新浪 -研发-运维部
刘明生

《SSD知识入门和提高》

讲师姓名：刘明生
邮箱　　：mingsheng@staff.sina.com.cn
新浪微博：@明生78
手机　　：13810097928

2012-02

讲师介绍

专注于服务器性能优化与设备选型，致力于为互联网公司提供高性价比服务器，降低TCO.

课程目标

了解使用SSD的原因

了解SSD基本原理

掌握SSD正确使用方法

学会为应用配置正确的SSD

课程结构

1.SSD & 机械硬盘 基本组成

2.SSD 读/写模式及底层工作原理

3.SSD 使用中的注意事项

4.SINA ERP中SSD机型讲解

使用SSD的原因

提升单机性能

=降低采购数量

=降低机房费用

=**降低项目总成本**

实现个人价值/
成就感

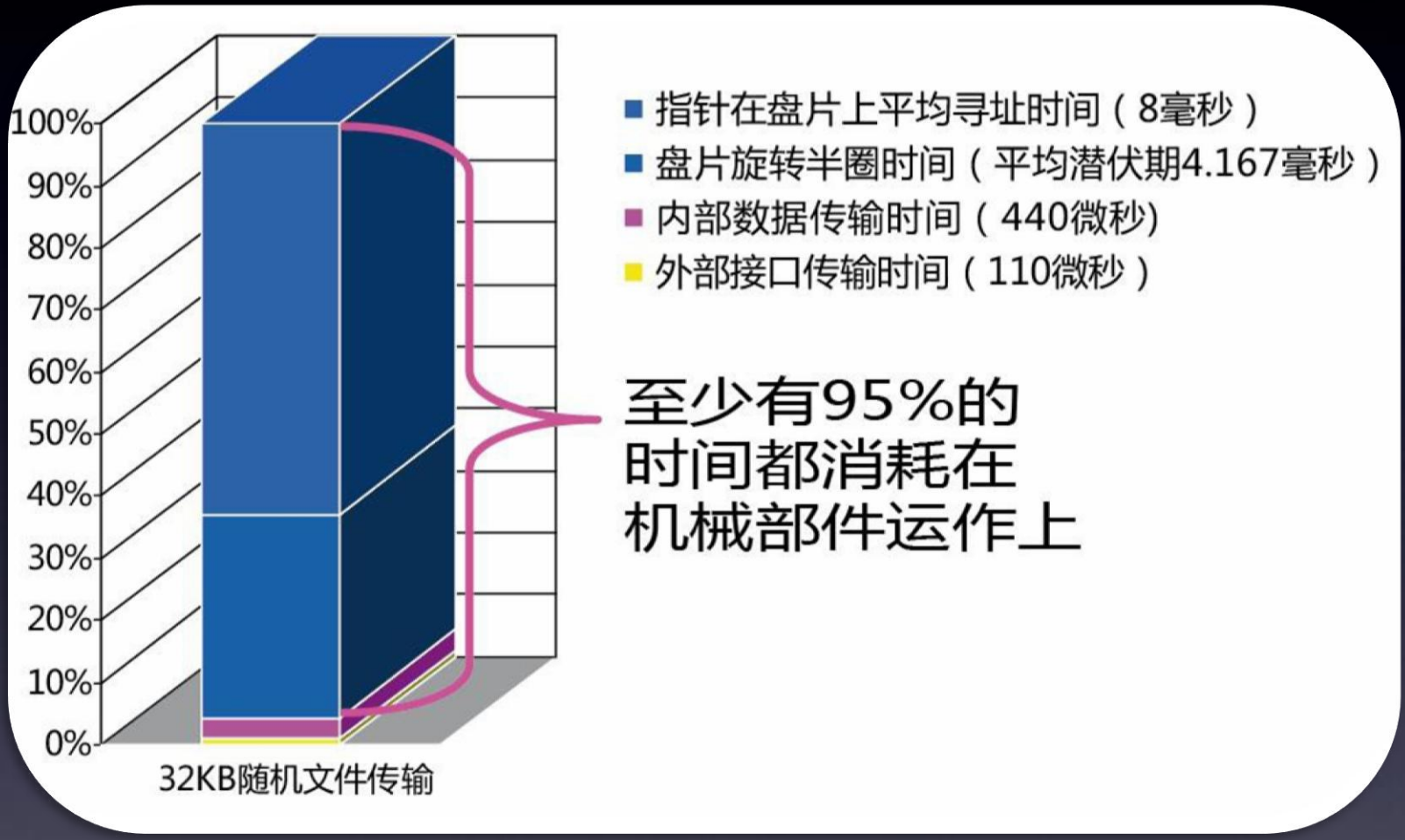
降低
本部门成本

降低
运维部门成本

降低
公司运作成本

硬盘历史

工作原理- 机械硬盘劣势



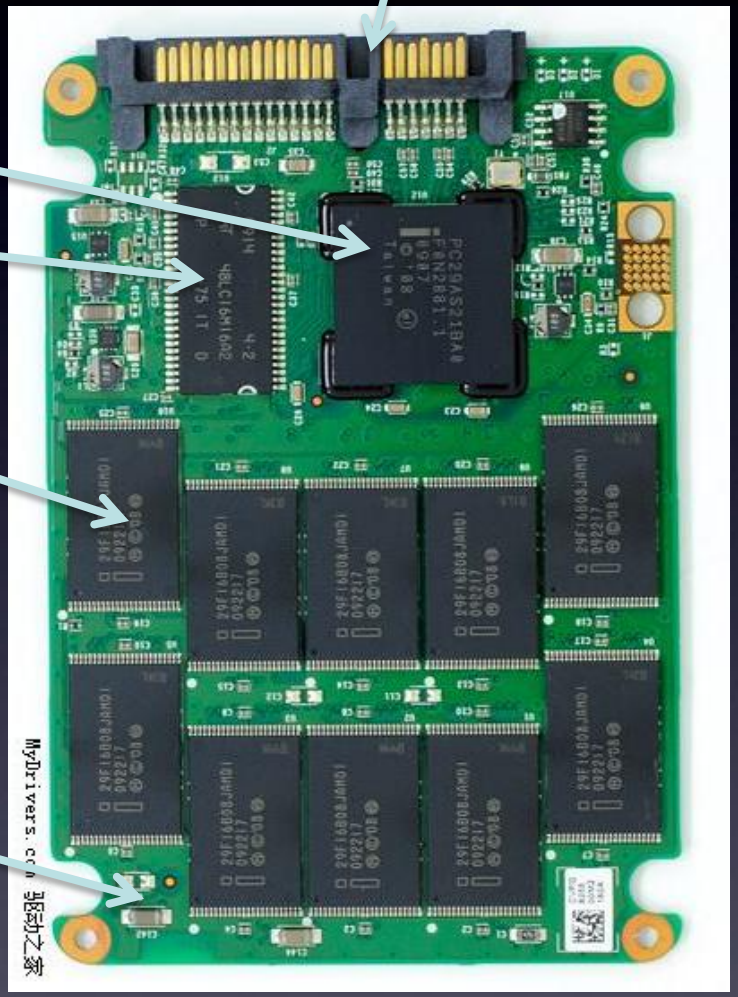
1秒=1000毫秒=1,000,000微秒=1,000,000,000 纳秒 =1,000,000,000,000 皮秒

SSD结构

一.主控

二.NAND

三.SSD底层技术

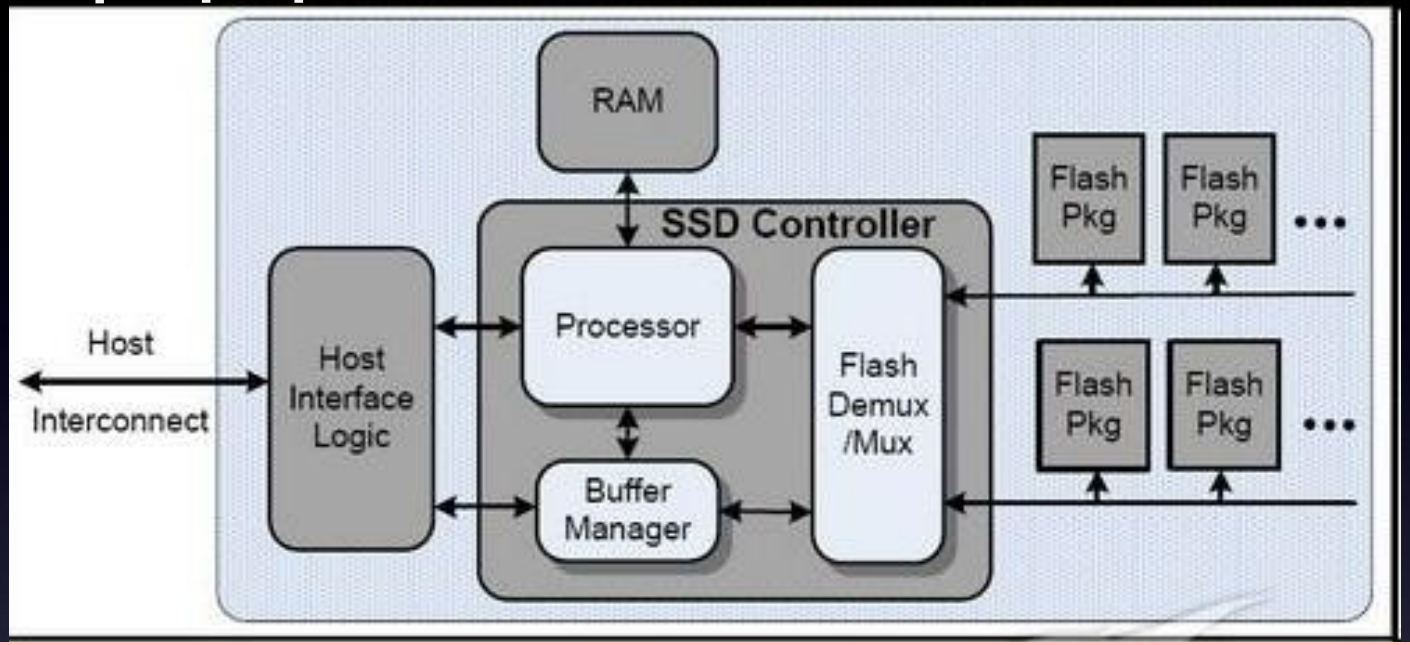


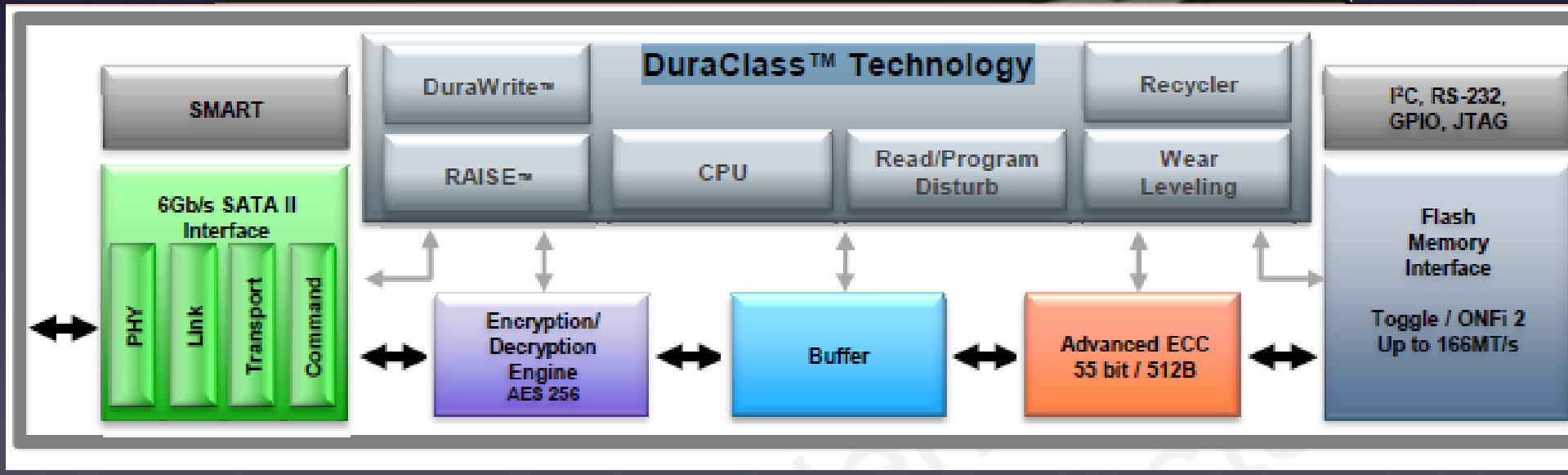
Flash 颗粒-Intel/Micron

主控

Flash NAND

SSD框图


RAM
SSD Controller
Processor
Buffer Manager
Host Interface Logic
Flash Demux /Mux
Host
Interconnect
Flash Pkg
Flash Pkg
Flash Pkg
Flash Pkg

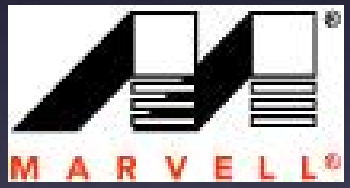

SMART
DuraClass™ Technology
DuraWrite™
RAISE™
CPU
Read/Program Disturb
Wear Leveling
Recycler
I²C, RS-232, GPIO, JTAG
6Gb/s SATA II Interface
PHY
Link
Transport
Command
Encryption/ Decryption Engine AES 256
Buffer
Advanced ECC 55 bit / 512B
Flash Memory Interface
Toggle / ONFi 2 Up to 166MT/s

