闪存存储特性以及文件系统应用

Greenliant®

李炫辉，13581768005，xli@greenliant.com

主要内容

- 闪存特性以及架构
- 利用闪存优化性能
- 当前闪存存在的问题

主要内容

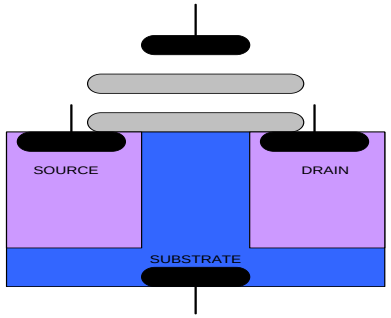
- **闪存特性以及架构**
- 利用闪存优化性能
- 当前闪存存在的问题

SSD和闪存

- 采用Flash Memory的存储设备，可以统称为闪存存储
- 固态硬盘（SSD）可以由volatile/non-volatile memory构成。
- 除了闪存以外，还有其它多种快速存储技术，如DRAM，NVRAM， MRAM and Spin-Torque（自旋力矩磁阻式随机存取内存）， Carbon Nanotube（ 碳纳米管 ）， Phase Change Memory（相变内存）， Memristor （ 忆阻器 ）等等。
- 决定快速存储大规模应用的主要因素是量产规模、稳定性以及经济性
- 闪存主要用于IO性能加速环境，如数据库加速、虚拟化、延时敏感型应用、Server SAN或SDS、大数据处理等等

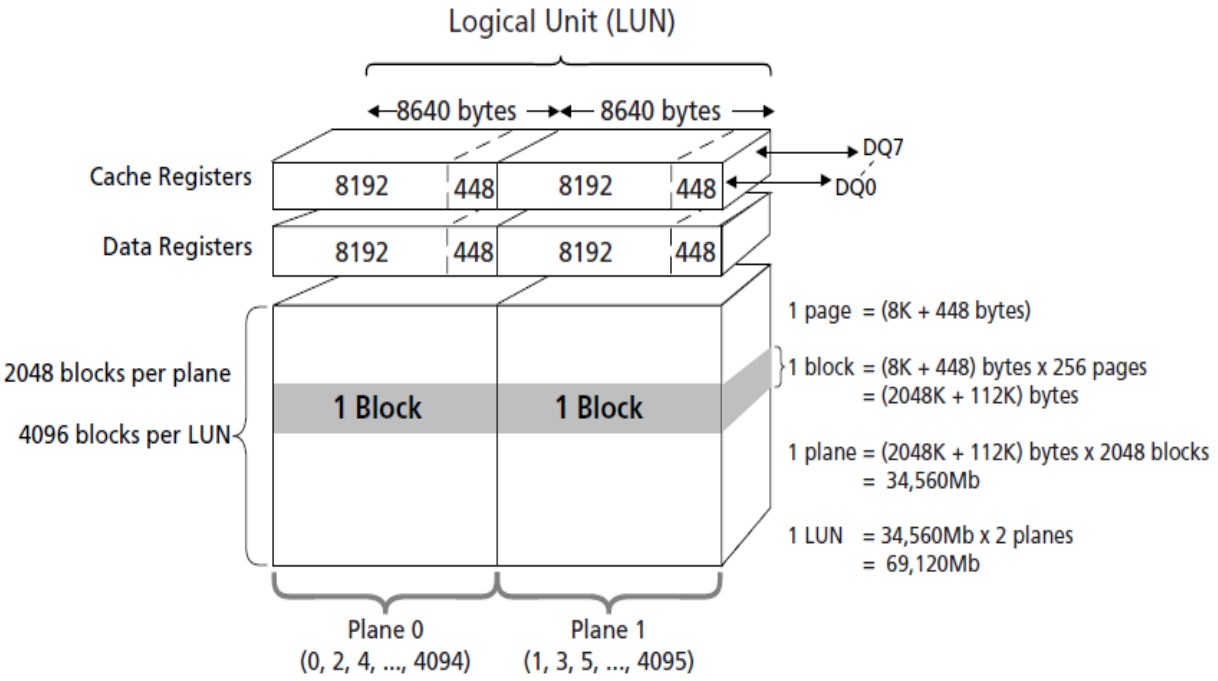
NAND 闪存颗粒

- 当前SSD主要使用NAND Flash：非易失存储介质，成本低



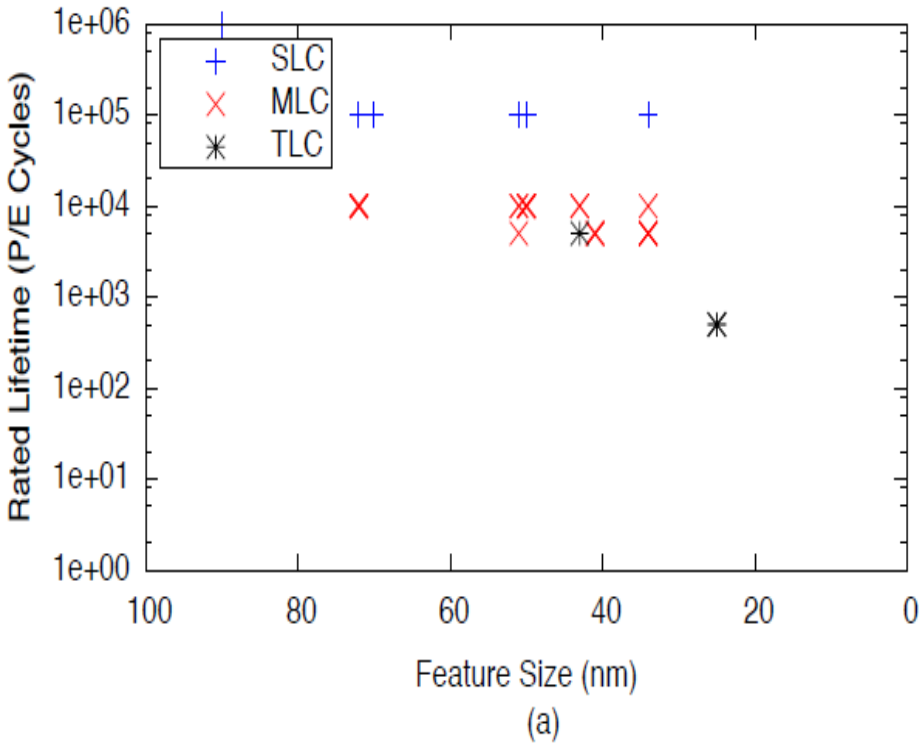
当前的挑战:

