


SSD通俗原理简介

 **tom-sun**
builder, 求真务实, 反对媚俗。
160 人赞同了该文章

1. 什么是SSD

SSD即Solid State Drive, 即固态硬盘的缩写。目前主流的SSD使用半导体闪存 (Flash) 作为介质的存储设备。SSD有别于HDD(Hard Disk Drive)机械硬盘。

SSD诞生于上世纪70年代。最早的SSD使用RAM, RAM掉电数据就会丢失。价格也特别贵。后来出现了基于闪存的SSD。闪存掉电之后数据不丢失。flash SSD慢慢取代了RAM SSD。此时HDD盘已占据了大部分的存储市场。到本世纪初。由于工艺的不断进步。SSD迎来了大发展。容量和性能不断提升。价格也不断下降。HDD的工艺和技术上已经很难有突破性的进展。SSD在性能和容量上还在不断突破。相对于HDD市场停滞。无论是企业级市场还是消费级市场SSD快速增长。SSD市场份额一直在扩大。相信不久的未来。SSD在在线存储领域取代HDD成为主流的存储设备。成为软件定义存储的主流设备。

在工作方式上。HDD使用磁盘。即磁性介质作为数据存储介质。在数据读取和写入上。使用磁头+马达的方式进行机械寻址。因为机械硬盘靠机械驱动读写数据的限制。导致机械硬盘的性能提升遇到了瓶颈。特别是HDD盘的随机读写能力。受其机械特性的限制。是一个巨大的瓶颈。SSD使用Flash作为存储介质。数据读取与写入通过SSD控制单元进行寻址。不需要机械操作。有着优秀的随机访问能力。

下图分别是HDD和SSD的组成。HDD使用机械SSD由主控、闪存、DRAM (可选)、PCB (电源芯片、电阻、电容等)、接口 (SATA、SAS、PCIe等)。



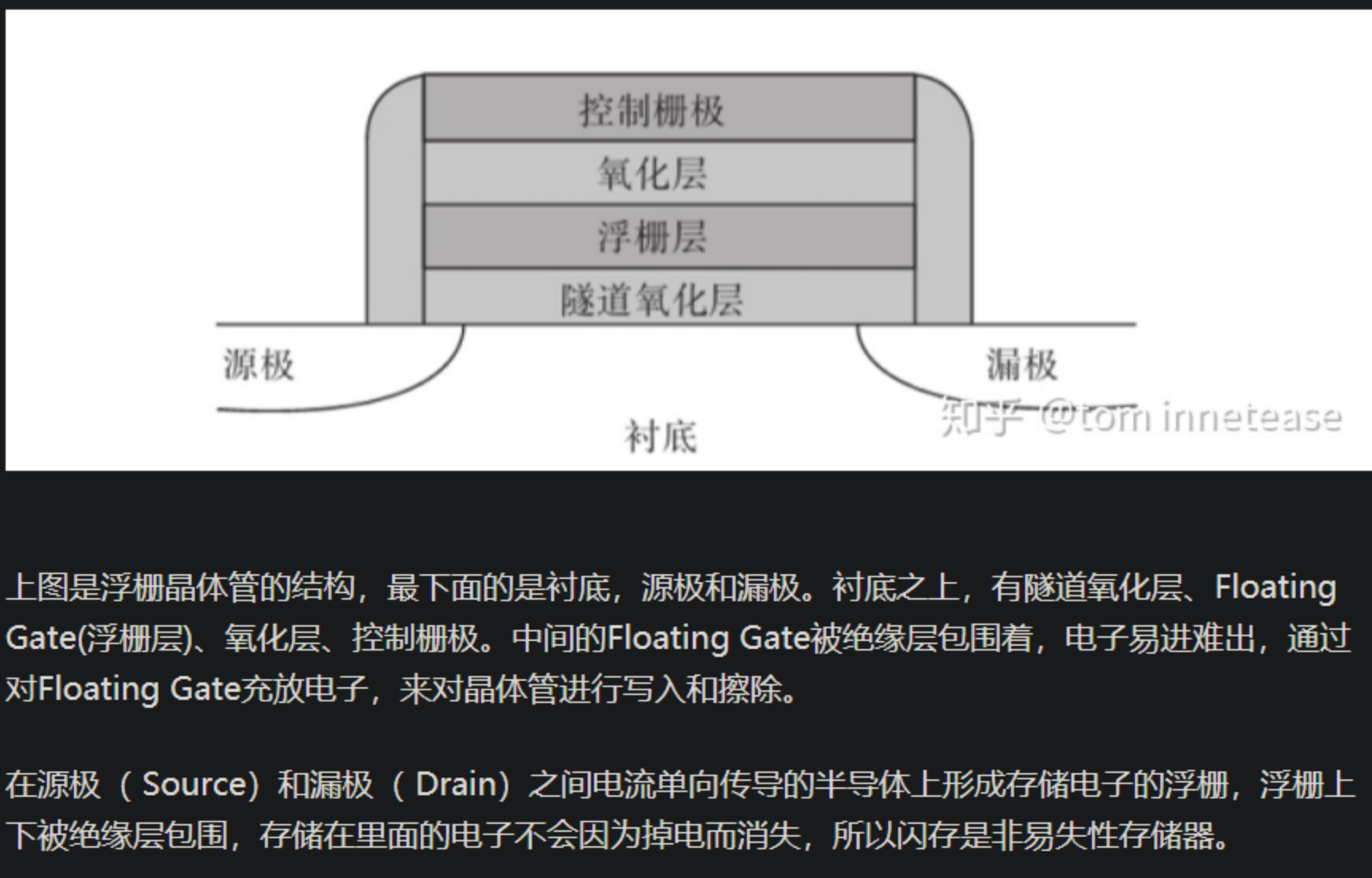
SSD分为前端、中间、后端。前端是接口和相关的协议模块 (STAT/SAS/PCIe等)。中间是FTL (Flash Translation Layer) 模块。后端是和闪存通信模块 (闪存通常ONFI或者Toggle协议)。