一、主控

Intel

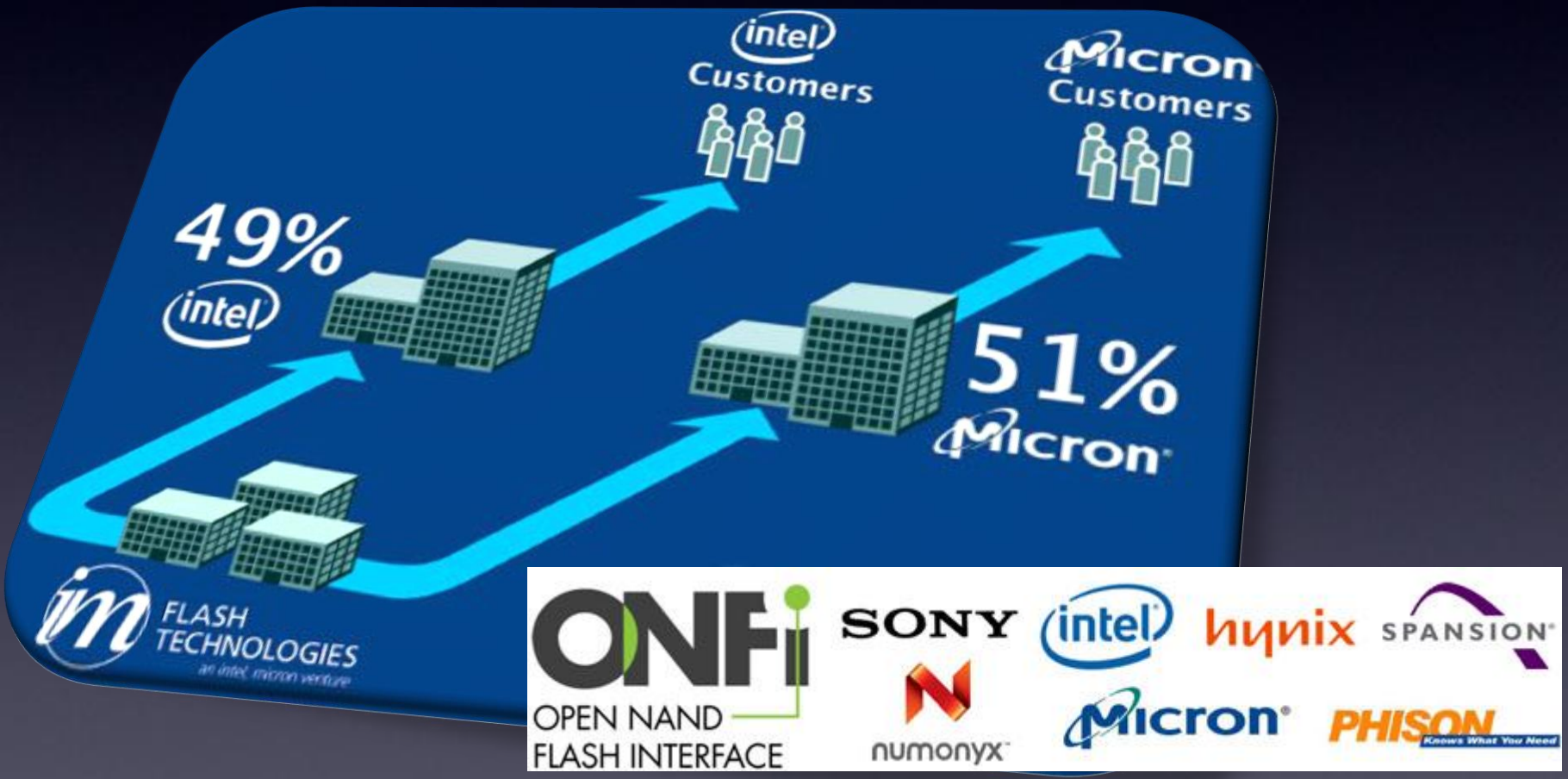
SandForce

Marvell



二、 Flash 颗粒

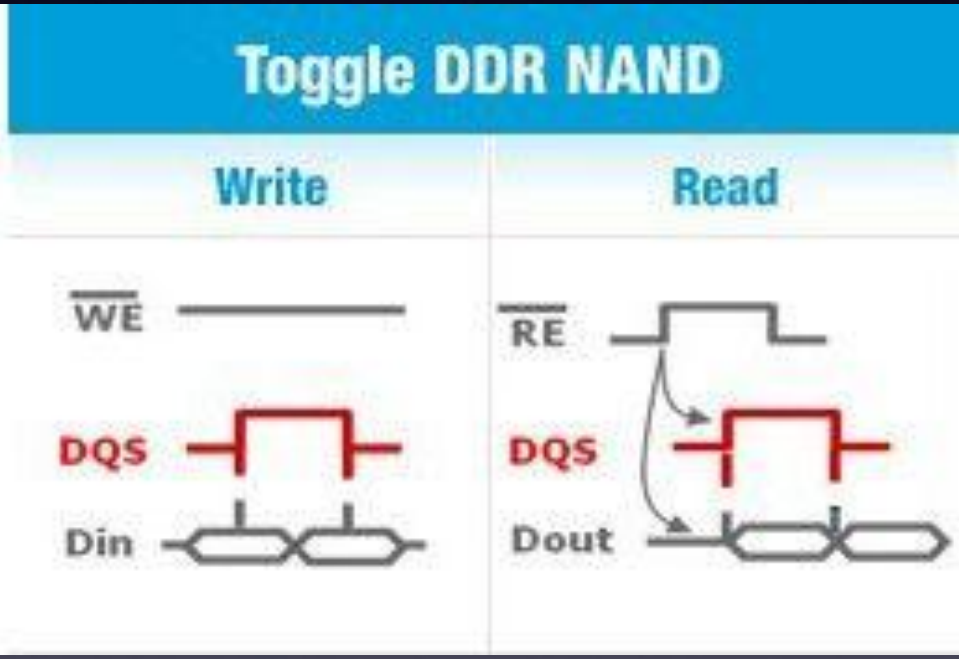
Intel/Micron合资企业IMFT



二、 Flash 颗粒

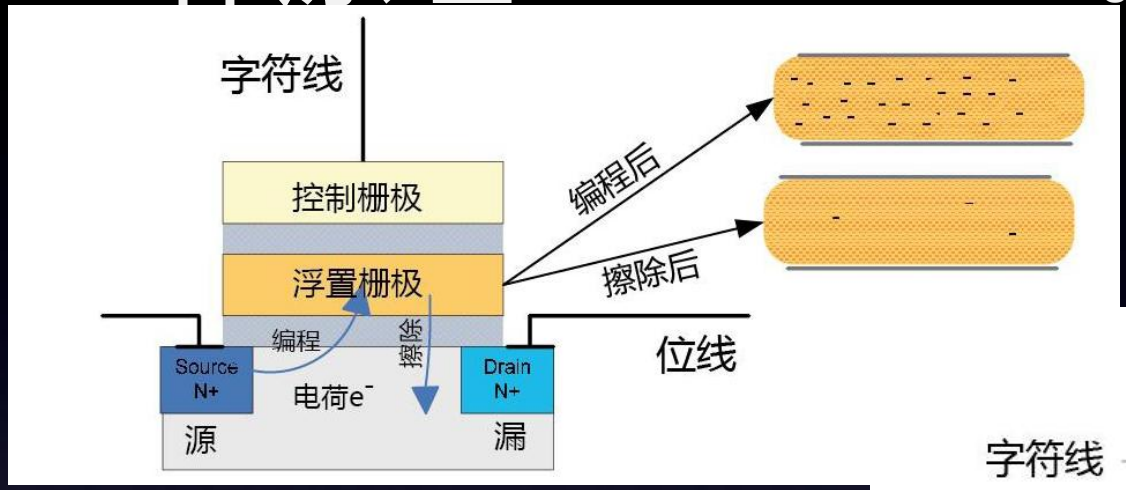
东芝、三星组成的Toggle DDR联盟

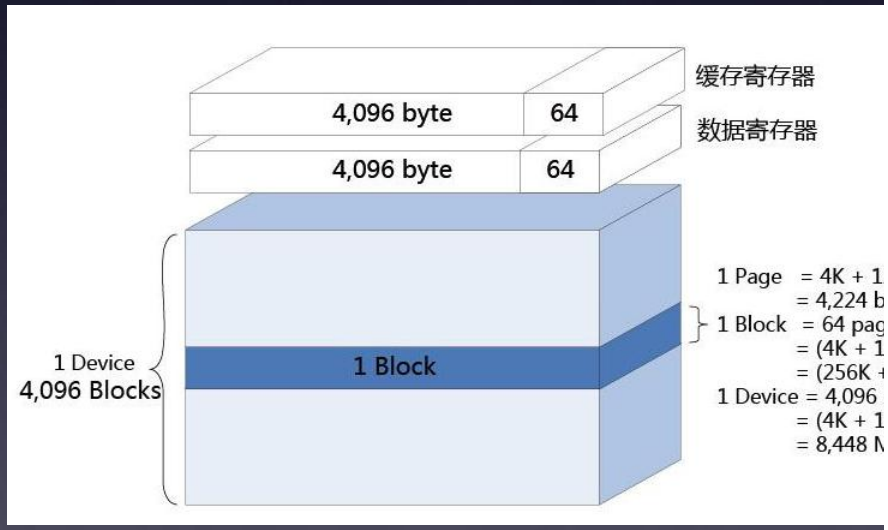
ONFi
OPEN NAND
FLASH INTERFACE

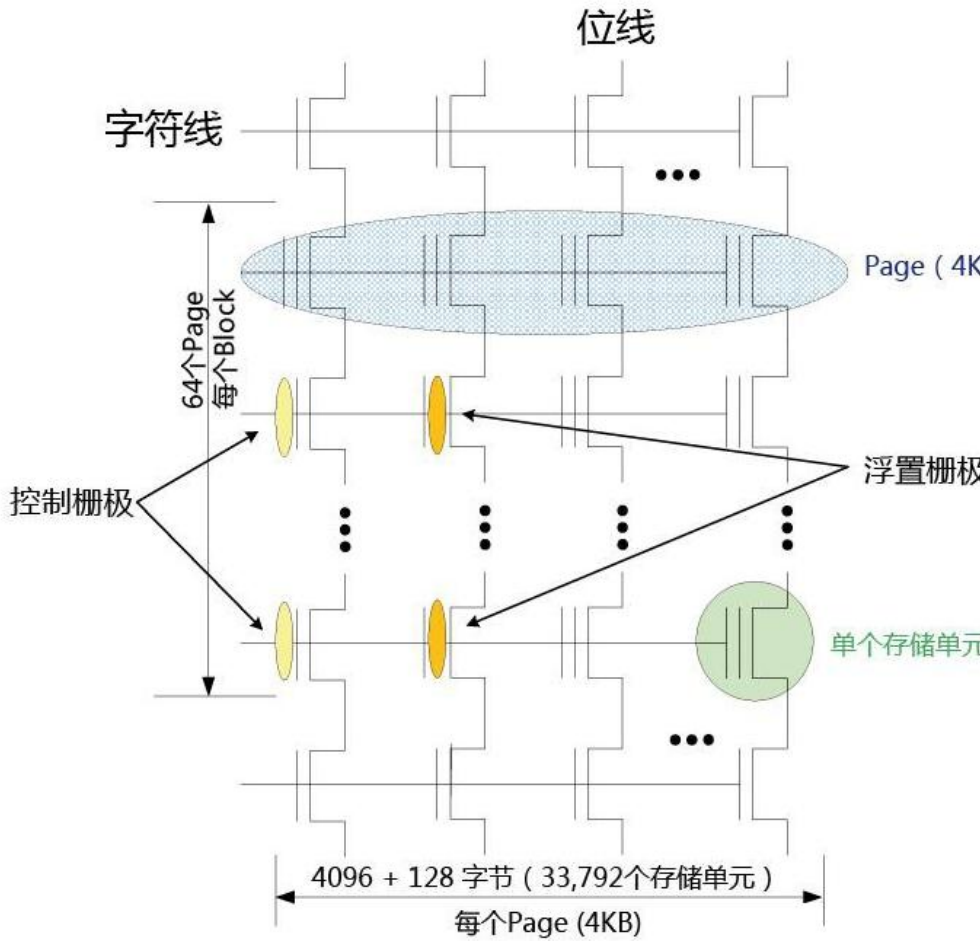


工作原理- Flash

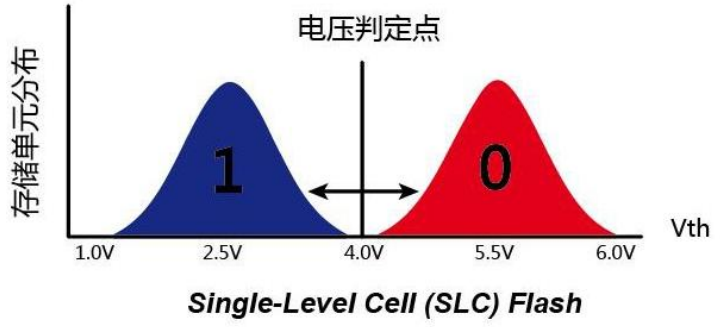
8Gb 50nm的SLC颗粒内部架构。





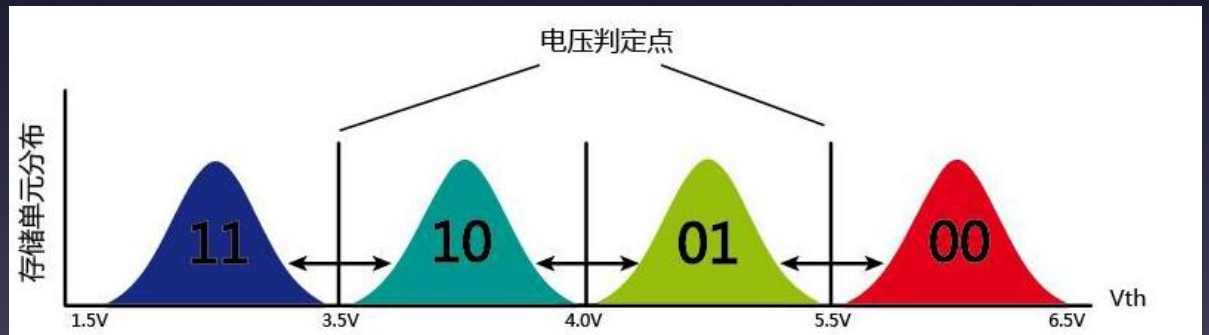


工作原理- MLC和SLC的区别



Single-Level Cell (SLC) Flash

值	状态
0	已编程
1	已擦除

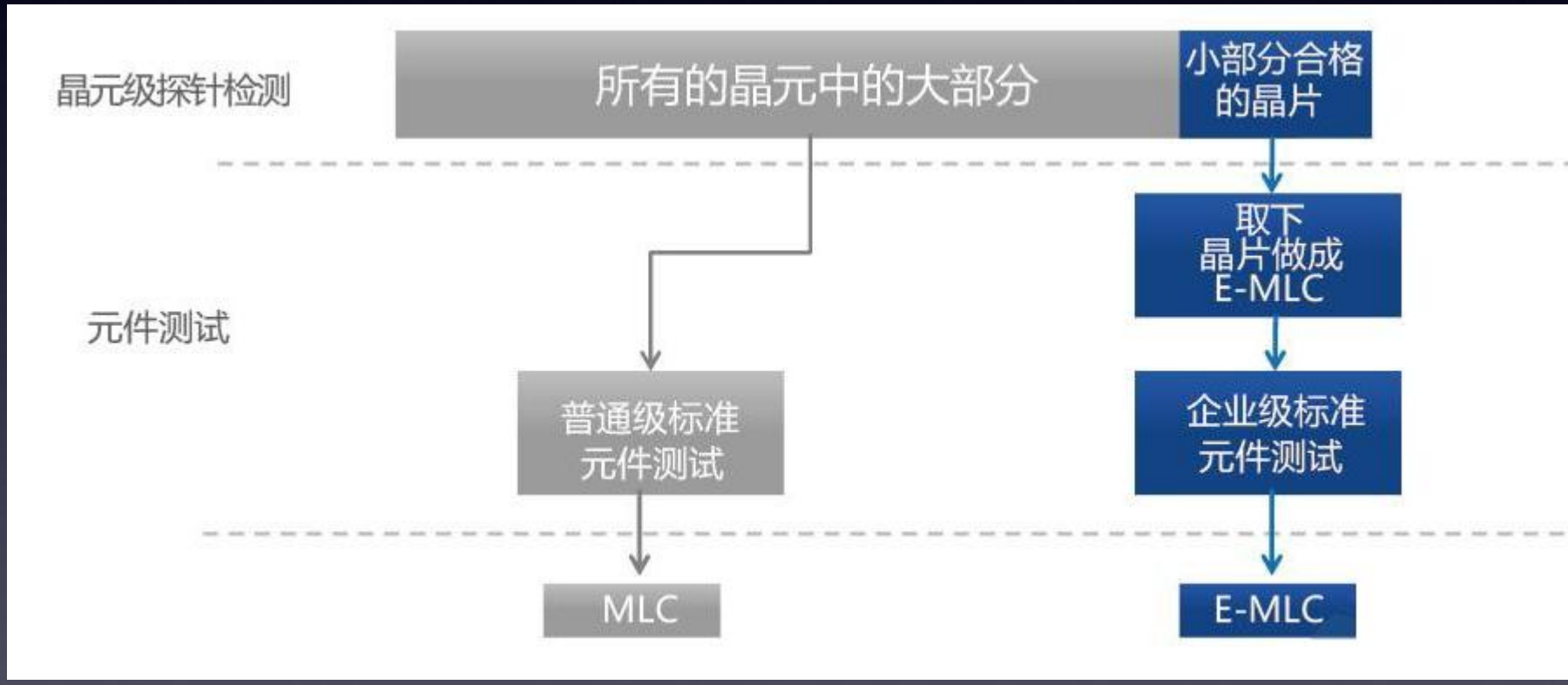


Multi-Level Cell (MLC) Flash

值	状态
00	完全编程
01	部分已编程
10	部分已擦除
11	完全擦除

工作原理- MLC和eMLC的区别

在NAND Flash工厂制造处理过程中，厂商把晶元上最好的那部分Flash晶片挑选出来并用企业级的标准来检测晶片的数据完整性和耐久度。检测完后，这些晶片被 取下来改变内部些许参数并进行之后的比标准MLC更苛刻的测试。当这些晶片通过测试后，就被定义为eMLC级别组，余下的就成为MLC级别组了。



工作原理- MLC和eMLC的区别

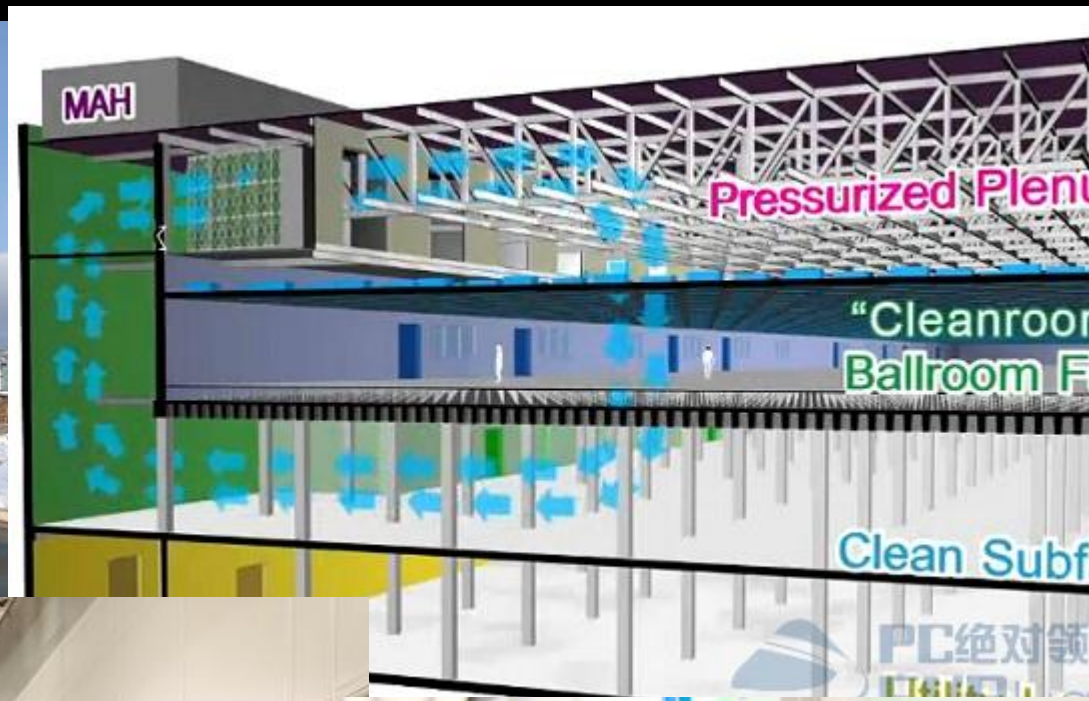
在NAND Flash工厂制造处理过程中，厂商把晶元上最好的那部分Flash晶片挑选出来并用企业级的标准来检测晶片的数据完整性和耐久度。检测完后，这些晶片被 取下来改变内部些许参数并进行之后的比标准MLC更苛刻的测试。当这些晶片通过测试后，就被定义为eMLC级别组，余下的就成为MLC级别组了。

应用程序组	使用环境 (通电时)	数据保存期 (断电后)	功能性出错 概率要求	UBER
消费级	40度下平均 每天8小时	30度下1年	≤3%	$\leq 10^{-15}$
企业级	55度下平均每 天24小时	40度下3个月	≤3%	$\leq 10^{-16}$

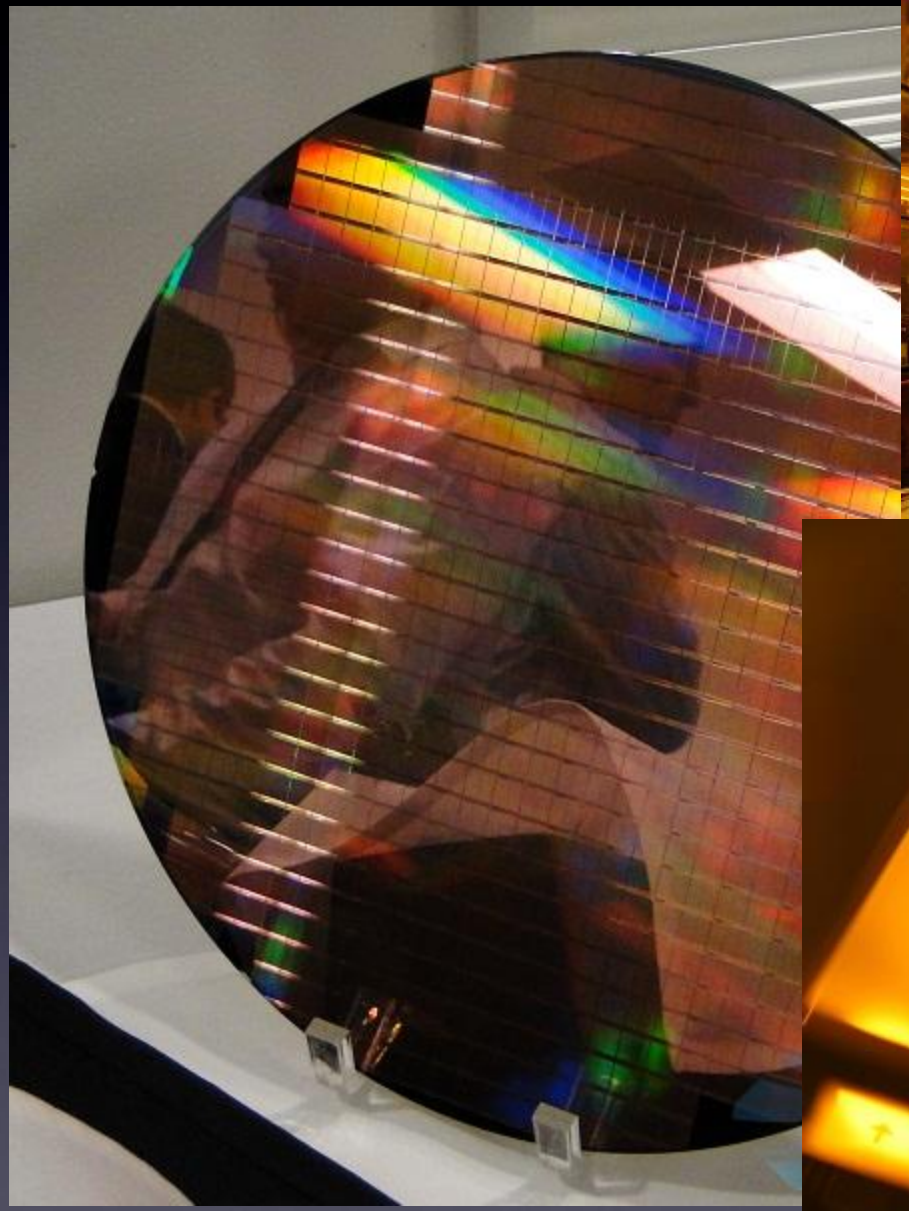
UBER(uncorrectable bit error rate)无法纠正位错误率