- 耐久性 ↓
- 数据保持性 ↓
- 读写干扰 ↑
- 制造工艺缺陷 ↑



Greenliant®

NAND Flash可靠性随制程工艺减小而降低



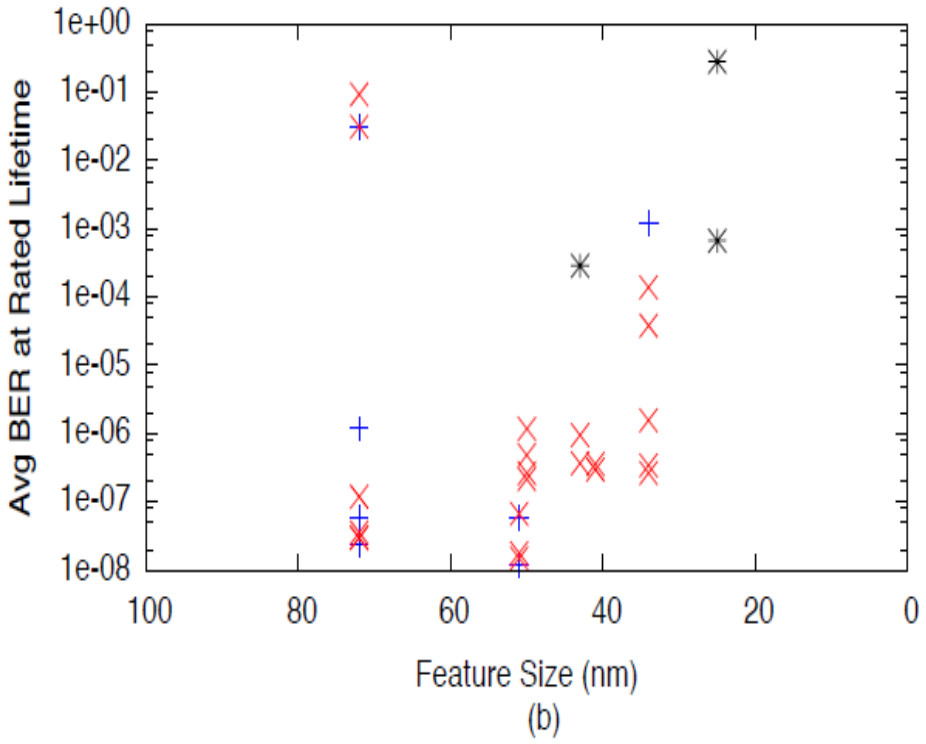


Figure 1: **Trends in Flash's Reliability** Increasing flash's density by adding bits to a cell or by decreasing feature size reduces both (a) lifetime and (b) reliability.

数据来源: nvsl.ucsd.edu

Greenliant®

闪存与磁盘系统的对比

	闪存	磁盘
访问机制	电子驱动	机械驱动
IOPS	几万到几十万	几十到几百
延时	几十微秒到一百多微秒	几毫秒
稳定性	随时间和数据擦写量增长存在波动和衰减	无衰减
擦写次数	SLC 10万次， MLC 3000次	无限制
MTBF	与擦写次数相关	200万小时
工作功耗	7到25瓦	7到8瓦

如何解决闪存设备的可靠性是闪存存储产品设计的关键

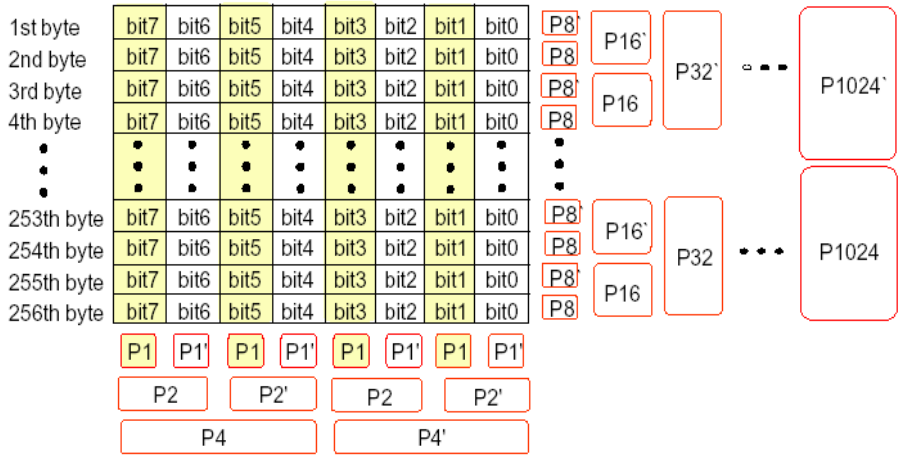
闪存控制器是闪存系统的核心

- 闪存控制器是联系主机和NAND Flash的桥梁
- 闪存控制器有以下闪存管理功能：
  - ✓ Error-correcting code （ECC校验）
  - ✓ Wear leveling（磨损平衡）
  - ✓ Bad block mapping（坏块管理）
  - ✓ Read/write disturb management（读写干扰管理）
  - ✓ Garbage collection（垃圾收集）
- 某些闪存控制器也有其它定制化功能，如
  - ✓ 加密
  - ✓ 安全擦除或自毁
  - ✓ 压缩或去重

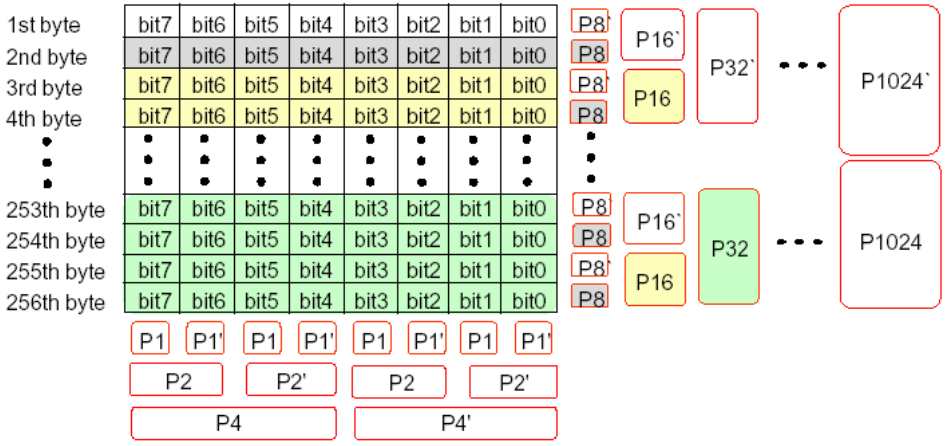
NAND ECC校验机制

- NAND Flash的工艺不能保证NAND的Memory Array在其生命周期中的可靠性，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策略，而管理坏区的前提是能比较可靠的进行坏区检测。
- 常用的ECC算法有Hamming、Reed-Solomon码、BCH、LDPC等。

