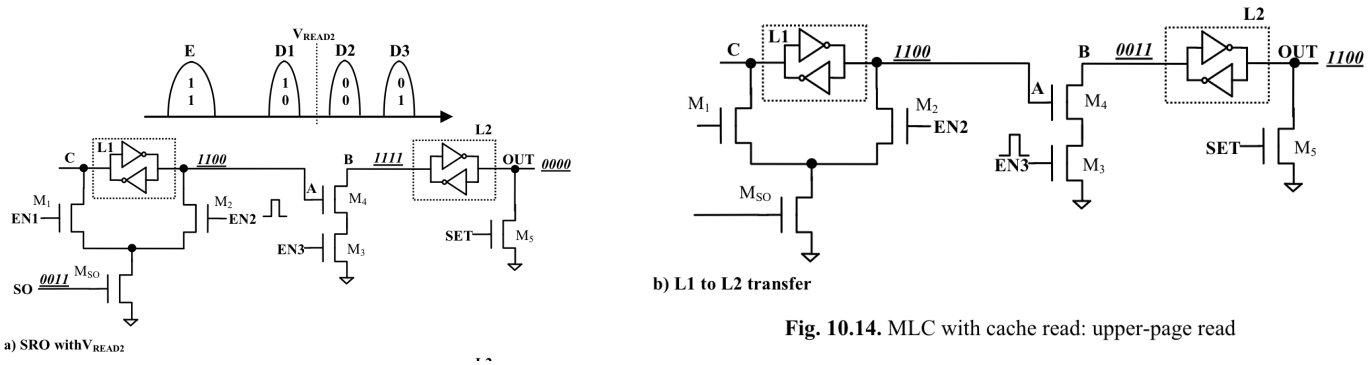
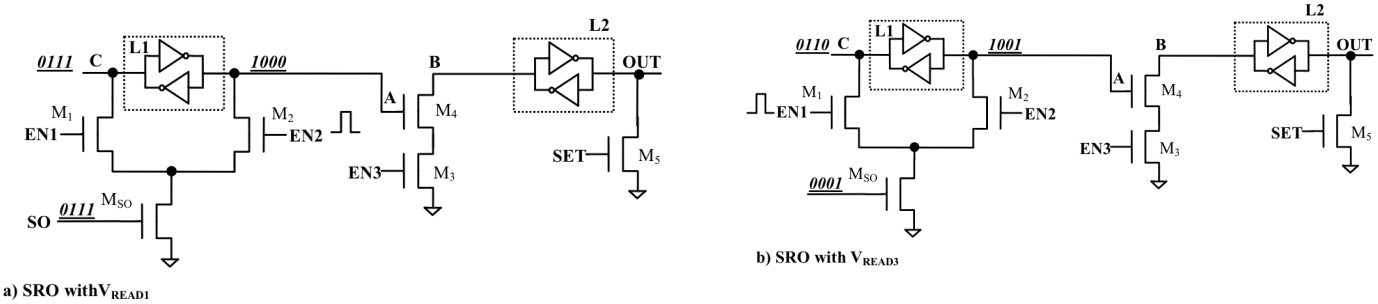
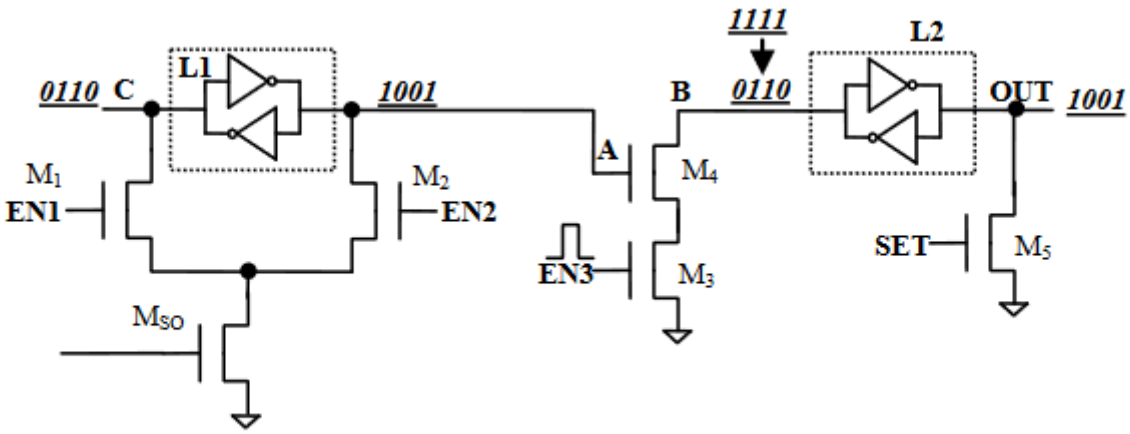


