GOPS 2016 Shenzhen

GOPS

2016 全球运维大会 深圳站

会议时间：3月25日-3月26日

会议地点：深圳·南山区 圣淘沙酒店(翡翠店)

主办单位：开放运维联盟 OOPSA Open OPS Alliance　高效运维社区 GreatOPS Community

指导单位：数据中心联盟 Data Center Alliance

协办单位：中国新一代IT产业推进联盟

GOPS 2016 Shenzhen

GOPS

2016 全球运维大会 深圳站

Flash & SSD 101

Zhichao Lv ,
Shannon Systems

What is Flash

- Obviously, The Flash we are talking about here is not an Player.
- Here, Flash is short for Flash Memory
- Flash memory is an electronic non-volatile computer storage media that can be electrically erased and reprogrammed.
- Introduced by Toshiba in 1984

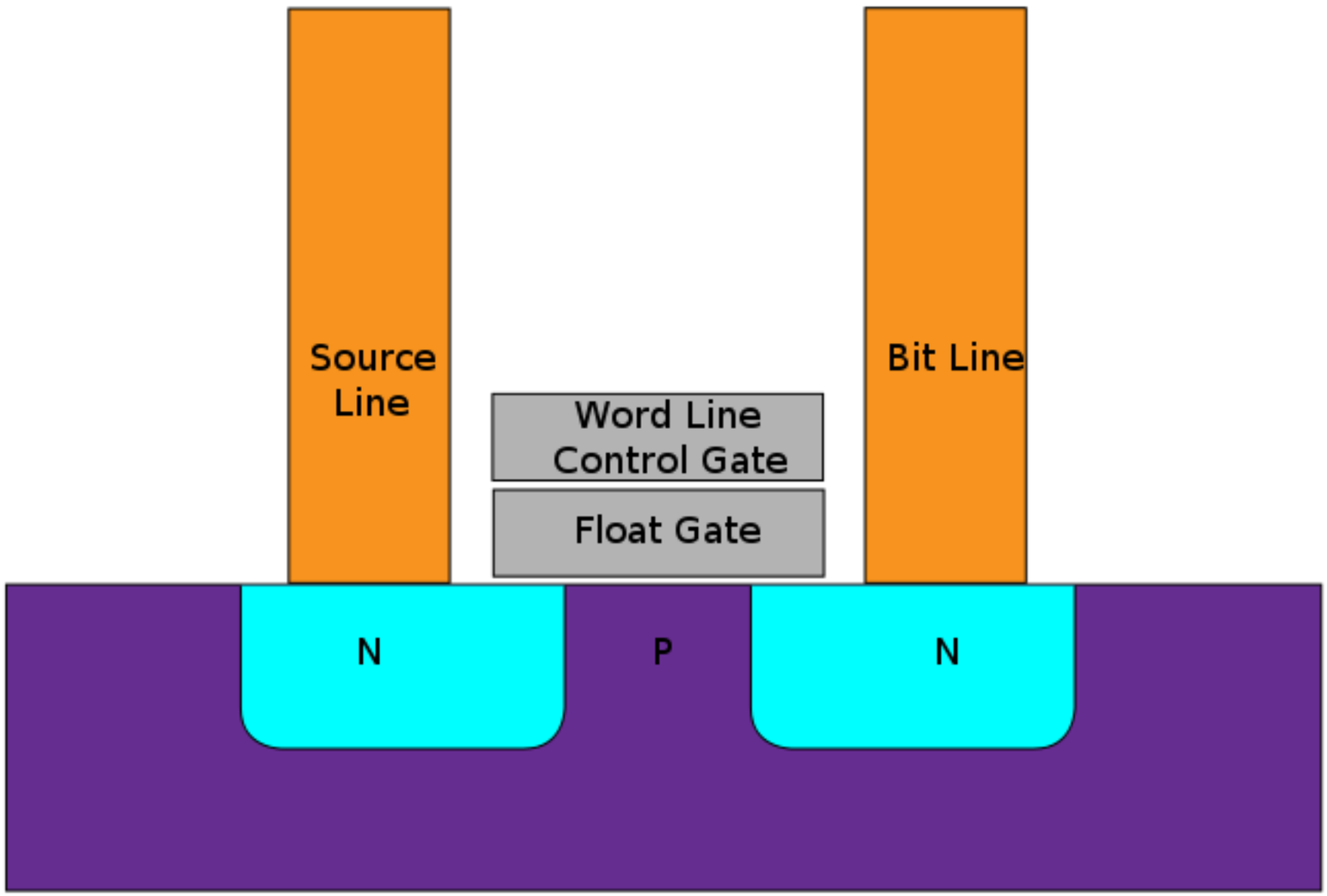
What is Flash

- Flash Memory uses Floating-gate(FG) transistor
- Categorized by the types of FG:
- NOR(或非门) Flash
- NAND(与非门) Flash
- Today, When we talk about Flash, mostly it refers to NAND Flash.
- NOR Flash 可以字节读取, 但是擦除很慢

Flash Cell

- Flash 媒介中存储信息的最小单元(NAND Flash)
- 但是不是人类对Flash 媒介可进行原子操作的最小单元.(NAND Flash)

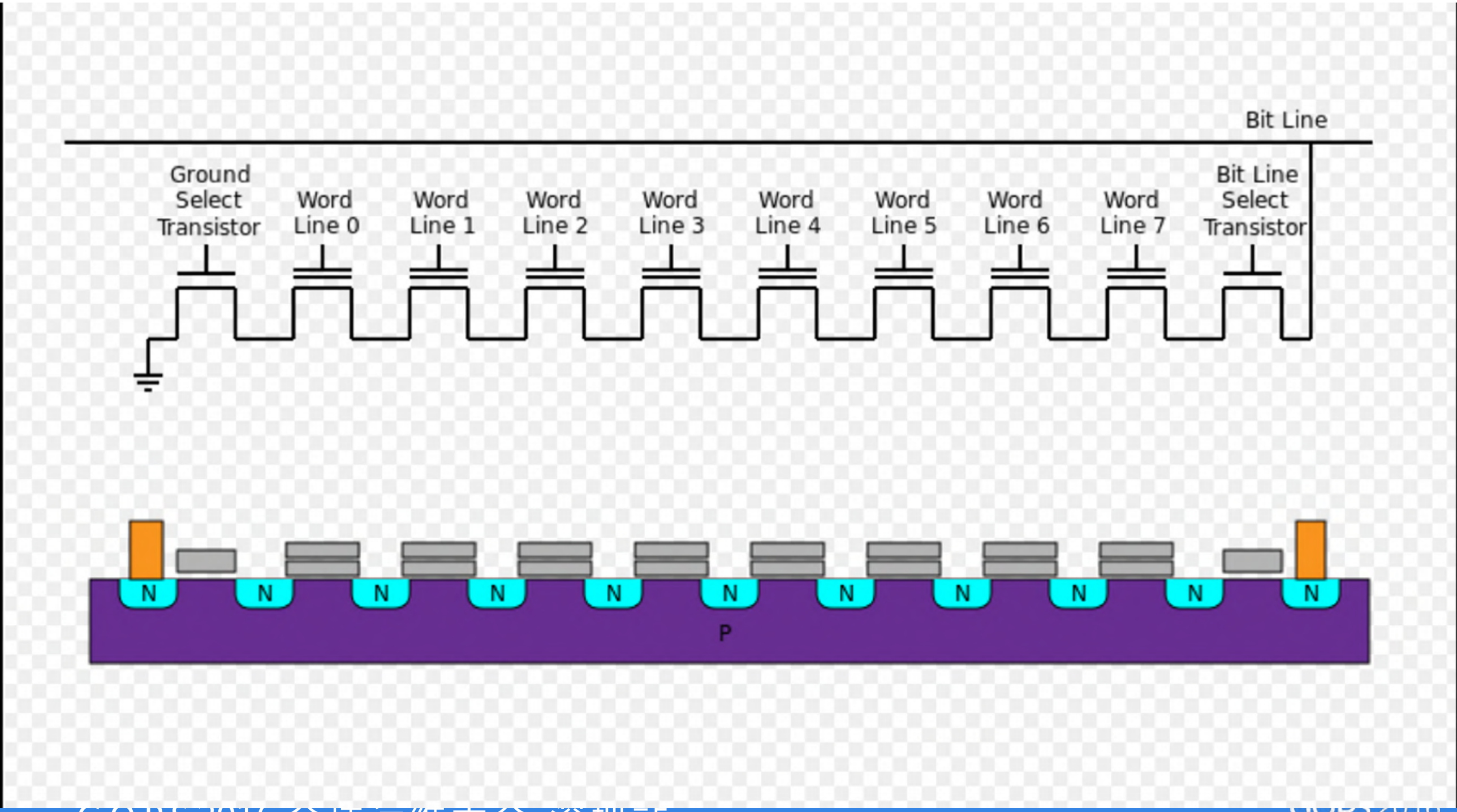
Flash Cell



Flash Cell

- SLC: Single-Level Cell 一个逻辑门存储1bit 数据(0,1)
- MLC:Multi-Level Cell 一个逻辑门存储2bit数据(00,01,10,11)
- TLC:Triple-Level Cell, And you got the idea.

Flash Cell


Bit Line
Ground Select Transistor
Word Line 0
Word Line 1
Word Line 2
Word Line 3
Word Line 4
Word Line 5
Word Line 6
Word Line 7
Bit Line Select Transistor
N
P

Flash Media 组织结构

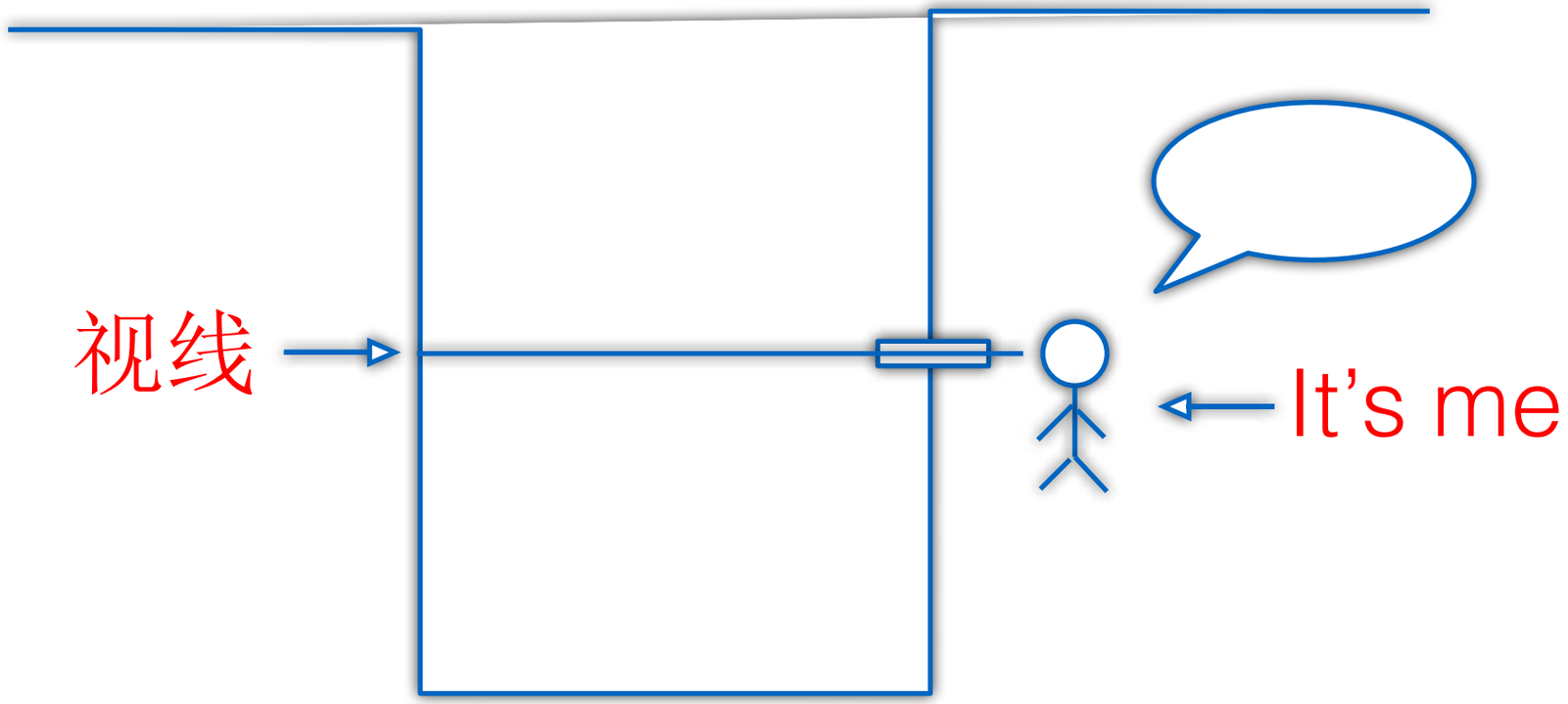
- 从小到大
- Cell
- Page/页 (8KB到16KB数据组成一页,不是Cell数量,目前用的最多的是16KB的)
- Block/块 (256～512页组成一擦除块,4~8MB)
- Die/Lun/芯片 (1024~4096个块组成一个芯片,4GB~16GB)
- Package/封装 (4～8芯片组成一个封装, 16GB~128GB)

