Violin MEMORY

White Paper

Flash Fabric Architecture™

Violin Memory's Flash Fabric Architecture is the foundation for the next generation of enterprise storage

Version 2.0

Abstract

Purpose-built flash technology impacts data center economics. The consistent low latency provided by Violin's Flash Fabric Architecture enhances virtualized environments and transforms enterprise computing. The FFA's parallelism and concurrent operations provide clear differentiation from SSD-based designs in terms of performance, resiliency, and density. Flash Fabric Architecture is found in Violin's 6000 All Flash Arrays, Flash Storage Platform 7000 series, and Windows Flash Arrays.

Table of Contents

02-0207 rev 1.1r 03/16

1. Introduction

Violin all-flash arrays deliver custom engineered hardware, firmware and software that functions as primary storage for enterprise data centers. Violin Memory has addressed the challenges in these demanding environments in a profound way. Unlike other companies rushing to enter the solid state storage marketplace by using commodity components to shorten their time to market, Violin identified the underlying storage-related challenges faced by our customers exploring virtualization, working with "Big Data analytics" and accelerating their mission critical applications. We targeted common data center technical challenges as our design criteria, then custom-built an integrated hardware, firmware and software solution to solve them, rather than cobbling together off-the-shelf components. The result is a purpose-built all-flash solution with the performance, reliability, and cost-effectiveness to perform as primary storage in your enterprise data center.

The key to building one of the most successful and powerful primary storage solutions on the market was our decision to custom-engineer the firmware, software and hardware components. This purpose-built approach enables a level of deep integration between our operating systems and Flash Fabric Architecture™ (FFA) hardware that off-the-shelf integration cannot achieve. In this white paper we introduce the FFA components, illustrate their unique capabilities, and highlight some of the advantages of our purpose-built chip-to-chassis engineering approach.

2. The Need for Purpose-Built Flash Storage

Successfully deploying NAND flash in enterprise storage environments requires purpose-built solution that can manage the reliability, density and performance challenges inherent in the 21st century data center.

The core storage technology implemented by Violin is NAND Flash. Each NAND package is based on 4 or 8 dies; each die is comprised of two planes, a number of blocks in each plane and a number of pages in each block. Data is stored in cells on a page.

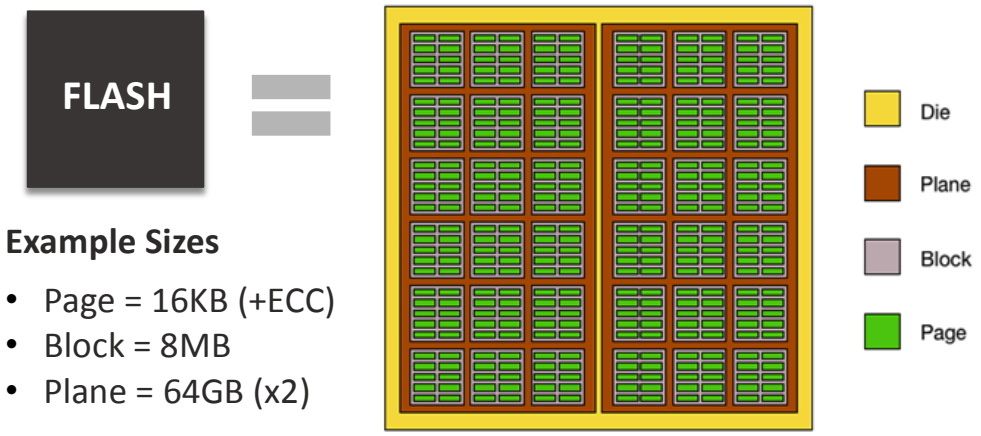

Figure 1: NAND Flash Die Layout

All flash supports the same basic set of operations: Read, Write, and Erase. The take-away from Table 1 is that Erase operations are relatively very slow: five milliseconds.

