固态存储技术

一、相关技术

1. **SSD： Solid State Disk** 固态盘
2. **SCM： Storage-Class Memory** 存储级内存

1.1 固态盘

分类：

- 基于闪存的固态盘，特点是数据能够持久保持，掉电也能保持数据，随机读性能好
- 基于新型NVM的固态盘，特点是数据能够持久保持，掉电也能保持数据，随机读写性能好
- 基于DRAM的固态盘，特点是读写速度快，但需要独立的电源来保持数据安全，需要备份硬盘来长久地存储数据
- 混合使用DRAM和闪存、DRAM和NVM进行存储的混合Cache结构的固态盘

1.1.1 固态盘(SSD)简介

- **SSD– Solid State Disk　固态盘**

1. 半导体存储设备
2. 块设备

SSD的种类

- 基于NAND FLASH的SSD

  基本存储介质是NAND FLASH。
- 基于3D Xpoint的SSD

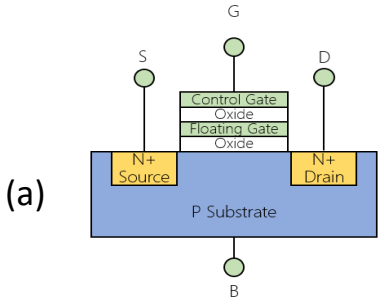
  基本存储介质是3D Xpoint。
- 基于DDR DRAM的SSD

  基本存储介质是DRAM。

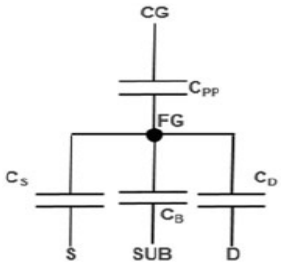
注：后面所提到的SSD均特指基于NAND FLASH的SSD

闪存单元结构

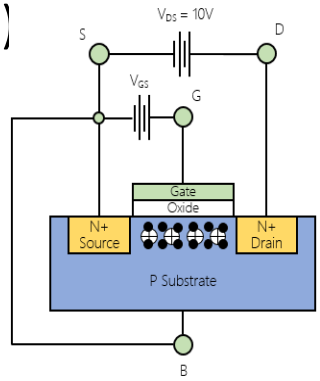
- 目前主流闪存单元分为浮栅单元和电荷俘获单元；
- 通过一定外加电场作用，使闪存单元俘获/排出电荷；改变存储单元阈值电压的高低，表示逻辑0、1；
  - 写：字线与位线间施加较大电压(20V)
  - 读：字线与位线间施加判断电压(≈2V)
  - 擦除：字线与位线间施加反向电压(20V)



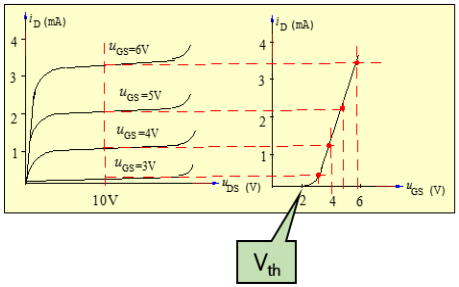
(a)



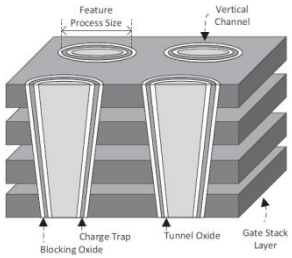
(b)



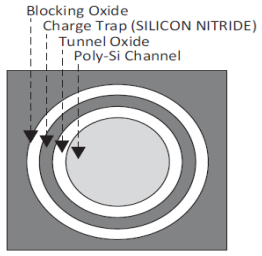
$i_D = f(u_{GS}) \mid u_{DS} = const$



阈值电压转移特性曲线



(c)

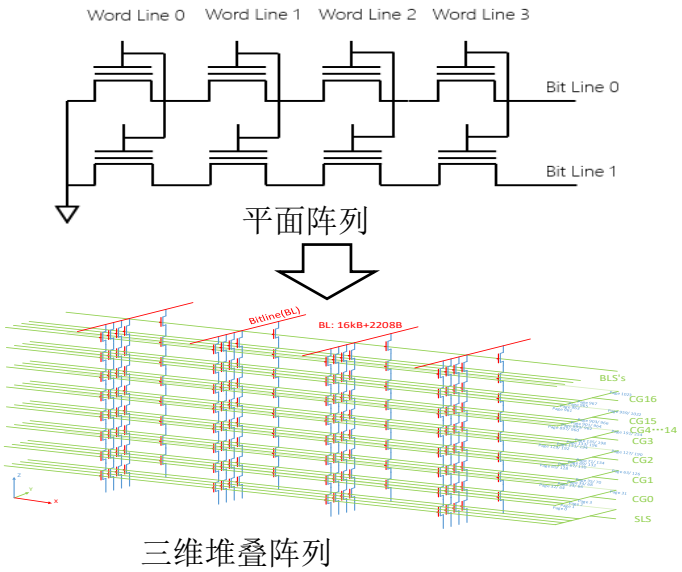


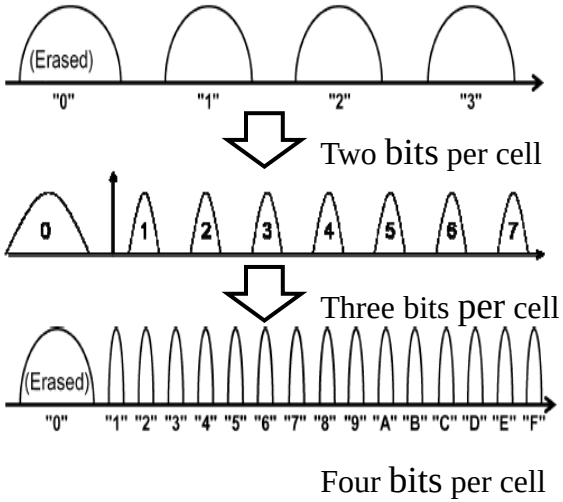
(d)

（a）浮栅单元结构
（b）浮栅单元电容模型
（c）电荷俘获单元闪存垂直通道
（d）电荷俘获单元结构

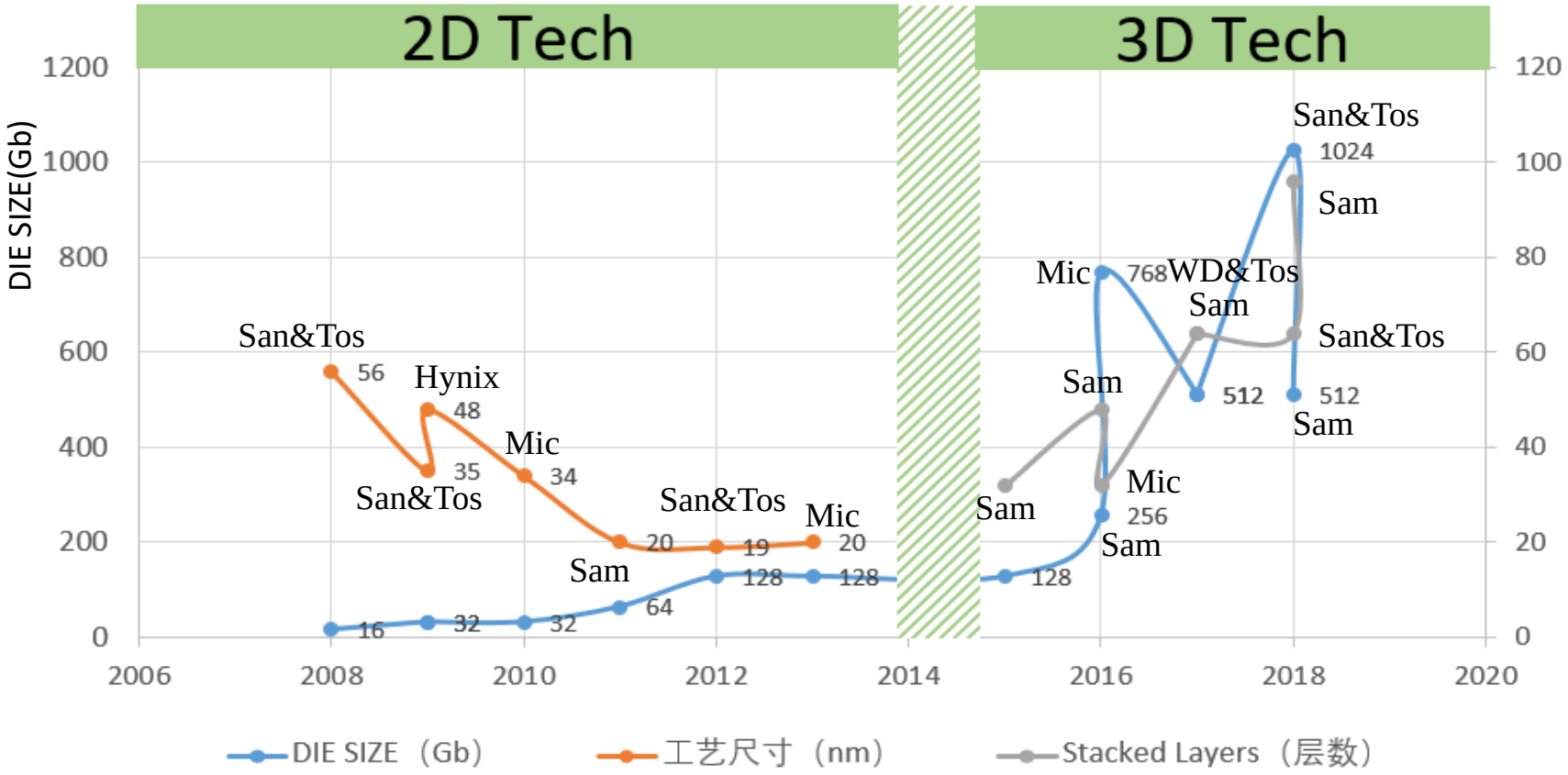
闪存阵列结构

- 从2D阵列转向3D堆叠阵列
  - 工艺尺寸缩小到达物理极限
  - 垂直堆叠进一步增加存储密度，堆叠层数逐年增加
- 单元存储更多逻辑比特
  - 存储密度进一步提高（MLC->TLC->QLC）
  - 阈值电压区间被压缩，可靠性和编程速度需要进一步研究





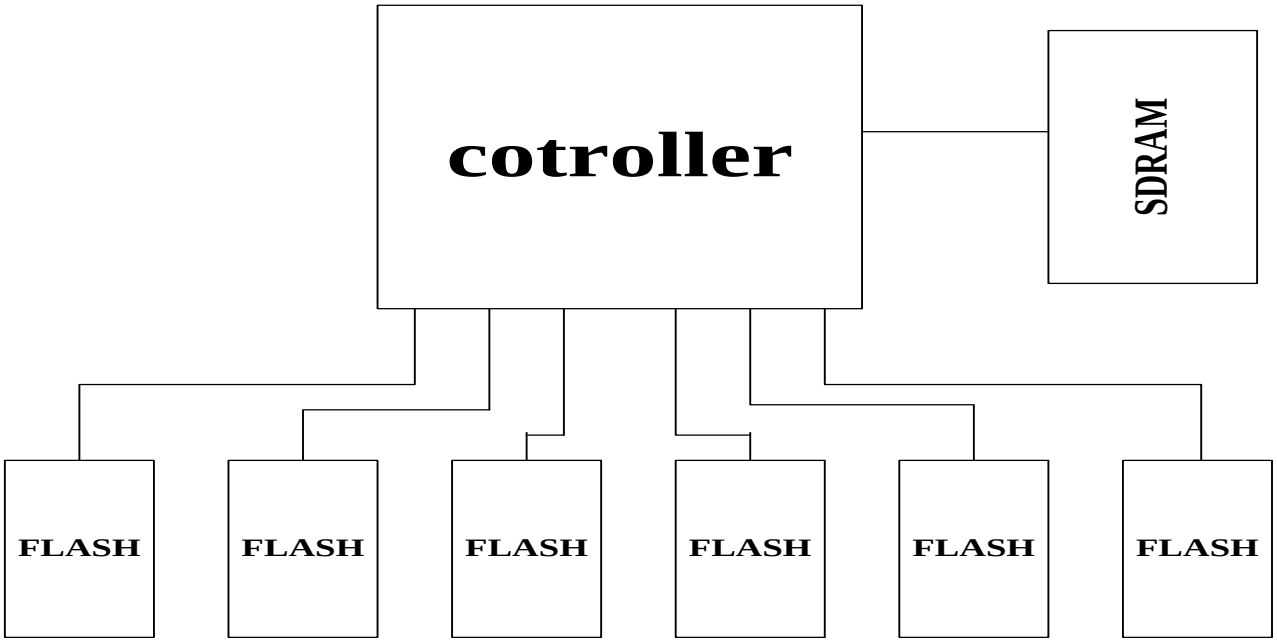
闪存芯片发展历程



图中缩写 Sandisk: San Toshiba: Tos Samsung: Sam Micron: Mic Western Digital : WD

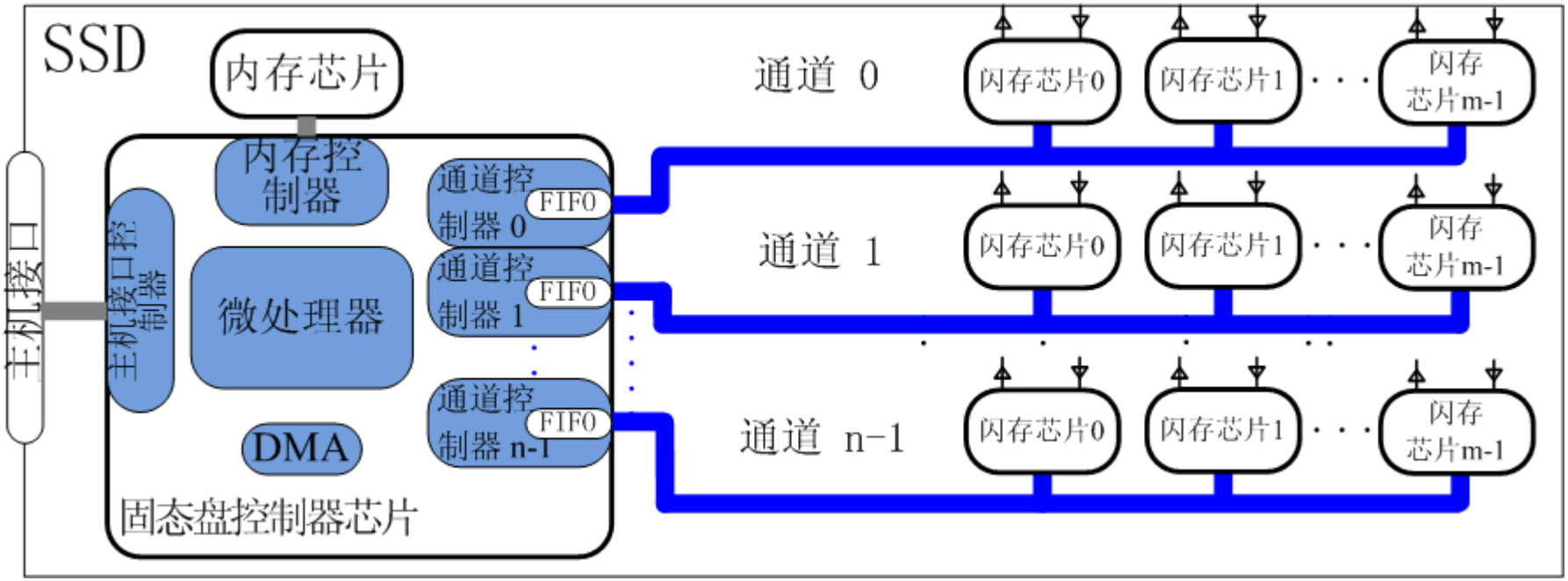
SSD

- 由NAND FLASH作为存储介质，由一个嵌入式控制器控制NAND FLASH的操作，RAM作为buffer，通过IDE,SATA,PCI-e等总线对外提供块接口