Flash Media 组织结构

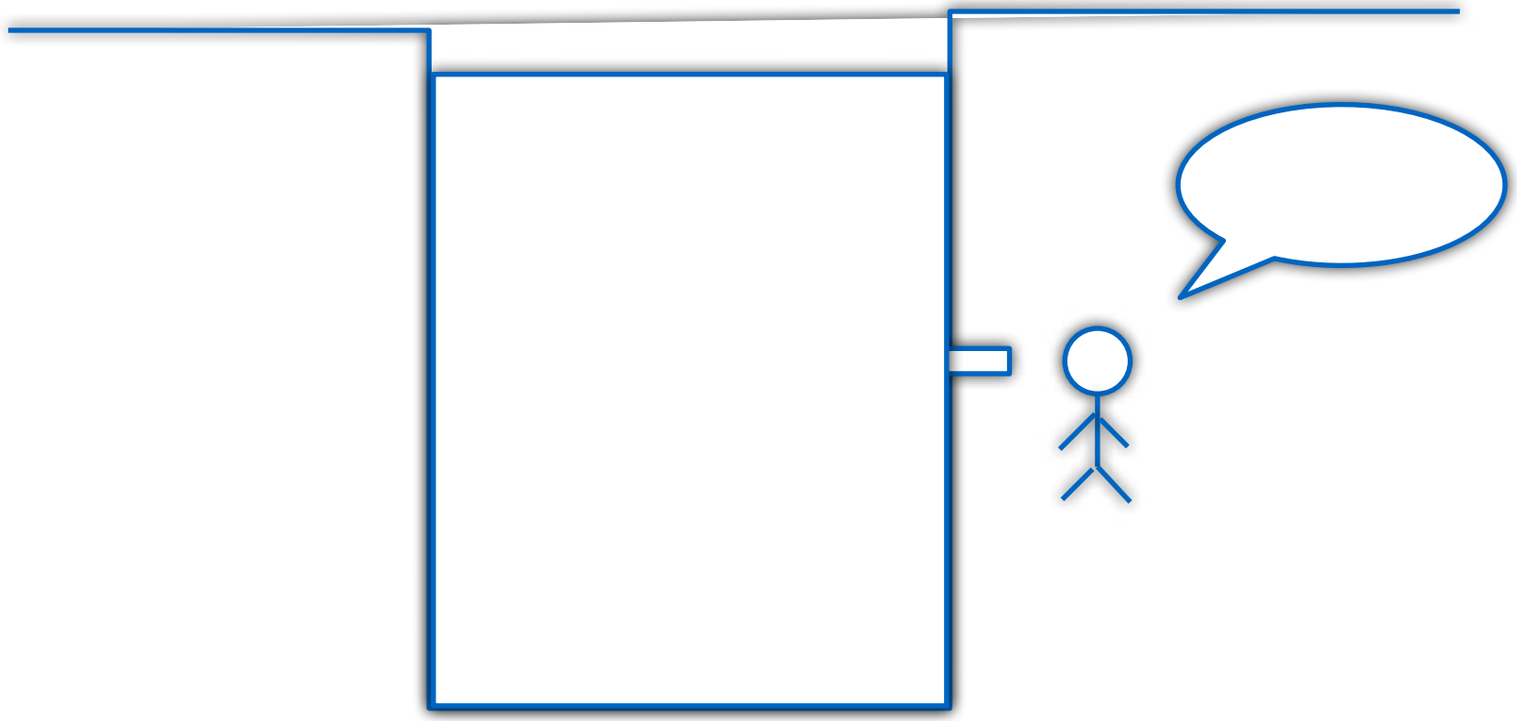
Flash Media 原子操作

- Read/读, 检测晶体管导通状态,导通为0,非导通为1. 原子操作单元Page/页(16KB) 40~80us
- Write/Program/写,对控制门加电压,电子通过隧道效应进入浮栅极(One Way). 原子操作单元:Page/页(16KB) 0.5~2ms
- Erase/擦除, 一个已经被Program的Page, 在被重新Program之前需要被擦除(Dirty->Clear),对浮栅极进行放电. 原子操作单元:Block/块(4MB) 1~4ms