Fig. 10.14. MLC with cache read: upper-page read

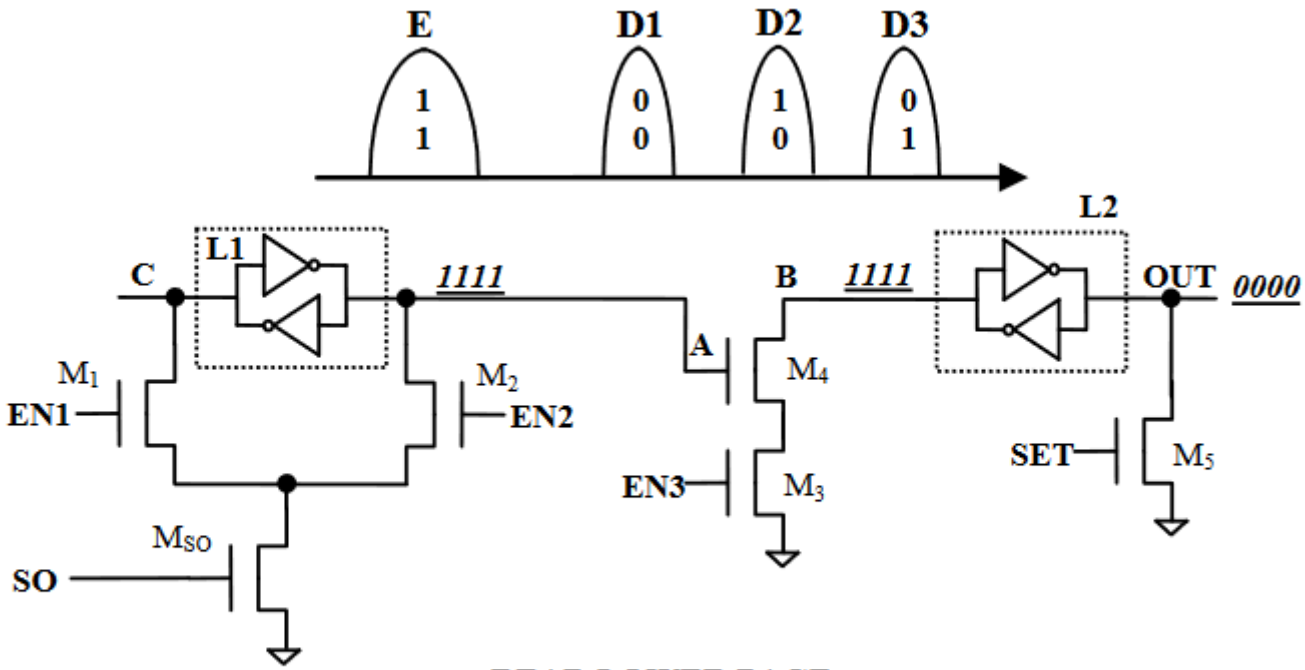
施加Vread1，SO=0111，M2打开，A=1000；施加Vread3，SO=0001，M1打开，C原本为0111，受到SO的影响为0110，所以A为1001。再由L1传输到L2中，输出到最后结果OUT。





c) L1 to L2 transfer

其他编码方式



READ LOWER PAGE

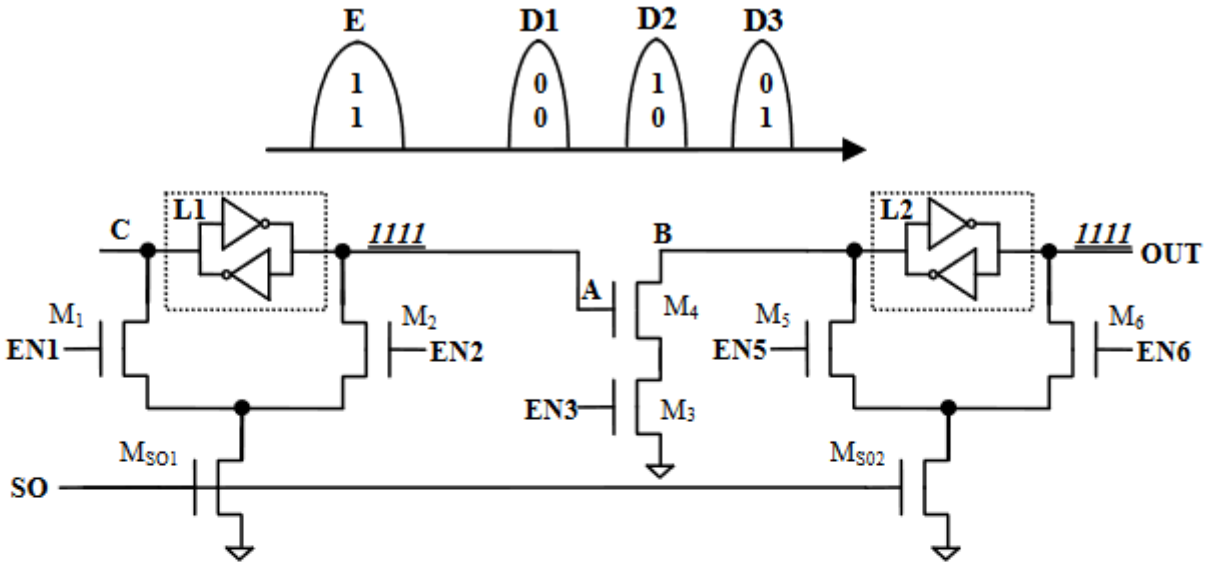
	SO	ENX	A	C	B	OUT
Start	xxxx	x	1111	0000	1111	0000
V_{READ1}	0111	EN2	1000	0111	1111	0000
V_{READ3}	0001	EN1	1001	0110	1111	0000
L1 TO L2 Transfer	xxxx	EN3	1001	0110	0110	1001