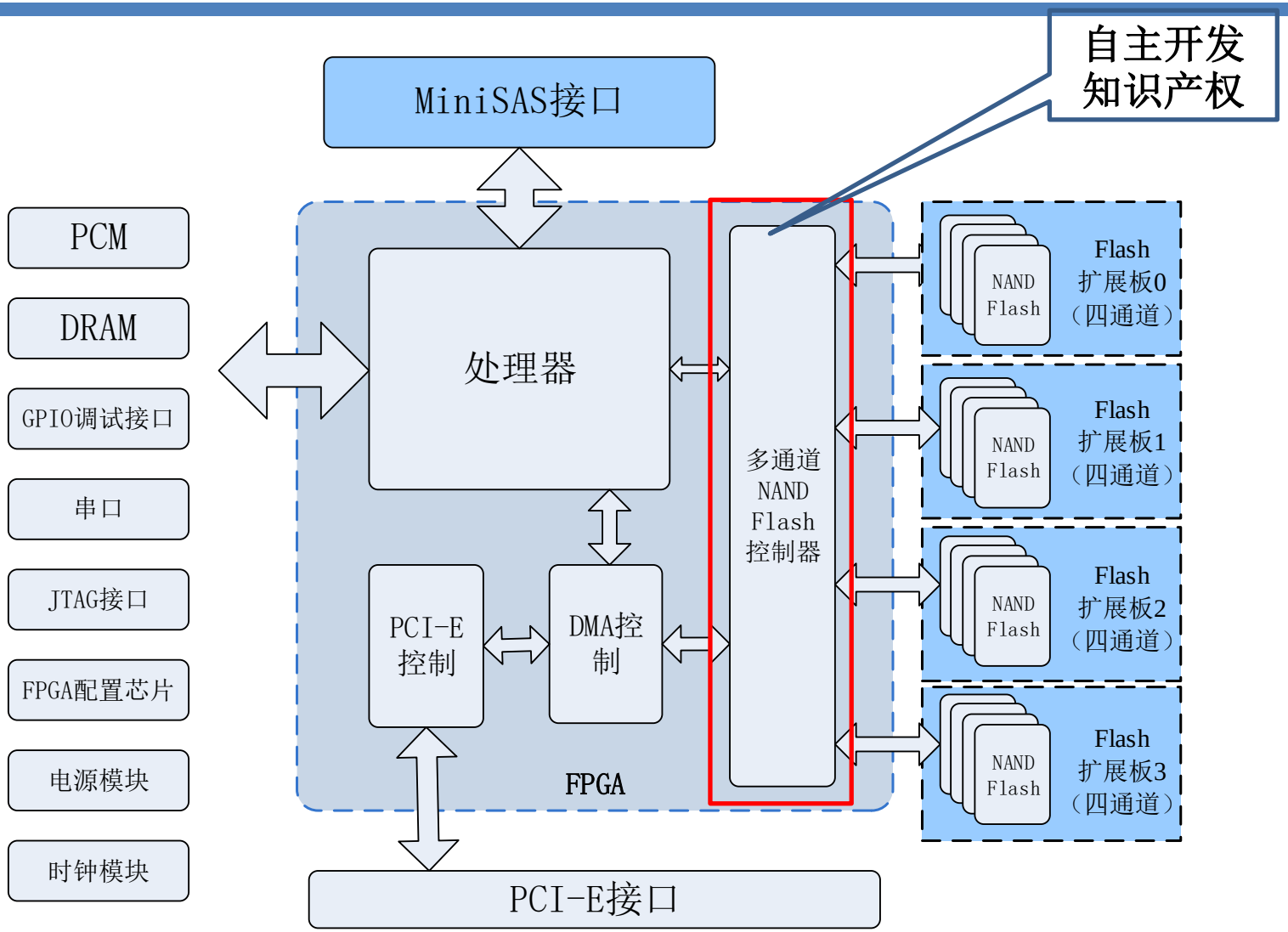


1.1.2 SSD构成

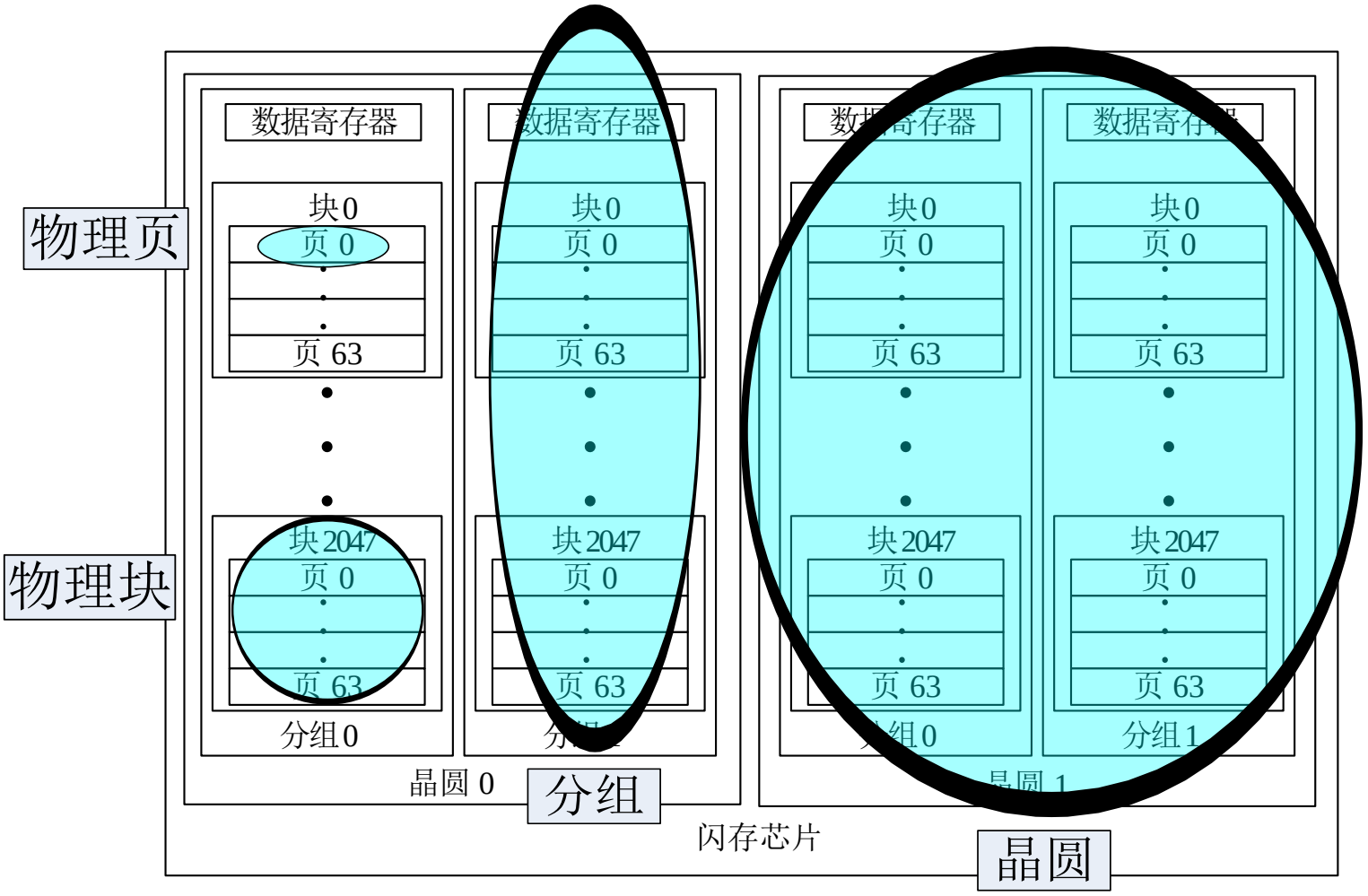
(1) 硬件结构



实例：PCIe接口SSD体系结构

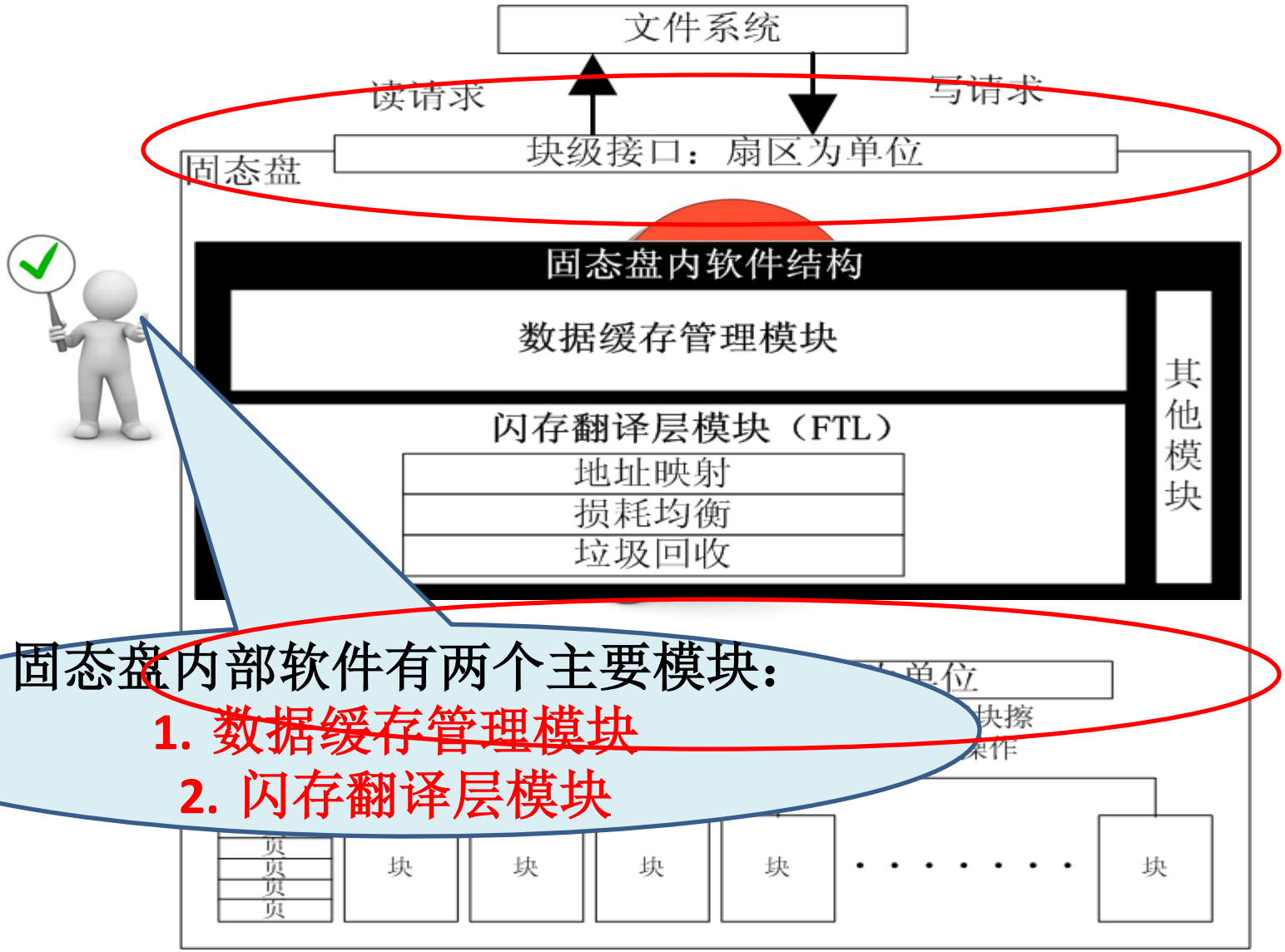


Flash芯片内部结构



例：三星的K9F1208U0M，每页528Bytes(512byte(Main Area)+16byte(Spare Area))，每32个page形成一个Block(32*528B)。有4096个block，故总容量为4096*（32*528B）=66MB，即实际可用容量64MB+ 2MBECC校验码。

(2) SSD软件结构



固态盘内部软件有两个主要模块：

1. 数据缓存管理模块
2. 闪存翻译层模块

(3) FTL(Flash Translation Layer)

- SSD是以硬盘的替代者的姿态出现，为了与现有系统无缝对接，SSD必须对外提供的是块接口，作为主机端，所看到的SSD是一个和HDD一样的块设备。
- 为了达到模拟块设备的目的，SSD中需要FTL作为中间层
- FTL: flash translation layer
- FTL从主机文件系统接收块级请求（LSN, size），经过FTL的处理，产生flash的各种控制命令

FTL

- FTL由三部分组成：
  - Address mapping (地址映射)
  - Wear leveling (损耗平衡)
  - Garbage collection (垃圾回收)

Address mapping (地址映射)

- 上层文件系统发送给SSD的任何读写命令包括两个部分（LSN，size）
- LSN是逻辑扇区号，对于文件系统而言，它所看到的存储空间是一个线性的连续空间。例如，读请求（260，6）表示的是需要读取从扇区号为260的逻辑扇区开始，总共6个扇区。
- 请求到达SSD后，需要经过地址转换，将逻辑扇区转换成NAND FLASH中的物理页号

<package，die，plane，block，page>

Address mapping (地址映射)

- 映射方式有很多种，常用的有三种：

✓页级映射

✓块级映射

✓混合映射

	性能	寿命	映射表大小	所需内存大小	成本
页级映射	好	长	大	大	高
块级映射	差	短	小	小	低
混合映射	较差	较短	较小	较小	较低

损耗平衡（Wear-Leveling）

- Flash中每个块都有一定的擦写次数限制。故不能让某一个块被写次数较多，而其他块被写的次数较少。
- 需要找一种方法：使flash中每个块被擦写的次数基本相同。

WL的基本方法

- 动态损耗平衡

  在请求到达时，选取擦除次数较少的块作为请求的物理地址。

- 静态损耗平衡

  在运行一段时间后，有些块存放的数据一直没有更新（冷数据），而有些块的数据经常性的更新（热数据）。那些存放冷数据的块的擦除次数远小于存放热数据的块。将冷数据从原块取出，存放在擦除次数过多的块，原来存放冷数据的块被释放出来，接受热数据的擦写。

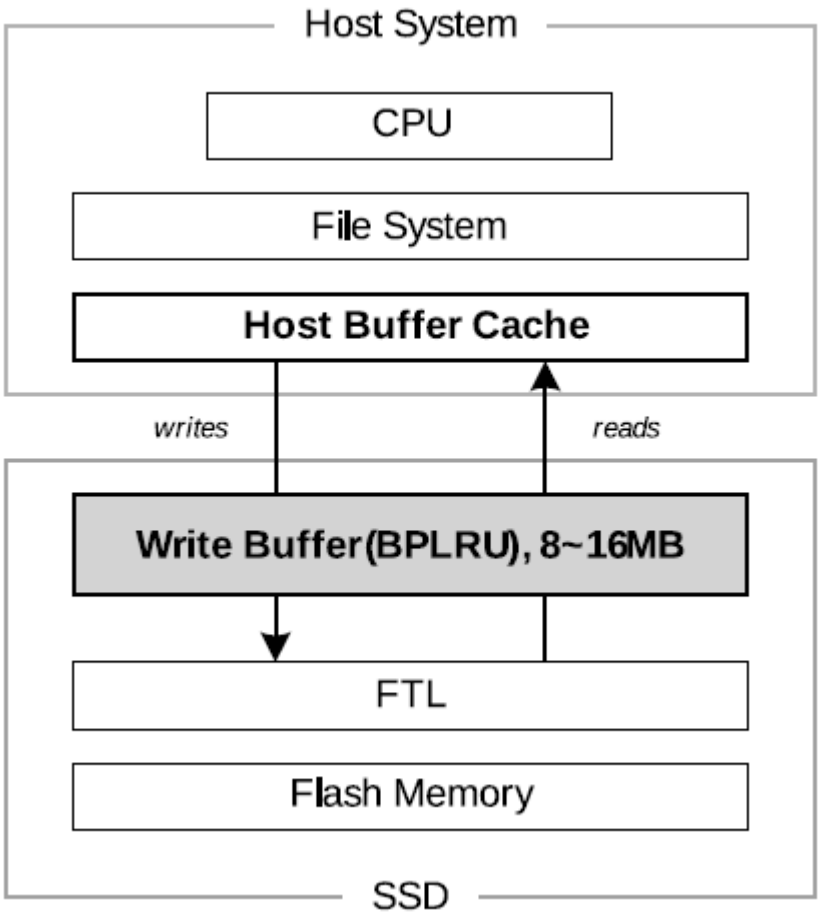
垃圾回收(Garbage Collection)

- 垃圾回收的目的

  SSD在使用过程中，会产生大量失效页，在SSD的容量到达一定阈值时，需要调用GC函数，清除所有失效页，以增加可用空间。

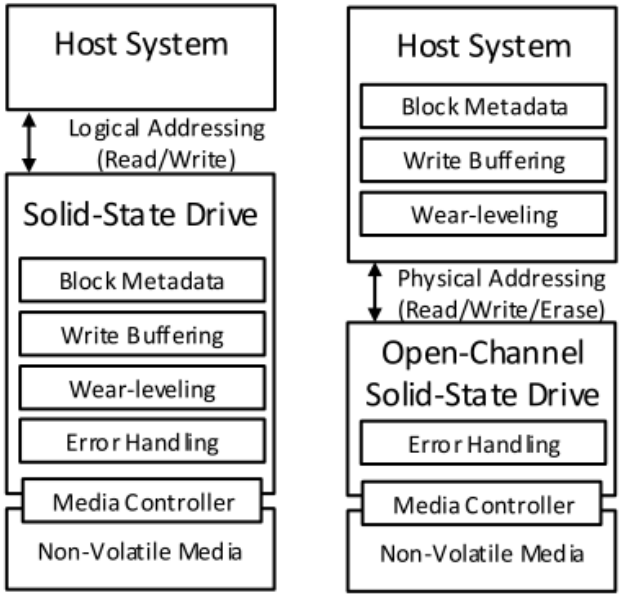
（4）SSD中buffer策略

- 好的buffer策略能够提高SSD的整体性能。



SSD研究现状

- 可靠性问题
  - 硬件减少编程干扰[IWM'13][ICCD'17]
  - 软件优化可靠性
    - 错误数据分散到无效数据[TOC'16]
- 大页问题
  - 提升大页使用效率[DATE'17]
- 垃圾回收优化研究
  - 分离flash页基于buffer中的脏位[DATE'18]
- Open Channel
  - lightNVM[FAST'17]
- 系统应用性能优化
  - KV SSD
    - 多日志结构，可变大小记录数据[HPCA'17]



