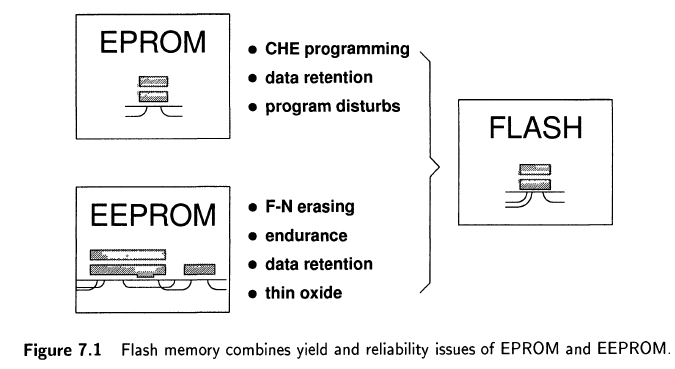
7. 闪存可靠性

文摘：参考主流技术，综述了影响快闪存储器成品率和可靠性的相关失效机制，指出了隧道氧化缺陷所起的主要作用。良好的测试方法与适当的产品设计相结合，在晶片级筛选隧道氧化物潜在缺陷是提高闪存可靠性的关键因素。本文讨论了由编程/擦除循环引起的器件性能下降，包括典型单元的行为、存储阵列分布的演变和单bit故障模式。氧化物陷阱被证明是不稳定擦除和单bit数据丢失的关键失效机制，应力导致的泄漏电流对数据保持的影响限制了隧道氧化物厚度的可扩展性。最后，简要分析了多级单元概念的可靠性影响。

7.1 引言

近年来，闪存已成为非易失性存储器中的明星，因为它提供了电擦除和重新写入的能力，迄今为止只有昂贵的EEPROM具有这一特点，其成本与EPROM相当。

不幸的是，除了具有如此吸引人的理想特性外，闪存也存在EPROM和EEPROM的良率和可靠性问题（图7.1），以及该技术特有的一些其他问题。写入操作涉及Fowler-Nordheim隧道效应和热载流子效应；必须在不降低数据保持性能的情况下实现编程/擦除循环耐久性；单晶体管结构使存储单元存在阵列干扰和过擦除问题。



这些问题使得Flash技术成为最难掌握的技术之一，需要非常精确的工艺优化和严格的工艺控制[1]。

从固有特性和缺陷密度方面来看，影响闪存良率和可靠性的最重要因素是隧道氧化物的质量。

隧道电流的均匀性决定了存储器阵列中擦除阈值（Vt）分布的宽度，以及在电流应力下的氧化物损耗，存储器耐久性与隧道氧化物的工艺条件密切相关。使隧道氧化物电导率局部增加的点缺陷是导致过擦除和编程干扰导致的单bit故障的原因，与闪存良率问题密切相关。

在过去的几年里，人们努力改进评估和监测隧道氧化物质量的方法，这与电测试方法和相关测试结构有关[2-4]。

此外，闪存的高可测试性允许在晶圆上筛选潜在缺陷，这些缺陷可能导致与编程干扰、数据保持和过早氧化击穿有关的单bit故障。

参考主流技术，本章将回顾影响闪存良率和可靠性的主要问题，详细讨论限制内存耐久性的故障机制，并介绍与多级存储相关的问题。

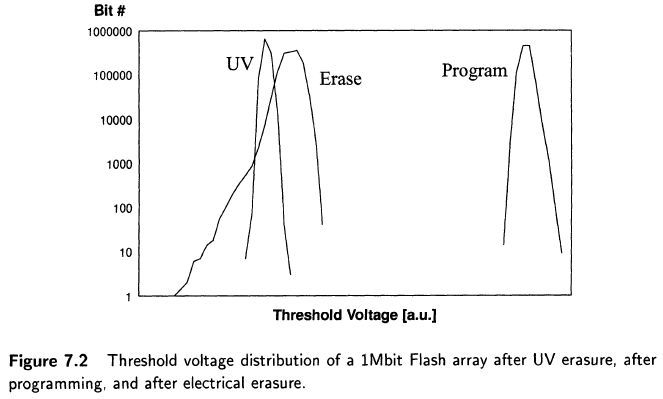
课题的广度和复杂性决定了演讲的组织方式；本章将根据作者的经验，深入了解关键概念和最相关的问题，而不是文献中报告的所有失效机制的冗长列表，强调对过程控制和产品设计与测试的实际影响。这就像是在闪存可靠性迷宫中的导游：导游的目的不是展示一切，而是向你展示出路。

7.2存储阵列阈值（Vt）分布和隧道氧化物“缺陷”

隧道氧化物“缺陷”的作用非常重要，如果不能很好地理解氧化物“缺陷”的含义以及它们如何影响存储器的功能性和可制造性，就无法开始讨论闪存的良率和可靠性。

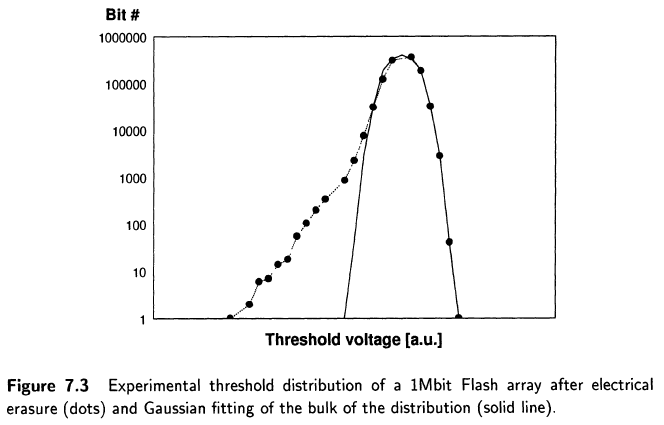
第二章详细介绍了闪存单元的编程和擦除机制，引入了过擦除的概念。本节将讨论存储阵列的擦除。

图7.2显示了存储器阵列中单元阈值电压的典型分布。紫外擦除分布非常窄且对称；精确分析了由于临界尺寸、厚度和掺杂的随机变化而产生的高斯分布，直接或通过耦合比导致阈值电压的离散。

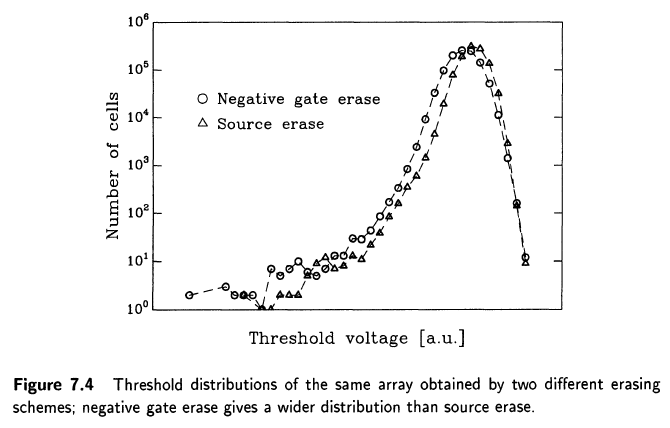


编程分布比紫外线擦除的分布更宽，但仍然是对称的；放大是因为大多数导致紫外线擦除单元Vt离散的参数也影响编程单元的阈值偏移。

擦除后阈值电压的分布要宽得多，且严重不对称。图7.3 表明，这种分布的主体（包括了超过99%的单元）是高斯分布，标准偏差大于编程单元的标准偏差；属于高斯分布的位称为“正常”位。分布的左侧由擦除速度高于平均值的单元的指数尾部组成：这些位将被称为“尾”位。



正常位的阈值电压的离散是由于耦合比变化引起的，Yoshillawa等人[5]对此进行了精确建模。通过比较两种不同擦除方法在同一阵列上获得的擦除阈值分布，该模型得到了很好的证实（图7.4）；负栅擦除方案得到的分布比源擦除方案得到的分布更宽。实际上，在负栅极擦除方案中，擦除速度同样受到源极和栅极耦合比的影响，而栅极耦合比对源极擦除速度的影响可以忽略不计。



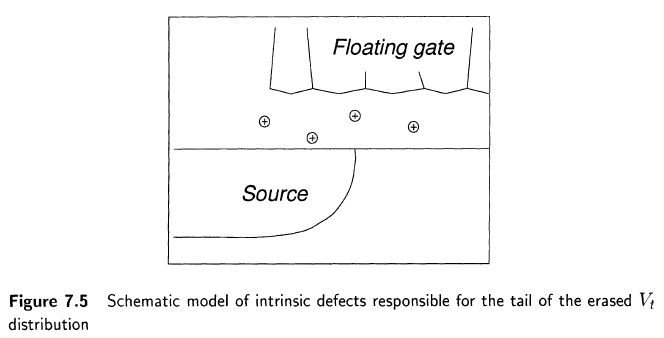
理解尾位的性质至关重要。由于在相同的外加电压下，这些位的擦除速度比正常位快，因此应该假定它们在某种程度上是“有缺陷的”。但是它们太多了，无法和外部缺陷联系在一起。

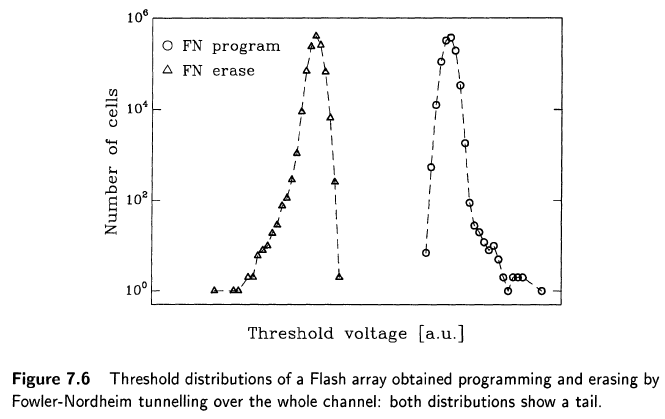
有两个固有的结构缺陷可用来解释尾位的性质。

Muramatsu及其同事[6]将擦除Vt分布中存在的尾部归因于注入电极的多晶结构；晶界处的势垒高度变化和/或底层氧化物中的掺杂将导致隧穿电流的局部增强。

Dumn及其同事[7]给出了另一种解释，他们将擦除尾部建模为纯粹由于隧道氧化物中随机分布的正电荷。该模型基于众所周知的类施主体氧化物陷阱的存在，模拟结果表明，由于靠近注入电极的基本正电荷的存在，隧道电流密度大幅增加（见第7.5.3.1节）。

图7.5所示的两个模型可能都有效；多晶硅晶粒尺寸对擦除Vt分布的影响已得到明确证明，但氧化物电荷也确实发挥了作用：图7.6显示了使用均匀通道FN/FN写入方案获得的Vt分布，即：向栅极施加正电压进行编程，向栅极施加负电压进行擦除：当电子从衬底注入时，负极是单晶，但也存在尾部！





由于擦除阈值分布的指数尾部主要与结构缺陷（即固有缺陷）有关，因此可以通过工艺优化（例如，在隧道氧化、浮栅和栅间介质工艺）将其最小化，但无法消除；产品设计时必须考虑到这种尾部的存在。

