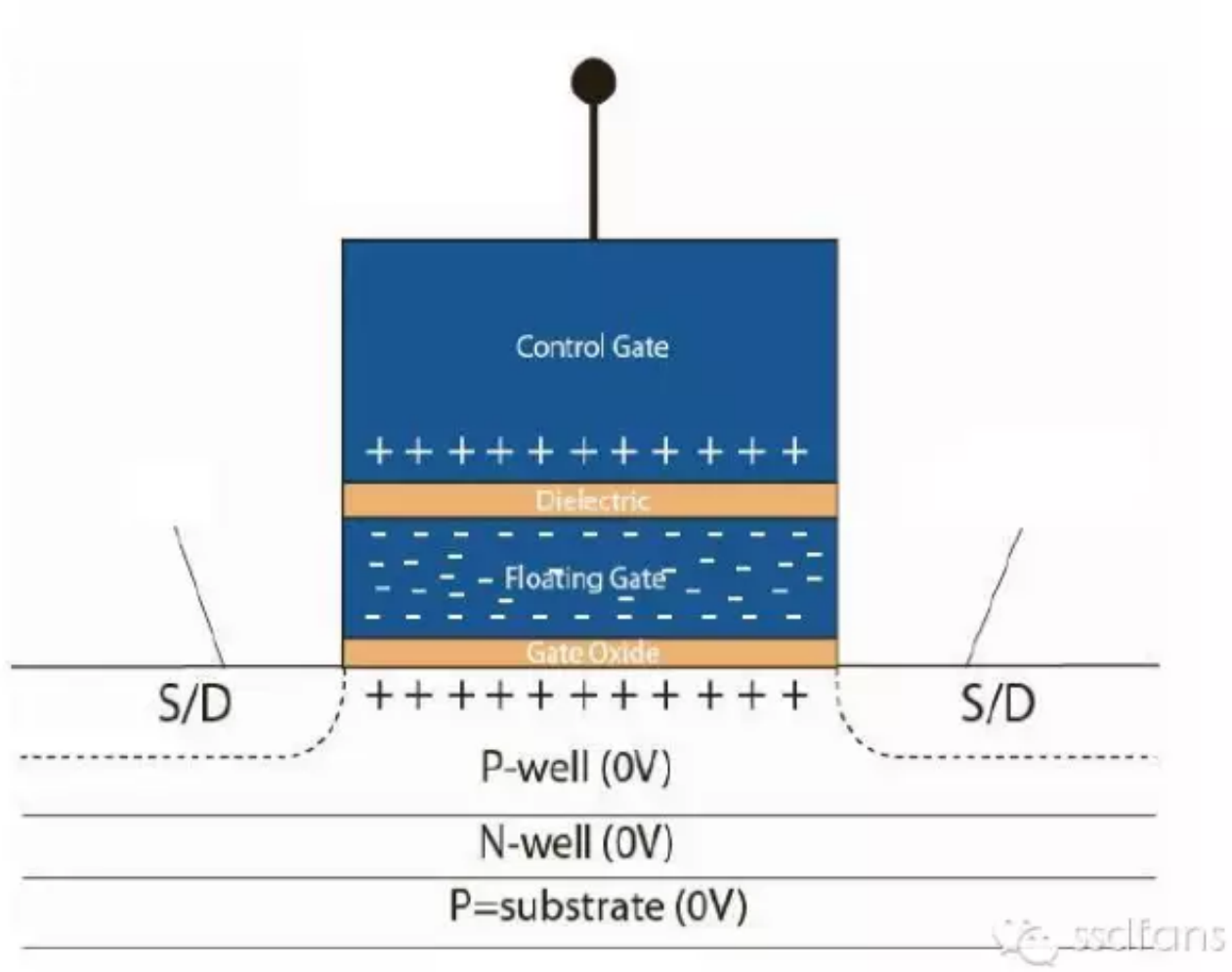
闪存基础

Original 2016-05-16 蛋蛋 ssdfans

目前绝大多数SSD都是以NAND FLASH为存储介质的。SSD工作原理很多都是基于NAND FLASH特性的。比如，NAND FLASH在写之前必须先擦除，而不能覆盖写，于是SSD 才需要垃圾回收（Garbage Collection，或者叫 Recycle）；NAND FLASH 每个块（Block）擦写次数达到一定值，这个块就不能用了（数据丢失，或者写

入不了），所以SSD 固件必须做 Wear Leveling，让数据平均写在所有块上，而不是盯着几个块拼命写（不然没几天SSD就报废了）。还有类似很多例子，SSD很多实现都是在为NAND FLASH服务的。所以，欲攻SSD，NAND FLASH首当其冲。 NAND FLASH是一种非易失性存储器，也就是说，掉电了，数据也不会丢失。NAND FLASH基本存储单元 (Cell) 是一种类NMOS的双层浮空栅 (Floating Gate)