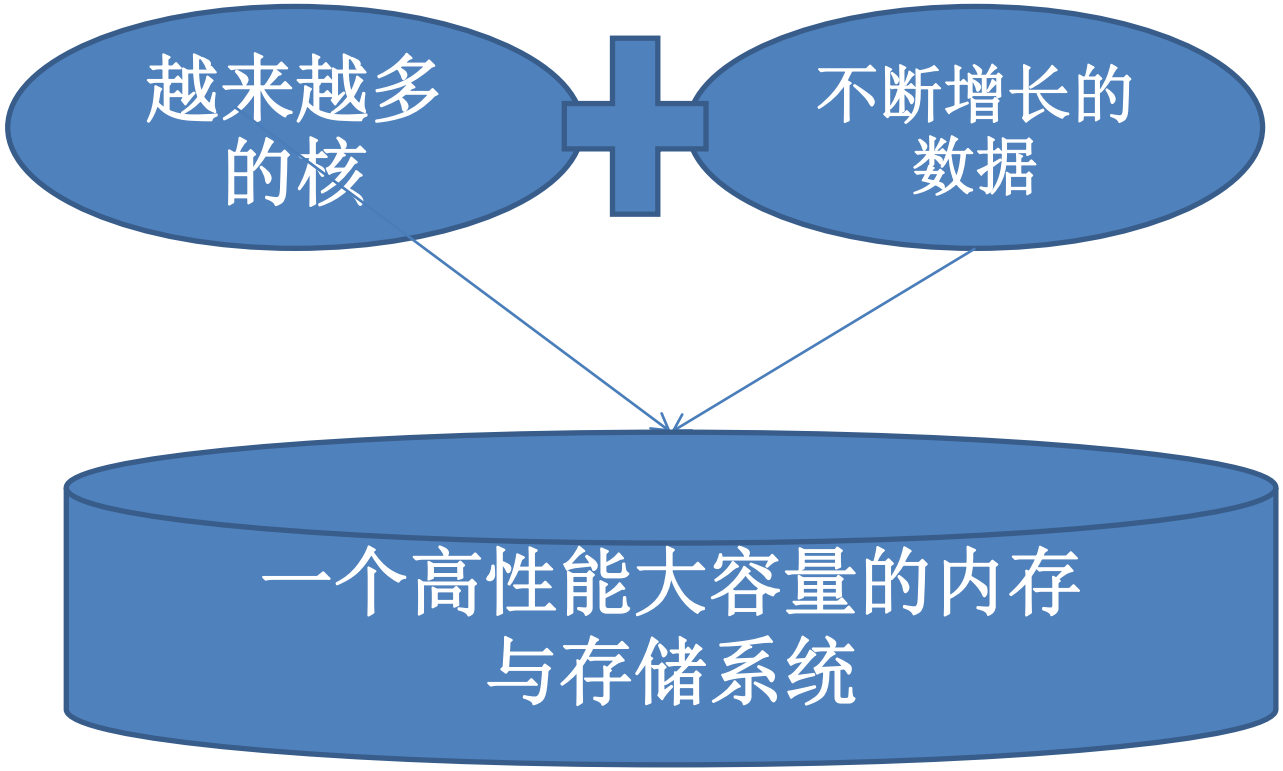
Open channel SSD

1.2 SCM (Storage-Class Memory)

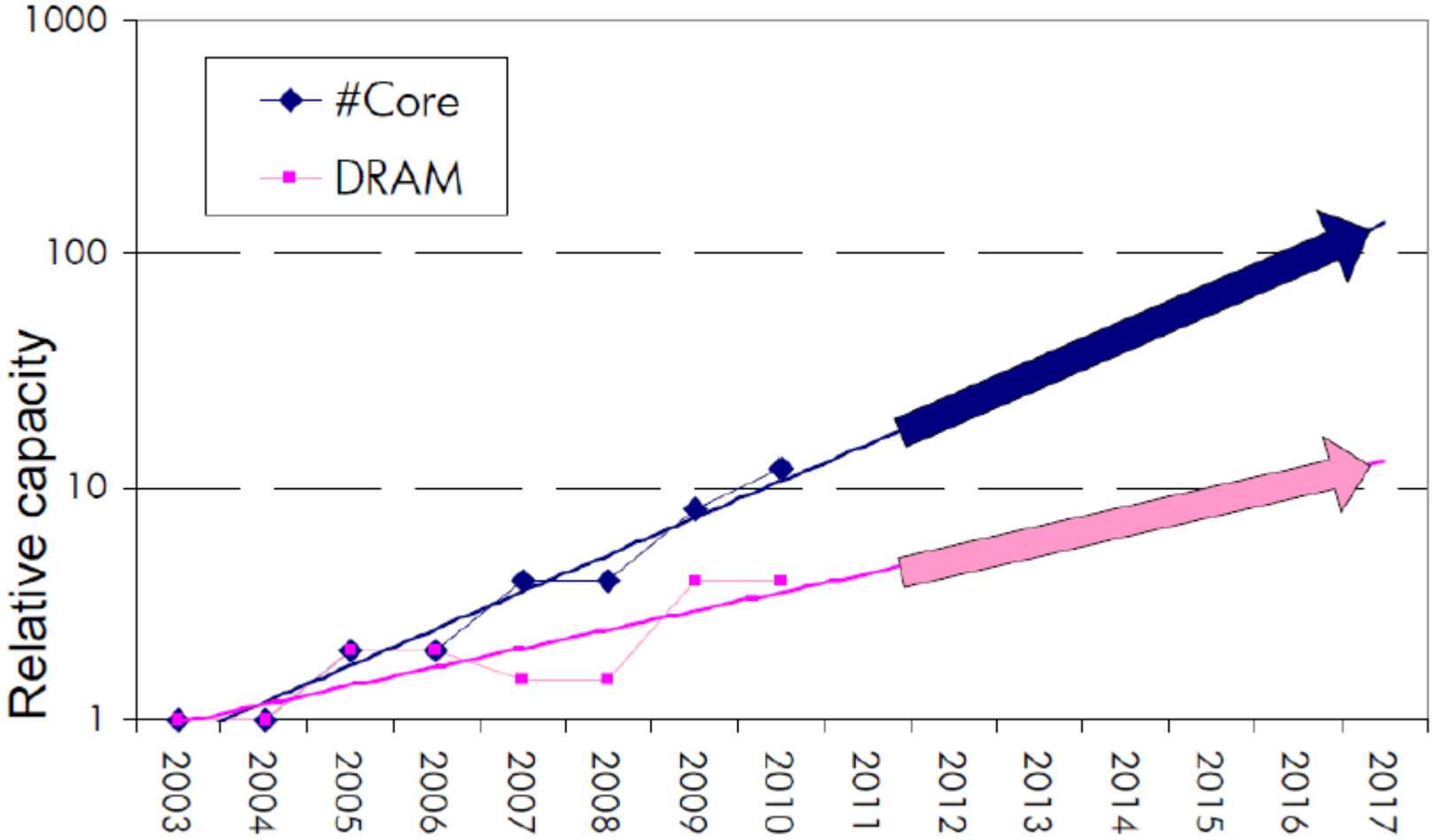
- SCM出现的前提：

1. 数据持续增长
2. 处理器的核越来越多

对高性能内存与存储系统需求

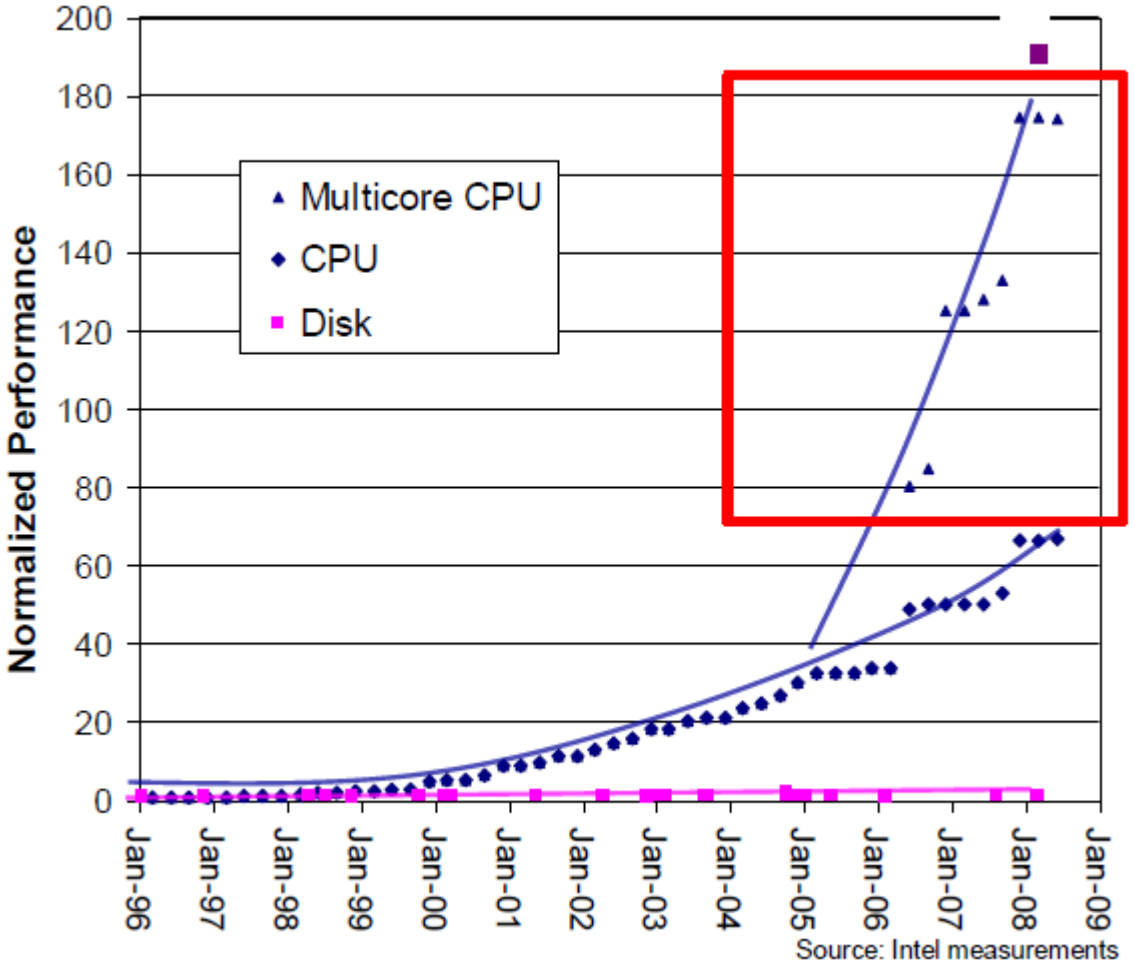


现状与趋势—每核拥有内存容量下降



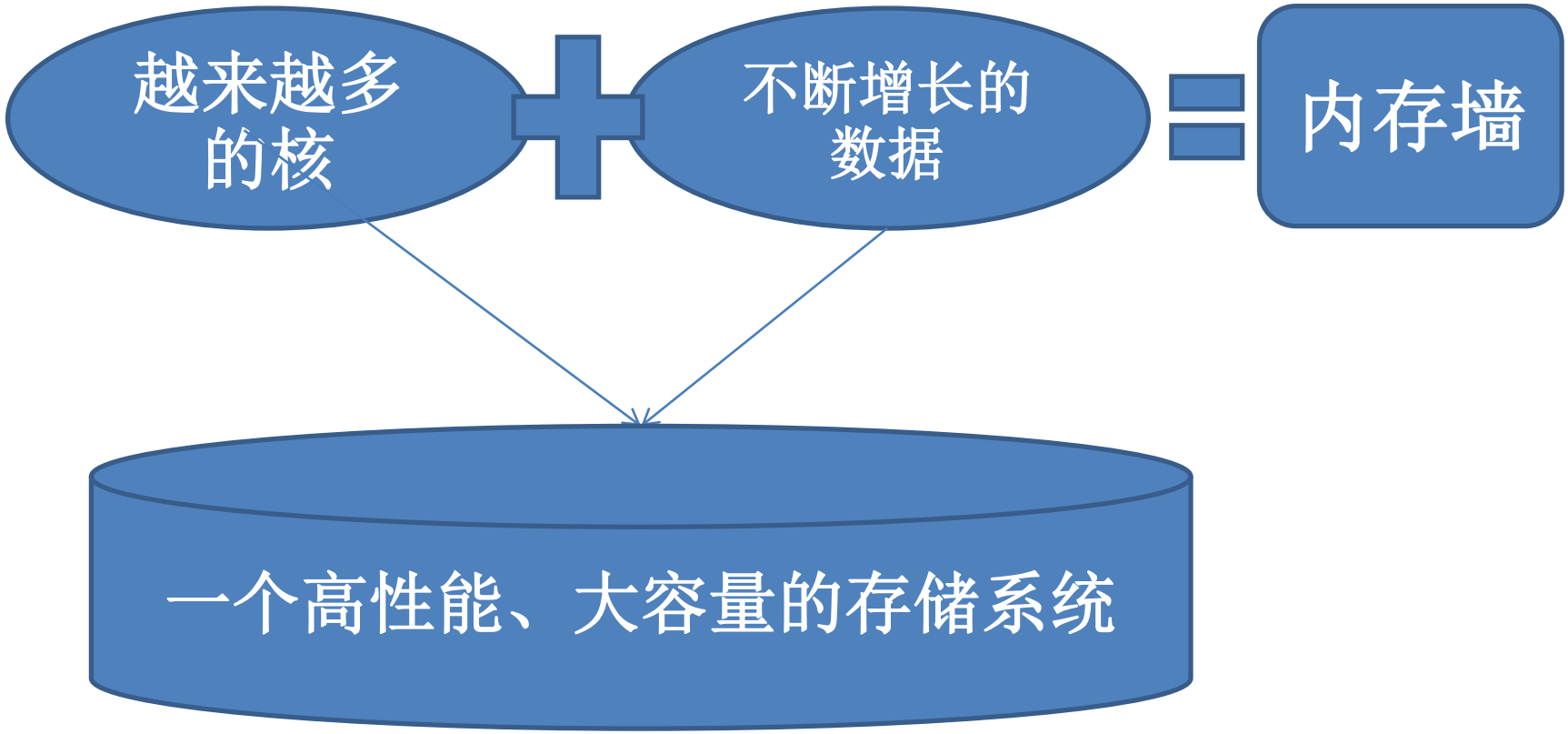
Source: Lim et al. ISCA 2009

现状与趋势-性能差距越来越大



Source: Intel measurements

对高性能内存与存储系统需求



现状与趋势—能耗正成为系统扩展的瓶颈

- **DRAM：不断刷新消耗了系统的巨大能耗：30-40%【1】。**
- **Disk: 磁盘存储子系统消耗了系统20-30%的能耗。**

[1]. Dhiman, G., R. Ayoub and T. Rosing. PDRAM: a hybrid PRAM and DRAM main memory system. in Proceedings of the 46th Annual Design Automation Conference. 2009. San Francisco, California: ACM.

[2]. Bisson, T., S.A. Brandt and D.D.E. Long. NVCache: Increasing the Effectiveness of Disk Spin-Down Algorithms with Caching. in Proceedings of the 14th IEEE International Symposium on Modeling, Analysis, and Simulation. 2006: IEEE Computer Society.

