IBM FlashCore™ Technology

Silverton Consulting, Inc.
StorInt™ Briefing

Introduction

IBM FlashCore™ technology innovations are fundamental to the FlashSystem 900 and help to make it the fastest and most reliable all-flash storage system on the market today. As NAND technology changes, flash storage systems must evolve over time. For example, IBM recently signed an agreement with Micron Technology that enables the use of new multi-level cell (MLC) NAND technology, which increases density and further reduces storage costs.

IBM has engineered FlashCore technology to maximize flash's lightning-fast I/O speed while at the same time supplying the most reliable flash storage available. FlashCore technology is integral to the following three fundamental components of IBM FlashSystem 900 storage:

- **Hardware-accelerated architecture** By using IBM developed hardware in the controller, the FlashSystem 900 can minimize or even eliminate the amount of software interaction during I/O activity to supply a controller with the highest I/O performance and the fastest response times for all-flash storage arrays.
- **BM MicroLatency™ module**. By using an IBM-designed, proprietary flash storage module, FlashSystem 900 storage can deliver the fastest performing and most reliable flash storage modules.
- **Advanced flash management**. By using IBM hardware and patented software algorithms to better manage native NAND reliability, FlashSystem 900 storage can provide the most reliable and most highly available flash data storage.

Figure 1 IBM FlashSystem 900

Hardware-Accelerated Architecture

The problem with other vendors' flash storage systems is that oftentimes the storage controllers' software causes I/O to slow down because NAND storage I/O is so fast. In contrast, FlashCore technology was designed to minimize (or even eliminate where possible) any software interaction during I/O activity. Minimization of software interaction starts in the FlashSystem 900 storage controller.

Engineered for Flash

Most all-flash storage systems on the market today were originally designed for rotating media. Rotating media can take tens of milliseconds to access a data block, which leaves a lot of spare instruction execution time to set up and terminate an I/O operation. NAND, on the other hand, processes an I/O operation in microseconds, which leaves little if any time for instruction execution or processing.

From the start, FlashCore technology was designed to use flash storage; as a result, everything from the packaging, data paths, and hardware selection to the software functionality evolved around the speed of flash storage. Taking advantage of the speed of flash requires the use of much more hardware functionality than that required by disk-based storage.

Hardware RAID

IBM FlashSystem 900 offers a system-level hardware RAID to provide supplementary data protection for failures affecting entire flash storage modules. Special-purpose hardware RAID provides fast parity generation during write operations and rapid parity use during rebuild operations. In the unlikely event that a flash controller or a flash storage module completely fails, system-level RAID 5 functionality can quickly reconstruct the inaccessible data onto a hot-spare flash module within the IBM FlashSystem. FlashCore technology hardware RAID also supplies a wider variety of RAID layouts that can support 3 data + 1 parity storage configurations all the way up to 10 data + 1 parity storage configurations.

FlashSystem 900 hardware RAID speeds up customer write operations and provides faster flash storage rebuilds. By providing smaller RAID group sizes, FlashSystem 900 also offers a much wider selection of storage capacities, allowing customers to more easily tailor their flash storage to meet application requirements.

Non-Blocking Crossbar Switch

Most storage systems today use PC-based Peripheral Component Interconnect Express (PCIe) buses to access host bus adapters (HBAs) for frontend connections to servers and use serial-attached SCSI (SAS) controllers for backend connections to storage. Unlike these storage systems, FlashCore technology controllers are built around a proprietary, redundant, non-blocking crossbar switch backplane. Crossbar switching equips the FlashSystem 900 with direct data paths between every host I/O interface and each flash storage module. FlashCore technology crossbar switching can provide higher internal data bandwidth than PCIe buses that also don't have to wait while other I/O operations complete, allowing more parallel and concurrent I/O activity. This feature eliminates the bus overhead required by other vendor flash storage systems. One crossbar switch backplane is located in each IBM FlashSystem 900 storage controller.

FlashCore technology crossbar switching means that the FlashSystem 900 can execute more I/O requests concurrently and that each individual flash I/O operation takes less time to complete. As such, it takes less IBM flash storage to provide equivalent performance to other vendors' all-flash storage arrays.

Hardware-Only Data Path

FlashCore technology features a hardware-only controller data path. Whereas non-IBM flash storage uses software execution to initiate, monitor and terminate data transfers, the IBM FlashSystem 900 uses purpose-built, dedicated field-programmable gate array (FPGA) hardware. As mentioned above, non-IBM flash storage systems typically use general purpose, Intel/x86 processor instruction execution to manage data transfer activities, which takes longer and adds I/O latency.

