LightNVM: The Linux Open-Channel SSD Subsystem

`NVM` `Open-Channel` `SSD` `Linux`

Bjørling, M., González, J., & Bonnet, P. (2017, February). LightNVM: The Linux Open-Channel SSD Subsystem. In FAST (pp. 359-374).

背景

SSD在数据中心和存储阵列中得到了广泛使用，并且应用开始对高稳定的时延有较大的需求。传统的SSDs提供块接口的访问，为用户提供很高的存储抽象，但难以获得可预估的性能和最优的资源利用，不能很好的满足应用需求。本文提出在Open-channel架构下（为主机开放闪存内部逻辑）进行SSD管理的讨论。文章提出LightNVM, 可以提供基于物理页地址的SSD IO接口，向应用暴露SSD内部的并发和存储介质特性,并且LightNVM能够与传统的存储栈进行整合。结果表明LightNVM有较小的主机端开销，能够调整减小读延迟的变化性，从而获得可预估的IO延迟。

传统架构SSDs存在的问题

传统SSDs应用于数据中心存在一些问题：Log-on-Log[37,57],长尾延迟（large tail-latencies）[15,23],不稳定的IO延迟（unpredictable I/O latency）,资源利用率低（resource underutilization）。这些问题并不是由于硬件的限制，而是软件层对SSD进行类似HDD的管理造成的。[5,52]
Open-channel SSDs架构的提出，为主机软件层和SSD硬件层的协同设计提供了可能。

Open-channel SSD的管理

- NAND Flash的特性
  - 多种制造工艺。SLC/MLC/TLC/QLC 3D NAND
  - 存储介质特性。channel,chip,die,plane,block,page; 读写以page为粒度，擦除以erase为粒度。page包括数据区和OOB区（记录ECC等）。读写延迟不对称，并且互相之间存在干扰。
  - 写的限制。写操作尽可能写满一个page;块内的page需要顺序写；页重写之前需要进行擦除；可擦除次数有限。QLC>TLC>MLC>SLC
  - 多种错误模式。比特错误；读写带来的cell之间的干扰；数据保持问题；写/擦除错误；硬件Die出错。
- NAND Flash的不同管理机制
  - 写缓冲区。写缓冲区在主机端和设备里都可以存在；设备里的写缓冲区可以由主机端控制以避免二者之间的不协调（如设备控制器进行写出操作，主机端却要读）。
  - 容错机制。ECC，RAID/RAIN, data-retention处理。
- 一些经验教训
  - 设备可用保证。年或DWPD；如果P/E cycles由主机管理，厂商则无法提供保证。
  - 将存储介质的特性暴露给主机，让开发者陷入一些细节，如ECC或电压阈值调控是低效的，限制了存储介质的抽象。
  - 写缓冲区的利用取决于应用场景。DRAM缓冲区可由设备上的SRAM或其他非易失存储器代替，从而大幅度减少电量消耗。
  - 应用不可见的磨损均衡是必须的。
- Open-channel架构

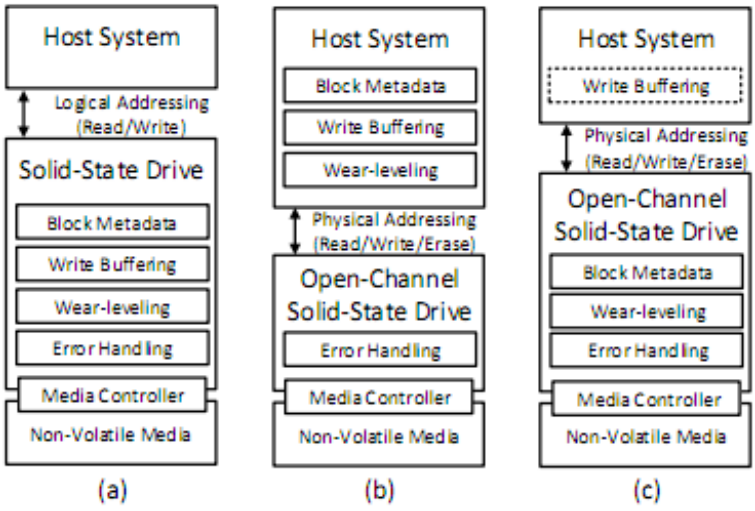

Figure 1: Core SSD Management modules on (a) a traditional Block I/O SSD, (b) the class of open-channel SSD considered in this paper, and (c) future open-channel SSDs.

基于物理页地址的I/O接口

- 地址空间。需要注意的是PU,也成为LUN，是内部并发的基础单位，每个channel包括多个LUN。

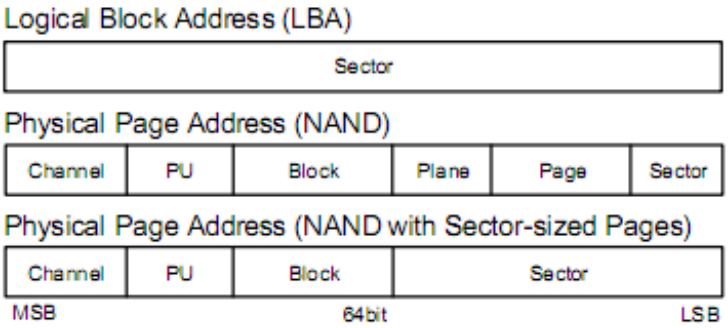

Figure 2: Logical Block Addresses compared to Physical Page Addresses for NAND flash.

- 几何布局和管理。LightNVM提供SSD的几何布局抽象，性能数据，介质页元数据（OOB）访问，COntroller的功能。
- 读/写/擦除操作。向量化的I/O操作，充分发挥闪存的并发性；支持不同的Plane操作模式，擦除/写操作延缓，有限的retry次数。

LightNVM 架构

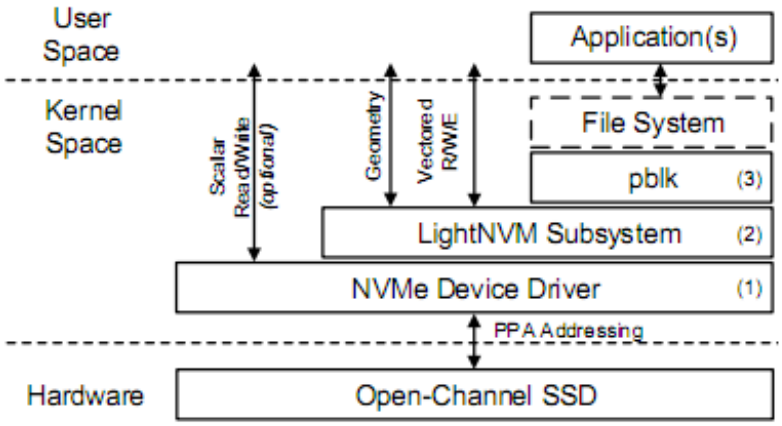

Figure 3: LightNVM Subsystem Architecture

- NVMe Device Driver. 通过IOCTL接口将设备以通用的Linux设备的形式提供给用户空间；

并提供给LightNVM以PPA I/O接口的操作。

- LightNVM子系统。提供设备的几何抽象和向量操作接口。
- 高层次的IO接口。pblk（Linux kernerl module）提供块接口；应用还可以创建接口访问LightNVM。

总结

LightNVM系统地对Open-channel SSDs 架构的设计问题进行了讨论，并且给出了一个标准化的实现，带动了Open-channel SSDs生态的发展； 另外还从实验的角度说明了Open-channel SSDs架构所呈现的低开销，可预计的IO延迟等。