SCM的提出

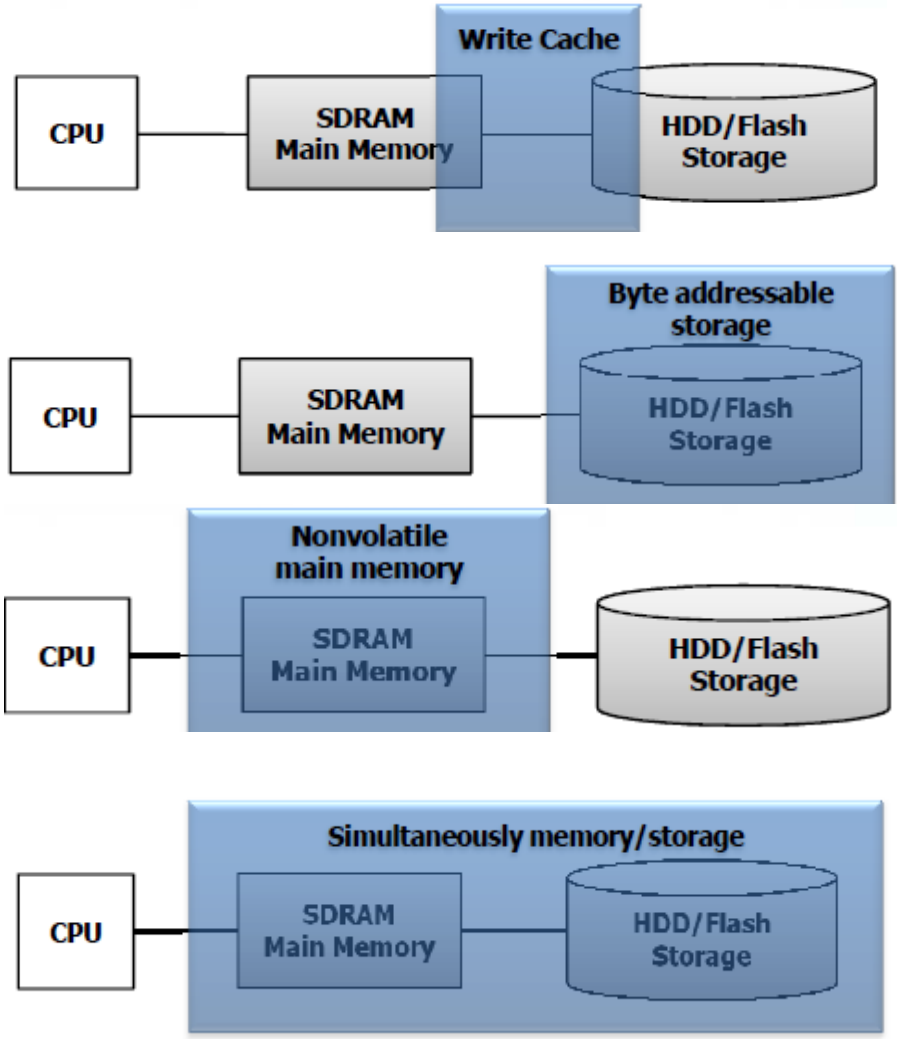
- 非易失
- 零或低空闲能耗
- 类似磁盘一样的容量
- 接近DRAM的存取延迟
- 字节级编址

将为未来Exa计算提供存储解决方案

SCM技术

- PCM
- STT-RAM
- MRAM
- FeRAM
- Memristor
- ……

集成SCM技术的四种策略



缓存策略（e.g.: flashcache@ISCA 2008）

存储替代策略（e.g.: Moneta@Micro 2010, ASPLOS 2012）

内存替代混合策略(e.g.: PDRAM@DAC 2009)

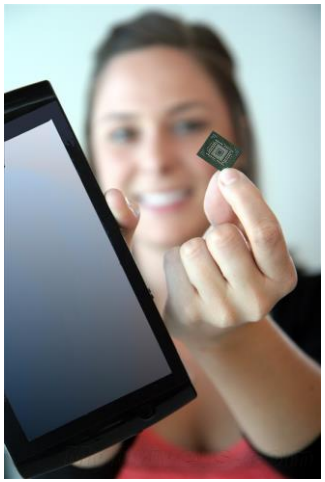
单级存储策略（e.g.: Mnemosyne@ASPLOS 2011）

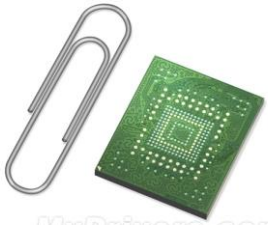
目前集中研究的方向

- **OS对SCM技术的支持，特别是内存数据结构的持久化。**
- **文件系统（SCMFS、BPFS）**
- **SCM的应用（系统恢复、检查点等）**

二、固态盘接口

- **SATA接口**
- **PCI-E总线的SSD接口标准**
- **NVMe接口**





SATA μSSD

- ATA-IO发布了面向嵌入式固态存储设备的新标准“SATA μSSD”
- SATA μSSD是一种嵌入式存储解决方案，移除了传统SATA界面的接口模块，而定义了一种新的电气针脚布局，从而在BGA封装的单颗芯片内实现SATA传输，并直接与主板相连。这也是迄今为止最为小巧的物理SATA实现方案，非常适合超轻薄笔记本、平板机等小型移动设备。
- SanDisk已经推出了基于该标准的iSSD系列固态硬盘产品，容量有8GB、16GB、32GB、64GB、128GB等，持续读取速度最高450MB/s，持续写入速度最高160MB/s，功耗最低10毫瓦。
- μSSD是目前所有硬盘接口中，最为小巧的一种接口形式，因此主要应用于UltraBook、平板电脑、小型移动设备上

SATA Express SSD

- SSD的快速发展对SATA接口速度提出了更高的要求
- SATA Express
  - 传输带宽将提升到8Gbps和16Gbps
  - 在SATA中融入PCI Express（PCI-E）技术
  - 定义了新的设备和主板连接器
  - 连接器支持新的SATA Express和现有的SATA设备。

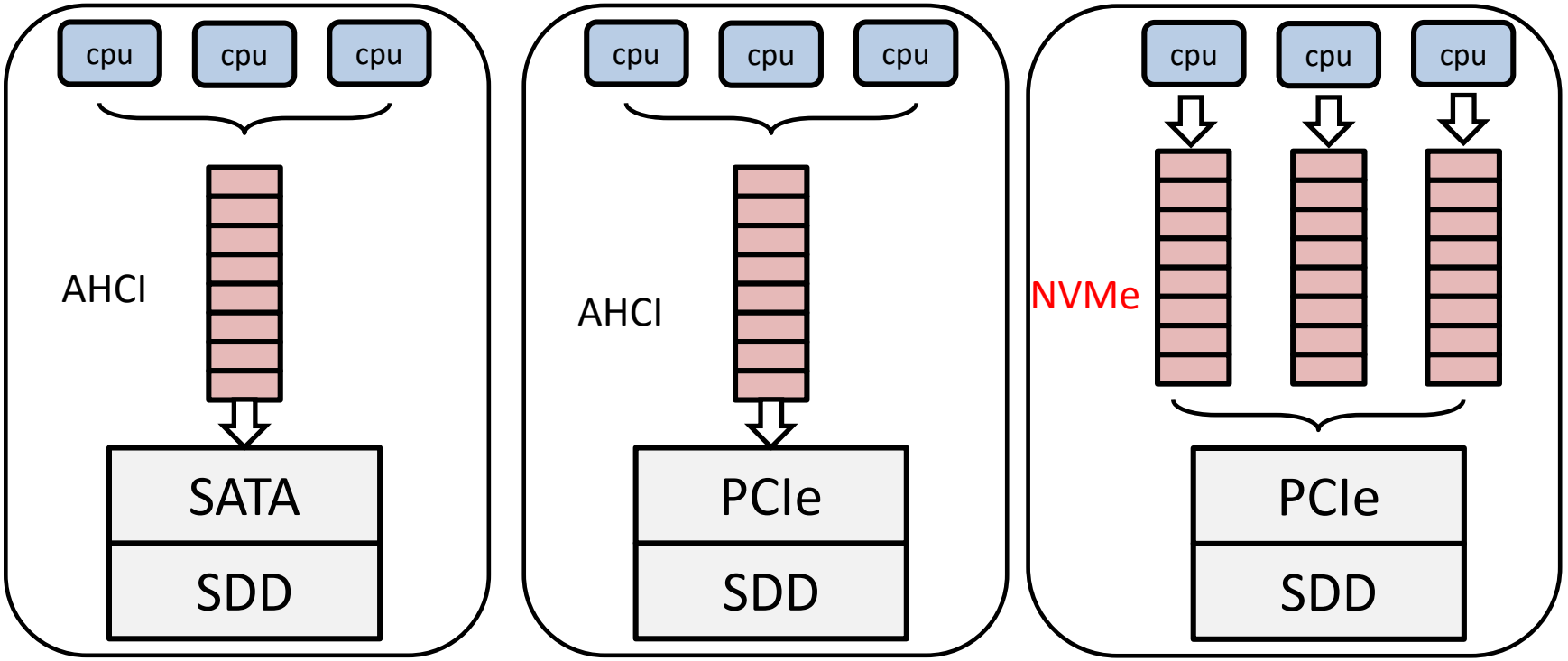
NVME Express

- **NVME Express是一个在工业界推广PCIe SSD的标准**
- **针对使用PCI Express SSD的企业和普通客户端系统开发的可扩展的主机控制芯片接口标准（包含优化的寄存器接口和指令集）**
- **由业界80多个公司成员合作开发，并由11发起企业主导。**

NVME Express

- 特点：采用了PCIe接口，定义了命令集以及命令发送机制，自定义了数据传输机制，采用了端到端的保护，采用了动态的电源管理策略，支持固件更新，支持SRIOV。
- 优点：采用了高速总线，能够提升读写性能，降低开销；提供了数据保护机制并有很强的容错能力。

SSD的接口发展


cpu
cpu
cpu
AHCI
SATA
SDD
cpu
cpu
cpu
AHCI
PCIe
SDD
cpu
cpu
cpu
NVMe
PCIe
SDD

SSD的接口发展

- AHCI vs NVMe

	AHCI	NVMe
最大队列深度	单命令队列 队列深度 32	65536个队列 队列深度 65536
寄存器访问	6~9次	2次
中断	单中断	2048个MSI-X中断
并行与多线程	需要加同步锁	不需要加锁
4KB命令处理效率	命令参数需要两次串行主机DRAM提取	获取命令参数 在一个64字节的提取

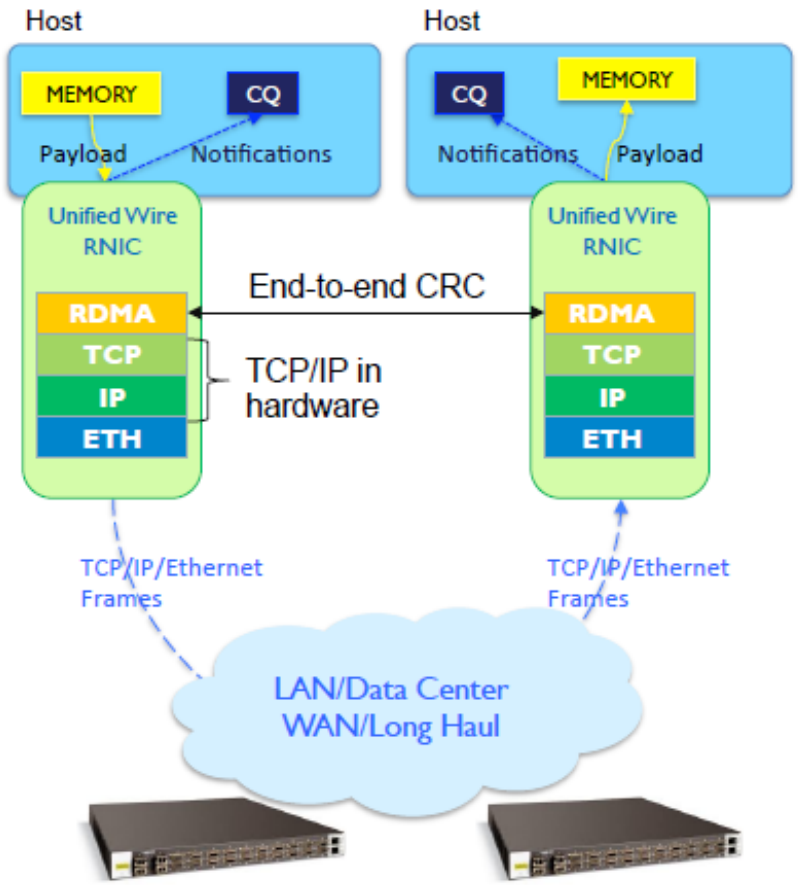
NVMe over Fabric VS NVMe over PCIe

- NVMe over Fabric在NVMe基础上的一些限制
  - 严格1：1对应SQ和CQ
  - 因为没有流量控制，所以CQ必须维持其最大队长
  - 元数据必须和数据在一起，目前不支持元数据和数据的分开传输的方式。
- NVMe over Fabric和NVMe协议的区别：

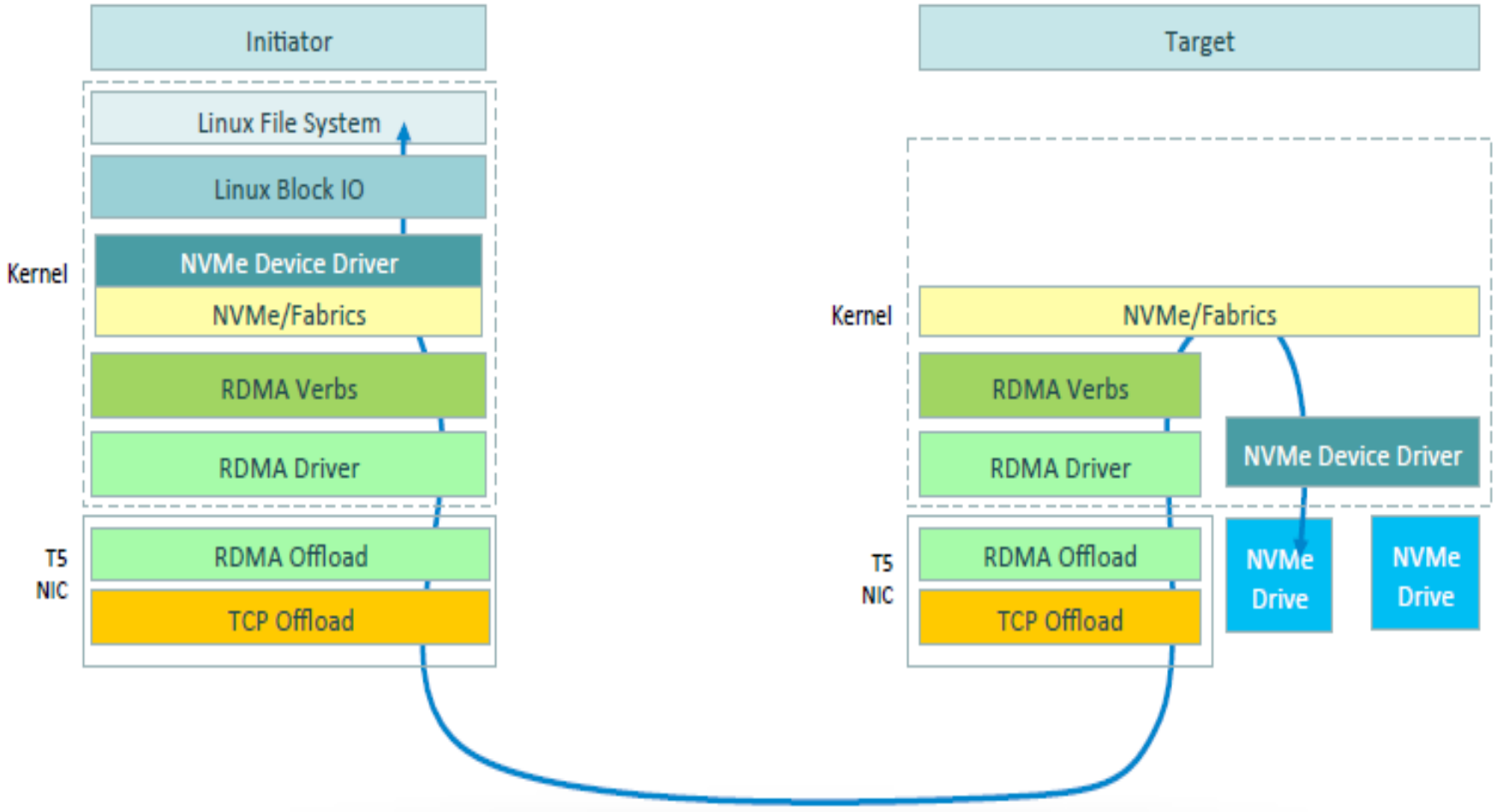
Differences	PCI Express	NVMe over Fabrics
Identifier	Bus/Device/Function	NVMe Qualified Name
Discovery	Bus Enumeration	Discovery and Connect commands
Queueing	Memory-based	Message-based
Data Transfers	PRPs or SGLs	SGLs only, added Key

iWARP RDMA over Ethernet

- Open, transparent and mature
  - IETF Standard(2007)
- Plug-and-play
  - Standard and proven TCP/IP
  - Built-in reliability congestion control,flow control
  - Natively routable
- Cost effective
  - Regular switches
  - Same network appliances
  - Works with DCB but NOT required
  - No need for lossless configuration
- No network restrictions
  - Architecture, scale, distance, RTT, link speeds
- Hardware performance
  - Exception processing in HW
  - Ultra low latency
  - High packet rate and bandwith



NVMe/Fabrics Layering



四、SSD评价指标

4.1 性能

- 对同一个SSD，采用不同的测试方法所表现的性能不一样。影响测试性能的因素很多，包括：读写比例，请求的数据块大小、测试时使用的是新SSD还是旧SSD、测试过程中是否调用了GC操作等等。
- 因此，在提供产品的性能指标时，应该有一系列的测试前提，如：读写比例（R/W：75/25, 50/50）;请求块大小（2KB、128KB）；测试 过程中是否调用过GC操作；保留空间是多少（20%）等。

四、SSD评价指标

4.2 寿命标准

- SSD的擦除次数是有限的，因此如何评价SSD的寿命很重要。
- JEDEC固态技术协会，全球微电子产业标准的领导制定机构，发布了两项新的固态硬盘标准：“JESD218 固态硬盘（SSD）要求与耐用性测试方法”及“JESD219固态硬盘耐用性工作负荷”。

JESD218和JESD219标准

- 耐用性分级与验证
  - JESD218标准还建立了一个固态硬盘耐用性评级，代表主机写入固态硬盘的兆兆字节数（TBW），做为按应用分类进行固态硬盘比较的标准。最终用户可以用耐用性评级标准来比较不同厂家的固态硬盘产品。 此外，该标准还确定了两种方法验证耐用性与数据保留 – 直接验证法与外推法。
- 工作负荷
  - 由于随着应用的发展工作负荷会随之改变，工作负荷的定义包含在一个单独的补充性标准中 -“JESD219 固态硬盘耐用性工作负荷”。 由于固态硬盘的工作负荷对需要写入硬盘的数据量有着实质性影响，因此需要确定一个标准工作负荷才能比较结果。 现在JESD219标准只规定了企业级的工作负荷；客户级在近期会增加进来。