在以下章节中将讨论尾位的可靠性；可以预料的是，如果作为个体，即不考虑在阵列中影响，这些尾位与正常位一样可靠。

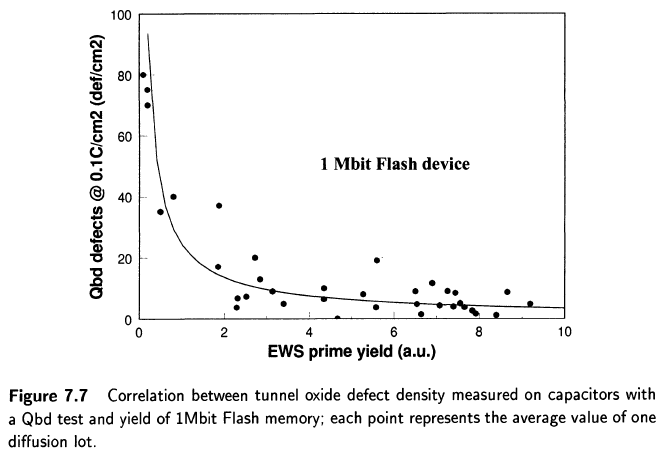
正常位和尾位不会覆盖全部存储单元。在一个非常“干净”的过程中，一些位的擦除速度比普通位快得多，大约为百万分之几甚至更少，它们也从分布的尾部突出；在需要共模擦除的阵列中，这些位通常会被过度擦除，这将引起稍后讨论的问题。

导致这些“快速擦除”位的物理缺陷可能不同于尾位的物理缺陷；由于它们显然不属于尾位的统计分布，而且数量很少，“快速擦除”位很可能是由于外部缺陷造成的，例如由污染引起的隧道氧化物变薄。尾位和快速擦除位之间的区别可能是任意的，在某种程度上无法定义一个清晰边界（例如，参见第7.5.3.1节中对不稳定位的讨论）；但这对于理解对良率和可靠性的影响非常重要。例如，已经证明的用于预氧化清洗的直流水或化学品中重金属或有机化合物的ppm级污染可以增加快速擦除位的数量，而不会影响已擦除Vt分布的指数尾部。此外，虽然尾位表现出良好的数据保持，但在数据保持测试中，快速擦除位通常表现出更强的电荷损失。

监控隧道氧化物中的外部缺陷密度有助于确保闪存生产的良率和可靠性。为了开发一种测试方法来测量氧化物缺陷并在过程控制策略中有效实施，人们付出了巨大努力。

指数斜率电流应力（ERCS）[3]已被证明是检测薄氧化物中缺陷的非常有效的测试。由于它提供了灵敏度和测试时间之间的最佳权衡，因此最适合密集的统计过程控制。该方法可以应用于简单的电容器测试结构，因此可以有效地实现工艺设备的短回路监测。

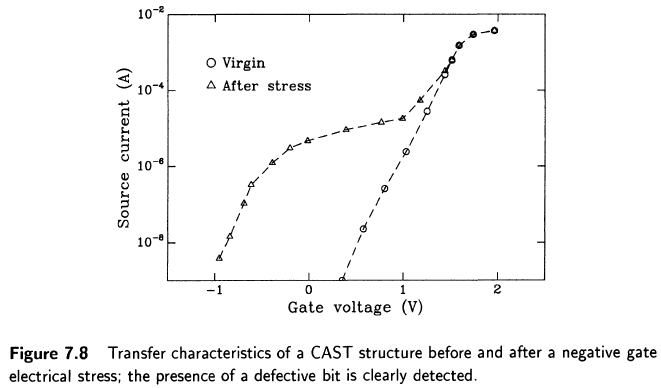
此外，已经证明，通过Qbd测量获得的缺陷密度与闪存的良率相关：图7.7显示，当Qbd缺陷密度增加时，1Mbit内存的良率降至零。由于高缺陷密度批次中的成品率损失主要是由于过擦除位造成的，这意味着导致过早氧化击穿的物理缺陷也会导致单个位表现为快速擦除，反之亦然，至少一些快速擦除位具有潜在缺陷，限制其耐久性。

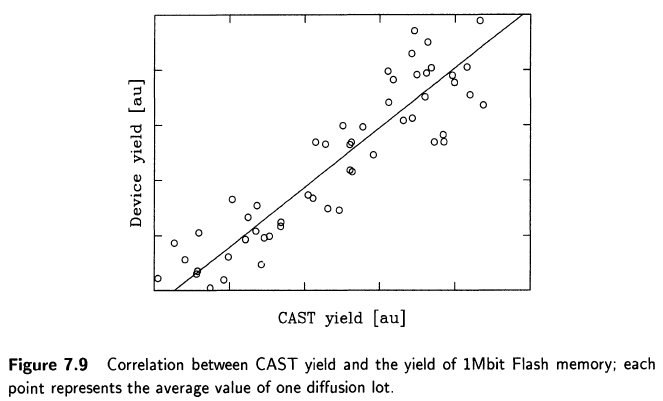


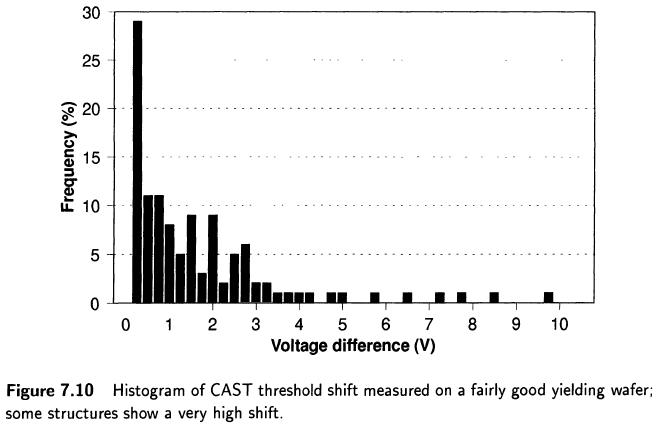
同一图显示，对于最低的Qbd缺陷密度值，不再与存储器的良率数据相关，这意味着，虽然Qbd测试检测到的所有缺陷也会导致存储器故障，但并非所有导致过擦除的缺陷都可以通过Qbd测试检测到。这种部分相关性易于解释，考虑到存储单元是一种非常灵敏的静电计，它可以显示比Qbd测试更不明显的氧化物缺陷。

为了提高氧化物缺陷的可检测性，专门为FLASH开发了一种测试方法，该方法积极利用了存储单元的固有灵敏度及其亚阈值I-V特性的斜率。

这种方法被称为单元阵列应力测试（CAST）[4]，所使用的结构是一个并行连接的存储单元阵列，以便它们可以同时承受应力和测试。这种测试结构完全类似于真实内存的阵列，并且可以以相同的方式擦除；通过分析CAST结构I-V特性曲线的亚阈值区域，可以发现快速擦除位的存在（图7.8）。CAST的良率已被证明与实际内存的良率有很好的相关性（图7.9）。







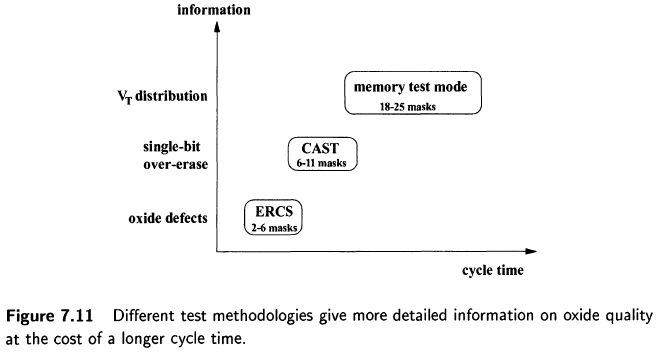
CAST结构为工艺开发提供了一个具有合理短周期的良好载体；此外，它不需要像真正的存储器一样需求巨大的设计工作和昂贵的测试设备。

对于本节的主题，CAST提供了一条真实存储器无法轻松获取的额外信息。由于快速擦除位独特且易于检测的特性，它们可以从已擦除阈值分布的指数尾部中脱颖而出，人们想知道，缺陷位距离主体分布多远仍可以被发现。在真实存储器上无法正常提供这个信息，因为在读模式下不可能将栅极偏置为负值，以及感测放大器饱和。

通过CAST可以有效地获得此类信息；图7.10显示了一个良率不错的晶片上缺陷结构的阈值Vt位移分布：这里有一些单个单元的结构，距离分布的主体有10V之多！同样的结果也可以以不同的方式读出：虽然施加到隧道氧化物上的电场在正常位上产生可检测的阈值Vt偏移在9-10MV/cm的范围内，但大多数缺陷位在低至2MV/cm的外加电场下显示出阈值Vt偏移！ CAST可以提供氧化物缺陷的电导率阈值的分布，这与从电容器击穿电场（Ebd）测量中获得的信息大不相同，也更有价值。

回到问题，所有快擦除位是否都属于已擦除Vt分布的尾部，在分布主体之外找到10V的指数尾部的概率将小于10-30！

作为本节关于氧化物缺陷的结论，值得再次强调持续控制和改善隧道氧化物质量对于闪存良率和可靠性的重要性。所述测试方法以及产品本身提供的信息都有助于开发和调高工艺的可制造性：如图7.11所示，它们以不同的加工周期为代价提供不同级别的信息。



7.3 主要良率和可靠性问题

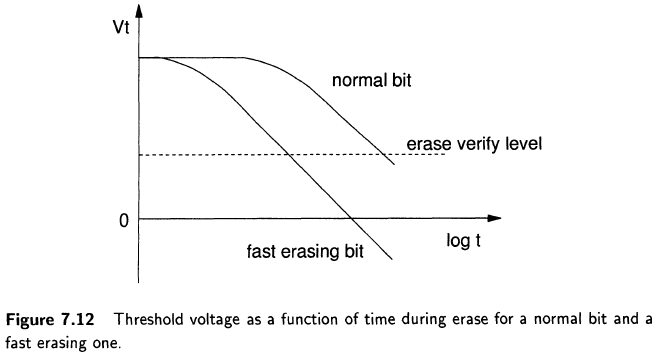
本节将说明有关闪存良率和可靠性的相关问题。

7.3.1 过擦除

与EPROM不同，闪存可能会被过擦除，因为它们的擦除机制不是自动停止的；只要将擦除电压施加到存储单元，电子就会不断地从浮栅中移除，如果不在正确的时间停止擦除脉冲，则阈值电压最终可能为负值。过擦除是故障的潜在原因，因为如果NOR类型内存阵列包含耗尽单元，则无法正确读取：连接到耗尽单元的同一位线的所有单元将被读取为“l”，而不管其实际内容如何。

为了在正确的时间停止擦除操作，使用了一种复杂擦除算法，该算法是一个短脉冲序列，然后在每个脉冲后验证闪存状态：一旦所有单元成功读取为“l”，该序列即结束。

由于阵列中的所有单元同时被擦除，并且擦除电压应用于所有单元，因此擦除最慢位所需的时间可能足以使最快位被过擦除。擦除算法只能确保在写入所有单元后立即停止擦除，但不能防止快速擦除位被过度擦除（图7.12）。



只有适当的工艺调整与适当的电路设计相结合，才能使擦除机制可靠地工作。在设计存储电路时，必须考虑擦除阈值Vt的分布，并且必须优化工艺以缩小其范围并使其处于控制之下。由于外部缺陷，必须使用冗余来修复快速擦除位。