READ UPPER PAGE

	SO	ENX	A	C	B	OUT
Start	xxxx	x	1111	0000	1111	0000
V_{READ1}	0111	EN2	1000	0111	1111	0000
V_{READ2}	0011	EN1	1011	0100	1111	0000
V_{READ3}	0001	EN2	1010	0101	1111	0000
L1 TO L2 Transfer	xxxx	EN3	1010	0101	0101	1010

Fig. 10.16. MLC with different distribution coding and cache read

MLC page buffer for reading two pages



READ LOWER AND UPPER PAGE

	SO	ENX	A	C	B	OUT
Start	xxxx	x	1111	0000	0000	1111
V_{READ1}	0111	EN2/EN6	1000	0111	0111	1000
V_{READ2}	0011	EN1	1011	0100	0111	1000
V_{READ3}	0001	EN2/EN5	1010	0101	0110	1001

LOWER PAGE　　UPPER PAGE

Fig. 10.17. MLC page buffer for reading two pages

矩阵占用面积的减少具有外围电路占用面积(特别是感测电路的面积)增加的缺点。事实上，感测电路内锁存器的数量（至少）等于单元内存储的位数 n。多个锁存器可以支持cache读，带来的代价是外围电路面积的增加。

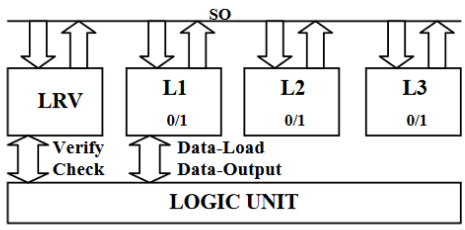


Fig. 16.12. 8LC sensing architecture

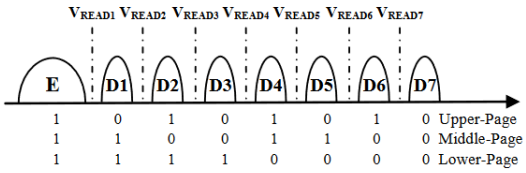