四、SSD评价指标

4.3 SSD能耗

- 产品标称上的功率不一定能够反映SSD真实的能耗。因为不同的SSD的内部结构可能有所差别，而且智能的功耗管理系统在SSD实际运行时会对能耗有影响。
- 因此，能反映能耗的指标是：完成相同的IO访问请求，所消耗的总能量，或者是单位能耗所能完成的IO访问数。

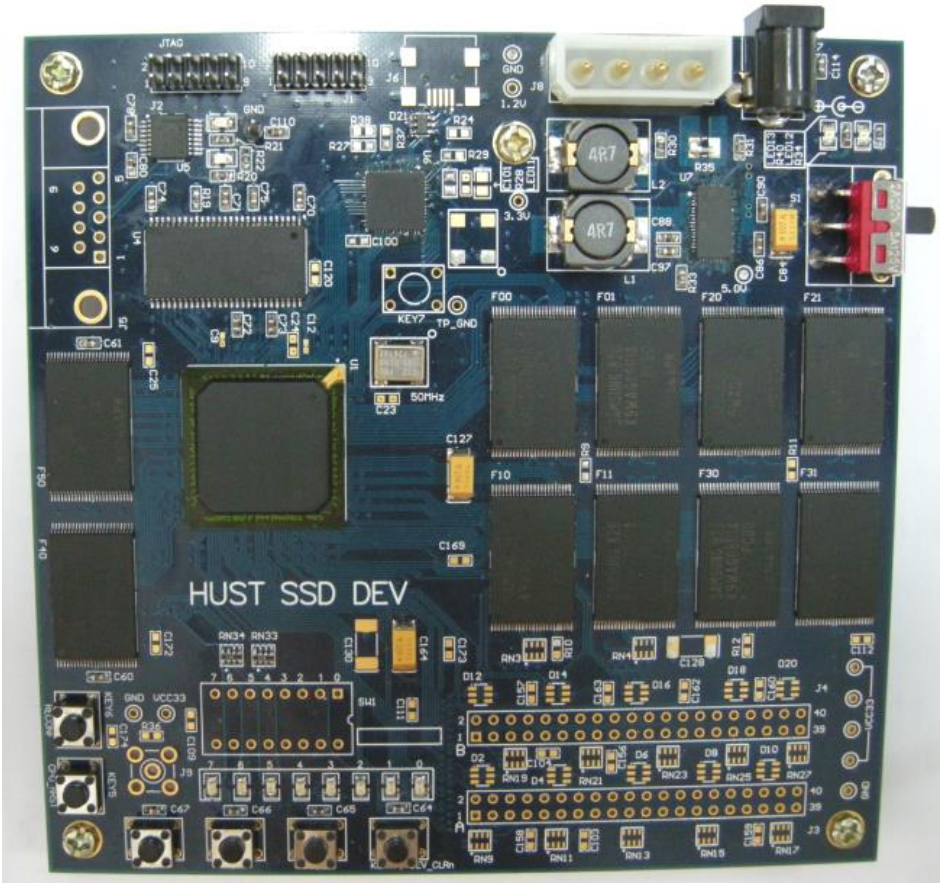
四、SSD评价指标

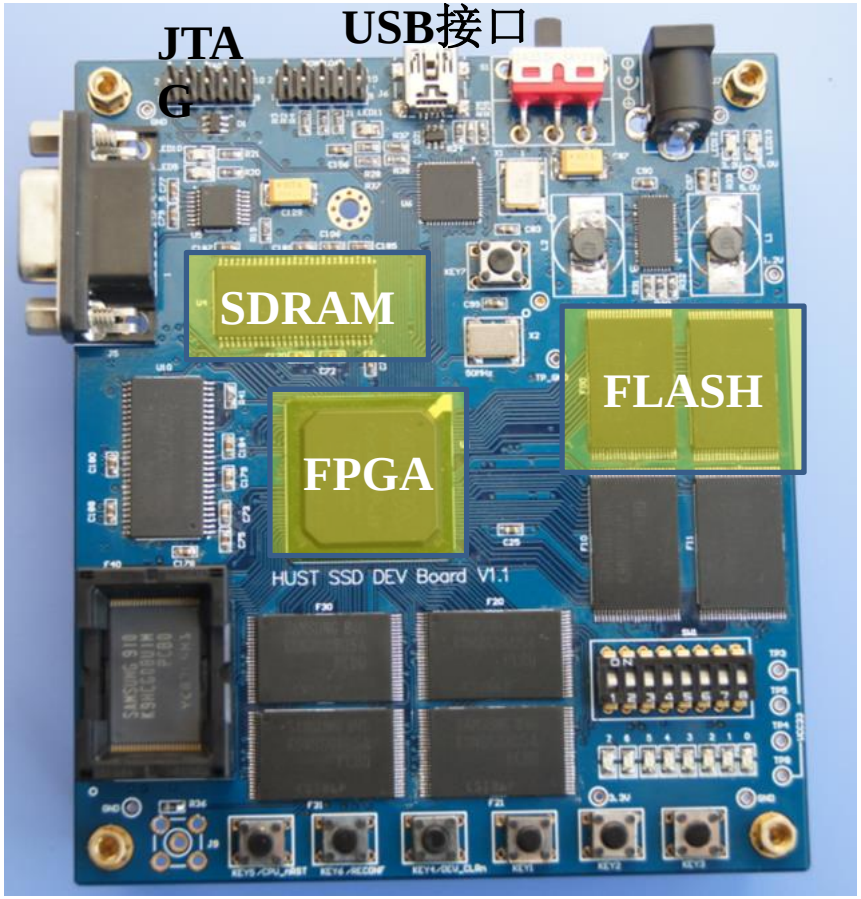
Latency	*50-10,000 ns*
Bandwidth	*>100 MB/s*
Endurance	*$>10^8$ cycles*
Hard error rate	*10^{-4} bits/terabyte*
Meantime between failure	*2 million hours*
Data retention	*Months-10 years*
Active power	*100 mW*
Standby power	*<1 mW*
Cost	*<$5/GB*

五、关键技术研究

- 相继开发了两款SSD原型系统
  - USB接口SSD原型系统的开发
  - 开发了自主知识产权的闪存控制器IP核
  - PCIe接口SSD原型系统的开发研究
- 开发了一套SSD模拟测试开发平台SSDsim

5.1 USB接口SSD原型系统





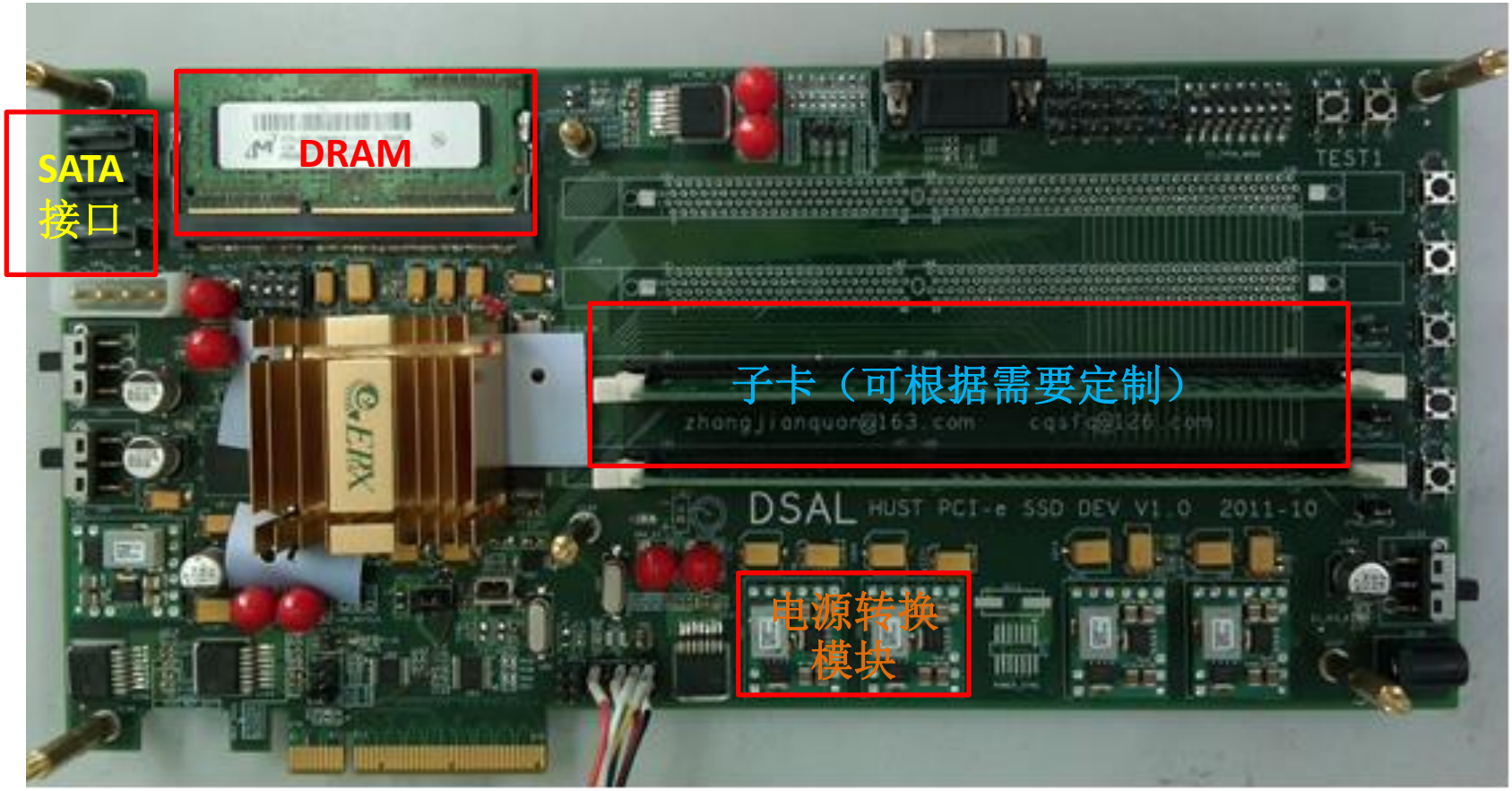
USB原型系统示意图

（左侧为1.0版本，右侧为1.1版本）

5.1 USB接口SSD原型系统

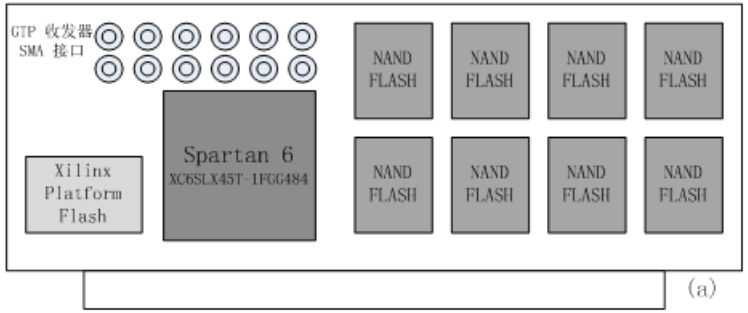
- 本系统以Cyclone II FPGA为核心，采用四通道并行架构，每通道相互独立
- 在FPGA中实现了具有自主知识产权的Nand Flash闪存控制器IP核
- 由于受USB接口限制，单通道读速度达10MB/s以上，写速度2MB/s以上
- 在此基础上开发PCIe接口SSD硬件原型系统

5.2 PCIe接口SSD原型系统



结构示意图

主卡模块结构示意图

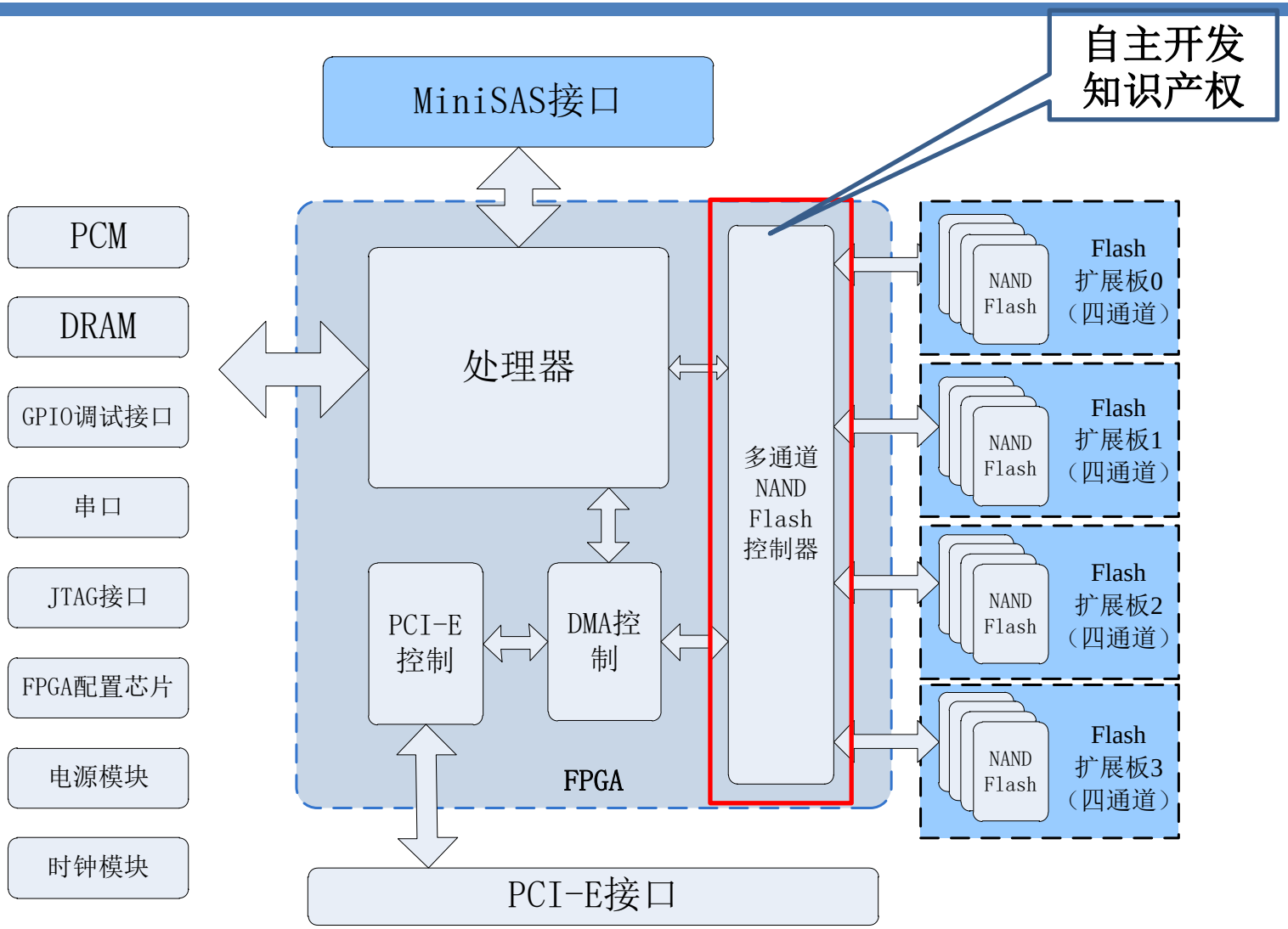


子卡模块结构示意图

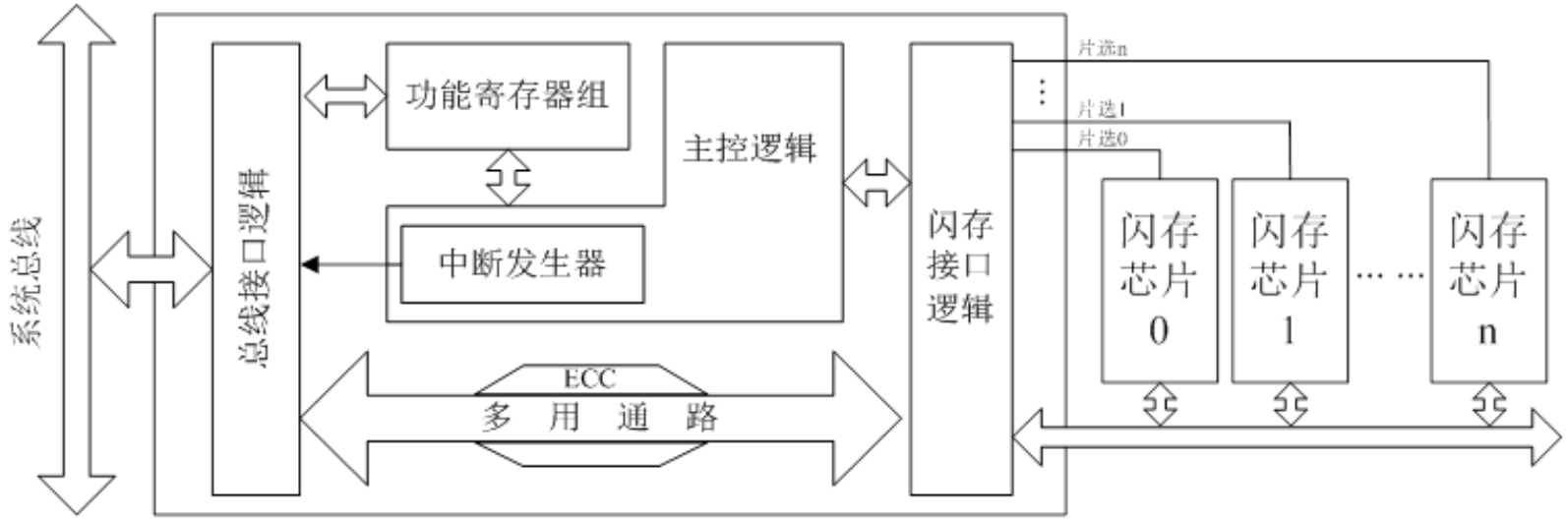
PCIe接口SSD原型系统硬件平台

- 该硬件平台以子母卡的形式存在，其中主卡以Xilinx公司的Virtex 6 FPGA芯片为核心，负责与主机间的通信与数据传输
- 子卡以Xilinx公司的Spartan 6芯片为核心，其上贴装闪存芯片，负责控制闪存的读写操作，以及与主卡之间的通信等
- 该原型系统与主机端的通信采用NVMe协议，紧跟行业标准与发展趋势