2. SSD的工作原理

2.1 SSD的存储单元

Flash的基本存储单元是——浮栅晶体管。



上图是浮栅晶体管的结构。最下面的是衬底。源极和漏极。衬底之上。有隧道氧化层。Floating Gate(浮栅层)。氧化层。控制栅极。中间的Floating Gate被绝缘层包围。电子隧进隧出。通过对Floating Gate充放电。来对晶体管进行写入和擦除。

在源极 (Source) 和漏极 (Drain) 之间电流传导的半导体上形成存储电子的浮栅。浮栅上下被绝缘层包围。存储在里面的电子不会因为掉电而消失。所以闪存是非易失性存储。

下图是浮栅晶体管的写和擦除的原理。

写操作如左图。是在上面的控制栅极加正电压Vpp。使电子通过隧效应进入浮栅。擦除操作正好相反。如右图。是在衬底加正电压Vpp。把电子从浮栅中吸出来。写入的过程是充电子的过程。如果写入的page之前已经写过。在写入之前。必须先对flash进行擦除0。清除浮栅中的电子。

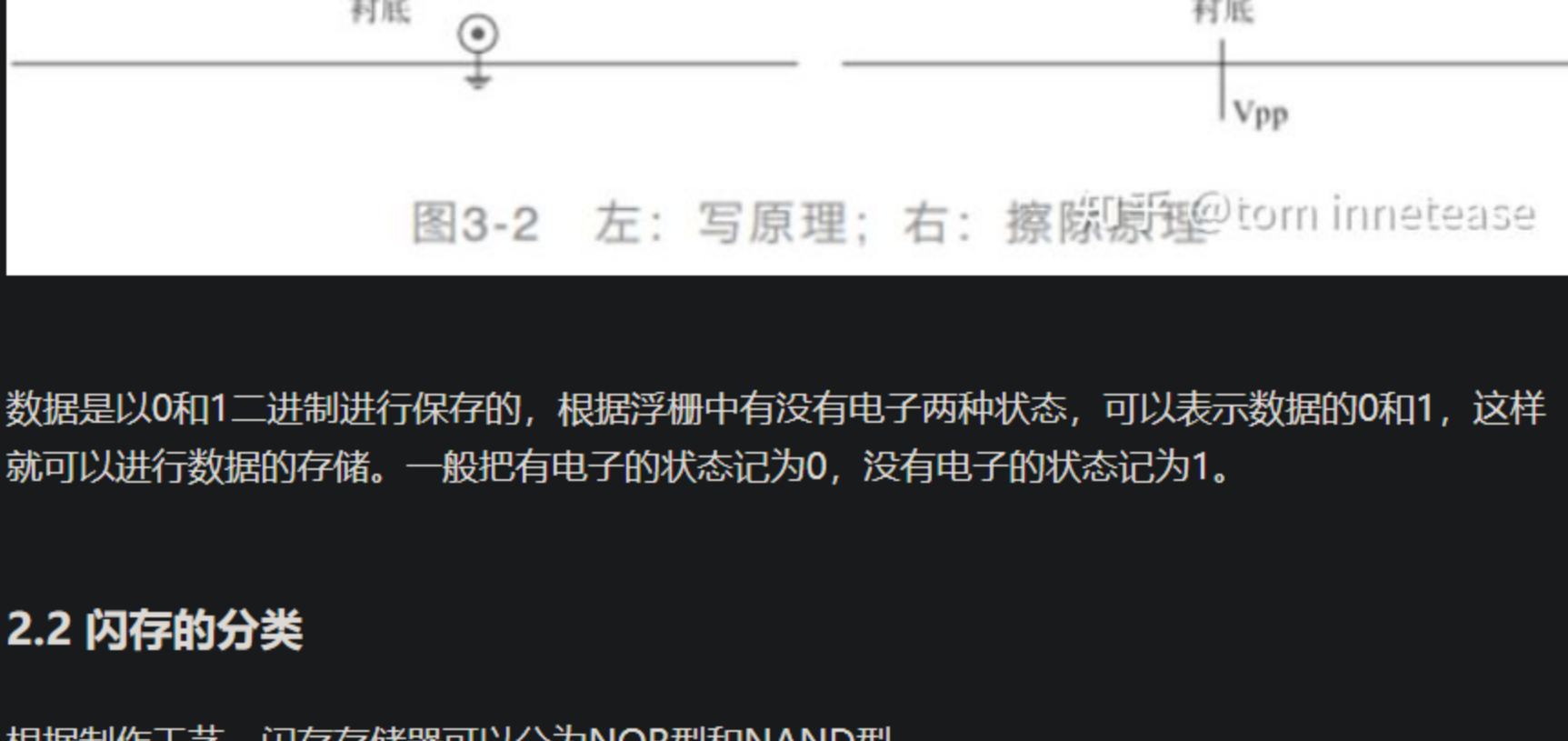


图3-2 左：写原理；右：擦除原理

数据是以0和1二进制进行存储的。根据浮栅中有没有电子两种状态。可以表示数据的0和1。这样就可以进行数据的存储。一般把有电子的状态记为1。没有电子的状态记为0。

2.2 闪存分类

根据制作工艺。闪存存储器可以分为NOR型和NAND型。

NOR型是为了替代EEPROM而设计。可以按位或者按字节进行访问。NOR型闪存芯片具有可靠性高。随机读取速度快的优势。但擦除和编程速度较慢。容量小。主要用于存储可执行的代码或代码。