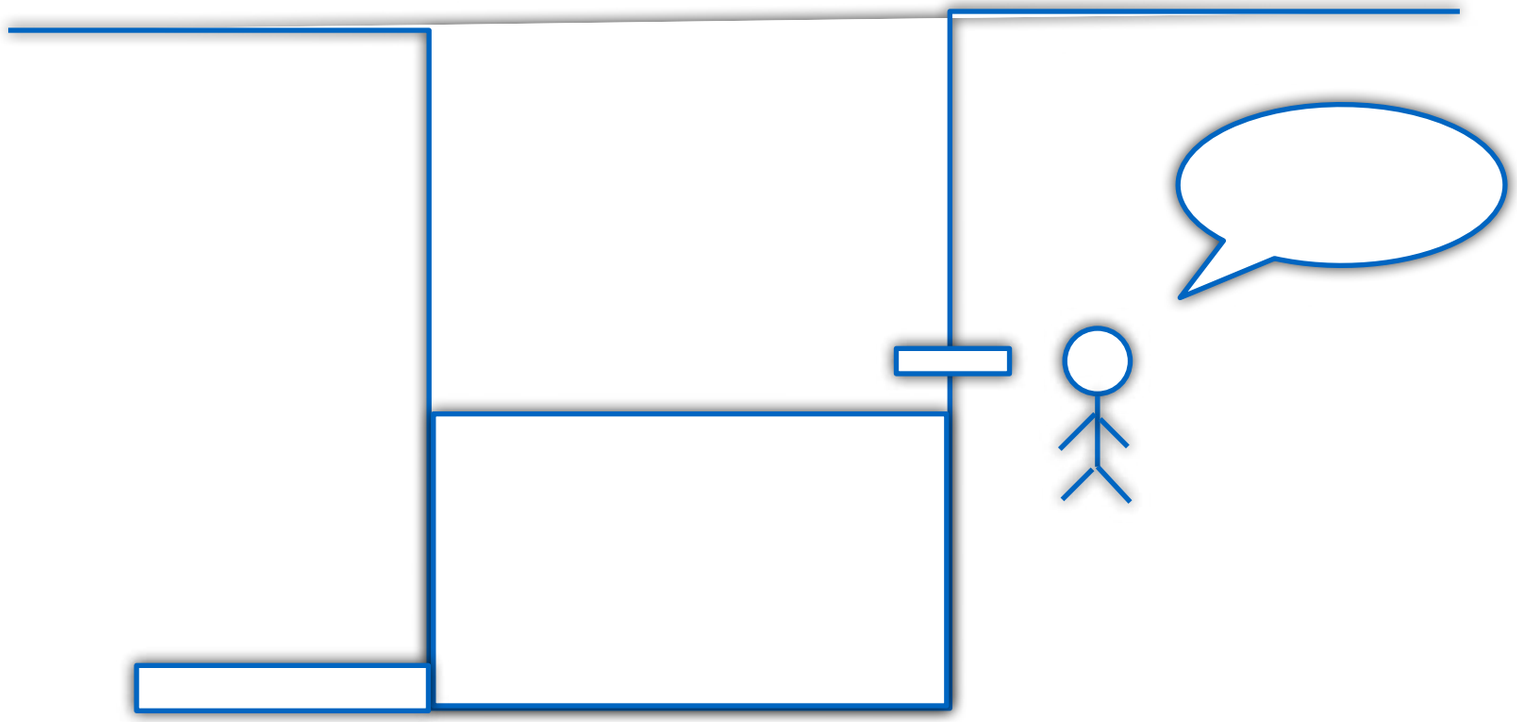
Flash Media 原子操作



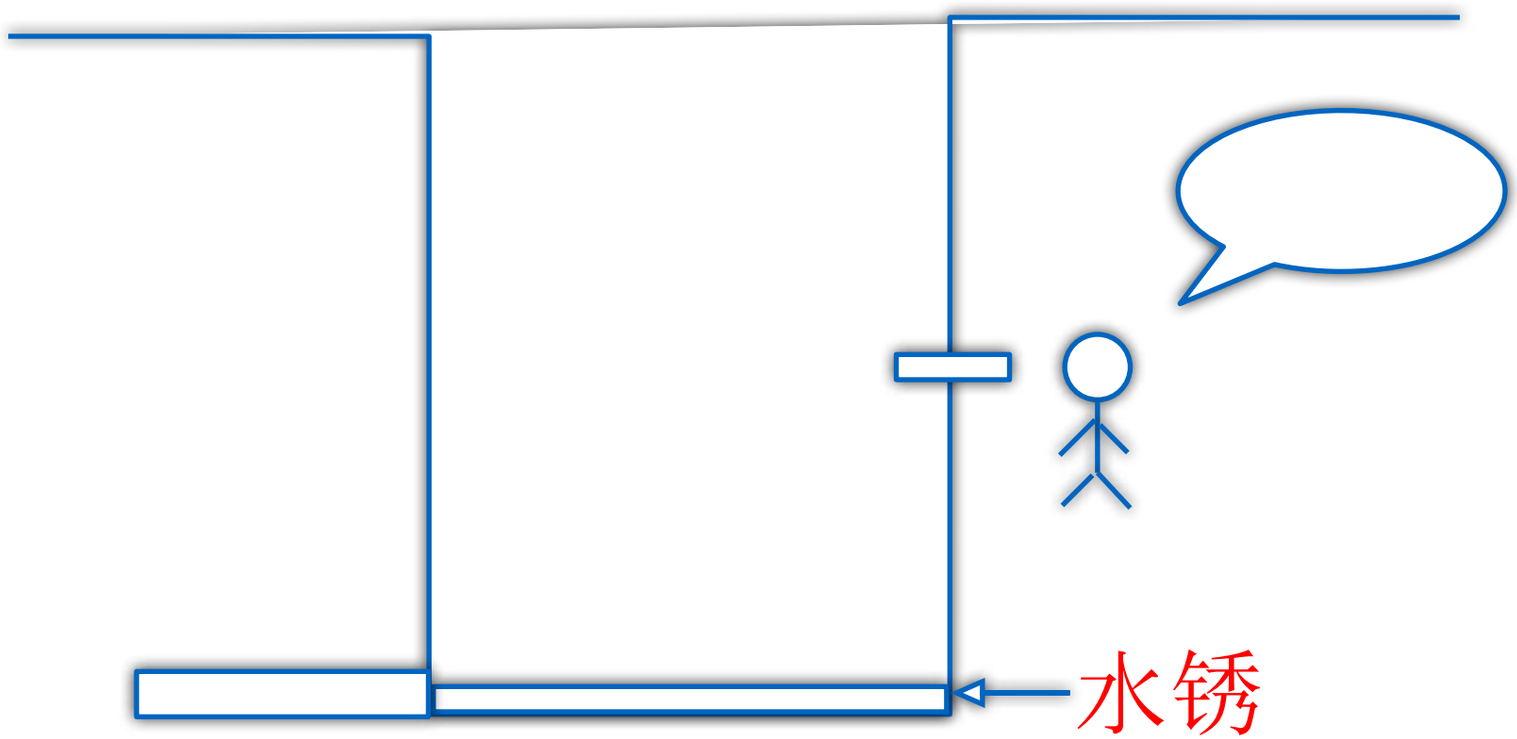
Flash Media 原子操作



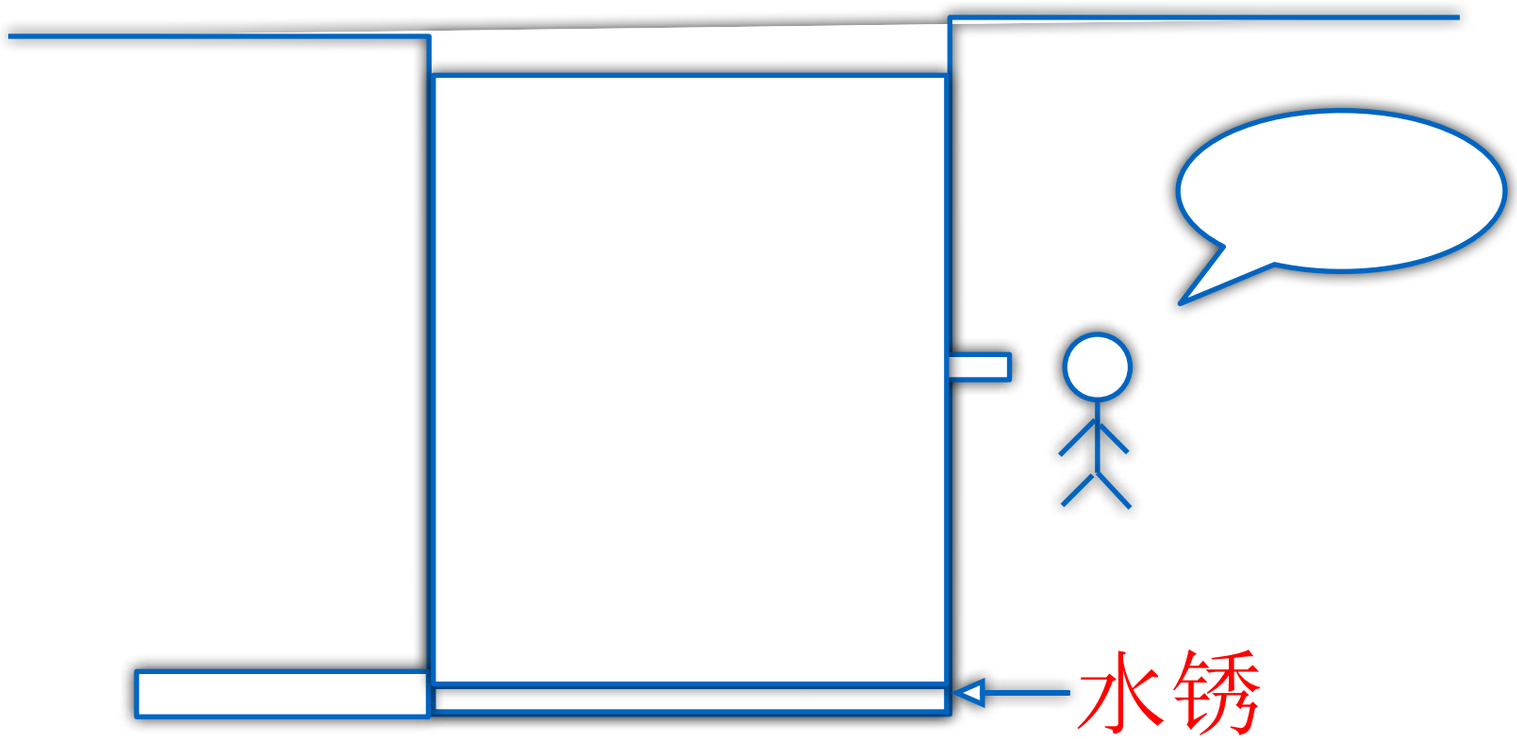
Flash Media 原子操作



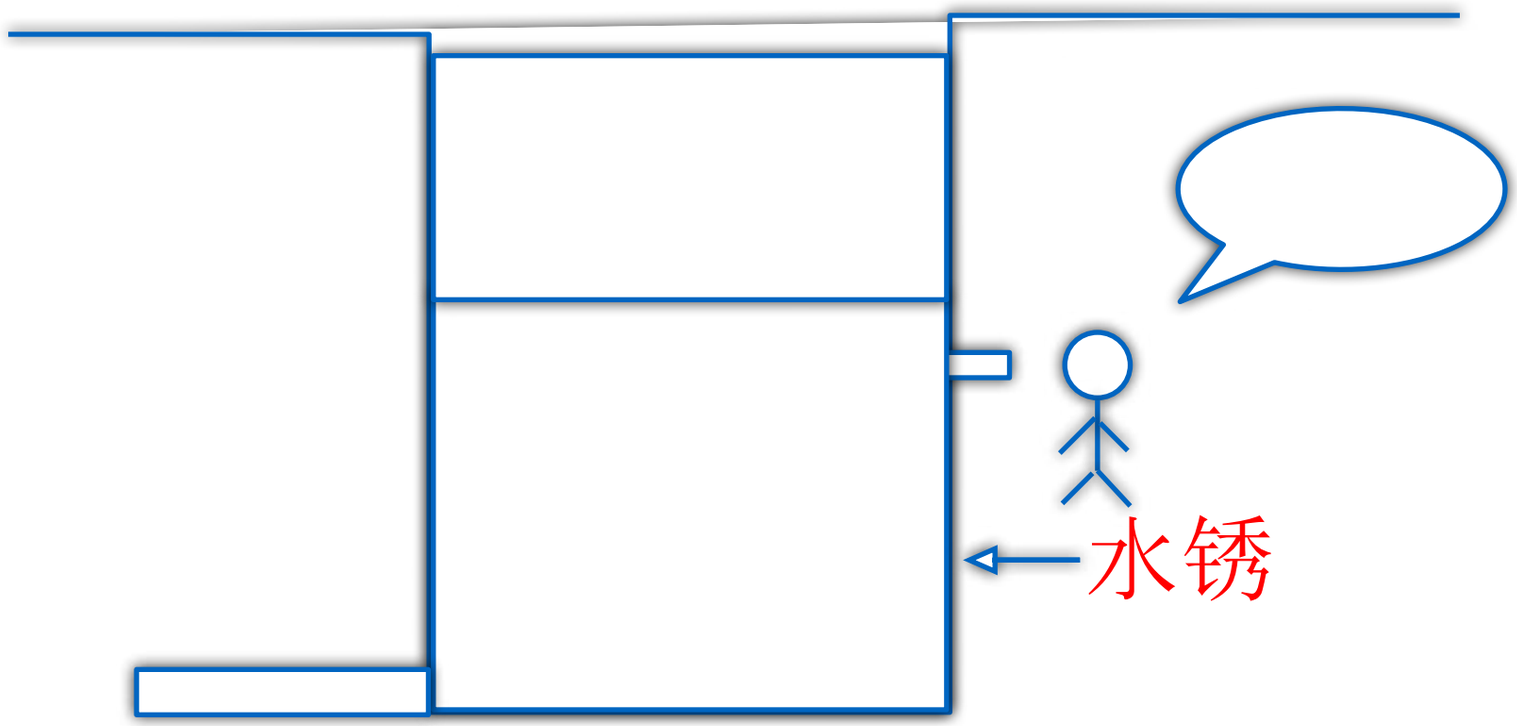
Cell 失效



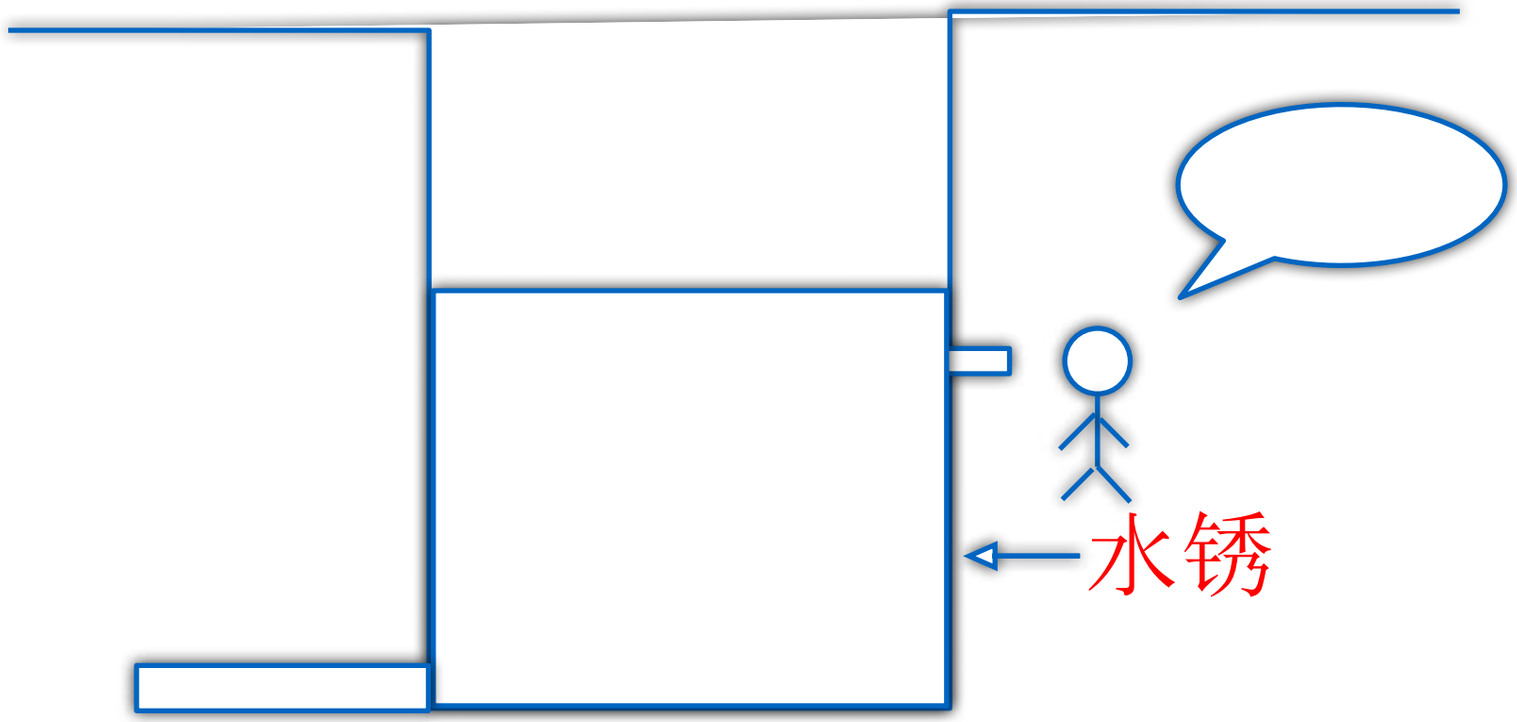
Cell 失效



Cell 失效



Cell 失效



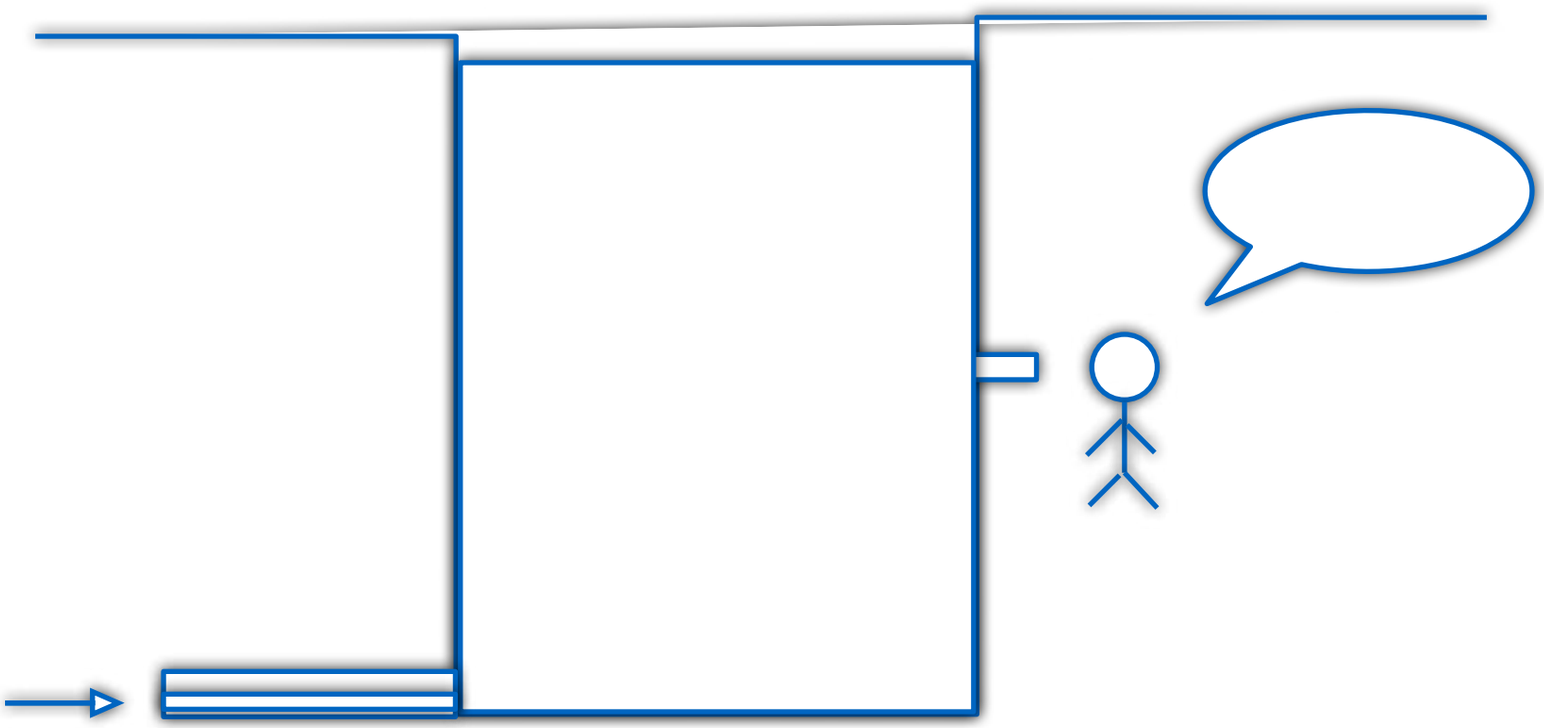
Flash Cell 失效

- 在Erase过程中,由于***效应,浮栅极与硅基层之间的绝缘层的绝缘能力逐渐下降(水锈)
- 很多次的Erase的积累会导致绝缘能力持续下降
- 进而导致Cell 不能有效储存电荷—>Cell失效
- 从一个新的Cell到Cell完全失效,Cell被Erase的次数
- Flash P/E Cycle (Program/Erase)

