面向闪存的存储系统构建与思考

姓名：谢丹华

学号：2017286190088

学院：测绘遥感信息工程国家重点实验室

目录

面向闪存的存储系统构建与思考

1. 闪存的发展与优势

1.1 闪存的发展趋势

在 1984 年，东芝公司的发明人舛冈富士雄首先提出了快速闪存存储器（此处简称闪存）的概念。与传统电脑内存不同，闪存的特点是非易失性（也就是所存储的数据在主机掉电后不会丢失），其记录速度也非常快。

虽然目前的价格还难以让用户接受，但是随着生产工艺的进步和产量的增加，其成本必然会大幅下降。相比传统的硬盘设备，Flash 硬盘将提供更快的平均存取速度，它不易损坏，有更小的体积，更轻的重量。随着技术发展，其容量会不断增大，数据分布更均匀，使用寿命加长，成为传统磁盘的有力竞争者。

1.2 闪存的优势

与传统硬盘相比，闪存的读写速度高、功耗较低。

1）闪存的体积小。并不是说闪存的集成度就一定会高。微硬盘做的这么大一块主要原因就是微硬盘不能做的小过闪存，并不代表微硬盘的集成度就不高。

2）相对于硬盘来说闪存结构不怕震，更抗摔。硬盘最怕的就是强烈震动。虽然我们使用的时候可以很小心，但老虎也有打盹的时候，不怕一万就怕万一。

3）闪存可以提供更快的数据读取速度，硬盘则受到转速的限制。

4）闪存存储数据更加安全，原因包括：

a. 其非机械结构，因此移动并不会对它的读写产生影响；

b. 广泛应用的机械型硬盘的使用寿命与读写次数和读写速度关系非常大，而闪存受影响不大；

c. 硬盘的写入是靠磁性来写入，闪存则采用电压，数据不会因为时间而消除。

5）质量更轻。

闪存及其他新型非易失存储器件与传统的磁盘和 DRAM 都有着相当大的差异，例如在易失性、寿命、读写性能、寻址、存储密度等方面表现出不相同的特征。

1.3 闪存在构建系统时的优势

1) 存储结构

闪存存储器相比于磁盘有延迟低和带宽高的特点。在带宽上，闪存设备中由于内部可实现不同级别并发，其带宽可高达数 3GB/s。在延迟上，闪存比磁盘的访问延迟低 6 个数量级，PCM 等持久性内存的访问延迟比磁盘低 12 个数量级；另一方面，闪存存储器体积更小，耗电更低，发热更少；在硬件接口上，由于传统 SATA 接口的速度限制不能发挥闪存的高并发性能，新的接口，例如 PCIe 和 DIMM 将被应用到高性能闪存设备上。

2) 系统软件

通过扩展现有的软件接口对系统进行优化，例如，使用 Atomic Write 来利用闪存的特性完成原子更新操作，使用 TRIM 来提供数据显示删除的语义，使用额外的数据标识来显示数据的冷热程度，等等。

3) 分布式协议

闪存的存储单元不可覆盖写，采用异地更新方式更新数据。异地更新方式天然地提供了数据多版本，利于实现数据一致性。相比传统系统软件的一致性实现，效率大幅提升。

2. 闪存对存储系统的变革

在近半个世纪里，磁盘占据着存储介质的主导地位。磁盘的机械式部件限制了其延迟和带宽性能的提升。现有存储系统中积累了大量针对磁盘机械式特性进行优化的设计。非易失性存储器与磁盘存储存在较大差异。非易失性存储器是电子式存储器件，数据读写无需机械式部件驱动，数据读写访问延迟可从毫秒级降至微秒甚至纳秒级。除此之外，非易失性存储在读写方式、读写对称性、耐久性等方面与传统磁盘存储相差较大。这些差异对现有存储系统的设计带来的巨大的挑战，包括存储体系结构、系统软件和分布式协议等方面。

