清華大學 Tsinghua University

大数据时代的存储系统若干变化的思考

舒继武

清华大学计算机系

shujw@tsinghua.edu.cn

http://storage.cs.tsinghua.edu.cn/~jiwu-shu/

主要内容

一 背景

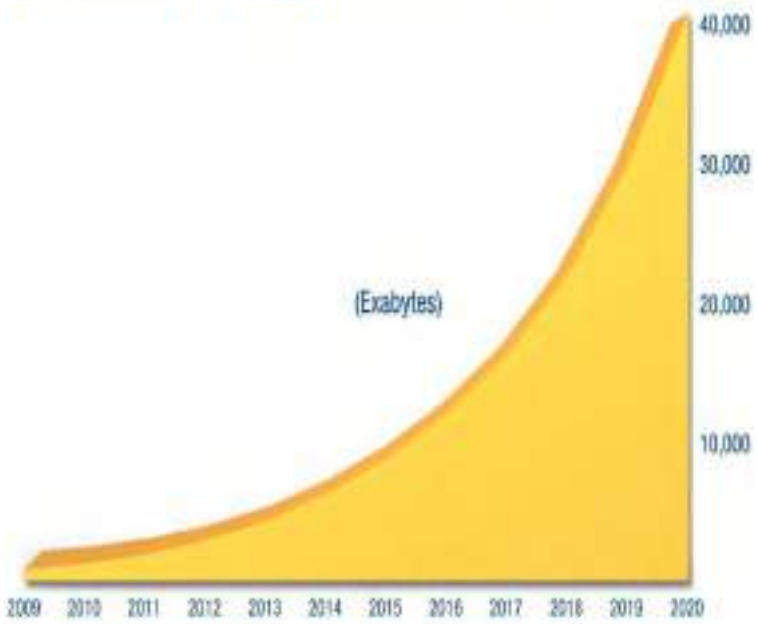
二 面向大数据的闪存存储系统

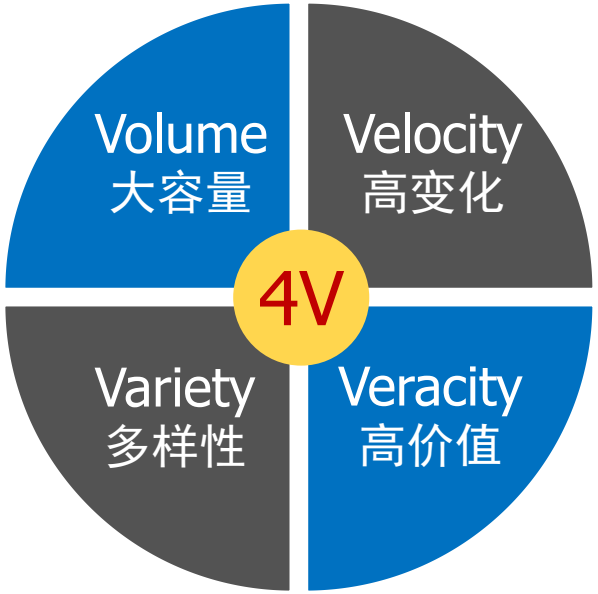
三 面向大数据的持久性内存存储系统

四 新型分布式存储系统

五 展望

1. 背景—大数据时代的数据特征





数据量呈现爆发性增长趋势

大数据，不仅是数量大

4V理论

Source: IDC 2012, IDC 2014.

1. 背景—存储系统面临严峻挑战

计算密集型负载 → 数据密集型负载：
如何支持越来越高的数据存储和处理需求?