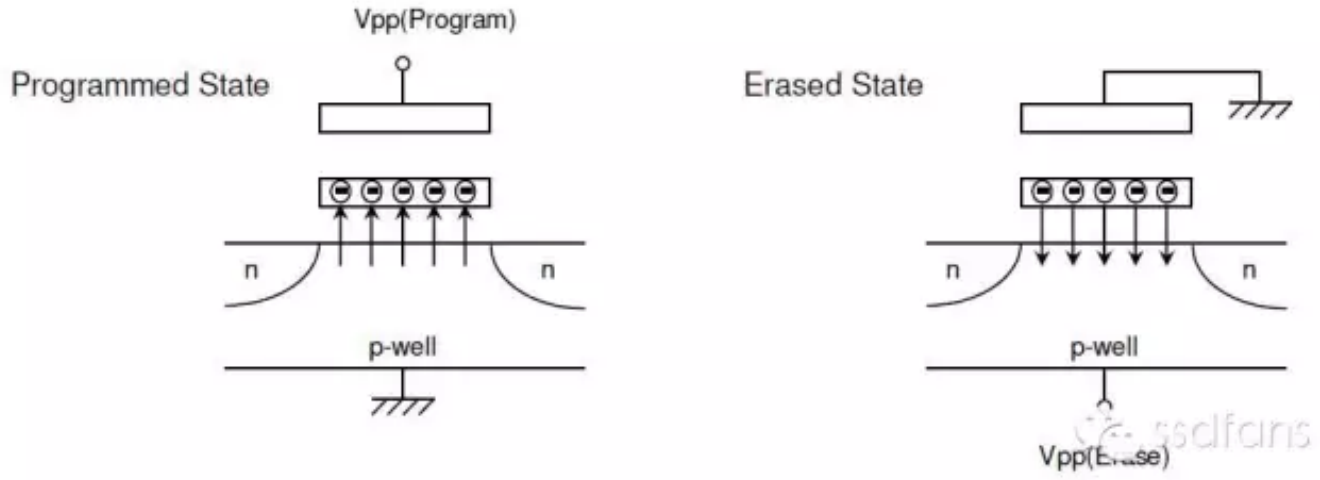
MOS管组成, 如下图所示。



在源极(Source)和漏极(Drain)之间电流单向传导的半导体上形成贮存电子的浮动栅（Floating Gate）。浮动栅包裹着一层硅氧化膜绝缘体。它的上面是在源极和漏极之间控制传导电流的选择/控制栅 (Control

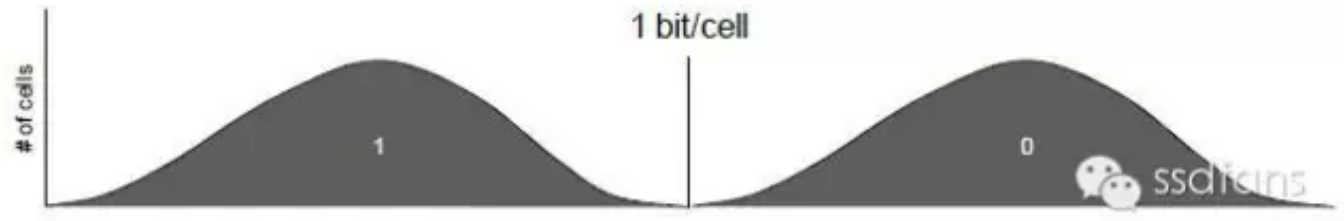
Gate)。里面的电子不会因为掉电而消失，所以NAND FLASH是非易失存储器。

写操作是在控制极加正电压，使电子通过绝缘层进入浮栅极。擦除操作正好相反，是在衬底加正电压，把电子从浮栅极中吸出来。如下图所示：

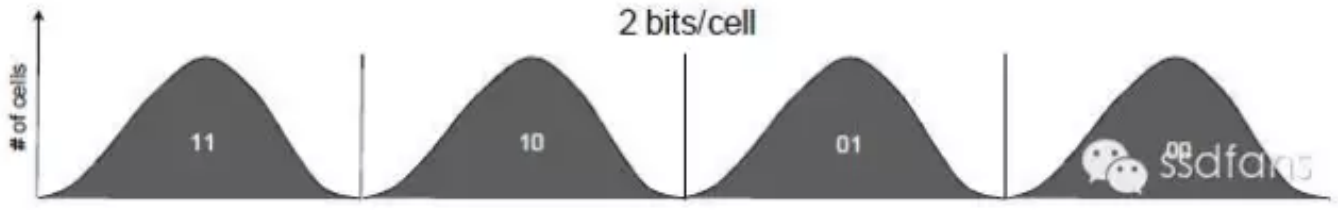


一个存储单元存储1bit数据的NAND FLASH，我们叫它为SLC (Single Level Cell)，2bit为MLC (Multiple Level Cell)，3bit为TLC (Triple Level Cell)。

对SLC来说，一个存储单元存储两种状态，浮栅极里面的电子多于某个参考值的时候，我们把它采样为0，否则，就判为1.

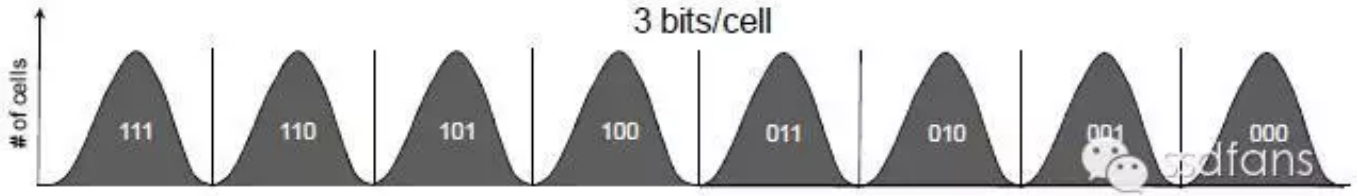


对MLC来说，一个存储单元存储四个状态，一个存储单元可以存储2bit的数据。通俗来说就是把浮栅极里面的电子个数进行一个划分，比如低于10个电子，判为0；11-20个电子，判为1；21-30，判为2；多于30个电子，判为3.