面对闪存存储系统在应用中的上述挑战，研究人员致力于对现有的存储 结构、系统软件以及分布式协议进行变革。例如，研究去掉闪存设备上的 FTL (flash translation layer)，设计基于裸闪存的存储系统；优化闪存集

群上的数据 迁移策略，延长闪存寿命;研究闪存与特定应用负载特征相结合的软件定义存储等。

2.1 存储结构的变革

1) 固态盘

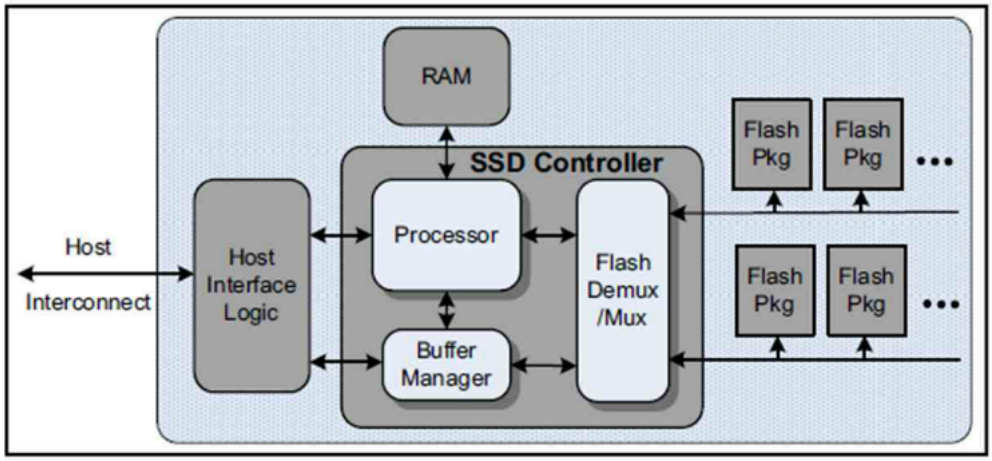

图 1 SSD 内部结构图

固态盘的方式是在设备内部通过闪存转换层(flash translation layer，FTL) 的转换可处理 SATA 命令，其外部的使用接口与传统磁盘没有差别。固态盘与传统磁盘在外部使用上差异极小，固态盘可以简单地替换磁盘，因而固态盘的形式为当前闪存存储的主要形式，包括个人笔记本、服务器、存储阵列等。

然而，固态盘的形式限制了闪存优势的有效发挥.在硬件接口上，由于闪存内部并发可以有效聚合访问带宽，SATA 接口的标准已经远不能满足要求，硬件接口成为闪存存储系统的瓶颈.在存储子系统上，以文件系统为代表的系统软件在存储管理上大多以磁盘为假设进行优化，较少考虑闪存特性，闪存优势难以得到充分利用。

2) PCIe 闪存卡

在主机端实现闪存转换层，能利用主机的冗余计算和缓存能力；利用异地的更新的特性，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。

将内部的设备信息开放给系统软件，使优化更具有针对性。

但它不可避免的也存在着一些问题：

难以实现存储容量的扩展和存储共享；相比于固态盘，体积和发热较大。

3) 闪存阵列

闪存加速卡难以实现存储容量的扩展及存储共享.类似于传统的磁盘阵列设计，多家存储厂商推出了闪存阵列，其中包括采用固态盘与传统阵列控制器的演进式设计，以及采用闪存芯片与全新阵列控制器的革新式设计.

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。

4) 基于闪存的分布式集群系统

FAWN (a fast array of wimpy nodes)是卡内基梅隆大学基于闪存介质构建的可扩展、低能耗、高性能的集群系统.与闪存加速卡和闪存阵列仅关注于 I / O 子系统的性能和可靠性的设计不同，FAWN 从集群整体设计的角度考虑闪存与处理器的匹配以降低系统整体能耗. FAWN 可提供每焦耳高

达 364 次查询请求，相比于桌面硬盘系统的每焦耳 1.96 次查询性能提升数百倍。

2.2 系统软件的变革

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，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