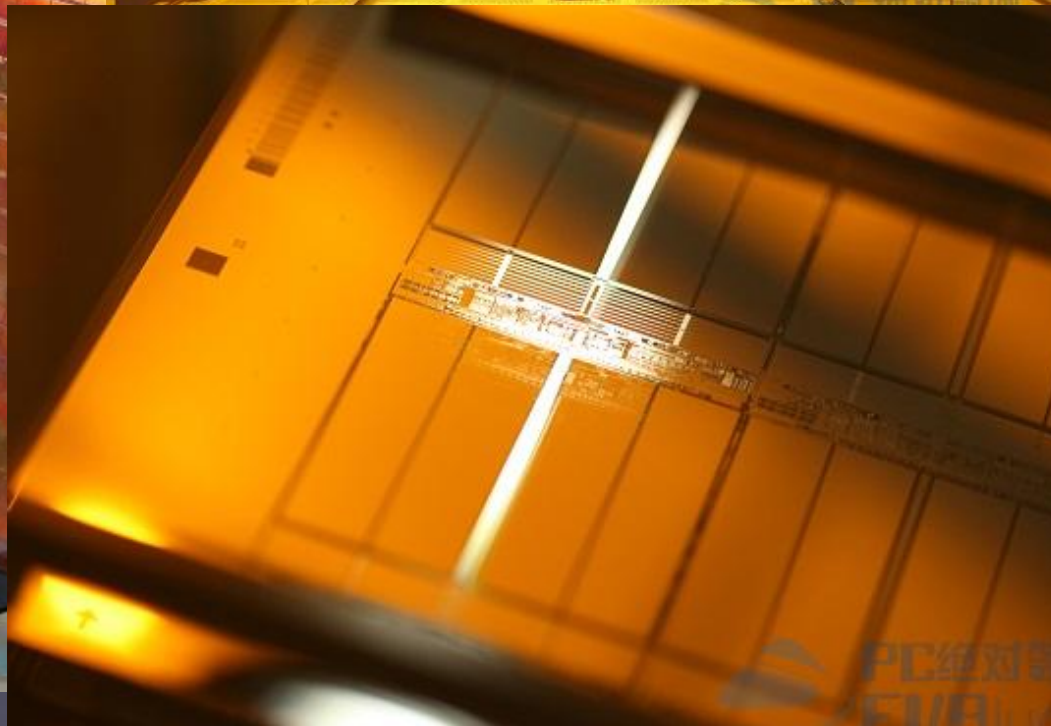
JEDEC 固态技术协会对消费级和企业级固态硬盘的标准

Flash 颗粒-Intel/Micron

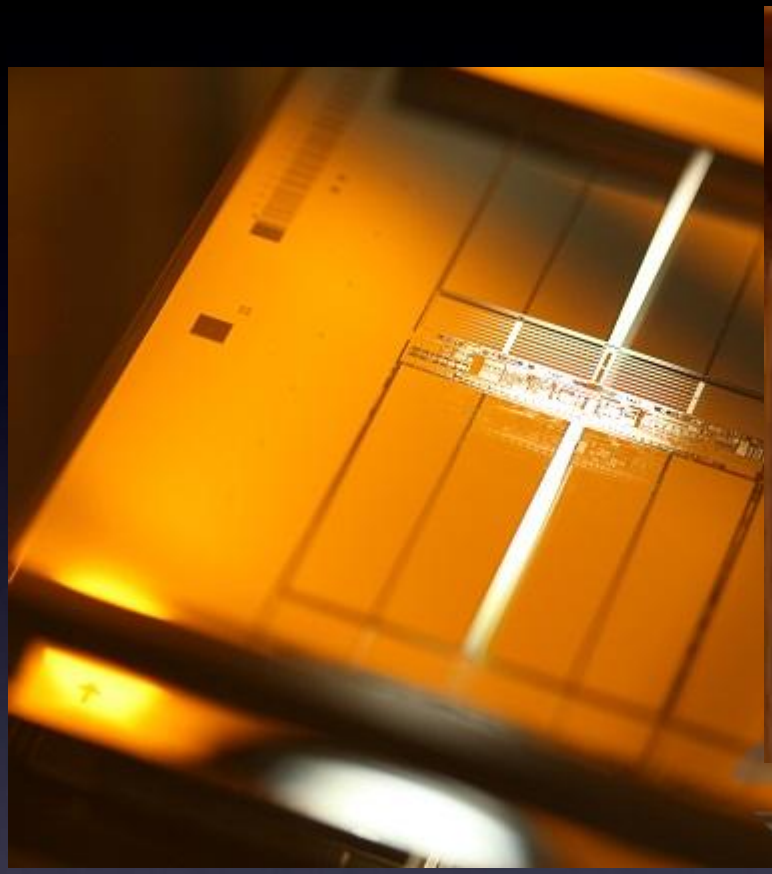


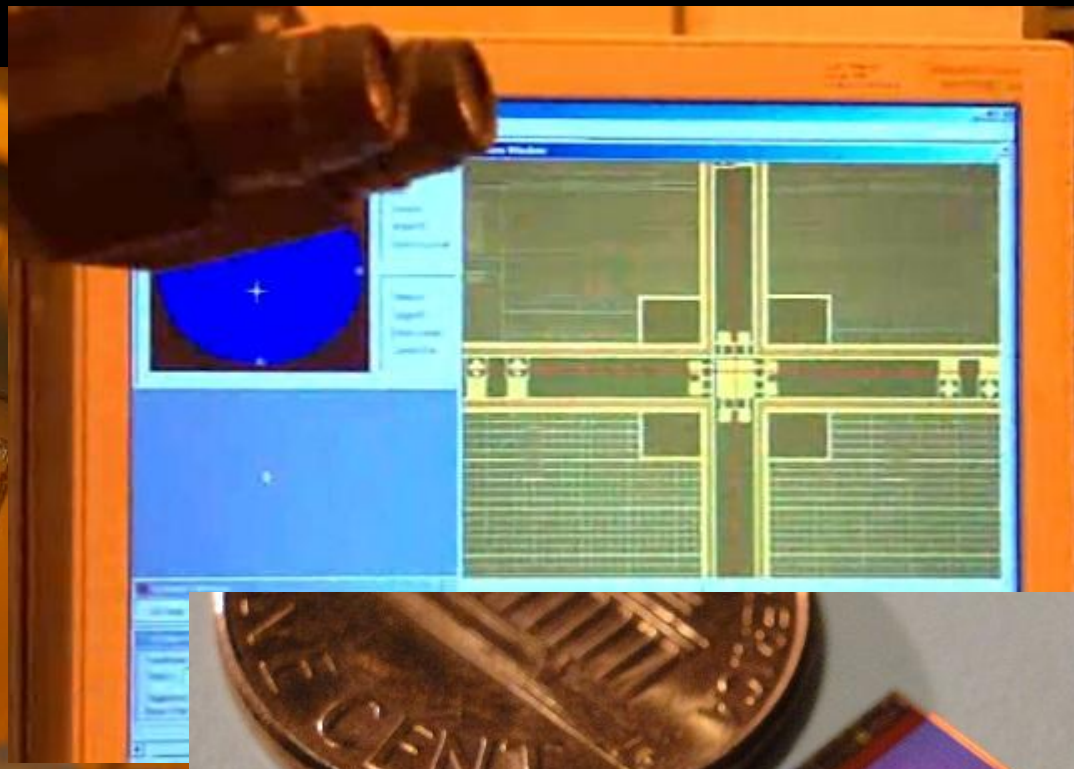
Flash 颗粒-Intel/Micron

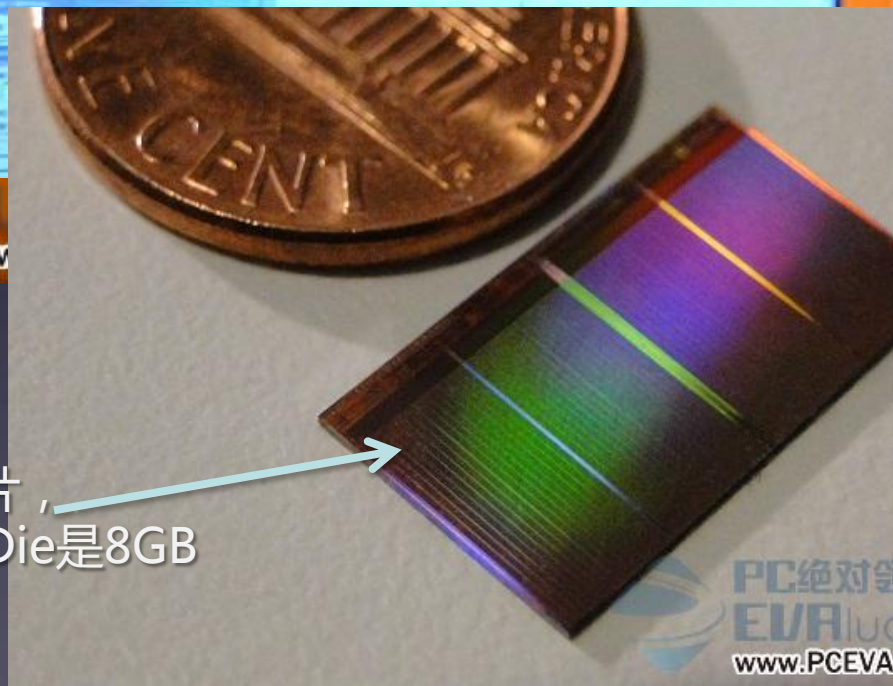




Flash 颗粒-Intel/Micron

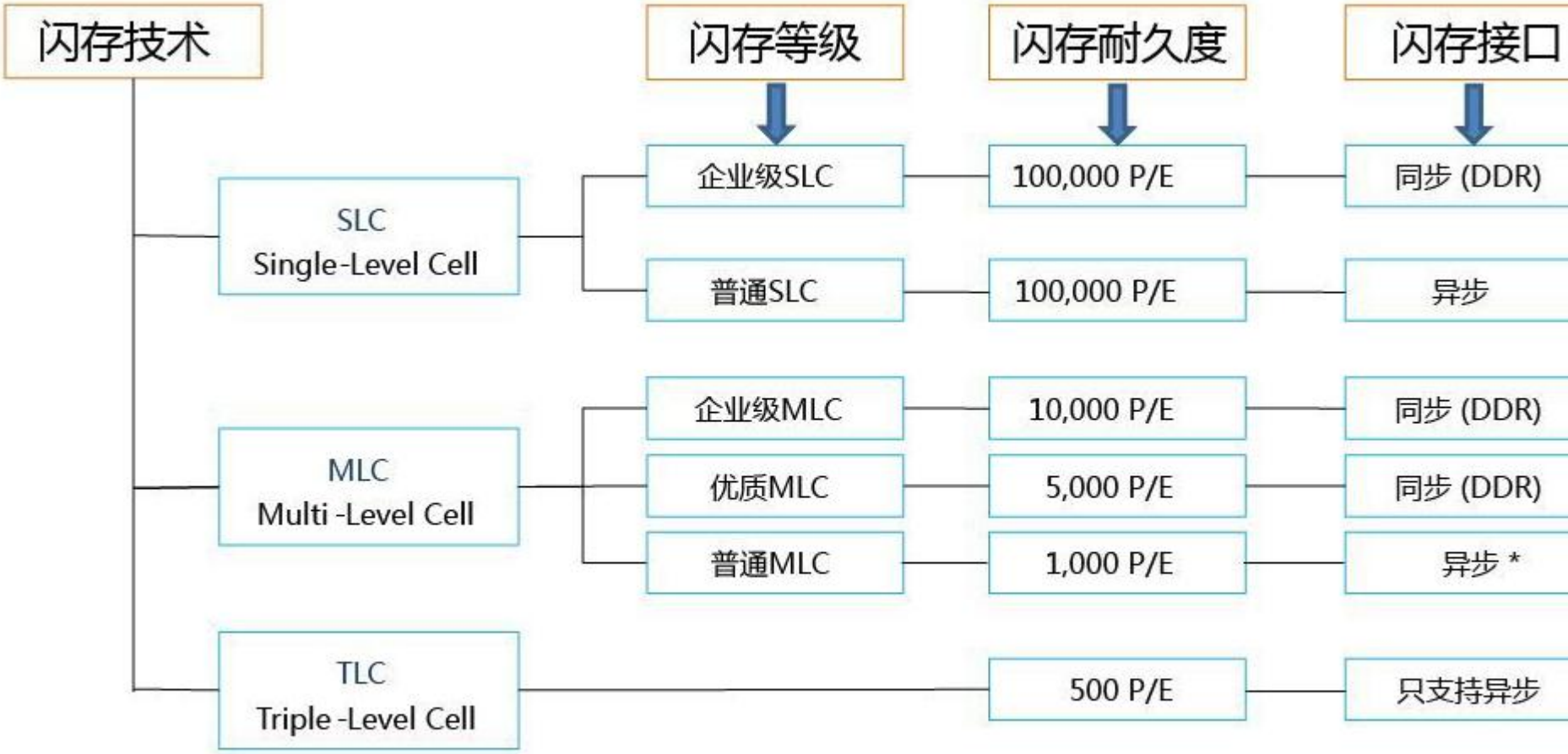






25nm制程NAND闪存芯片，
单个长方形为2GB，单个Die是8GB

Flash 颗粒-Intel/Micron



Intel 25nm 闪存产品分级

Flash 颗粒-Intel/Micron

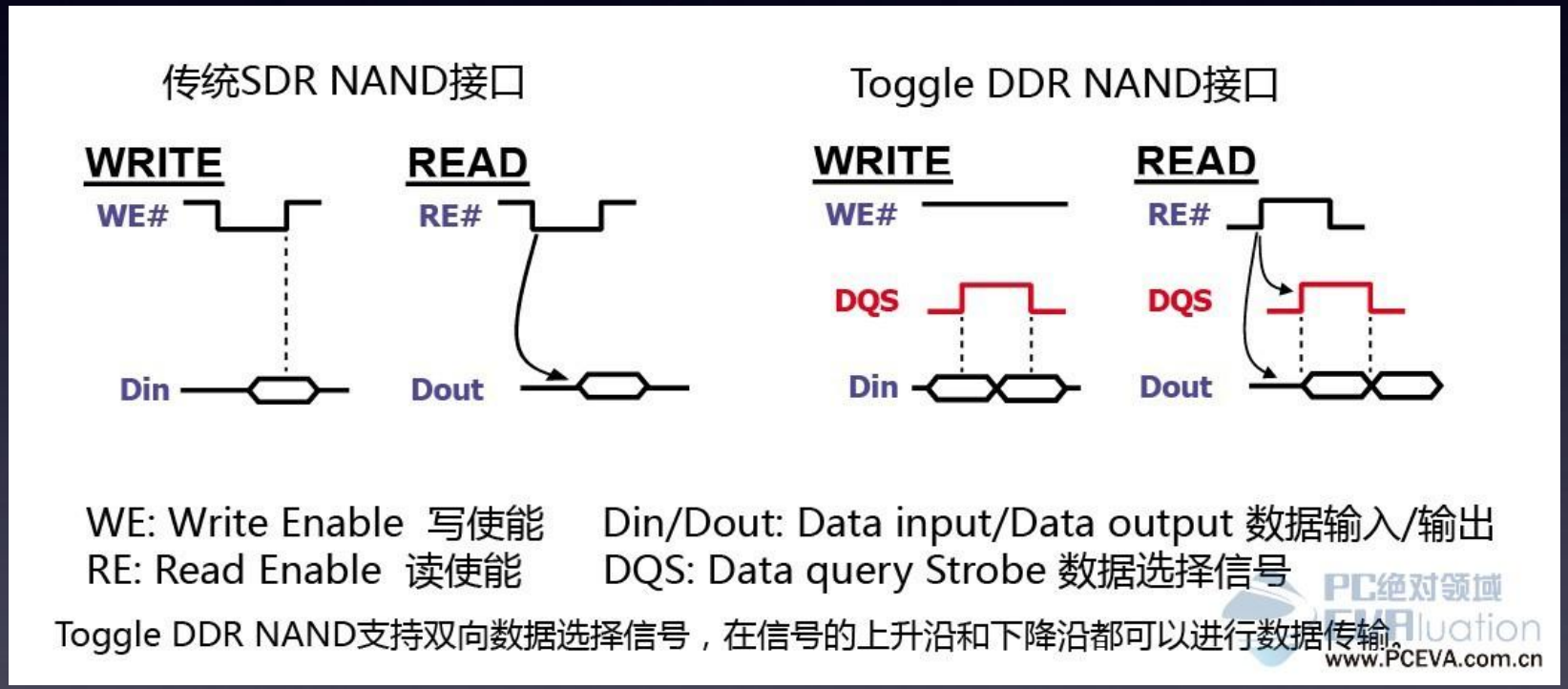
	JS29F32B08 JAME1	JS29F32B08 JCME1		JS29F32B08 JCME3	JS29F32B08 JCME2	PF29F32B08 NCME1		
	JS29F16B08 CAME1	JS29F16B08 CCME1		JS29F16B08 CCME3	JS29F16B08 CCME2	PF29F16B08 MCME1	PF29F16B16 NCNE1	PF29F16B NCNE2
	JS29F64G08 AAME1	JS29F64G08 ACME1	JS29F64G08 AAME2	JS29F64G08 ACME3	JS29F64G08 ACME2	PF29F64G08 LCME1	PF29F64G16 MCNE1	PF29F64G MCNE2
ONFI标准	2.2	2.2	2.2	2.2	2.2	2.2	2.2	2.2
芯片类型	MLC	MLC	MLC	E3 MLC	ccMLC	eMLC	eSLC	eSLC
块大小	256 pages	256 pages	256 pages	256 pages	256 pages	256 pages	128 pages	128 pag
页大小	8KB	8KB	8KB	8KB	8KB	8KB	8KB	8KB
同步模式	DC:11xx 同步	同步		同步	同步	同步	同步	同步
异步模式	DC:10xx 异步		异步					
耐久度	1000	1000	5000	3000	5000	10,000	100,000	100,000
接口通道	Single	Single	Single	Single	Single	Single/Dual x8	Dual x8	Dual x8
DieStacks	SDP/DDP/QDP	SDP/DDP/QDP	SDP	SDP/DDP/QDP	SDP/DDP/QDP	SDP/DDP/QDP	DDP/QDP	DDP/QD
封装类型	TSOP	TSOP	TSOP	TSOP	TSOP	152b BGA	152b BGA	100b BG
封装尺寸	12*20	12*20	12*20	12*20	12*20	14*18	14*18	12*18

产品价格 → 低 → 高

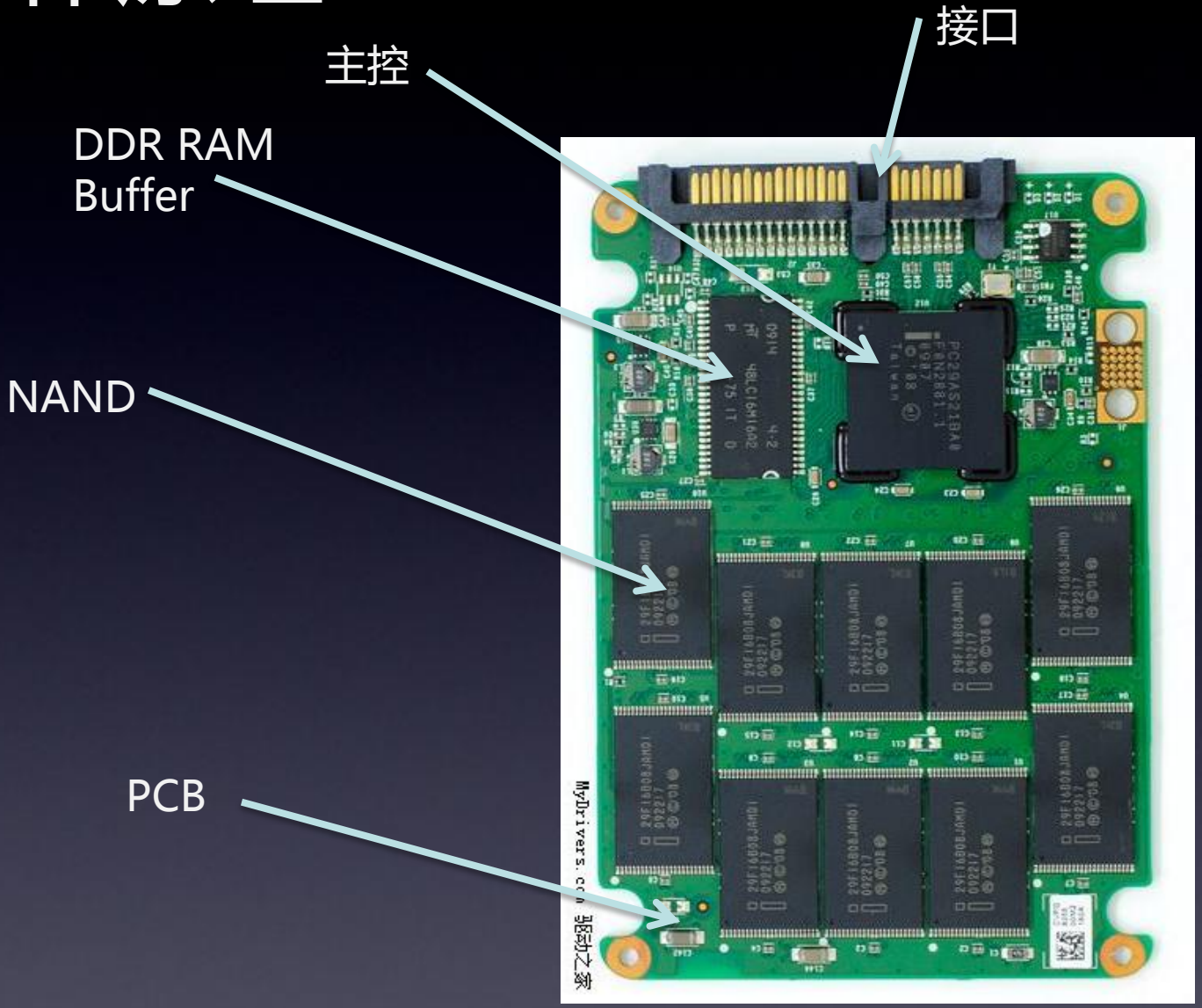
注：ccMLC=Client Compute Nand MLC；eMLC=Enterprise Compute Nand MLC；eSLC=Enterprise Compute Nand SLC；DC=生产日期代码