NAND闪存容量大。按页进行读写。容量大。适合进行数据存储。本文介绍都是基于NAND flash。

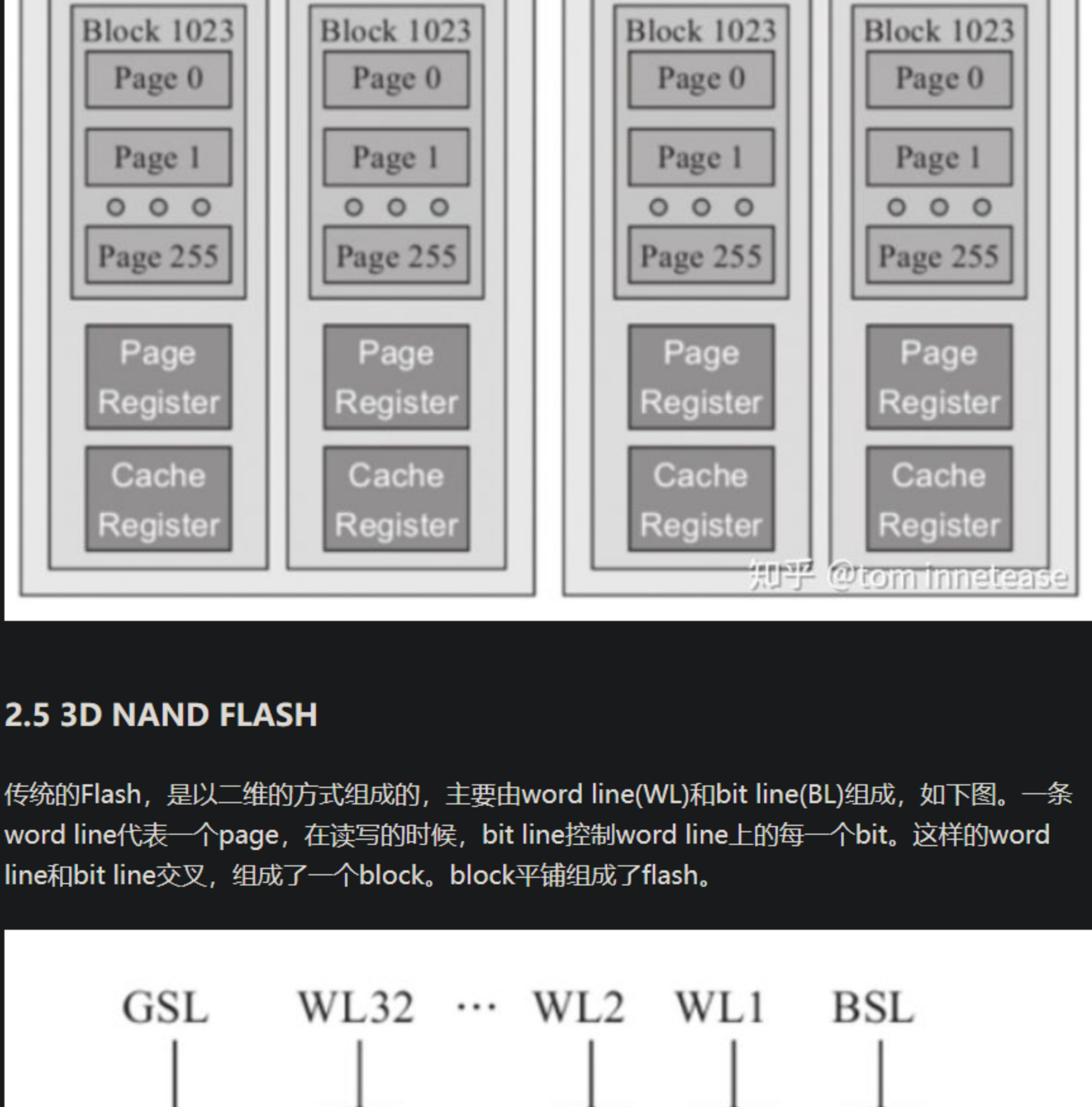
2.3 NAND flash的擦除、读、写

- 擦除。在衬底施加电压足够长的时间。把电子从浮栅中吸出来。擦除之后。整个block的数据都变成了1。由于一个block共用一个衬底。所以在擦除时。一次擦除一个block。即擦除的单位为block。
- 写。写的过程是对浮栅充电子。也称为编程。写之前需要先进行擦除。由于擦除之后。数据都变成了1。所以需要把要写入的浮栅进行充电。
- 读。读取的时候。对晶体管施加一个低电压。如果浮栅中没有电子。那么管子就是导通的。读到1；读的时候。有电子。管子不导通。读到0。读取数据时。因为是否有电子会影响到管子的导通性。所以可以利用电流感测浮栅即电子抽取量的多寡。靠感测强度转换成二进制的0与1。