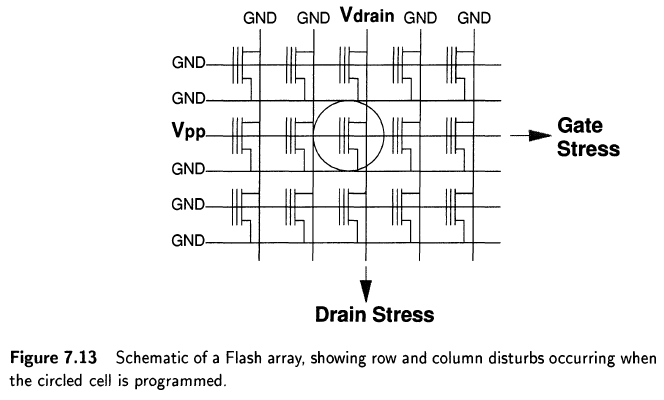
单元过擦除通常是闪存最关键的良率问题。已有不同的软编程和自收敛技术，可以通过恢复过擦除，甚至通过收敛分布[8，9]来解决该问题，但是，所有提出的方法都是耗时耗能的，并且它们也并非没有可靠性问题：大量使用这些技术来系统地缩小擦除Vt分布仍有待证明。无论如何，利用这些技术来恢复严重耗尽的位是不明智的；由于它们与外部缺陷相关，因此检测和修复此类位更可靠。

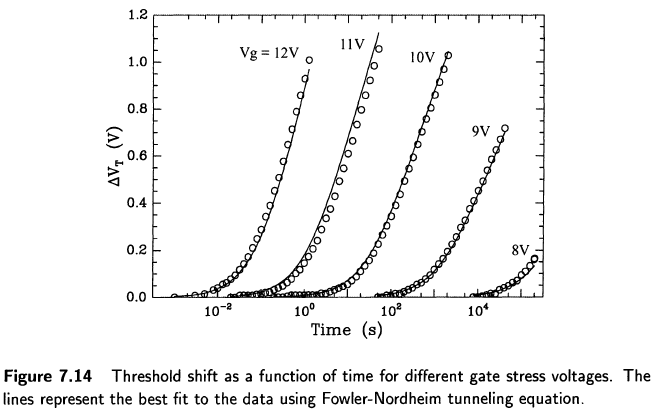
7.3.2编程干扰

被称为“编程干扰”的故障机制涉及在对闪存阵列中的其他单元编程时，由于施加在这些单元上的电应力而导致写入单元的数据损坏。由于栅极氧化物较薄，闪存比EPROM对编程干扰更敏感。

必须考虑两种类型的编程干扰：行干扰和列干扰。

行干扰是由于在同一字线上对其他单元进行编程时，施加到单元格的栅极应力引起的。参考图7.13，图中为一部分阵列的示意图，假设由圆标识的单元已被编程。向选定行施加高压，该行的所有单元必须承受栅极应力，而不会丢失数据。根据存储在单元中的数据，数据丢失可能是由栅极氧化物中的泄漏（由于“l”上的栅极应力而产生的电荷增益）或由层间电介质中的泄漏（由于“0”上的栅极应力而产生的电荷损失）引起的。图7.14显示了栅极应力对擦除正常单元的影响；电荷增益是由于从通道到浮栅的Fowler-Nordheim隧道效应。

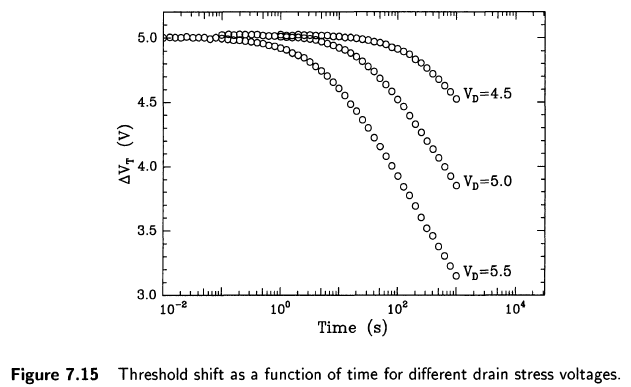




最坏情况下的应力时间取决于阵列架构和产品规格。假设字节编程的最大编程时间为50µs，并假设每行有128字节；最坏情况下的压力时间是

tstress =128 X 50µs=6.4ms。

列干扰是由于在同一位线上编程其他单元时施加到单元的漏极应力引起的。再次参考图7.13，所选列上的所有单元（正在编程的单元除外）都将栅极接地，并向漏极施加相当高的电压。在这种情况下，编程单元可以通过Fowler-Nordheim隧道效应从浮栅到漏极（软擦除）来释放电荷；图7.15显示了漏极应力对编程正常单元的影响。实际上，漏极应力与源极擦除的正常条件非常相似，只是加了较低的电压。



最坏情况下的应力时间再次与闪存架构和规格有关。假设每列有1K个单元，最大编程时间为50µs，最坏情况下的应力时间为

tstress=1024×50µs=51ms。

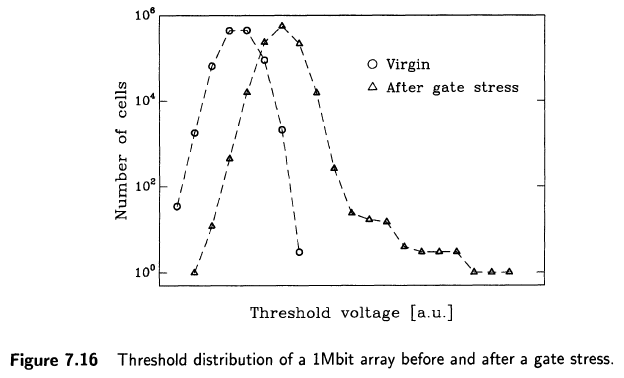
值得注意的是，计算应力时间时基于单个阵列考虑，例如大块存储器。这种情况对于扇区化的存储器来说更为关键，根据架构的不同，最坏情况下应力时间的计算需要考虑编程/擦除周期的数量，应力时间会增加几个数量级。

例如，一个闪存的扇区由不同扇区共享的列和字线组成。在计算最坏情况下的栅极应力时间时，必须考虑一个扇区必须保留其数据，而另一个扇区的写入时间与指定的持续时间相同的情况。假设8个扇区共用一条字线，每个扇区内每行有32个字节，每个字节的最大编程时间再次为50µs，编程/擦除周期数为105，则最坏情况下的应力时间为

tstress =（32×50µs）×（1+7×105）=l.6ms+1120s。

很明显，编程干扰在不同扇区之间的传播可能会对产品可靠性产生巨大影响，设计者必须认真考虑。防止干扰传播的最有效方法是在分割位线和分割字线架构中使用选择晶体管，以将每个扇区与其他扇区完全隔离；但无论是在阵列效率还是在处理复杂度方面，这都是一个昂贵的解决方案。另一种方法是使用衬底或源偏压来减少未选择扇区上的应力。

编程干扰确实是闪存中的一个重要问题；单元和电路的设计必须具有安全裕度，不仅与正常位的应力敏感性有关，也要考虑到分布的尾部（图7.16）。尽管如此，即使产品和工艺经过精心设计，栅氧化层或层间电介质中的缺陷也可能由于编程干扰而导致单bit故障。



7.3.3 读干扰

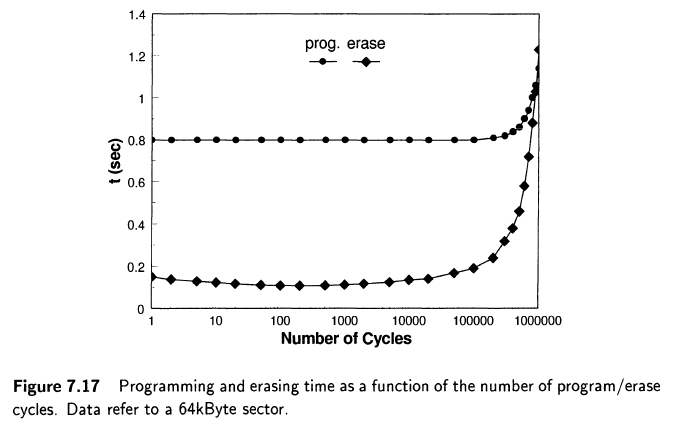
当读单元时，栅极电压和漏极电压分别施加到选定的行和列。这种情况会对处于擦除状态的单元造成两种读干扰。被读出的单元如果漏极电压不够低，则可能会由于CHE注入而发生不必要的编程。此外，所选行上处于擦除状态的所有单元都受到低压栅极应力的影响，该应力可诱导从通道到浮栅的隧道电流，再次导致不必要的编程。在这两种情况下，氧化物电流都非常低，但最坏情况下的读干扰时间与器件寿命具有相同的数量级。

在这里，可以通过适当的单元和电路设计来实现安全裕度，主要是在氧化物厚度和施加的读电压方面。当我们考虑编程/擦除循环后单元的读干扰特性时，会出现一个主要问题。循环引起的隧道氧化物退化可能会对读出的抗干扰产生不利影响，尤其是对于非常薄的氧化物来说，将在第7.5.3节讨论。

7.3.4编程/擦除耐久性

闪存的耐久性可能受到两种故障模式的限制：1）由于氧化老化而降低编程/擦除效率，这会导致参数失效；2） 由于隧道氧化缺陷导致的单bit故障。

第一种失效模式与相当均匀且可重复的存储单元损耗性能有关；第7.5.1节将详细讨论导致这种退化的机制。在103-104个编程/擦除周期中，编程时间和擦除时间都随着周期数的增加而增加（图7.17），并最终超过规范限制。通过适当的器件工程和优化隧道氧化工艺，可以减少循环磨损。然而，一旦工艺和产品符合给定的耐久性规范，就不会因批次间的差异而产生重大问题。



实际上，耐久性问题大多与单bit故障有关。

第7.5节将广泛涵盖由编程/擦除循环引起的故障模式，第7.5.3节将专门讨论单bit故障。

就本节而言，只需强调，在任何更复杂的失效机制（如第7.5节中讨论的失效机制）之前，隧道内氧化物的外部缺陷可能会由于过早的氧化物击穿或数据保持的退化而导致单bit耐久性失效。这些故障模式历来是EEPROM的主要可靠性问题，迫使在1Mbit甚至更低容量上引入高密度存储器的片上纠错。事实上，闪存在生产中的复杂度为8Mbit和16Mbit，无需片上纠错；这并不是因为闪存对隧道氧化物中的缺陷不敏感，而是因为相反的原因，如第7.5节所述。

7.5编程/擦除周期引起的故障模式

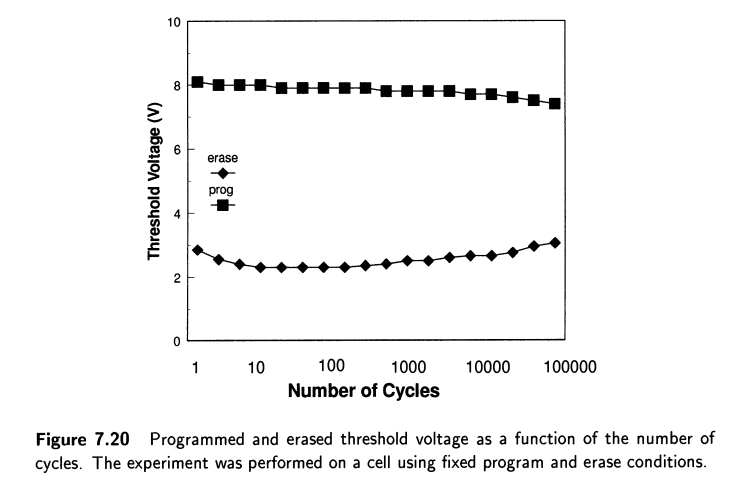
在上一节中，我们已经看到了适当的晶圆级测试在消除潜在缺陷的影响方面的有效性，大大提高了产品可靠性。