依次类推TLC，它的一个存储单元有8个状态，可以存储3bit的数据，它在MLC的基础

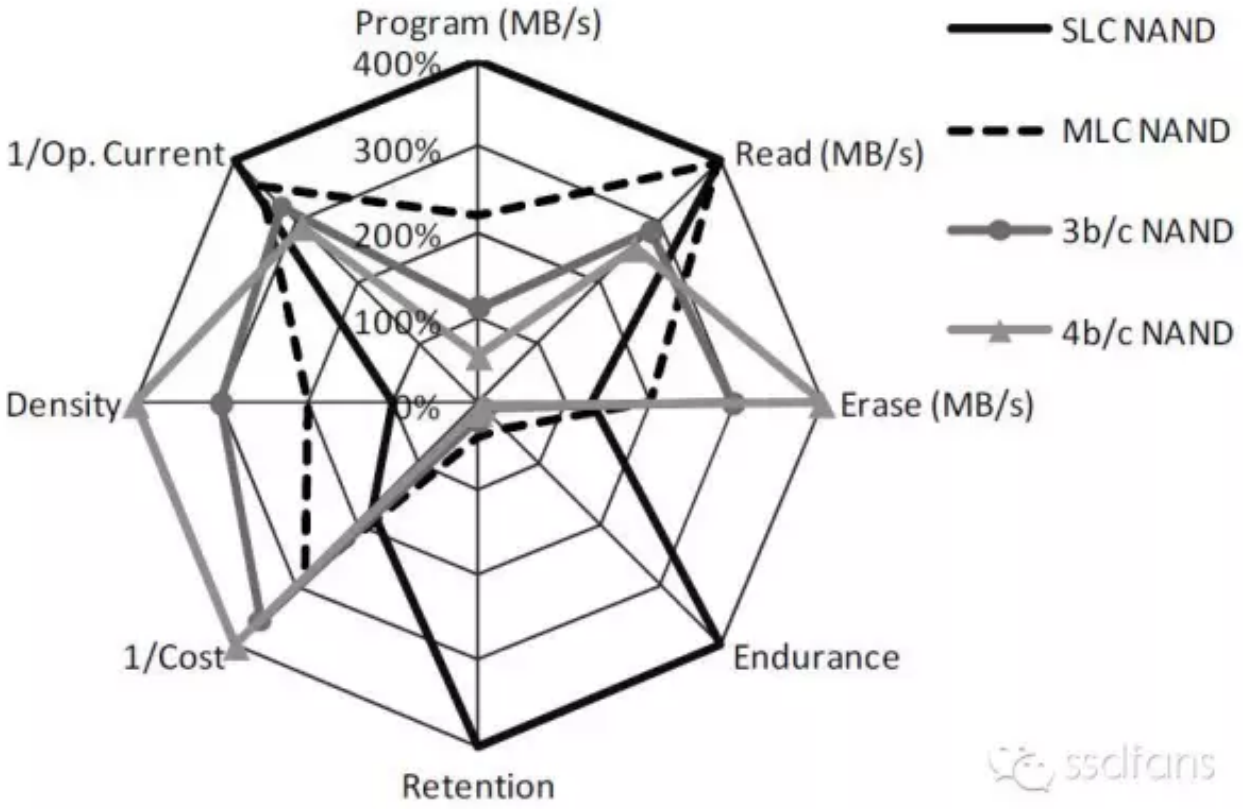
上对浮栅极里面的电子数又进一步进行了划分。



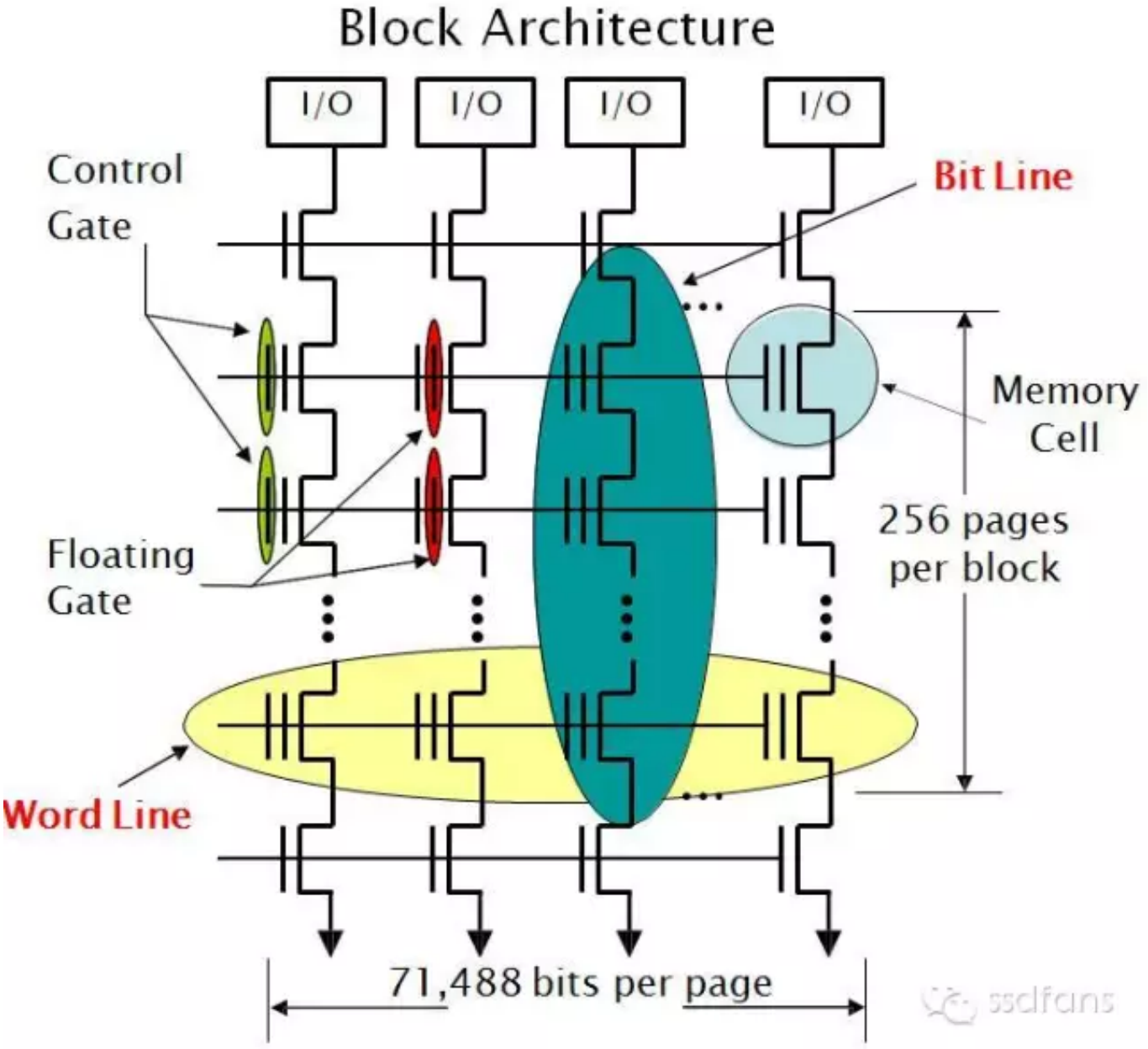
同样面积的一个存储单元，SLC，MLC和TLC，依次可以存储1,2,3bit的数据，所以在同样面积的LUN上，NAND FLASH容量依次变大。

但同时，一个存储单元电子划分的越多，那么在写入的时候，控制进入浮栅极的电子的个数就要越精细，所以写耗费的时间就加长；同样的，读的时候，需要尝试用不同的参考电压去读取，一定程度上加长读取时间。所以我们会看到在性能上，TLC不如

MLC,MLC不如SLC.