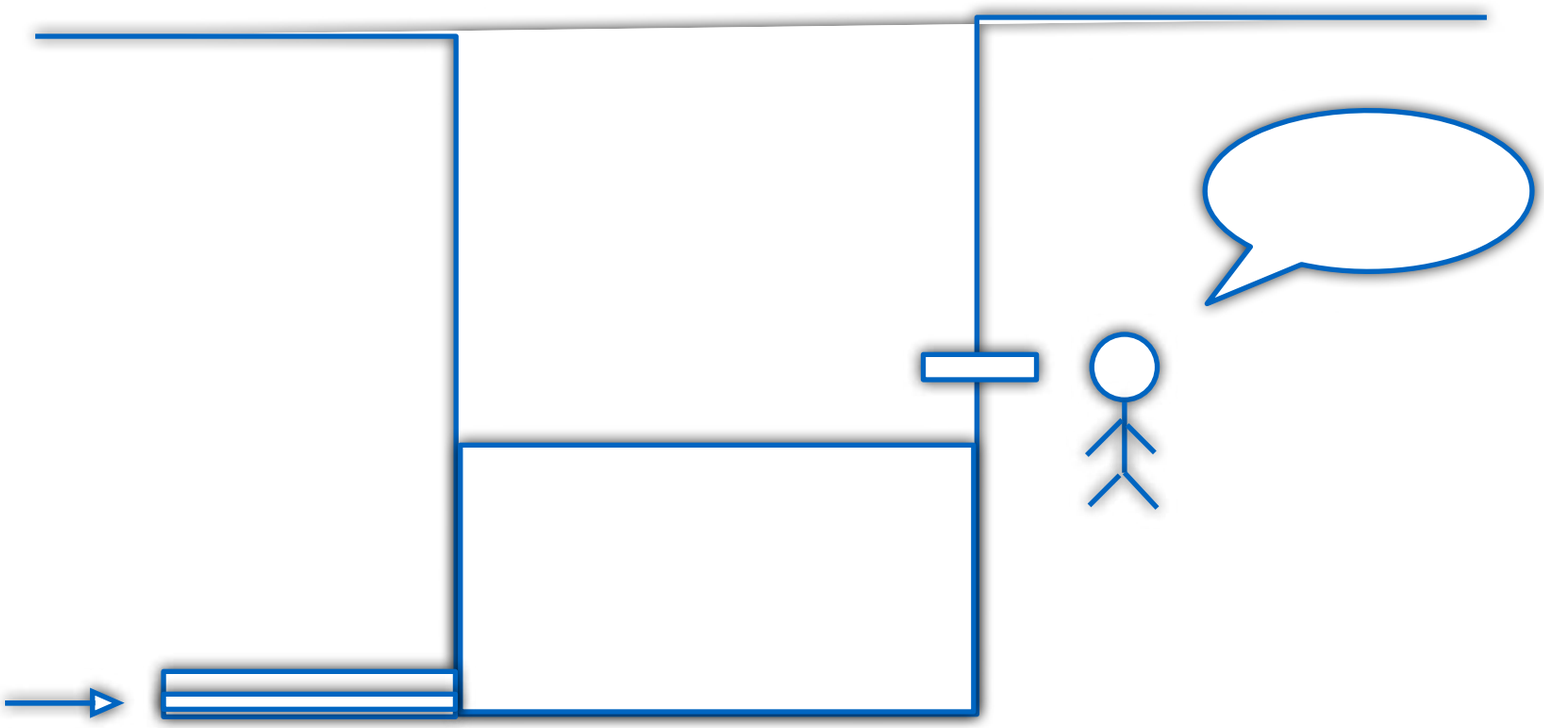
Flash Cell 失效

- MLC
- 典型P/E Cycle:5000次
- P/E Cycle 决定Flash Media 的寿命.
- 注意P/E Cycle,不决定Flash存储产品的寿命(WA,WL)
- 对Flash Media进行读操作不影响Flash Media的寿命

Flash 数据持久性



Flash 数据持久性



Flash 数据持久性

- 由于存在浮栅极与基极之间的漏电效应
- 如果Flash长期不加电,浮栅极中的电荷会逐渐减少.
- 减少到阈值之下,会发生数据错误.
- 新的 P/E Cycle=0 10年
- 寿命殆尽的 P/E Cycle 接近5000的 3个月

Flash 数据持久性

- 由于存在浮栅极与基极之间的漏电效应
- 如果Flash长期不加电,浮栅极中的电荷会逐渐减少.
- 减少到阈值之下,会发生数据错误.
- 新的 P/E Cycle=0 10年
- 寿命殆尽的 P/E Cycle 接近5000的 3个月

Flash 数据持久性

- 没事U盘往电脑上插一下
- 没事U盘往电脑上插一下
- 没事U盘往电脑上插一下
- 重要事情说三遍

Flash 产品简介

- SSD is short for Solid State Disk
- Flash Media 是一种很不稳定的存储媒介
- Don’t Panic!
- 做Flash 产品其实是在做Flash Media管理软件
- Driver/FW(主控)

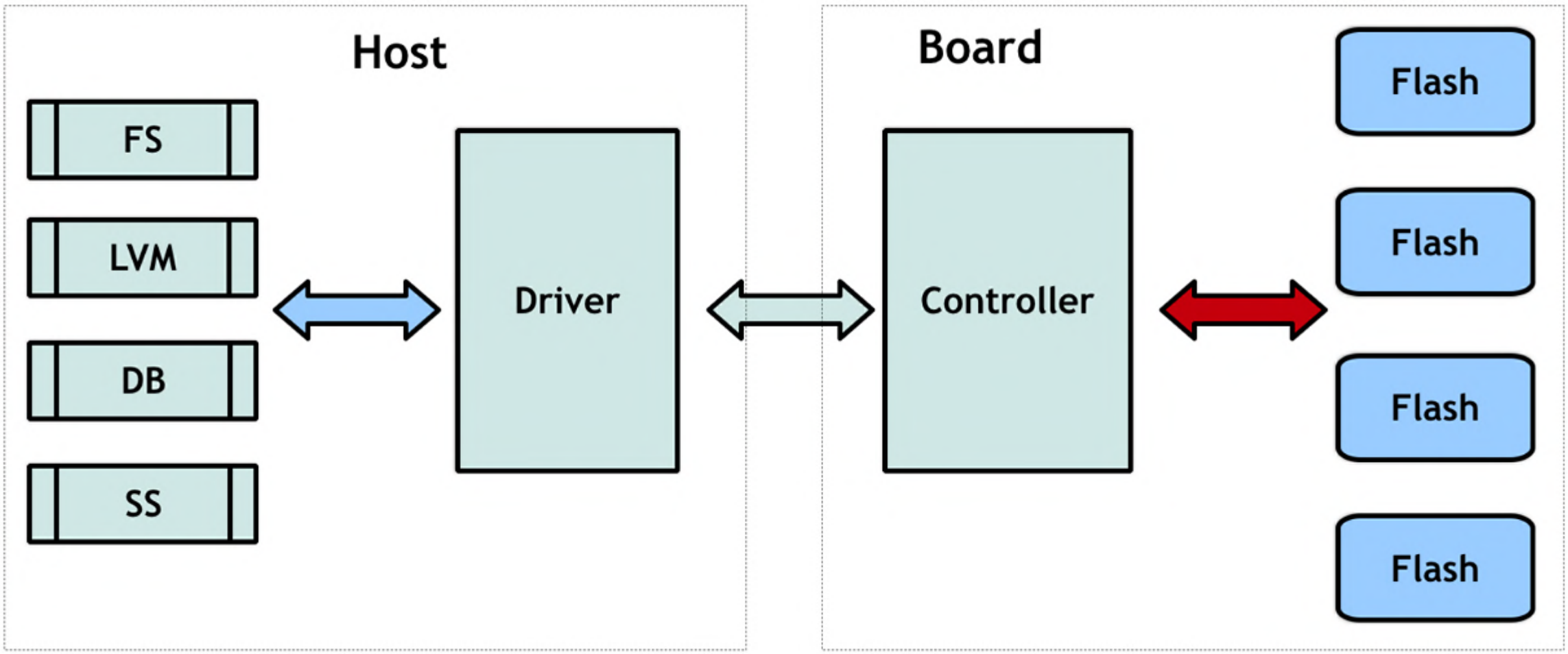
Flash 产品简介

- SSD is short for Solid State Disk
- Flash Media 是一种很不稳定的存储媒介
- Don’t Panic!
- 做Flash 产品其实是在做Flash Media管理软件
- Driver/FW(主控)

Flash 产品简介

- FTL: Flash Translation Layer
- 由Intel 提出
- Mapping Table + Flash Media管理
- 物理地址与逻辑地址的映射表
- 操作系统看到的是逻辑地址(从硬件的角度来看)

Flash 产品简介


Host
FS
LVM
DB
SS
Driver
Board
Controller
Flash
Flash
Flash
Flash

FTL,RAID,OP,GC,WL

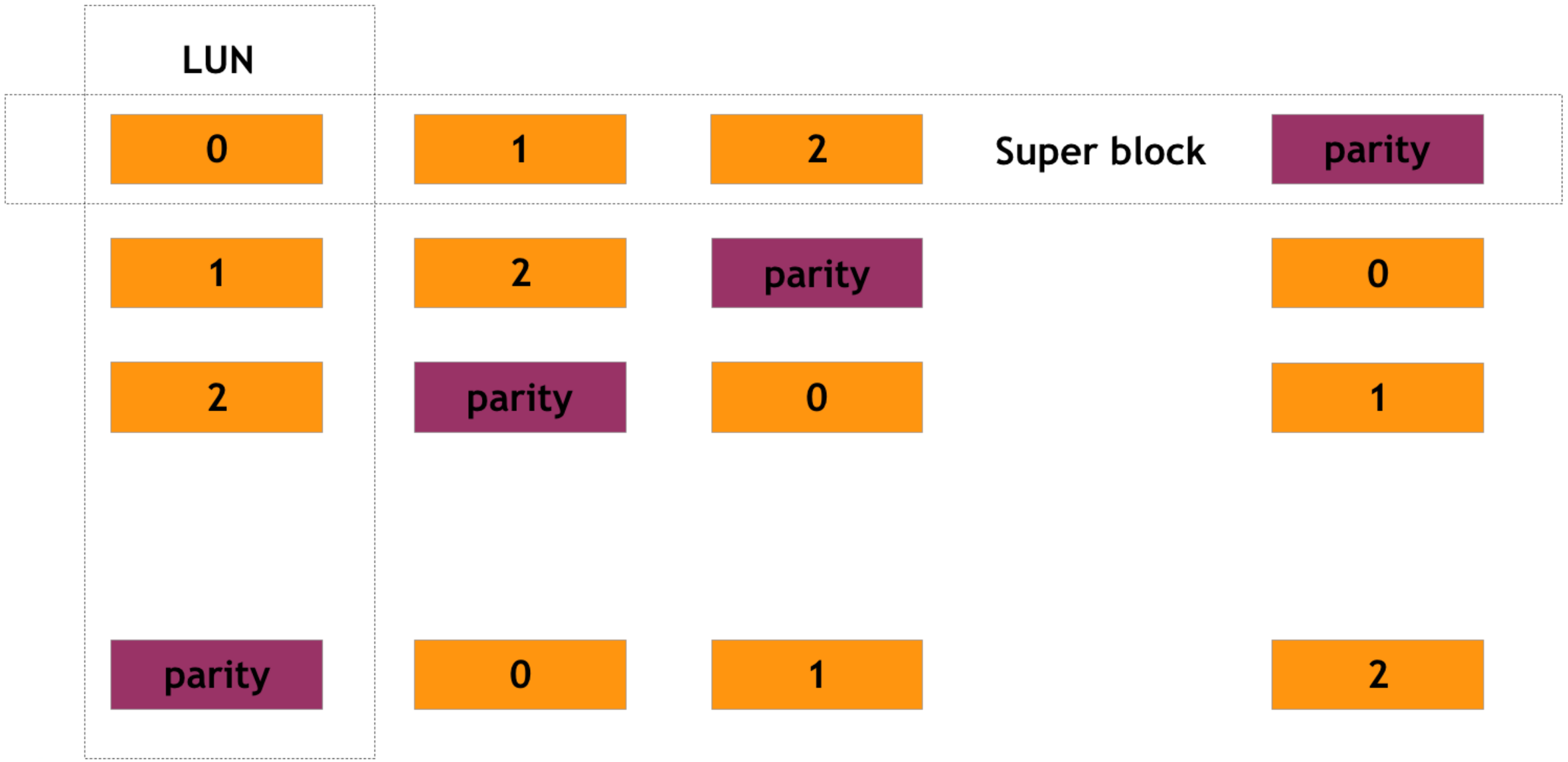
- Mapping Table
- 一个大的数组,存储地址之间的影射关系
- 存储在DRAM中为了保证速度
- 要有持久化机制