ECC的列校验和生成规则



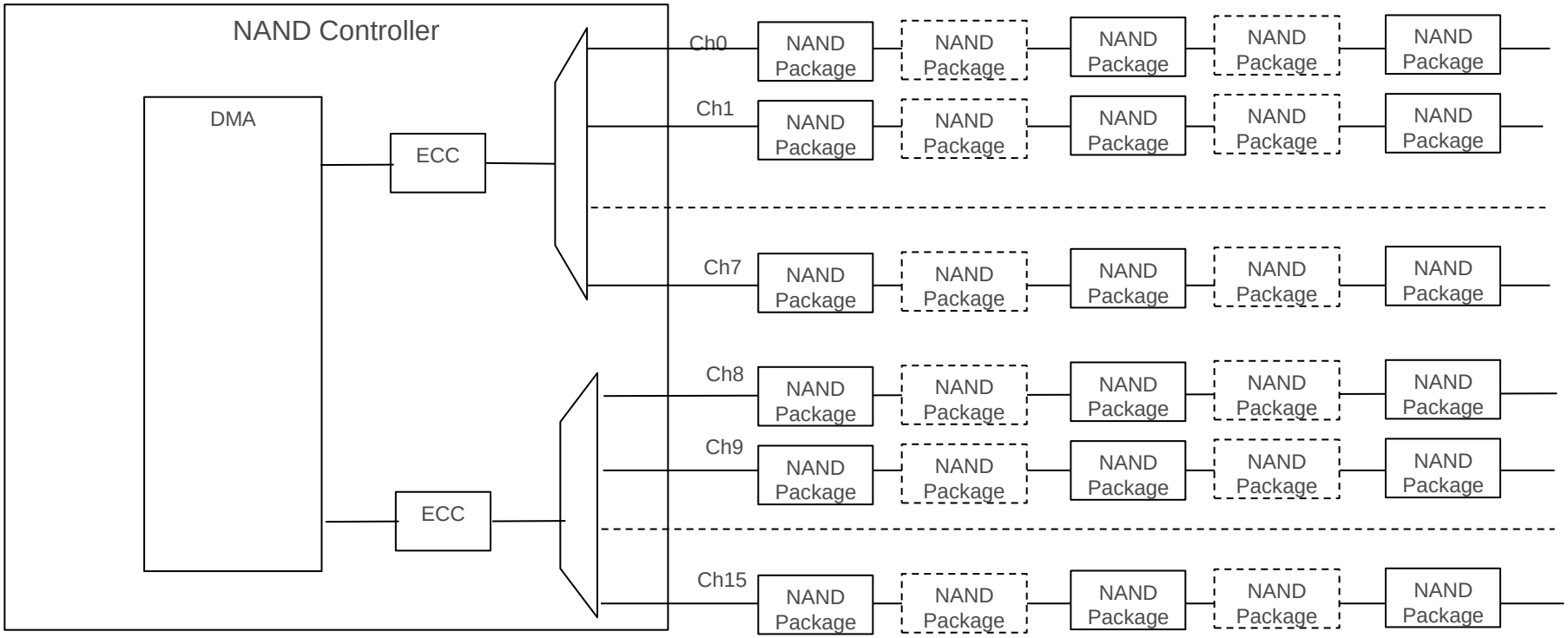
ECC的行校验和生成规则



图表引自http://blog.163.com/starjj_embeded/blog/static/204500051201221702924742/

Greenliant®

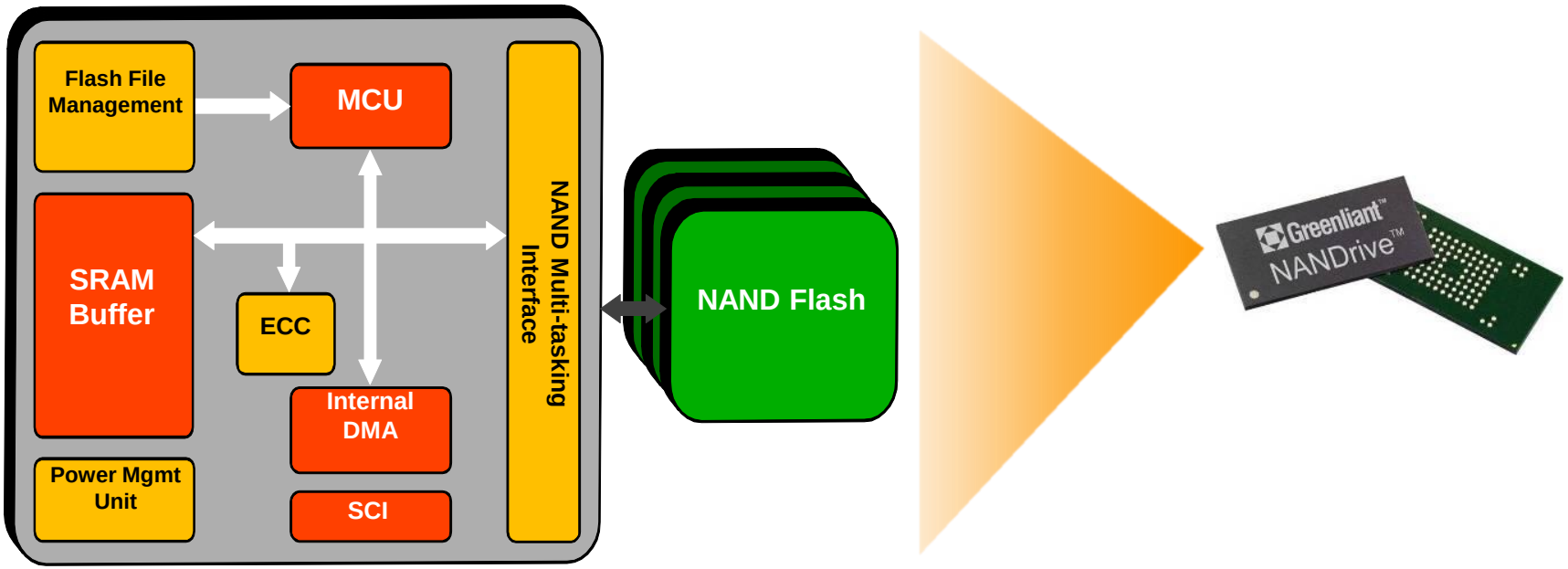
常见闪存卡ECC模块设计



潜在的问题：性能衰减和可靠性下降。
当Flash设备随着使用时间和数据量的增长，坏块会逐渐增加，会产生大量的ECC Error，这时设备性能和可靠性会大幅度下降，对应用性能和数据安全带来影响。

