IBM FlashCore™ Technology

Silverton Consulting, Inc.

StorInt™ Briefing

引言

IBM FlashCore™ 技术创新是 FlashSystem 900 的基础，FlashSystem 900 也因此而成为当今市场上最快、最可靠的全闪存存储系统。随着 NAND 技术的不断发展，闪存存储系统也必须与时俱进。举例来说，IBM 近期与 Micron Technology 签署了一项协议，藉此 IBM 将可使用全新的多级单元 (MLC) NAND 技术，这将有助于提升密度配置并进一步降低存储成本。

IBM 已对 FlashCore 技术进行了重新设计，旨在最大程度地提升闪存的 I/O 速度，同时提供最可靠的闪存存储解决方案。FlashCore 技术是 IBM FlashSystem 900 存储产品下列三大基础组件的重要组成部分：

- **硬件加速的架构。**FlashSystem 900 在控制器中采用 IBM 开发的硬件，可最大程度地减少甚至消除 I/O 活动中的软件交互量，从而为全闪存存储阵列提供具有最高 I/O 性能及最快响应时间的控制器。
- **BM MicroLatency™ 模块。**FlashSystem 900 采用 IBM 专有的闪存存储模块，可提供性能最快和最可靠的闪存存储。
- **高级闪存管理。**FlashSystem 900 采用 IBM 硬件及专利软件算法，确保更好地管理原生 NAND 的可靠性，从而提供最可靠、最具可用性的闪存数据存储。

图 1 IBM FlashSystem 900

硬件加速的架构

其他供应商的闪存存储系统都存在的一个共性问题是，存储控制器的软件会经常由于 NAND 存储 I/O 过快而导致 I/O 速度下降。相反，FlashCore 技术可最大程度地减少（甚至完全消除）I/O 活动过程中的任何软件交互。这一设计是 FlashSystem 900 的首创。

专为闪存而设计

目前市场上的所有闪存存储系统最初几乎都专为旋转式存储媒介而设计。旋转式存储媒介在访问数据块时耗时长达几十毫秒，这会留出大量的多余指令执行时间，用以设置和终止 I/O 操作。另一方面，NAND 处理 I/O 操作仅需数微秒即可完成，留出的指令执行或处理时间非常短暂。

FlashCore 技术从最初开始便是专为使用闪存存储而设计；因此从封装、数据路径、硬件选择到软件功能，均围绕闪存存储速度而设计。若要充分利用闪存的速度，相比基于磁盘的存储，需要更多的硬件功能。

硬件 RAID

IBM FlashSystem 900 可提供系统级的硬件 RAID，为可能影响整个闪存存储模块的故障提供额外的数据保护。专门设计的硬件 RAID 可在写入操作过程中提供快速的奇偶生成，并在重建操作过程中提供快速的奇偶使用。一旦某个闪存控制器或闪存存储模块完全失效（不过出现这种情况可能性非常小），借助系统级 RAID 5 的功能，IBM FlashSystem 可快速在其内部的热后备闪存模块上重建不可存取的数据。FlashCore 技术的硬件 RAID 还可提供更多种类的 RAID 布局，可始终支持“3 数据 + 1 奇偶”的存储配置，甚至高达“10 数据 + 1 奇偶”的存储配置。

FlashSystem 900 的硬件 RAID 可提升客户写入操作速度，并提供更快的闪存存储重建。FlashSystem 900 的 RAID 分组更小，因此可提供更广泛的存储容量选择，便于客户根据其应用需求轻松定制闪存存储。

无闭塞纵横制交换

目前的大多数存储系统都采用外围装置互连高速 (PCIe) 总线来访问前台连接所用的主机总线适配器 (HBA)，同时采用串行 SCSI (SAS) 控制器来控制存储的后台连接。与这些存储系统不同的是，基于 FlashCore 技术的控制器围绕专用的冗余式无闭塞纵横制交换基架而构建。通过纵横制交换功能，可为 FlashSystem 900 在每个主机 I/O 界面与每个闪存存储模块之间提供直接的数据路径。FlashCore 的纵横制交换功能，与 PCIe 相比可提供更高的内部数据带宽，如此便无需在其他 I/O 操作完成时等待，从而允许更多的并行和并发 I/O 活动。该功能可消除其他闪存存储系统所必需的总线开销。IBM FlashSystem 900 的每个存储控制器中均配有一个纵横制交换基架。

借助 FlashCore 技术的纵横制交换功能，FlashSystem 900 可同时执行更多的 I/O 请求，且每个闪存 I/O 操作的耗时更少。因此，仅需更少的 IBM 闪存存储便可提供与其他供应商的全闪存存储阵列同等的性能。

