未来的云存储：从Open Channel SSD到ZNS

Original 可可 SSDFans Yesterday

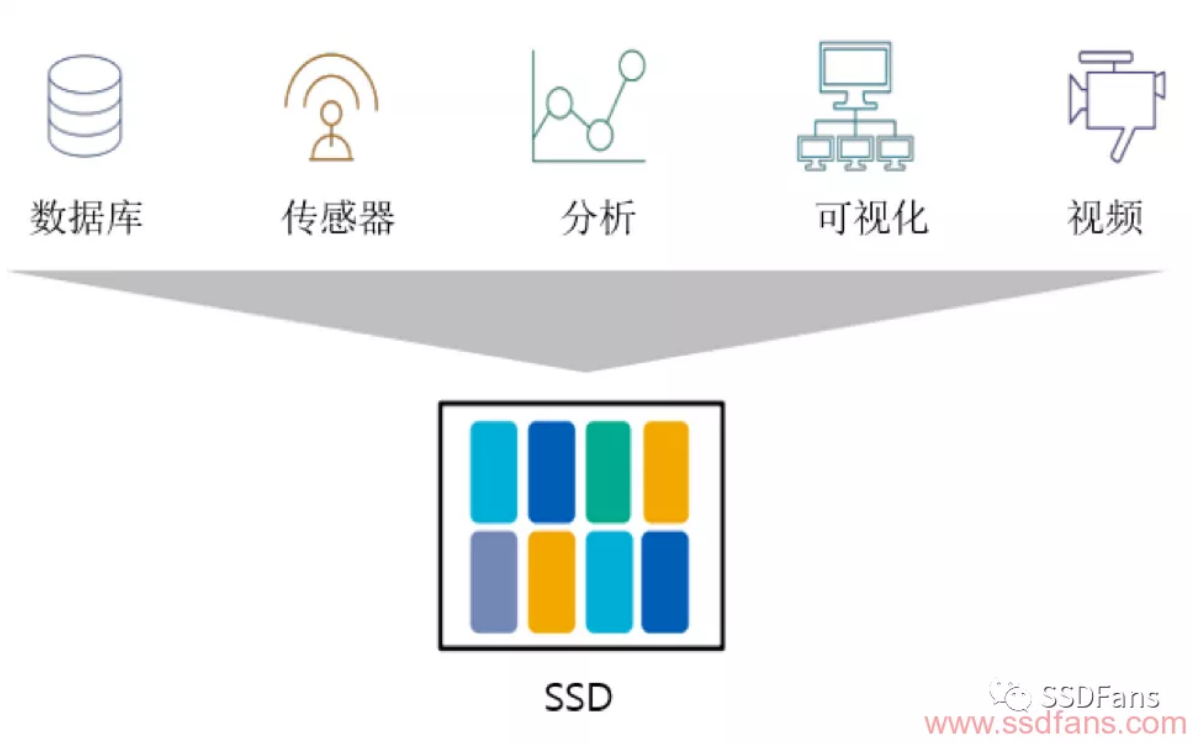
可可读OpenChannelSSD之六

可可读OpenChannelSSD之一_简介
可可读OpenChannelSSD之二_PPA接口
Linux竟然为Open Channel SSD开小灶?
管理闪存芯片，Open Channel SSD靠的是……
什么是Virtual OC SSD?

关于在GitHub上的Open Channel SSD 的开源项目好久没人更新过，在qemu搭建的平台上改进过，经常会出现bug, 对内核版本，qemu版本，系统版本要求相当高了。

虽然有很多论文已经发表出来了，但是讲真的觉得很多都是理论假设，并不能真的得到相应的实验结果，因此在学相关方面内容一定要看顶会论文，毕竟阿里的内部技术并不是公开的。