2.4 闪存的内部组织架构

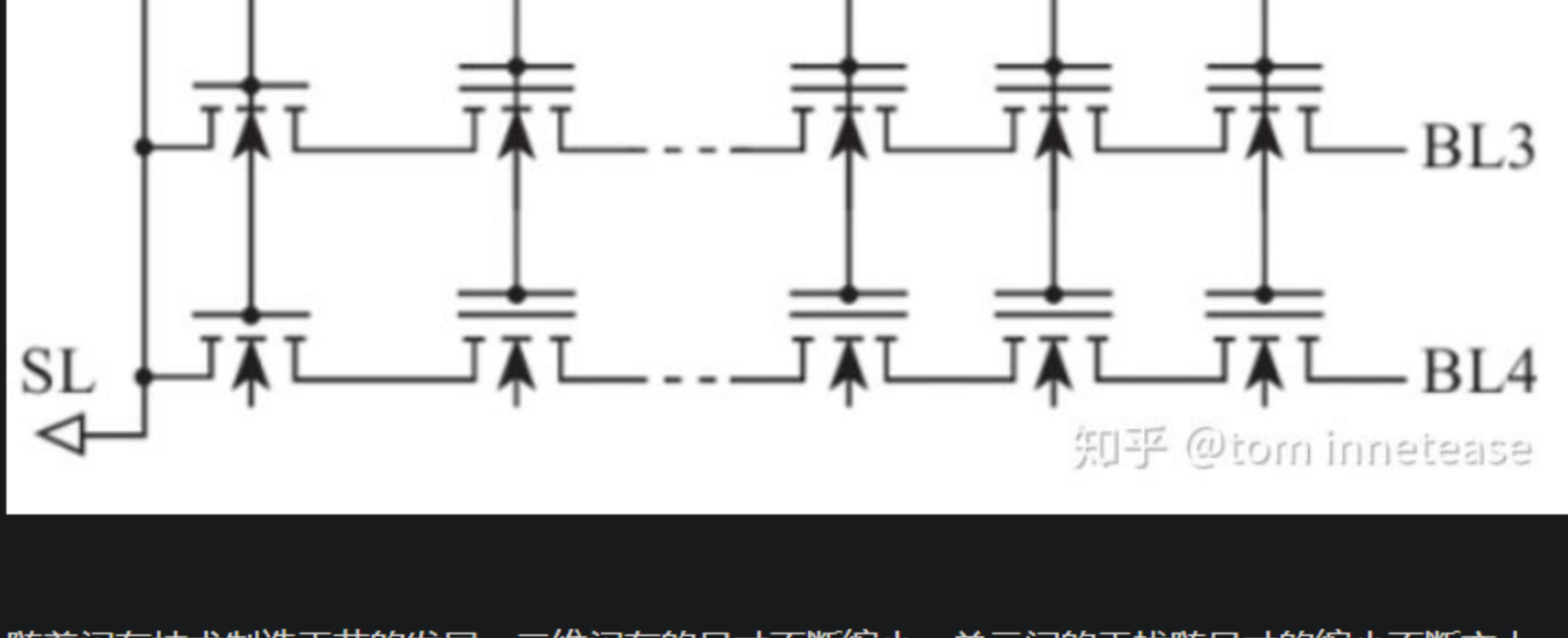
如下图的闪存的内部组织架构:

- 1个Chip/Device -> 多个DIE或者LUN
- 1个DIE/LUN -> 多个Planes
- DIE/LUN是接收和执行内存命令的基本单元
- 1个Plane -> 上千个Blocks
- 每个Plane有块切分寄存器。一个Page register。一个Cache register
- 1个Block -> 上百个Pages
- Block是擦除的基本单位
- 1个Page -> 一般是4KB 或者 8 KB + 几百个字节的数据空间
- Page是读或者写的基本单位
- Cells -> flash存储信息的基本单位。根据每个cell可以保存Flash 2bit, 3bit可以分为SLC, MLC, TLC