硬件数据路径

FlashCore 技术采用硬件控制器数据路径。非 IBM 闪存存储系统使用软件执行来启动、监视和终止数据传输，而 IBM FlashSystem 900 则使用专门设计的专用现场可编程逻辑闸阵列 (FPGA) 硬件。如上所述，非 IBM 闪存存储系统通常采用通用的 Intel/x86 处理器指令执行来管理数据传输活动，不仅耗时且会增加 I/O 延迟。

此外，FPGA 还会生成并确认**数据路径校验**。FlashCore 可在内部将数据路径校验附加于传输的所有数据之上，确保系统可快速识别并纠正各种数据传输错误。

借助 FPGA 托管的控制器数据路径，FlashCore 技术可实现业内最低的闪存存储 I/O 延迟。[1] 通过 FPGA 数据路径硬件与无闭塞纵横制交换功能，IBM FlashSystem 900 存储系统相比其他非 IBM 闪存存储系统，可在更少的时间内执行更多的 I/O 活动。

[1] 请参阅 Silverton Consulting 发布的有关存储基准性能调查结果的最新 Storage Intelligence 报告，其网址为：http://silvertonconsulting.com/cms1/dispatches/。

单箱高可用架构

借助 FlashCore 技术，闪存存储系统的可用性和可维修性迈上了一个新台阶。举例来说，当前的 FlashSystem 900 存储系统是一款全模块化的存储解决方案，所有的关键非被动组件均包含在现场可更换单元 (FRU) 或模块之内。在 FlashSystem 900 中，下列组件采用完全冗余式设计，可在需要时进行热插拔：

- **闪存存储模块**。FlashSystem 900 MicroLatency 模块可通过单元前端访问。一旦出现故障，可轻松更换这些模块，而不会影响存储操作。

图 2 IBM FlashSystem 900 后视图

- **两套界面、RAID 控制器、基架和管理控制器**。FlashSystem 900 采用冗余式控制器 FRU 或容器，所有这些组件均包含其中。控制器容器可从系统后部访问且可热插拔，确保不中断的持续可用存储操作。
- **双电源、电池和风扇模块**。一旦发生故障，可访问并热插拔冗余电源、电池和风扇，而不会影响系统运行。

FlashSystem 900 继承了 IBM 高可维修性全闪存存储产品可长时间运行的优势，确保在硬件故障维修期间，缩短维修时间并减少应用中断。

并发代码加载与维护

FlashCore 技术的先进功能支持无中断代码升级及其他软件维护活动。换句话说，除了在上述可用性和可维修性方面的提升，FlashSystem 900 还可实现**无中断（并发）的代码加载**，可在更新或修改系统代码时确保数据与 I/O 的可存取性。

与上述的高可用性和可维修性硬件类似，借助 FlashCore 技术的并发代码加载功能，FlashSystem 900 可在软件升级期间继续确保 I/O 存取的运行。借助高可用性的控制器硬件及并发代码加载功能，FlashSystem 900 存储系统可在所有硬件维修或代码更改期间始终保持在线，确保在其维护期间应用可继续使用。

IBM MicroLatency™ 模块

在 FlashSystem 900 中，IBM 设计了 MicroLatency 闪存模块，用以在控制器级补充硬件加速的架构，从而为当今市场上 NAND 数据存储领域提供了最短的 I/O 响应时间。

IBM 并未采用标准的 SSD 存储，而是设计了专有的 MicroLatency 模块。这些模块使用业内的标准 NAND 芯片，但可提供相比其他 SSD 存储更高的 I/O 性能。

此外，通过专门设计的闪存存储模块，IBM 还可提供更高密度的存储，且相比标准 SSD，闪存存储更多，同时单个软件包中的闪存芯片控制器更多。通过这些独特设计，FlashSystem 900 的客户仅需更少的机架和占地面积，便可获得更高的 I/O 性能。

并行设计

每个 5.7 TB 的 MicroLatency 模块配有 4 个控制器和 64 个闪存芯片（每个控制器对应 16 个芯片）。FlashCore MicroLatency 模块的每个闪存芯片和闪存控制器可支持多个并发操作。

图 3 IBM FlashSystem 900 闪存模块

每个闪存控制器最多可并行对其闪存存储进行 40 次直接内存存取操作；若扩展至完整的 FlashSystem 900 存储系统，在 57 TB 系统中最多可同时实现 1,760 次的 NAND 存取操作。 因此，即便在存在大量读写 I/O 工作量的情况下，FlashSystem 900 存储系统也能维持高速的 I/O 性能。

当从每秒单次 I/O 操作增加至每秒数百万次 I/O 操作时，FlashSystem 900 的客户可以获得同样高速的 I/O 响应时间。而其他的全闪存存储系统，若要在如此大范围的 I/O 活动中实现同等响应速度，必然会增加 I/O 延迟，且需要更多的 SSD 及更多的闪存控制器和闪存芯片。