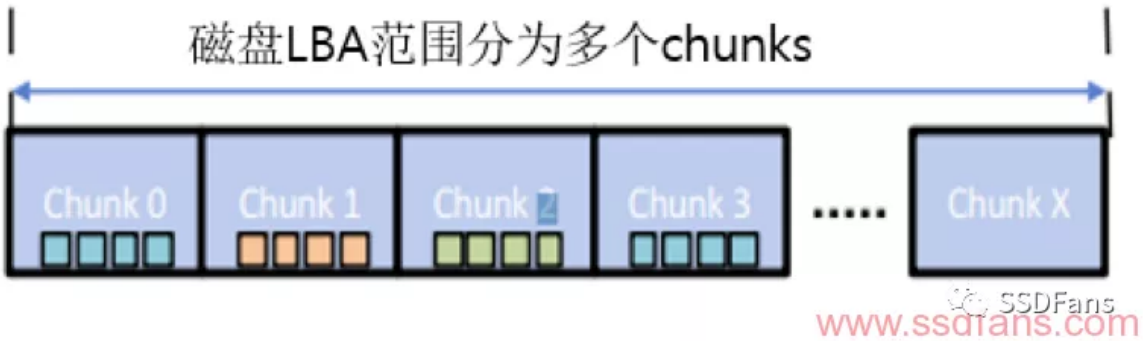
如今企业级要实现云存储效率需要单个SSD满足许多不同的工作负载，而工作负载现在可以说是无处不在。在应用共享SSD的时候，负载之间干扰造成延时忽高忽低，最坏时延巨幅升高。 保证为每一个硬盘用户提供稳定的服务质量，才能体现出云环境的服务质量。



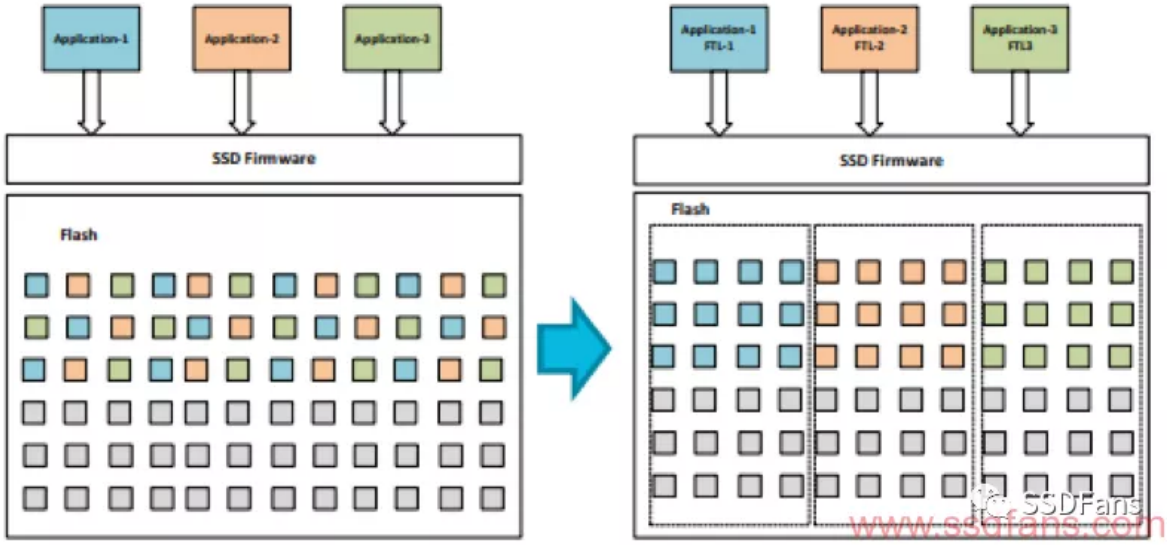
传统SSD把内部的FTL交给主机处理是Open-Channel SSD的主要功能，让用户自制属于自己的SSD。

Open-Channel SSD提出了chunk和PU的概念。

- Chunks特点：
  - 在LBA范围内顺序写入;
  - 需要重置才能重写;
  - 借鉴HDD的SMR规范（ZAC / ZBC）；
  - 针对SSD物理限制进行了优化:使写入与介质对齐



- Parallel Units特点：
  - Host可以对单独的工作负载进行direct I/O;
  - 单个或者多个die实现条带化;
  - 并行单元继承了底层介质的吞吐量和延迟特性;
  - 与NVMe中的I / O确定性相似的概念;