Flash 颗粒
三星+TOSHIBA

老牌NAND制造厂商三星的接口标准为OneNAND,而东芝的接口标准为LBA-NAND，这两家全球份额加起来接近70%



三、工作原理

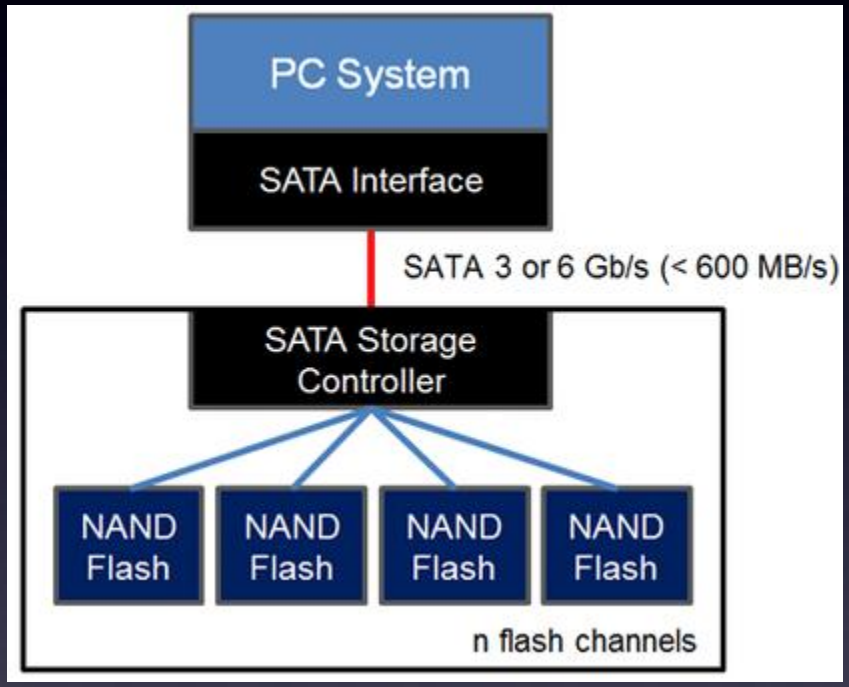


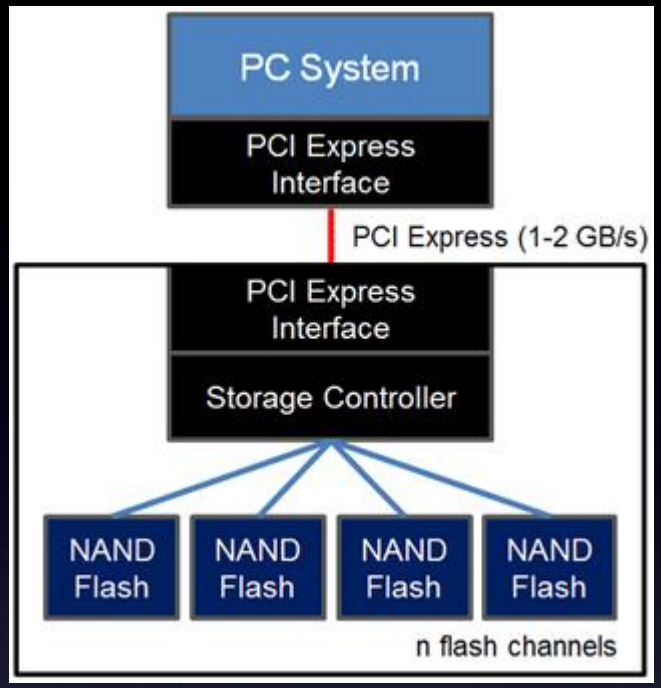
PCI-E接口产品

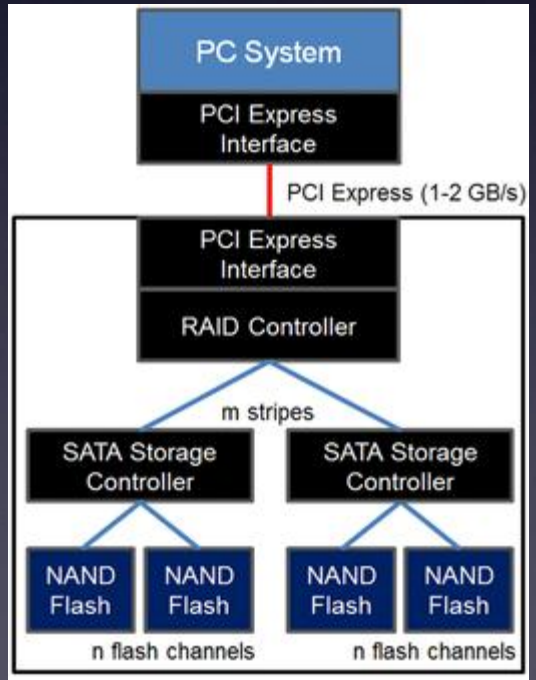
工作原理

1.读/写如何实现的
2. LBA(Logical Block Addressing)逻辑块寻址
3. FTL(Flash translation layer)闪存转换层
4. WL(Wear leveling) 磨损平衡
5. GC(Garbage collection) 垃圾回收
6. OP（Over-provisioning）预留空间
7. WA（Write amplification）写入放大
8.TRIM
9. BBM（Bad block management）坏块管理
10. ECC - 校验和纠错
11. Interleaving NAND 交叉存取技术
12. 品质和稳定性

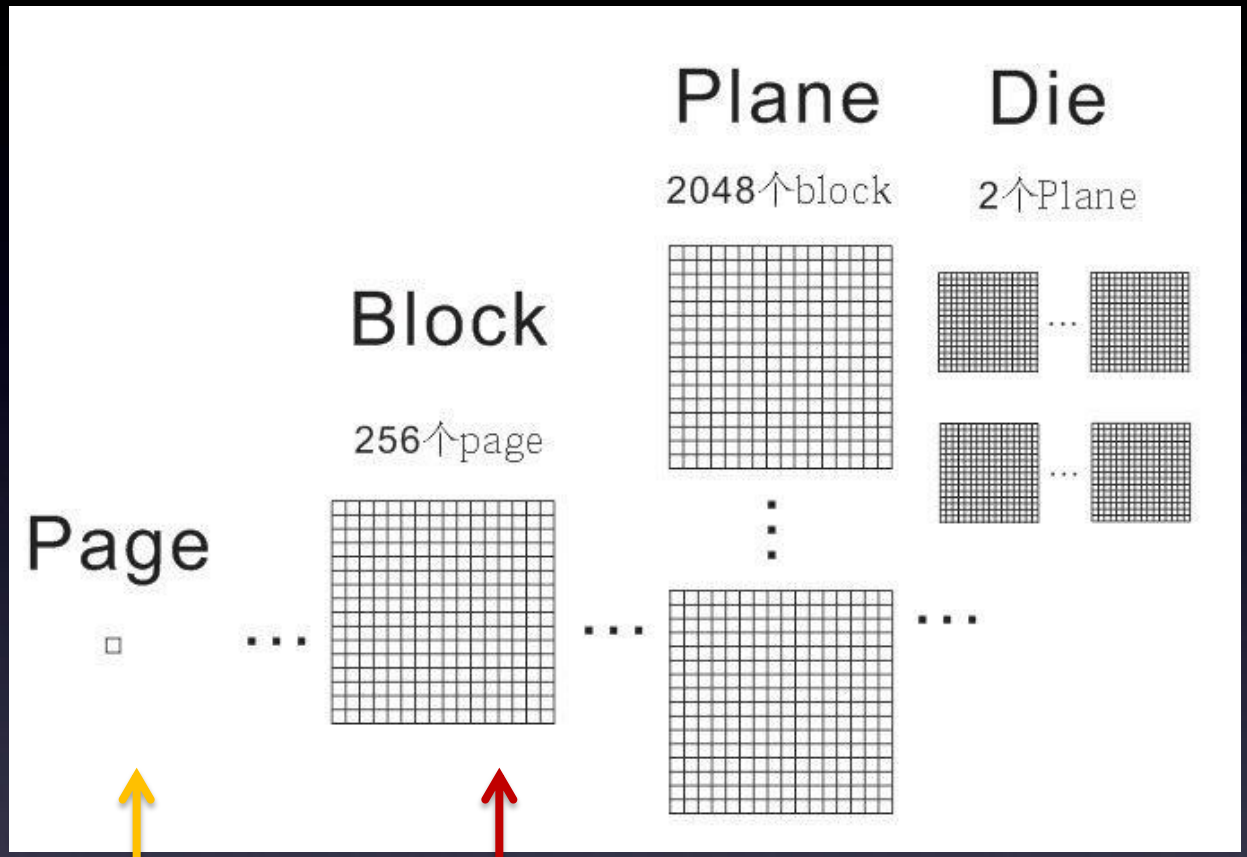
工作原理-1


PC System
SATA Interface
SATA 3 or 6 Gb/s (< 600 MB/s)
SATA Storage Controller
NAND Flash
NAND Flash
NAND Flash
NAND Flash
n flash channels


PC System
PCI Express Interface
PCI Express (1-2 GB/s)
PCI Express Interface
Storage Controller
NAND Flash
NAND Flash
NAND Flash
NAND Flash
n flash channels


PC System
PCI Express Interface
PCI Express (1-2 GB/s)
PCI Express Interface
RAID Controller
m stripes
SATA Storage Controller
SATA Storage Controller
NAND Flash
NAND Flash
NAND Flash
NAND Flash
n flash channels
n flash channels

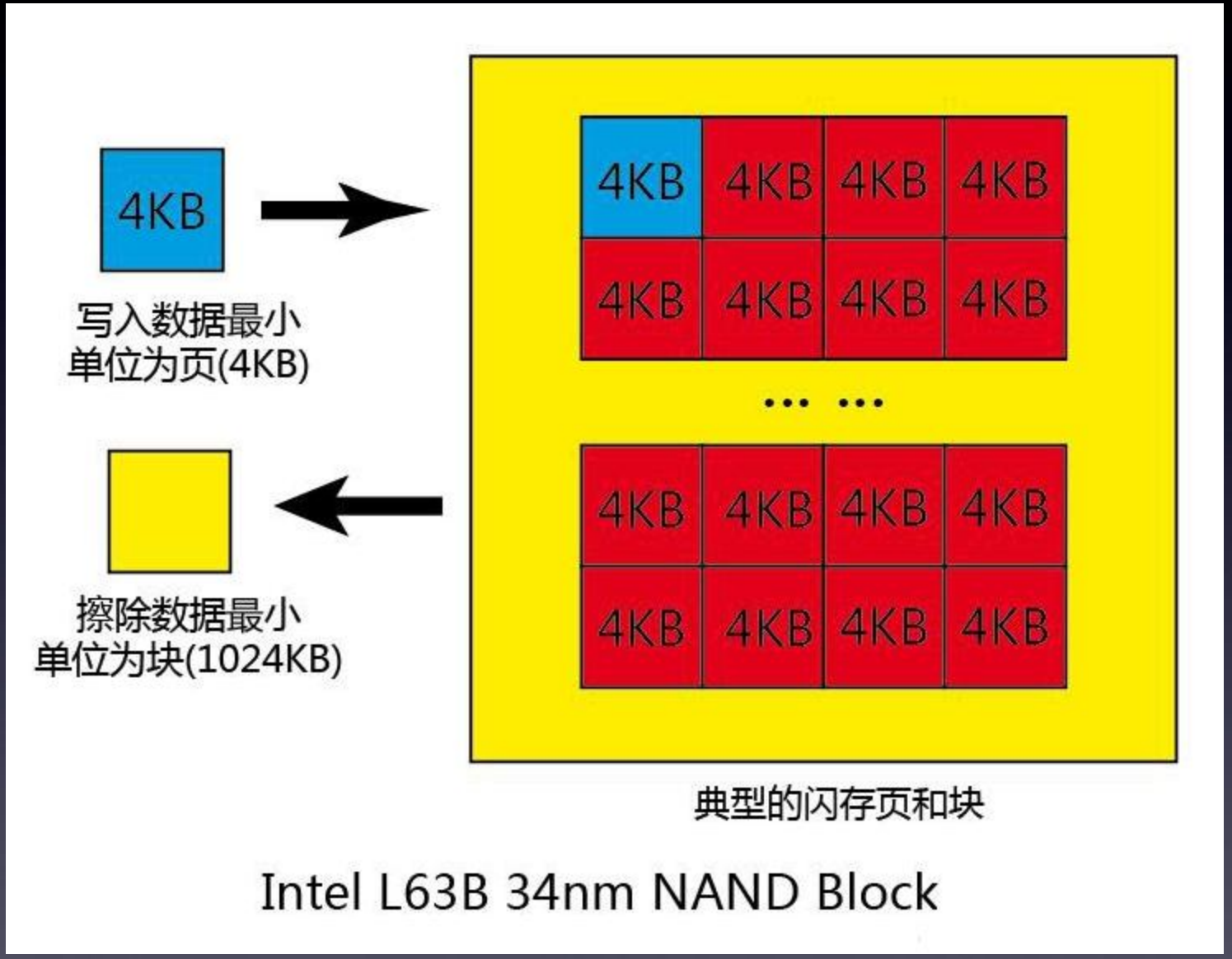
工作原理-NAND



最小
读
写
单位

最小
擦除
单位

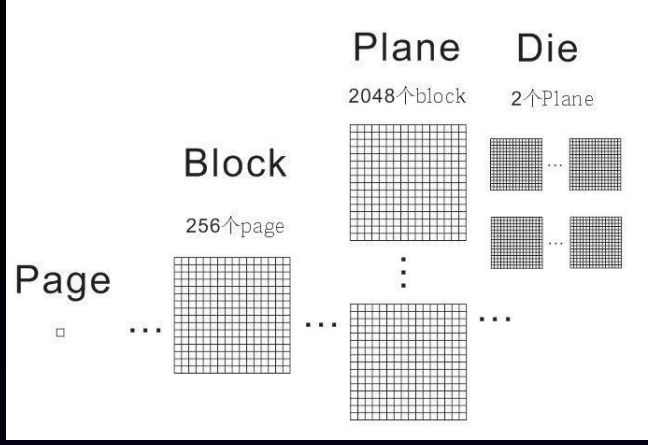
工作原理-NAND



Intel L63B 34nm NAND Block

工作原理

Page
- 4 Kbytes （intel X25-M G2）
- 8 Kbytes （intel 320 系列）



Block容量=4 KB * 256 = 1 MB

Plane容量= 1 MB × 2048 = 2 GB

Die 容量= 2G × 2 = 4 GB

工作原理-容量组成

Intel X25-V 40GB

1个 Flash Chip

Die1 4GB

+

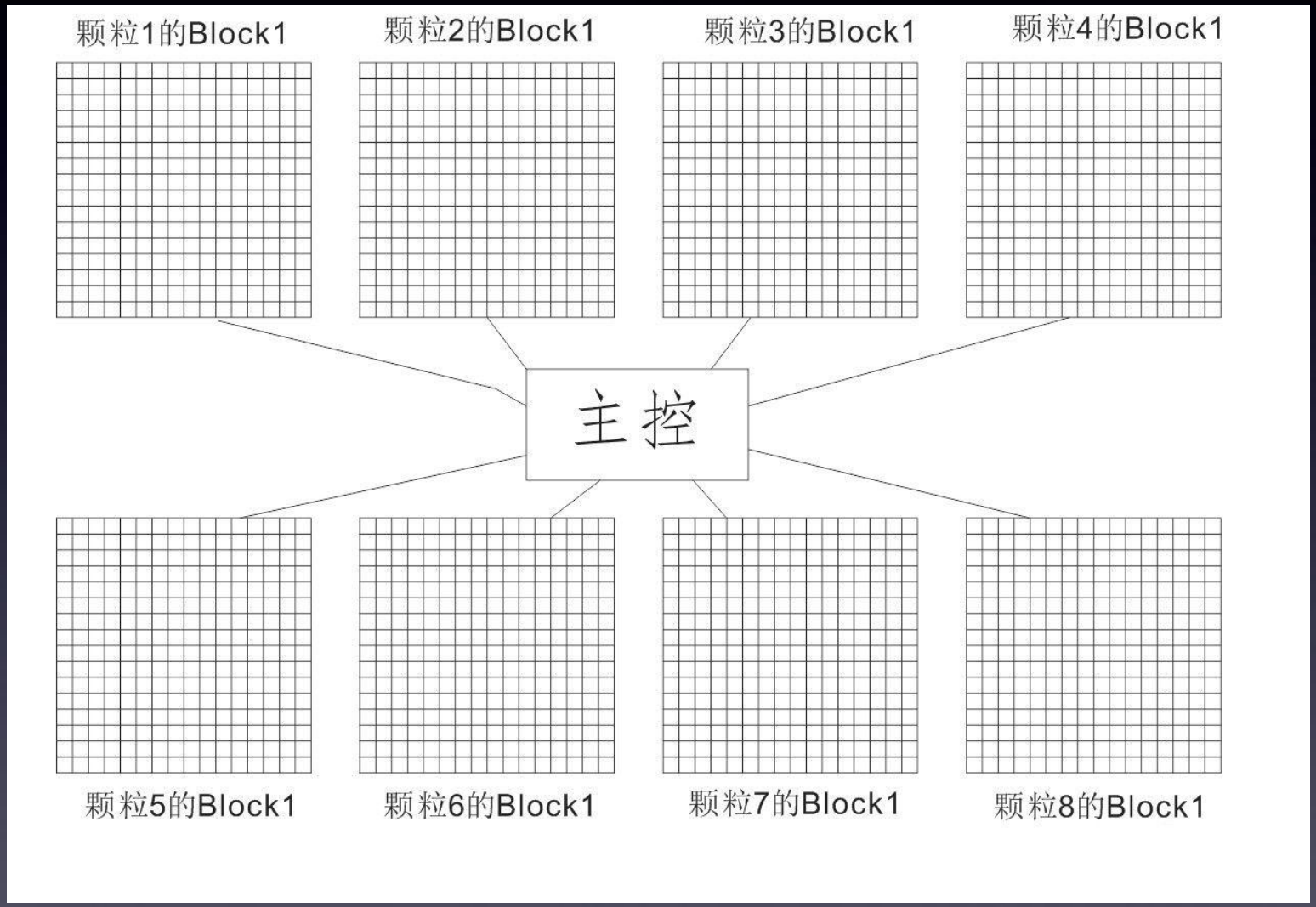
Die2 4GB

= 8 GB

*5

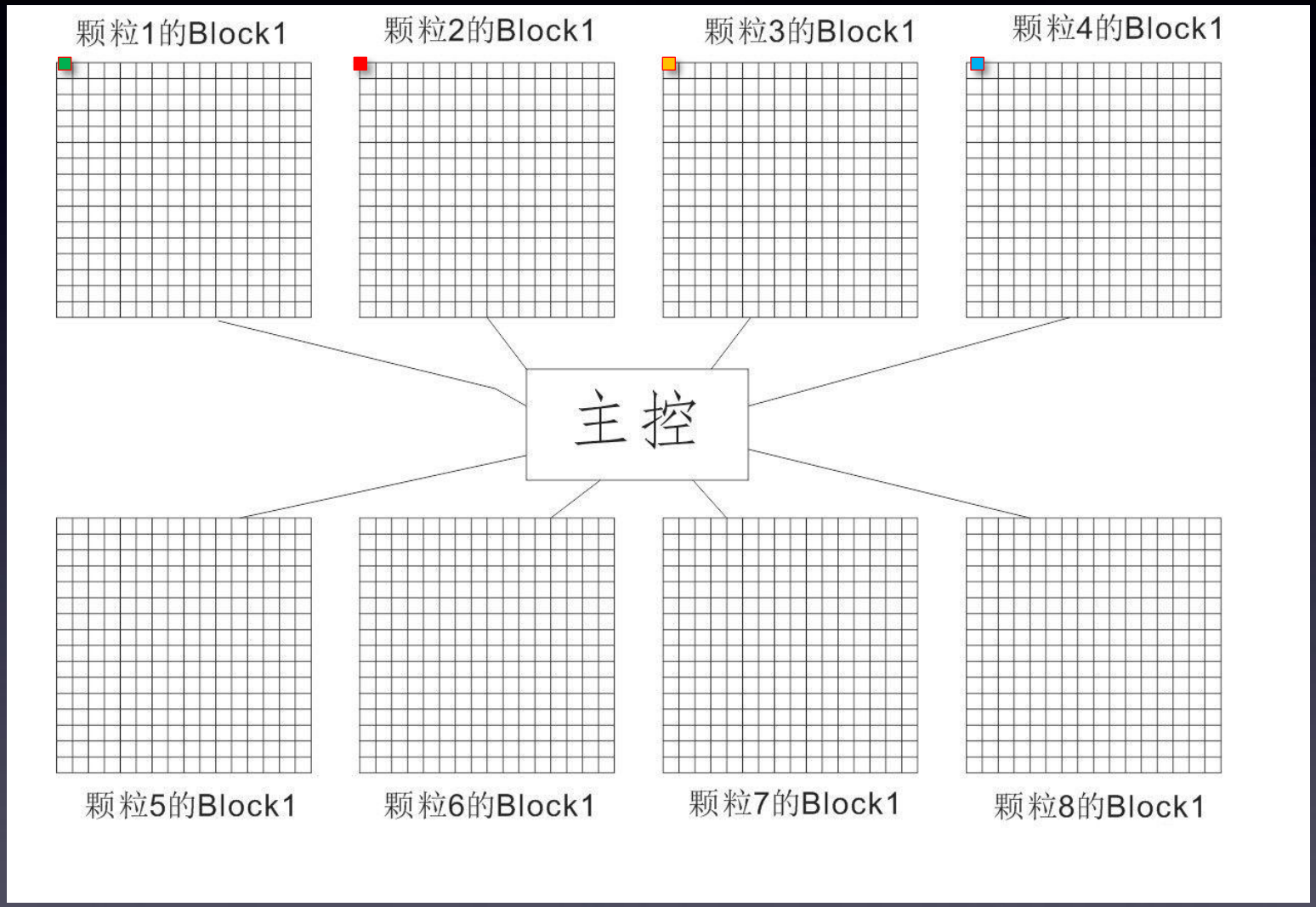
= 40 GB

工作原理-写入


颗粒1的Block1
颗粒2的Block1
颗粒3的Block1
颗粒4的Block1
主控
颗粒5的Block1
颗粒6的Block1
颗粒7的Block1
颗粒8的Block1