PCIe接口SSD体系结构



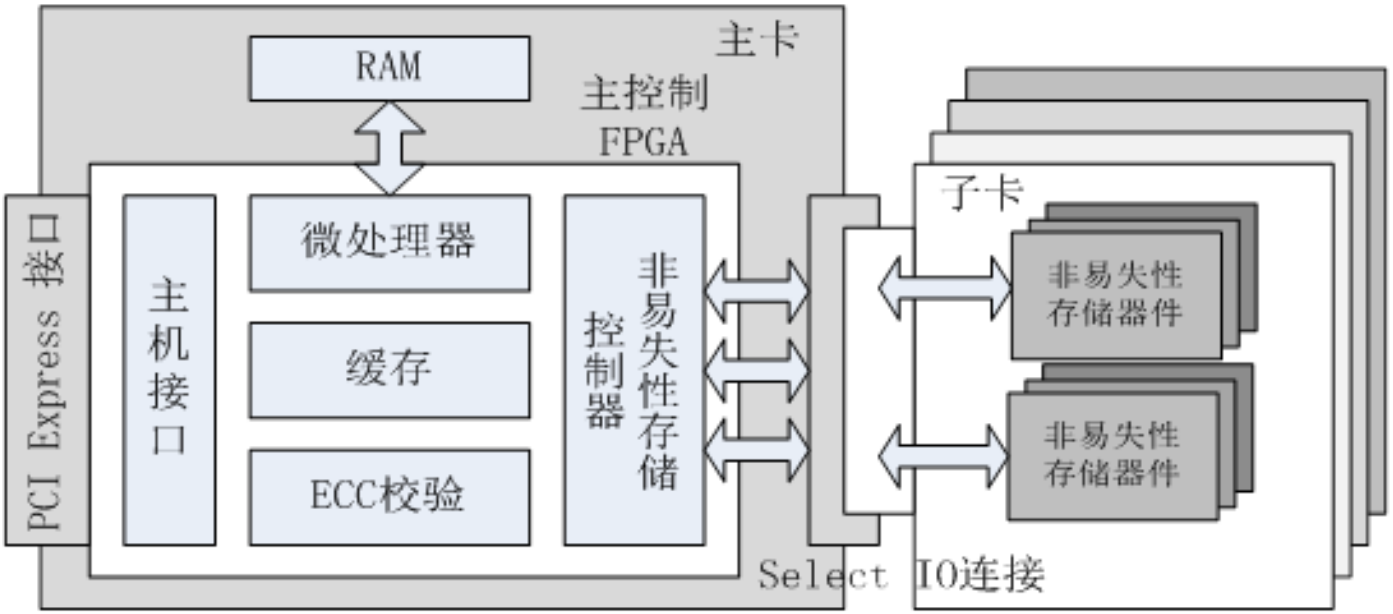
多通道闪存控制器架构


系统总线
总线接口逻辑
功能寄存器组
主控逻辑
中断发生器
ECC
多 用 通 路
闪存接口逻辑
片选n
片选1
片选0
闪存芯片0
闪存芯片1
闪存芯片n

PCIe SSD硬件平台扩展研究

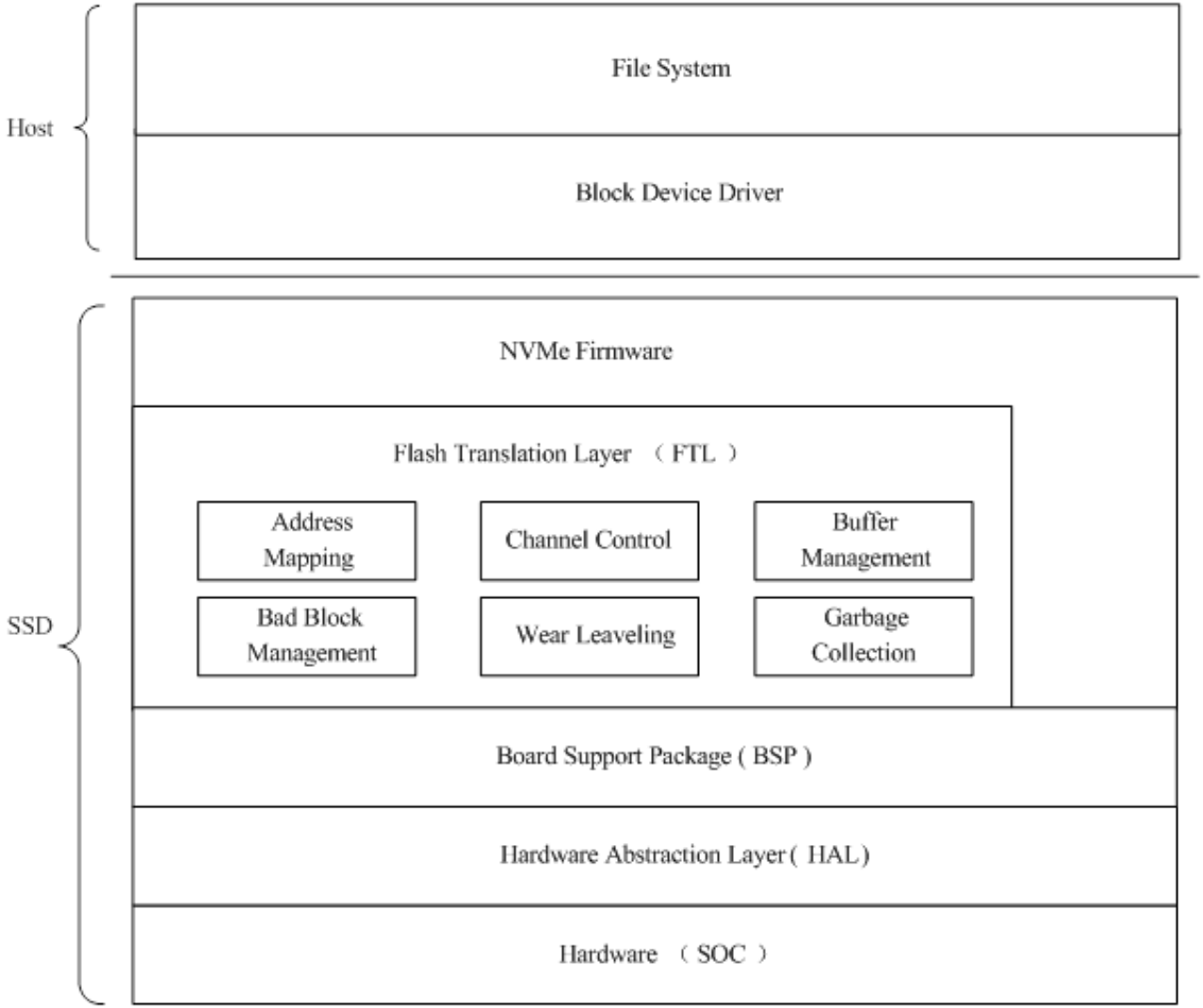
- 该研究平台可以通过变换不同的子卡进行不同的研究
  - 将子卡上的闪存芯片替换为PCM相变存储器芯片开展相变存储器相关研究
  - 将基于闪存的子卡与基于PCM的子卡同时插入主卡，构建混合SSD存储系统
  - 通过主卡上的SATA接口外接硬盘构建混合存储系统

相变存储器研究



在PCIe SSD的原型系统中，将载有闪存的子卡替换为PCM子卡，并将FPGA中的闪存控制器替换为相变存储控制器。用来进行相变存储器研究以及相变存储器与闪存的混合固态存储的研究。

5.3 NVMe –SSD系统软件架构



5.4 固态盘的高性能闪存转换层研究

1. 设计前提
2. 隐藏翻译过程映射算法核心思想
3. 系统测试

设计前提

- 闪存转换层分三种类型：页级映射，块级映射，混合映射。

	性能	寿命	映射表大小	所需内存大小	成本
页级映射	好	长	大	大	高
块级映射	差	短	小	小	低
混合映射	较差	较短	较小	较小	较低

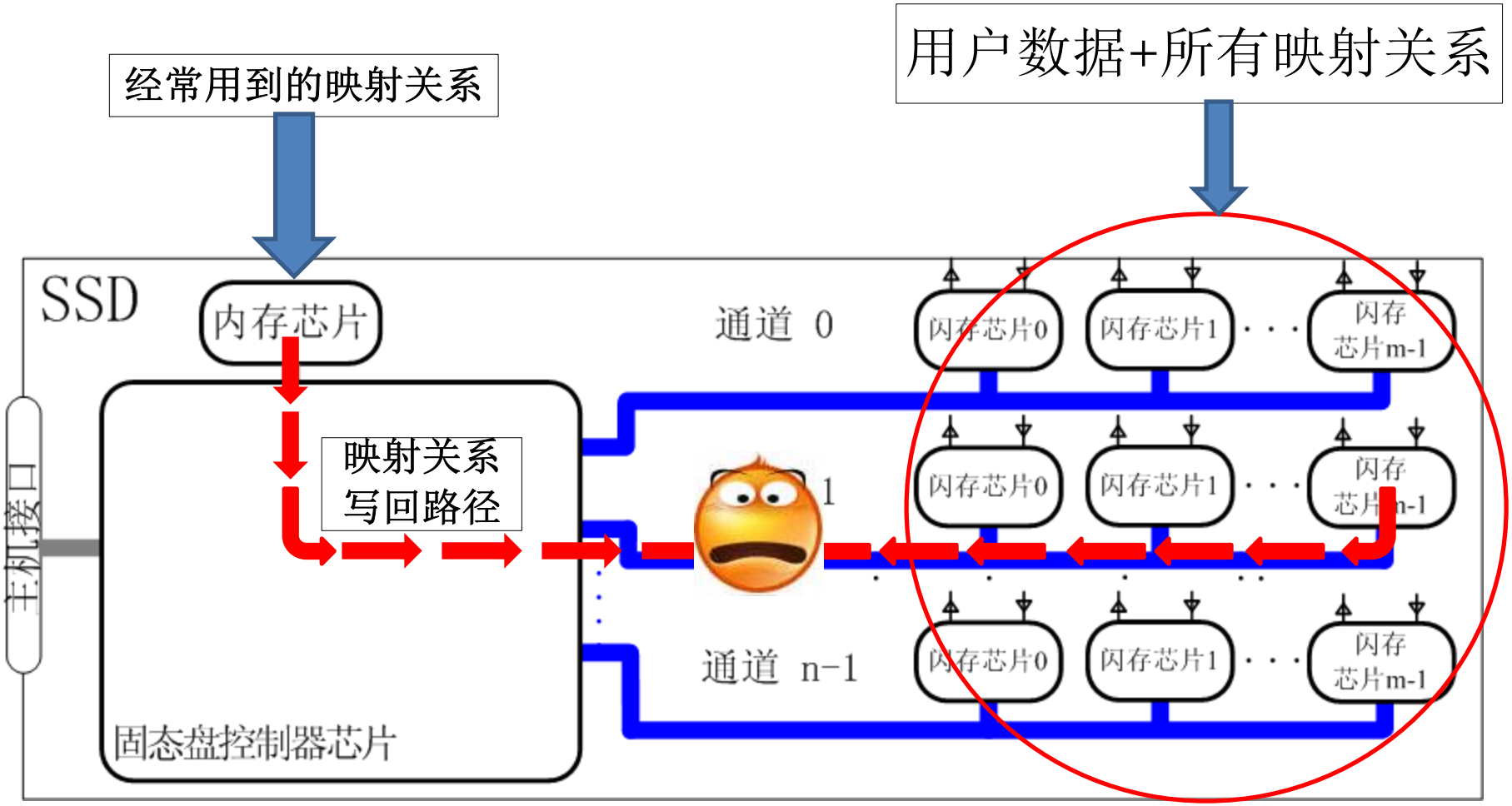
- 页级映射算法的性能最佳，因此在高性能的固态盘设计中，大多采用或者基于页级映射。为了减少映射表大小，DFTL被提出来了。DFTL是基于页级映射 的映射算法，是目前性能、寿命、成本综合最优的闪存转换层算法。

设计前提

- **DFTL是基于负载的局部性原理，将经常访问的数据的映射关系存放在内存中，通过这种方式减少映射关系占用内存的容量。DFTL依赖于局部性，当负载的局部性下降，将导致系统性能急剧下降。**

负载特征	网页搜索	金融2	金融1	邮件服务器
局部性	2.5%	76.7%	65.9%	47.1%
读操作比例	99.0%	82.3%	23.2%	45.4%
请求间隔时间	3.0毫秒	11.1毫秒	8.2毫秒	<1毫秒(96.0%)
性能损失	18.8%	8.3%	21.8%	57.1%