Moreover, **data path checksums** are generated and validated by FPGAs. FlashCore technology appends these data path checksums to all data being transferred internally so that the system can quickly identify and correct any data transmission errors.

FlashCore technology FPGA-managed controller data paths provide the lowest latency flash storage I/O in the industry today.[1] FPGA data path hardware, together with non-blocking crossbar switching, enables IBM FlashSystem 900 storage to execute more I/Os in less time than other non-IBM flash storage systems.

[1] See Silverton Consulting's latest Storage Intelligence dispatch on storage benchmark performance results, available at http://silvertonconsulting.com/cms1/dispatches/.

© 2015 SILVERTON CONSULTING, INC. ALL RIGHTS RESERVED

Single-Box Highly Available Architecture

FlashCore technology creates a whole new level of availability and serviceability for flash storage systems. For instance, the current FlashSystem 900 is a fully modular storage solution with all key, non-passive components contained within field-replaceable units (FRUs) or modules. The following FlashSystem 900 components are fully redundant and can be hot-swapped when needed:

- **Flash storage modules**. FlashSystem 900 MicroLatency modules are accessible from the front of the unit. Whenever a failure occurs, these modules can be easily replaced with no impact to storage operations.

Figure 2 IBM FlashSystem 900 Back

- **Dual sets of interfaces, RAID controllers, backplanes, and management controllers**. FlashSystem 900 has redundant controller FRUs or canisters that include all of these components. The controller canisters can be accessed and hot-swapped from the rear of the system, providing non-disruptive, continually available storage operations.

- **Dual power supplies, batteries and fan modules**. Whenever a failure occurs, redundant power supplies, batteries and fans can be accessed and hot-swapped without impacting system operations.

FlashSystem 900 brings the long-running tradition of IBM's highly serviceable storage offerings to all-flash storage, which should result in shorter service actions and less application disruption during the repair of hardware failures.

Concurrent Code Load and Maintenance

FlashCore technology provides advanced functionality to enable non-disruptive code upgrades and other software maintenance activities. That is, in addition to the enhanced availability and serviceability hardware described above, FlashSystem 900 offers **non-disruptive (concurrent) code load**, which maintains data and I/O accessibility while system code is being upgraded or modified.

Similar to the highly available and serviceable hardware described above, FlashCore technology concurrent code load enables FlashSystem 900 to continue to provide I/O access during software upgrades. Together, highly available controller hardware and concurrent code load enable FlashSystem 900 storage to stay online during any hardware service or code change activities, ensuring that applications can continue to be used during FlashSystem 900 maintenance.

IBM MicroLatency™ Module

IBM has engineered the FlashSystem 900's MicroLatency module's flash storage to complement the hardware-accelerated architecture at the controller level in order to offer the fastest I/O response time for NAND data storage available today.

Rather than making use of standard SSD storage, IBM designs its own MicroLatency modules. These storage modules use industry-standard NAND chips but offer much higher I/O performance than that available from other SSD storage.

© 2015 SILVERTON CONSULTING, INC. ALL RIGHTS RESERVED

Furthermore, by using a purpose-built flash storage module, IBM also offers higher density storage with more flash storage than standard SSDs, along with more flash chip controllers in a single package. Thus, FlashSystem 900 customers receive higher I/O performance while needing less rack and floor space for their flash storage capacity.

Figure 3 IBM FlashSystem 900 Flash Modules

Parallel Design

Each 5.7TB IBM MicroLatency module has four controllers and 64 flash chips (16 per controller). The FlashCore MicroLatency module offers multiple concurrent operations per flash chip and flash controller. Every flash controller can perform up to 40 direct memory access operations to its flash storage in parallel, which, when scaled up to a complete FlashSystem 900 storage system, means up to 1,760 simultaneous NAND access operations in a 57TB system. FlashSystem 900 storage can thus sustain quick I/O performance even under heavy read and write I/O workloads.

Consequently, customers can enjoy the same fast I/O responsiveness as the FlashSystem 900 goes from one I/O operation per second to millions. To provide similar responsiveness across such a wide range of I/O activity, other all-flash storage systems would either suffer increased I/O latency or require more SSDs with more flash controllers and more flash chips.