FTL,RAID,OP,GC,WL

- 数据保护
- RAID

FTL,RAID,GC,WL

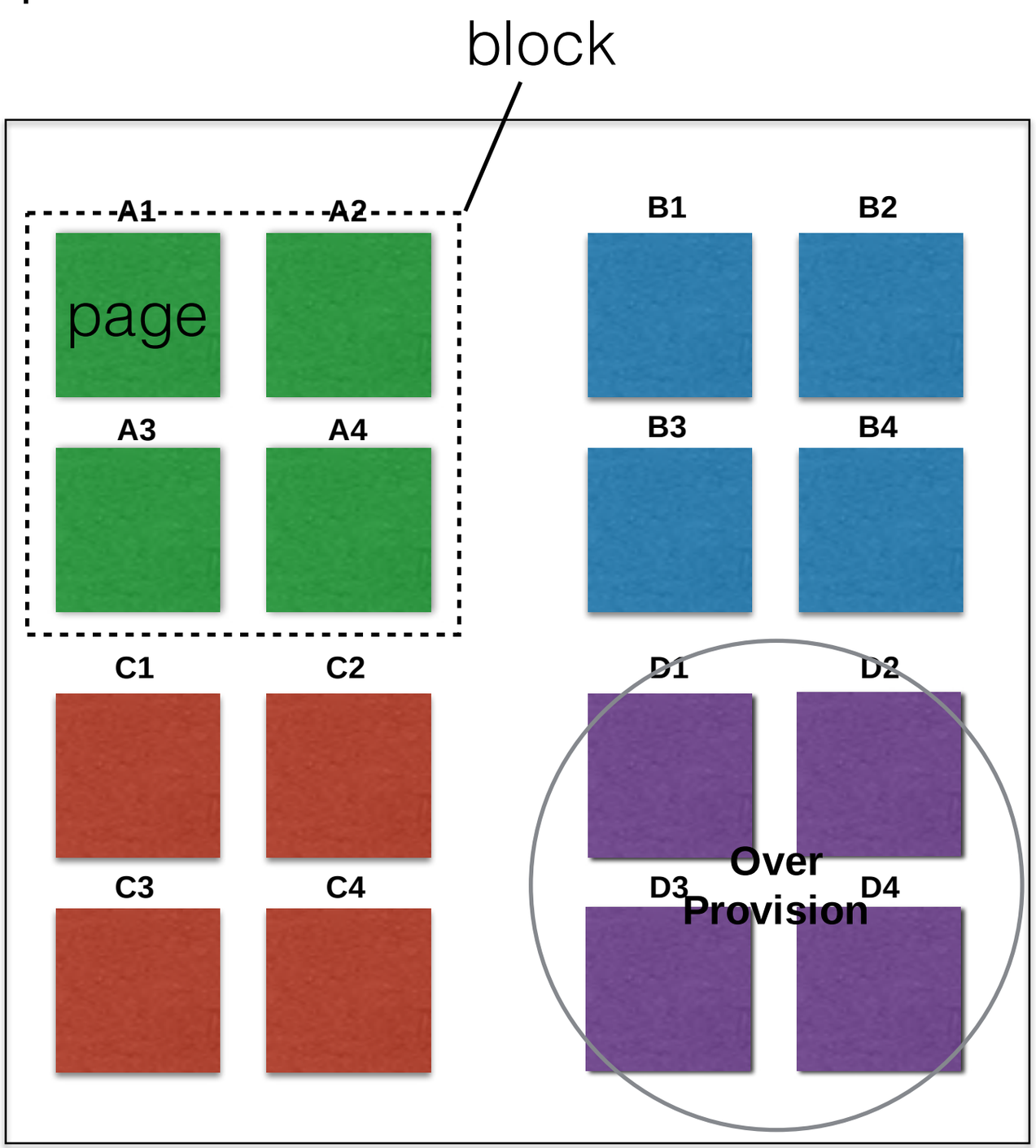
RAID组织



FTL,RAID,OP,GC,WL

- OP过量提供
- OP目的, 保证Flash产品可运行, 更好的性能与更好的寿命
- GC, WL
- 垃圾回收与磨损均衡
- 垃圾回收目的,保证性能
- 磨损均衡目的,保证寿命

Simple SSD



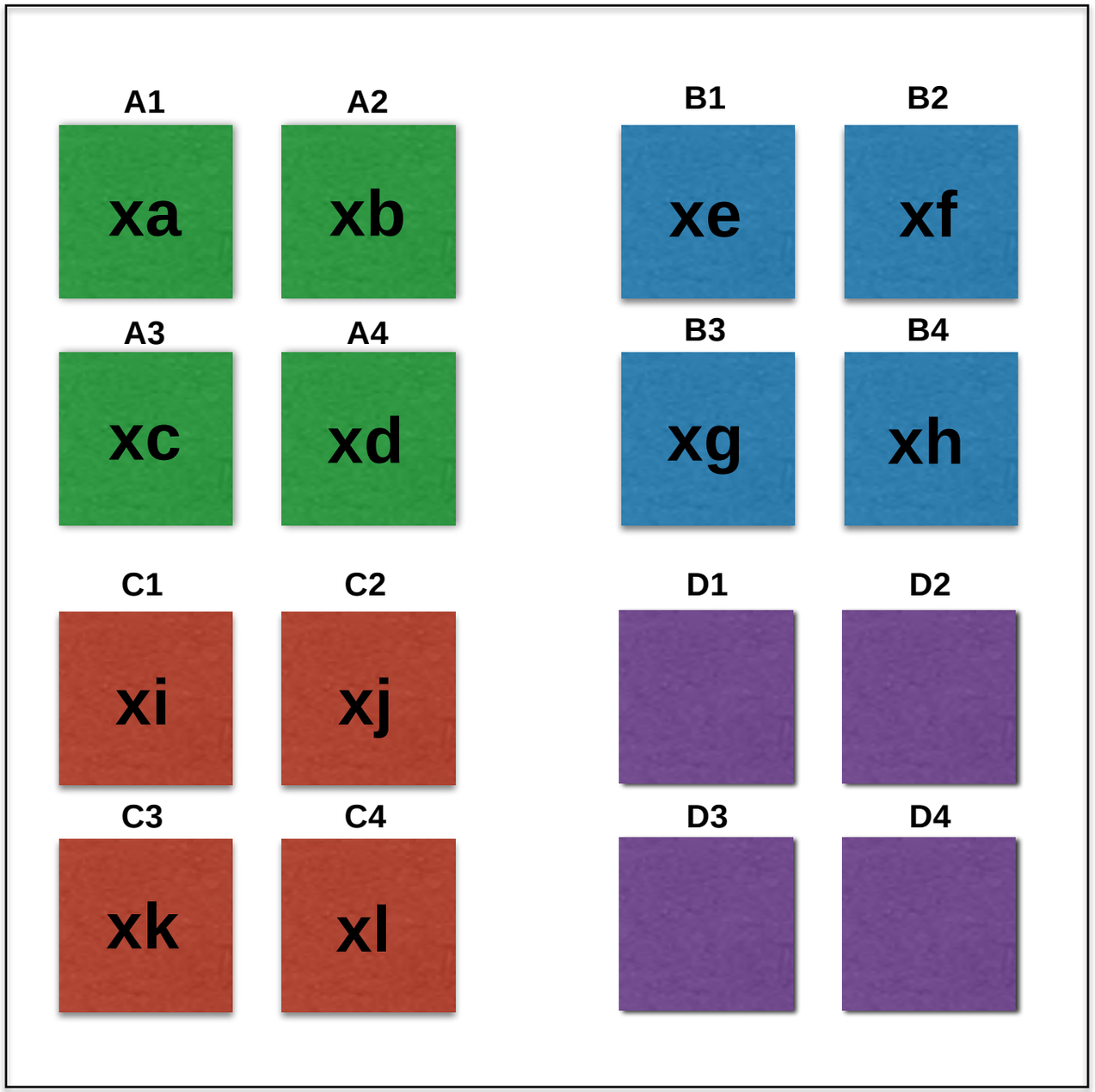
FTL(Flash Translation Layer)