数据路径中的 FPGA

FlashCore 技术的硬件数据路径可全程对 MicroLatency 模块进行扩展。换句话说，MicroLatency 中的 I/O 数据传输操作由专用的 FPGA 处理，而无需依赖于通用的微处理器指令执行。借助控制器中的 FPGA 以及 MicroLatency 模块中的 FPGA，即便在超大负载条件下，FlashSystem 900 存储系统也可提供超低延迟的 I/O 性能。

因此，相比其他全闪存阵列存储系统，FlashSystem 900 存储系统可更快完成 I/O 操作，且响应时间非常稳定。

分布式 RAM

FlashCore 技术不使用基于 DRAM 缓存的传统控制器。在以往，企业存储系统一直使用控制器级的 DRAM 缓存作为待存取数据的暂存区并用于存储频繁存取的数据，以便实现相比磁盘或 SSD 读写数据更快的 I/O 性能。 FlashCore 技术存储系统没有磁盘或 SSD；相反，FlashSystem 900 存储系统能以与当今企业存储系统中 DRAM 缓存相当的速度运行。不过，FlashCore 技术使用位于 MicroLatency 模块级的分布式 RAM 来存储元数据，诸如查找表格及闪存寻址和转换活动所需的其他信息等等。MicroLatency 模块还使用此类分布式 RAM 中非常少量的一部分作为输入数据的写入缓冲器。

相比其他全闪存阵列，FlashCore 技术系统可使用更少的 DRAM 实现同等效果。此外，FlashSystem 900 系统在更新和管理控制器级缓冲层时，无需额外的处理时间或开销，这意味着 FlashCore 技术系统使用较少的指令执行便可处理每次 I/O 操作，从而提升 I/O 活动的速度。

高速界面

MicroLatency 模块与专用的高速纵横制基架采用高密度针脚连接。 这种连接方式可消除 MicroLatency 模块级 I/O 活动的任何总线指令处理开销或串行传输延迟，同时也是对 FlashSystem 900 控制器级纵横制交换的高并行性能的补充。

借助 FlashCore 技术的控制器纵横制交换功能，MicroLatency 模块的高密度针脚连接可实现更快的数据传输，同时允许从主机界面到 NAND 芯片的传输全程进行更多的并发 I/O 操作，反之亦然。藉此，IBM FlashSystem 900 的客户便可从其闪存存储容量中获取最高级别的性能。

线速静态数据加密

每个 MicroLatency 模块中的专用芯片可提供 AES 256 基于硬件的静态数据加密。硬件加密和解密以内部数据路径线速进行，且运行过程中不会影响 I/O 延迟。这样，客户可使用静态数据加密功能对其 FlashSystem 900 存储系统中的信息进行加密，而不会导致任何性能降级，从而更轻松部署闪存存储的数据安全设施。

高级闪存管理

除了硬件加速的架构及 IBM MicroLatency 模块外，FlashCore 技术还通过专门设计的硬件和专利算法大大延长了 NAND 内存的使用寿命，从而显著提升了 NAND 的可靠性。借助这些技术，IBM FlashSystem 900 可以称得上是市场上最可靠的闪存存储系统。

IBM Variable Stripe RAID™

FlashCore 技术在闪存芯片级采用“可变条带 RAID (Variable Stripe RAID)。可变条带 RAID 是指通过使用 IBM MicroLatency 模块内部的专用闪存控制器，在 NAND 内存芯片之间采用“15 数据 + 1 奇偶”配置的 RAID 5 实施（旋转式奇偶）。 一旦出现闪存（芯片或子芯片）故障，便会在之前预留的（超容量）存储区重建数据，受影响的 RAID 条带会收缩为“14 数据 + 1 奇偶”（或“13 数据 + 1 奇偶”、“12 数据 + 1 奇偶”等）RAID 分组。收缩 RAID 分组条带大小的方法是 IBM 的业内首创，可更好地保留闪存存储的可用性，而对数据保护或系统功能几乎没有影响。

由于很多竞争对手的闪存系统在模块或 SSD 中并无 RAID 保护，因此，可变条带 RAID 闪存芯片或子芯片数据保护要优于当前的行业实践。在模块中仅使用系统级 RAID 5 的竞争对手不必保持闪存容量和性能以及可变条带 RAID。

FlashCore 数据保护的两个组件（MicroLatency 模块级的可变条带 RAID 及系统级的硬件 RAID）可相互独立运行，但若相互结合，可提供协同式系统容错功能，以修复多个闪存内存故障。此外，借助专为可变条带 RAID 预留的空间及系统级 RAID 专用备件，即便在出现闪存故障时，可用系统容量亦不会减少。