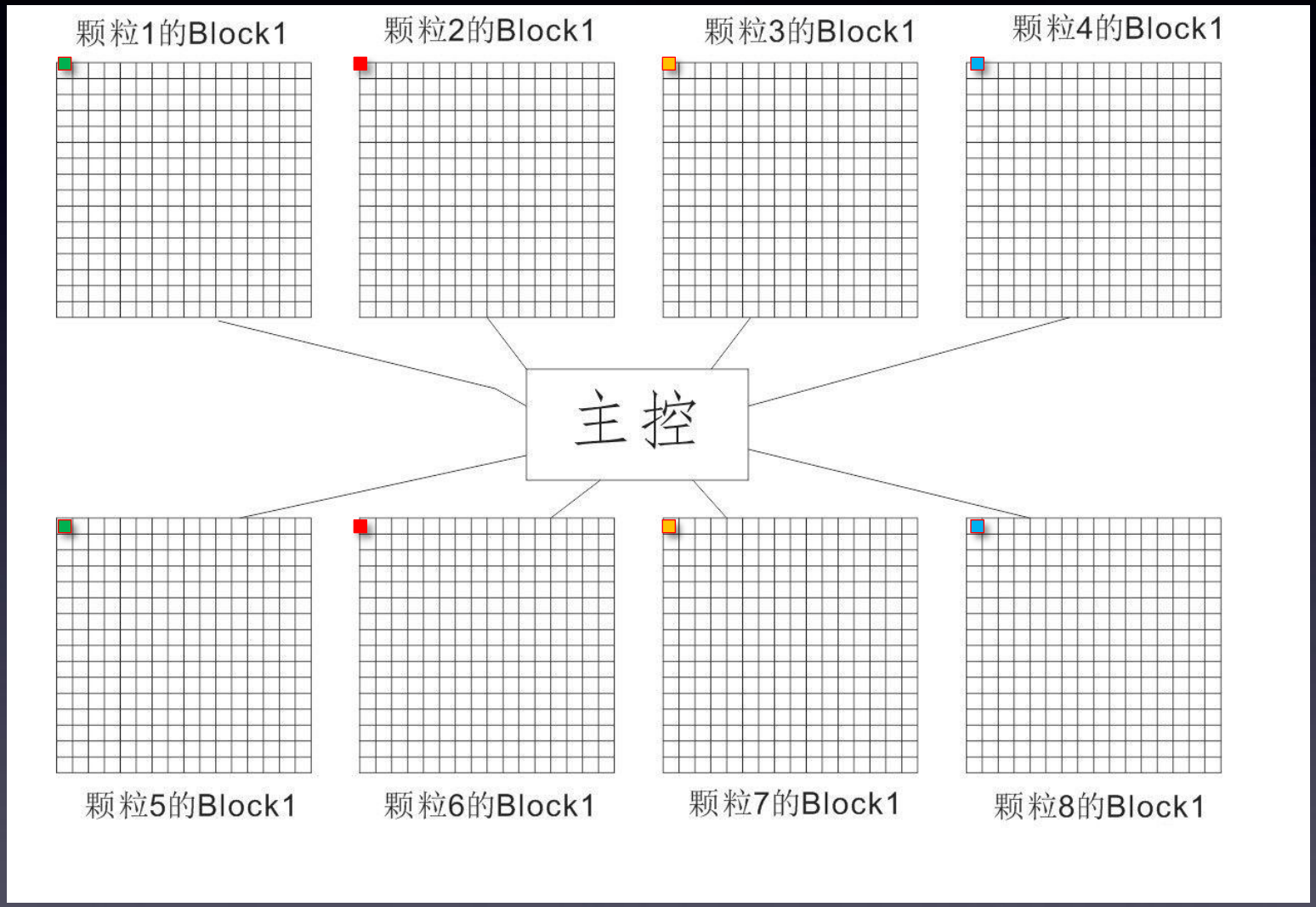
工作原理-读取

OS


颗粒1的Block1
颗粒2的Block1
颗粒3的Block1
颗粒4的Block1
主控
颗粒5的Block1
颗粒6的Block1
颗粒7的Block1
颗粒8的Block1

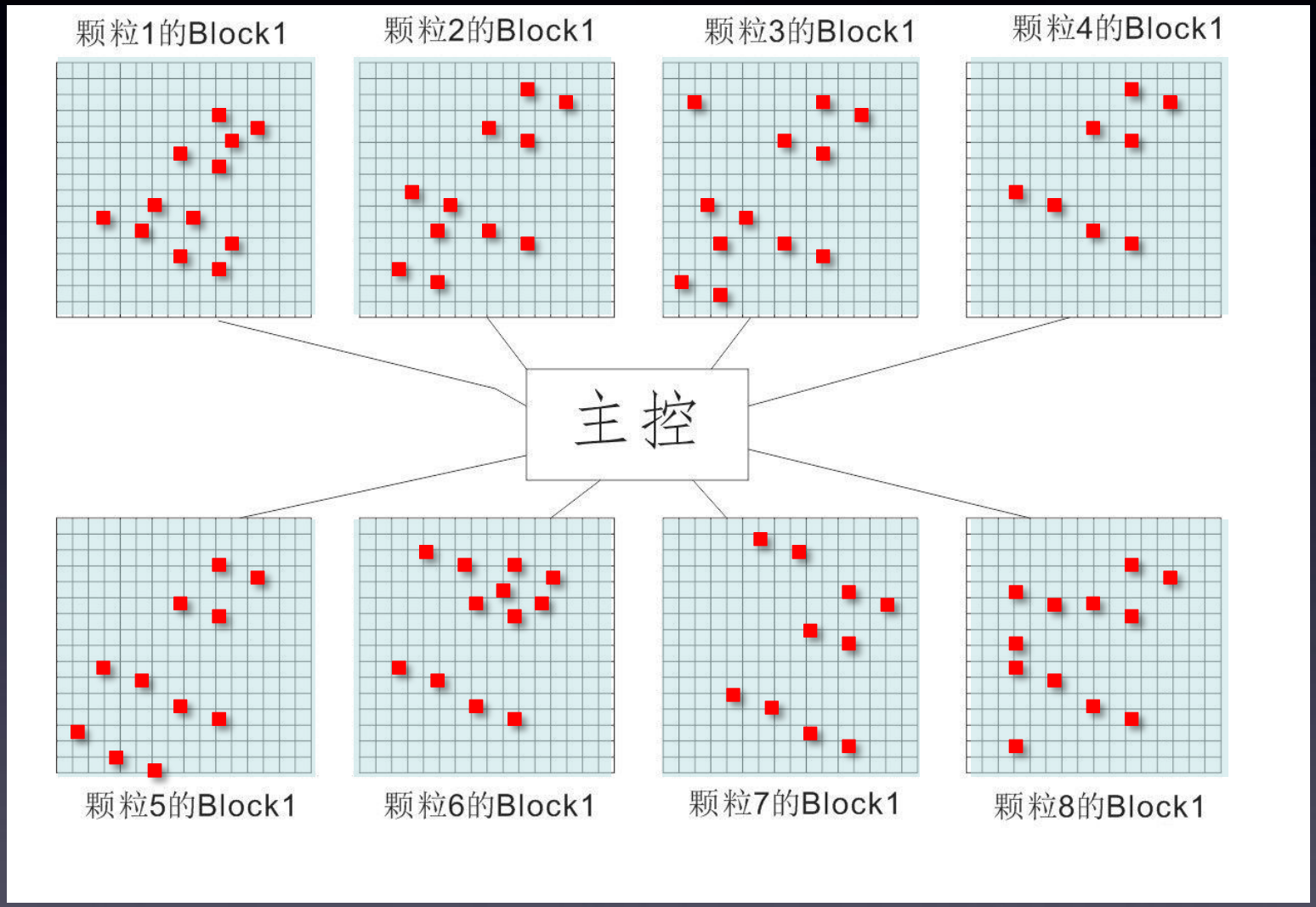
工作原理-读取-最大性能

OS



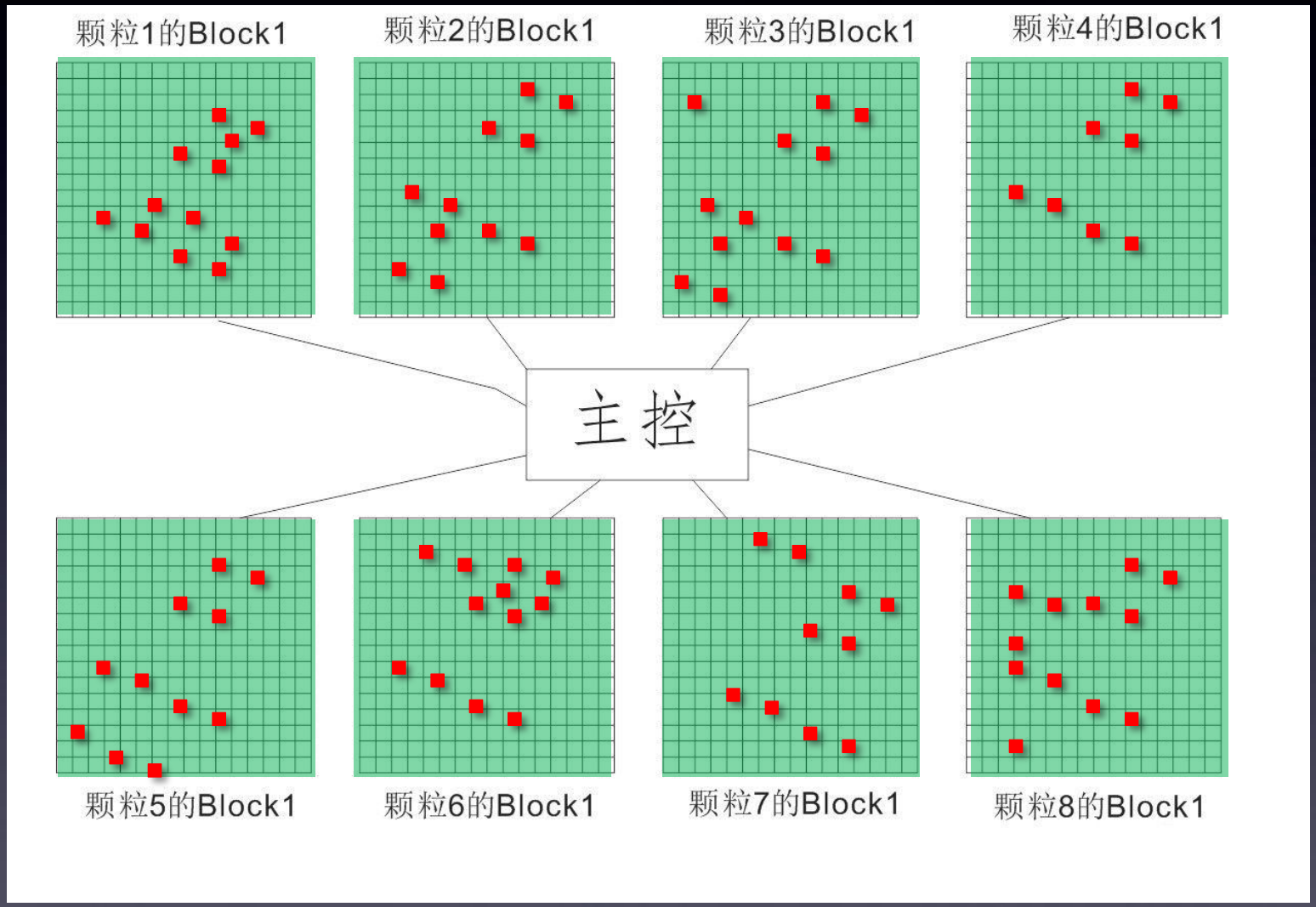
工作原理-删除-OS视角

有效数据块　　无效块（因删除产生的）



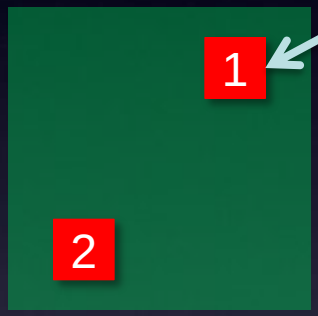
工作原理-删除-SSD视角

有效数据块　　无效块（因删除产生的）



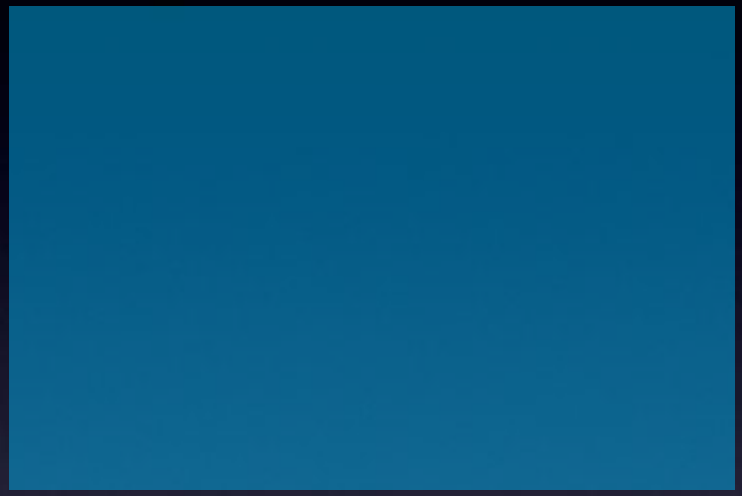
工作原理-数据更新

有效数据块　　无效块（因删除产生的）



Page-1
4 KB

Block-1
1 MB

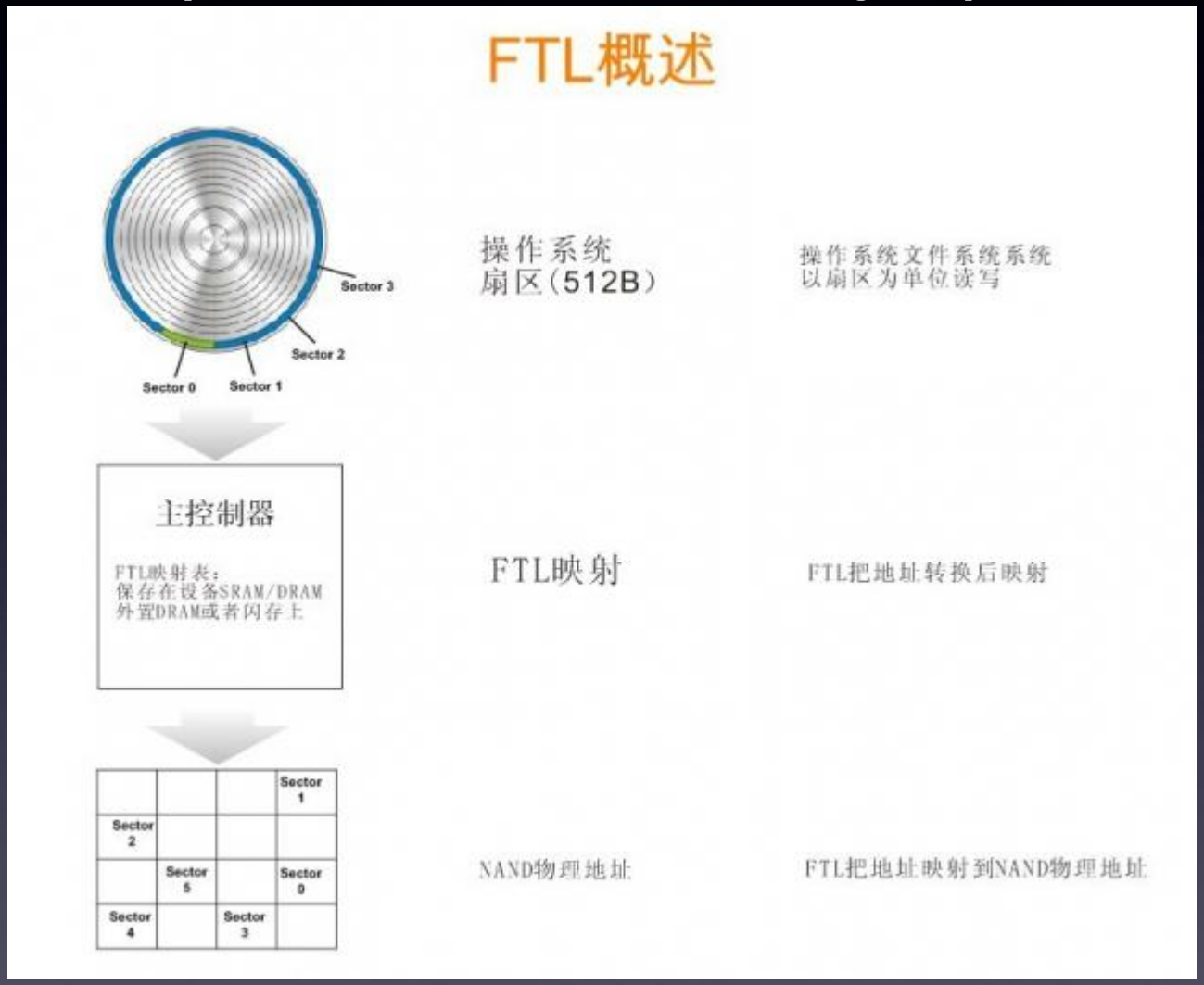


SSD 主控

NEW Block
1 MB

工作原理-3

- FTL(Flash translation layer) 闪存转换层



工作原理-4 - WL(Wear leveling) 磨损平衡

WL(Wear leveling) 磨损平衡 - 确保闪存的每个块被写入的次数相等的一种机制。

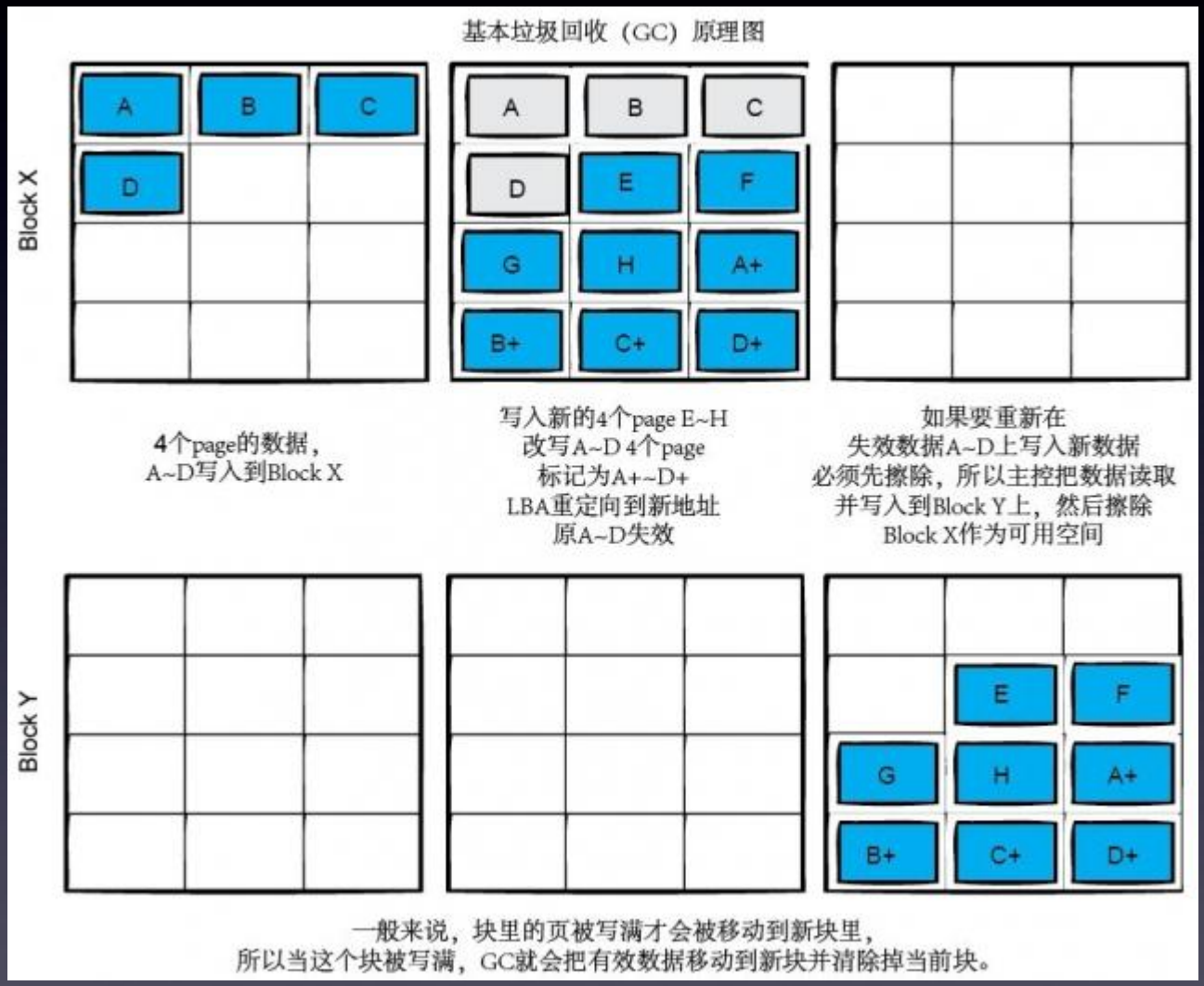
通常情况下，在NAND块里的数据更新速度是不同的：有些会经常更新，有些则不常更新很明显，那些经常更新的数据所占用的块会被快速的磨损掉，而不常更 新的数据占用的磨损就小得多。为了解决这个问题，需要让每个块的编程次数尽可能保持一致：这就是要对每个页的读取/编程操作进行监测，在最乐观的情况 下，这个技术会让全盘的颗粒物理磨损程度接近并同时报废。

磨损平衡技术依赖于逻辑和物理地址的转换：也就是说，每次主机上的应用程序请求相逻辑页地址时，内存控制器动态的映射逻辑页地址到另一个不同的物理页地 址，并把这映射的指向存放在一个特定的”映射表“里。而之前过期的物理页地址就被标记为”无“并等待之后的擦除操作。这样一来，所有的物理块就能被控 制在一个相同磨损范围，同时”老化“。