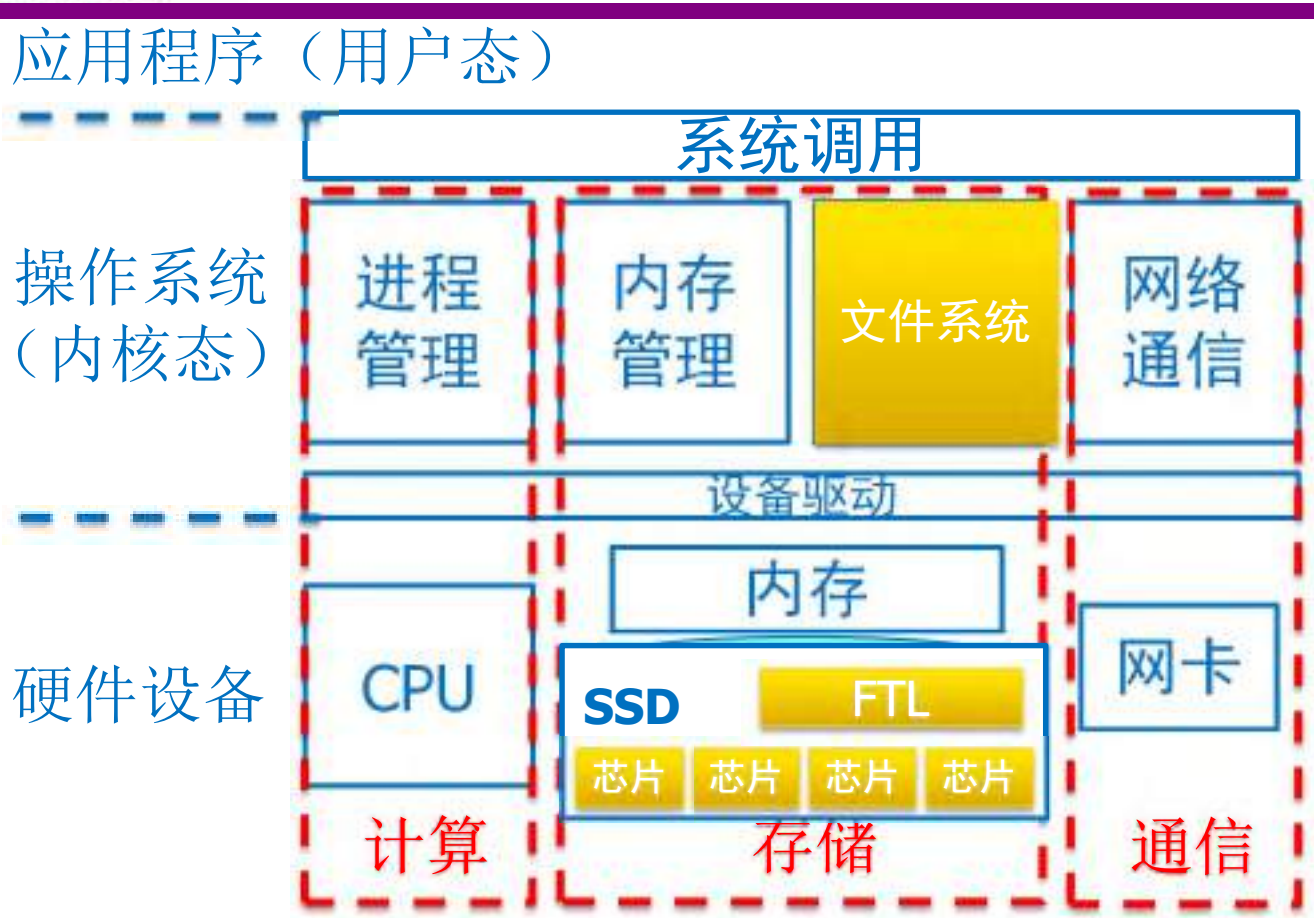
如何构建面向大数据的高效率存储系统？

固态盘和持久性内存等应用越来越广：
如何高效地发挥这些新型高速存储介质的优势？

固态盘

持久性内存

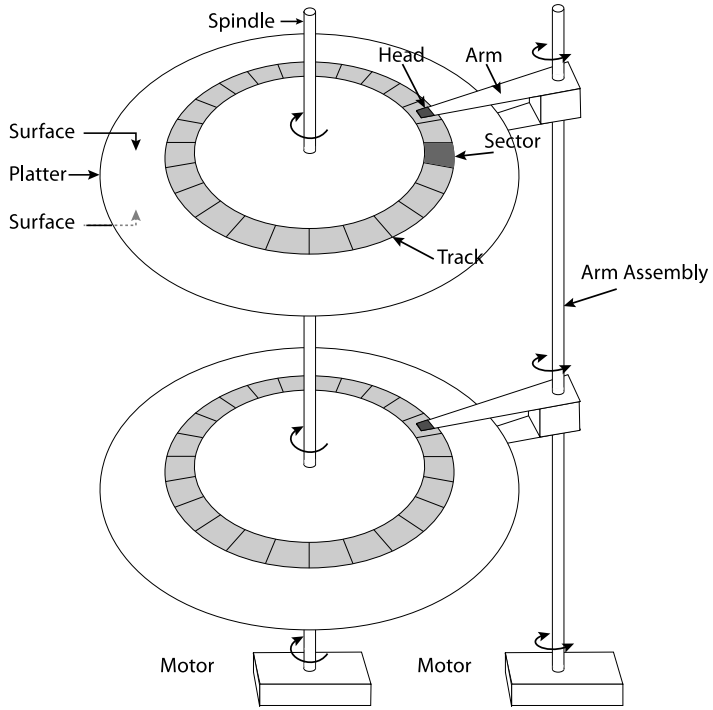
1. 背景—磁盘的局限（1）

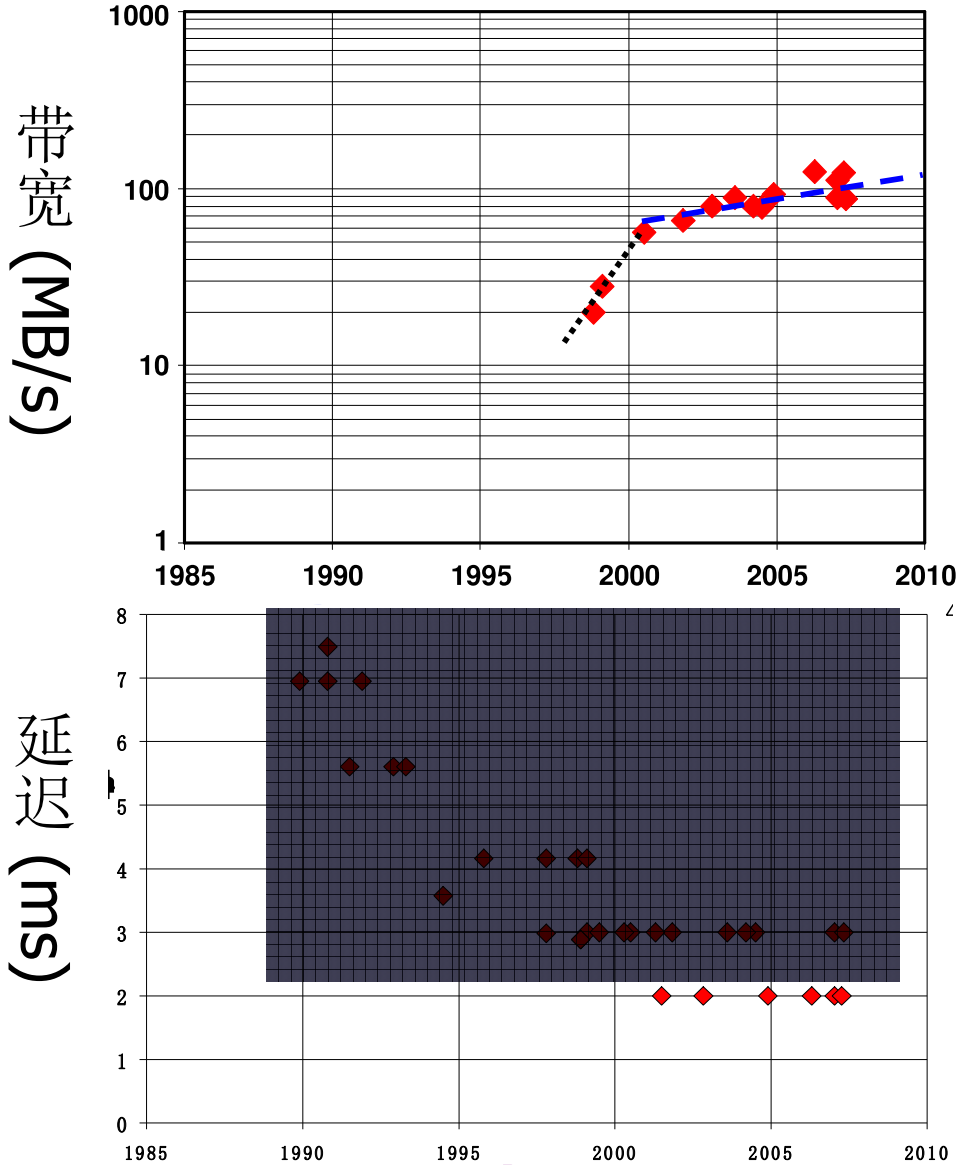


- 在计算机系统的三大基石（计算、存储、通信）中，唯有存储仍包含机械式部件操作
- 电子式存储器件—闪存/持久性内存—是计算机发展的趋势

1. 背景—磁盘的局限（2）

- 机械式外存 — 磁盘
- 性能局限性
  - ✓ 带宽、延迟





图片来源：FAST'10 Tutorial

1. 背景—磁盘的局限（3）

- 能耗与体积局限性

Extrapolation to 2020

(at 70% CGR → need **2 GIOP/sec**)

磁盘存储：
5,000,000 块
超过 2,000 平方米
能耗 22 兆瓦

图片来源：Storage-class memory: the next storage system technology, IBM Journal

1. 背景—固态盘（1）

- 读写性能：带宽与延迟

类型	设备型号	读带宽	写带宽	读延迟	写延迟
磁盘	Seagate Savvio	202	202	2.000	2.000
SATA 固态盘	Intel X25-E	250	170	0.075	0.085
PCIe 固态盘	FusionIO ioDrive Octal	6,000	4,400	0.030	0.030
		30X	20X	1/100	1/100

- 体积与能耗

磁盘存储：
5,000,000 块
超过 2,000 平方米
能耗 22 兆瓦

1/1000
1/10000

SCM存储：
8,000 个模块
1+平方米
能耗 1千瓦

主要内容

一 背景

二 面向大数据的闪存存储系统

- 2.1 存储结构的变革
- 2.2 系统软件的变革

三 面向大数据的持久性内存存储系统

四 新型分布式存储系统

五 展望

2. 用闪存构建大数据存储的优势（1）

- 以目前广泛应用的闪存为例
  - ✓ 低延迟、高带宽、随机读写性能高：

设备	随机读	随机写	读带宽	写带宽
Fusion-IO ioDrive2	480K IOPS	490K IOPS	3GB/s	2.5GB/s
Samsung 840	100K IOPS	78K IOPS	540MB/s	450MB/s
Seagate Barracuda	52 IOPS	47 IOPS	156MB/s	156MB/s

  - ✓ 成本“优势”

	容量	RMB/GB	RMB/IOPS
Samsung 840	256 GB	5.86	0.015
Seagate Barracuda	3 TB	0.26	50

2. 用闪存构建大数据存储的优势（2）

- 以目前广泛应用的闪存为例

✓ 可靠性高：

HDD MTTF = 500K Hours.
SSD MTTF = 2M Hours.

source : SanDisk SSD: A More Reliable Alternative to the Laptop HDD， 2007

✓ 低能耗： [SOSP'09 FAWN: A Fast Array of Wimpy Nodes. CMU]

System	Watts	QPS	Queries /Joule
2TB HDD	20	250	12.5
32GB SSD	15	35K	2.3K

2. 用闪存构建大数据存储的优势（3）

- 用于大数据存储的闪存与HDD的综合对比

	DRAM	PCM	NAND Flash	HDD
Read Energy	0.8J/GB	1J/GB	1.5J/GB	65J/GB
Write Energy	1.2J/GB	6J/GB	17.5J/GB	65J/GB
Idle Power	~ 100mW/GB	~ 1mW/GB	1-10mW/GB	~10W/TB
Endurance	∞	10^6 - 10^8	10^4 - 10^5	∞
Page Size	64B	64B	4KB	512B
Read Latency	20-50ns (64B) ~ 1-3us (4kB)	~ 50ns (64B) ~ 3us (4KB)	~25us (4KB)	~ 5ms (512B) ~ 40ms (4KB)
Write Latency	20-50ns (64B) ~ 1-3us (4kB)	~ 1us (64B) ~ 64us (4KB)	~ 500us (4KB)	~ 5ms (512B) ~ 40ms (4KB)
Erase Latency	N/A	N/A	~ 2ms	N/A

Source: “Rethinking Database Algorithms for Phase Change Memory”, CIDR 2011, Intel

2. 用闪存构建大数据存储的优势（4）

- 存储结构
  - ✓ 带宽更高 (3 GB/s VS 150 MB/s)
  - ✓ 体积更小、耗电更低、发热更少 (0.068w VS 8.77w)
  - ✓ 新的硬件接口 (PCIe、DIMMs …)
- 系统软件
  - ✓ 异地更新
  - ✓ OOB的利用
  - ✓ 新的软件接口：TRIM、atomic-write ...
- 分布式协议
  - ✓ 异地更新的日志记录方式 → 更加简单的一致性方式[1]
  - ✓ 更低的延迟 (25 us VS 40 ms) → 新的传输路径

1:CORFU: A Shared Log Design for Flash Clusters. NSDI'12, Microsoft

主要内容

一 背景

二 面向大数据的闪存存储系统

2.1 存储结构的变革

2.2 系统软件的变革

三 面向大数据的持久性内存存储系统

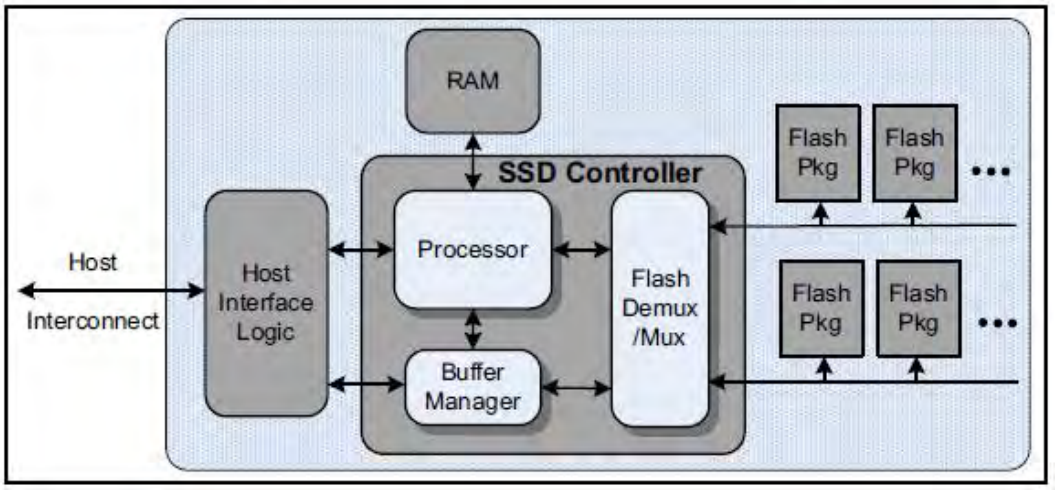
四 新型分布式存储系统

五 展望

2.1 存储结构的变革（1）

- 闪存存储的分类
  - ✓ 固态盘

- 存在的问题



Source: Desgin Tradeoffs for SSD Performance. USENIX'08, UWM

1.固态盘的形式限制了闪存优势的发挥，SATA接口成为存储速度的瓶颈。

2.以文件系统为代表的系统软件，大多以磁盘为假设进行优化，较少考虑闪存特性，其优势难以得到充分发挥。

3.在等待闪存读写时，不能利用主机空闲的计算和内存资源。

2.1 存储结构的变革（2）

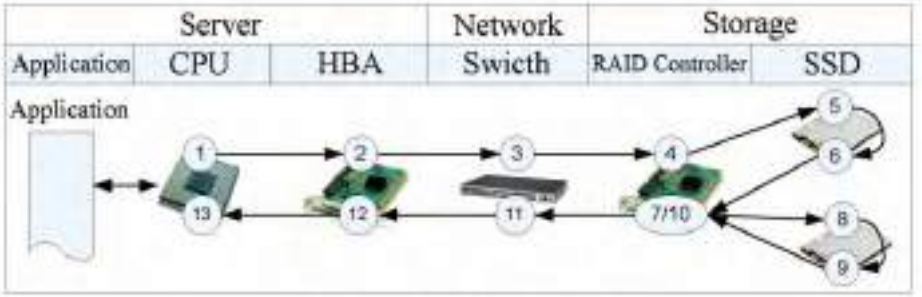
- 闪存存储的分类

✓ PCIe 闪存卡

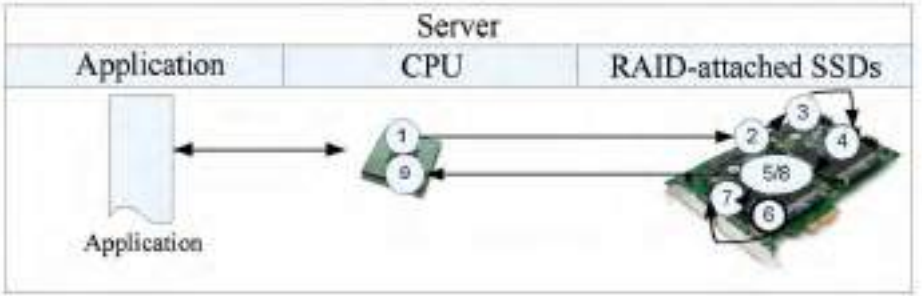
1、在主机端实现闪存转换层
2、能利用主机的冗余计算和缓存能力
3、利用异地更新的特性，可实现多闪存页的原子写操作，避免log带来的重复写，延长寿命
4、将内部的设备信息开放给系统软件，使优化更有针对性