可靠性问题与编程/擦除周期引起的故障有关，因为它们在晶圆测试时无法充分覆盖；本节将专门介绍限制闪存寿命的故障机制

7.5.1存储单元固有耐久性

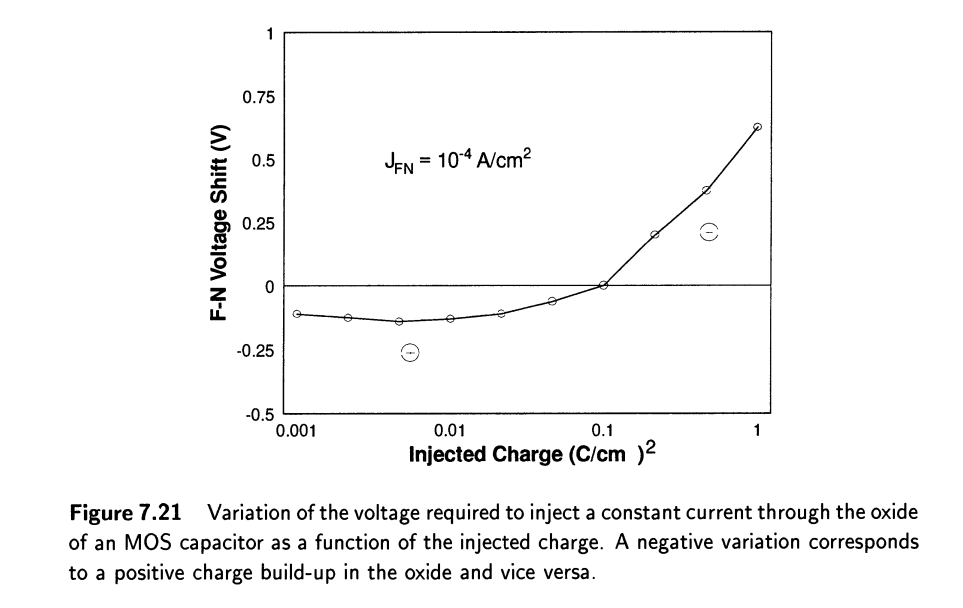
编程擦除循环引起的单元性能磨损是由于氧化物退化引起的。图7.20显示了单一单元耐久性操作后，试验的典型结果。即使用恒定脉冲进行实验后，编程和擦除阈值水平的变化，给出了氧化物退化的度量。

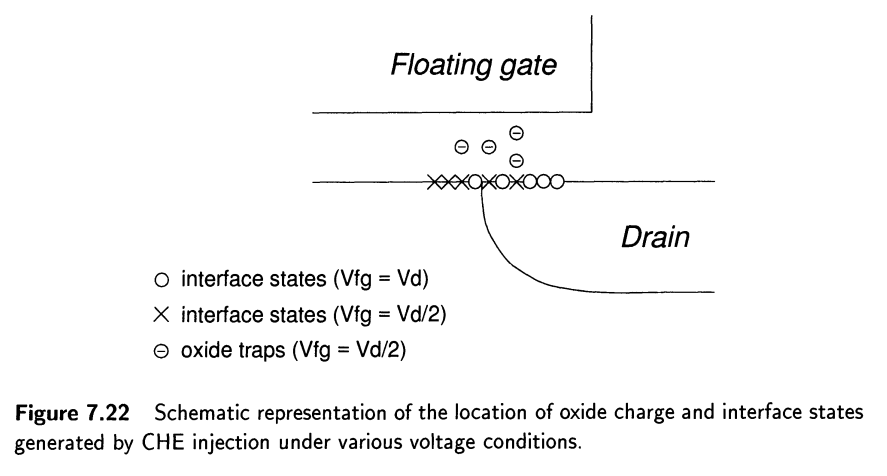
擦除阈值电压（Ve）的演化过程与EEPROM单元中典型的演化过程相似；它反映了众所周知的隧道氧化物中净固定电荷的动力学，作为注入电荷的函数，如图7.21所示。 因此，Vte的初始降低是由于正电荷的堆积，从而提高了隧道效率，而Vte的长期增加是由于负陷阱的产生。

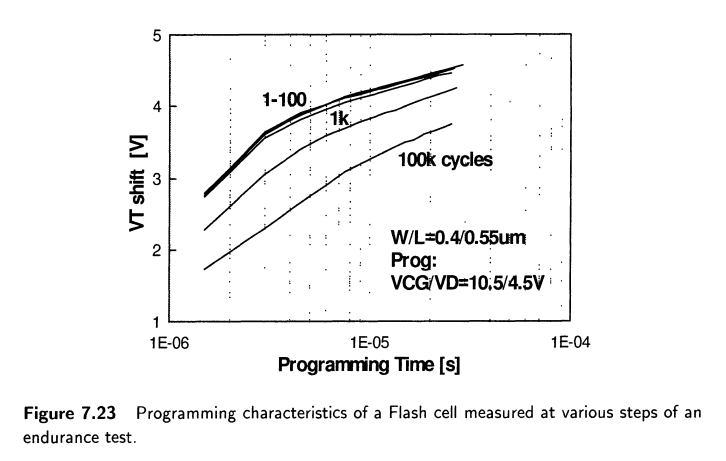


Yamada等人很好地解释了高循环次数下编程阈值电压的降低.这归因于存储单元漏极侧的氧化物陷阱和界面态生成（图7.22），这是CHE编程固有的一种渐变机制。根据Yamada及其合著者的模型，当浮栅几乎为中性且单元在三极管模式下工作时，对应于编程开始的应力状态反应在漏极区域上产生界面状态。

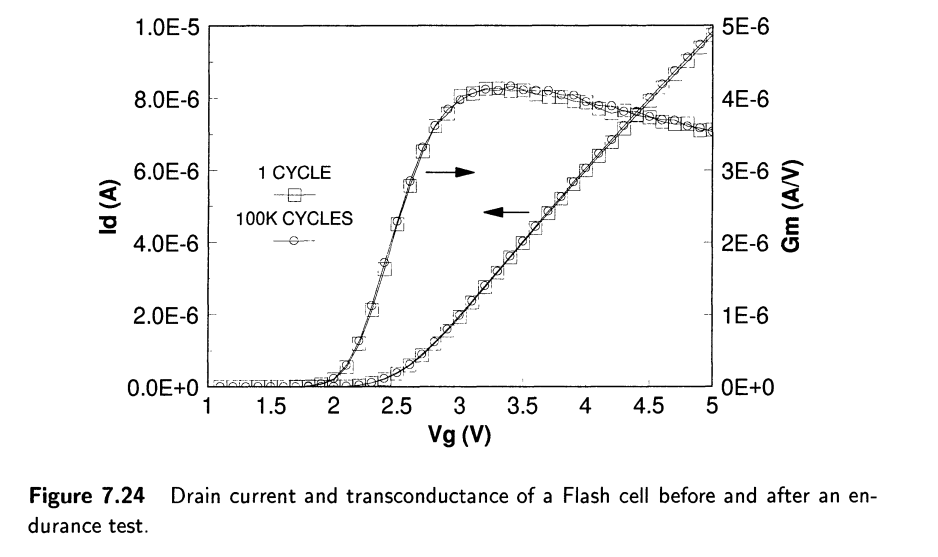
编程结束时，浮栅带负电，单元在五极管模式（Vfg~Vd/2）下操作，在通道上产生界面态和体氧化物电荷。这种情况与单元性能的下降最为相关：界面状态减少了电子平均自由程，影响了热载流子的产生机制，电子被困在隧道中。氧化物改变了注入点的电场，降低了编程效率。图7.23清楚地显示了编程/擦除循环导致的编程曲线退化。







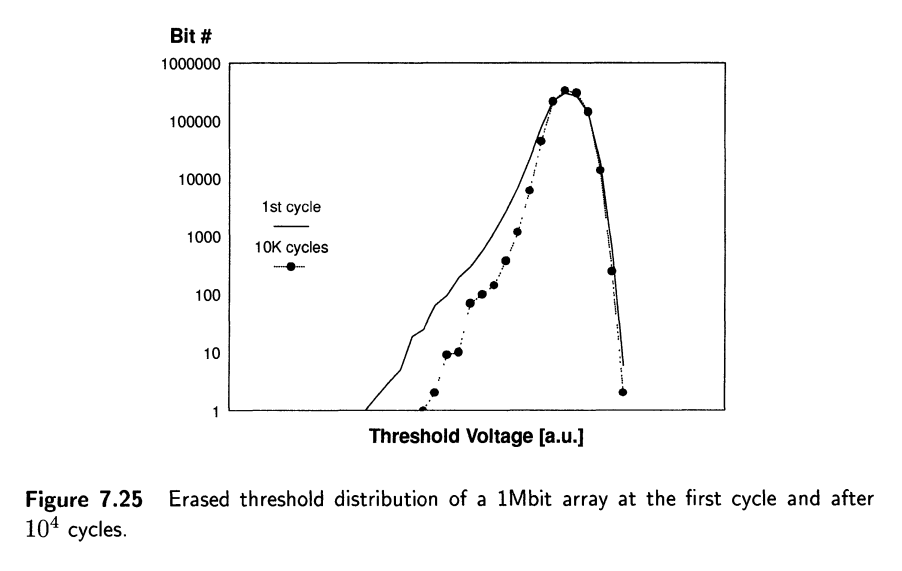
然而，一个设计合理的存储单元可以达到10万-100万次编程/擦除周期，写入性能下降可以接受，跨导没有明显的降低。图7.24.