NAND FLASH就是由成千上万这样的存储单元按照一定的组织结构组成的。

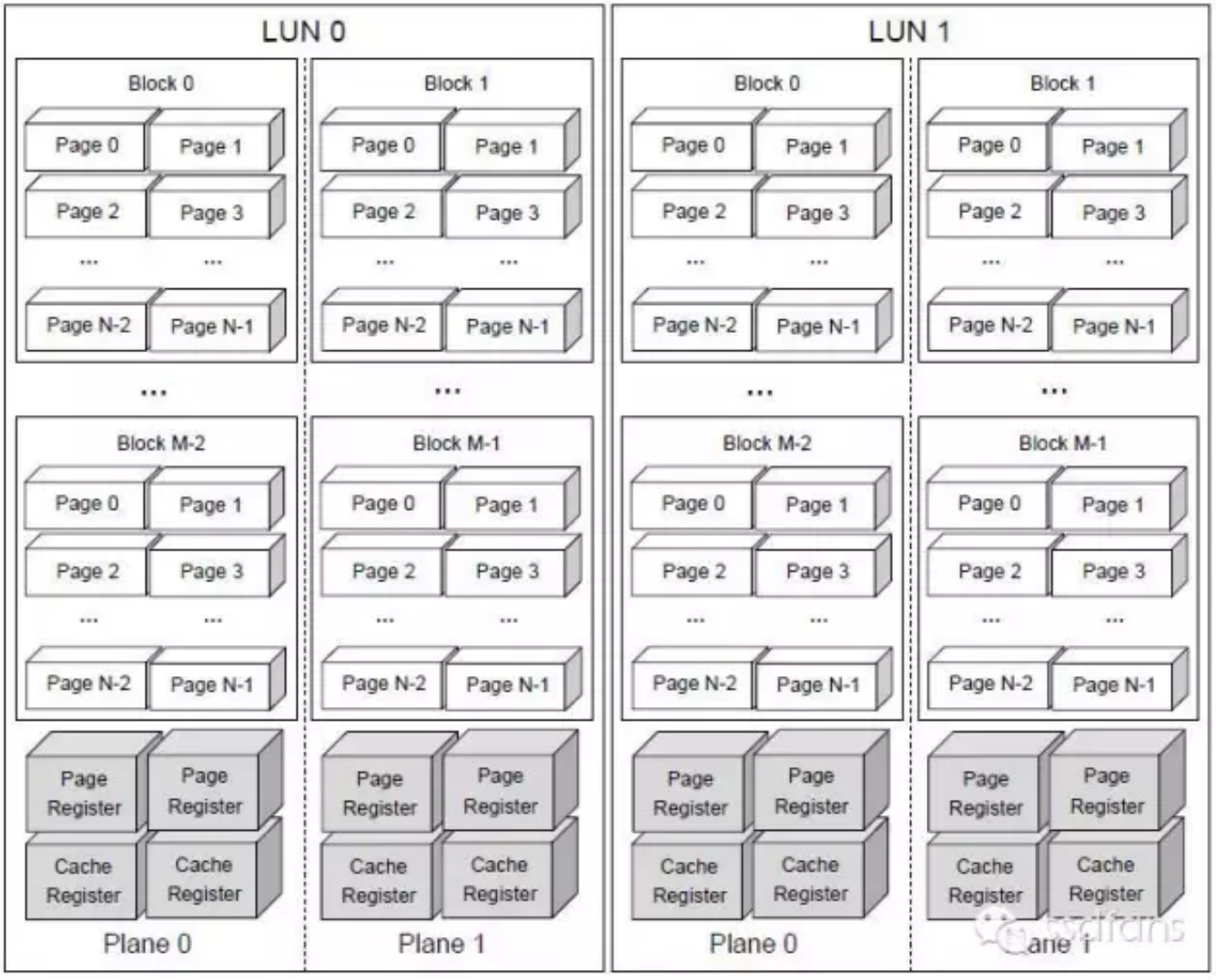


上图是一个FLASH Block的组织架构。一个WordLine对应着一个或若干个Page，取决于SLC,MLC或者TLC。对SLC来说，一个WordLine对应一个Page；MLC则对应2个Page，这两个Page是一对：Lower Page

和Upper Page；TLC对应3个Page。一个Page有多大，那么WordLine上面就有多少个存储单元（Cell），就有多少个Bitline。一个Block当中的所有这些存储单元（Cell）都是共用一个衬底的。

一个NAND FLASH内部存储组织结构是这样的：一个Device有若干个DIE（或者叫LUN），每个DIE有若干个Plane，每个Plane有若干个Block，每个Block有若干个Page。每个Page对应着一个Wordline，

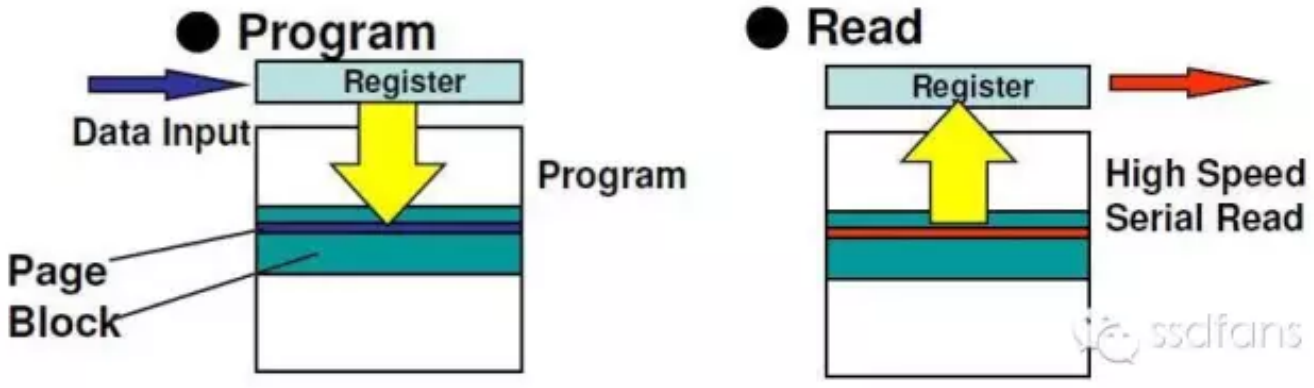
由成千上万个存储单元构成。



DIE/LUN是接收和执行FLASH命令的基本单元。上图中，LUN0和LUN1可以同时接收和执行不同的命令。但在一个LUN当中，一次只能执行一个命令，你不能对其中的某个Page写的同时，又对其他Page进行读访问。