存在的问题

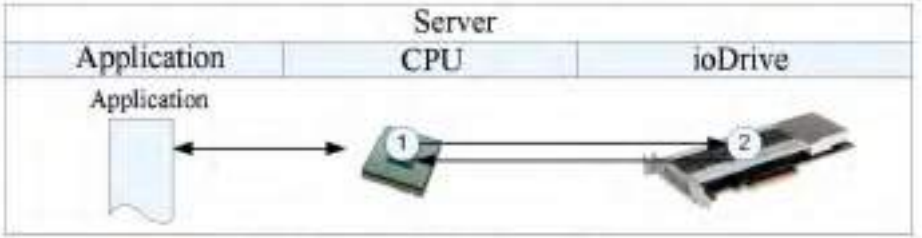
- 难以实现存储容量的扩展及存储共享
- 相比于固态盘，体积和发热较大



(a) I/O path of the disk array



(b) I/O path of the PCIe SSD



(c) I/O path of the FusionIO ioDrive

Source: Survey on Flash-Based Storage Systems. JCRD'13, THU

2.1 存储结构的变革（3）

- 闪存存储的分类

✓ 闪存阵列

基于传统存储阵列的演进式设计

SSD → HDD

优化的存储控制器

存在的问题

- 结构复杂，价格昂贵
- 相比磁盘阵列，有效容量较低

基于全闪存阵列的革新式设计

闪存芯片 → HDD

全新的阵列控制器

2.1 存储结构的变革（4）

- 闪存存储的分类

✓ 基于闪存的分布式集群系统

FAWN[1]：从集群整体设计的角度考虑闪存与处理器的匹配，以降低整体能耗。

低频率CPU　　普通CPU

364 query/J　　>>　　1.96 query/J

Gordon[2]：发挥闪存芯片间的并发特性，匹配处理器和内存芯片的性能与能耗。

已经应用到San Diego 超算中心：

300TB容量，340Tfps

1: FAWN: A Fast Array of Wimpy Nodes. SOSP'09, CMU

2: Gordon: Using flash memory to build fast, power-efficient clusters for data-intensive applications. ASPLOS'09, UCSD

主要内容

一 背景

二 面向大数据的闪存存储系统

2.1 存储结构的变革

2.2 系统软件的变革

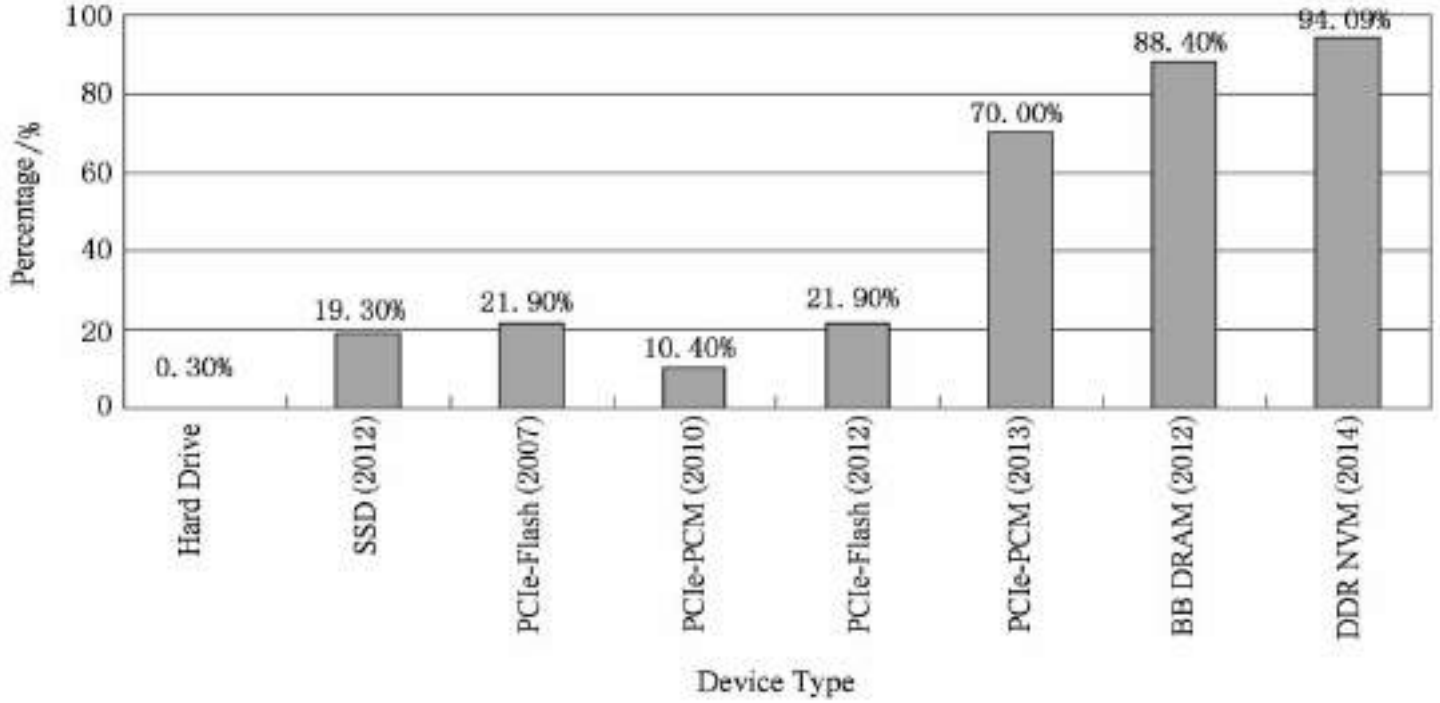
三 面向大数据的持久性内存存储系统

四 新型分布式存储系统

五 展望

2.2 系统软件的变革（1）

随着存储介质访问延迟的越来越低，软件开销所占比例越来越高。报告[1]指出，传统磁盘存储系统中，软件开销所占比例为0.3%，PCIe闪存卡系统中软件开销占21.9%，随着NVM的发展，预计软件开销比例将高达94.09%。



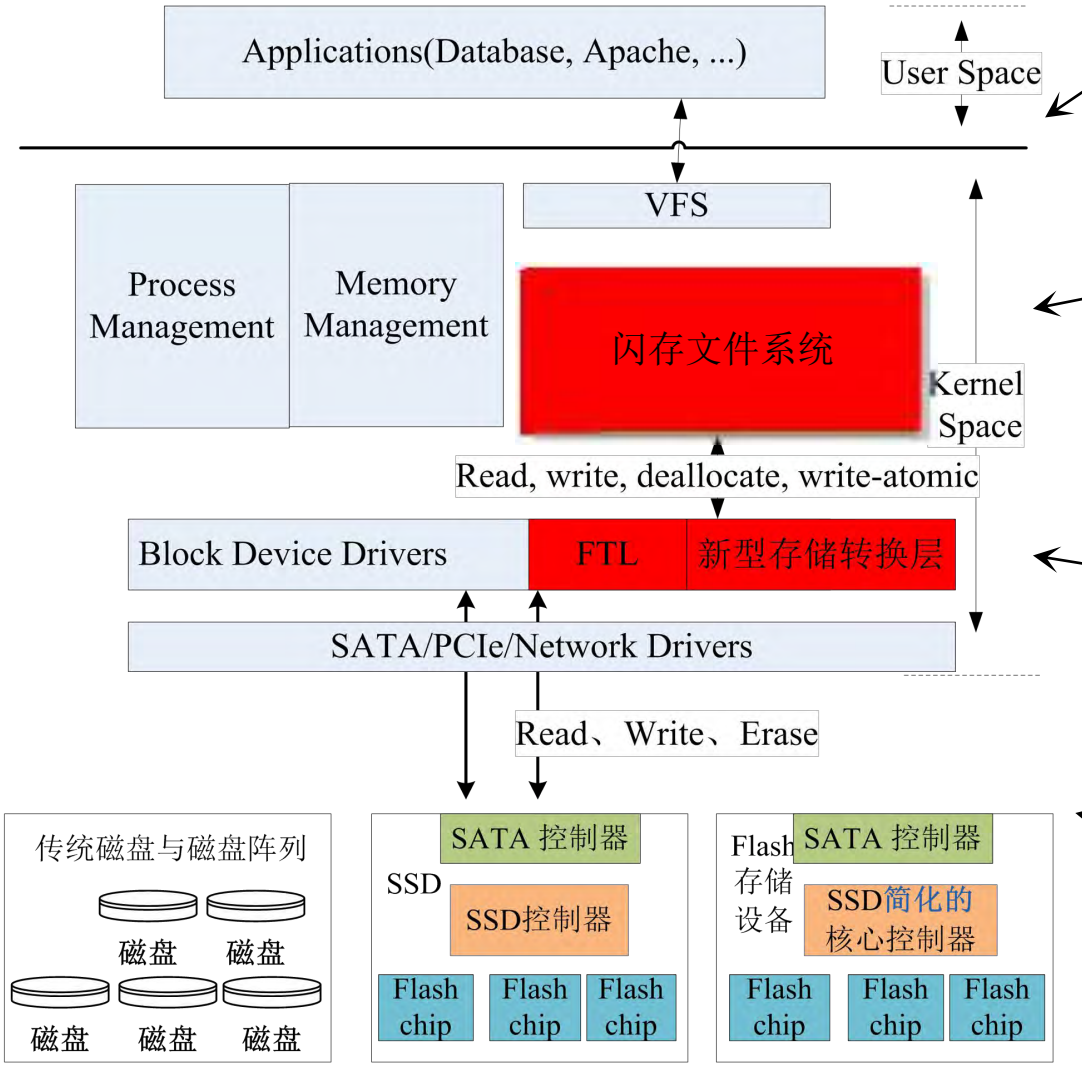
1: Redrawing the boundary between software and storage for fast non-volatile memories.[OL] 2012-9-1, UCSD