IBM 专有的 ECC

IBM FlashCore 存储系统使用强大的**错误校正码 (ECC)** 算法，可对在闪存内存中存取的数据提供保护。对于每个新一代的 NAND 技术，制造商至少需要最低水平的 ECC 算法，才能满足其闪存可靠性规范。相比其 NAND 供应商所需的 ECC 算法，FlashCore 技术实施更强大的 ECC 算法，可实现更高的闪存可靠性。

此外，借助 IBM FlashCore 的特定创新，可使用硬件而非软件处理大多数 ECC 活动。许多系统依赖于 ECC 硬件检测，但可能会使用软件功能来纠正位错误。不过，采用软件纠正的方法在修复位错误时需要更长的时间，较高密度的 NAND 芯片更容易出现这种情况。借助采用 FlashCore 技术设计的硬件 ECC，FlashSystem 900 存储系统可充分利用这种高密度但更易失的 NAND 内存，但不会导致过度的性能降级。

因此，IBM FlashSystem 900 的客户不仅可以获得最新一代 NAND 技术成本更低、密度更高的优势，还可以获得高 I/O 性能的闪存存储功能。

IBM 优化的超容量算法

FlashCore 技术在用户可存取的数据空间之外引入了额外的预留闪存容量。IBM FlashSystem 900 使用该**超容量 NAND 空间**作为闪存单元失效时的备用系统容量。此外，NAND 内存技术仅可将数据写入或编程到已擦除的数据块中，而且无法直接重写；但通过超容量算法，IBM FlashSystem 900 存储系统可提供更多的已擦除 NAND 内存块，以供数据写入之用。大多数闪存存储系统在闪存模块中使用同一级别的超容量算法。

借助经 FlashCore 技术优化的超容量算法，FlashSystem 900 可实现高可用性的闪存存储及更快的写入 I/O 性能。从完全空白的系统到写满客户数据的系统，FlashSystem 900 均可维持高写入性能。

耗损均衡

FlashCore 技术还使用**耗损均衡**算法在系统内的更多闪存内存中分布写入活动，以避免由于单个位置写入活动过于频繁而导致 NAND 芯片过早抹除。在 NAND 内存耐久性有限的前提下，任何闪存存储解决方案必须将写入活动或编程/擦除循环分布在尽可能多的 NAND 位置上。

借助 FlashCore 耗损均衡算法，连同之前所述的经优化的超容量算法，便可充分利用更多的 NAND 存储空间，从而更好地维持 FlashSystem 900 闪存存储系统的寿命周期。

写入缓冲器与硬件卸载

在结合耗损均衡算法的基础上，FlashCore 技术还采用了专门设计的硬件闪存转换层，以便将新数据块依次写入闪存内存中的邻近位置。此外，写入 MicroLatency 模块分布式 RAM 之中的数据还可通过硬件处理写入到闪存内存中，而无需使用软件功能。

因此，在将数据写入到 NAND 内存位置的整个过程中，FlashCore 技术均可在控制器级和 MicroLatency 模块级完成所有的硬件托管数据传输。借助 FlashSystem 900 存储系统，客户便可尽可能快速地进行写入操作，确保最高水平的写入性能。

IBM 垃圾回收

FlashCore 存储技术还包括 IBM 专有的**垃圾回收**、**重定位和块挑选**算法，不仅可提升闪存的耐久性，还可以降低写入延迟。大多数闪存存储垃圾回收算法采用对称设计，所有的数据块和存取操作会采用同一处理方式。FlashCore 技术更进一步，使用详细的 NAND 块特性数据来确定每个数据块的健康情况，并将之与接下来的写入活动相匹配。IBM FlashSystem 900 的垃圾回收算法能够考虑多个属性，以减少过多的写入活动（写入扩增）并尽可能延长每个 NAND 数据块的寿命。

结合优化的超容量算法及耗损均衡算法，IBM 的高级垃圾回收算法不仅可提升 MicroLatency 模块的闪存耐久性，还可为 FlashSystem 900 存储系统提供非常高的写入 I/O 吞吐量。

总结

IBM FlashCore 技术的高级硬件及专有算法旨在提供当今市场上速度最快、性能最可靠的闪存存储系统。此外，IBM FlashSystem 900 存储系统还采用了业内最领先的 NAND 技术，以期为客户提供最具成本效益的闪存存储系统。

为了借助当今的 NAND 技术实现性能更高、可靠性更强、成本更低等综合优势，IBM 现已推出了 FlashSystem 900 这款硬件更为密集的存储解决方案，从控制器到 MicroLatency 模块，再到闪存芯片控制器等等不一而足。尽管所有硬件均需特别设计，但借助目前最高密度的 NAND 技术，就像 IBM FlashSystem 900 的硬件密集型设计已经实现了性能与可靠性的最大化。