	Read Ops	Write Ops	Erase Ops
MLC	50 µs	1400 µs	5000 µs

Table 1: NAND Flash Read, Write, and Erase command response times

However, there are several issues with flash devices (chips):

- Flash physics requires some special handling of Erase operations.
- Writes ("flash programs") are done at the page level, are sequential and relatively slow (1,400µsec or 1.4ms).
- Erases require a whole block to be erased and take considerably more time (often requires over 5,000µsec or 5ms), during which nothing can be read or written.
- Reading is very fast (<100µsec) and can be either random or sequential. However, only a single page can be read at a time.
- Cells can be corrupted by repeated reading (Read Disturb).
- A single block can be erased only so many times before it wears out. (program/erase cycles or P/E)

2.1. Read/Write Speed: Write Cliff

Flash management issues are most obvious when measuring the sustained random Write performance of certain flash-based devices. Initially, the performance is good when the pages are empty, but drops dramatically over a "Write Cliff" when pages are recycled in a process known as Garbage Collection(GC) or grooming. Both solid-state disk (SSD) and in-server PCIe flash cards suffer from significant Write Cliff performance degradation.

When SSDs hit their Write Cliffs they begin cycling through already programmed flash blocks. At this point, new Writes get stuck behind extra Write and Erase operations from data cycling thus causing a dramatic drop in overall performance, up to 60%. The Write Cliff phenomenon doesn't just affect Write latencies. Because of the way flash dies work, Erase operations also get in the way of Read operations on blocks within the same flash die. Erase operations and their associated massive latency spike can seriously degrade both Read and Write latencies.

Sustained Write performance is a function of write amplification; which in turn is a function of outstanding Write workload and of flash over-provisioning (formatting level). Write amplification is the ratio of total flash Writes to user Writes. Violin's FFA, when faced with extreme write activity can avoid write cliff through alternative formatting levels. Violin Arrays are shipped with 84% formatting as standard but can be changed in the field, typically before provisioning. By reducing the formatting to 78% or in more extreme cases to 65% there is sufficient capability to effectively eliminate any performance impact from GC. SSD-based arrays cannot provide this capability and will experience Write Cliff impact on an active fully-loaded system. The Violin Array is designed to support mixed and multiple workloads. Formatting is a significant way to ensure that the Violin Array can meet customer requirements, even with heavy Write workloads.

3. The Violin Flash Fabric Architecture

Performance and reliability are engineered into the Violin FFA, from the chip to the chassis. For NAND flash to be viable in data center applications, it requires a very different set of attributes compared with PC or consumer devices. Sustained and predictable performance is required for enterprise applications. Violin Memory has solved the inherent "behavioral issues" of NAND flash through innovative design and chip-to-chassis integration of hardware components known as the FFA. Our FFA enables thousands of flash devices to operate efficiently together, masking chip level issues, and delivers reliable and sustained system performance.

Violin's FFA is woven from multiple layers of innovative technologies, the result of a purpose-built flash system approach with patented flash optimization algorithms implemented in hardware, operating at line rate.

- At the architecture's core lies a resilient, highly available deep mesh of thousands of flash dies that work in concert to continuously optimize performance, latency, and longevity.
- **Violin Intelligent Memory Modules (VIMMs)** organize this mesh of individual dies into intelligent flash management units. VIMMs provide a hardware-based Flash Translation Layer with GC, wear leveling, and error/fault management.
- Each VIMM is integrated into the patented **Violin Switched Fabric (vXF)**, designed from the ground-up for power efficiency and performance.
- Finally, VIMMs and the vXF layer work in conjunction with **vRAID**, Violin's patented hardware-based RAID algorithm specifically designed to increase reliability and reduce latency.

Collectively, the mesh of flash dies organized into VIMMs integrated into vXF and overlain by vRAID make up the Violin Flash Fabric Architecture.

Violin's FFA enables a profoundly reliable, highly available storage at silicon speeds offering multiple industry leading benefits:

- Spike-free Low Latency – FFA delivers spike-free and predictable latency that is lower than HDD or SSD solutions.
- High Bandwidth - A single Violin all-flash array supports over 8,000 flash die and 500 independent flash interfaces. This provides the bandwidth needed for outstanding flash performance which cannot be matched by SSD-based systems.
- Extreme Reliability – All active components of the FFA are redundant and hot-swappable for enterprise grade reliability and serviceability.