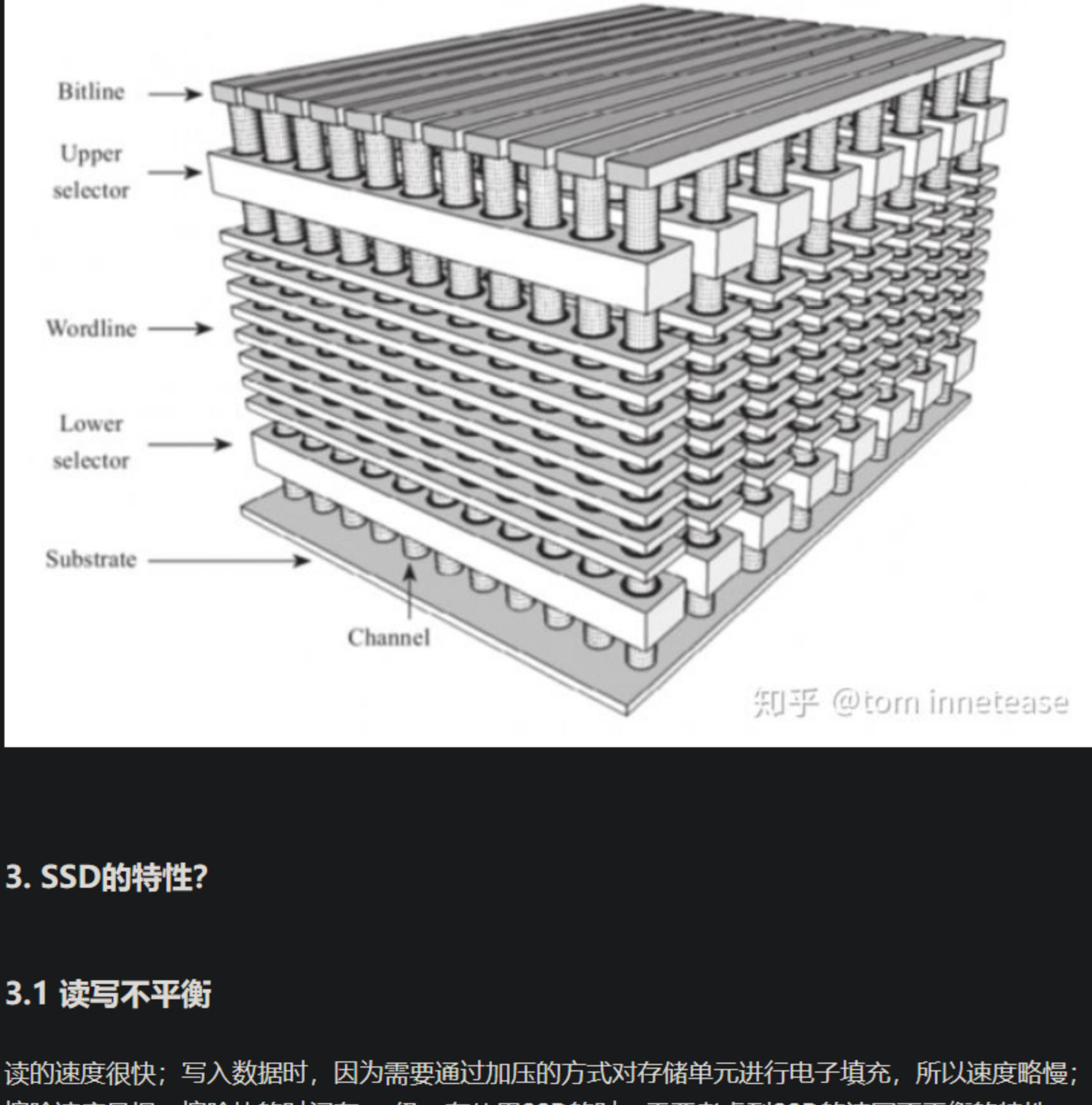
2.5 3D NAND FLASH

传统的Flash。是以二维的方式组成的。主要由word line(WL)和bit line(BL)组成。如下图。一条word line代表一个page。在该半的时候。bit line控制word line上的每一个bit。这样的word line/bit line交叉。组成了一个block。block平面组成了Flash。



随着闪存制造工艺的发展。二维闪存尺寸不断缩小。单元间的干扰随尺寸的缩小不断变大。3D NAND Flash技术的出现。有效的解决了单元干扰的问题。

下图是一种3D闪存的立体图。在这种三维闪存中。flash堆叠了起来。如果2D NAND Flash比作平房。那么3D NAND Flash可以看成是楼房。3D NAND Flash可以通过提高flash的层数在单位面积上建更多的晶体管。3D NAND Flash在单位面积堆叠更多的存储单元。在降低每bit成本上很有优势。



3. SSD的特性?

3.1 读写不平衡

读的速度很快。写入数据时。因为需要通过加电压的方式对存储单元进行电子填充。所以速度较慢。擦除速度较慢。擦除块的时间在ms级。在使用SSD的时候。需要考虑到SSD的读写不平衡的特性。

3.2 先擦后写

Nand Flash的写入以page为单位。擦除以block为单位。在Page页写入之前。必须要把page页所在的block块擦除。这个是由Nand Flash的工作原理决定的。

3.3 快速页写

一个Wordline对应着一个或若干个Page。具体层数多少取决于层SLC、MLC或者TLC。对SLC来说。一个Word line对应一个Page; MLC则对应2个Page。这两个Page是一一对 (Lower Page和Upper Page) ; TLC对应3个Page (Lower Page、Upper Page, Extra Page。不同厂家厂家叫法不一样) 。一个Page有多大。那么Wordline上面就有多个存储单元。就有多个bitline。写入以页为单位。