2.2 系统软件的变革（2）

- 通知机制

- 存取路径
  - ✓ 在块设备层，基于磁盘的IO调度策略，并不适用于闪存的特性。
  - ✓ 在文件系统层，通过文件系统与闪存设备间的新软件接口，由闪存设备自主选择数据写入的物理地址，再通知文件系统更新记录，减少了重复映射的管理开销。
- 软件接口
  - ✓ TRIM：提供了数据显示删除的语义。
  - ✓ atomic-write：提供原子写操作。
  - ✓ PTRIM：提供持久性删除语义。
  - ✓ EXIST：检查数据页存在性。
- 事务闪存接口
  - ✓ AtomicWrite / TxWrite事务接口

2.2 系统软件的变革（3）



- 基于闪存的KV Store
  - Eurosys'14
  - FAST'16
  - ASPLOS'14
  - ATC'15
  - FAST'17
  - CODE'17
  - DATE'14
- 基于闪存的文件系统
  - FAST'13
  - FAST'10
  - FAST'14
  - FAST'12
  - FAST'15
  - FAST'16
  - DATE'14
  - ATC'16
- 基于闪存的事务管理
  - SOSP'09
  - HPCA'11
  - ICCD'13
  - TC'16
- 分布式闪存的研究
  - NSDI'12
  - SOSP'13
  - ICCD'13
  - IPDPS'14

2.2 系统软件的变革之文件系统（1）

闪存介质的访问，呈现低延迟、读写不对称的特点，随机读写性能较硬盘提升很高。传统针对硬盘优化设计的软件系统直接用于闪存时，一方面带来了不必要的冗余功能，另一方面隐藏了闪存可能带来的优势。

- **传统文件系统存在的问题**

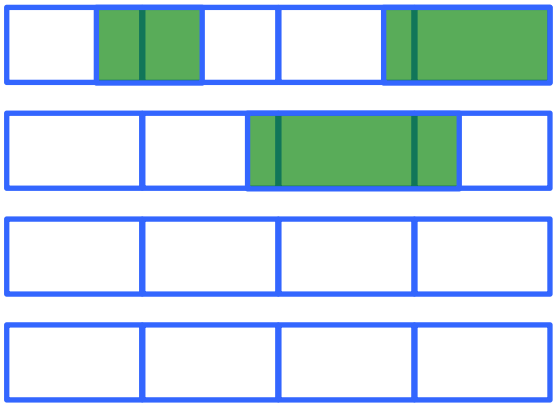
1.冗余工作：文件系统中从文件逻辑块到设备物理块的映射，与FTL中逻辑地址到物理地址的映射 → 双层映射

2.语义缺失：设备不能理解数据页面间的关系，难以优化数据分布和感知文件系统的操作 → 数据删除

3.特性缺失：闪存设备的异地更新与文件系统的原子性操作 → 两次写

2.2 系统软件的变革之文件系统（2）

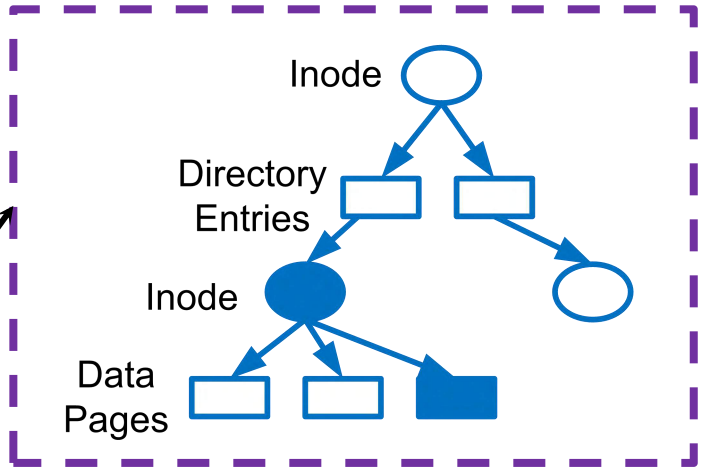
- SSD文件系统的潜在问题一：优化错配
- 文件系统对磁盘读写等优化并不适用于闪存存储
- 闪存在读写特性等方面与传统磁盘存在较大差异
  - ✓ 闪存的随机读性能较好，但随机写性能较差
  - ✓ 闪存的边界不对齐读写对性能影响较大
  - ✓ ……

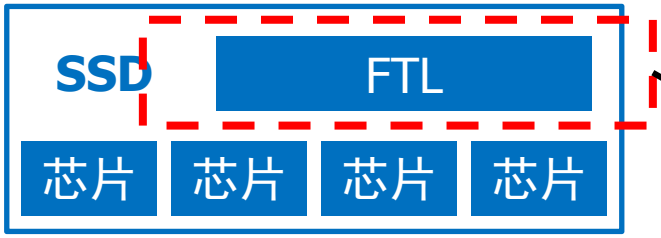


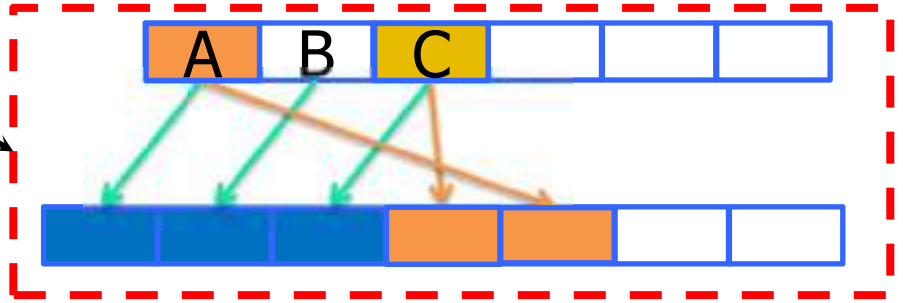
- 传统的顺序写入方法也不完全适用
- 文件系统的写入还影响垃圾回收效率，以及闪存写入寿命
- ……

2.2 系统软件的变革之文件系统（3）

- SSD文件系统的潜在问题二：功能冗余
- 文件系统的存储管理与FTL存储管理存在冗余
  - ✓ 地址映射关系的冗余
  - ✓ 空闲空间管理的冗余
  - ✓ ……

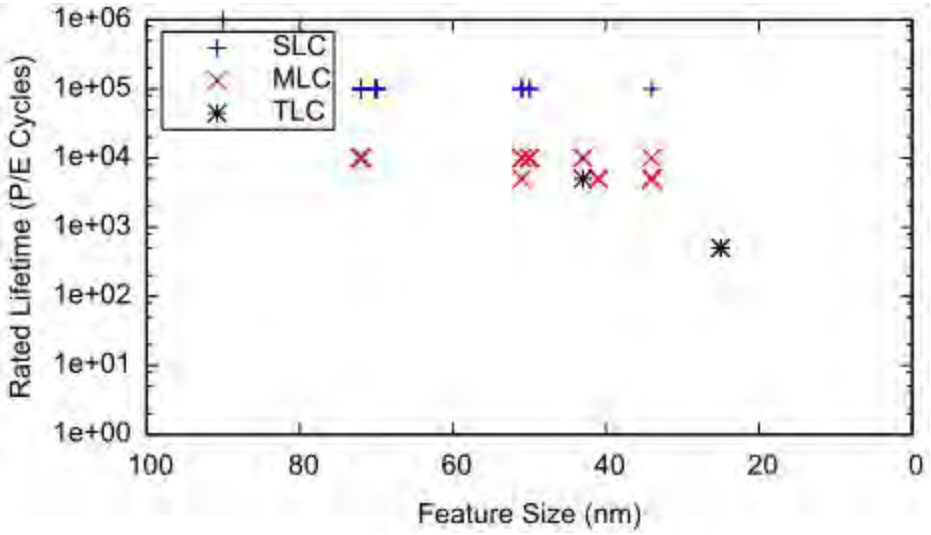






2.2 系统软件的变革之文件系统（4）

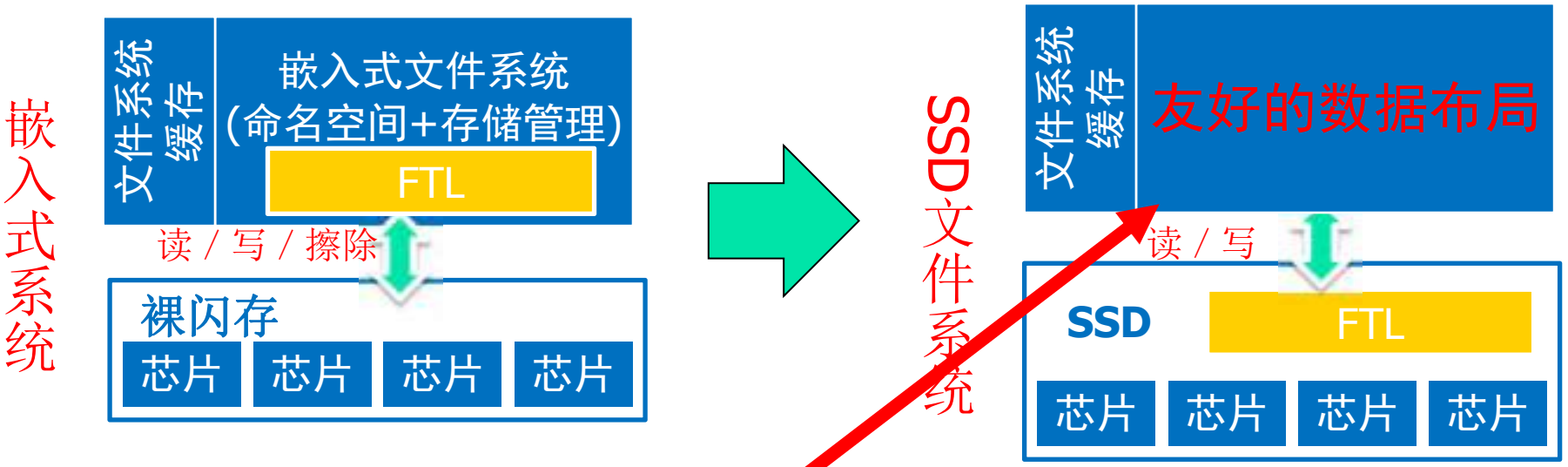
- SSD文件系统的潜在问题三：维度缺失
- 耐久性问题
  - ✓ 随着写入次数增加，闪存单元可靠性降低
  - ✓ 寿命次数：SLC(100,000)-> MLC (10,000)-> TLC(1,000)