一个LUN又分为若干个Plane，一般为1个或者2个，现在也有4个Plane的NAND了。每个Plane都有自己独立的Cache Register或者 Page Register，一般情况下，两个Register内容都是一样的，其大小等于一个Page的大小。Host在写某个Page的时候，它是先把数据从Host传输到该Page所对应Plane的Cache Register当中，然后再把整个Cache Register当中的数据写到NAND FLASH阵列；读的时候类似，它先把这个Page的数据从FLASH阵列读取到Page Register，然后再按需传给host。这里按需是什么意思？就是我们读取数据的时候，没有必要把整个Page的数据都传出来给Host，按需选择数据传输。但要记住，无论是从FLASH 阵列读数据到Page

Register，还是把Page Register的数据写入FLASH阵列，都是以Page为单位！



我们通常所说的FLASH读写时间，是不包含数据从NAND与HOST之间的数据传输时间。FLASH写入时间指是一个Page的数据从Cache Register 当中写入到FLASH阵列的时间，FLASH读取时间是指一个Page的数据从FLASH阵列读取到Page Register的时间。对现在的MLC NAND FLASH来说，写入时间一般为几百个微秒甚至几毫秒，读取时间为几十微秒。 NAND FLASH一般都支持Multi-Plane或者说Dual-Plane操作。

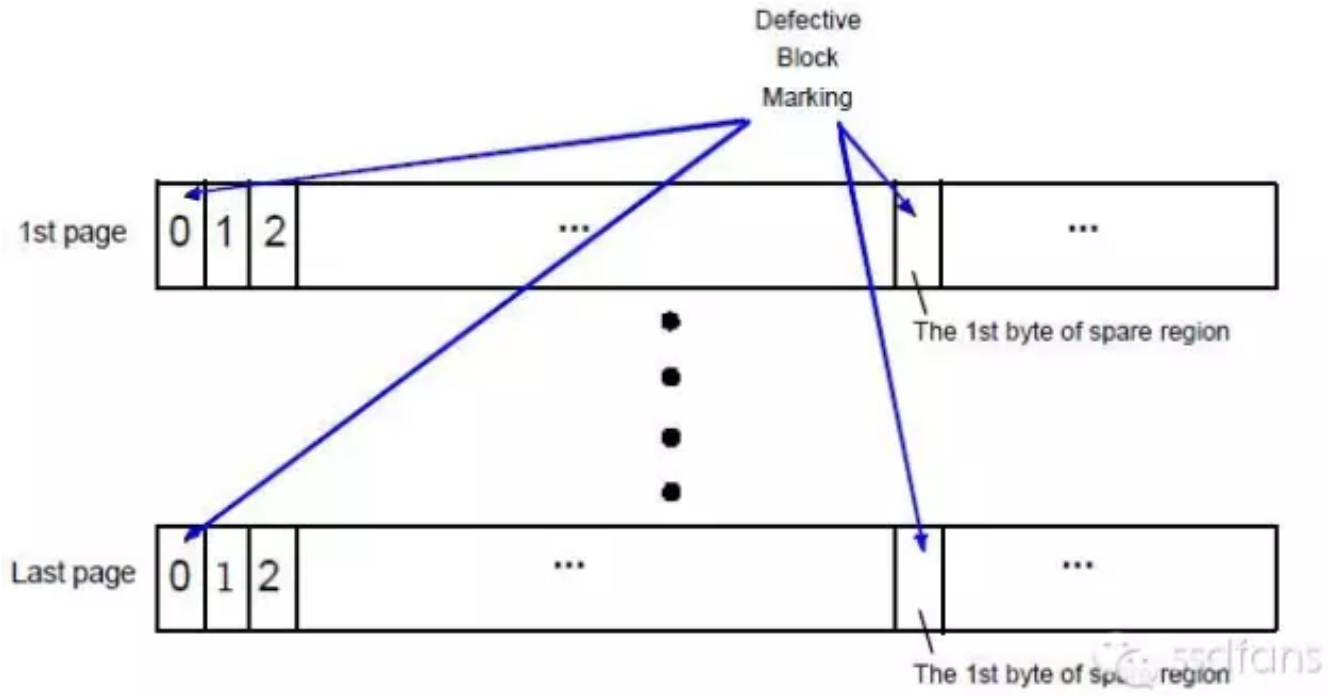
那么什么是Dual-Plane操作呢？对写来说，HOST先把数据写入到上第一个Plane的Cache Register当中，数据hold在那里，并不立即写入到FLASH阵列，等HOST把同一个LUN上的另外一个或者几个Plane上的数据传输到相应的Cache Register当中，再统一一起写入FLASH阵列。假设写入一个Page的时间为1.5ms，一个Page的传输时间为50us：如果按原始的Single Plane操作，写两个Page需要至少3ms+20us；但如果按照Dual-Plane操作，由于隐藏了一个Page的写入时间，写入两个Page只要1.5ms+20us，缩减了几乎一半的时间，写入速度几乎翻番。对读来说，使用Dual-Plane操作，两个不同Plane上的Page数据会在一个NAND读取时间加载到各自的Page Register当中，这样用一

个读取时间读取到两个Page的数据，读取速度加快。考虑读取时间和数据传输时间相当，假设都是50us，Single Plane读取传输两个Page需要50us*4=200us，Dual-Plane则需要50us*2+50us=150us，时间为前者的75%，读取速度也有大的提升。