不难看出，Open-Channel SSD实现了I/O分离，可预测性延迟的特点，FTL功能移至Host端负责数据管理以及I/O调度。

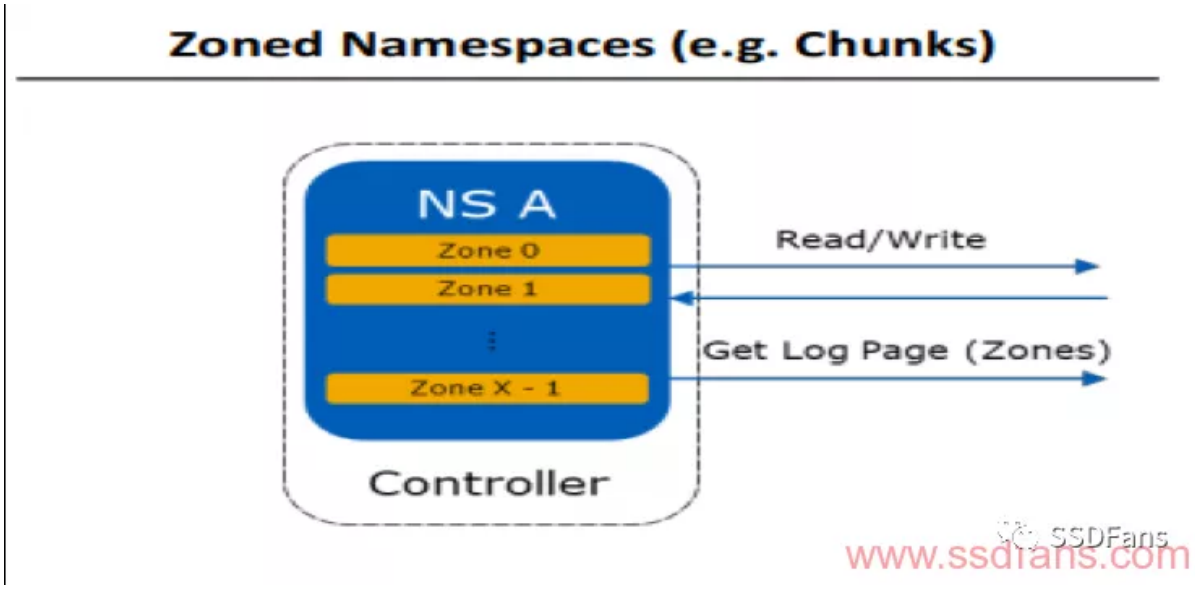
但是实际情况，Open-Channel Specification 仅仅定义了Open-Channel涉及的最为通用的部分。不同厂商的SSD产品特性不同，它们或者难以统一，对定制化应用和工作负载的需求，依旧欠缺灵活性。

关于Zoned Namespaces (ZNS)

采用Open-Channel SSD架构有阿里，微软等，将这个架构成为NVMe标准规范一部分的概念，提供灵活的定制化需求是一个热点研究。西部存储将功能驱动到解决关键OCSSD用例的NVMe中，提出了ZNS的概念。

- 它是NVMe工作组中的技术提案
  - 相对于正常的NVMe Namespace, Zoned Namespace将一个Namespace的逻辑地址空间切分成一个个的zone。Zone的基本操作有Read, Append Write， Zone Management 以及Get Log Page。

- 将zone接口标准化是为了：
  - 减少设备端的WAF;
  - 减少OP;
  - 减少SSD的DRAM，这是SSD中代价最高的部分;
  - 改善延迟和吞吐量;
  - 适用软件生态系统