文件系统对文件的管理既包括文件目录树的维护信息，也包括对于文件逻辑块到存储设备物理块的映射，同时也包括存储设备空间管理。闪存设备内部需要提供地址映射以实现数据块的异地更新。文件系统中从文件逻辑块到设备物理块的映射，与闪存设备转换层中逻辑地址到物理地址的映射，构成了闪存系统的双层映射。双层映射的出现既增加了元数据的管理与存储开销，也可能造成双层的优化之间存在冲突。

2) 语义缺失

闪存设备转换层封装了对闪存介质的操作，向上层提供了统一的块设备接口。设备内部接收到以页面为单位的数据，却不能理解数据页面之间的关系，也就难以优化数据页面的读写顺序与分布。除此之外，闪存设备也难以感知文件系统的操作，不能及时处理，以至于影响下次关联操作。例如，文件系统对于文件的擦除操作仅修改对应元数据，而数据的擦除操

作直到下次写操作才被闪存设备感知，影响闪存设备的垃圾回收效率。

3) 特性缺失

闪存设备提供了异地更新的特性，数据的更新会写入新分配的闪存页，旧闪存页一直保留到垃圾回收的时刻。而文件系统、数据库管理系统或者其他上层应用为提供操作的原子性，通常使用 WAL(write ahead logging)的方式，先申请新的存储空间记录数据更新作为 redo 日志，然后在原地进行数据的更新。这样的原子性以两次写操作为代价保证，既增加了延迟，也损害了闪存寿命。同样，文件系统为容错而采用多版本信息也不能直接索引到旧的闪存页。

随着存储介质访问延迟越来越低，软件开销所占比例越来越高。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，预计软件开销比例降高达 94.09%。

1) 通知机制

随着闪存设备的延迟越来越低，传统基于中断的软硬件通知机制开销增大，低延迟访问造成的频繁中断，上下文切换代价已经超出了循环等待的代价。

2) 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。

3) 软件接口

在软件接口上，传统的基于块的 10 接口，由于缺乏丰富的语义支持，上层软件不能发挥闪存的特性，例如异地更新、内部并发等;而闪存设备也不能对 10 请求做优化，例如有效的冷热分组、高性能的请求调度等。一些研究工作通过扩展现有的软件接口对系统进行优化，例如，使用 Atomic Write 来利用闪存的特性完成原子更新操作，使用 TRIM 来提供数据显示删除的语义，使用额外的数据标识来显示数据的冷热程度，等等。

4) 文件系统

传统文件系统过多地针对磁盘设计。一方面，文件系统的自身设计与磁盘特性密切相关，例如，传统文件系统大多假设文件系统的随机读写性能较差，继而采取缓存、预取、数据分组、碎片整理等机制，甚至用对磁盘磁道对齐等方式提高系统性能，这些优化对闪存作用有限，甚至额外的软件管理引入造成数据写放大、处理延迟增加等问题。

另一方面，文件系统与闪存设备的自管理机制，在很多方面产生了功能冗余， 例如重复的空间管理、垃圾回收和地址映射等，这些冗余会造成元数据管理的双倍开销、文件系统优化失效以及额外的写放大等问题。