Fig. 16.21. 8LC read levels

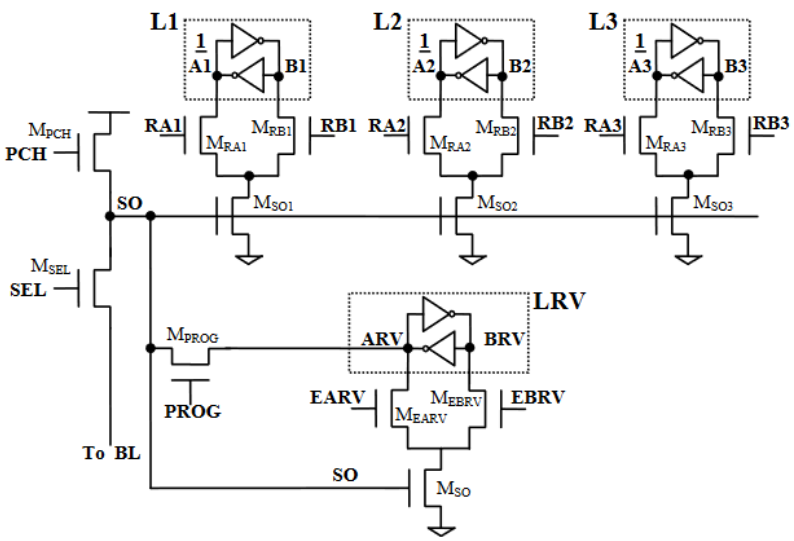
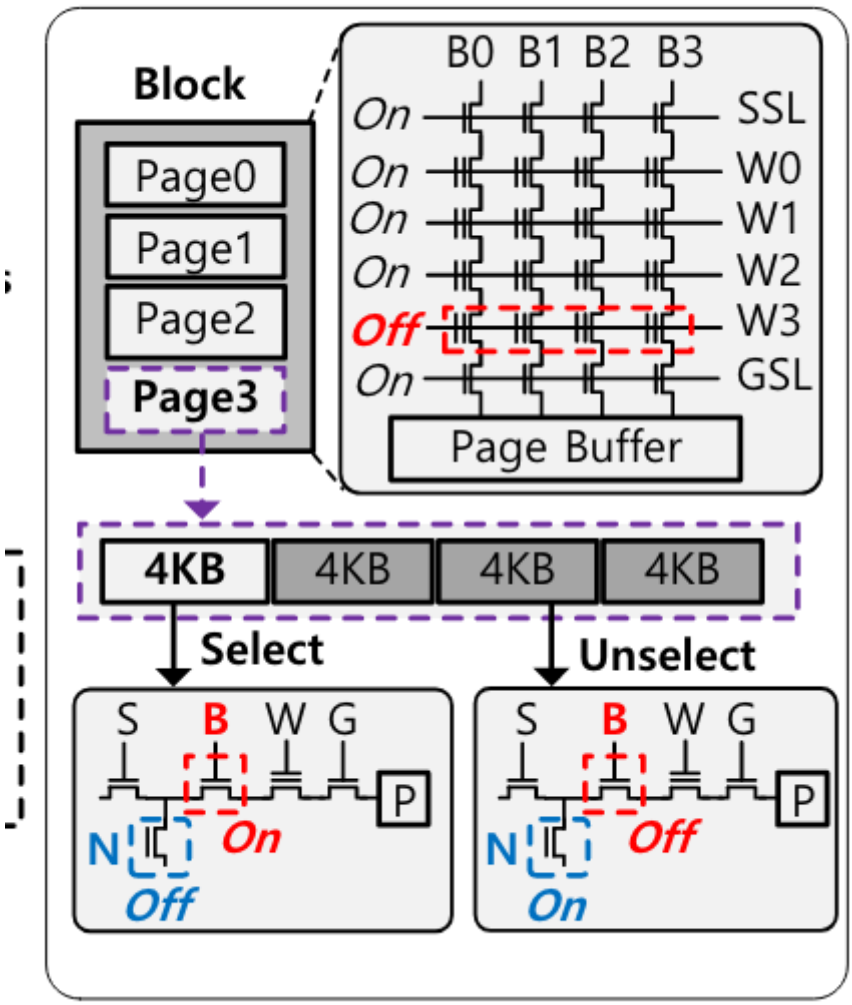


Fig. 16.20. 8LC read circuit

Table 16.4. Enabled transistors after the SRO at V_{READi}

	V_{READ1}	V_{READ2}	V_{READ3}	V_{READ4}	V_{READ5}	V_{READ6}	V_{READ7}
M_{RX3}	M_{RA3}	M_{RB3}	M_{RA3}	M_{RB3}	M_{RA3}	M_{RB3}	M_{RA3}
M_{RX2}		M_{RA2}		M_{RB2}		M_{RA2}	
M_{RX1}				M_{RA1}			

部分读功能



为了实现部分读取操作，添加了两个晶体管，即用于选择位线的 BLC 晶体管（红色 B）和用于对位线放电的 NLO 晶体管（蓝色 N）。对于选定的部分单元，BLC晶体管导通，使得单元中的电流可以对页缓冲器充电。在未选择的部分单元 NLO 晶体管导通且电流放电时，BLC 晶体管关闭。因此，电流无法对页缓冲器充电。结果，仅读取选定的部分单元。仅读取选定的部分单元。由于激活的位线和单元的数量减少，部分读操作与正常读操作相比可以减少20%的延迟，并减少40%的电流消耗。

擦除

擦除操作通过向基板施加高电压 (> 20 V) 将电子从单元的电荷陷阱中排出，从而降低单元的 VTH 电平。由于编程和擦除操作是单向的，因此需要先擦除页才能编程数据（先擦除后写入）NAND 闪存芯片以块粒度执行擦除操作（出于成本原因）。这会导致高擦除带宽，因为一个块由数百（例如，576 [37]）或数千（例如，1,472 [44]）页组成，但也导致擦除延迟 tBERS 比编程延迟 tPROG 长得多（例如，3.5 ms 与 660 µs）