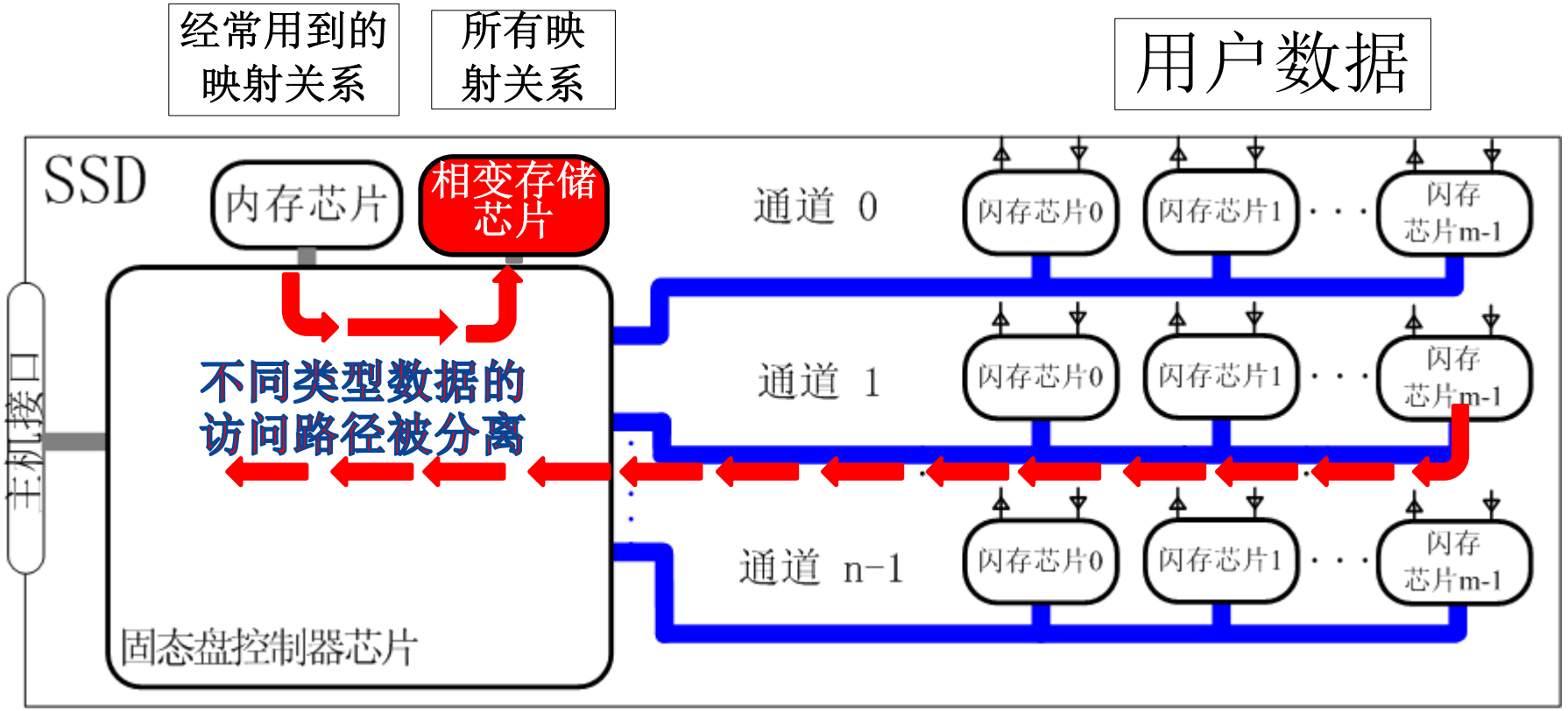
DFTL出现性能下降的根本原因



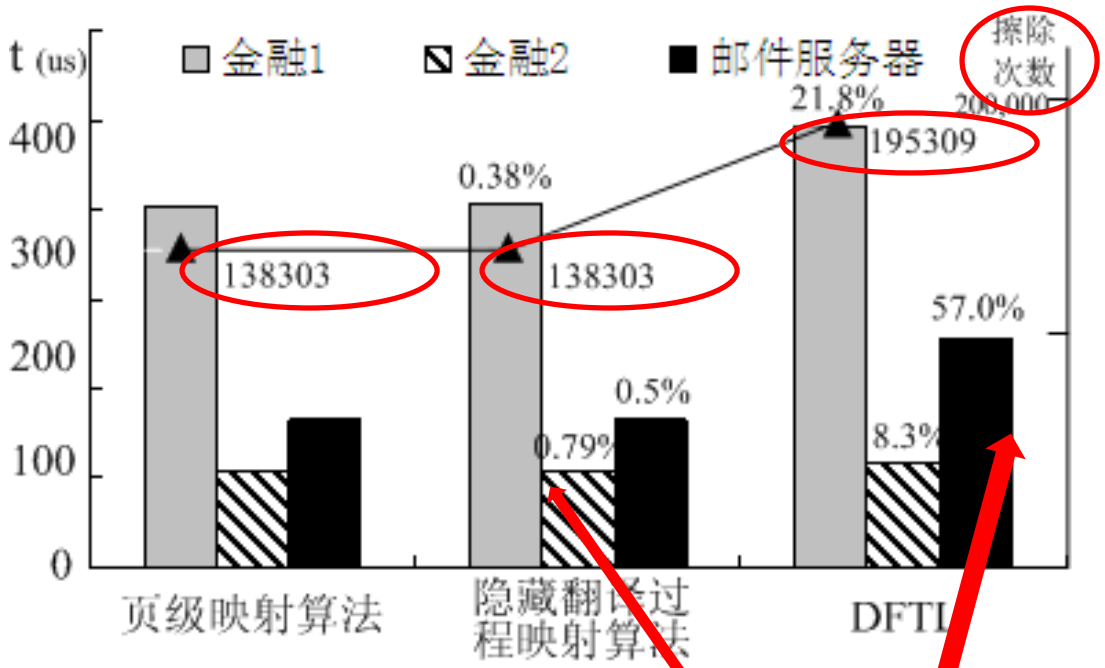
- 映射关系数据的访问路径与用户数据的访问路径是同一条路径，因此产生冲突。

隐藏翻译过程映射算法的核心思想

- 将映射关系数据的访问路径与用户数据的访问路径进行分离。



系统测试

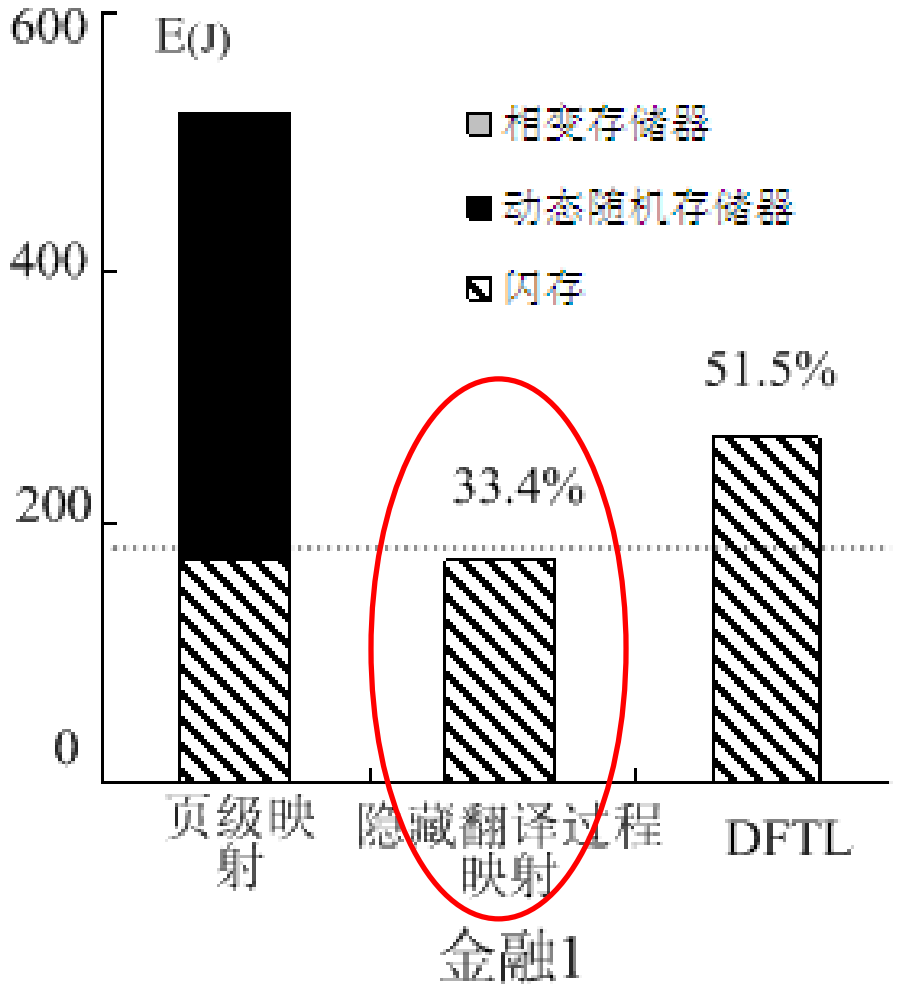


与页级映射相比的性能差距

映射算法	最小性能差距	最大性能差距
隐藏翻译过程映射算法	0.38%	0.79%
DFTL	8.3%	57.0%

隐藏翻译过程映射算法的性能、寿命基本接近或等于页级映射算法

系统测试



相比页级映射、DFTL，隐藏翻译过程映射算法的能耗最低。

实验结论：隐藏翻译过程映射算法提供了优异的性能、寿命和能耗结果。

5.5 面向3D闪存的性能优化

背景

闪存芯片容量随着工艺进步成倍增长，闪存块容量也随之成倍增长。

- 介质层：闪存页大小从2KB持续增长到16KB，甚至更大；
- 文件系统层：文件系统块大小仍然以4KB为主；

- 优势
  - 降低存储成本
  - 更大的吞吐量
- 劣势
  - 上下层存储单元大小不匹配
  - 存储空间与传输时间浪费

Year	2006	2007	2008	2010	2014	2017	2019
Process/layers	90nm	72nm	50nm	25nm	16nm	64层	128层
Block Size	0.125MB	0.25MB	0.5MB	2MB	6MB	12MB	24MB
Page Size	2KB	2KB	4KB	8KB	8KB	16KB	16KB+

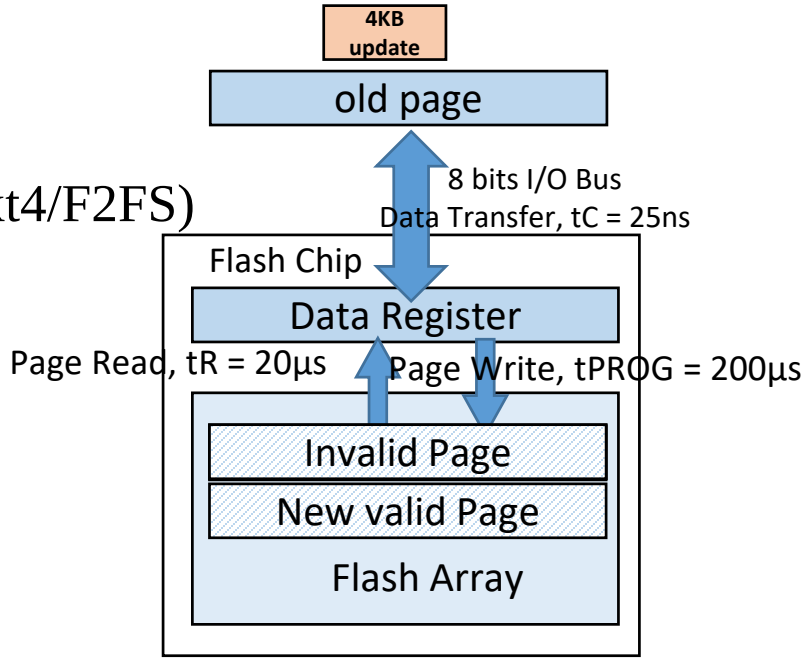
Evolution of High-Density Micron Flash Devices

5.5 面向3D闪存的性能优化

背景

上下层存储单元大小不匹配问题：

- 闪存页(page) VS 文件系统块(FSB, e.g. Ext4/F2FS)
  - Flash page (16KB) vs FSB/cluster (4KB)
  - 性能下降
  - 存储空间浪费



- **传输放大：Transfer Amplification = transferred data size / request data size ×100%**

  = (16KB + 16KB) / 4KB ×100% = 800%

- **写放大：Write amplification = write data size / request data size ×100%**

  = 16KB / 4KB ×100% = 400%

5.5 面向3D闪存的性能优化

现有工作

上下层存储单元大小不匹配问题：

- Sector-log[SAC'11]利用缓存进行小写合并，合并成闪存页大小后写入闪存中，需要额外子页映射表存储这部分元数据信息，并且引入额外数据整理开销；
- BPLRU[FAST'08]优化传统LRU算法，以闪存块为粒度将数据刷回介质，并且在刷回时进行页合并，最后采用LRU补偿方案提升顺序写入时LRU的效率；
- ……

- 现有工作主要从优化缓存的角度出发，但是由于负载的时间、空间局部性不同，设备DRAM容量限制，引入的额外管理和数据迁移开销等原因，优化的程度是受限的
- 我们基于闪存重复编程特性来解决不匹配的问题，避免额外写放大和对缓存大小的依赖

5.5 面向3D闪存的性能优化

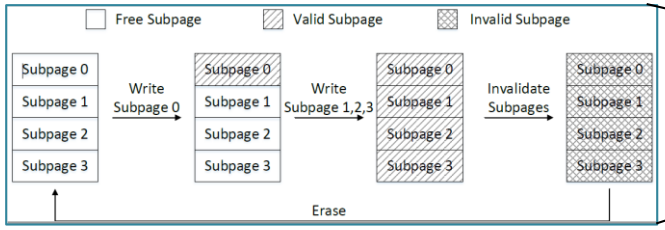
闪存重复编程特性分析

- MLC闪存的重复编程方法：
  - 利用闪存中的双模特性，闪存芯片可以在默认模式（MLC/TLC模式）和SLC模式之间进行切换；
  - ✓ SLC模式很好的支持单元状态单向从“1” → “0”的转化
  - ✓ SLC模式绕开了随机化模块，用户写入的数据即是存储到阵列的数据
  - ✓ SLC模式有性能和可靠性的优势
  - × 双模式转换不当会导致数据崩溃
  - × 带来数据碎片化的问题，降低读性能

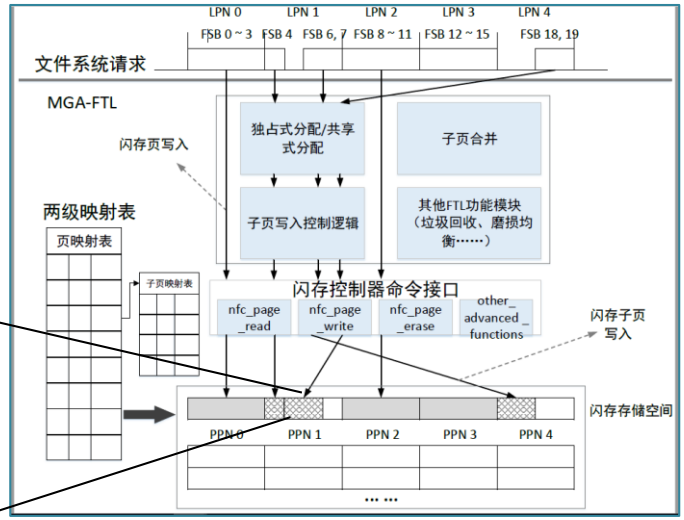
5.5 面向3D闪存的性能优化

面向SLC闪存的映射粒度自适应FTL（MGA-FTL）技术

- 核心思想
  - 采用更细粒度的子页响应小写请求
  - 擦除之前实施多次子页写
- FTL模块
  - 两级映射表
  - 闪存页状态转换
  - 分配策略



Multiple subpage writes

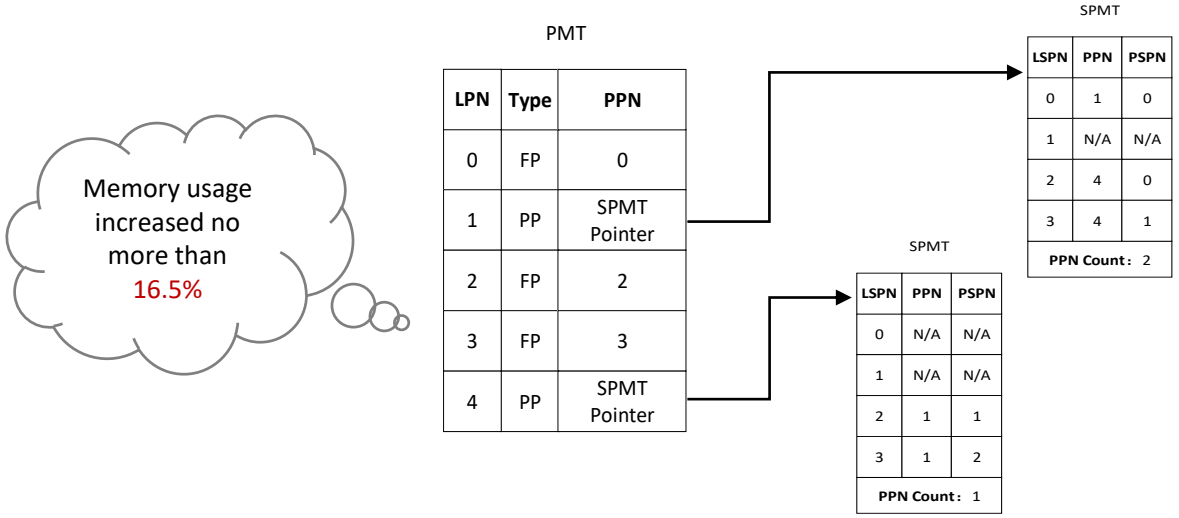


The architecture of MGA-FTL

5.5 面向3D闪存的性能优化

两级映射表

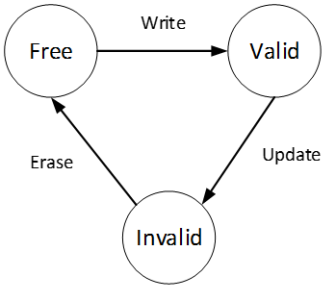
- 采用页级和子页级混合的两级映射表方案
  - 只使用子页级映射表会成倍增加映射表占用DRAM的开销， (1 TB SSD → 2GB mapping tables)
  - 初始只创建页映射表，只在进行了子页小写时，才创建对应子页表项



5.5 面向3D闪存的性能优化

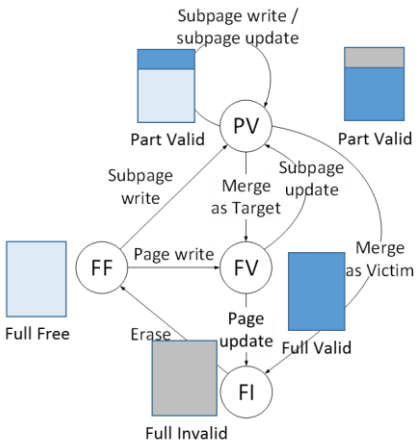
闪存页状态转换

- 已有的三种闪存页状态 (free/valid/invalid) 不能表示一个闪存页中部分子页被写入的情况
  - ✓ 增加一个新的页状态：部分有效(Partially Valid, PV)
  - ✓ 原先3种状态随之变化为：完全空闲(Fully free, FF), 完全有效(Fully valid, FV), 完全无效(fully invalid, FI) 以及新添加的 部分有效PV



The original state machine

扩展状态转换机



5.5 面向3D闪存的性能优化

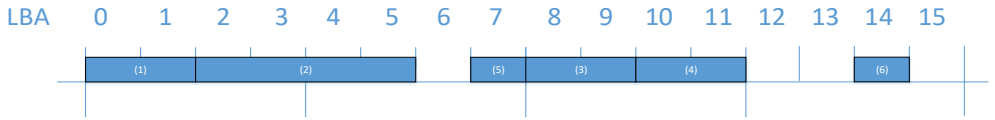
分配策略

- 主要思想
  - 使用子页响应小写请求
    - 冗余数据传输减少
    - 一个物理页中的数据并不是逻辑连续的
- 分配策略
  - 独占式分配, 一个物理页只能存储属于相同逻辑页的数据，更少的数据碎片化
  - 共享式分配, 一个物理页可以存储不同逻辑页的数据，获得更高的空间利用率