Greenliant®

分布式ECC设计架构—NAND Package封装Flash Controller



Greenliant®

闪存与磁盘系统地址映射比较

PBA

0	1	…	…	…	…	…	…	…	…

LBA

0	1	…	…	…	…	…	…	…	…

PBA

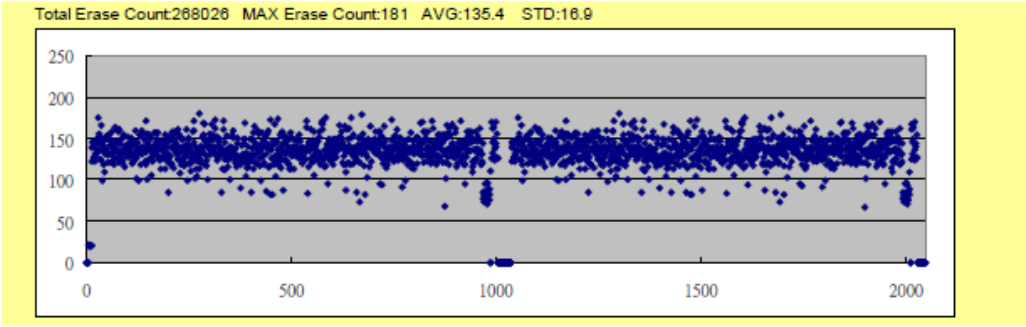
CHS	CHS	…	…	…	…	…	…	…	…

LBA

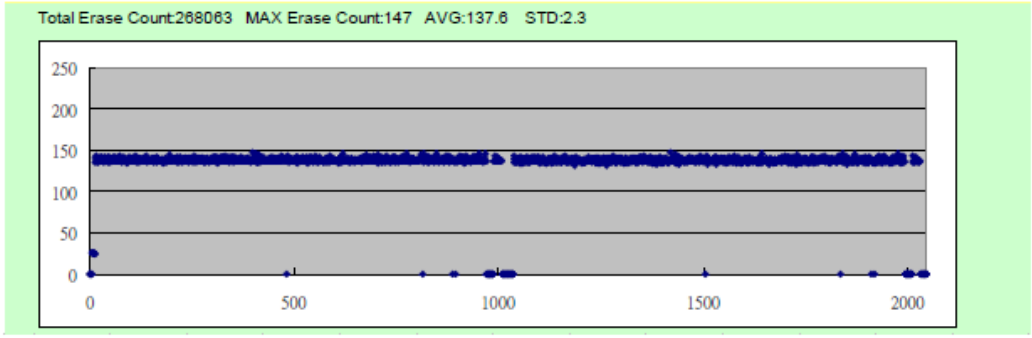
0	1	…	…	…	…	…	…	…	…

Wear Leveling（磨损平衡）

- Wear Leveling可以改善闪存的耐久性，但是会影响闪存的性能
- 动态磨损平衡，静态磨损平衡，全局磨损平衡



Without Wear-Leveling



With Wear-Leveling

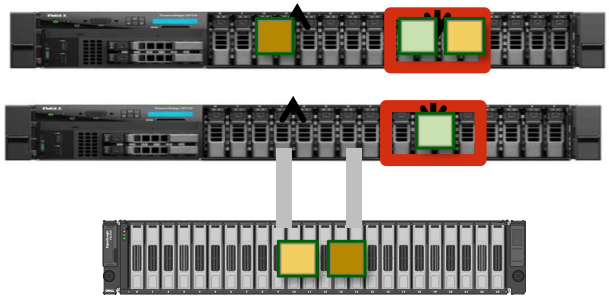
Greenliant®

主要内容

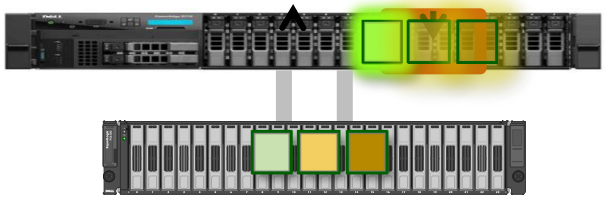
- 闪存特性以及架构
- **利用闪存优化性能**