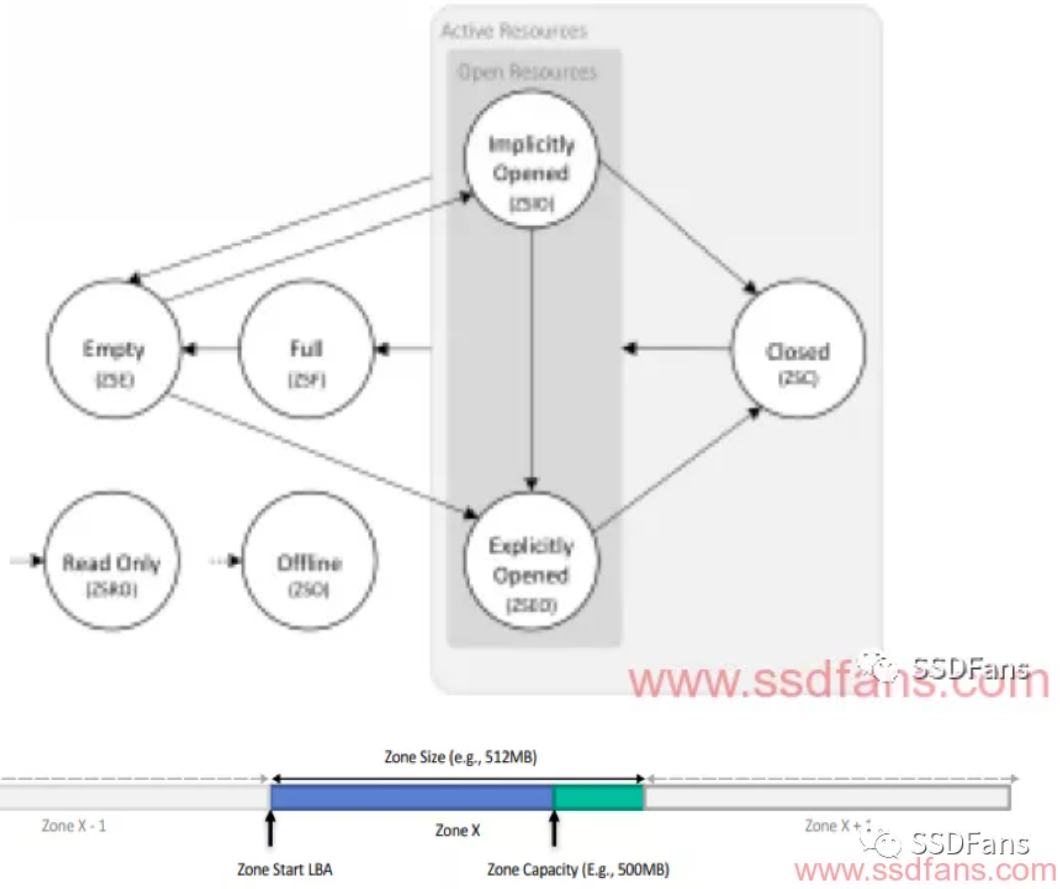
怎么来理解?
ZNS与SMR的ZBC / ZAC相似

- 存储空间被分成多个zone
- 每一段zone内都是顺序写入的
- 它是针对SSD优化的接口
  - 与介质特征保持一致（Zone的大小和Nand的块大小一致，Zone的容量与介质大小一致）
  - 减少NAND介质擦除周期