LBA	PBA
1	A1
2	A2
3	A3
4	A4
5	B1
6	B2
7	B3
8	B4
9	C1
10	C2
11	C3
12	C4

Physical Capacity = 16page
User Capacity = 12page

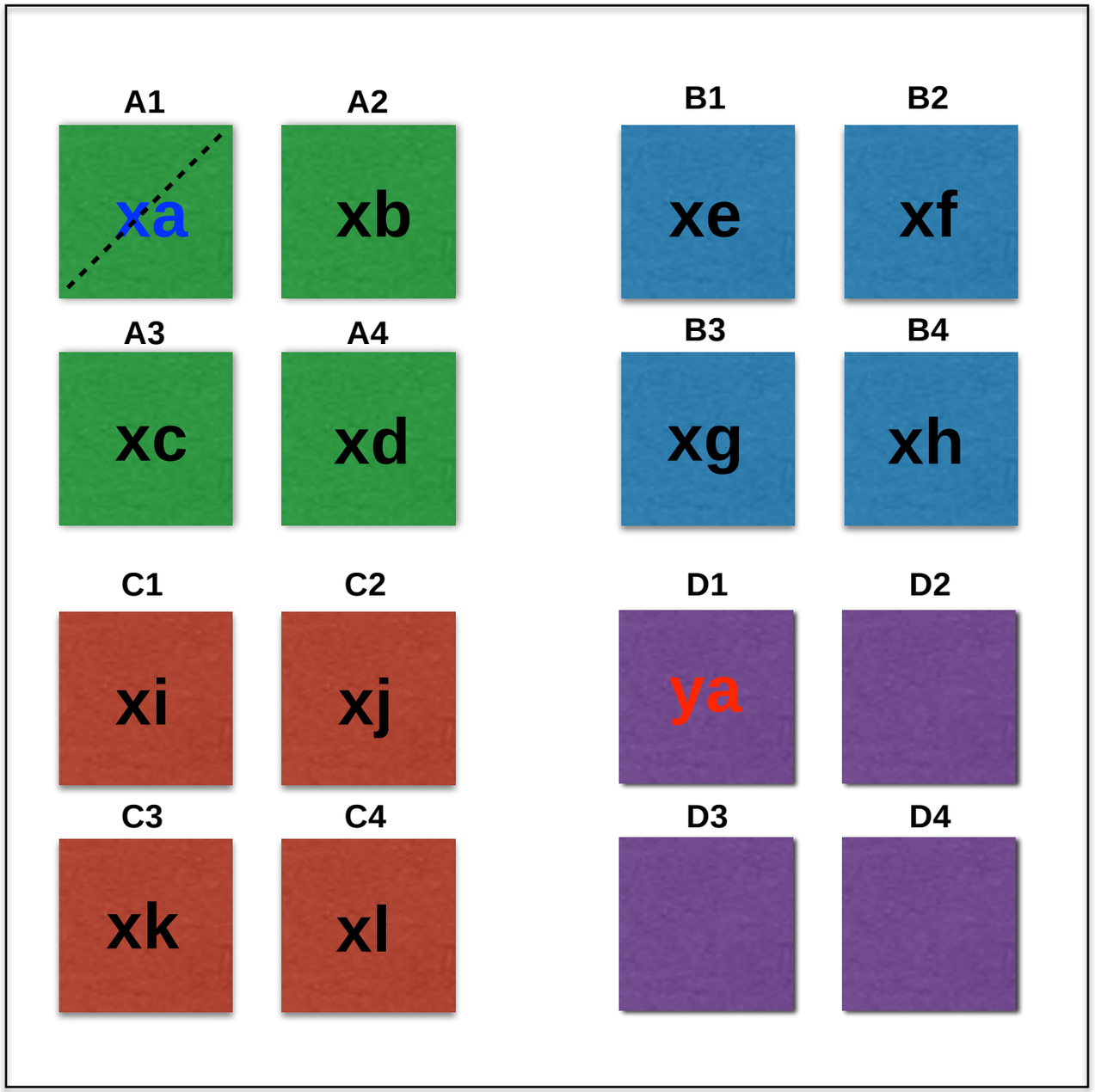
OP = 25%

Whole Disk Write



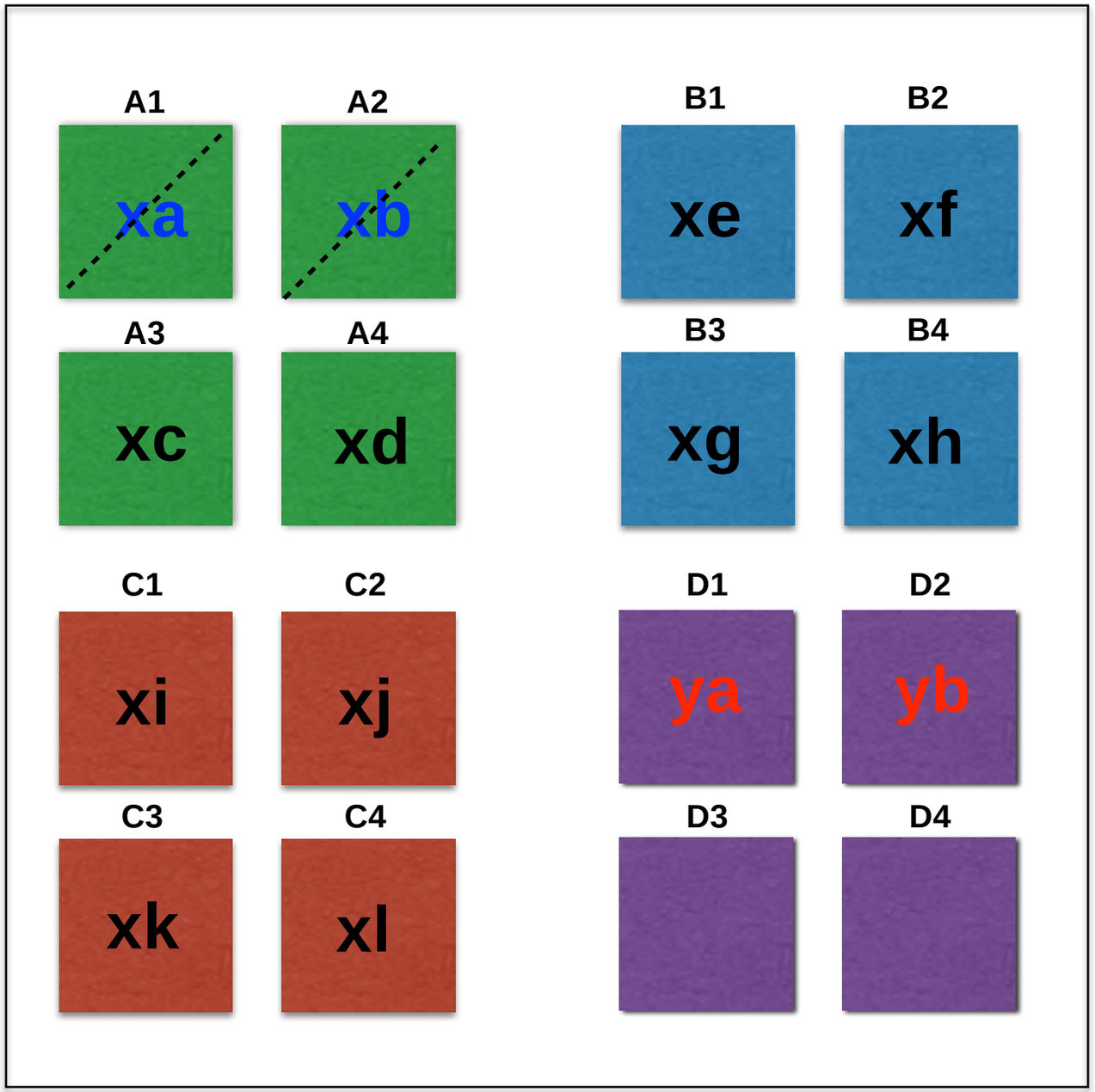
LBA	PBA	Data
1	A1	xa
2	A2	xb
3	A3	xc
4	A4	xd
5	B1	xe
6	B2	xf
7	B3	xg
8	B4	xh
9	C1	xi
10	C2	xj
11	C3	xk
12	C4	xl

LBA 1 xa -> ya



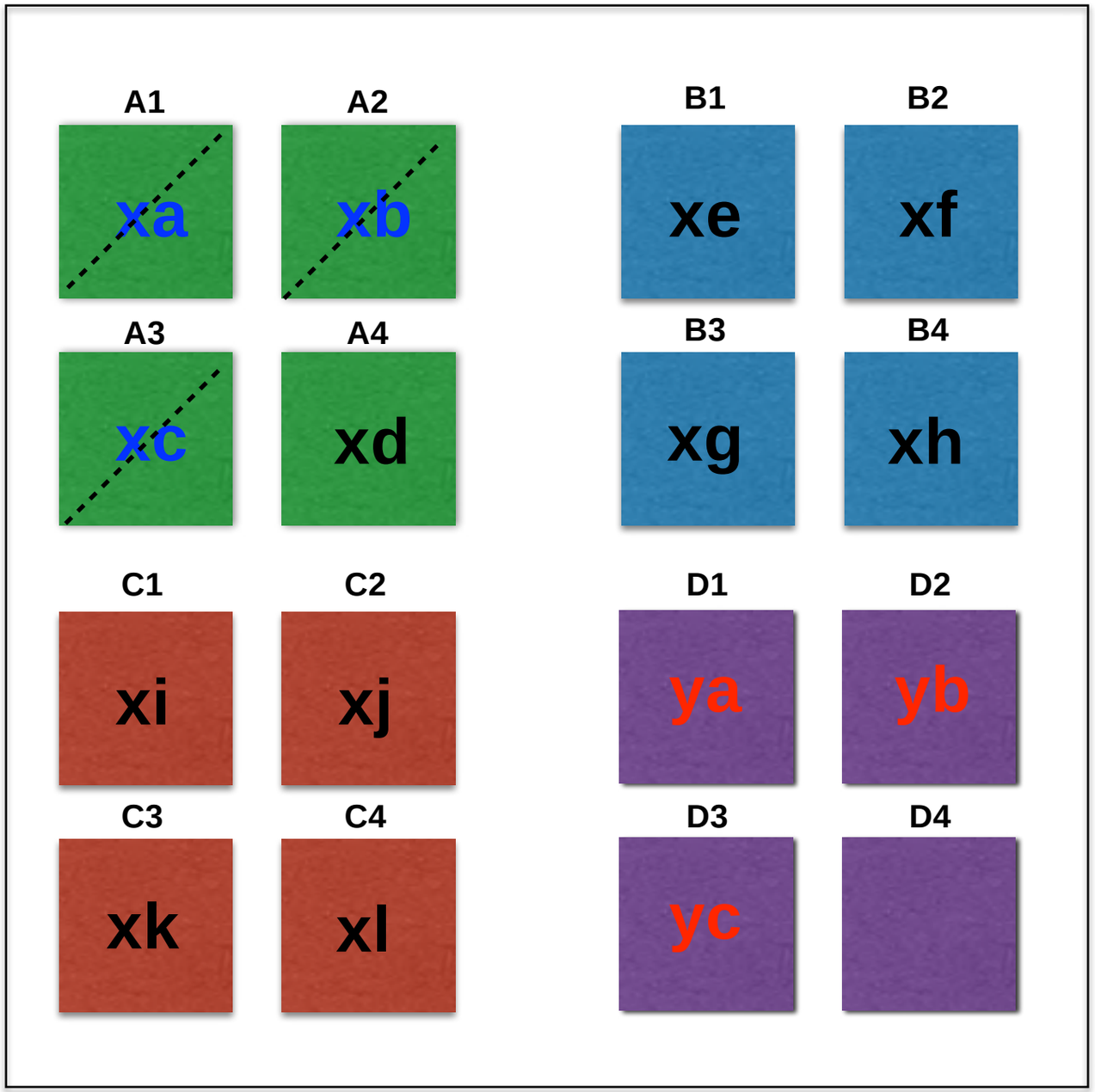
LBA	PBA	Data
1	A1->D1	xa ->ya
2	A2	xb
3	A3	xc
4	A4	xd
5	B1	xe
6	B2	xf
7	B3	xg
8	B4	xh
9	C1	xi
10	C2	xj
11	C3	xk
12	C4	xl

LBA 2 xb -> yb



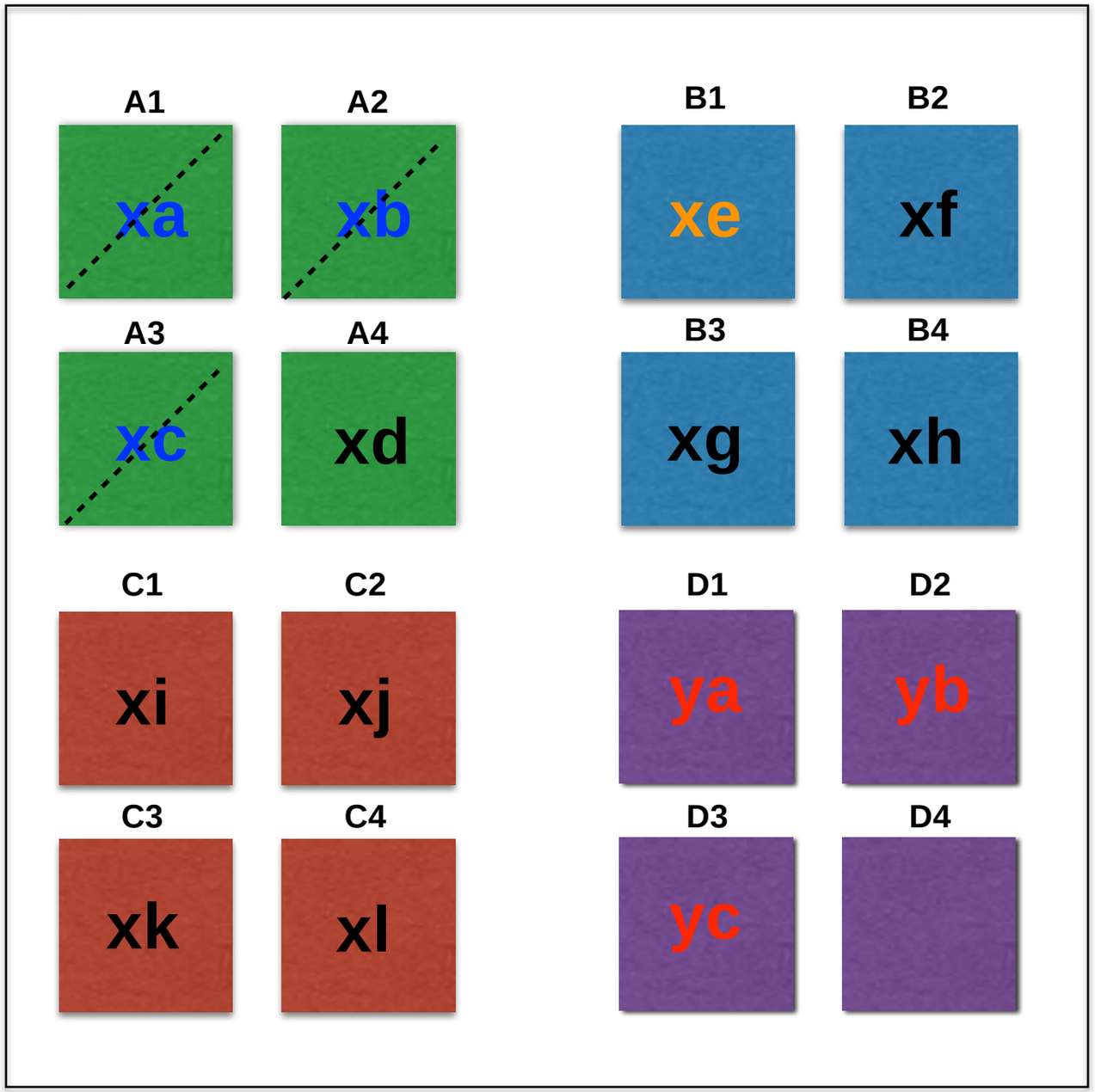
LBA	PBA	Data
1	D1	ya
2	A2->D2	xb->yb
3	A3	xc
4	A4	xd
5	B1	xe
6	B2	xf
7	B3	xg
8	B4	xh
9	C1	xi
10	C2	xj
11	C3	xk
12	C4	xl

LBA 3xc -> yc



LBA	PBA	Data
1	D1	ya
2	D2	yb
3	A3->D3	xc->yc
4	A4	xd
5	B1	xe
6	B2	xf
7	B3	xg
8	B4	xh
9	C1	xi
10	C2	xj
11	C3	xk
12	C4	xl