图片来源：The Bleak Future of NAND Flash Memory, FAST'12

- 文件系统管理是否引入额外的数据量的写入？
- 如何控制文件系统自身的写入数据量？

2.2 系统软件的变革之文件系统（5）

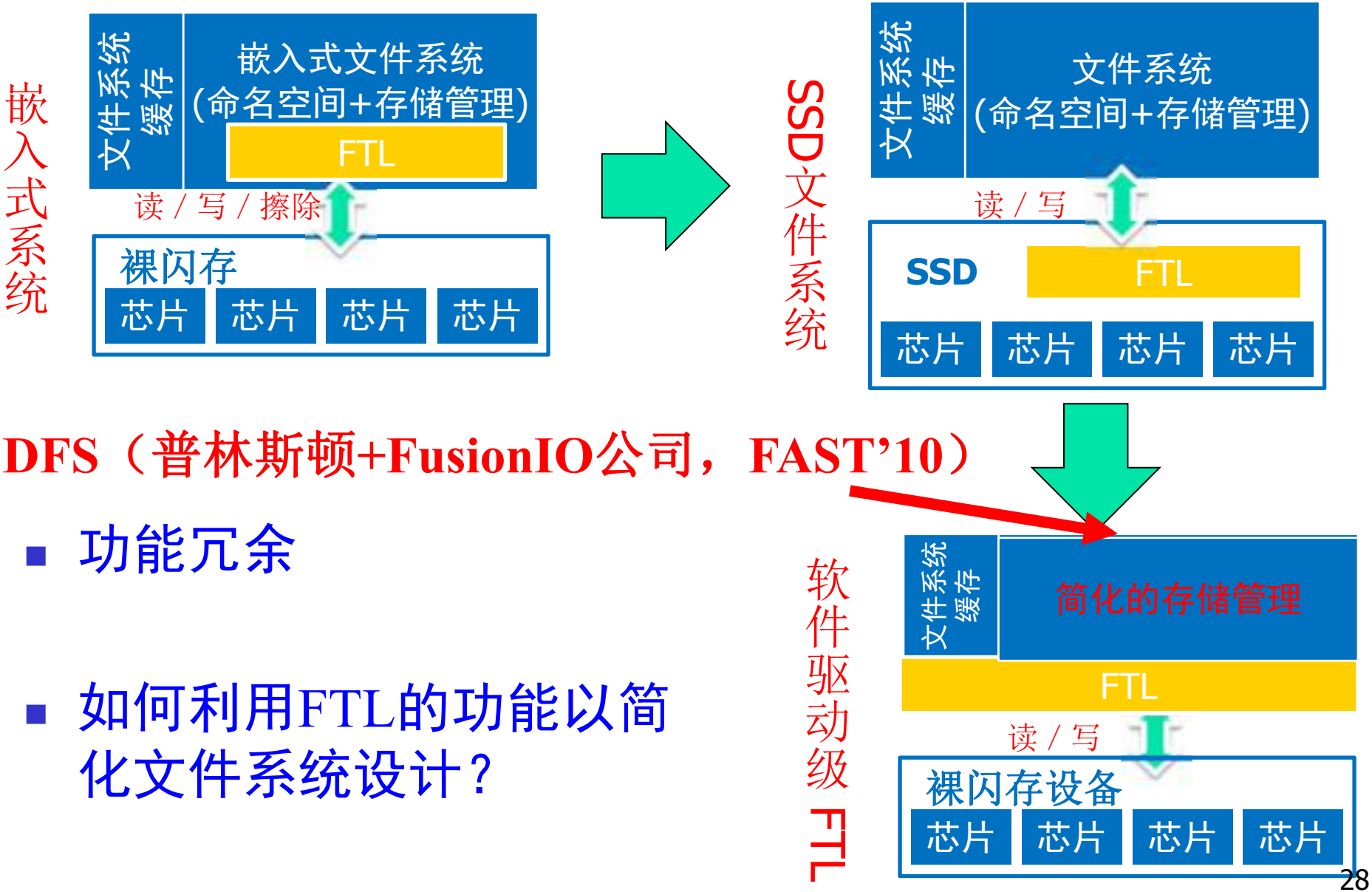


- 优化错配

F2FS（三星公司，FAST'15）

- 如何根据SSD特性设计适用于SSD的文件系统？

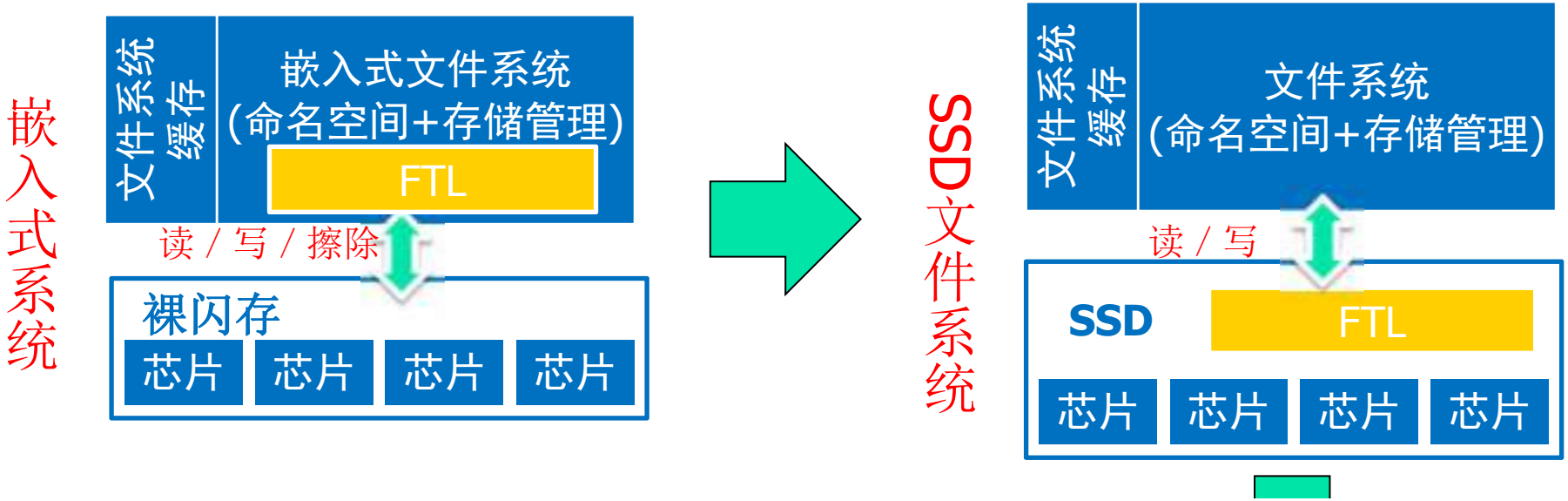
2.2 系统软件的变革之文件系统（6）



DFS（普林斯顿+FusionIO公司，FAST'10）

- 功能冗余
- 如何利用FTL的功能以简化文件系统设计？

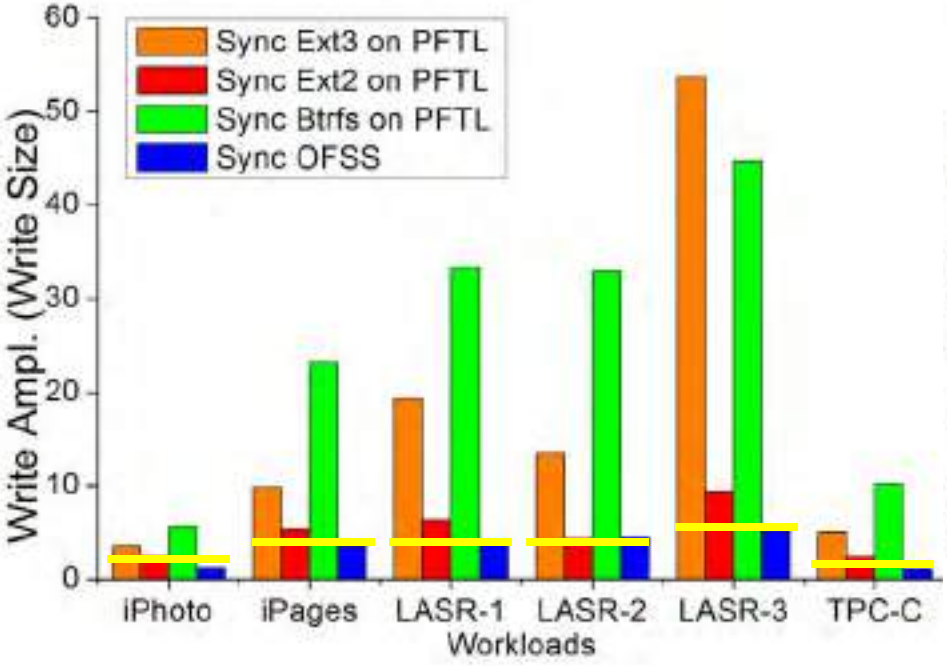
2.2 系统软件的变革之文件系统（7）

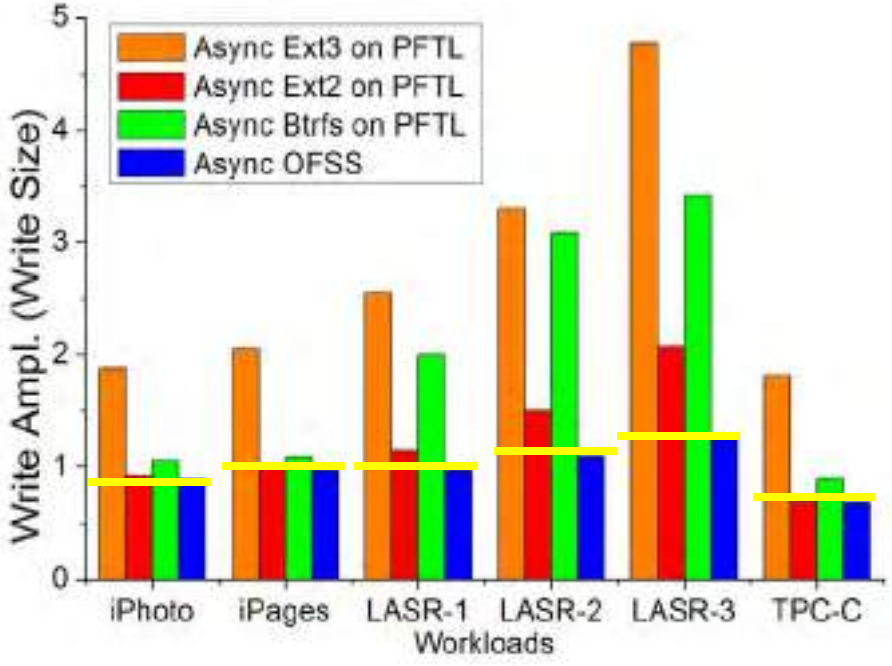


- 优化错配 **ReconFS (清华大学, FAST'14)**
- 如何利用闪存特性设计新目录树管理？
- 维度缺失 **OFSS (清华大学, FAST'13)**
- 如何在文件系统中考虑耐久性？
- 语义隔离 **ParaFS (清华大学, ATC'16)**
- 如何充分发挥闪存的内部并发特性？