一个Block当中的所有这些存储单元都是共用一个衬底的。所以对衬底施加电压。上面所有浮栅的电子都会被吸出来。所以擦除是以块为单位的。

3.4 寿命有限

每个NAND Block都有读写次数的限制。当超过这个次数时。该Block可能就不能用了。浮栅吸收不进电子 (写失败) 或者浮栅吸收的电子很容易释放出来 (读时翻转。0->1)。或者浮栅吸收的电子出不来 (擦除失败)。这个最大读写次数按SLC、MLC、TLC依次递减。数据的读写次数可达十万次。MLC一般为几千到几万。TLC降到几千到几千。

3.5 SLC/MLC/TLC

根据一个存储单元可以存储多少bit的数据。闪存单元可以分为SLC(Single Level Cell)、MLC(Multiple Level Cell)、TLC(Triple Level Cell)。

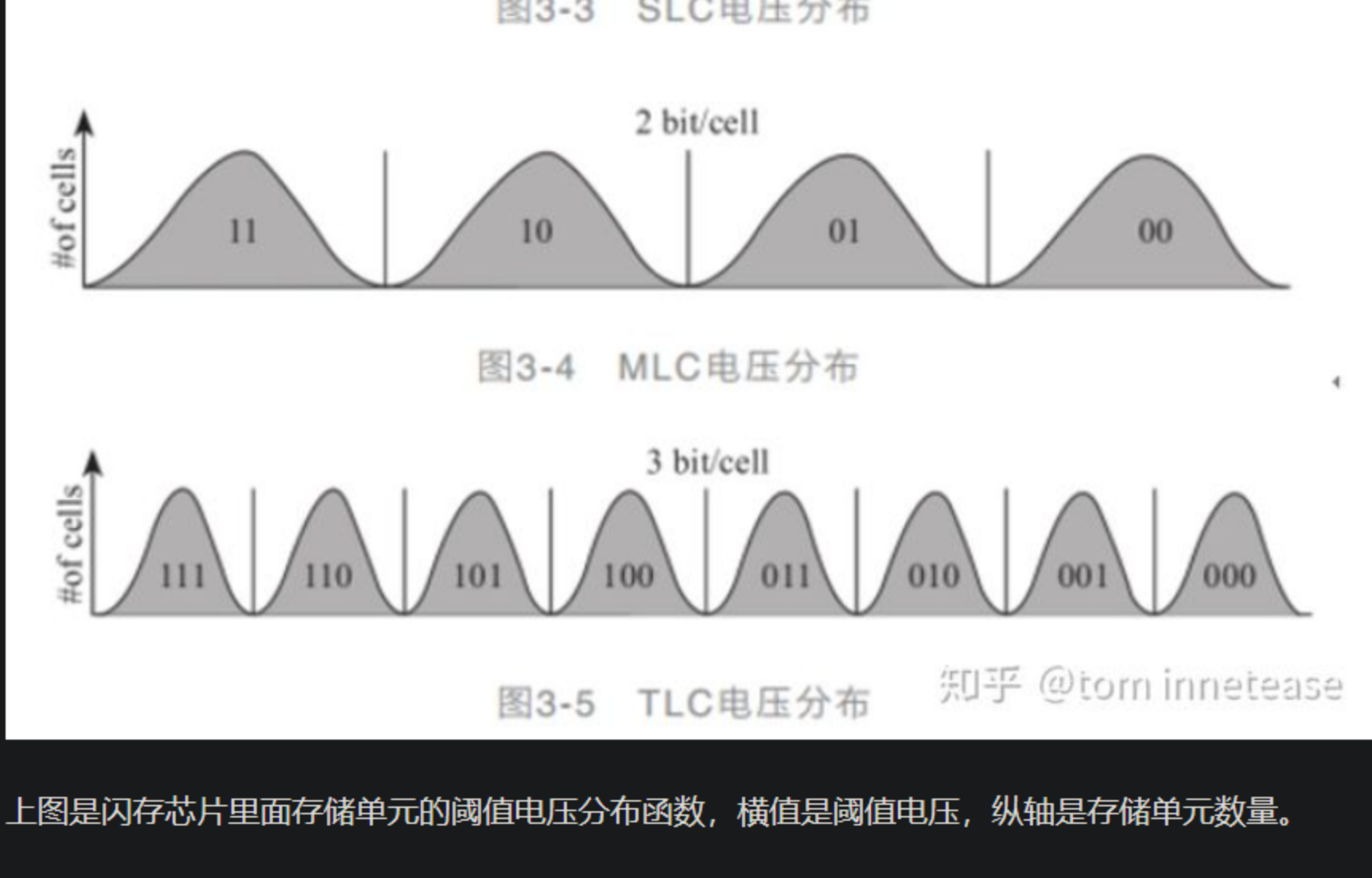


图3-3 SLC电压分布

图3-4 MLC电压分布

图3-5 TLC电压分布

上图是闪存芯片里面存储单元的阈值电压分布函数。横轴是阈值电压。纵轴是存储单元数量。

一个存储单元电子划分得越多。那么在写入的时候。控制电子浮栅吸收的电子个数就越来越精确。所以写的时间就越长。同样的。读的时候。需要尝试用不同的参考电压去读取。一定程度上加快了读取时间。在性能上。TLC不如MLC。MLC不如SLC。在寿命上。SLC > MLC > TLC。在价格上SLC > MLC > TLC。下表是SLC、MLC、TLC的一些参数比较。数据比较旧了。但是对比趋势不变。现在又出现了一种QLC。一个闪存单元可以存储4个bit的数据。

闪存类型 SLC MLC TLC 每单元bit数 128bit 256bit 512bit 每单元状态数 248 1024 1024 读取时间(us) 305075 150000 150000 写入时间(us) 15003000 4500

4. SSD的FTL

FTL即Flash Translation Layer的缩写。是SSD的一个重要组成部分。实现了以下功能:

- **Interface Adapter:** 在内部FTL中主要关联eMMC/SCSI/SATA/PCIe/NVMe等接口。而在外部FTL中主要关联Linux Block Device。