磨损平衡算法分静态和动态。动态磨损算法是基本的磨损算法：只有用户在使用中更新的文件占用的物理页地址被磨损平衡了。而静态磨损算法是更高级的磨损算 法：在动态磨损算法的基础上，增加了对于那些不常更新的文件占用的物理地址进行磨损平衡，这才算是真正的全盘磨损平衡。目前家用SSD主控支持静态磨损算 法的有：SandForce主控和Marvell 88SS9174-BJP2主控（也就是镁光C300用的主控制器）

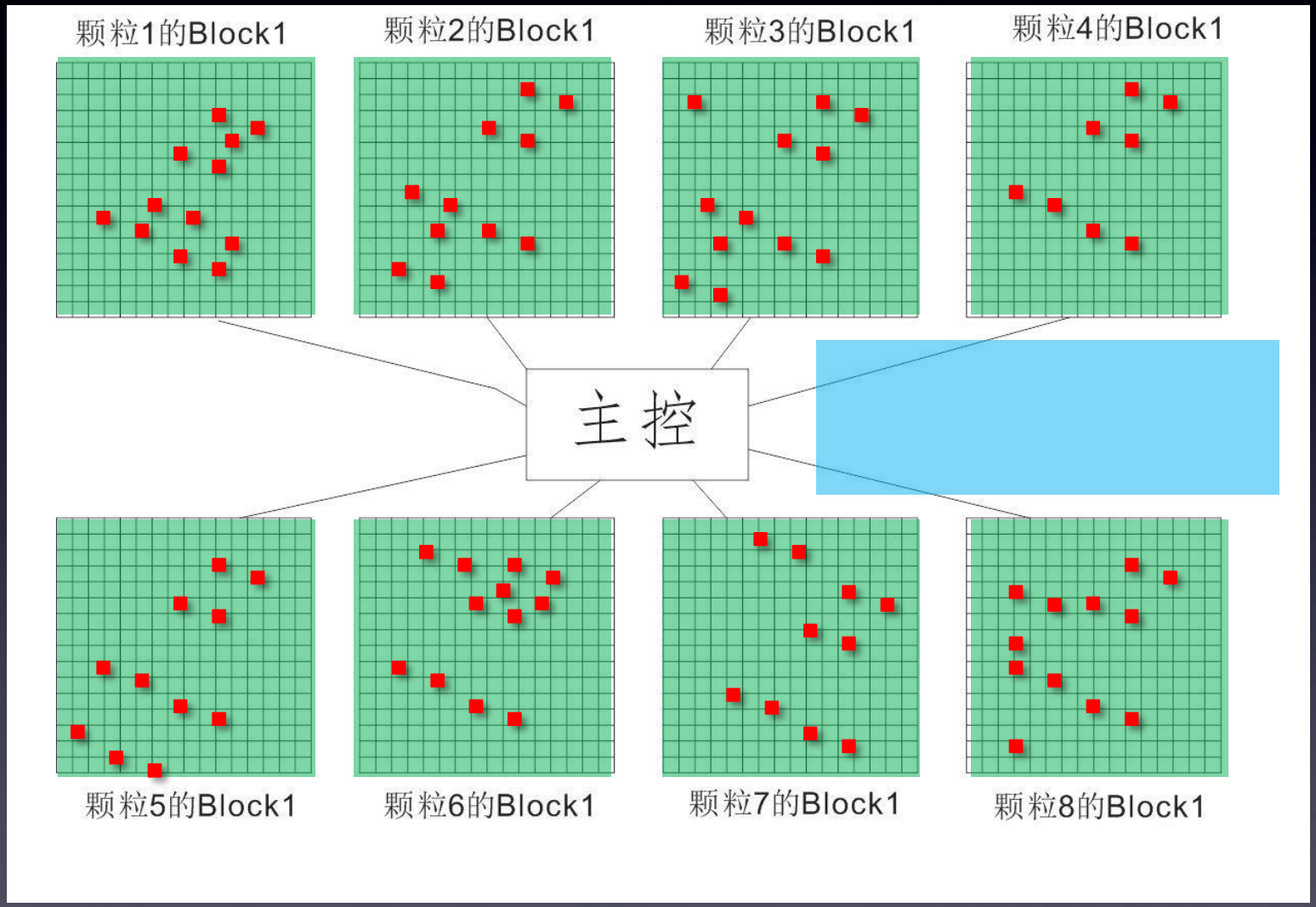
工作原理-5 GC(Garbage collection) 垃圾回

- NAND颗粒“清洁工”



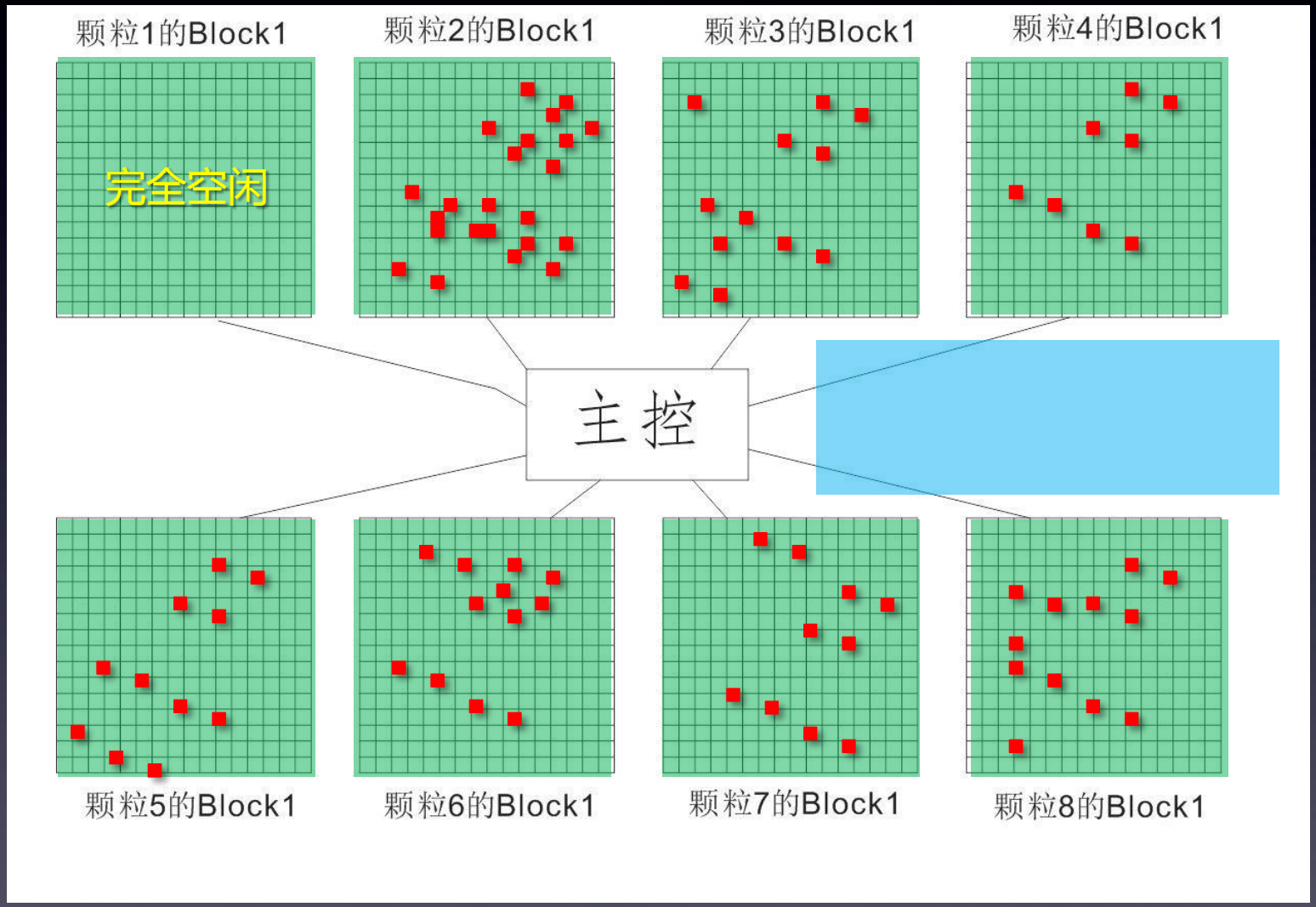
工作原理-删除-GC

有效数据块　无效块　预留块



工作原理-删除-GC

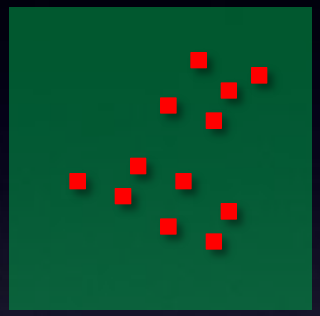
有效数据块　无效块　预留块



工作原理-删除-GC的时机

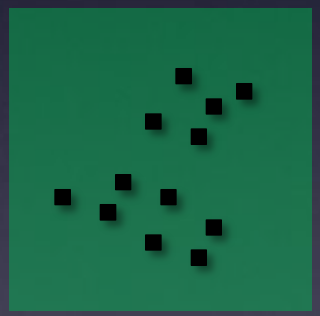
有效数据块

需要删除的数据Page



TRIM后马上GC

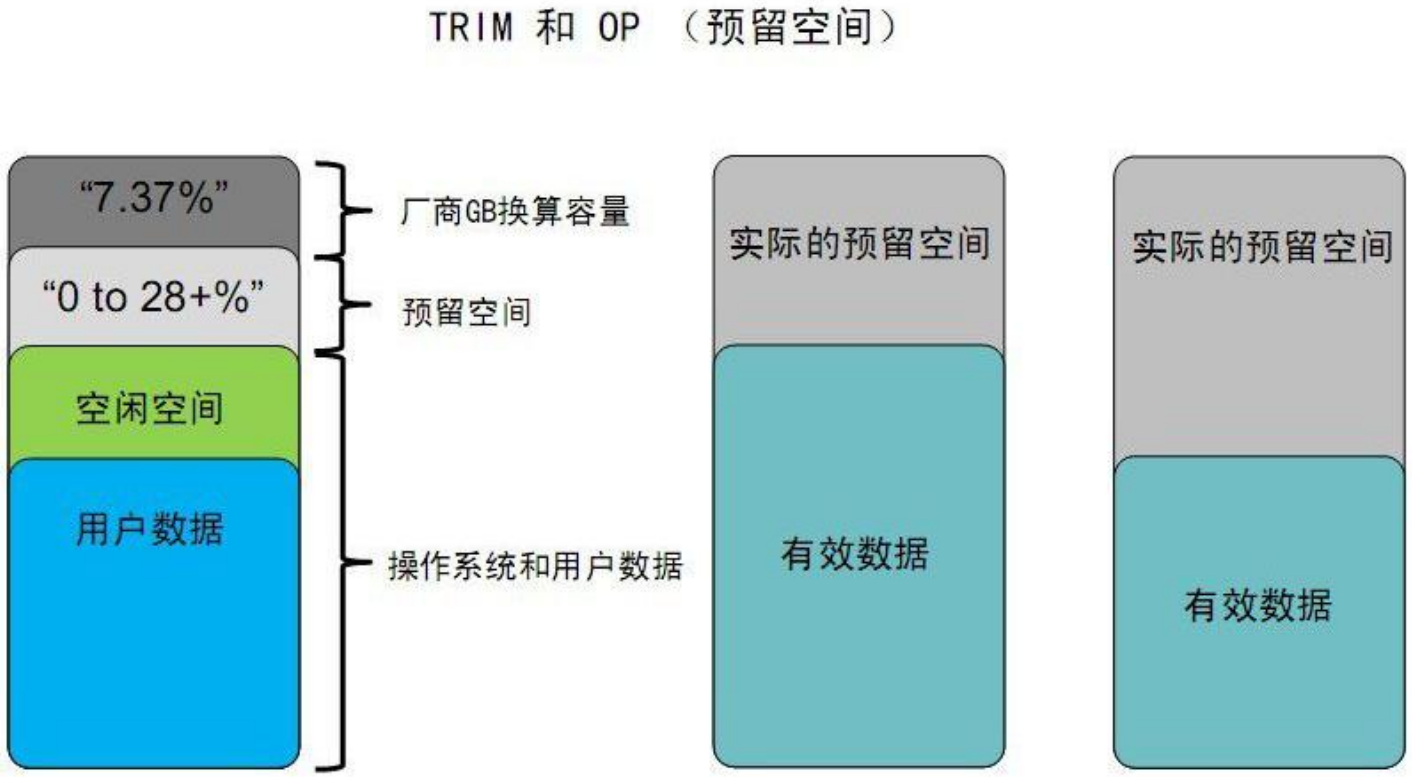
SSD 主控



TRIM后等待最佳时机再GC

SSD 主控

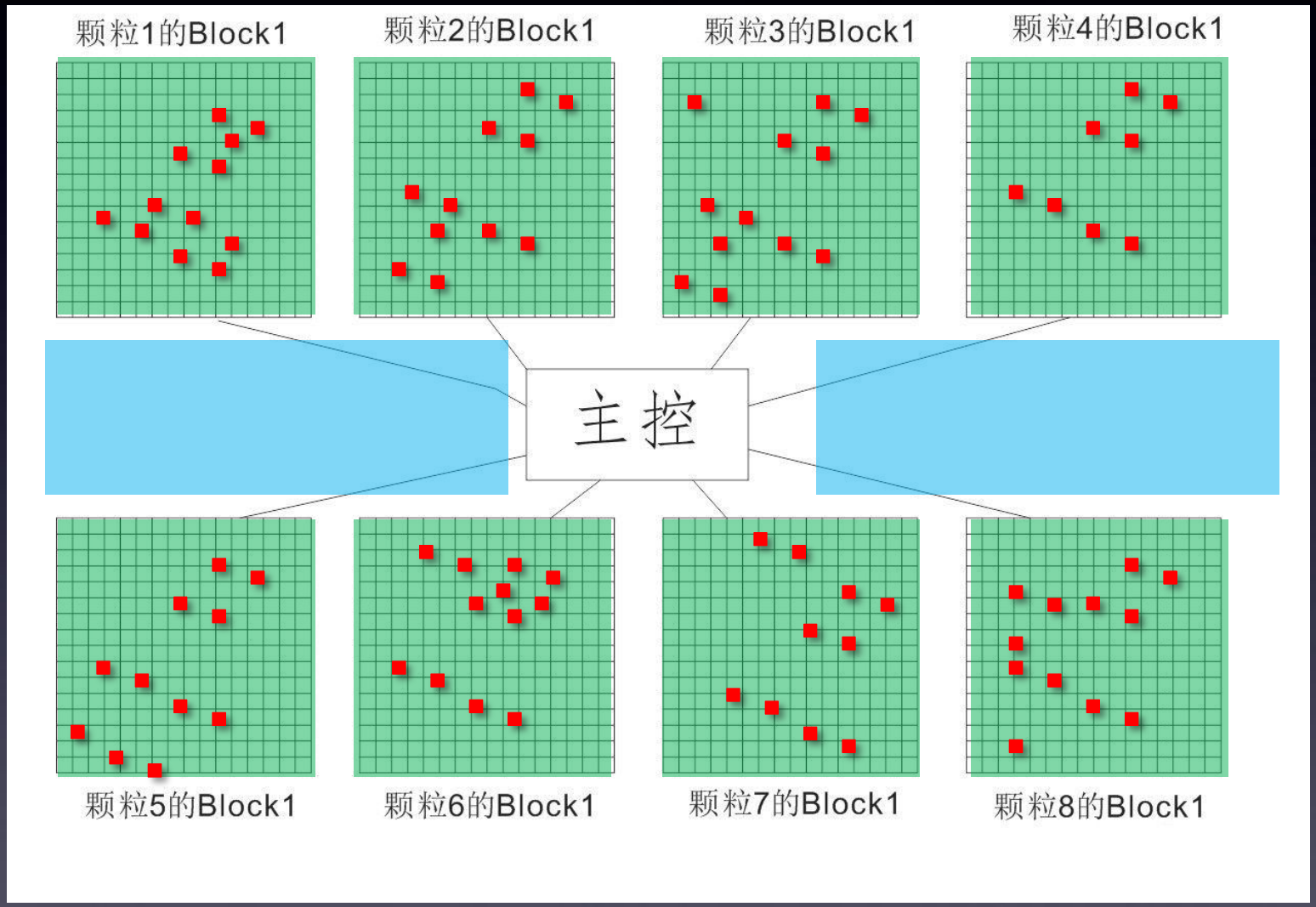
工作原理- 6
（Over-provisioning）预留空间



- 写入放大越低
- 性能越强
- 寿命越长

工作原理-OP（Over-Provisioning）

有效数据块　无效块　预留块



工作原理-7

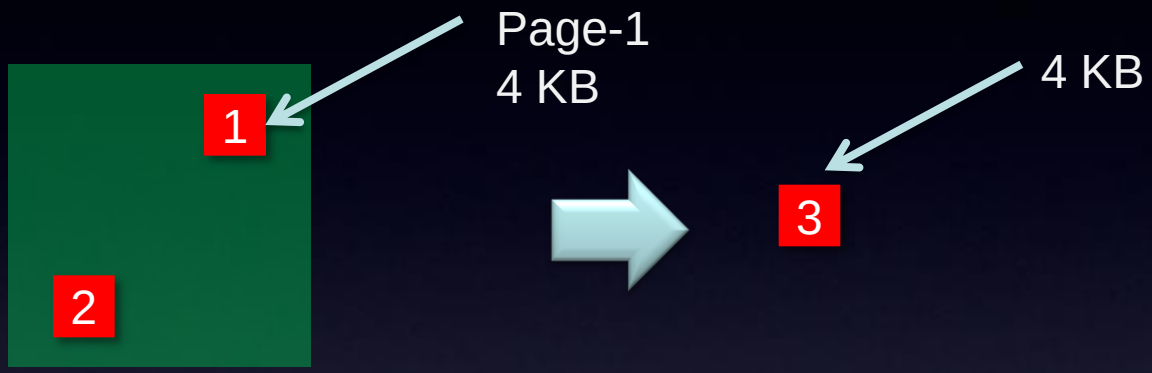
- WA（Write amplification）写入放大

最简单的例子，比如我要写入一个4KB的数据，最坏的情况就是，一个块里已经没有干净空间了，但是有无效数据可以擦除，所以主控就把所有的数据读到缓存， 擦除块，缓存里更新整个块的数据，再把新数据写回去，这个操作带来的写入放大就是: 我实际写4K的数据，造成了整个块（1024KB）的写入操作，那就是256倍放大。同时带来了原本只需要简单的写4KB的操作变成闪存读取 (1024KB)，缓存改（4KB)，闪存擦(1024KB)，闪存写(1024KB)，造成了延迟大大增加，速度慢是自然了。所以说写入放大是影响 SSD随机写入性能和寿命的关键因素。

用100%随机4KB来写入SSD，目前的大多数SSD主控，在最坏的情况下写入放大可以达到20以上。如果是100%持续的从低LBA写到高LBA的 话，写入放大可以做到1，实际使用中写入放大会介于2者之间。用户还可以设置一定的预留空间来减少写入放大，假设你有个128G的SSD，你只分了64G 的区使用，那么最坏情况下的写入放大就能减少约3倍。

工作原理-写入放大

有效数据块　　无效块（因删除产生的）



$$\text{写入放大} = \frac{\text{实际擦写数据量}}{\text{更新数据}} = \frac{1024\ \text{KB}}{4\ \text{KB}} = 256$$

工作原理-8 - TRIM指令

Trim - 一个ATA指令，由操作系统发送给SSD主控制器，告诉它哪些数据占的地址是”无效“的。

要明白什么是Trim和为什么它很重要，需要先知道一点文件系统的知识。

当你在电脑里删除一个文件的时候，操作系统并不会真正的去删除它。操作系统只是把这个文件地址标记为“空”，可以被再次使用，这意味着这个文件占的地址已 经是“无效”的了。这就会带来一个问题，硬盘并不知道操作系统把这个地址标记为”空“了，机械盘的话无所谓，因为可以直接在这个地址上重新覆盖写入，但是 到了SSD上问题就来了。