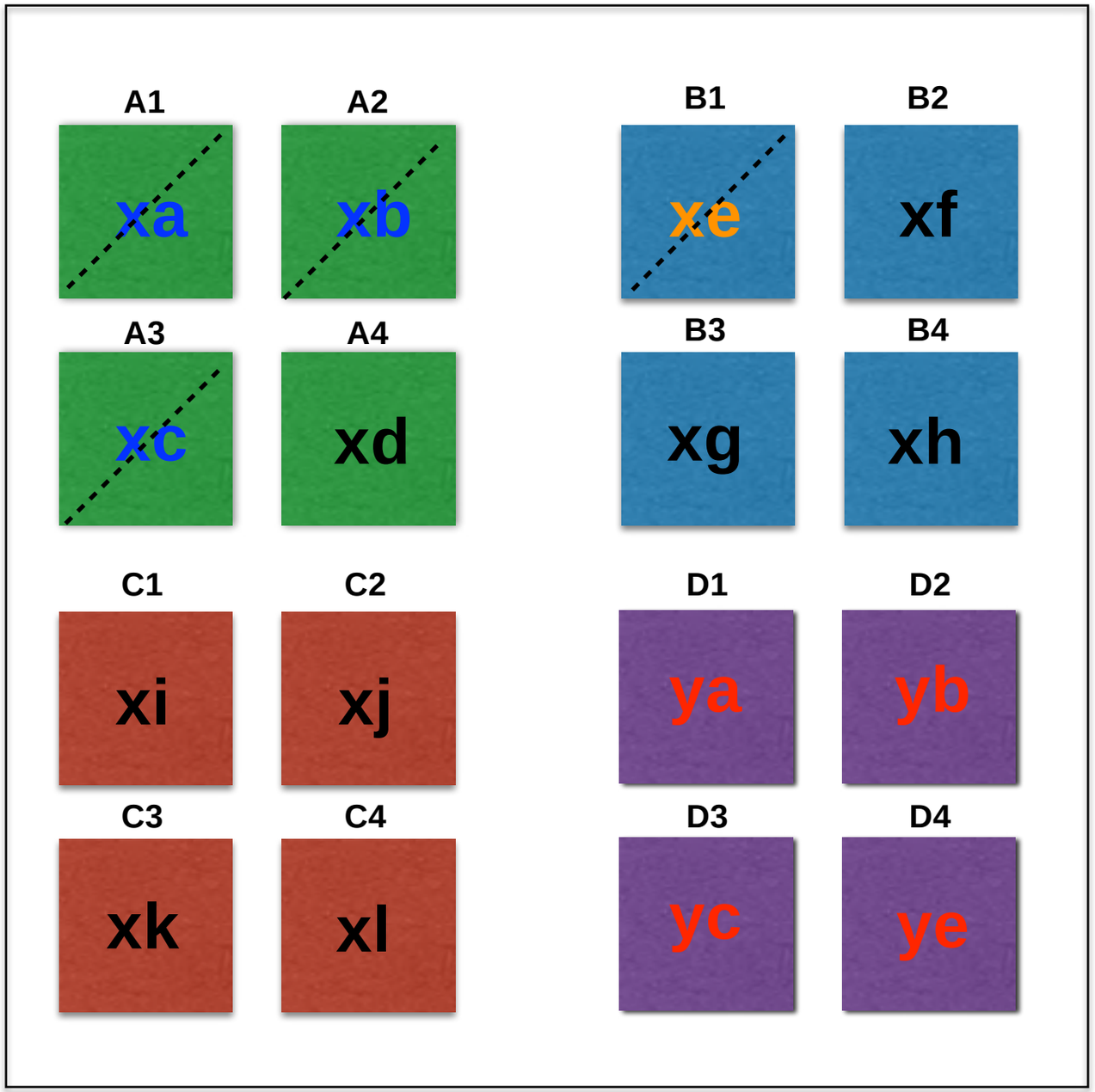
LBA 5 xe -> ye



LBA	PBA	Data
1	D1	ya
2	D2	yb
3	D3	yc
4	A4	xd
5	B1	xe
6	B2	xf
7	B3	xg
8	B4	xh
9	C1	xi
10	C2	xj
11	C3	xk
12	C4	xl

LBA 5 xe -> ye

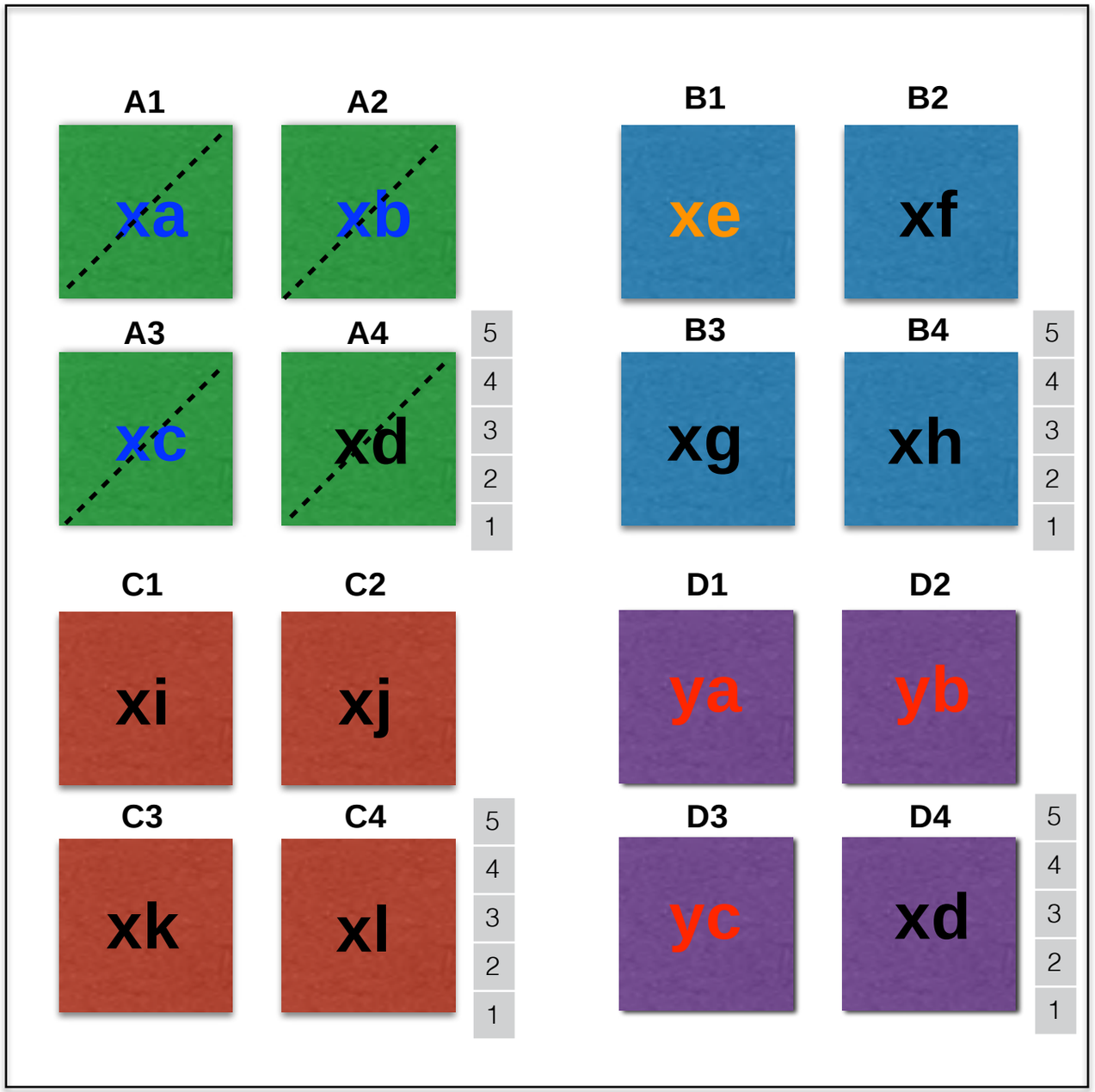
LBA 5 xe -> ye (error)



LBA	PBA	Data
1	D1	ya
2	D2	yb
3	D3	yc
4	A4	xd
5	B1->D4	xe->ye
6	B2	xf
7	B3	xg
8	B4	xh
9	C1	xi
10	C2	xj
11	C3	xk
12	C4	xl

OP空间内至少要有一个free block

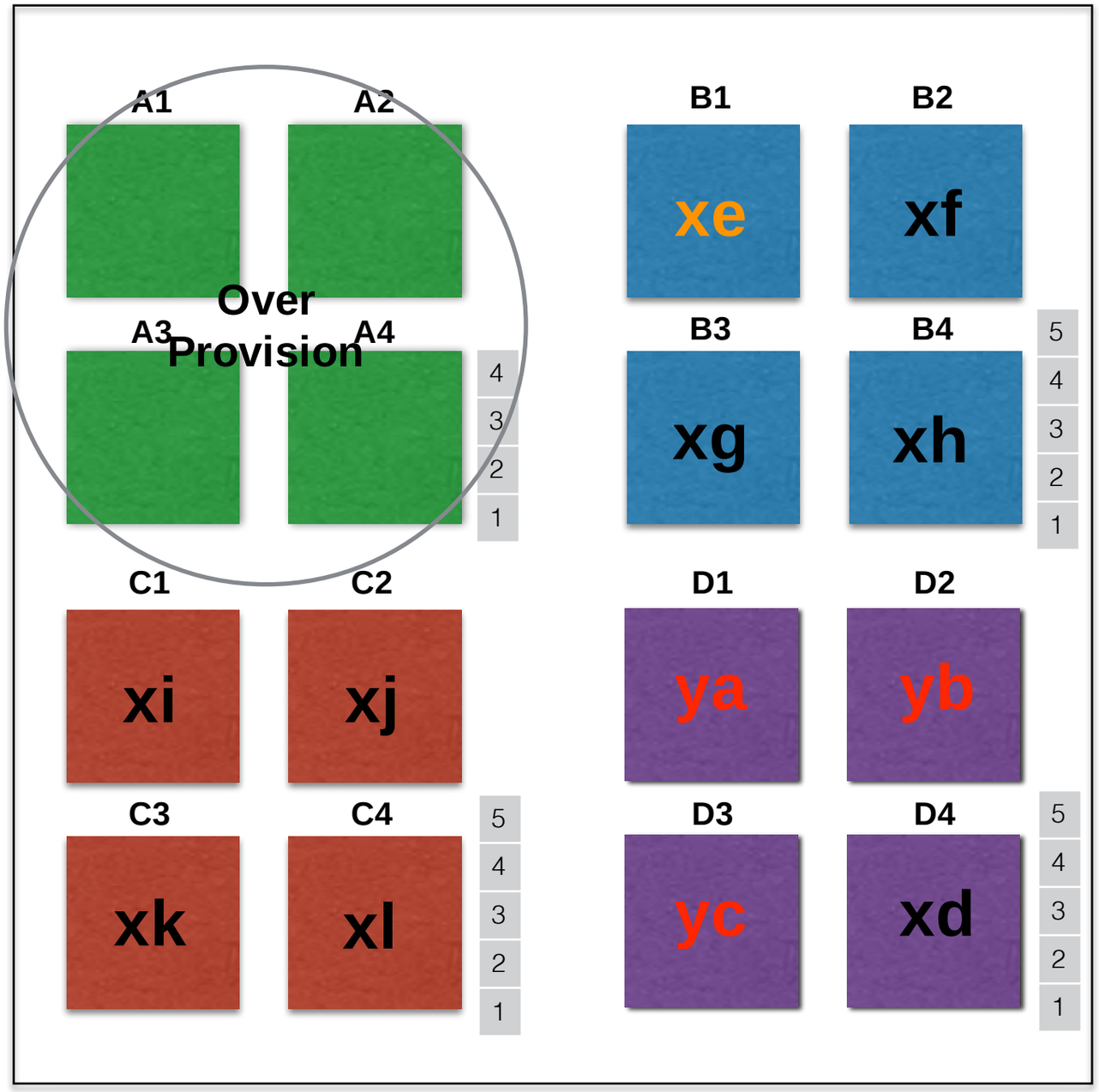
Garbage Collection



LBA	PBA	Data
1	D1	ya
2	D2	yb
3	D3	yc
4	A4->D4	xd
5	B1	xe
6	B2	xf
7	B3	xg
8	B4	xh
9	C1	xi
10	C2	xj
11	C3	xk
12	C4	xl