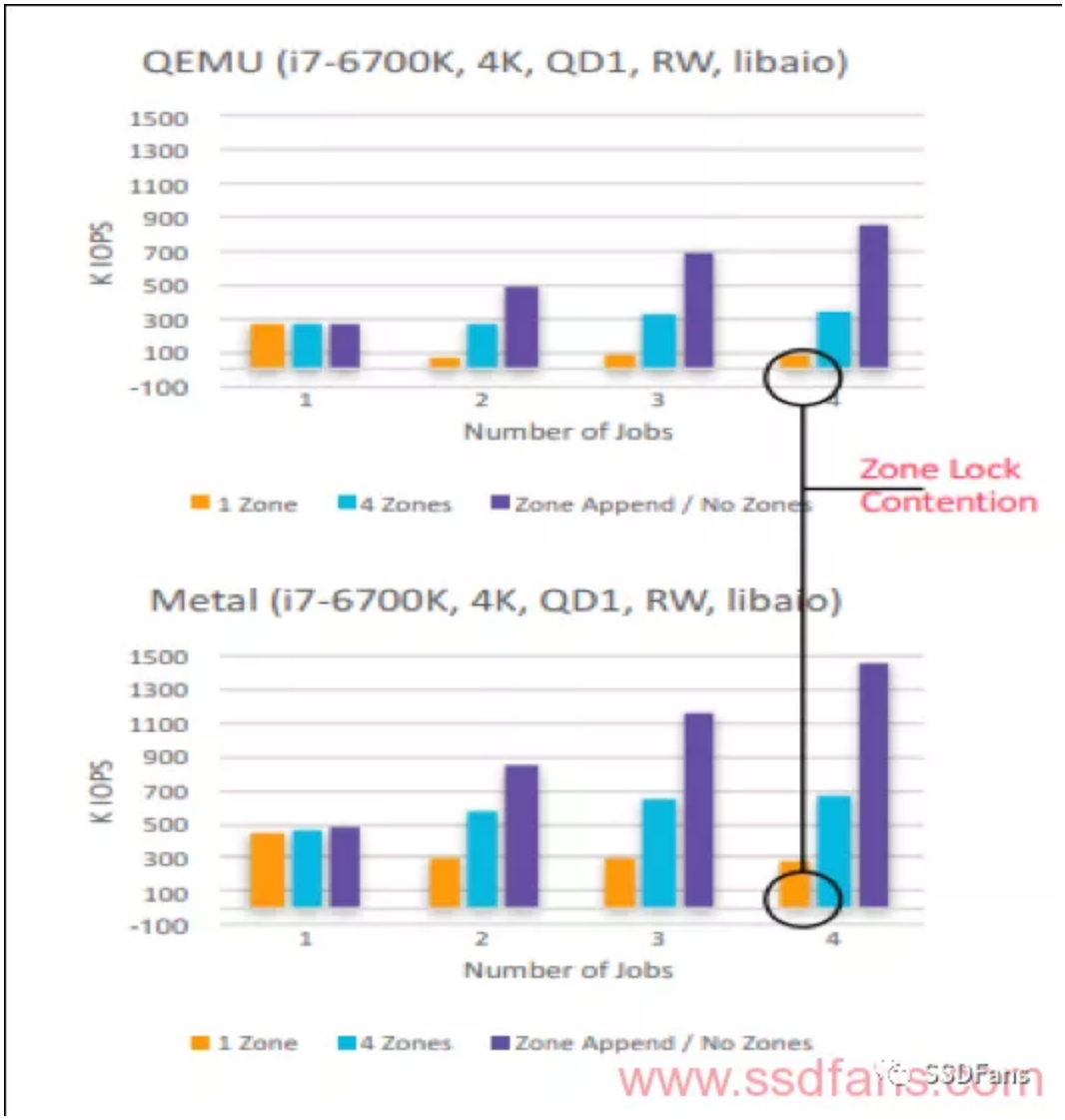
关于Zone的信息：

- Zone 状态转换
  - Empty, Implicitly Opened, Explicitly Opened, Closed, Full, Read Only, Offline
  - Empty -> Open -> Full -> Empty -> ....
- Zone Reset
  - Full -> Empty
- Zone 大小 和 Zone容量

- Zone 大小是固定的
- Zone容量是在一个Zone内的可写区域
- 与Open-Channel相比，最大的区别就是在Zoned Namespace中，Zone的地址是LBA（Logical Block Address, 逻辑块地址）Zone Namespace就可以避免Open-Channel里繁琐的各类地址转换。

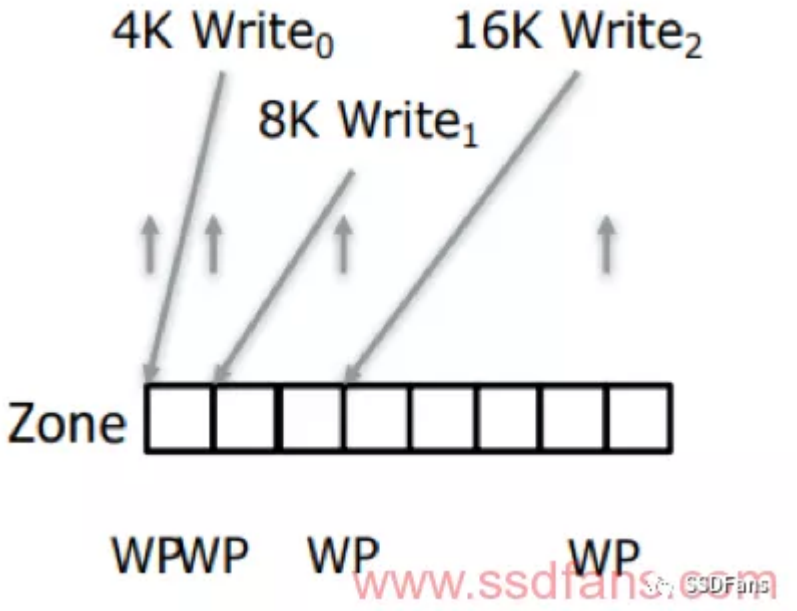


考虑到对一个区域（zone）多个写入的可伸缩性低（如下图为例），ZAC/ZBC 需要严格的写入顺序，限制写入性能还会增加host的开销，因此，软件生态系统，HBA等面临巨大挑战。