- 当前闪存存在的问题

采用闪存存储优化系统性能

所有的数据都在闪存存储



混合模式： 一部分数据在闪存存储



缓存架构： 缓存热数据在闪存存储

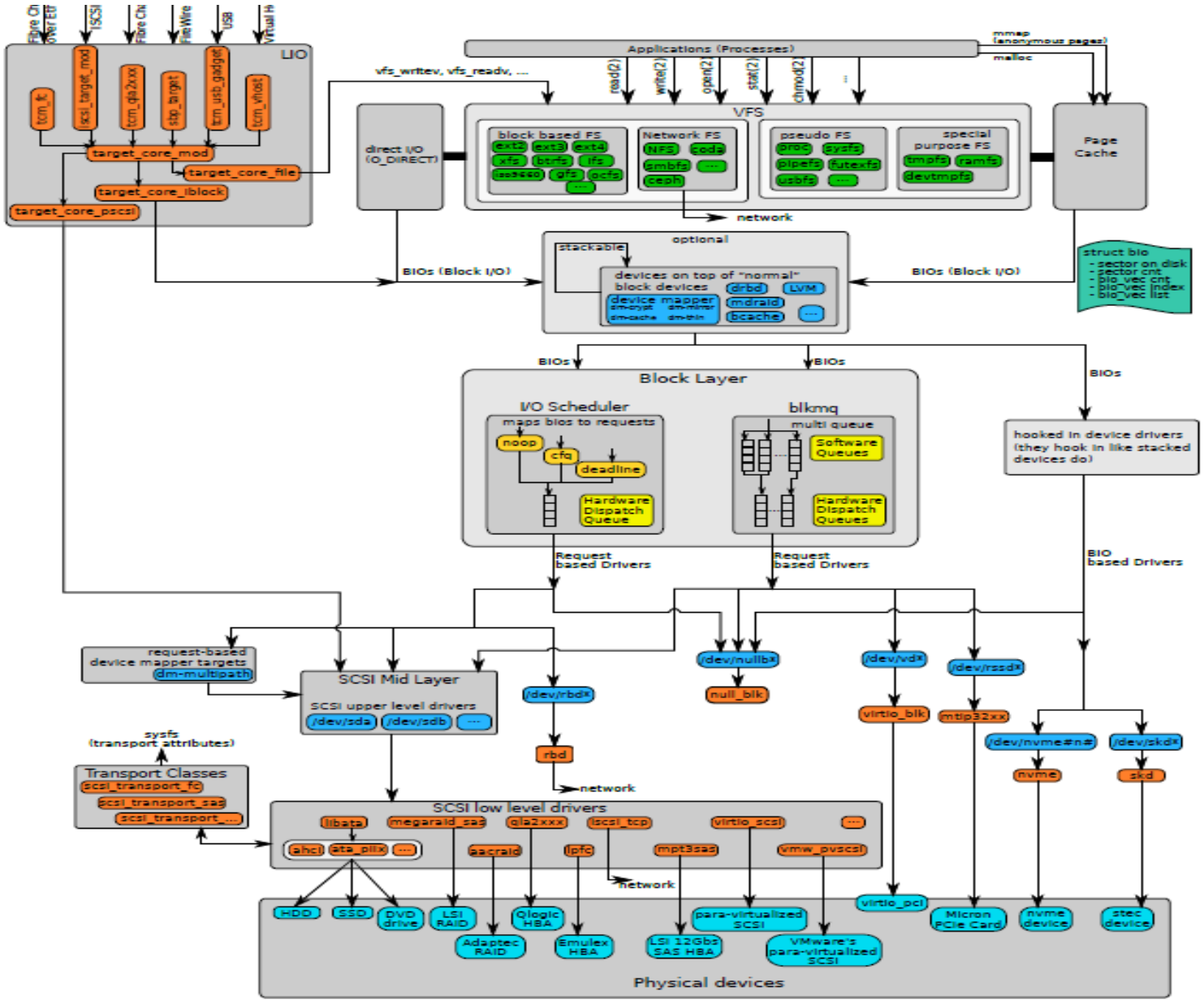
TempDB/Scratch

Index/Transaction

Data

Greenliant®

Linux Storage Stack Diagram version 3.17

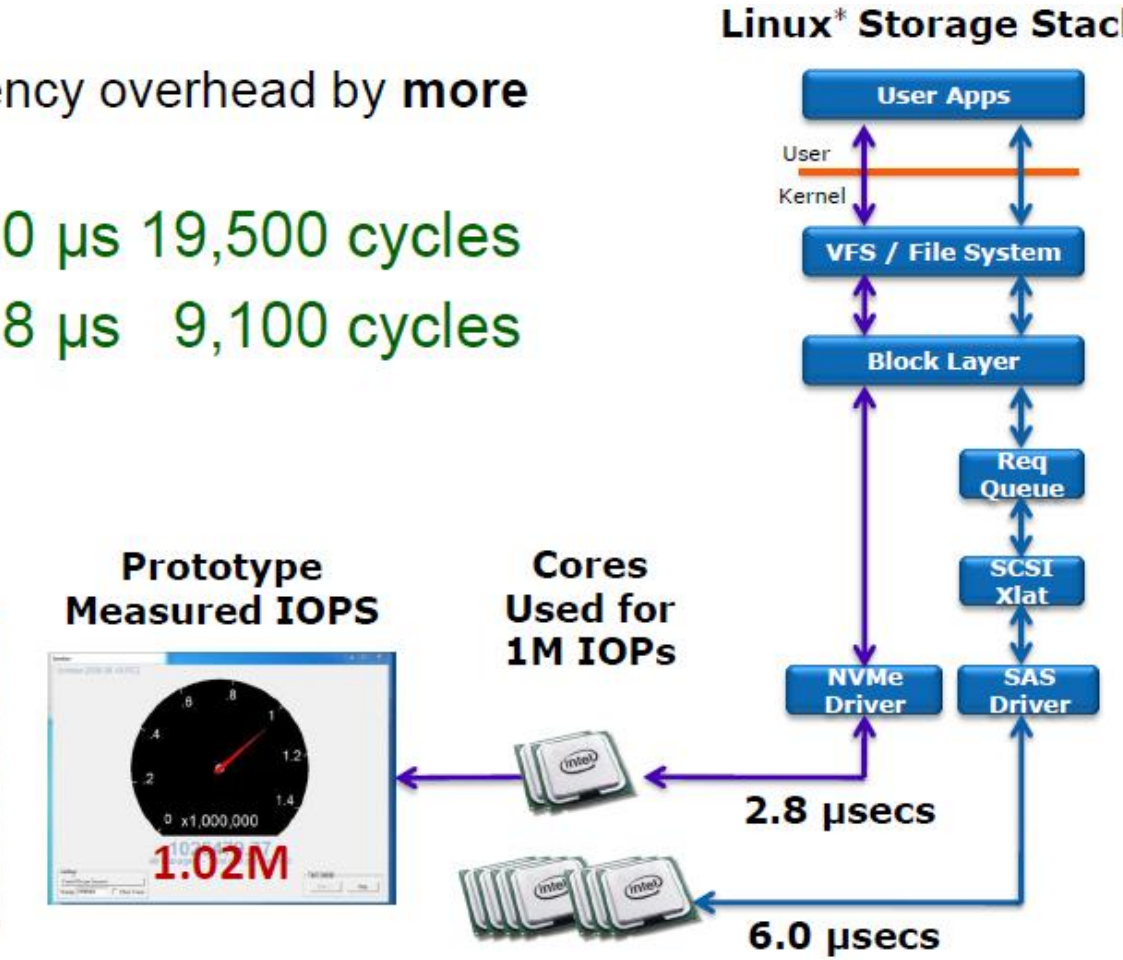


NVMe优化访问堆栈提升性能

NVM Express reduces latency overhead by **more than 50%**

SCSI/SAS: 6.0 µs 19,500 cycles

NVMe: 2.8 µs 9,100 cycles



Measurement taken on Intel® Core™ i5-2500K 3.3GHz 6MB L3 Cache Quad-Core Desktop Processor using Linux RedHat® EL6.0 2.6.32-71 Kernel.