life(A)=5
life(B)=5
life(C)=5
life(D)=5

Garbage Collection

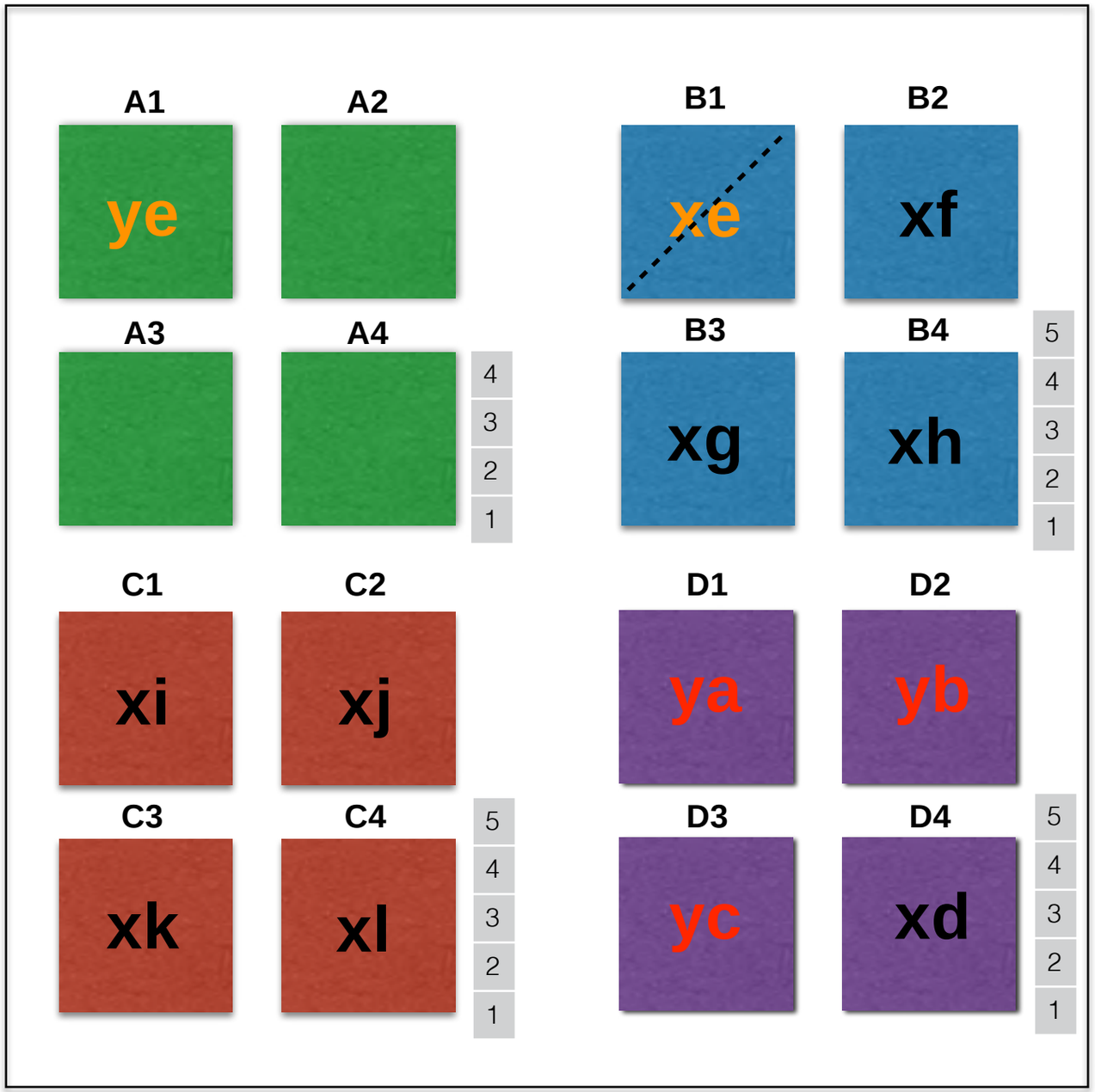


LBA	PBA	Data
1	D1	ya
2	D2	yb
3	D3	yc
4	D4	xd
5	B1	xe
6	B2	xf
7	B3	xg
8	B4	xh
9	C1	xi
10	C2	xj
11	C3	xk
12	C4	xl

life(A)=4
life(B)=5
life(C)=5
life(D)=5

GC是制造free block的过程

LBA 5 xe -> ye

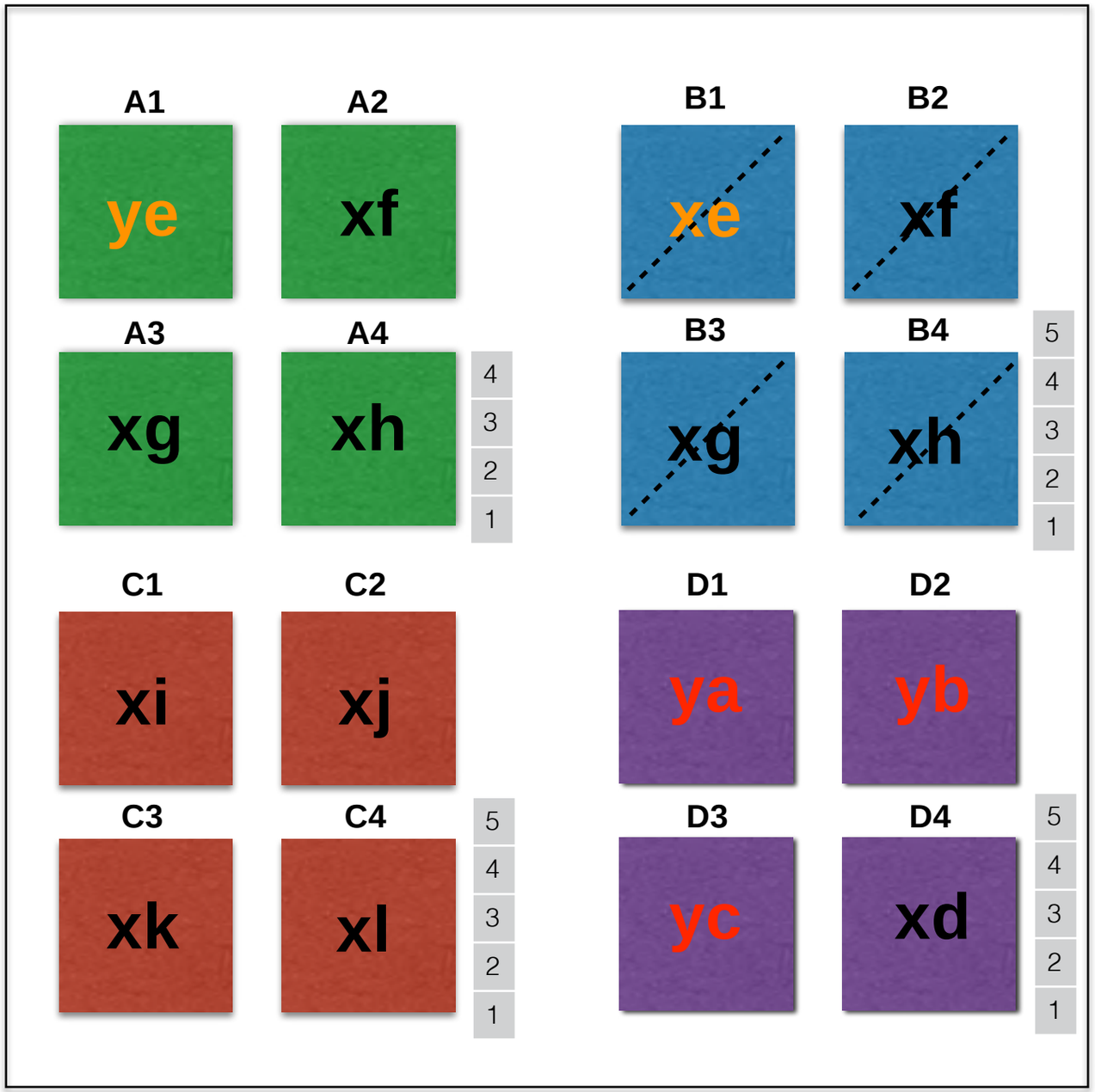


LBA	PBA	Data
1	D1	ya
2	D2	yb
3	D3	yc
4	D4	xd
5	B1->A1	xe->ye
6	B2	xf
7	B3	xg
8	B4	xh
9	C1	xi
10	C2	xj
11	C3	xk
12	C4	xl

life(A)=4
life(B)=5
life(C)=5
life(D)=5

$$\frac{\text{Flash write} = 17 \text{ page}}{\text{Host write} = 16 \text{ page}} = 1.0625 \text{ (Write Amplifier)}$$

LBA 1 ya -> za

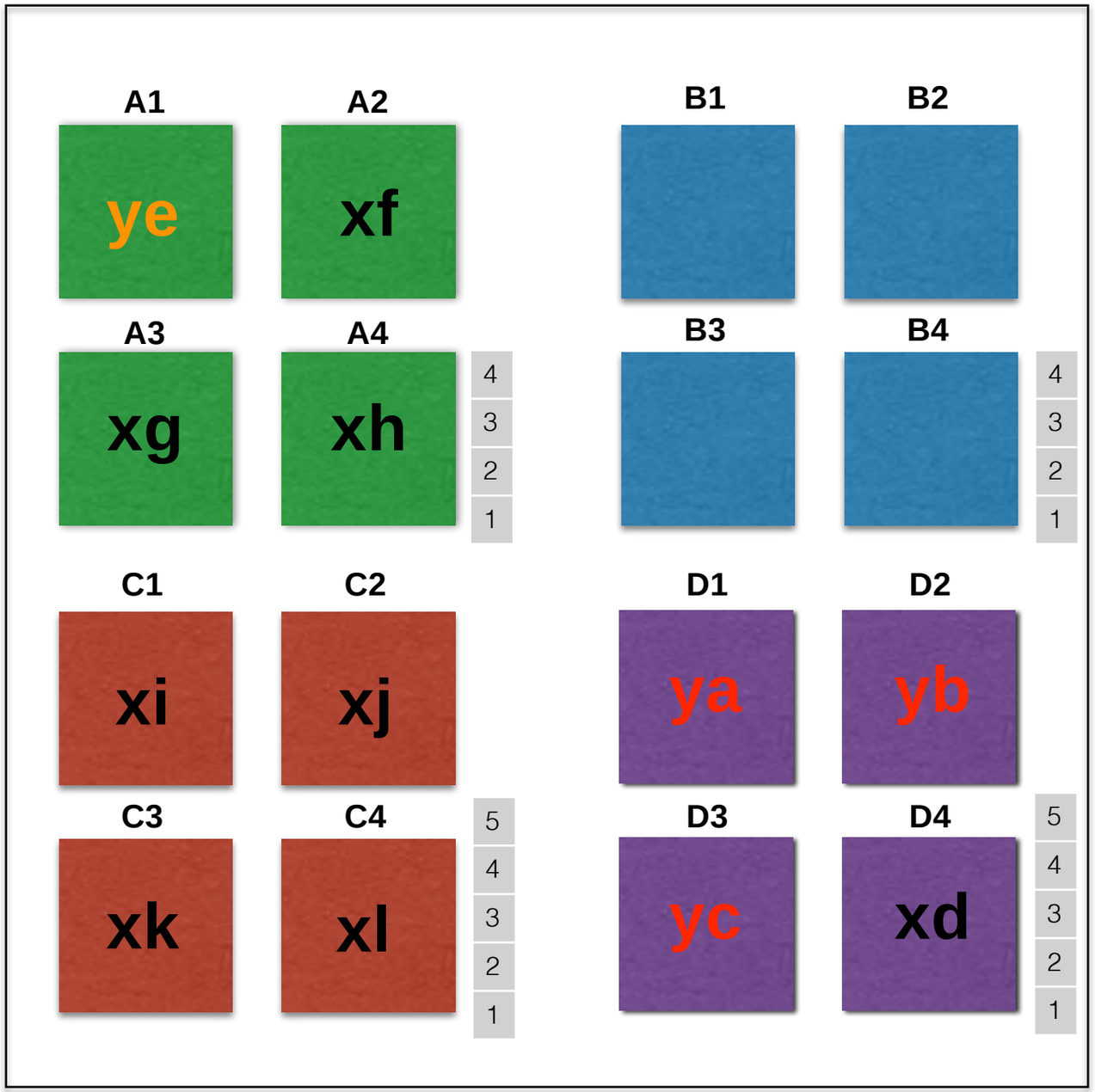


LBA	PBA	Data
1	D1	ya
2	D2	yb
3	D3	yc
4	D4	xd
5	A1	ye
6	B2->A2	xf
7	B3->A3	xg
8	B4->A4	xh
9	C1	xi
10	C2	xj
11	C3	xk
12	C4	xl

LBA 1 ya -> za

life(A)=4
life(B)=5
life(C)=5
life(D)=5

Wear Leveling



LBA	PBA	Data
1	D1	ya
2	D2	yb
3	D3	yc
4	D4	xd
5	A1	ye
6	A2	xf
7	A3	xg
8	A4	xh
9	C1	xi
10	C2	xj
11	C3	xk
12	C4	xl

life(A)=4
life(B)=4
life(C)=5
life(D)=5

FTL,RAID,OP,GC,WL

- GC/WL 异同
- 共同点
- 都是将旧的擦除块里的数据搬到新块里
- 垃圾被去除
- 旧块被释放
- 不同点
- 被迁移块的选择

谢谢