The FFA is found in Violin's 6000 All Flash Arrays, Windows Flash Arrays and Flash Storage Platform 7000 Series and is comprised of up to 64 VIMMs and four active/active vRAID Control Modules (VCM). All inputs/outputs (I/O) are processed through Array Managers before being handed off to the VCMs, which implement vRAID and orchestrate flash management operations across all of the flash dies. Background GC performs various flash optimization tasks, one of which is proactively erasing flash blocks so they are ready for new incoming Writes. Performing GC involves reorganizing data placement and erasing flash blocks within VIMMs.

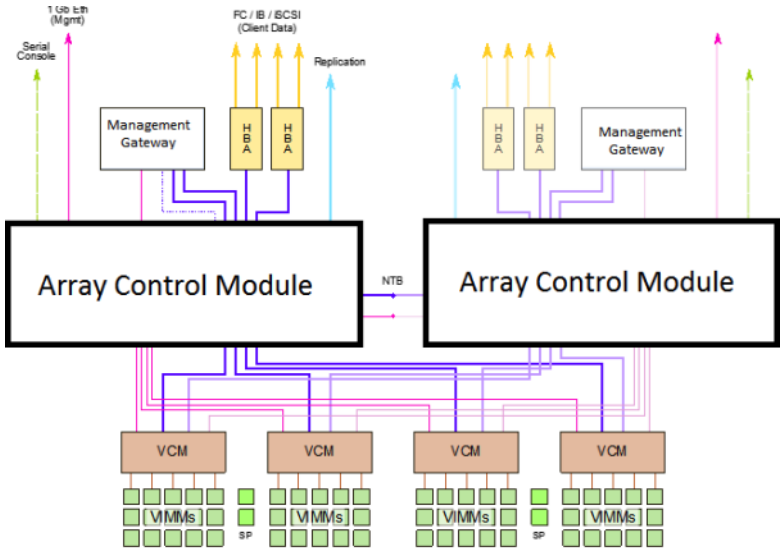

Figure 1: Flash Fabric Architecture

In addition to the highly parallel architecture formed around the four VCMs, FFA also employs concurrent operations to deliver consistent low latency. For incoming Reads, we do not have the luxury of choosing where these reads can occur. So VCMs ensure that only one VIMM in each protection group is performing GC at any given point; this leaves four VIMMs available so that vRAID rebuilds can be performed to serve any Read request without delay. We call this "Erase Hiding." Because these algorithms are implemented in the FFA at line rate, we lead the industry in delivering sustainable spike-free microsecond latencies for mixed and multiple workloads. The Flash Storage Platform introduced another concurrent operation. For incoming writes we have "Write Hiding" which is implemented using similar principles that further reduces Read latency enabling concurrent Read and Write handling in the same vRAID stripe.

3.1. NAND Flash Memory Chips

At the heart of the Flash Fabric Architecture lie thousands of individual Toshiba flash dies. Violin and Toshiba have formed the Violin/Toshiba Strategic Supply and Roadmap Agreement, ensuring that Violin always has high priority access to supplies of Toshiba flash chips. Our engineers collaborate to develop the most reliable and high performance flash technologies possible. Also, our unique roadmap sharing arrangements provide Violin engineers with deep understanding of Toshiba flash architectures, allowing us to optimize the performance and reliability of our FFA through perfect integration of our VIMM, vXF, and vRAID layers. Our relationship with Toshiba allows us to use specific controller designs for cMLC parts without the massive overprovisioning and still provide enterprise level reliability.

3.2. Violin Intelligent Memory Modules

Violin VIMMs are the core building block of the FFA, designed from the ground up to be the highly resilient, hot swappable technology implementing our proprietary fabric-level flash optimizations algorithms.

Array operations (erase/program/read) are done at the die level, and a single VIMM contains up to 128 flash dies. A 64 VIMM Violin array thus contains more than 8,000 flash dies, managed as a single system by vRAID in the VCMs. Optimizing flash endurance, data placement, and performance across such a large number of dies is the key to Violin Memory's unique ability

to deliver sustainable performance, ultra-low latency, and industry leading flash endurance. While a commodity SSD contains a flash controller optimizing flash across tens of dies within the SSD, the Flash Fabric Architecture can leverage thousands of dies within which to implement optimization decisions.