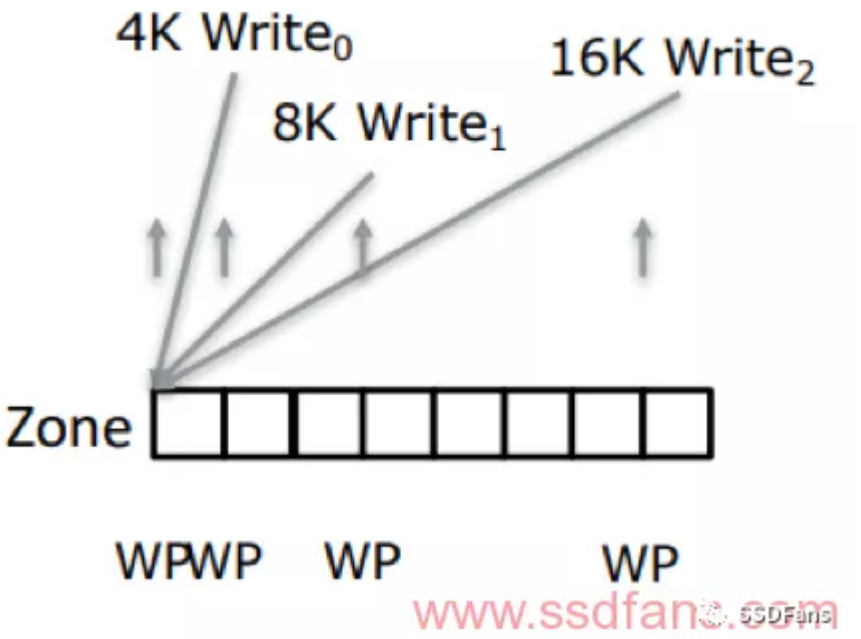


所以引入追加区域 (Zone Append)，将数据追加到一个区域而不定义偏移量，由驱动器返回将数据写入该区域的位置。

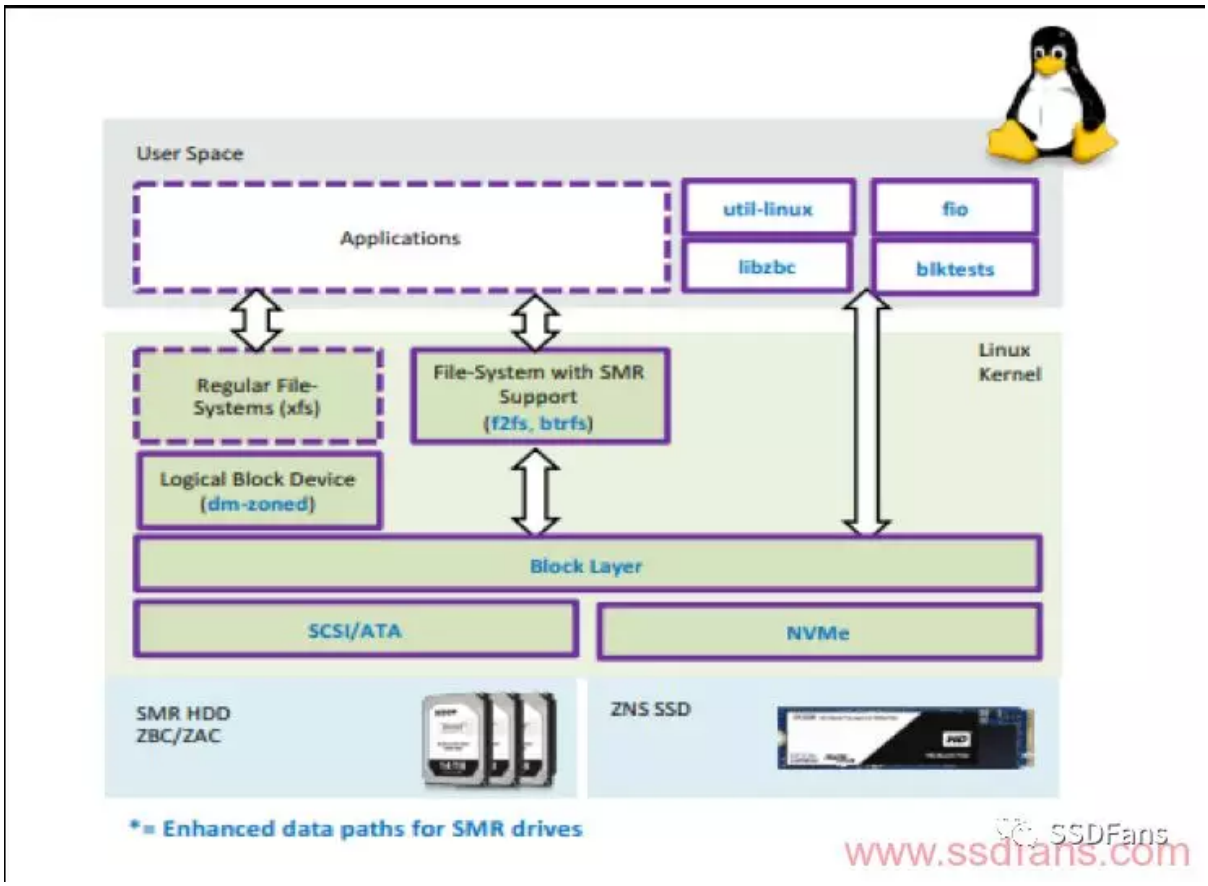
Zone Write 示例：
3x Writes (4K, 8K, 16K) – Queue Depth = 1