7.5.2分布图形尾巴的bit行为

在讨论了单个细胞的耐久性之后，现在让我们考虑一个记忆阵列 。正常位的行为反映了单个处理单元的行为，如上一节所述：关注来自分布图形尾巴处的bit。

通常，编程/擦除循环不会导致擦除的Vt分布的扩大;相反，增加了循环次数（图7.25），尾巴变得更窄。尾部位的加速磨损与这些位擦除速度更快的模型一致，因为由电场峰值引起的局部较高电流注入。事实上，较大的电流密度对应于负陷阱的较高生成速率，即更快的老化。因此，产生的负电荷部分平滑了电场的峰值，使电流注入更加均匀，并且尾部比特的擦除速度更类似于正常比特的擦除速度。



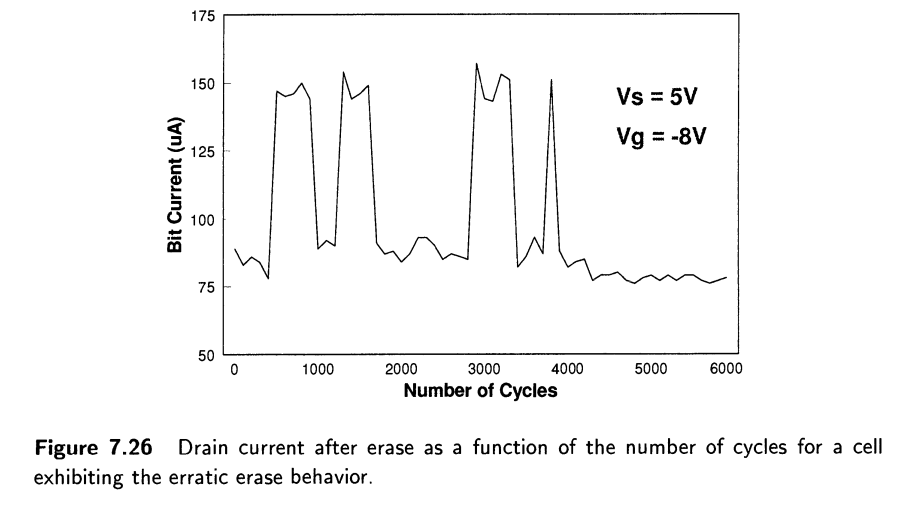
7.5.3单比特失效机制

到目前为止，我们对闪速存储器耐久性的讨论给出了一个非常令人满意的画面：外部氧化物缺陷在硅片分选时可以有效地检测到，性能磨损是可重复的，并且与产品规格兼容，擦除的Vt分布不会随循环而恶化，相反收窄了分布图形。

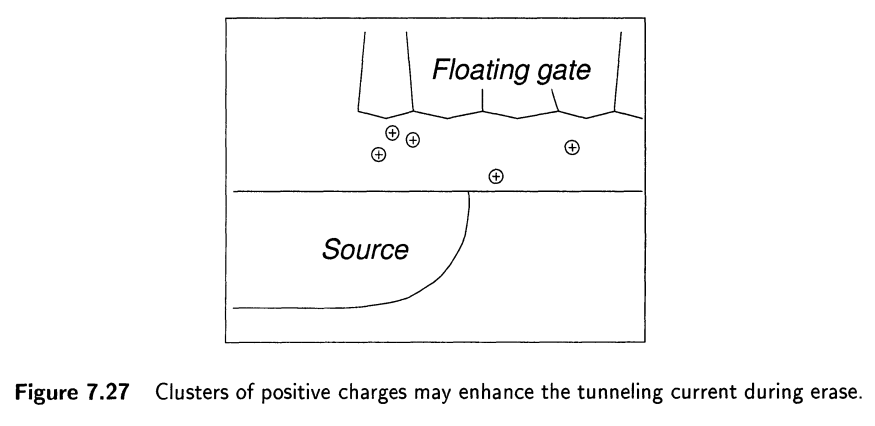
因此，我们应该预计，在设计得当的闪存中，不会发生持久性故障。实际上，耐久性失效通常是由单个位引起（非常复杂的机制造成的）。本节将重点介绍。

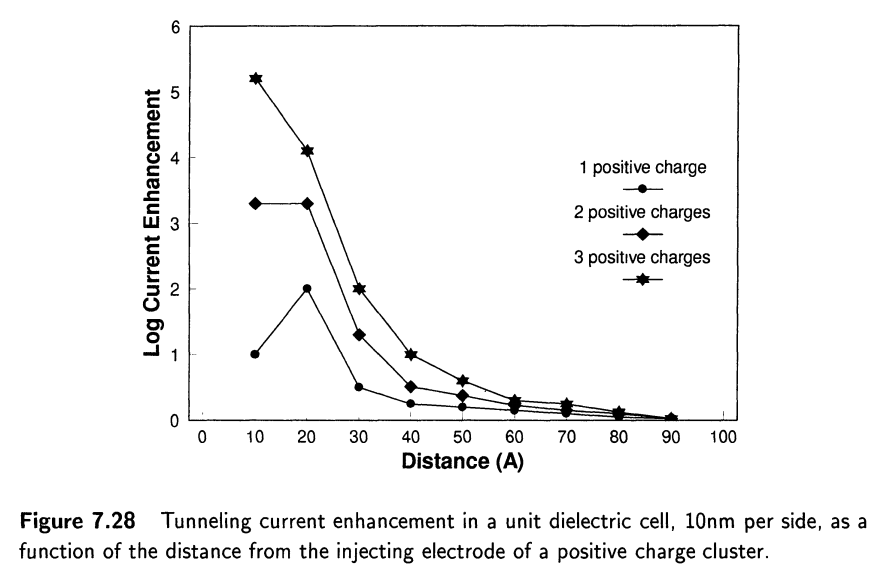
7.5.3.1不稳定的擦除现象

到目前为止，编程/擦除循环中最相关的单位失败机制是不稳定擦除，这是Ong等人首次提出的，并由Dunn等人进一步研究。 不稳定位在擦除过程中表现出不稳定和不可预测的行为：它们的Vte（位电流）于不同周期，在两个或多个不同值之间随机变化（图7.26），在Vt的窗口间来回移动，分布可以到尾部的最低部分；不稳定的位可能会导致过擦除失效。

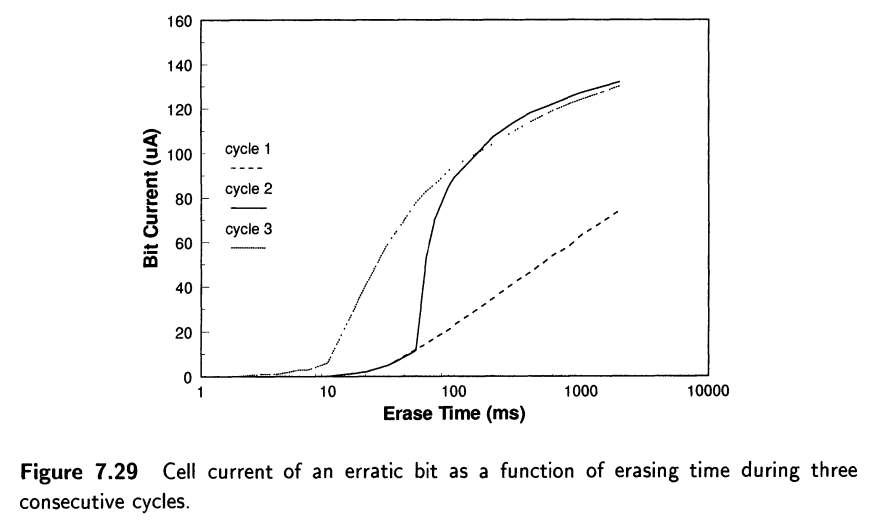


这种行为被归因于隧道氧化物中的空穴陷阱：空穴陷阱的统计分布给出了具有三个或更多正电荷簇的极低但有限的概率（图7.27），其电场相互重叠，从而产生隧道电流的局部增加。在这种情况下，单个正电荷的捕获/解捕获会导致擦除速度和阈值水平发生可检测的变化。该模型已通过WKB计算7、14得到定量证实； 7.28显示了由单、双和三个正电荷引起的电流密度增加，这是它们与注入电极距离的函数：电流密度的增加可以达到4或5个数量级。





所述模型的实验验证如图7.29所示。该图显示了在三个连续周期内测量的不稳定位的擦除曲线，即单元电流随擦除时间的变化。在第一个周期中，单元表现为正常位，在擦除结束时，电流达到已擦除Vt分布中单元的典型值。在下一个周期中，单元开始擦除，与第一个周期一样；然后，电流以更高的速率突然增加，导数明显不连续，即隧道电流：；在第二个循环结束时，单元被过度擦除，其电流达到与耗尽单元对应的值。



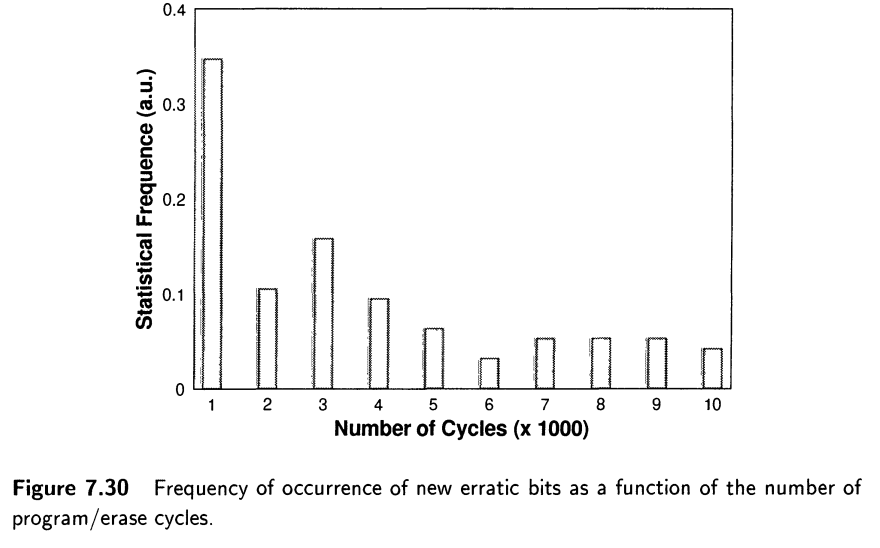
在第三个周期中，单元从一开始就表现为一个快速擦除位，在结束时，单元电流与第二个周期中的电流相同。曲线如图所示。7.29可以用所提出的模型清楚地解释：在第二个循环期间，一个孔被集群的一个陷阱捕获导致位不稳定，而集群的其他陷阱已经带正电荷，这决定了第二个循环中隧道电流的突然增加。值得注意的是，图.7.29以某种方式表示工业产品检测到的单个基本电荷的捕获事件的图景。

为了更好地理解该机制并解决不稳定位问题，应回答两个相关问题。

1.不稳定比特的“生成速率”与周期数的关系是什么？

2.负栅极擦除方案会消除不稳定的位现象吗？

第一个问题显然与所涉及陷阱的性质、它们在循环过程中的生成速率以及它们的空穴捕获截面有关。7.30显示了前10^4个周期中新不稳定位的发生率直方图；它是通过对1Mbit闪存进行大样本测试循环测试中找出包含的不稳定位，并统计其中有多少在前1000个循环中开始显示不稳定行为，有多少在后1000个循环中开始显示不稳定行为，依此类推。



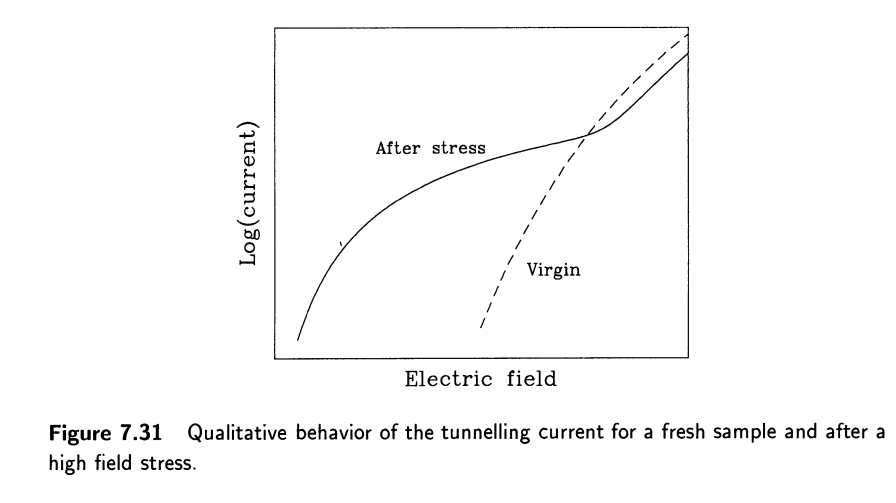
随着周期数的增加，不稳定位发生率降低，但即使在10^4个周期后也不会变为零。这种行为与已知的应力隧道氧化物中正电荷的演化非常一致：正电荷在电流应力的初始阶段堆积得最快，但不会饱和。出于实用目的，图7.30证明即使10^3循环测试也不能有效筛选不稳定的位。