针对这些问题，面向闪存构建文件系统也是近年来的研究热点。

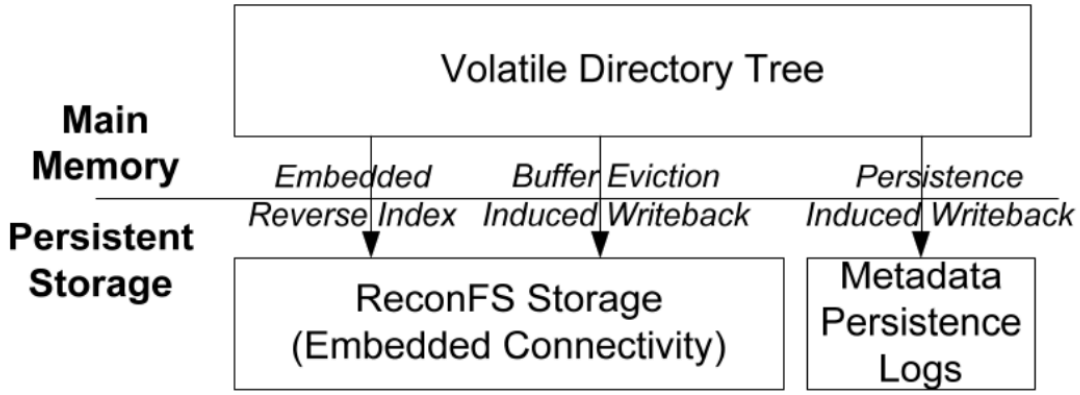

图 2　可重构文件系统 ReconFS

ReconFS 是清华大学研制的对目录树管理进行优化的闪存文件系统。ReconFS 是闪存颠覆了传统文件系统目录树构建的一个例子。闪存文件系统命名空间(目录树)构建了文件系统的结构，但目录树结构的维护引入了较大的元数据开销，具体表现为写入频繁和分散小写的特征。元数据的频繁、分散小写不仅损害了存储系统的性能，也加速了闪存磨损，降低了闪存系统寿命。针对此问题 ReconFS 提出了新型文件系统命名空间的设计，并基于此实现了可重构文件系统 ReconFS。可重构文件系统 ReconFS 通过分离易失性目录树与持久性目录树的管理，并通过嵌入式索引机制与日志持久化机制，提高了闪存文件系统目录树的一致性与持久性。在系统意外掉电时，由于闪存存储的高 IOPS 与高带宽，可重构文件系统 ReconFS 在数秒内完成文件系统目录树的重构。

2.3 分布式协议的变革

与磁盘相比，闪存具有随机访问速度快、访问低延迟、读写不对称以及有限的寿命等特点，给分布式协议的设计带来了革新的机遇。

一部分研究工作，致力于对现有技术的改进，例如改进分布式存储系

统上 客户端的缓存机制。

另一部分研究工作，致力于设计新的协议以发挥闪存设备的特性。CORFU 利用多闪存卡构建了低延迟、高带宽的共享日志型、分布式存储系统。通过将闪存设备直连到网络，CORFU 去除了服务器，降低了集群复杂性、访问延迟和能耗。Tango 则实现了基于 CORFU 的共享数据结构，能保证集群系统中元数据操 作的一致性、持久性、原子性和隔离性。

3. 总结与思考

软件系统的设计是发挥闪存优势的关键。高效的软件设计既是闪存软件系统中所必需的，也是其中的难题与挑战。

软件管理已经成为全闪存阵列中中的核心。包括数据删冗、数据压缩以提升软件的寿命；全局垃圾回收以解决低延迟的问题；提供数据副本保证软件的高可靠性；冷热分组以延长软件寿命等管理机制。软件系统设计的矛盾已经转化成低的软件开销和复杂的软件管理功能之间的矛盾。

系统构建的变革可能是颠覆式的。相对本地文件系统或操作系统的功能冗余、特性缺失、频繁中断、应用失配等缺陷，分布式系统从低延迟、高宽带到共享访问再到新的分布协议如 CORFU 和 TANGO 的出现实现了一系列的蜕变。

所以，结合 Flash 和硬盘两种技术优势的产品的混合式硬盘需要应运而生。混合硬盘里面既有如今标准硬盘使用的传统磁盘，也有闪存芯片。闪存芯片用来存储一些需要快速读与写的数据，其他要求不高的数据被写

入到磁盘，这意味着可以迅速读取数据，不必等待硬盘工作。将大大提高存储效率。

另外，存储的数据分布与访问模式将越来越收到关注。在存储技术发展中，一个基本问题是数据在存储介质上如何分布，能提高数据的访问速度。同样，什么样的访问模式，可以提高数据访问的速度。这些问题都是业界应该关注并解决的问题。

总之，闪存存储器对存储结构、系统软件和分布式协议的革新是接近于颠覆性的。这些颠覆性的研究对于存储系统，乃至计算机系统，将产生较大的变革。