Data is written to and read from VIMMs. Each VIMM operates its own instance of a Flash Translation Layer. Data is written to VIMMs through the Logical Page Address (LPA), which assigns the page to a Physical Page Address (PPA) within the flash of the VIMM. Metadata maps between logical addresses and physical addresses. DRAM on the VIMMs keeps the entire flash virtualization table in memory, ensuring exactly one flash access for a 4KB user read. DRAM isn't used for caching, only for buffering which gives the system predictable performance. Each VIMM includes:

- High-performance logic-based flash controller
- Management processor
- DRAM for metadata
- NAND flash for storage.

Each VIMM is designed to enable the scalability of large arrays of flash storage. The advantages of this architecture over a simple PCIe flash card are significant:

- Integrated GC for sustained Write performance
- Low latency access to DRAM metadata and flash storage
- Safe access and local storage of metadata for fault recovery
- Integrated monitoring and management of flash storage health
- Error Correction Code (ECC) correction for maximum bandwidth
- Hot swap and redundancy management
- 3 port design so that single port failures do not impact data access.

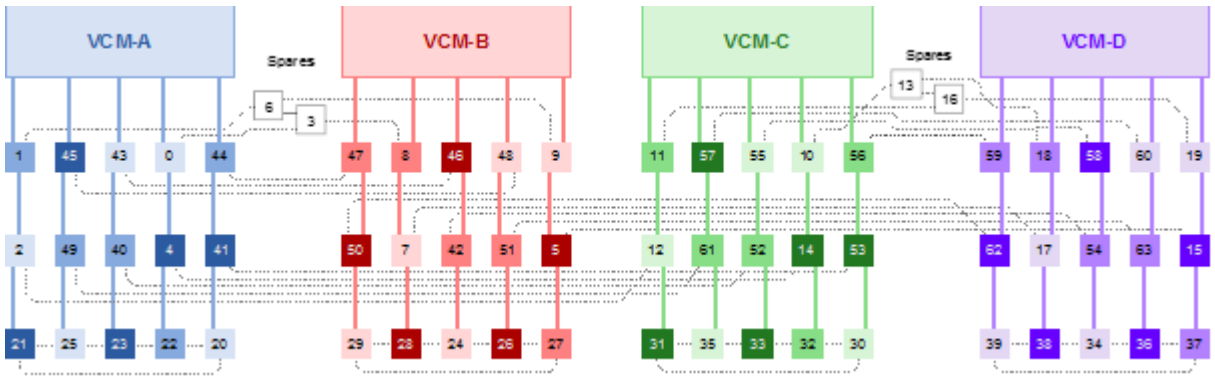

Figure 2: VIMM Tree Topology showing 3 port design

Unlike most SSDs and PCIe cards, a failed flash device does not cause a VIMM to lose data or be removed from service. ECC and vRAID protection manages data against loss.

VIMMs work in conjunction with Violin Memory's patented vRAID to ensure a very high degree of system availability and data integrity using multiple techniques, including:

- Data protection**:** Robust ECC and Cyclic Redundancy Check (CRC) algorithms detect and correct bit errors in the system.
- Data validation**:** Extra data is stored with each block of flash so that invalid data is detected rather than passed on.
- Data scrubbing: All data in the system is read on a weekly basis and scanned for errors. Any errors found are then repaired. Violin does this without any noticeable impact to user performance. It greatly reduces data loss rates and increases data endurance.
- Flash wear leveling: FFA evenly distributes data Reads and Writes to all the flash devices in all the modules. No specific Logical Unit (LUN) is tied to a specific module and hence active LUNs do not wear out specific flash devices.
- Flash monitoring: All Read, Write, and error statistics are captured and reported. Any VIMM behaving below specification is automatically removed from service and the error events logged. The data from that VIMM is moved to a spare VIMM according to the vRAID algorithm to rebuild the data. This is performed in the background without administrative intervention or any significant impact on access to user data. The system may have one to four spares and hence replacement of the module is not an urgent requirement and may be hot-swapped monthly or quarterly.
- Flash recovery: in the unlikely event of a VIMM failure, the VIMM self-scrubs and automatically requests a RAID rebuild of any unreadable addresses to protect against block or whole-plane failure.