2.2 系统软件的变革之文件系统（8）

- 大数据存储系统对闪存的使用寿命要求更高
  - ✓ 采用开放通道闪存可以有效提升闪存的使用寿命

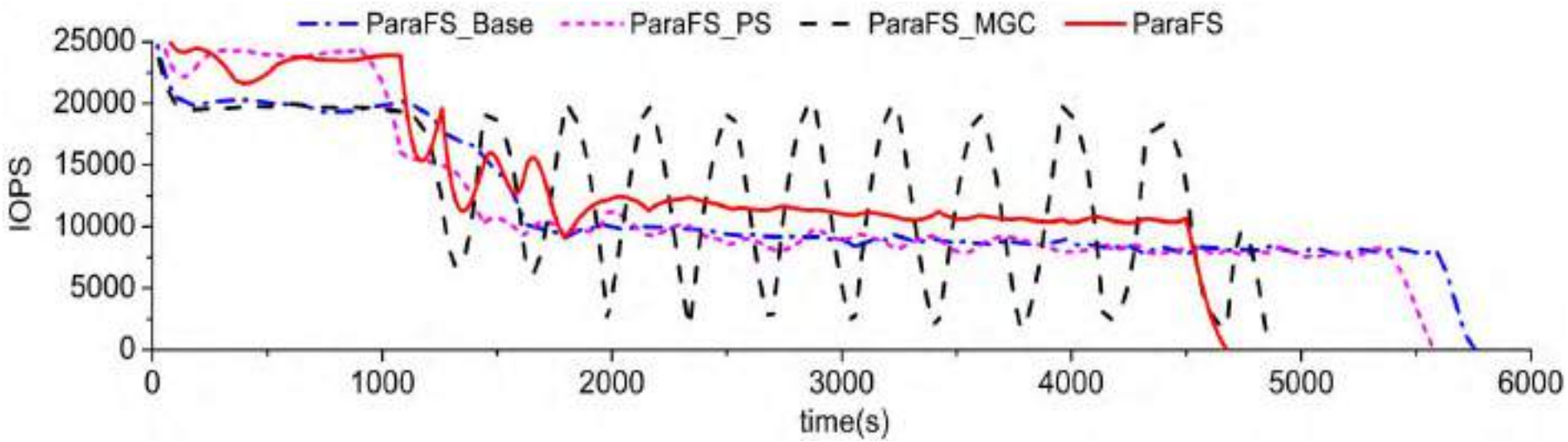




- 使用寿命提升20%至6.7倍

2.2 系统软件的变革之文件系统（9）

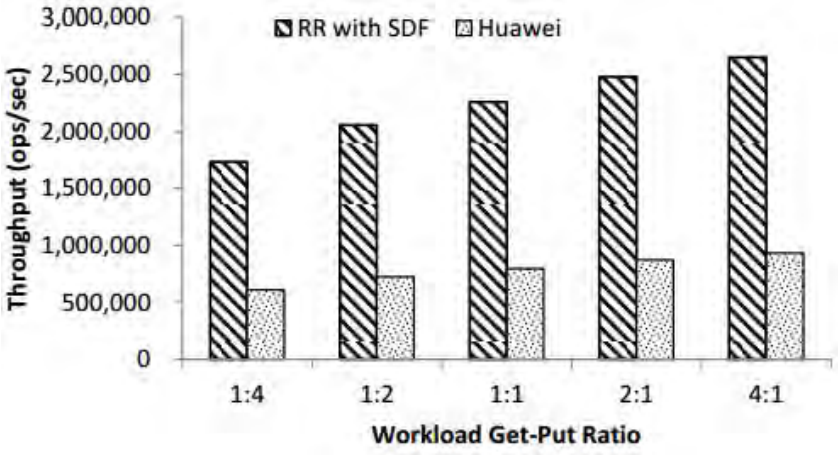
- 大数据应用场景对存储性能的稳定性要求更高
  - ✓ 采用开放通道闪存可以更好地保证稳定的存储性能

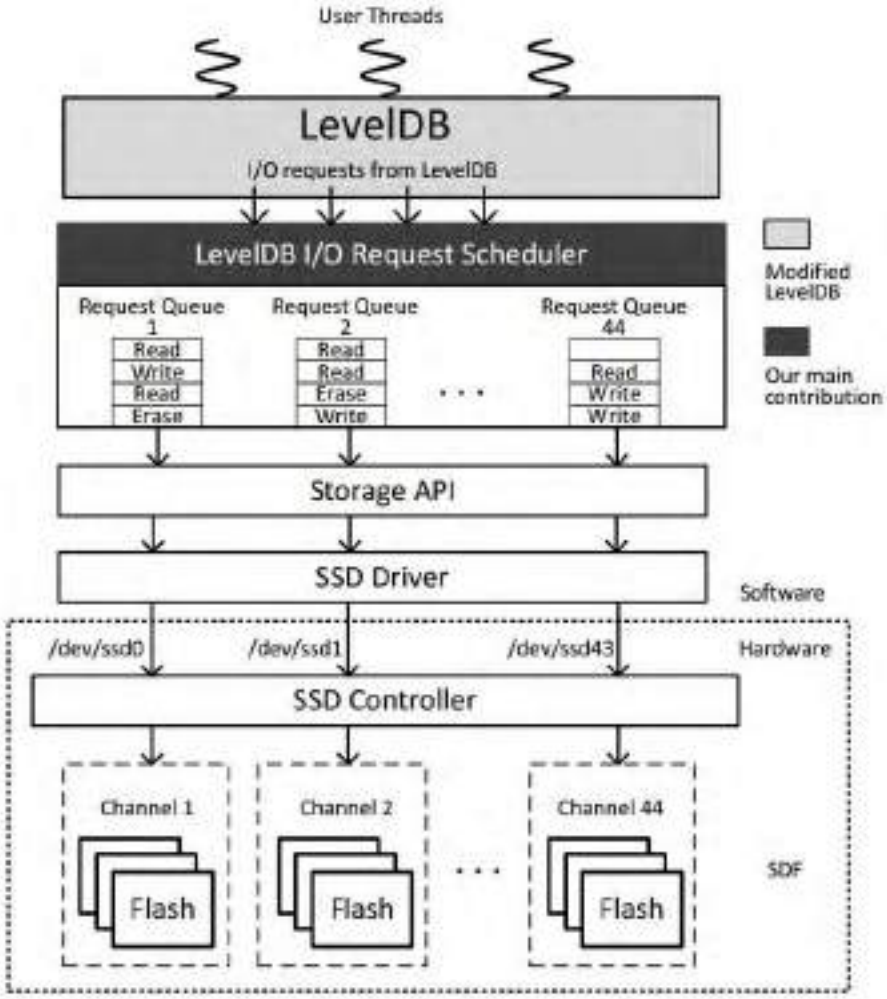


- 存储性能稳定性更好
- 应用执行时间更短

2.2 系统软件的变革之KV数据库（1）

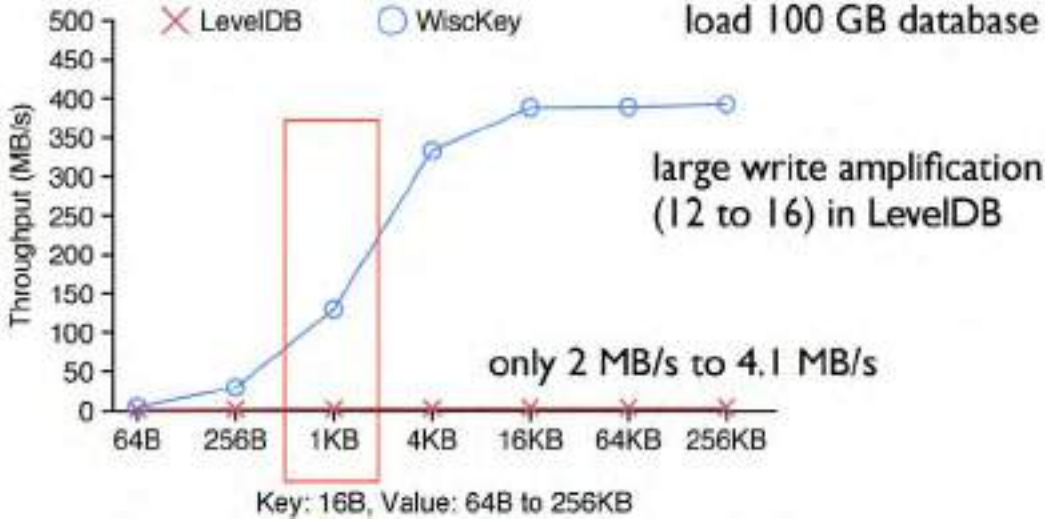
- LOCS—基于裸闪存的KV数据库（Eurosys'14）
  - ✓ 数据库直接管理裸闪存，绕过文件系统
  - ✓ 并发感知的请求调度
  - ✓ 发挥设备的内部并发特性

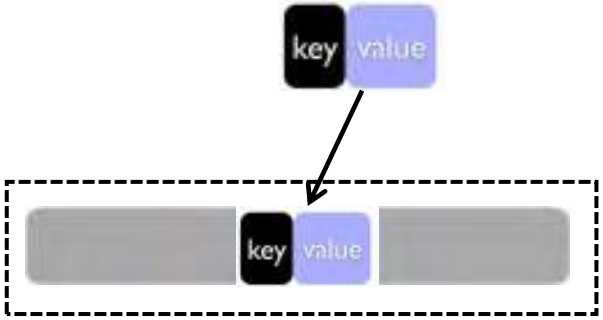




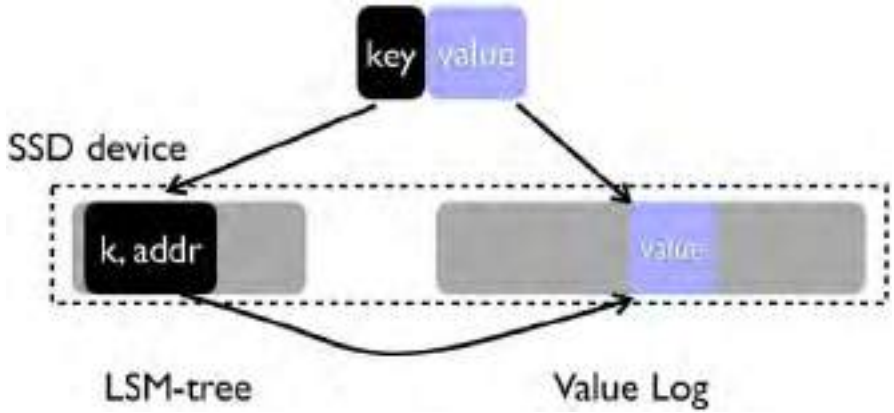
2.2 系统软件的变革之KV数据库（2）

- WiscKey—Key Value分离存储的LSM-Tree（FAST'16）
  - ✓ LSM-tree 的压缩操作会产生严重的写放大。
  - ✓ Key Value分隔存储。使用LSM-tree存储key，使用Log存储value。
  - ✓ 发挥闪存设备，随机读的性能高的特性。