- **Address Translation:** 地址映射。也可以叫做mapping。负责逻辑地址和物理地址之间的映射。多技术模块难以说机制为核心进行。众所周知。Nand Flash具有写时擦除的特性。因此写入数据时不得不不断地擦除。
- **Garbage Collection:** 垃圾回收。简称GC。回收并更新产生的数据所占空间的回收工作。
- **Wear Leveling:** 磨损均衡。简称WL。避免某一个Nand Block很快坏去。使所有Block的PE Cycle均衡发展。因为flash的读写次数是有限的。如果不进行磨损均衡。整个SSD的有些block可能读写次数不平衡而很快坏去。
- **Power Off Recovery:** 掉电恢复。简称POR。正常掉电。SSD会把缓存中的数据刷新到闪存。重新加载保存的数据即可。如果异常掉电。因为某些人为或自然外力的原因导致数据没有成功写入到Nand中。掉电恢复要恢复到掉电前的安全状态。比如恢复RAM中的数据以及Address Translation中的映射表。
- **Parallelization and Load Balancing:** 在前面的2.4节的闪存的内部组织架构介绍中。可以知道SSD中存在一定的并行性。利用这些并行性可以提供SSD的并发请求处理能力。提高其性能。
- **Cache Manager:** Cache不仅可以存放用户数据。也可以存放FTL Metadata。对系统的整体性能有着天然的提升。
- **Error Handler:** 处理读写操作中遇到的Fatal Error或ECC Error状况。以及Bad Block或Weak Block的管理。

在下一篇文章《带你了解SSD(2)-FTL》中。会详细的介绍FTL的基本功能。

Notes

作者: 网易存储团队攻城略地

如有理解上和描述上有错误或错误的地方。欢迎共同交流。参考已经在参考文献中注明。但仍有可能有疏漏的地方。有任何侵权或者不明确的地方。欢迎指出。必定及时更正或者删除。文章供于学习交流。转载注明出处

5 参考文献

[1] SSDFans深入解读SSD固态硬盘核心技术、原理与实践[M].北京: 机械工业出版社, 2018.6

[2] codecapsule.com/2014/02/...

[3] ssdfans.com/blog/2018/0_那些事 (0) 之写在后面的话/

编辑于 2018-08-30

[固态硬盘](#) [云存储](#) [数据备份 \(广义\)](#)