Greenliant®

基于PCIe的NVMe协议

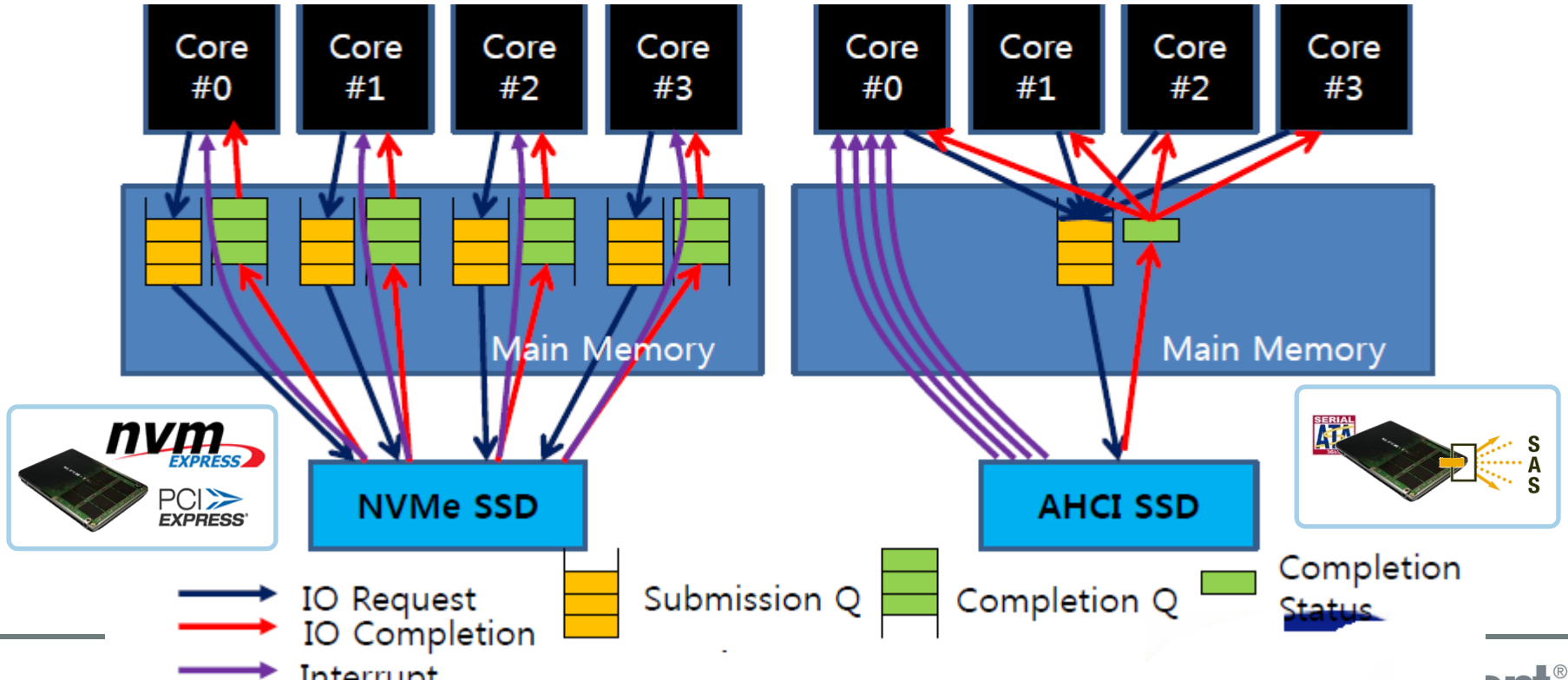
在PCIe的新世代IO逻辑层协议

NVMe协议

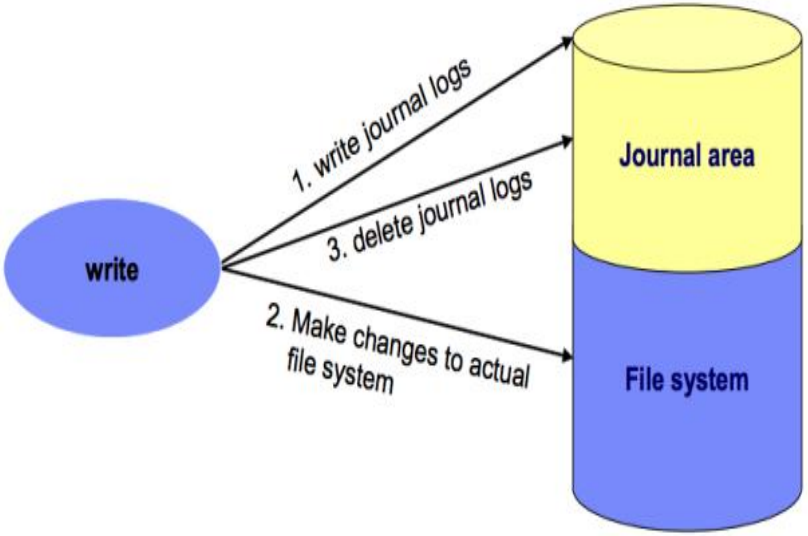
NVMe比AHCI协议的优点

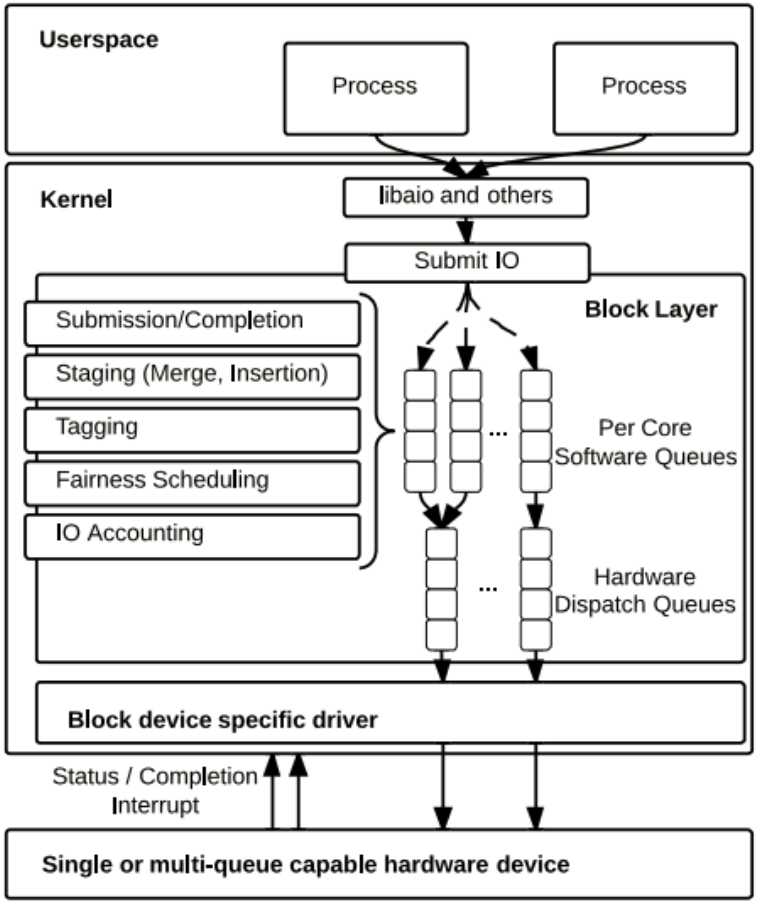
- 无锁式IO并发
- 可横向扩展
- 更低的写延迟

AHCI协议



Linux 文件系统





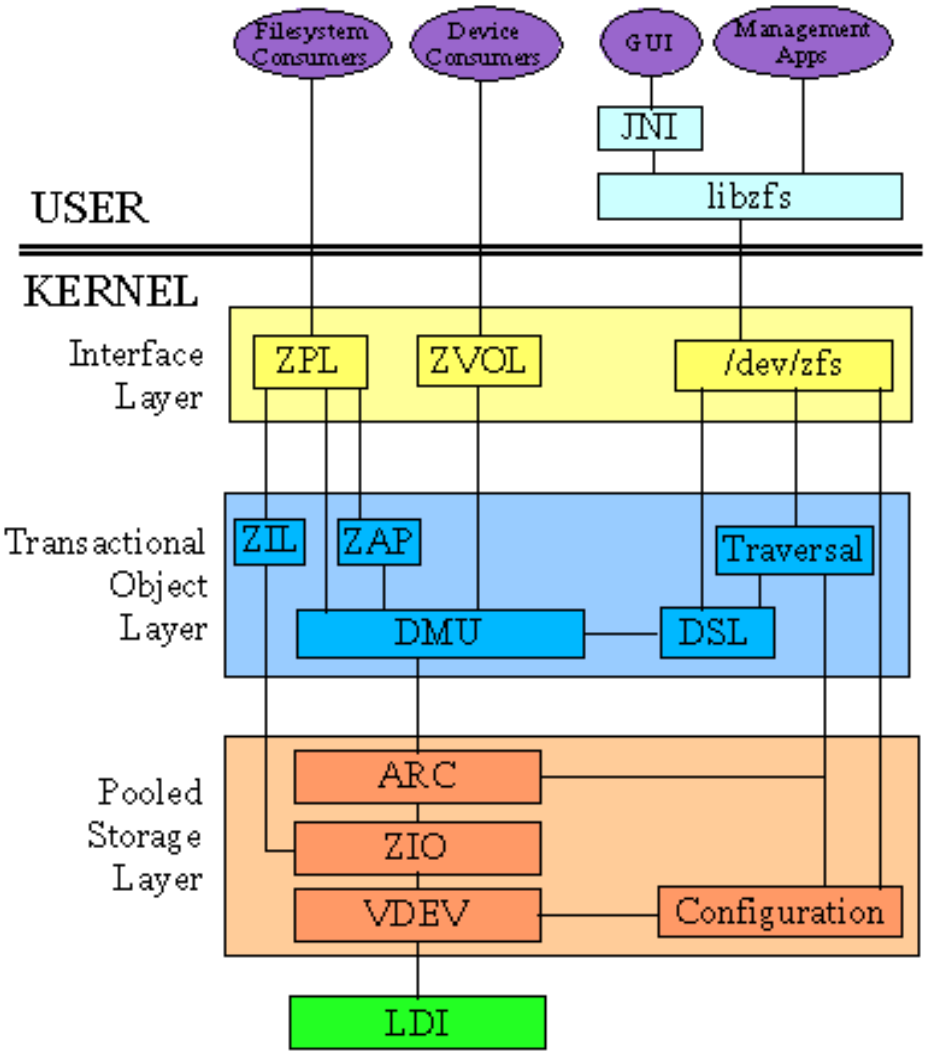
将Journal 放置到闪存上
对数据安全不要求的情况下可以禁用Journal或设置Writeback模式
调整IO scheduler
禁用merge/rotational等
考虑Btrfs

Greenliant®

ZFS文件系统架构

ZFS使用日志机制，ZFS intent log (ZIL) 处理同步写，ZIL处理同步写越快，系统性能越好。将闪存作为ZFS日志设备，可以大幅提升同步写性能。

ZFS的ARC(Adjustable Replacement Cache)读缓存淘汰算法可以优化系统的读性能，而SSD可以作为二级缓存L2ARC设备提供更大的缓存空间



Greenliant®

主要内容

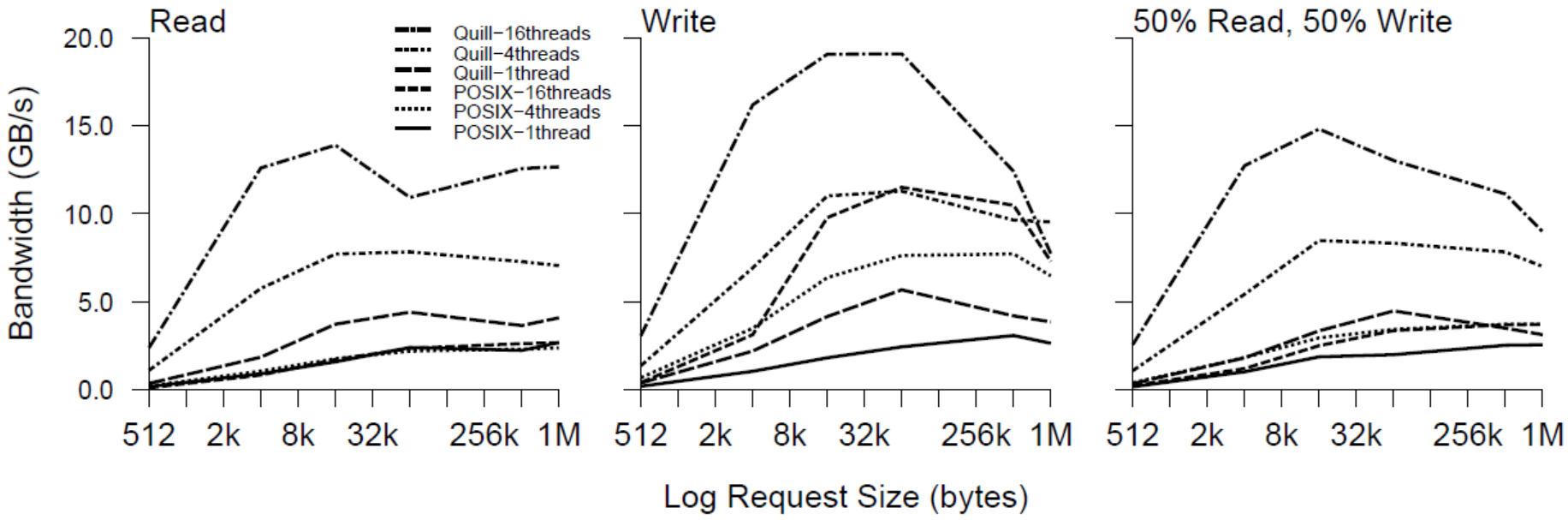
- 闪存特性以及架构
- 利用闪存优化性能
- **当前闪存存在的问题**

操作系统和应用的存储设计均是针对磁盘访问特性进行优化的

- 由于传统磁盘的机械设计，需要通过磁头的旋转进行寻道和数据IO操作，限制了磁盘的性能，因此操作系统和应用的底层算法大部分都是针对如何减少IO操作和寻道时间所设计的
- 磁盘存储架构下为了提高IO性能采用了缓存设计，从而又增加了复杂的数据保护的处理机制。

  VFS—FS—Cache—Abstract Lay of Block device—IO scheduler—Physical device driver， Index算法等等
- 应用的并发度

透明地绕过文件系统使用非易失性存储器



数据来源: nvsl.ucsd.edu

Greenliant®

将闪存设备作为磁盘的缓存会加速设备磨损

- 通过将SSD作为传统磁盘的读写缓存以加速IO性能，通过LFU，LRU，MRU等缓存算法，会带来大量的碎片IO操作，加速SSD设备的磨损。

In a write intensive Virtual SAN environment, the drive may wear out sooner than expected. User should collect more information about the drive and planned workload to determine whether the drive is suitable for deployment."