Zone Append 示例：
3x Writes (4K, 8K, 16K) – Queue Depth = 3



ZNS 指的是具有ZAC / ZBC软件生态系统的Synergy。现有的ZAC/ZBC-aware文件系统和设备映射都“工作正常”，支持ZNS只需要很少的更改；重用已应用于ZAC / ZBC硬盘驱动器（SMR）的现有工作；直接与文件系统集成（没有host-side FTL;1TB的介质设备不需要1GB 的DRAM; 能更好地利用SSD）；代码已经在大厂商生产中使用，并且可以在Linux生态系统中使用。



ZNS使用现有的存储堆栈：

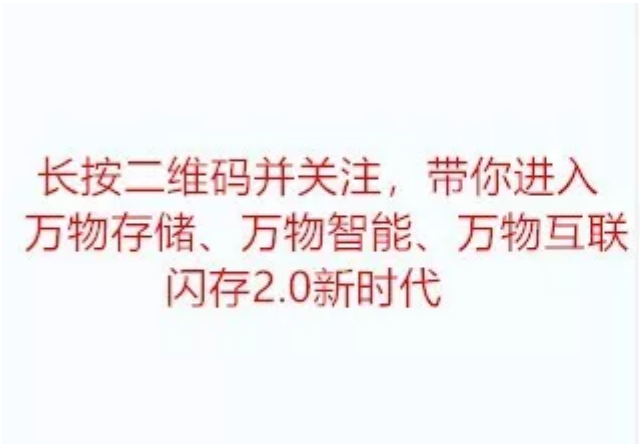
- 用户空间库
  - Libzbd
  - Nvme-cli
  - Blktests
  - Util-linux (blkzone)
  - fio
  - libzns
- 内核空间库
  - NVMe对Zones的支持
  - XFS, Btrfs, F2FS, dm-zoned, etc...
- 具有ZNS支持的Qemu

ZNS是为了满足多数应用对QOS及Latency需求的基础，然而却不如Open-Channel灵活。

高端微信群介绍

创业投资群	AI、IOT、芯片创始人、投资人、分析师、券商
闪存群	覆盖5000多位全球华人闪存、存储芯片精英
存储群	全闪存、软件定义存储SDS、超融合等企业级存储
AI芯片群	讨论AI芯片和GPU、FPGA、CPU异构计算
5G群	物联网、5G技术与产业讨论
第三代半导体群	氮化镓、碳化硅等化合物半导体讨论
存储芯片群	DRAM、NAND、3D XPoint等各类存储介质和主控讨论
汽车电子群	MCU、电源、传感器等汽车电子讨论
光电器件群	光通信、激光器、ToF、AR、VCSEL等光电器件讨论
渠道群	存储和芯片产品报价、行情、渠道、供应链

想加入这些群，长按或扫描下面二维码加nanoarchplus为微信好友，介绍你的姓名-单位-职务，注明群名，拉你进群。



Read more