NAND FLASH的擦除是以Block为单位的。为什么呢？那是因为在组织结构上，一个Block当中的所有存储单元（Cell）是共用一个衬底的（Substrate）。当你对某衬底施加强电压，那么上面所有浮栅极的电子都被吸出来了。每个NAND Block都有擦写次数的限制，当超过这个次数时，该Block可能就不能用了：浮栅极充不进电子（写失败），或者浮栅极的电子很容易就跑出来（比特翻转，0->1)，或者浮栅极里面的电子跑不出来（擦除失败）。这个最大擦写次

数按SLC,MLC,TLC依次递减：SLC的擦写次数可达十万次，MLC一般为几千到几万，TLC降到几百到几千。随着NAND FLASH工艺的不断进步（现在已进入1Xnm时代），NAND FLASH容量不断加大，但性能与可靠性却在变差。要克服NAND FLASH的这些不利因素，对SSD固件算法带来了更多更大的挑战。 FLASH Block不一定要达到寿命才不能用。一块FLASH，刚出厂的时候就会有坏块，这些坏块叫出厂坏块。有些厂商会在该Block的某几个Page当中加入坏块标记（如下图所示），用户在使用前，应该按照FLASH DATASHEET把这些坏块挑出来建立坏块表，避免以后使用这

些坏块。



也有一些FLASH厂商会直接告诉你哪些块是出厂坏块，这些信息存储在FLASH的某个地方，用户只需读取这些信息即可，无需对整个FLASH的所有Block进行坏块扫描。

用户在时候过程中，一个Block，即使未达到最大使用寿命，也有可能变坏。FLASH是允许有一定的坏块率的。质量好的

FLASH，坏块率是很小的；质量差的FLASH,坏块产生频繁。所以在挑选SSD的时候，尽量挑选知名主流的FLASH厂商生产的FLASH,质量有保证。

对MLC来说，擦除一个Block的时间大概是几个毫秒。 NAND FLASH的读写则是以Page为基本单元的。一个Page大小主要有4KB,8KB,16KB。对MLC或者TLC来说，写一个Block当中的Page，应该顺序写：Page0，Page1，Page2，Page3，...；禁止随机写入，比如：Page2，Page3，Page5，Page0，...，这是不允许的。但对读来说，没有这个限制。SLC也没有这个限制。 HOST是通过一系列FLASH命令与NAND通讯的。每个FLASH，都定义了其支持的命令，以MICRON 某型号的FLASH

为例，它定义了如下命令：

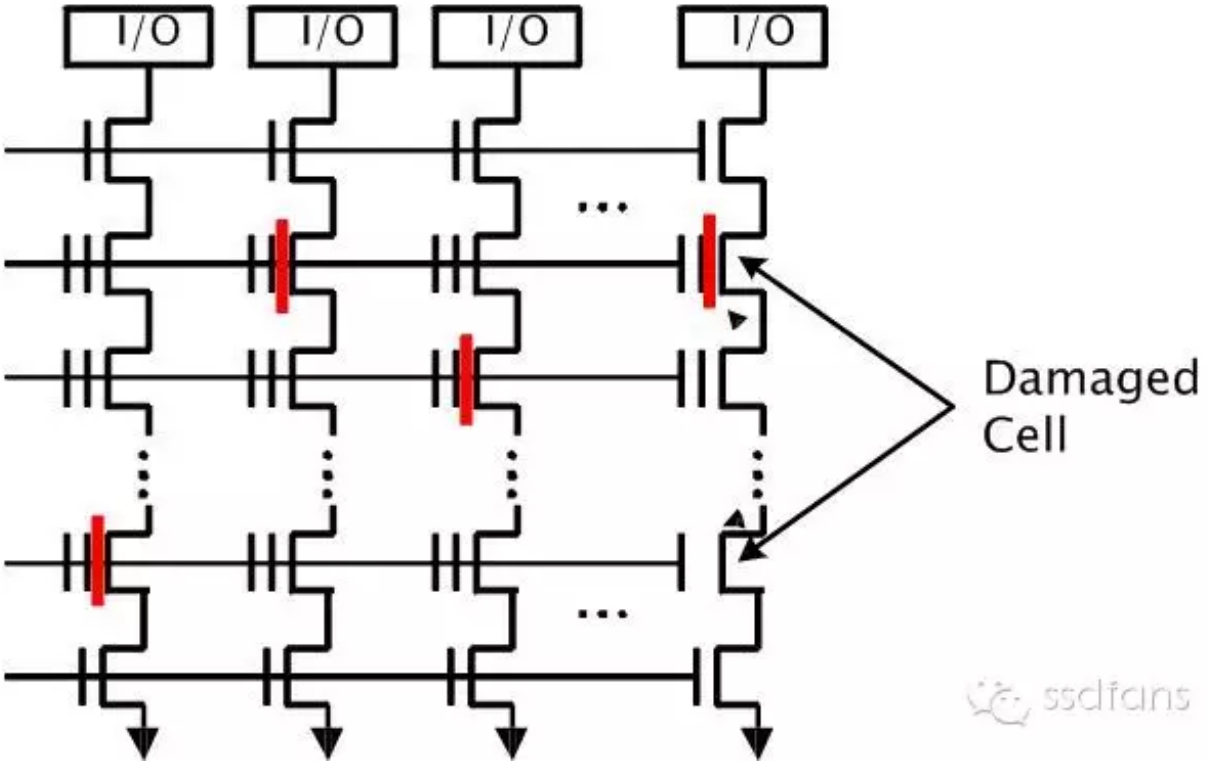
Command	Command Cycle #1	Number of Valid Address Cycles	Data Input Cycles	Command Cycle #2	Number of Valid Address Cycles #2	Command Cycle #3	Valid While Selected LUN is Busy[1]	Valid While Other LUNs are Busy[2]	Notes
Reset Operations									
RESET	FFh	0	–	–	–	–	Yes	Yes	
SYNCHRONOUS RESET	FCh	0	–	–	–	–	Yes	Yes	
RESET LUN	FAh	3	–	–	–	–	Yes	Yes	
Identification Operations									
READ ID	90h	1	–	–	–	–			3
READ PARAMETER PAGE	ECh	1	–	–	–	–			
READ UNIQUE ID	EDh	1	–	–	–	–			
Configuration Operations									
VOLUME SELECT	E1h	1	–	–	–	–			
ODT CONFIGURE	E2h	1	4	–	–	–			
GET FEATURES	EEh	1	–	–	–	–			3
SET FEATURES	EFh	1	4	–	–	–			4
Status Operations									
READ STATUS	70h	0	–	–	–	–	Yes		
READ STATUS ENHANCED	78h	3	–	–	–	–	Yes	Yes	
Column Address Operations									
CHANGE READ COLUMN	05h	2	–	E0h	–	–		Yes	
CHANGE READ COLUMN ENHANCED (ONFI)	06h	5	–	E0h	–	–		Yes	
CHANGE READ COLUMN ENHANCED (JEDEC)	00h	5	–	05h	2	E0h		Yes	
CHANGE WRITE COLUMN	85h	2	Optional	–	–	–		Yes	
CHANGE ROW ADDRESS	85h	5	Optional	11h (Optional)	–	–		Yes	5
Read Operations									
READ MODE	00h	0	–	–	–	–		Yes	
READ PAGE	00h	5	–	30h	–	–		Yes	6