FPGAs in the Data Path

The FlashCore technology hardware-only data path extends all the way through to the MicroLatency module. That is, I/O data transfer operations in the MicroLatency Module are processed by dedicated FPGAs and do not depend on general-purpose microprocessor instruction execution. FPGAs in the controller, together with FPGAs in the MicroLatency Module, result in FlashSystem 900 storage that can supply very-low-latency I/O performance under extreme load.

As a result, FlashSystem 900 storage I/O operations complete faster with less variability in response time than other all-flash array storage.

Distributed RAM

FlashCore Technology does not use traditional controller-based DRAM cache. Historically, enterprise storage systems have used controller-level DRAM caching as a staging area for data being accessed and as a means of holding frequently accessed data to provide faster I/O performance than reading from or writing to disk or SSD. FlashCore technology systems have no disk or SSDs; instead, FlashSystem 900 storage operates at speeds (reads and writes from flash storage) approaching DRAM caching in enterprise storage systems today. However, FlashCore Technology does use distributed RAM located at the MicroLatency module level to hold metadata such as lookup tables and other information required for flash addressing and translation activities. The MicroLatency module also uses a very small amount of this distributed RAM as a write buffer for incoming data.

FlashCore technology systems can make due with much less DRAM than other all-flash arrays. Moreover, FlashSystem 900 systems don't need the additional processing time or overhead to update and manage a controller-level caching tier, which means that FlashCore technology systems use less instruction execution to process each I/O operation, resulting in faster I/O activity.

© 2015 SILVERTON CONSULTING, INC. ALL RIGHTS RESERVED

High-Speed Interface

A high-density pin connection interfaces the MicroLatency module to the proprietary high-speed crossbar backplane. This connection eliminates any bus instruction processing overhead or serial transport delays for I/O activity at the MicroLatency module level and supplements the high parallelism of the crossbar switching at the FlashSystem 900 controller level.

Together with FlashCore technology's controller crossbar switching, the MicroLatency Module high-density pin connection supplies a higher speed data transfer with more concurrent I/O operations all the way from the host interface to the NAND chips and back again. As a result, IBM FlashSystem 900 customers can extract the highest level of performance possible from their flash storage capacity.

Line Speed Data-at-Rest Encryption

A dedicated chip inside each MicroLatency Module provides AES 256 hardware-based, data-at-rest encryption. Hardware encryption and decryption occur at internal data path line speed and have no impact on I/O latency when in operation. In this way, customers can use data-at-rest encryption for their FlashSystem 900 information without suffering any performance degradation, making it much easier to deploy data security for their flash storage.

Advanced Flash Management

In addition to the hardware-accelerated architecture and IBM MicroLatency Modules, FlashCore technology strengthens NAND reliability by using special-purpose hardware and patented algorithms to extend the life of NAND memory. Such technologies make IBM FlashSystem 900 storage some of the most reliable flash storage available.

IBM Variable Stripe RAID™

FlashCore technology uses patented Variable Stripe RAID at the flash chip level. Variable Stripe RAID is a 15-data plus 1-parity RAID 5 implementation (rotating parity) across NAND memory chips, using dedicated flash controllers inside IBM MicroLatency Modules. When a flash (chip or sub-chip) failure occurs, the data is rebuilt on a previously reserved (over-provisioned) storage area, and the affected RAID stripe shrinks to become a 14-data plus 1-parity (or 13-data plus 1-parity, 12-data plus 1-parity, etc.) RAID group. Shrinking RAID group stripe size is unique in the industry and can better retain flash storage availability with little to no impact on data protection or system functionality.

Variable Stripe RAID flash chip or sub-chip data protection is superior to current industry practice, as many competitors' flash systems have no RAID protection within modules or SSDs. Competitors using only system-level RAID 5 across modules do not necessarily preserve flash capacity and performance as well as Variable Stripe RAID.

The two components of FlashCore data protection – MicroLatency module-level Variable Stripe RAID and system-level hardware RAID – operate independently, but together they provide synergistic system fault tolerance to mend multiple flash memory failures. Further, reserved space for Variable Stripe RAID and dedicated spares for system-level RAID mean there is no reduction in usable system capacity when flash failures do occur.

IBM-Engineered ECC

IBM FlashCore storage uses a strong **error correcting code (ECC)** algorithm to protect data as it is accessed in flash memory. For each new generation of NAND technology, manufacturers require a

© 2015 SILVERTON CONSULTING, INC. ALL RIGHTS RESERVED

minimum level of ECC algorithm to meet their flash reliability specifications. FlashCore technology implements a more powerful ECC algorithm than that required by its NAND suppliers, which provides even more flash reliability.