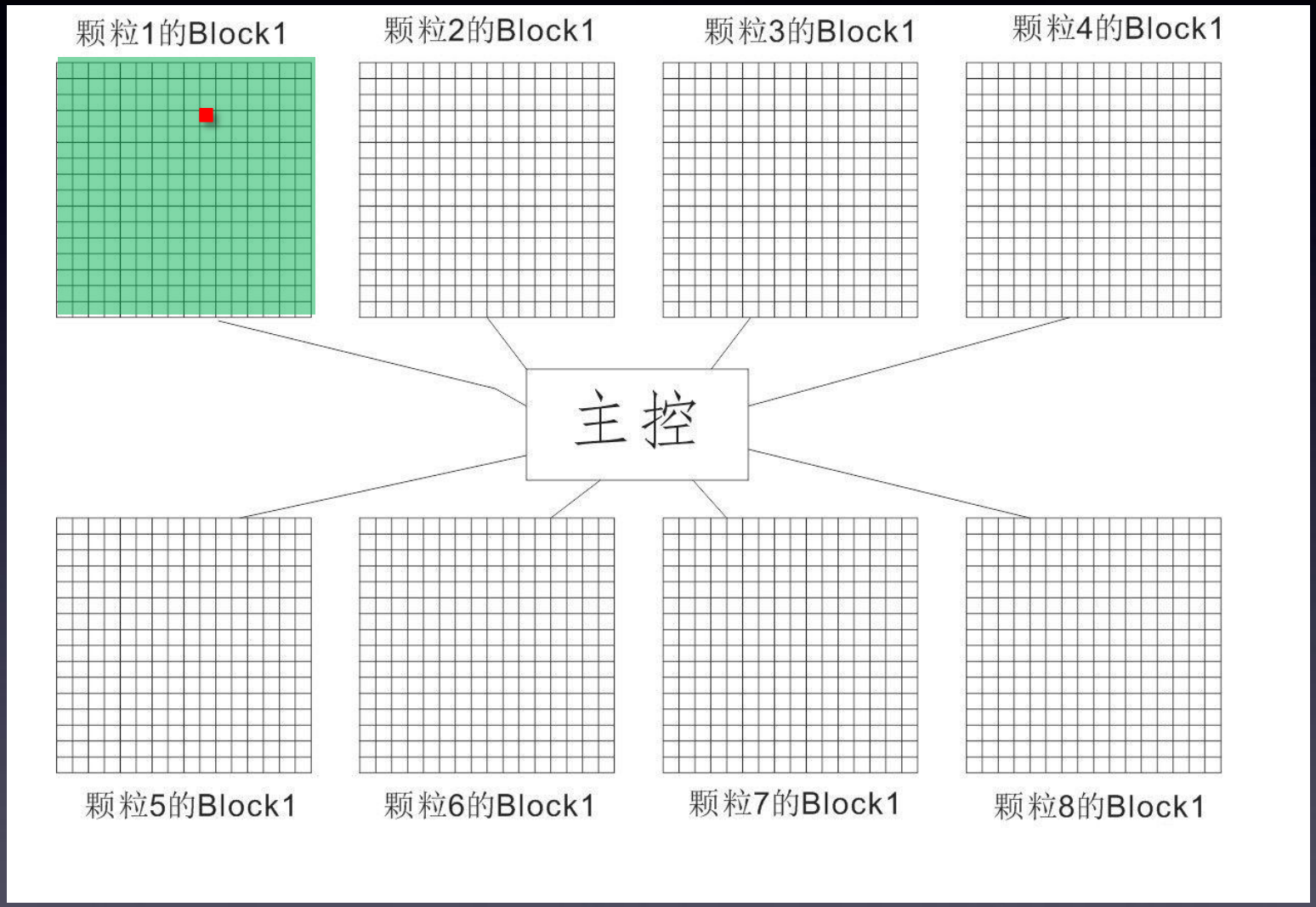
NAND需要先擦除才能再次写入数据，要得到空闲的NAND空间，SSD必须复制所有的有效页到新的空闲块里，并擦除旧块（垃圾回收）。如果没有Trim，意味着SSD主控制器不知道这个页是”无效“的，除非再次被操作系统要求覆盖上去。

Trim只是条指令，让操作系统告诉SSD主控制器这个页已经”无效“了。Trim会减少写入放大，因为主控制器不需要复制”无效“的页（没Trim就是”有效“的）到空白块里，这同时代表复制的"有效”页变少了，垃圾回收的效率和SSD性能也提升了。

Trim能大量减少“有效”页的数量，它能大大提升垃圾回收的效率。

工作原理-删除-TRIM原理

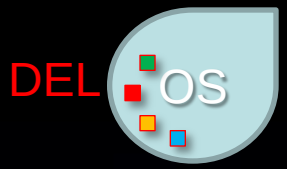
有效数据块　需要删除的数据Page

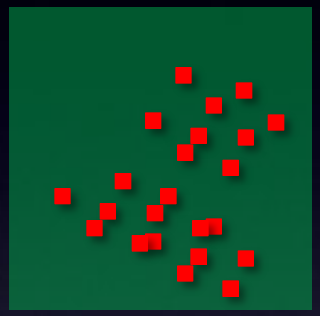


工作原理-删除-TRIM优势

有效数据块

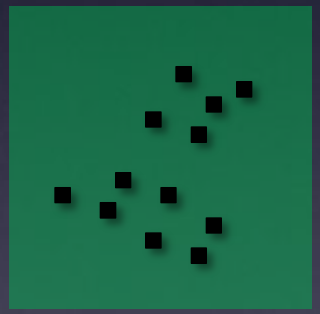
需要删除的数据Page





不使用TRIM

SSD 主控

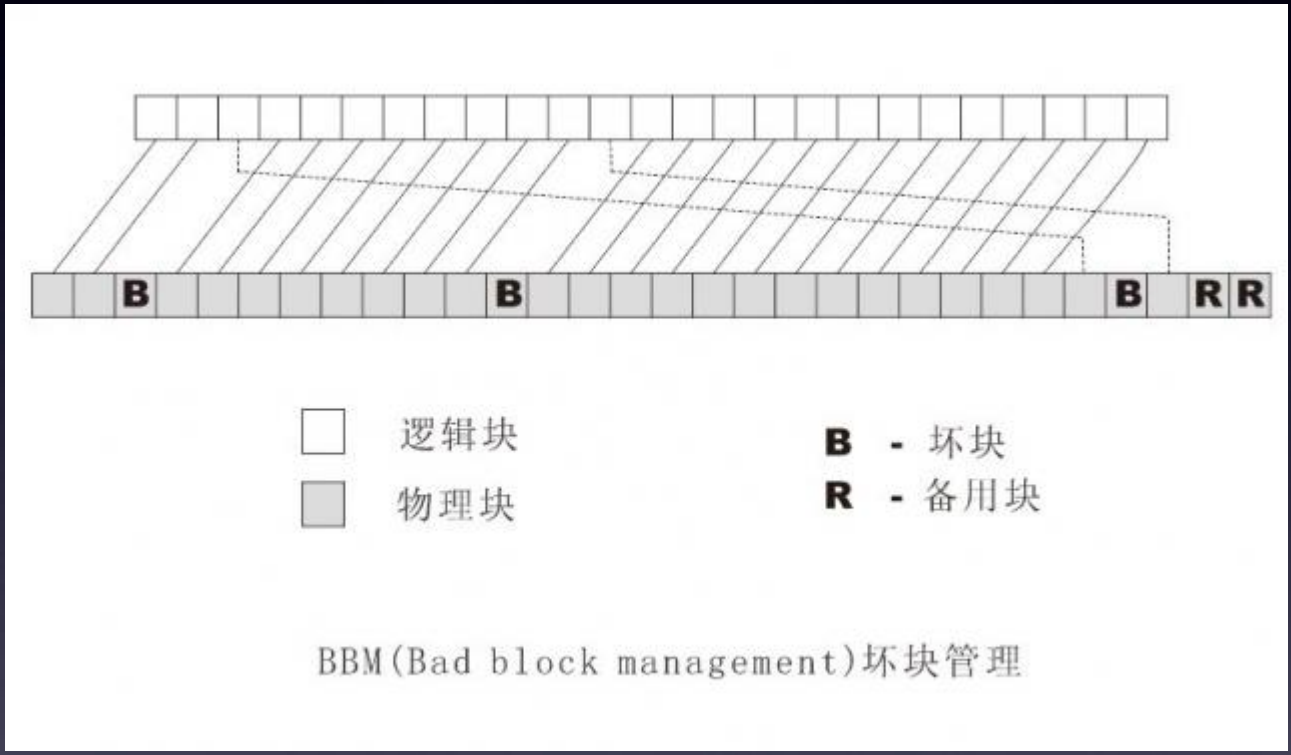


使用TRIM

SSD 主控

工作原理-9

- BBM（Bad block management）坏块管理



工作原理-10- ECC - 校验和纠错

ECC的全称是Error Checking and Correction，是一种用于Nand的差错检测和修正算法。由于NAND Flash的工艺不能保证NAND在其生命周期中保持性能的可靠，因此，在NAND的生产中及使用过程中会产生坏块。为了检测数据的可靠性，在应用 NAND Flash的系统中一般都会采用一定的坏区管理机制，而管理坏区的前提是能比较可靠的进行坏区检测。如果操作时序和电路稳定性不存在问题的话，NAND Flash出错的时候一般不会造成整个Block或是Page不能读取或是全部出错，而是整个Page中只有一个或几个bit出错，这时候ECC就能发挥 作用了。不同颗粒有不同的基本ECC要求，不同主控制器支持的ECC能力也不同，理论上说主控越强ECC能力越强。

工作原理-11

- Interleaving - NAND 交叉存取技术

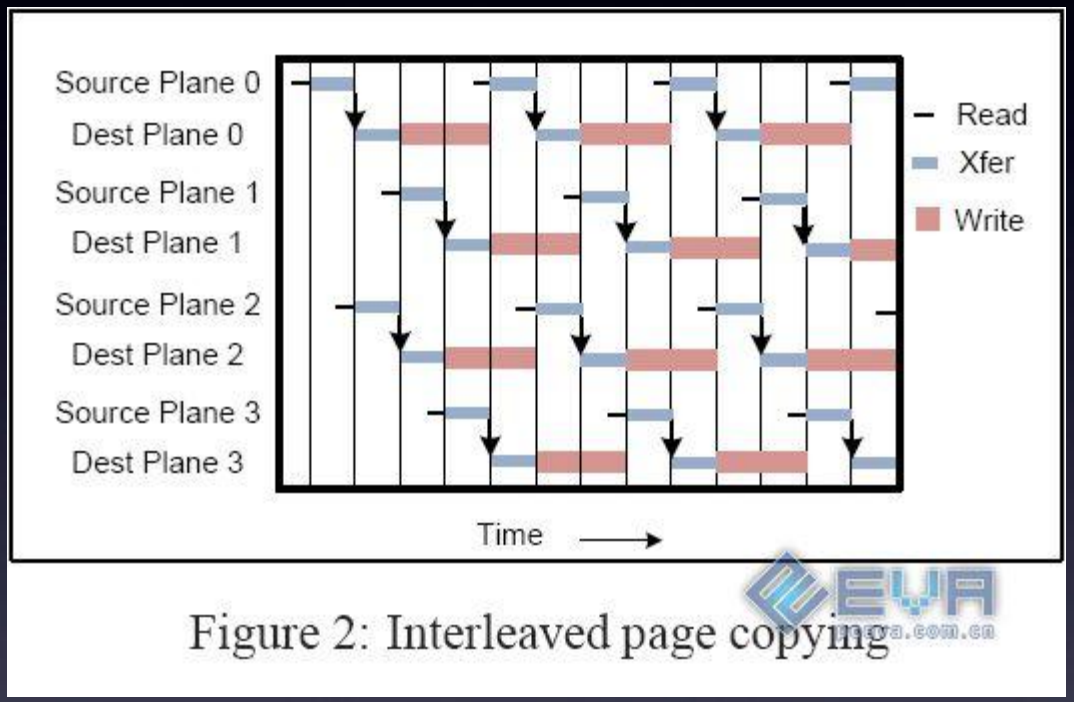
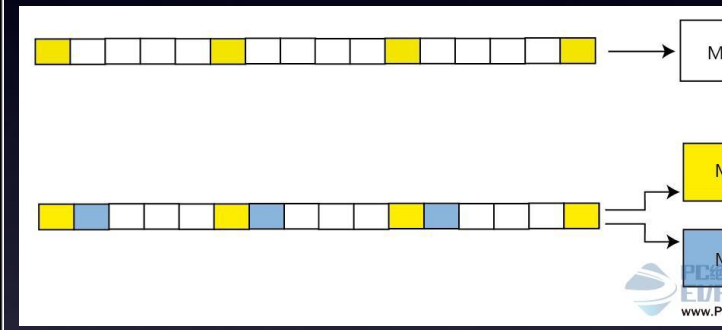


Figure 2: Interleaved page copying



交错操作可以成倍提升NAND的传输率，因为NAND颗粒封装时候可能有多Die,多Plane（每个plane都有4KB寄存器），Plane操作时候 可以交叉操作（第一个plane接到指令后，在操作的同时第二个指令已经发送给了第二个plane，以此类推），达到接近双倍甚至4倍的传输能力（看闪存 颗粒支持度）。

工作原理- SSD品质和稳定性

1.设备用料的好坏：
很多人搞不明白Micron, Crucial, Lexar, SpecTek,Numonyx等名字之间的关系，这里稍微解释下：

Micron Technology 镁光科技 – 总公司 - 负责研发生产，主要企业市场和OEM笔记本市场。

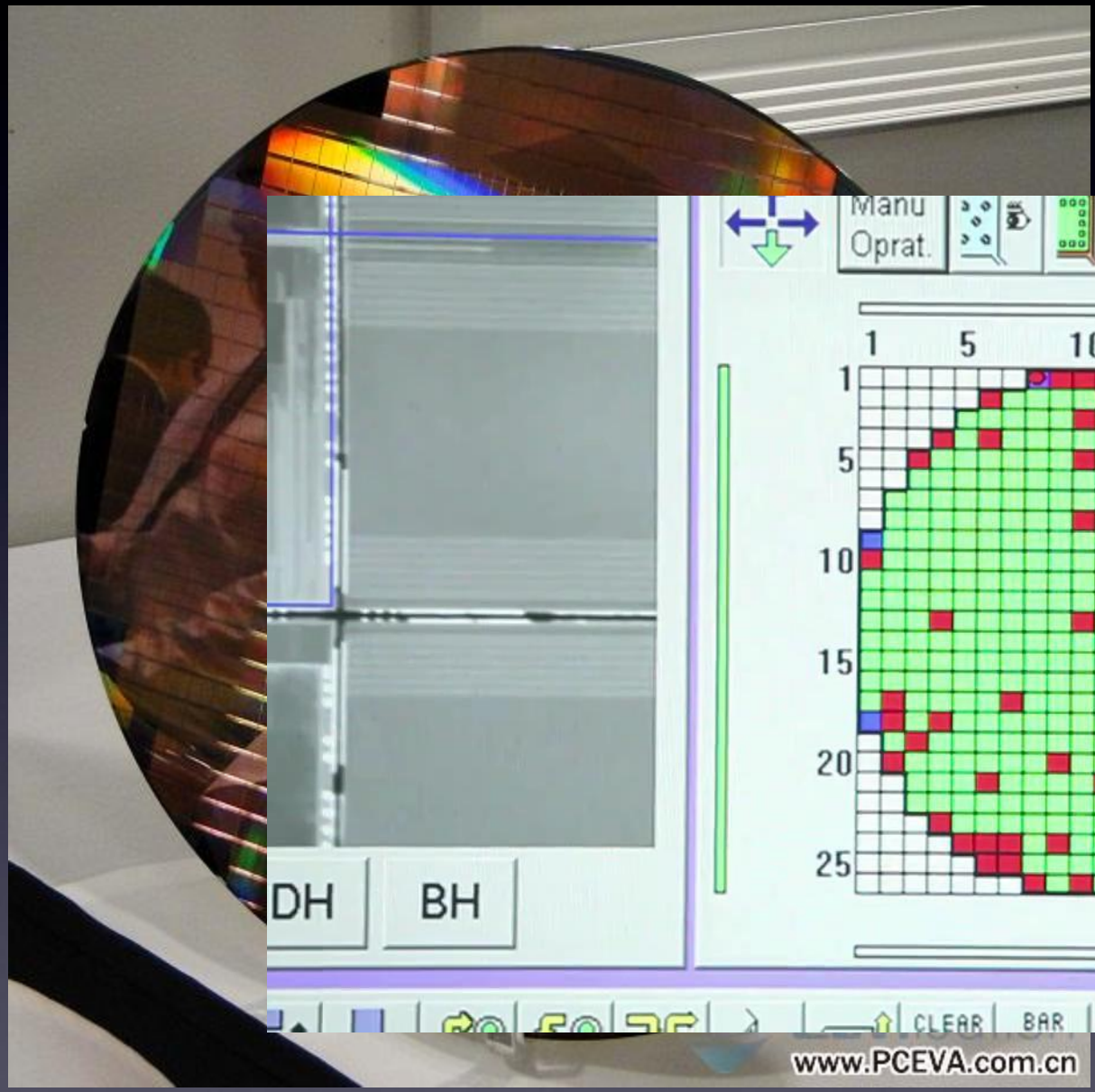
Lexar 雷克沙 – 镁光科技子公司（06年被收购） - 负责零售渠道。

SpecTek - 镁光科技全资子公司 – 88年开始负责元件恢复和测试，96年开始使用自己品牌。

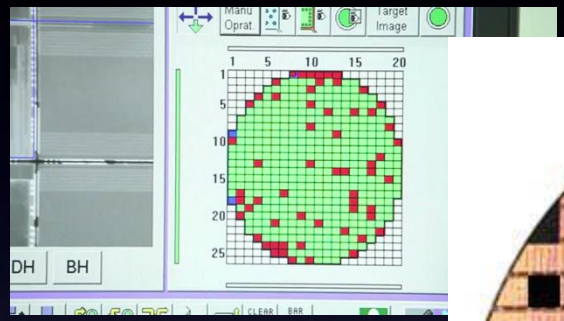
Numonyx 恒忆 - 镁光科技子公司 （10年被收购）– 负责嵌入式设备的市场研发生产。

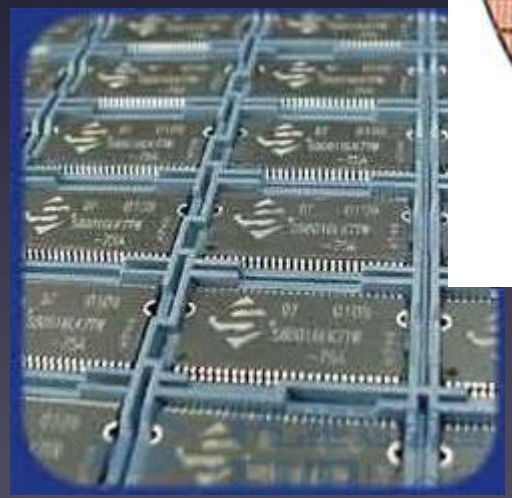
Crucial - 雷克沙子品牌 – 内存和固态硬盘外加小部分闪存盘品牌，主要网购渠道。