3.3. vRAID: Hardware-based Data Protection and Performance Enhancement

Other solid-state storage solutions and architectures, such as SSDs and PCIe cards, employ central processor units (CPU) and software to perform RAID, page mapping, and GC. Violin implements these functions in hardware to reduce latency and dramatically increase sustained random Write IOPS from less than 10,000 to more than one million. Violin hardware-based controller functions perform CRC analysis on each read to maintain data integrity. If any error is detected, the address is automatically RAID rebuilt. SSD-based designs can't match this level of protection without a large performance penalty. This capability is also built into Violin's non-disruptive upgrade (NDU) technology that knows where spare VIMMs are available to ensure that all Writes are being written with parity and then automatically incorporates them back into the VIMM after the NDU is complete. In short, Violin's hardware-based controller functions enable Violin arrays to optimize feature delivery, not performing low level operations of the array.

Low-latency flash vRAID, Violin's patented flash technology designed specifically to enhance NAND flash system performance, provides full RAID data protection and a fundamentally more efficient and higher performance solution. Existing RAID 5 and 6 solutions rely on Read-Modify-Write operations that are unsuited to flash. Unlike inefficient RAID 1 (50% efficient) solutions, Violin's vRAID enables 80% usable capacity and bandwidth.

vRAID guarantees spike-free latency under load by making sure there aren't any Reads blocked by Erases. Notably, the microsecond latency of the Violin 6000 Series All Flash Array is 80% lower than Tier-1 storage cache (DRAM) and significantly improves metrics such as file Read and Write, response, and query times. Violin designed flash controller functions enable vRAID to provide constant low latency, not just "average" low latency. Low average latency with spikes in performance results in much lower application performance than consistent low latency provided by Violin. With traditional SSD-based designs, GC is performed in the foreground and impacts application performance. vRAID enables background GB to avoid latency spikes that can impact application performance.

SSD-based RAID groups do not effectively accommodate SSD maintenance. When SSDs go through garbage collection, there is no ability to take the overall workload of the array into consideration. This means that when two SSDs in the same RAID group are going through GC simultaneously, I/O operations are suspended until the GC is complete. This results in latency spikes that prevent consistent performance, and makes SSD-based arrays a poor fit for consolidation of mixed workloads.

vRAID delivers fabric-level flash optimization, dynamic wear leveling, and advanced ECC for fine grained flash endurance management, as well as fabric orchestration of GC to maximize system level performance. vRAID also protects the system against VIMM failures.

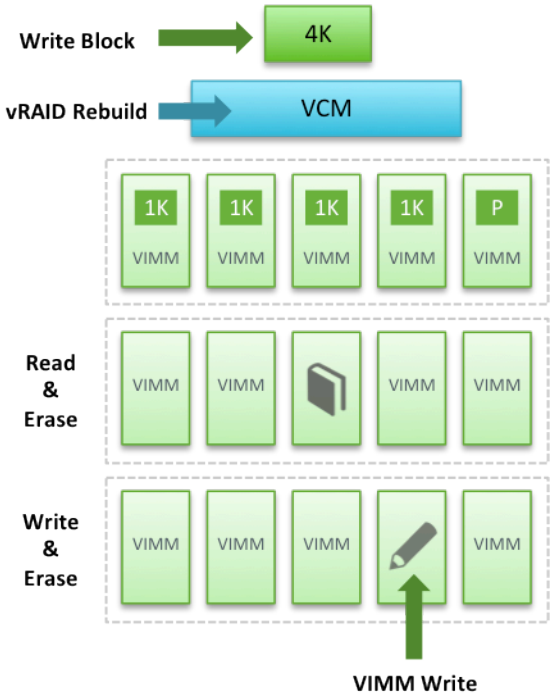

Figure 3: vRAID Concurrent Writes & Reads

vRAID is implemented within the Violin Control Module (VCM). Each VCM manages 15 VIMMs (4 VCMs for a total of 60 VIMMs plus 4 hot spares in a fully populated Violin array). Note that when VCMs fail, VIMMs are automatically reallocated across the remaining VCMs in the system such that a single VCM is able to manage all VIMMs in a system, though at the cost of degraded performance. For the purpose of data placement and protection, VIMMs are organized in groups of 5: any incoming 4KB Write request is allocated to a VCM which transfers the Write request to the appropriate group of 5 VIMMs to handle the Write. 4KB is written across 5 VIMMs as 4 * 1KB of data, plus 1KB of parity data.