In addition, specific IBM FlashCore innovations provide the capability to handle most ECC activity using hardware rather than software. Many systems rely on ECC hardware detection but may correct bit errors using software functionality. However, software correction takes longer to fix bit errors, which occur more frequently with higher density NAND chips. With FlashCore technology-designed hardware ECC, FlashSystem 900 storage can take advantage of this high-density but more volatile NAND memory without suffering any undue performance degradation.

IBM FlashSystem 900 customers can therefore benefit from the lower cost and higher density of the latest-generation NAND technologies while gaining high I/O performing flash storage.

IBM-Optimized Over-Provisioning

FlashCore technology incorporates additional reserved flash capacity beyond user-accessible data space. IBM FlashSystem 900 uses this **over-provisioned NAND space** as spare system capacity for failing flash cells. Moreover, because NAND memory technology can write or program only to erased blocks and cannot be overwritten directly, over-provisioning helps IBM FlashSystem 900 storage write performance by supplying more erased NAND memory blocks for write activity. Most flash storage uses some level of over-provisioning in the flash modules.

FlashCore technology-optimized over-provisioning helps FlashSystem 900 supply highly available flash storage and faster write I/O performance. FlashSystem 900 high write performance starts with a completely empty system and maintains that level of performance even when the storage system fills up with customer data.

Wear Leveling

FlashCore technology also uses **wear leveling** to distribute writes across more flash memory within a system and to eliminate premature NAND chip wear-out due to high write activity at any one location. Given NAND memory's endurance limitations, any flash storage solution must spread write activity or program/erase cycles across as many NAND locations as possible.

Together with the optimized over-provisioning described above, FlashCore wear leveling can take advantage of even more NAND storage to better preserve FlashSystem 900 flash storage lifetime.

Write Buffer and Hardware Offload

In conjunction with wear leveling, FlashCore technology uses a specially designed hardware flash translation layer so that new data blocks are written sequentially to contiguous locations in flash memory. Further, the data written to MicroLatency Module-distributed RAM is also written to flash memory via hardware processing and doesn't use software functionality.

As a result, all of the FlashCore Technology's hardware-managed data transfer at the controller and MicroLatency Module level is carried through all the way to the process of writing data to NAND memory locations. The FlashSystem 900 thus processes customer write operations in the quickest fashion possible to provide the highest level of write performance.

TWITTER.COM/RAYLUCCHESI | RAYONSTORAGE.COM |
+1-720-221-7270 | SILVERTONCONSULTING.COM
© 2015 SILVERTON CONSULTING, INC. ALL RIGHTS RESERVED

IBM Garbage Collection

FlashCore storage includes IBM-proprietary **garbage collection**, **relocation and block-picking** algorithms that not only prolong flash endurance but also decrease write latency. Most flash storage garbage collection algorithms are symmetrical and treat all blocks and all accesses the same. FlashCore technology algorithms go further and use detailed NAND block characterization data to determine the health of each block and match it to incoming write activity. IBM FlashSystem 900 garbage collection takes into account several attributes to reduce excess write activity (write amplification) and get the most life out of every NAND block.

In combination with optimized over-provisioning and wear leveling, IBM's sophisticated garbage collection works to extend MicroLatency Module flash endurance while at the same time providing high write I/O throughput for FlashSystem 900 storage.

Summary

IBM FlashCore technology's advanced hardware and proprietary algorithms have been designed to provide the fastest performing and most reliable flash storage available today. Moreover, IBM FlashSystem 900 storage includes the industry's most advanced NAND technology to supply customers with the most cost-effective flash storage.

To achieve this combination of higher performance, greater reliability and lower cost from today's NAND technology, IBM has developed a much more hardware-intensive FlashSystem 900 design, starting with the controller, flowing through the MicroLatency module, and stretching all the way to the flash chip controllers. Although all this requires extra engineering, hardware-intensive designs, like in the IBM FlashSystem 900, achieve the highest performance and best reliability from today's highest density NAND technology.

Silverton Consulting, Inc., is a U.S.-based Storage, Strategy & Systems consulting firm offering products and services to the data storage community.

QRcode: SilvertonConsulting.com

Disclaimer: This document was developed with International Business Machines Corporation (IBM) funding. Although the document may use publicly available material from various sources, including IBM, it does not necessarily reflect the positions of such sources on the issues addressed.

© 2015 SILVERTON CONSULTING, INC. ALL RIGHTS RESERVED