Command	Com-mand Cycle #1	Number of Valid Address Cycles	Data Input Cycles	Com-mand Cycle #2	Number of Valid Address Cycles #2	Com-mand Cycle #3	Valid While Se-lected LUN is Busy[1]	Valid While Oth-er LUNs are Busy[2]	Note s
READ PAGE MULTI-PLANE	00h	5	–	32h	–	–		Yes	
READ PAGE CACHE SEQUENTIAL	31h	0	–	–	–	–		Yes	7
READ PAGE CACHE RANDOM	00h	5	–	31h	–	–		Yes	6,7
READ PAGE CACHE LAST	3Fh	0	–	–	–	–		Yes	7
Program Operations									
PROGRAM PAGE	80h	5	Yes	10h				Yes	
PROGRAM PAGE MULTI-PLANE	80h or 81h	5	Yes	11h				Yes	
PROGRAM PAGE CACHE	80h	5	Yes	15h				Yes	8
Erase Operations									
ERASE BLOCK	60h	3	–	D0h				Yes	
ERASE BLOCK MULTI-PLANE (ON-FI)	60h	3	–	D1h				Yes	
ERASE BLOCK MUL-TI-PLANE (JEDEC)	60h	3	–	60h	3	D0h		Yes	
ERASE SUSPEND	61h	3	–	–	–	–	Yes	Yes	
ERASE RESUME	D2h	–	–	–	–	–		Yes	
Copyback Operations									
COPYBACK READ	00h	5	–	35h				Yes	6
COPYBACK PRO-GRAM	85h	5	Optional	10h				Yes	
COPYBACK PRO-GRAM MULTI-PLANE	85h	5	Optional	11h				Yes	

不同的FLASH，所支持的命令有所差异。用户应该严格按照FLASH DATASHEET与FLASH通讯。

谈谈NAND FLASH的一些特点，或者说它作为存储介质面临的挑战。

1. Block具有一定的寿命，不是长生不老的。前面提到，当一个Block接近或者超出其最大擦写次数时，导致存储单元的永久性损伤，不能再使用。随着NAND工艺不断向前，这个擦写次数也变得越来越小。

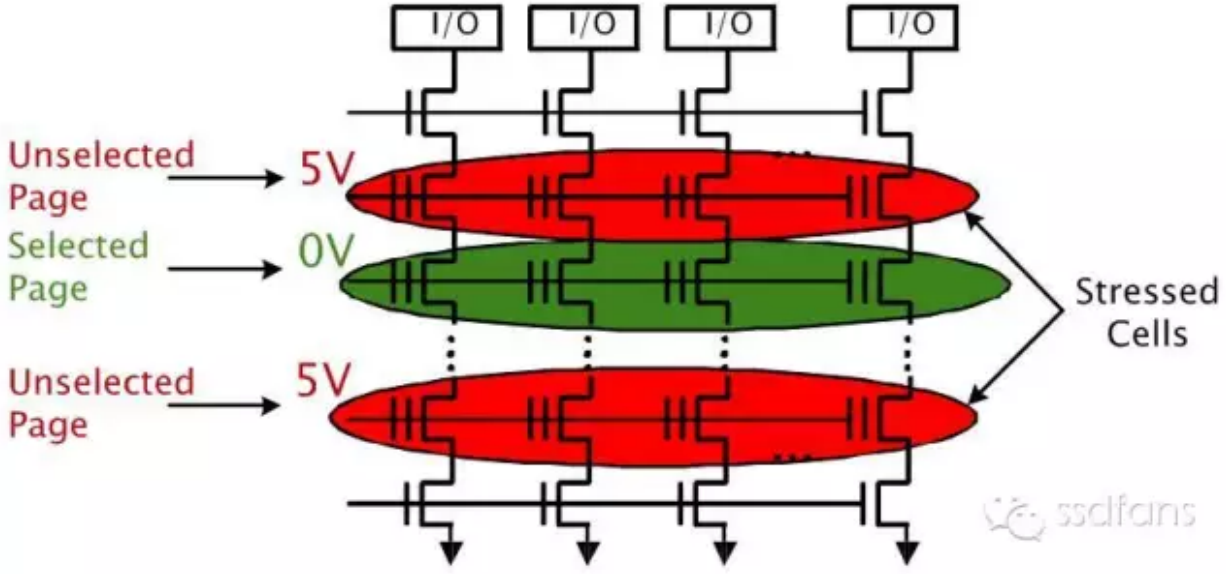


1. 在NAND当中的存储单元中，先天就有一些是坏掉的，或者说不稳定的。并且

随着NAND的不断使用，坏的存储单元越来越多。所以，用户写入到NAND的数据，必须有ECC保护，这样即使其中的一些比特发生反转，读取的时候也能通过ECC纠正过来。一旦出错的比特超过纠错能力范围，数据就丢失，对这样的Block，我们应该废弃不再使用。