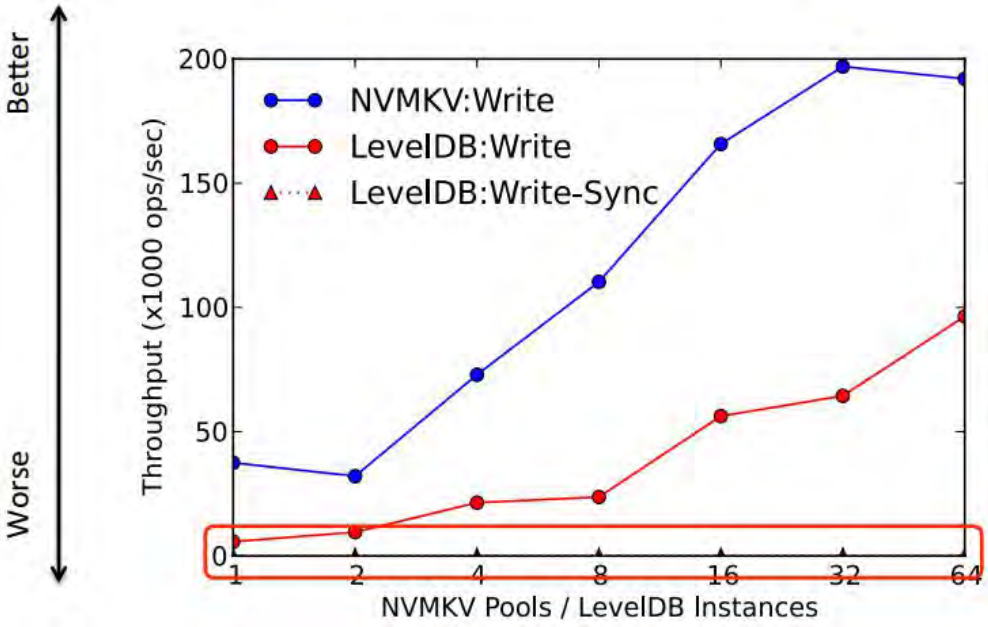
LSM-tree

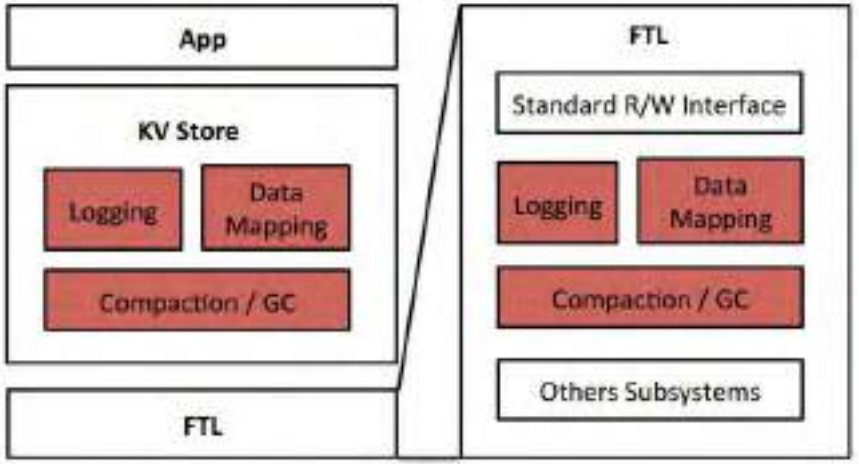


WiscKey

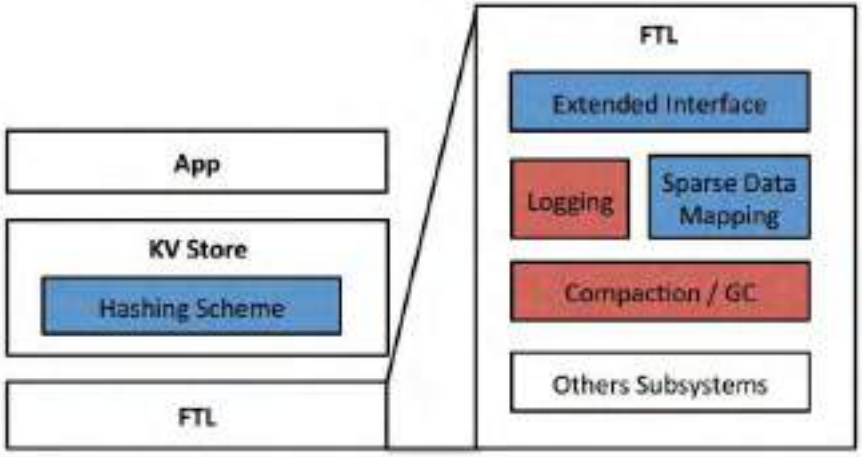
2.2 系统软件的变革之KV数据库（3）

- NVMKV—基于闪存扩展接口的KV缓存系统（ATC'15）
  - ✓ 闪存的FTL与KV存储系统存在着功能上的冗余。
  - ✓ 通过扩展FTL的接口，将与FTL冗余的功能移植到FTL中来完成。





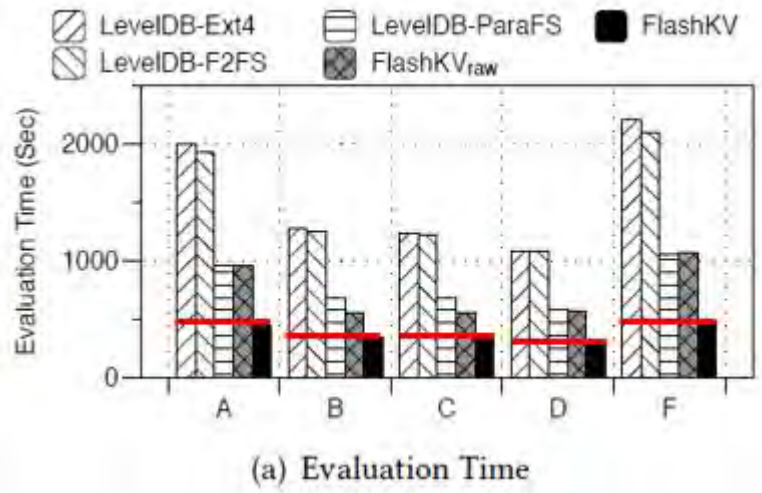
传统的KV系统



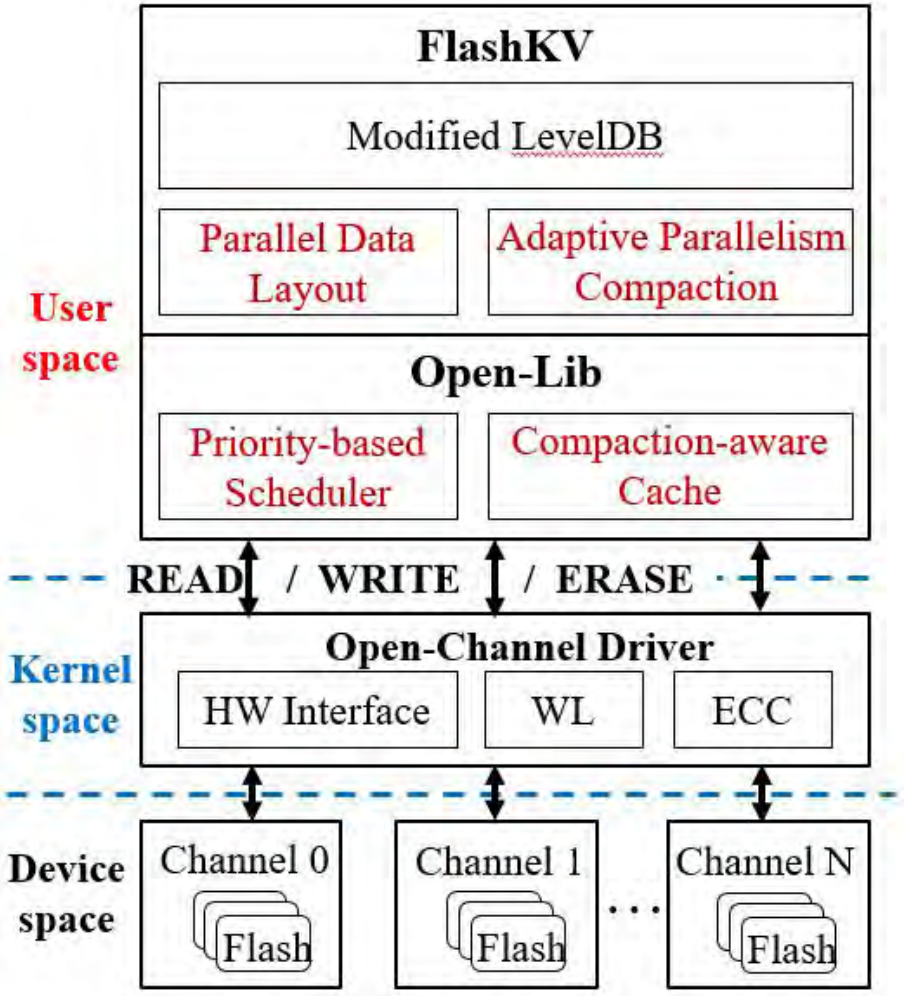
NVMKV

2.2 系统软件的变革之KV数据库（4）

- FlashKV—基于开放通道闪存的KV系统（CODES+ISSS'17）
  - ✓ 减少KV存储系统、文件系统与闪存的FTL三者之间功能上的冗余，消除KV系统与闪存之间的语义隔离。
  - ✓ 利用KV系统和闪存的特性，优化了数据布局，提出了自适应的压缩策略、压缩感知的缓存算法和基于优先级的请求调度机制。