第二个问题与陷阱的起源有关。如果它们是源结耗尽区雪崩倍增产生的热空穴，负栅擦除方案将显著降低热空穴生成率，实际上消除了不稳定的位问题。事实并非如此：7.26的数据。来自使用负栅极擦除方案的存储器（Vo=-8v，Vs=5V）

总之，由于不稳定行为是由于固有氧化物缺陷的统计波动引起的，因此可以通过工艺优化减少不稳定位的发生，但不能完全消除；已经开发了设计解决方案来解决电路级的问题。

7.5.3.2编程/擦除循环后的单位数据丢失。

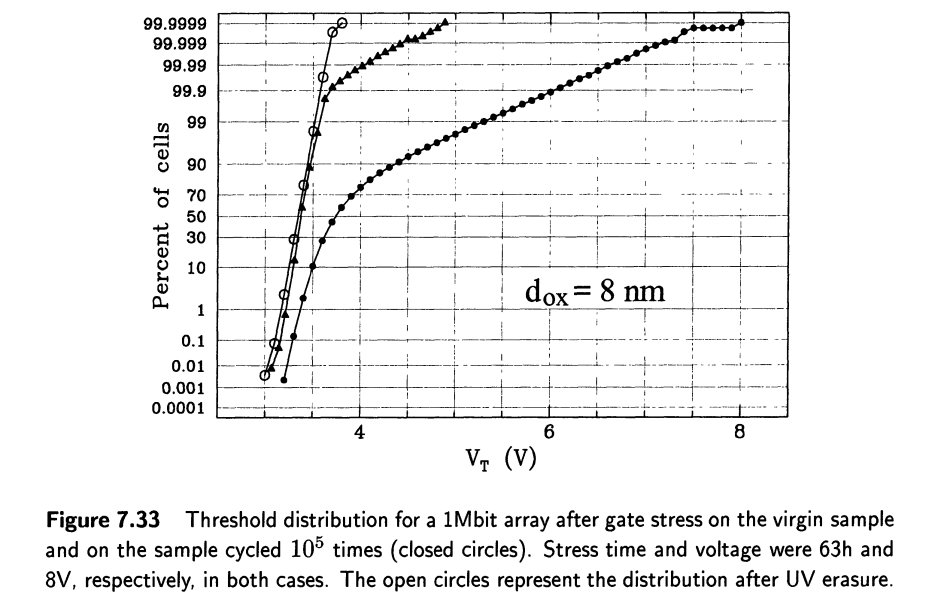
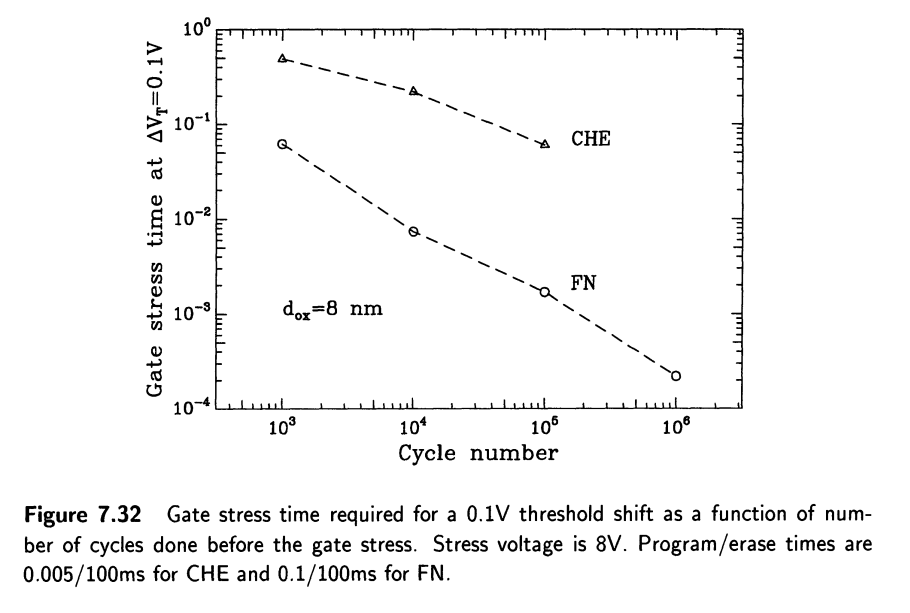
在第7.5.1节中，我们讨论了编程/擦除循环产生的氧化物电荷对单元性能的影响，重点关注了写入时间。然而，最关键的氧化层退化的可靠性问题与数据保留有关。

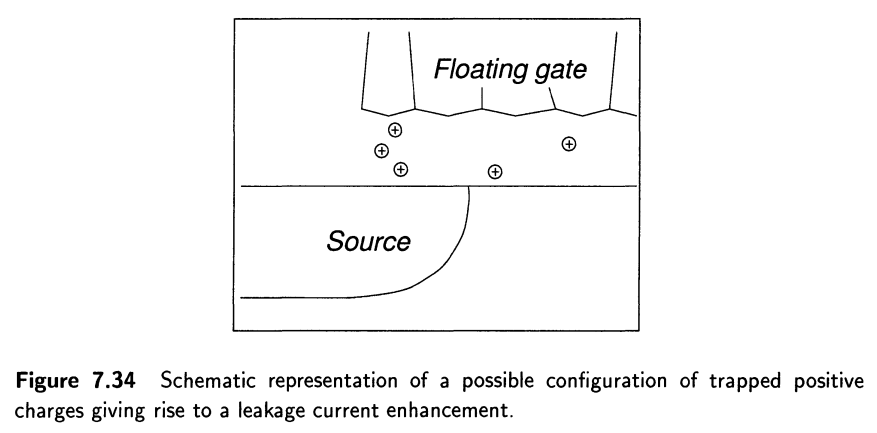


已知薄氧化物上的高场应力会增加低电场的电流密度;在低场时，导致IV曲线与理论Fowler-Nordheim特性明显偏差的多余电流分量被称为应力感应泄漏电流（SILC）。SILC显然与应力诱导的氧化物缺陷有关，就传导机制而言，它被归因于陷阱辅助隧穿（详见第4章）。控制SILC的主要参数是应力场、应力期间注入的电荷量和氧化物厚度。对于固定应力条件，泄漏电流随着氧化物厚度降低到10nm以下而强烈增加。

图7.31中比较了氧化物电流作为强应力薄氧化物电场函数的定性行为。在高场下，电流减小，这是导致长时间循环后擦除时间增加的众所周知的现象。在低场下，会出现相反的情况：随着应力时间的增加，电流的增加远高于新氧化物的电流

这种效应可以在闪存单元中观察到，作为循环后对低压栅极应力的增强敏感性。7.32显示了具有薄隧道氧化物的单个处理单元对栅极应力的敏感性，作为编程/擦除周期的函数：增加周期数会恶化栅极应力的影响。该图还报告了使用双Fowler-nordheim写入方案循环的单元的结果，显示了由于更高的场和该写入方案固有的双向应力而增强的退化。





再一次，主要关注点不是来自一个典型的单元，而是来自分布的尾部，如图所示7.33。其中比较了循环前后在1mb阵列上进行的低压栅极应力实验结果。不同单元的Vt位移的不同幅度可以再次用负责隧穿电流增强的氧化物陷阱的随机空间分布来解释，就像不稳定位一样;图7.34给出了可能会局部增强泄漏电流的正电荷协调示意图。Yamada等人报告的异常SILC电流的不稳定行为加强了与不稳定位现象的对比。

SILC对闪存可靠性的第一影响可能是增加编程干扰灵敏度。实际上，在编程干扰的情况下，电场足够高，以至于落在图中电流交叉点附近的区域7.31，因此在循环前后的编程干扰特性没有发现很大差异。

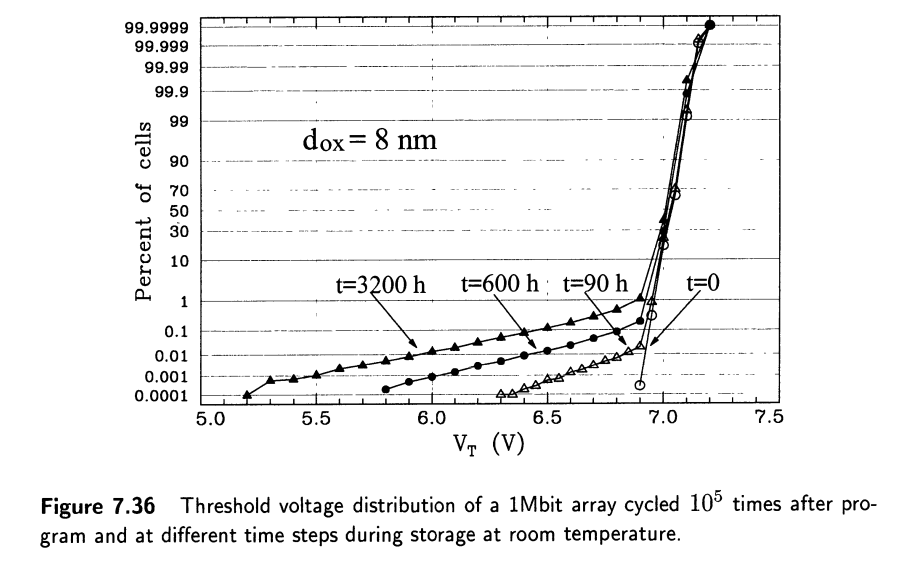
SILC对读取干扰和数据保留的影响与隧道氧化物厚度密切相关对于非常薄的隧道氧化物（8nm以下），即使在与典型的读取或数据存储条件一样低的电场值下，SILC也不可忽略。读取模式下擦除单元的隧道氧化物中的电场在2-3MV/cm范围内，而对于未偏置的编程单元，其在1-2MV/cm范围内；在这种情况下，循环存储单元必须将其数据保留数年。许多论文已经发表了关于在闪存中循环后的读取干扰的论文，这些论文基于使用非常薄的隧道氧化物的FN编程。

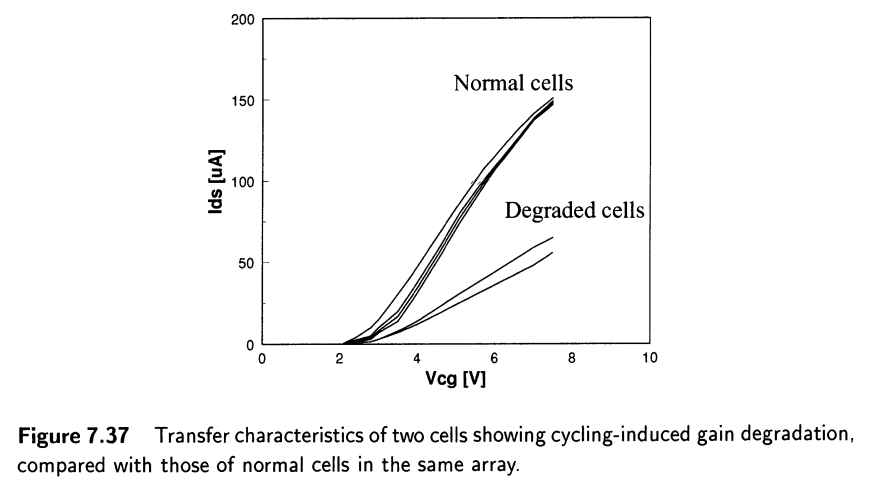
根据电容上获得的SILC数据，Runnion等人[18]已经表明，在相当于10^5个周期的应力后，5-7nm范围内的隧道氧化物泄漏在很大程度上超过了10年数据保留或读取干扰抗扰性的要求。

如果这是隧道氧化物缩放的极限，人们可以从给出平均SILC值的电容中估计，Yamada等人[16]已经表明，由于单个单元表现出更高的SILC，真实内存上的情况可能会更糟（图7.35）.