闪存的性能

- 闪存的性能依赖于很多因素，如主机配置，主机访问模式，块大小，闪存架构等等。
- 闪存的性能需要较长时间的评测

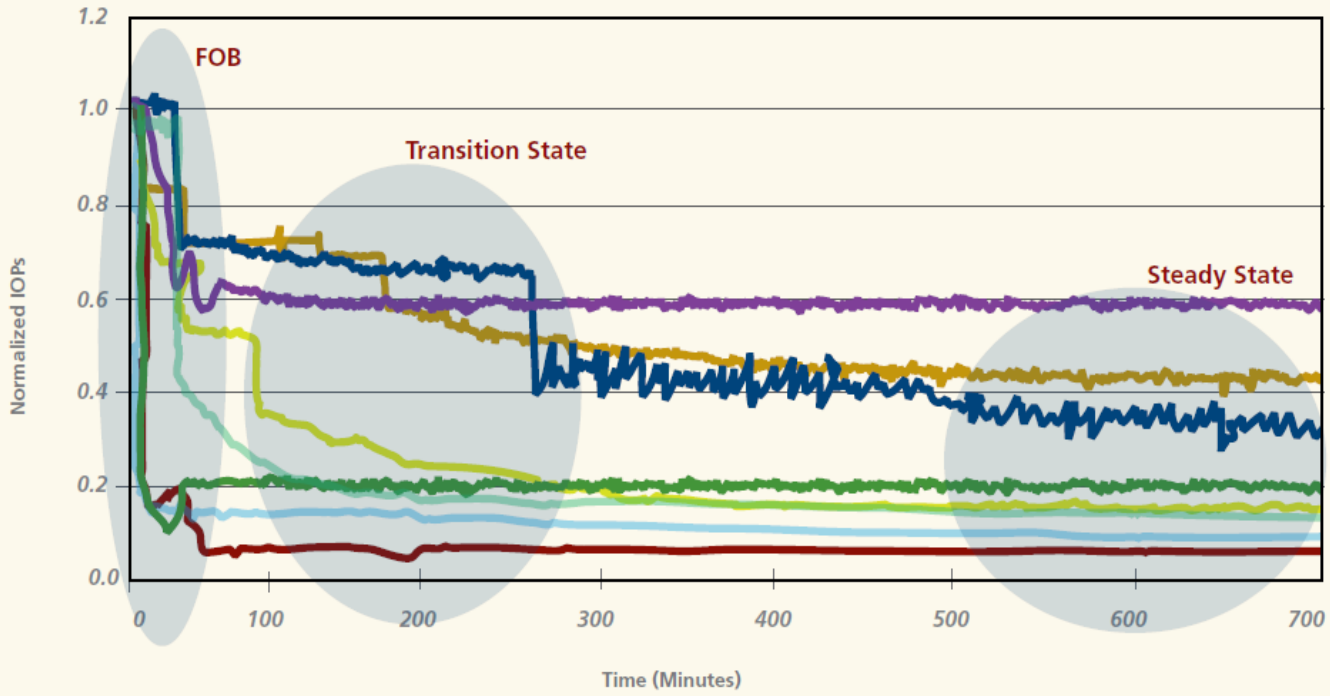
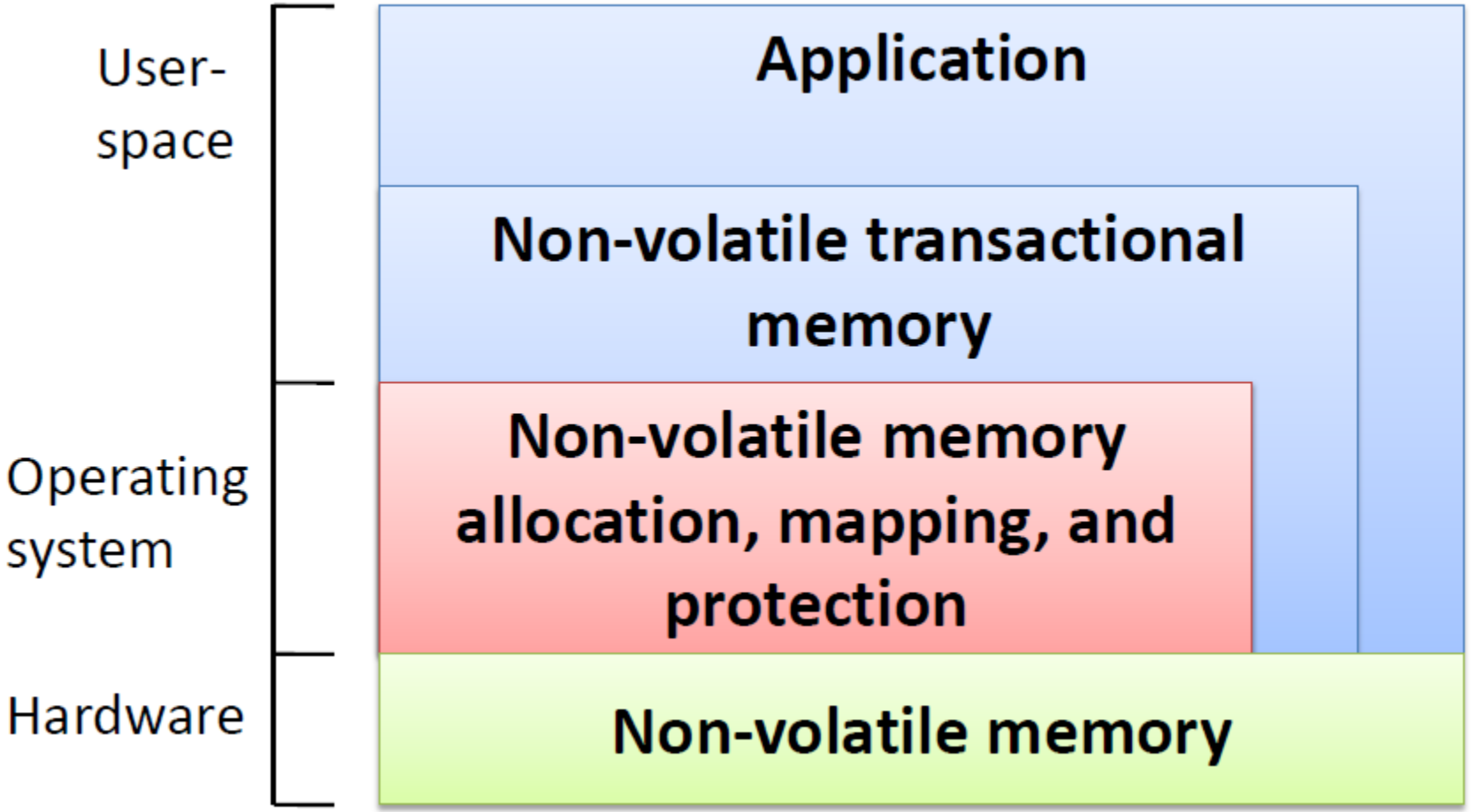


Figure 1: Performance States

Greenliant®

面向NVMe闪存设备的访问机制



Greenliant®

谢　　谢

Greenliant®