5.5 面向3D闪存的性能优化

分配策略

- **举例：**



文件系统请求 (LBA, SIZE, R/W)	写请求 Page cache op	闪存物理写操作 (LPN, <State, [LSPN]>, R/W)
① (0, 2, W)	① (0, 4, W)	① (0, Full, W)
② (2, 4, W)	② (4, 2, W)	② (1, <Part, 0, 1>, W)
③ (8, 2, W)	③ (8, 4, W)	③ (2, Full, W)
④ (10, 2, W)	④ (7, 1, W)	④ (1, <Part, 3>, W)
⑤ (7, 1, W)	⑤ (14, 1, W)	⑤ (3, <Part, 2>, W)
⑥ (14, 1, W)		

LPN = LBA / subpage_num
LSPN = LBA % subpage_num

LPN	State	PPN
0	Full	0
1	Part	Point
2	Full	2
3	Part	Point

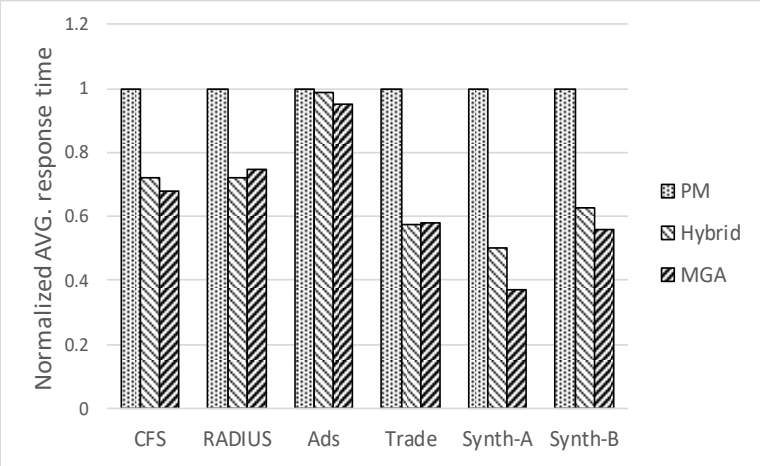
Exclusive Allocation
Shared Allocation

LSPN	PPN	PSPN
0	1	0
1	1	1
2	1	3
3	1	2

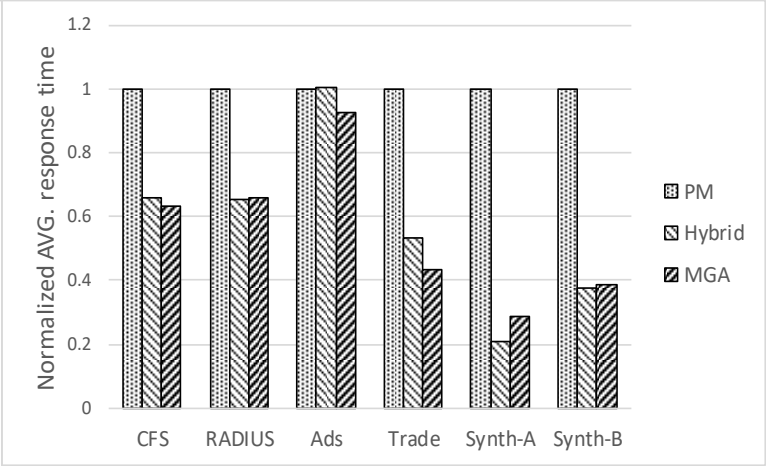
LSPN	PPN	PSPN
0	3	0
1	N/A	N/A
2	N/A	N/A
3	N/A	N/A

5.5 面向3D闪存的性能优化

实验评估



(a) 8KB Page Avg. Response Time

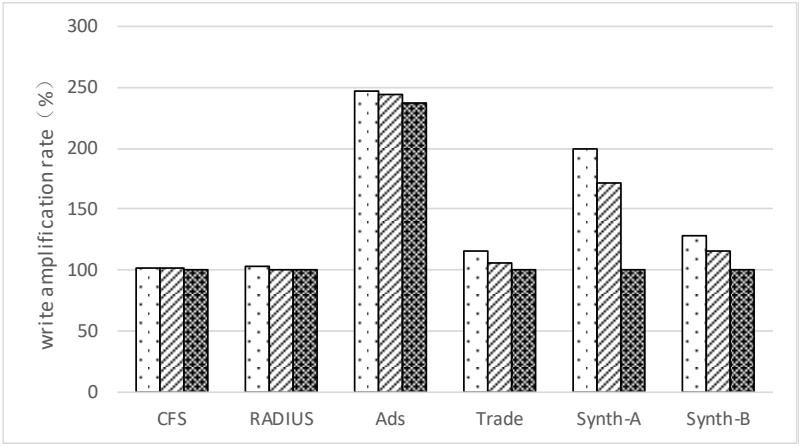


(b) 16KB Page Avg. Response Time

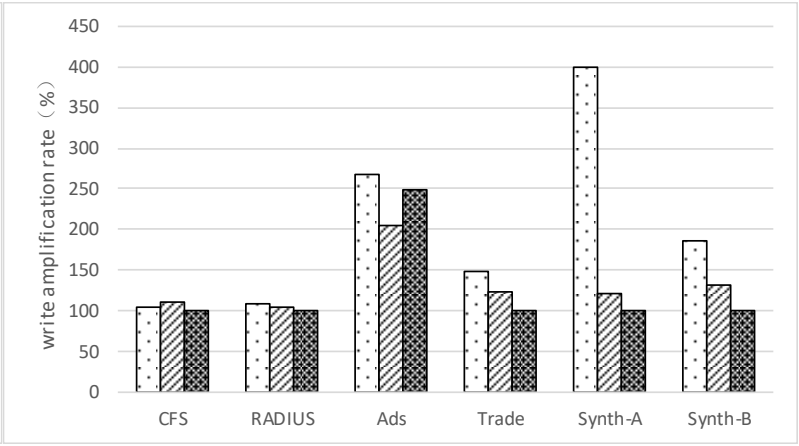
- 真实负载:
  - 降低 **8%-53%** 平均响应时间
- 合成负载:
  - 降低 **72%**平均响应时间
  - 小请求为主的负载效果更佳显著

5.5 面向3D闪存的性能优化

实验评估



(a) 8KB Page Write Amplification



(b) 16KB Page Write Amplification

- 对于**8KB**大小闪存页，减少了 **7.42%** 的写放大
- 对于**16KB**大小闪存页，减少了**30.83%** 的写放大

Y. Feng, D Feng et al., “Mapping granularity adaptive FTL based on flash page re-program” in Proc. DATE, 2017.

5.6 固态盘模拟器SSDsim的设计实现

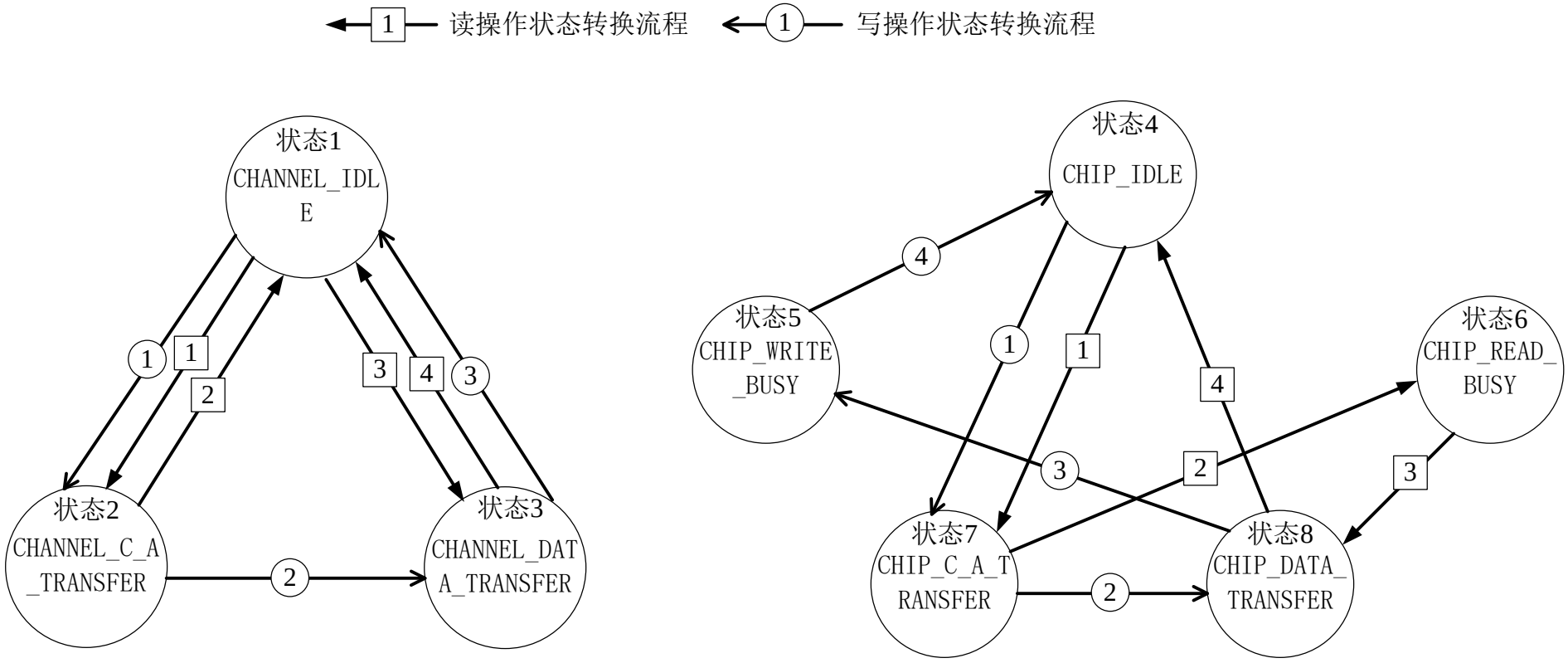
SSDsim是一个固态盘模拟器

针对现有开源固态盘模拟器的缺陷，SSDsim增加了以下功能：

1. **数据缓存区的模拟**
2. **能耗结果的模拟**
3. **闪存高级命令的模拟**

固态盘模拟器SSDsim的设计实现

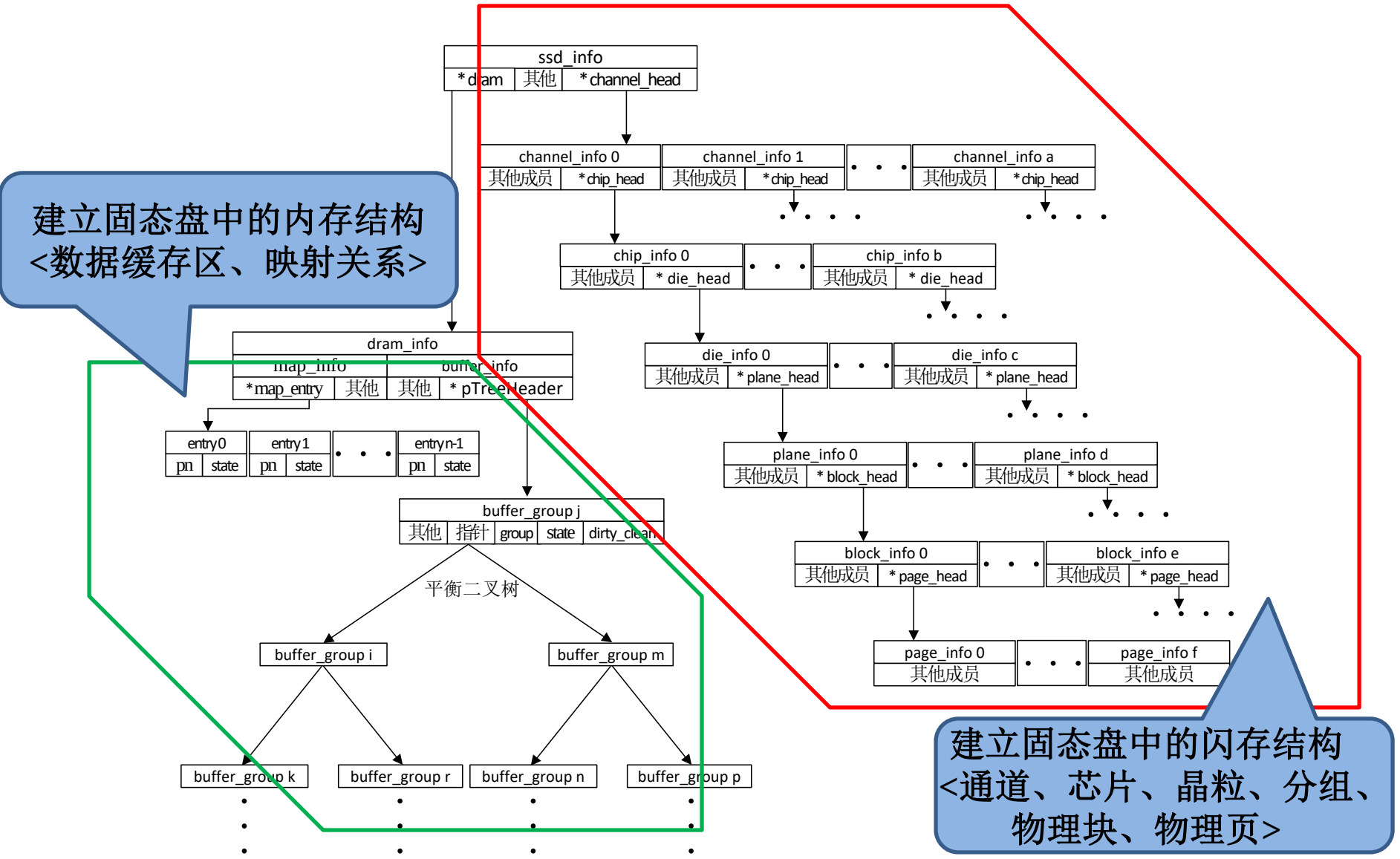
- **SSDsim是基于事件驱动的固态盘模拟器**



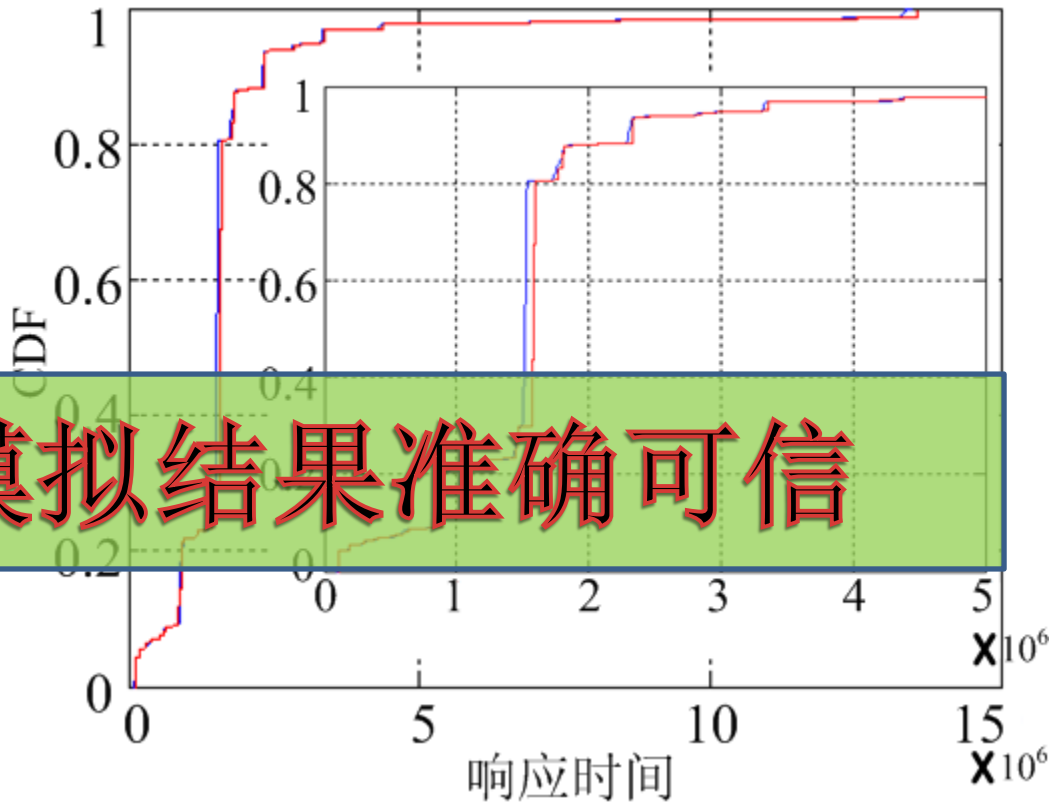
固态盘通道状态变化

闪存芯片状态变化

固态盘模拟器SSDsim的设计实现



固态盘模拟器SSDsim的验证



SSDsim的模拟结果准确可信

用一个负载在真实系统上测试得到的响应时间结果，与同一个负载在SSDsim上模拟得到的响应时间结果进行对比

固态盘模拟器SSDsim

- **SSDsim是一款事件驱动、模块化、可配置、高准确性的固态盘模拟器，为固态盘的研究提供了一个方便快捷的测试工具。**
- **目前SSDsim已经作为开源工具，可以从网上自由下载，网址为：**

  http://storage.hust.edu.cn/SSDsim/