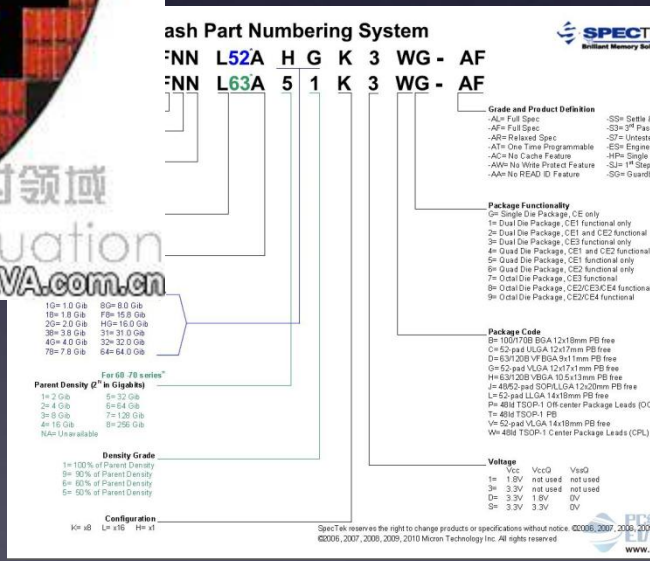
工作原理- SSD品质和稳定性



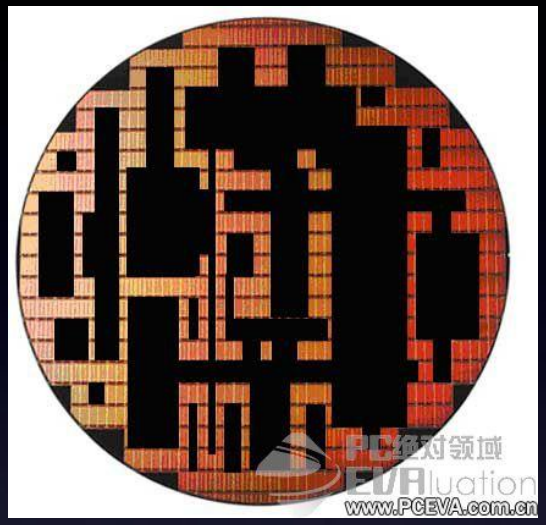
工作原理- SSD品质和稳定性



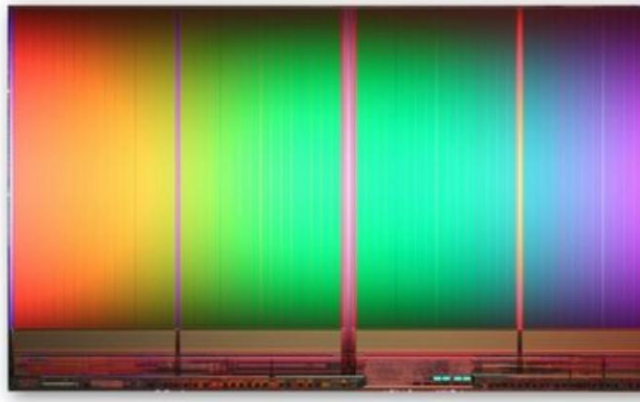




工作原理- SSD品质和稳定性



Downgrade Flash Wafer。



合格品

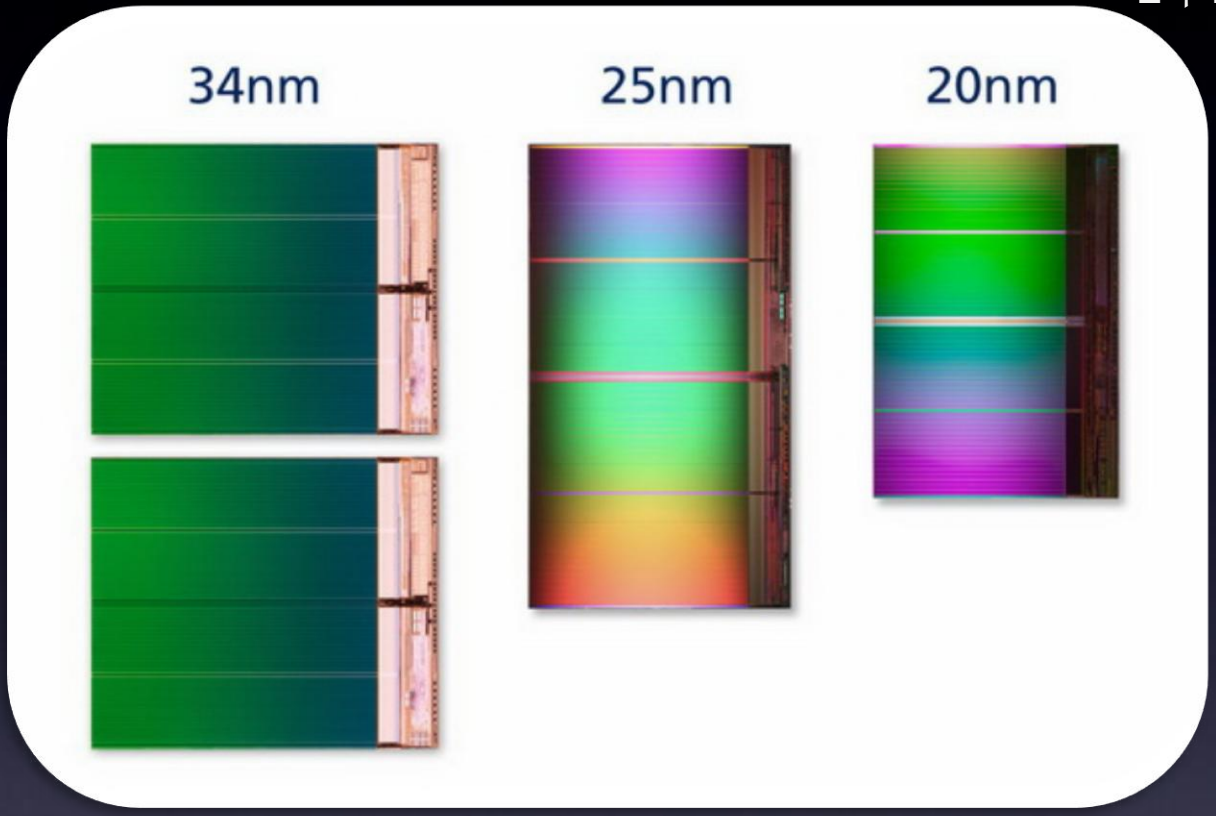
黑片

白片

次品

工作原理- SSD品质和稳定性

1个P/E(Program/Erase cycles)

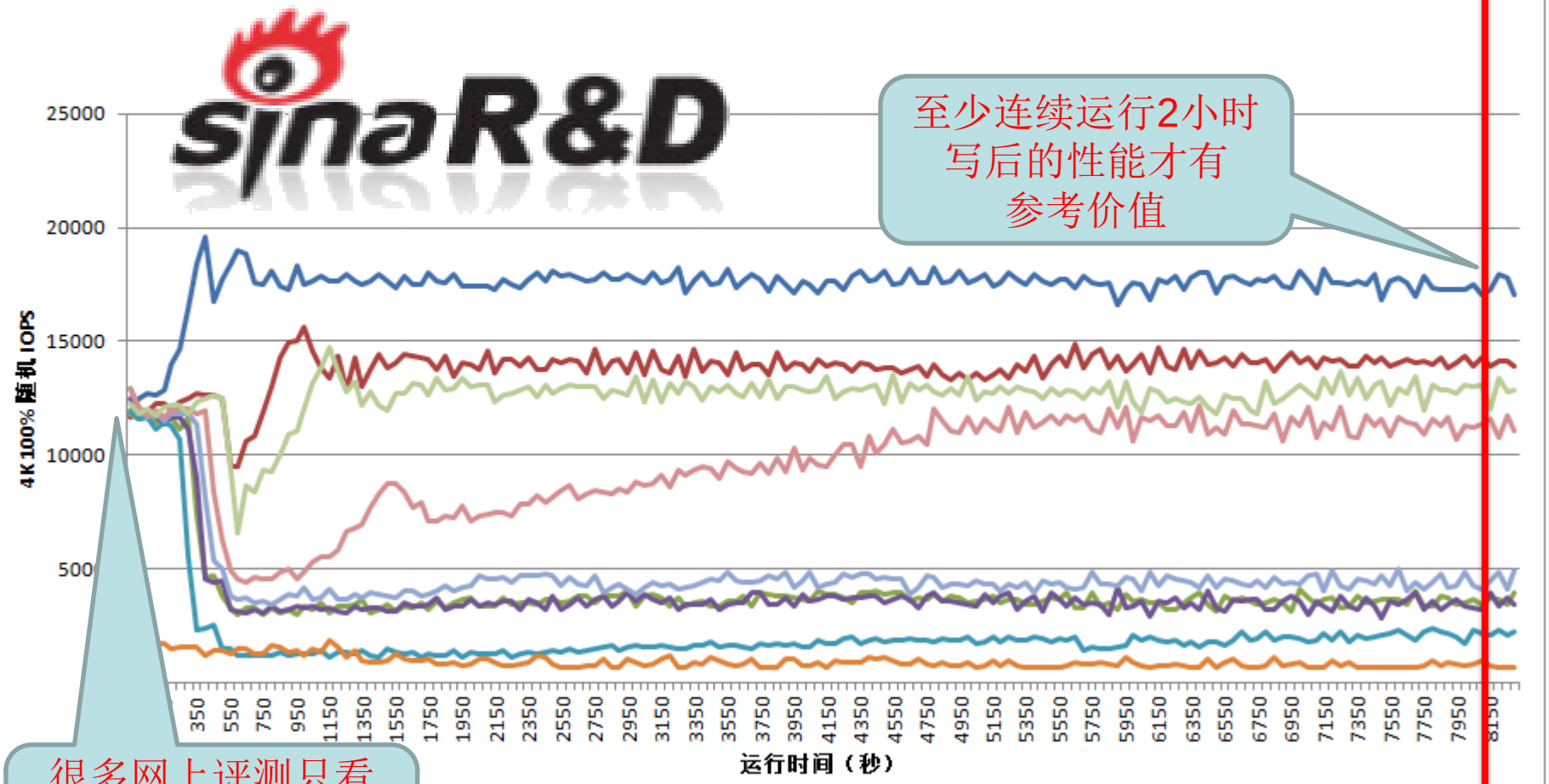


5000 P/E ➡ 3000 P/E ➡ 1000 P/E

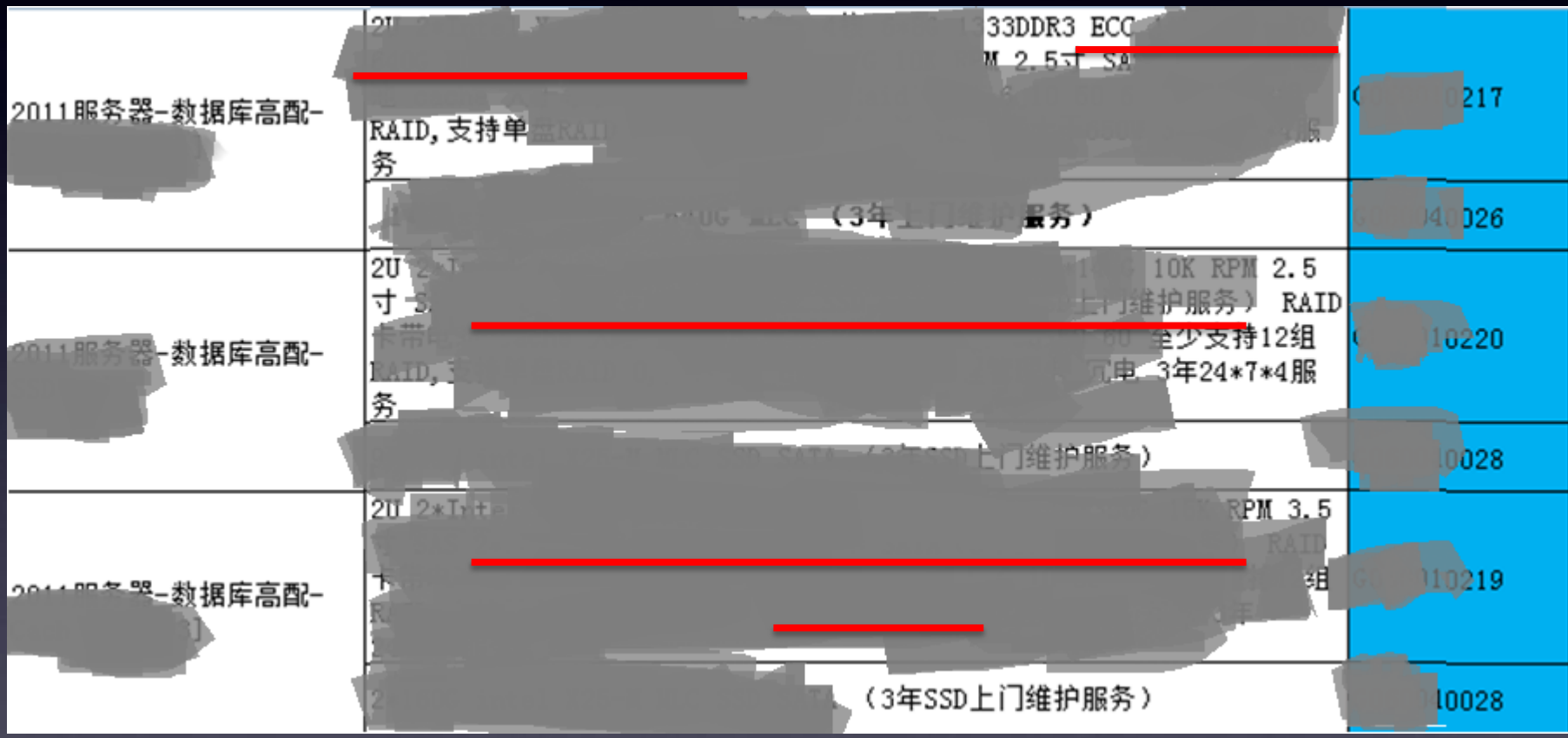
获得更高可靠性的方法是不创新

正确的测试方法

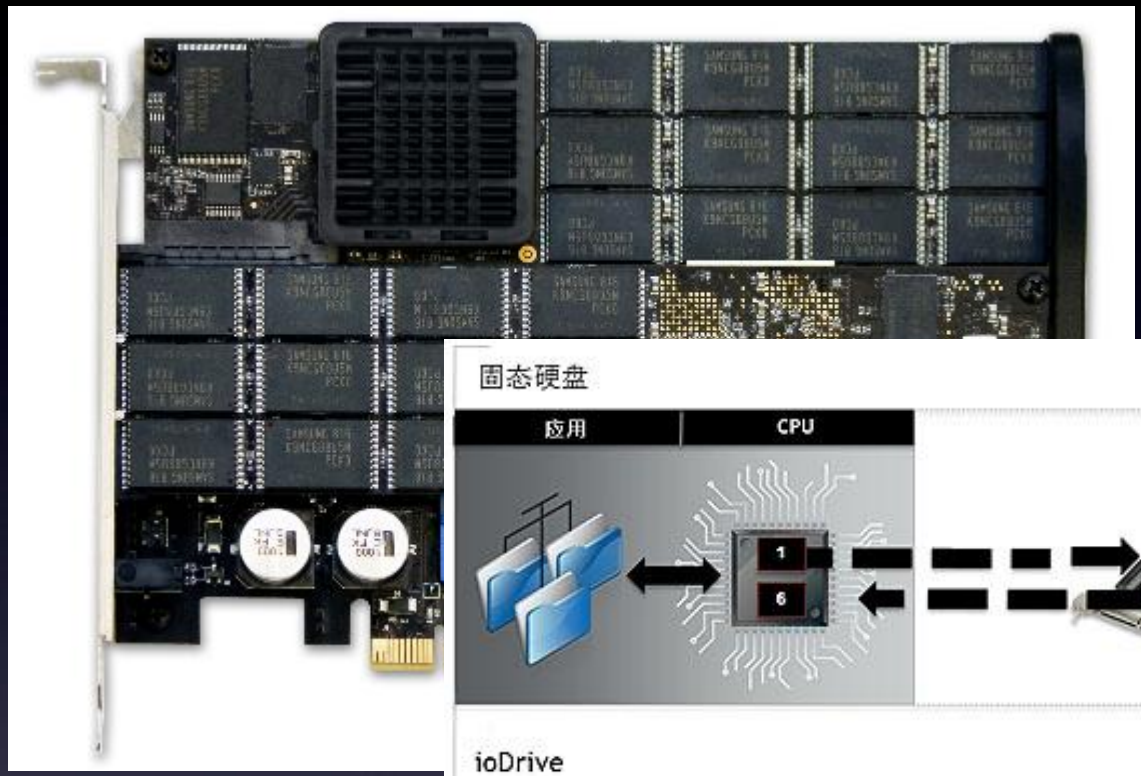
100% 全盘 4k 随机写性能与时间的关系

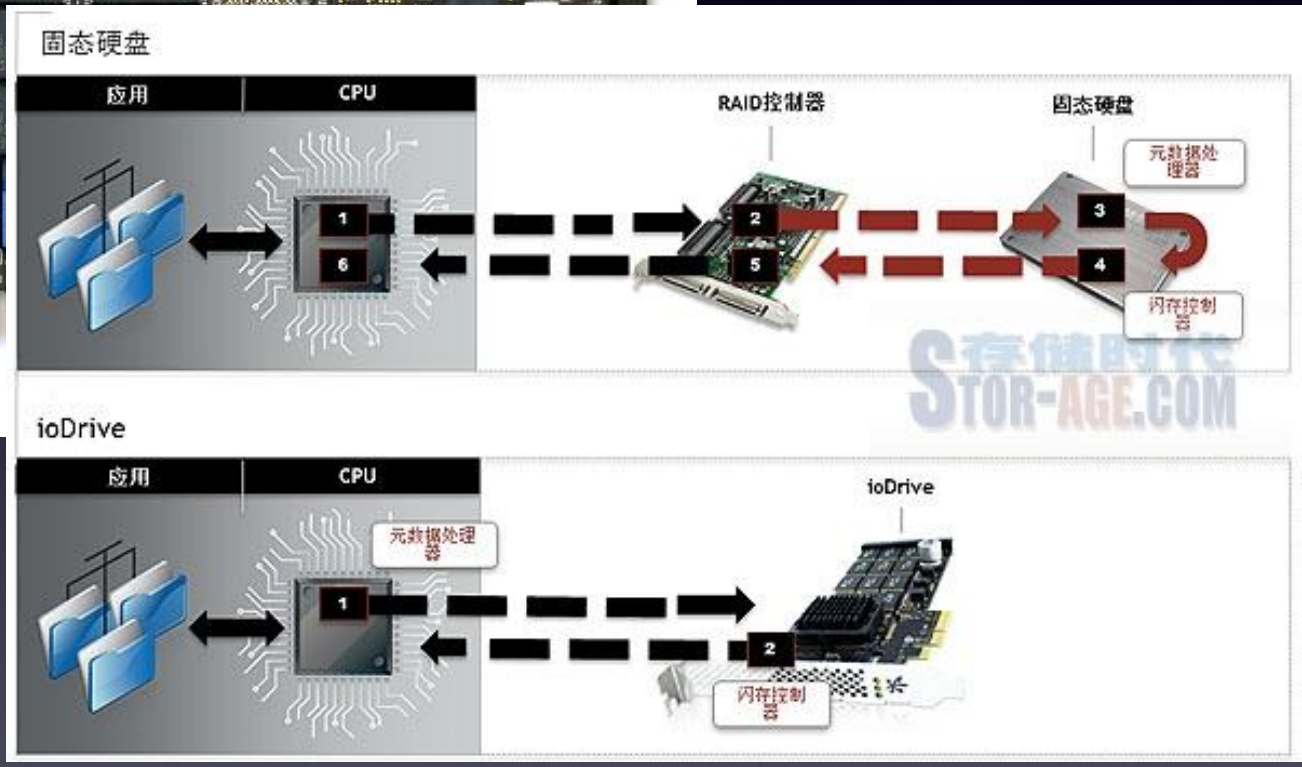


在ERP中选购SSD设备



在ERP中选购SSD设备





在ERP中选购SSD设备

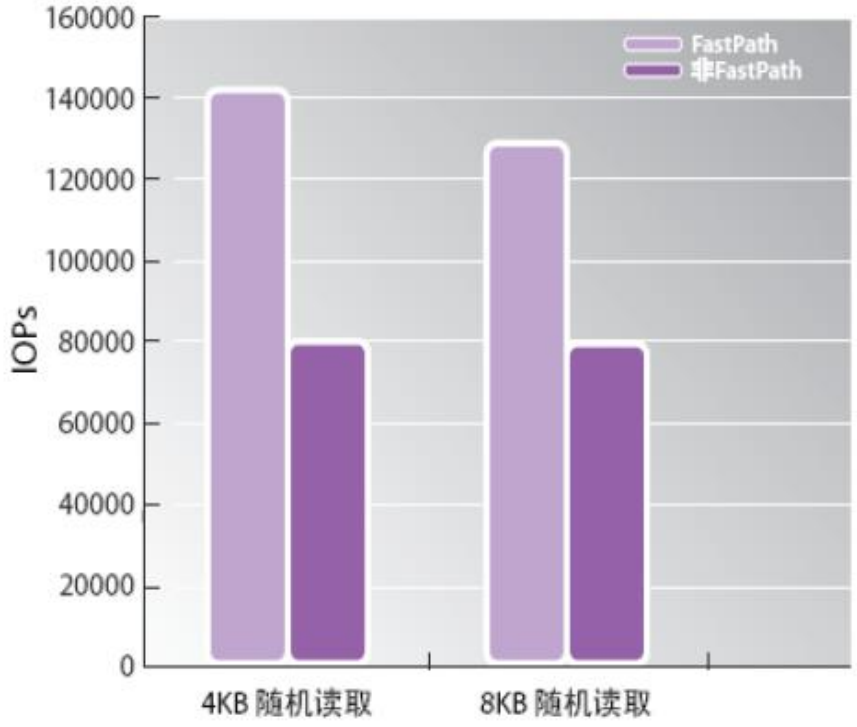


图 2：RAID 5 配置中的读取性能

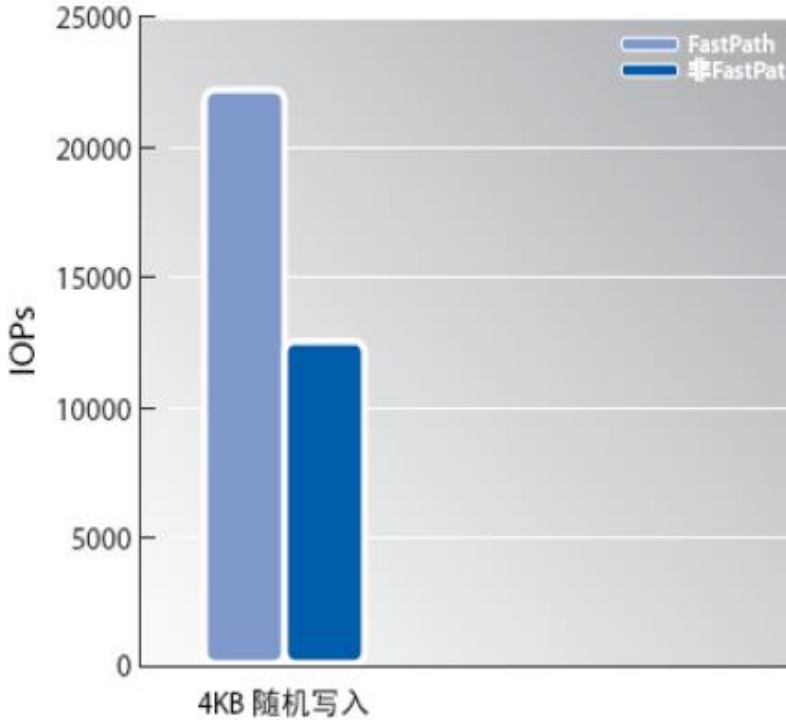


图 3：RAID 5 配置中的写入性能

多块SSD硬盘盘组成RAID 5 后的性能
随机读>14万 iops

问答时间

讲师姓名：刘明生
邮箱　　：mingsheng@staff.sina.com.cn
微博　　：@明生78
手机　　：13810097928