不过，用于设置隧道氧化物的最小可靠厚度，我们不能参考典型单元，但必须为分布的尾部提供裕度；图7.36显示了10 ^ 5个周期后，在具有8nm隧道氧化物厚度的1mbit阵列上进行的室温保持测试的结果：虽然阵列几乎没有任何可检测的阈值偏移，但有一个单元尾部存在松散电荷。

循环后的数据保留问题肯定会限制非常薄的氧化物的隧道氧化物厚度缩放。泄漏单元的数量变得如此之大，以至于即使是错误纠正技术也无法解决该问题.





7.5.3.3增益退化。

我们在第7.5.1节中看到，一个适当设计的存储单元在10^5个周期后不会出现任何跨导衰减。然而，由于增益下降，单个比特可能在耐久性测试中失败。

图7.37显示了两个退化位的I-V特性以及未通过循环修改的同一阵列中其他位的正常特性。这种退化位引起的故障模式可能不同。它们会导致擦除失败，因为它们会强制整个阵列被过度擦除。它们可能会导致编程失败，因为它们的编程性能也会降低。它们可能在高温保持测试中失败，因为烘焙部分恢复了其原始特性，并且在烘焙后位不再读取为“0”。

跨导降低是由于热空穴产生的界面态引起的。有不同的机制可以导致异常高的热空穴注入速率，从而导致单个位的跨导降低。

空穴最可能的来源是擦除本身。如第4章所详细讨论的，在擦除过程中，源区的带间隧穿产生空穴电子对；空穴向衬底漂移，并从反向偏置源结耗尽区的电场中获得能量：通道上方隧道氧化物中的电场有利于热空穴的注入。源结设计旨在最小化电场并防止雪崩倍增，但其加速了氧化物退化。此外，电路设计通过限制擦除时提供给源的电流，有助于最小化雪崩倍增的风险。因此，在精心设计的存储器中，热空穴的产生被最小化，并且循环不会降低存储器单元的跨导。

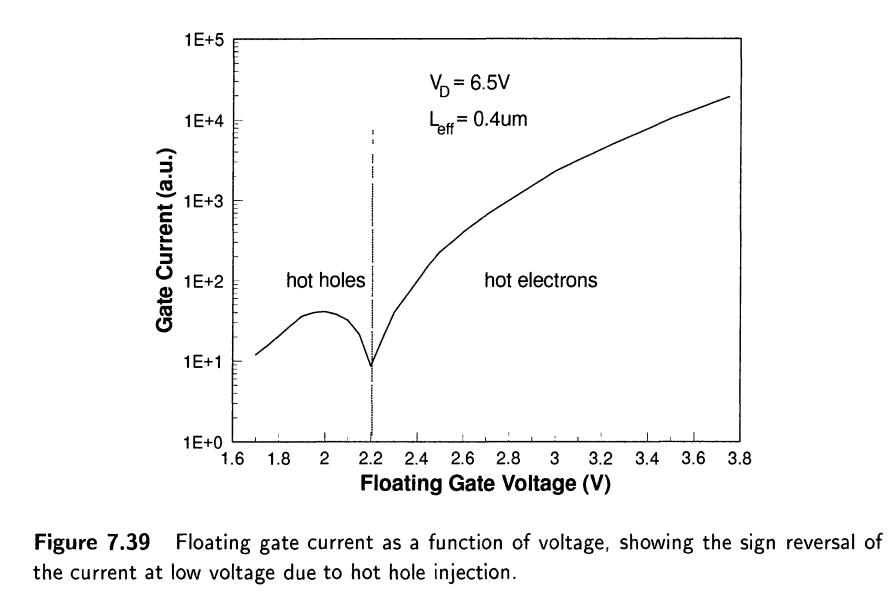
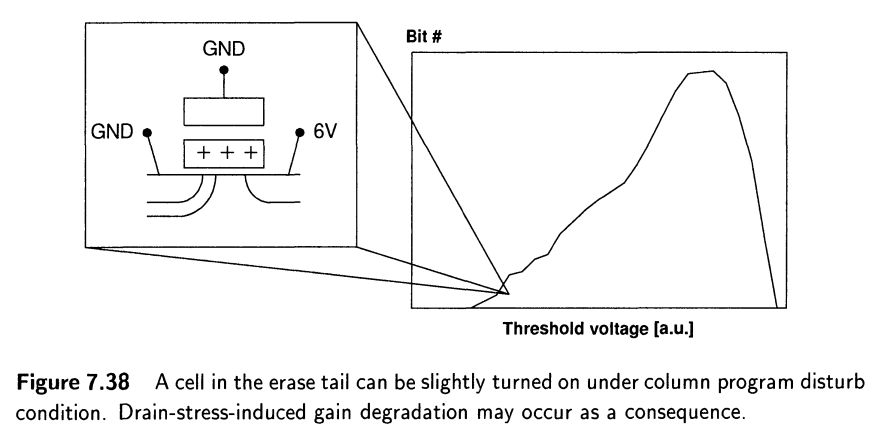
然而硅缺陷或污染可以局部显著降低源结击穿电压；因此，由于局部增强的结电场，缺陷单元在擦除时可能会出现比平均值高得多的热空穴生成。源极电流控制对于防止缺陷单元中的雪崩倍增无效，因为它限制了阵列的总电流，但无法避免单个电池吸收的电流比平均值高出几个数量级。该模型已通过擦除操作期间退化位的聚光发射得到实验验证。

工艺和掺杂剖面优化可以有效地最小化与上述机制相关的耐久性故障率。此外负栅极擦除方案通过显著降低源结反向偏置大大有助于解决该问题.

第二种机制与编程操作有关，它可以在单个单元中引起增强的热空穴注入，是编程干扰和过擦除的复杂组合。

它可能会影响电擦除后电压最低的单元（即尾端位），这是由于这些单元在同一位线上的其他单元编程时所经历的漏极应力（图7.38）。

擦除Vt分布最低边缘的位使浮栅带正电。在漏极应力条件下，浮栅中的正电荷和漏极耦合效应的组合可以稍微开启该位，即使栅极接地。高漏极电压和低浮栅电压是热空穴注入的最有利条件（图7.39）。漏极应力诱导的热空穴注入在每个周期产生的界面损伤的累积效应最终导致单元跨导降低。



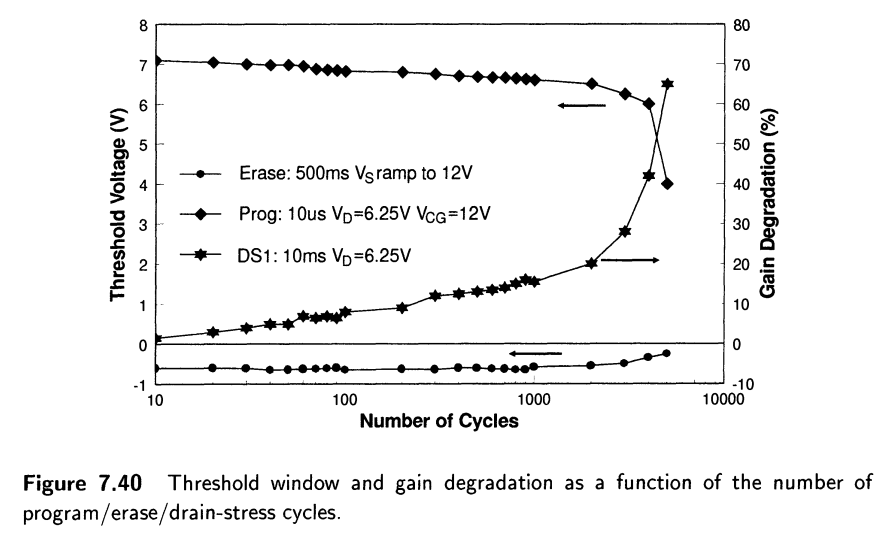
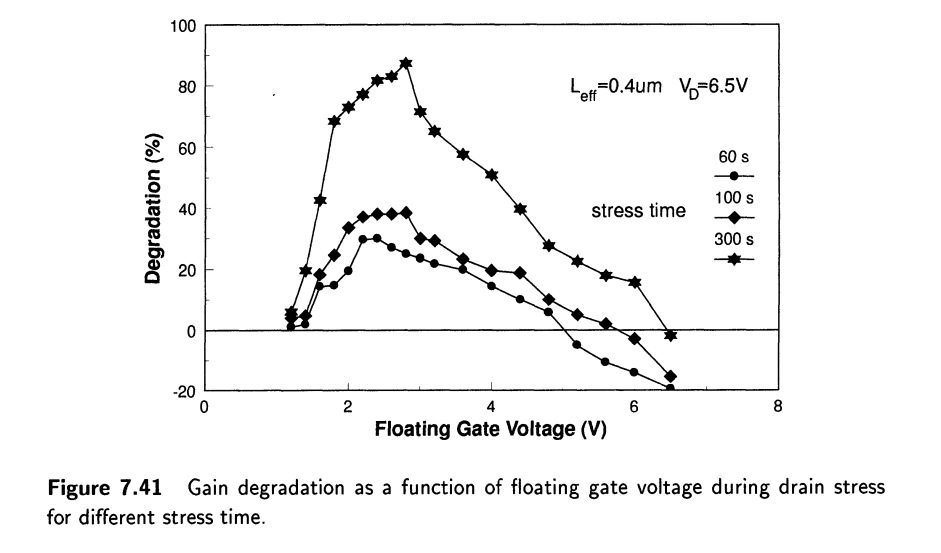