(a) Evaluation Time



主要内容

一 背景

二 面向大数据的闪存存储系统

三 面向大数据的持久性内存存储系统

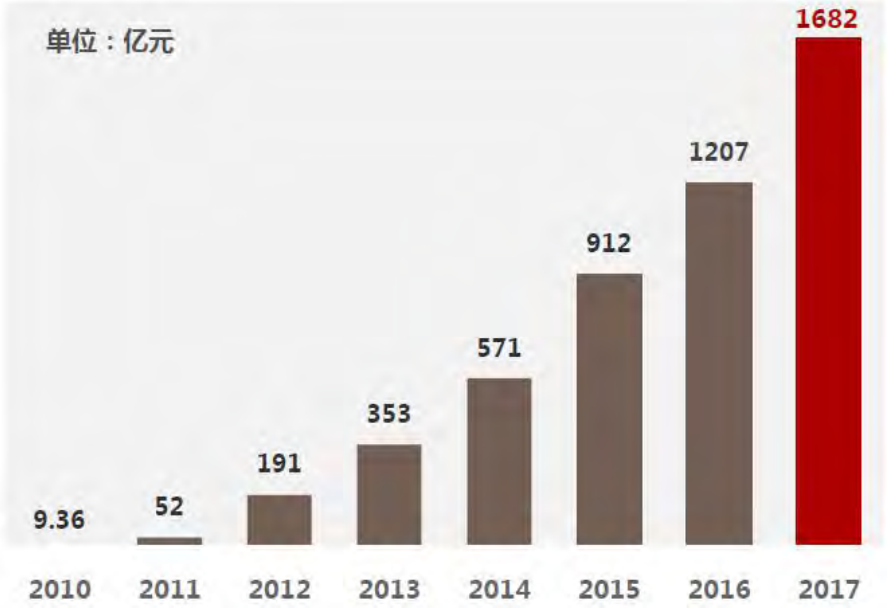
- 3.1 持久性内存编程模型
- 3.2 持久性内存系统软件

四 新型分布式存储系统

五 展望

3. 背景—大数据处理时效性越来越高

2017年双11天猫整体交易再创新高



双11支付总数大幅增长

数据来源：天猫-2017.11.12；

12306 9.56 million tickets/day

每天超过30亿次搜索

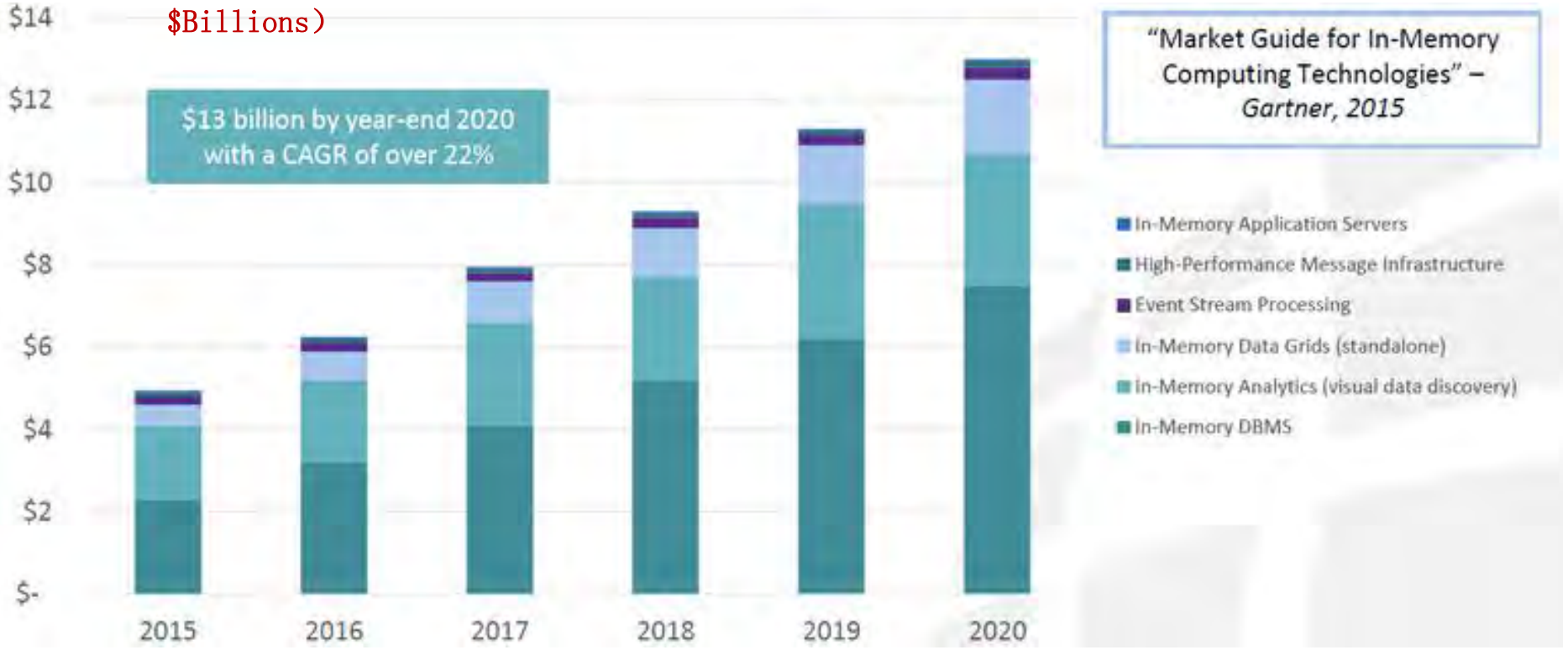
Google

陈海波-CCF ADL；Google

3. 背景—内存计算的兴起

- 内存代价的降低（每12个月降低32%[1]）
- 支持数据的近实时计算与分析

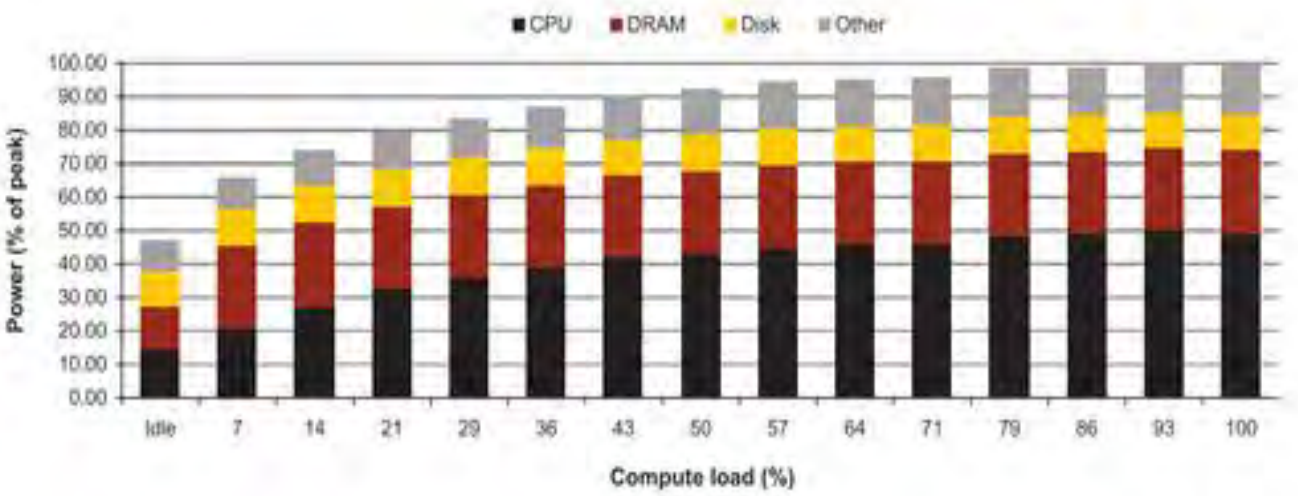
内存计算技术市场的快速发展[2] （in $Billions）

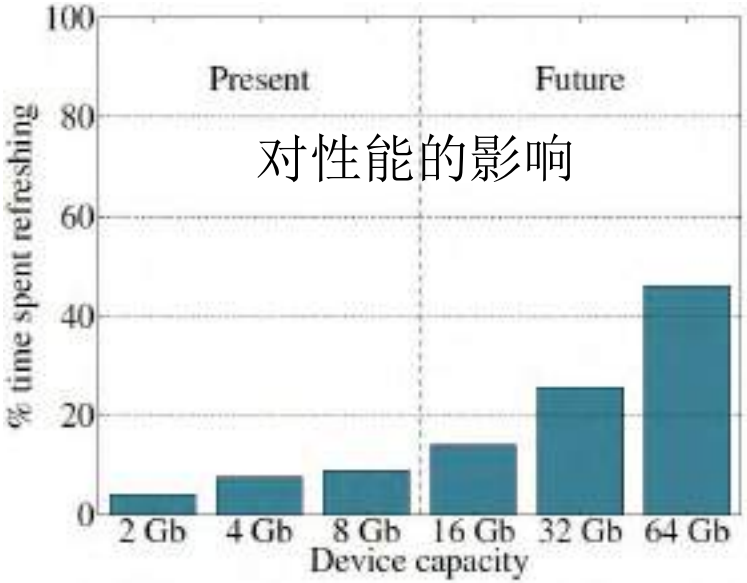


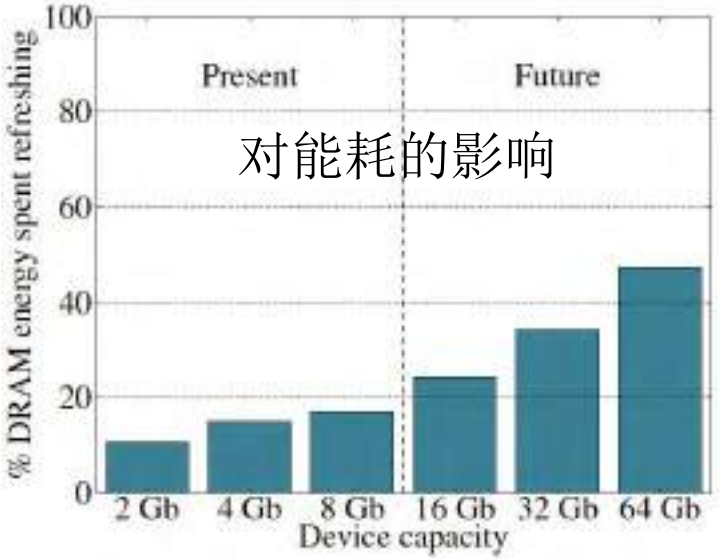
1. Gartner's "Weekly Memory Pricing Index, 21 December 2012", G00247628
2. Gartner's "Market Guide for In-Memory Computing Technologies", 2015

3.背景—内存刷新带来的问题

- 高能耗
- 性能降低
- 性能难预测
- 限制扩展性







数据来源: 金海-2017云栖大会的PPT