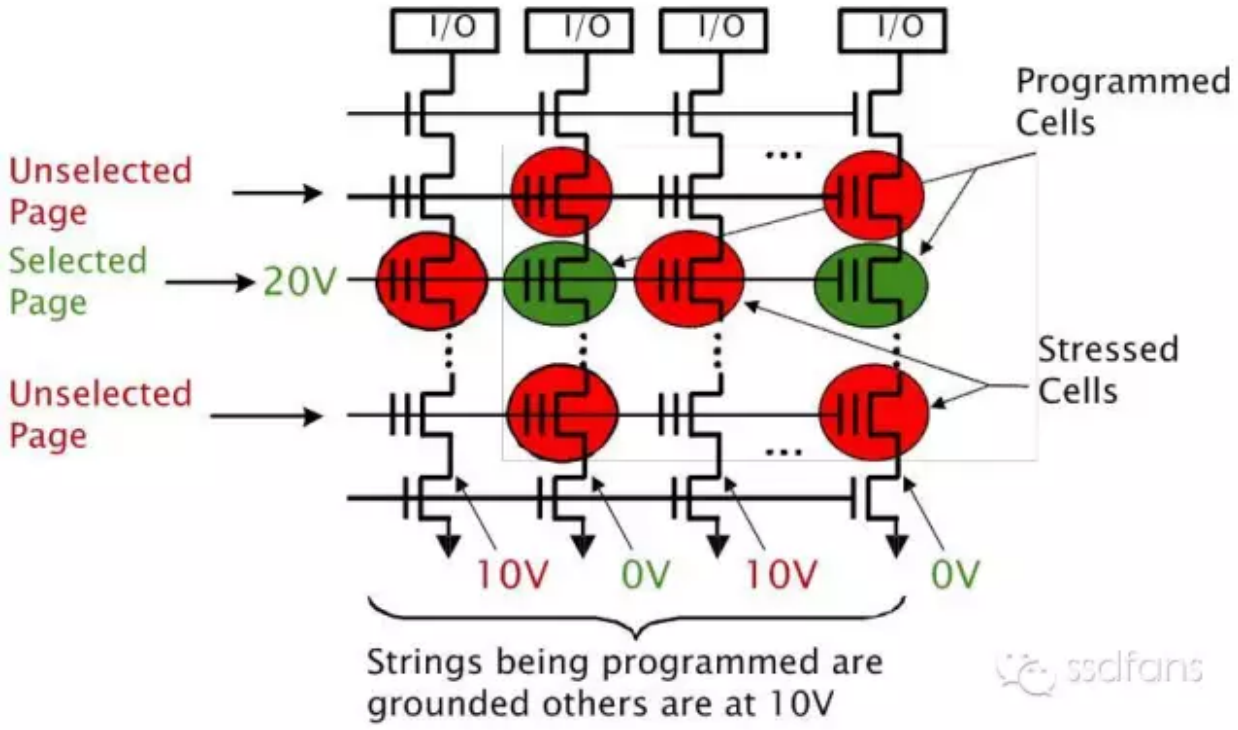
2. FLASH先天有坏块，也就是说有出厂坏块。并且，用户在使用的时候，也会新添坏块，所以用户在使用FLASH的时候，必须有坏块管理机制。
3. 读干扰（**Read Disturb**）。什么意思？从NAND读取原理来看，当你读取一个Page的时候，Block当中未被选取的Page控制极都会加一个正电压，以保证未被选中的MOS管是导通的。这样问题就来了，频繁的在一个MOS管控制极

加正电压，就可能导致电子被吸进浮栅极，形成轻微的Program。从而最终导致比特翻转。但是，这个不是永久性损伤，重新擦除Block还能正常使用。注意的是，Read Disturb影响的是同一个Block中的其它Page，而非读取的Page本身。



4. 写干扰（**Program Disturb**）。除了Read Disturb会导致比特翻转，Program Disturb也会导致比特翻转。还是要回到FLASH内部的Program原理

上来。



我们写一个Page的时候，数据0和1混合的。由于对擦除过的Block，其所有的存储单元初始值就 是1，所以Program的时候，只有写0的时候才真正需要Program。如上图所示，绿色的Cell是写0，需要Program的，红色的代表写1，并不需要Program。我们这里把绿色的Cell称之为Programmed Cells，红色的Cell叫Stressed Cells。写某个Page的时候，我们是在其 WordLine的

控制极加一个正电压(上图是20V),对于Programmed Cells所在的String，它是接地的，不需要Program Cell所在的String，它是接一正电压（上图为10V）。这样最终产生的后果是，Stressed Cell也会被轻微Program。与Read Disturb不同的是，Program Disturb 影响的不仅是同一个Block当中的其它Page，自身Page也受影响。相同的是，都是不期望的轻微Program导致比特翻转，都非永久性损伤，经擦除后，Block还能再次使用。

5. **电荷泄漏**。存储在NAND FLASH存储单元的电荷，如果长期不使用，会发生电荷泄漏。不过这个时间比较长，一般十年左右。同样是非永久性损伤，擦除后Block还能使用。

上面说的这些，是所有NAND面临的问题，包括SLC，MLC和TLC。对MLC来说，又有其特有的一些问题。

1. 正如前面提到的，MLC最大擦写次数变小。这样，就更需要Wear Leveling技术来保证整个存储介质的使用寿命。
2. 对MLC来说，一个存储单元存储了两个比特的数据，对应着两个Page：Lower Page和Upper Page。假设Lower Page先写，然后再写Upper Page的过程中，由于改变了整个Cell的状态，如果这个时候掉电，那么之前写入的Lower Page数据也丢失。一句话，写一个Page失败，可能会导致另外一个Page的数据损坏。
3. 前面说到，不能随机写。不能先Program Upper Page，然后再

Program Lower Page，这点就限制了我们不能随意的写。

4. 写Lower Page时间更短，写Upper Page时间更长。所以会看到有些Page写入速度快，有些Page写入时间慢。读取时间对Lower Page和Upper Page来说都差不多。

想要每天看一条SSD文章吗？扫一扫，微信关注我们！或者微信搜索公众号ssdfans关注。

转载请注明来自SSD技术学习网，本文地址：

Read more