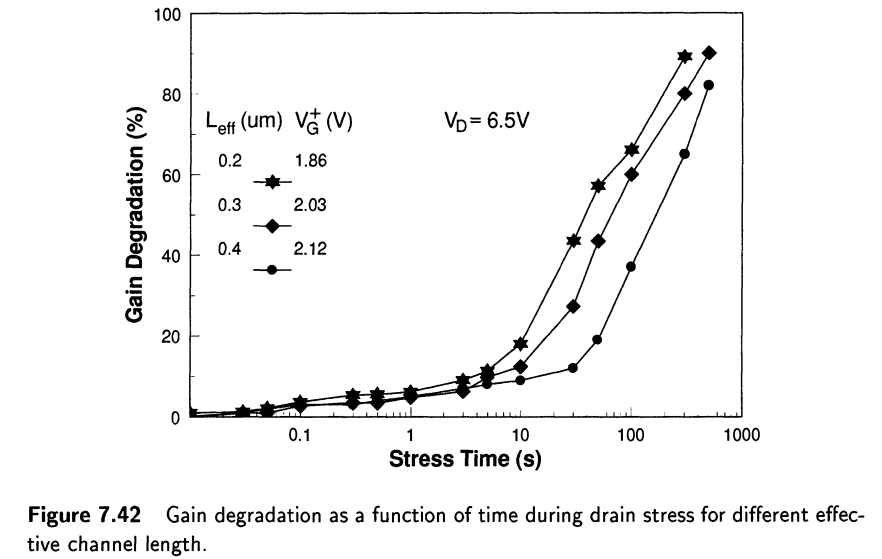
图7.40显示了在有意过度擦除的单元上进行耐久性测试时，在每个周期引入漏应力（Va=6.25Vt=10ms）的显著效果。人们对接触浮栅器件在不同偏置条件下的退化进行了深入研究。浮栅电压等于或略高于栅极电流的空穴和电子分量相互补偿的值时，退化速率（图7.41）最快，因此图中的净电流归零（图7.39）。

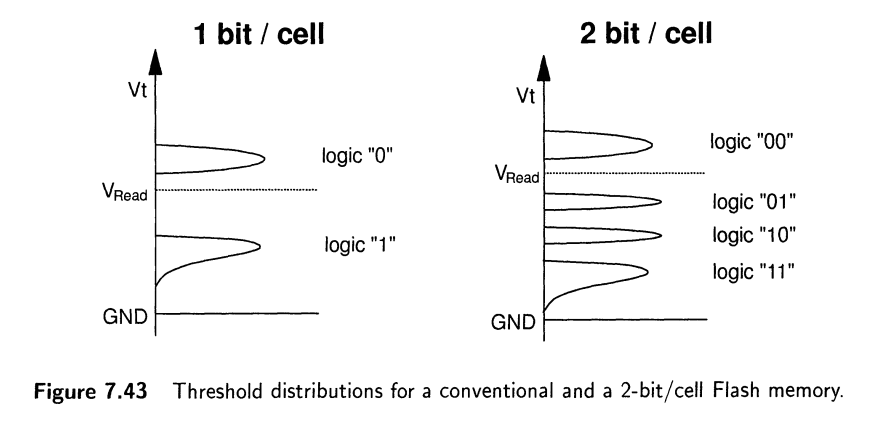
在单单元循环实验中再现的应力条件对应于擦除Vt分布尾部最低边缘的位的极端情况。

它们可能在正常工作条件之外，这取决于耗尽位的最小可接受V和漏极电压限制。相反，这些条件很可能对应于一些自收敛方案的典型情况，这些方案用于回收耗尽比特，因此其可靠性非常值得怀疑。

器件尺寸（Leff）和操作条件（Vd、Vfg）对性能降低速率有非常关键的影响。受这种性能降低速率的影响，Leff较短的单元在应力时间和循环方面更快（图742）。为了缩小单元尺寸，必须仔细考虑这一机制：必须优化存储单元架构，例如结设计或氧化物对热载流子应力的鲁棒性，以减少漏极应力退化的影响。







7.6多位存储可靠性

基于模拟编程非易失性存储单元,每个单元存储一个以上位的概念是已被证实可能的。

在标准数字闪存中，每个单元存储一位和两个Vt范围，对应于两个逻辑状态1（已擦除单元）和0”（已编程单元） . 从概念上讲，没有什么可以阻止将更多Vt级别定义为不同的单元状态，以便在每个单元中存储更多位：4Vt级别允许存储2位/单元，8Vt级别允许存储3位/单元等等。一般规则是需要2^n个级别Vt来存储n位/单元。

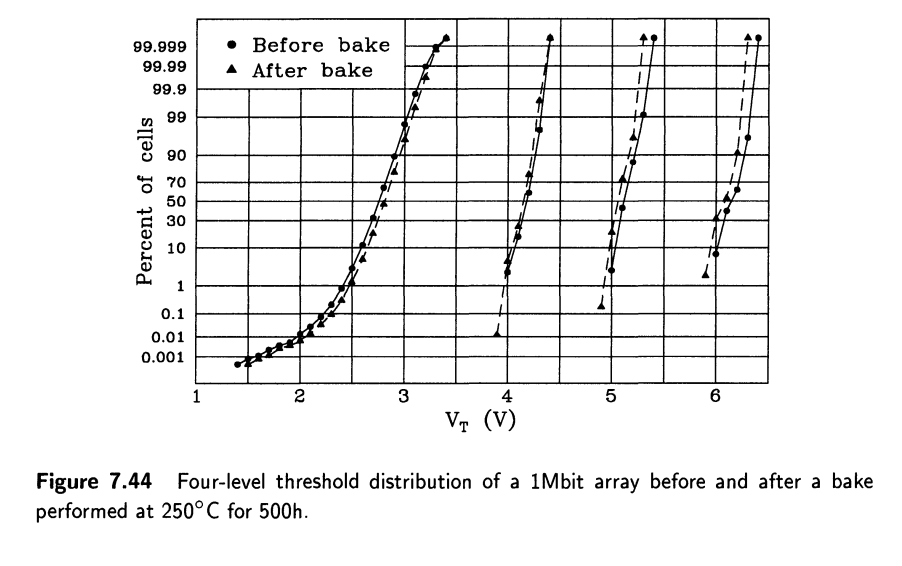
在图7.43中，比较了标准和2位/单元闪存的Vt分布。分配电平的Vt范围在较低的一侧受到限制，因为所有阈值必须高于最小正值，以避免读取或编程期间的位线泄漏。在上侧，Vt范围受第二高电平必须为0的条件限制。在读取过程中低于栅极电压0.5~1V，以便与最高电压正确区分。由于可靠性问题（读取干扰），尤其是在低压设备中，由于电压升压远高于Vcc的缺点，读取电压远高于6V是不现实的。因此，定义更多级别需要更紧密的Vt分布和级别之间更少的裕度。通过适当的编程算法，可以大大减少编程级别的宽度。

最精确的方法是阶梯栅极电压斜坡，由一系列具有恒定漏极电压的编程脉冲组成，每个脉冲的栅极电压增加恒定量。每个脉冲后都有一个程序验证操作。在低栅极电压下的初始阶段后，每个脉冲后的Vt：偏移等于栅极电压步长，因此可以获得相同阶数的Vt分布，可以实现非常小的分布，但以编程速度为代价，因为需要大量的编程/验证步骤。类似的考虑也适用于噪声裕度，通过使用高精度电路可以降低噪声裕度，但代价是访问时间。必须在准确性和速度之间进行最佳权衡，这会导致数据保留和干扰的可靠性裕度较小（约数百mv）。

在目前的知识状态下，没有预见到多位闪存的特定故障机制，但必须仔细考虑之前提出的每种故障模式的影响。

就编程/擦除循环导致的性能下降而言，在处理多位存储时不应该出现大的问题。除了Vt窗口增加导致通过隧道氧化物的电荷量略大这一小影响外，与传统存储器一样，erase time随着周期数的增加而增加。在编程中也有类似的考虑，另外需要考虑的是，精确编程需要较低的Vg步长，这被认为是导致注入效率下降的一个关键条件。

增益退化和不稳定擦除问题与传统内存相同，至少在CHE编程内存的情况下是这样的。隧道电流的不稳定行为对于那些使用FN编程的内存可能是一个问题，因为在编程期间电流的突然增加可能导致过编程。



数据保留显然是多级存储的关键问题。内在行为与3位甚至4位/单元的要求兼容，如图7.44所示，其中报告了加速保留试验的结果。对于激活能高于0.6ev的泄漏机制，试验条件（250度下500小时）对应于100度下10年以上。值得注意的是，Vt偏移与编程Vt大致呈线性，因此电平之间的间距仅减少最大偏移的一小部分。通过与编程参考单元进行比较，可以有效利用这一事实执行读取操作，编程参考单元）经历与阵列单元相同的电荷损耗。第7.5.3.2节讨论的一个主要问题是编程/擦除循环后的单位数据丢失。这里的问题是，与V窗口增加相关的，存储期间隧道氧化物中的裕度减少和更高的电场。虽然在传统闪存中，故障前的电场可能非常小，但在多级闪存中，编程到最高电平的电池的电场必然要高得多。

预计对编程干扰的敏感性提高有三个原因：余量减少、Vt窗口最大和编程时间增加。最坏的情况是编程到最高水平的单元上的列干扰，因为相对于中间状态的Vt偏移较大。此外，编程漏极电压与传统闪存中相同，但编程时间显著增加。相反，行干扰不应受到编程时间增加的影响，至少对于阶梯编程算法而言，因为在大多数时间内，栅极电压相当低，只有最后几个脉冲会导致干扰。

主要的可靠性问题是读取干扰，这会影响处于最低Vt状态的单元。除了裕度降低外，相对于传统闪存的更高读取电压对读取干扰抗扰性有很大影响，尤其是对于非常薄的隧道氧化物。对于具有8nm隧道氧化物的原始单元，读取电压应低于7V以保证10年的寿命，但最坏的情况显然是循环单元，其中SILC对氧化物厚度和读取电压都有更严格的限制。参考数据表明，即使氧化层厚度为10nm，在10个编程/擦除周期后，6.5V的读取电压也能够在几天的时间尺度内在一些异常单元上诱导数百mv的Vt偏移。即使这种异常SILC发生的概率非常低（ppm级或更低），在大密度多级闪存中，也可能对可靠性产生很大影响。

从以上考虑，我们得出结论，尽管多层闪存具有良好的内在行为，但其实际利用将受到单位故障对产品可靠性的影响的限制。最有可能的是，多级概念最初将用于容错应用（音频、视频）或与纠错技术相结合。

7.7 结论

随着人们对闪存可靠性问题的理解，闪存在半导体市场上的成功也与日俱增。在其推出的早期阶段，系统制造商普遍关注从众所周知的EPROM转向有吸引力但刚刚诞生的闪存，这种关注主要是关于可靠性和可制造性。接下来的几年不知何故证实，掌握闪存生产并非易事，在90年代上半年，闪存市场曾两次陷入供应短缺。现在，闪存作为一种成熟可靠的技术被广泛接受。

认识到隧道氧化物质量的关键重要性，并学习如何改进和监测隧道氧化物质量，对于实现工业制造标准和可靠性标准至关重要。

对失效机制的广泛研究为改进存储单元和产品设计奠定了坚实的知识基础。

导致编程/擦除循环中器件性能磨损的内在退化机制占主导地位；通过适当优化单元结构，可以将其影响降至最低，从而将耐久极限推到10^5-10^6范围内。

已经确定并深入研究了一种相当棘手的故障机制，不稳定擦除：解决方案已在工艺和电路层面上得到解决。

应力引起的泄漏电流对数据保留的影响将限制隧道氧化层厚度的可扩展性，除非使用纠错技术或系统内数据刷新数据来解决该问题

证明了多级存储的可行性，评估每个单元存储超过一位的闪存的可靠性将是未来几年的挑战。