This data placement algorithm is applied across the entire array. Data is placed across the entire set of VIMMs. The net result is automated, granular, LUN-wide striping across all available flash dies in the Fabric (all 8,000 of them). This wide stripping happens at all levels in the Fabric, across VCMs, across VIMMs, and across flash dies inside each VIMM. All operations are implemented in hardware, at line speed, ensuring that any data can be read from the Array with the lowest levels of latency. As a result, any LUN in the array has access to the full system bandwidth, by default. The system is simply fast.

The combination of VIMMs and vRAID resolve issues of flash errors, reliability, and wear to deliver an enterprise-grade all-flash solution with a far greater life expectancy than typical HDD arrays. For example, Violin All Flash Arrays using SLC NAND flash can sustain a write rate of 8TB/hour – greater than 2GB/s – for over 10 years. The truth is, in all of the time that Violin Memory

has been shipping flash arrays, no customer has worn out a VIMM. They have failed due to component failures, but typical enterprise use has resulted in wear rates of significantly less than 10% per year.

The real test for storage systems comes from handling a combination of media errors and module failures. As an example, 30 SATA disk drives in a RAID-5 configuration have long RAID rebuild times and high error rates that lead to a Mean Time to Data loss (MTDL) of less than a few years. The MTDL of a VIMM is estimated at 200 years, about 20 times higher than a rotating HDD. The RAID rebuild time for a Violin array is typically between one and 24 hours, depending on user load and NAND type. The RAID-5 rebuild time for large HDD-based arrays is measured in days, especially under load.

3.4. Violin Switched Fabric

Violin has developed the industry's first flash switched fabric. This patent-pending architecture is called Violin Switched Fabric (vXF) and is implemented within our FFA. With our vXM, each VIMM is part of a switched array that supports large topologies with fault tolerance. Unlike other flash interconnects, vXF was designed from the ground up for power efficiency and performance. The vXF is the industry's first flash solution developed specifically for applications with large datasets. SSD-based architectures often use SCSI commands to move data inside the array. Violin created the Violin Memory Channel (VMC) as an alternative which is designed for low latency, high reliability flash transactions. The result is higher performance and reliability compared to SCSI-based systems. These large dataset applications including databases, image, video, scientific, and web content can leverage multiple vXF technology benefits:

- Unmatched scalability
- Highest performing I/O
- Fault tolerance
- Ultra-green power savings

Previous solutions have not scaled because they relied on classical linear bus topologies. Bus speeds limit the number of registered devices that can be supported in a single channel to fewer than eight. As bus speeds increase, the number of registered devices can drop to just one or two. The vXF does not suffer from these limitations.

vXF protects high speed data in a number of ways. VMC monitors channel sequence numbers to immediately detect loss and causes automatic hardware retransmission. In addition, VMC creates a multidirectional CRC that protects both the transmission on each channel as well as across the RAID stripe.

3.5. Integrated Chassis-No Battery Required

Since Violin builds our own flash and RAID controllers within an integrated chassis, we don't need external battery backup units (BBU). In case of power failure, the power supply notifies the system which causes the VIMMs to immediately write out any data in their buffers. We don't allow any more data in buffers than we can offload in the time provided by our power control unit capacitance reserve. The result is protection from unexpected power outage without the need for batteries and all the service issues that they create.

The array managers plug into a PCIe switch fabric of our own design that allows us to arrange the CPU connectivity to the HBAs, the RAID controllers and the other array managers (remember it's an HA configuration) in ways that can't be done in a standard off-the-shelf server, enabling us to build systems optimized for low latency and optimal data throughput.

4. Conclusion

Architecture does make a difference. By choosing to custom engineer our arrays from the chip to the chassis – from NAND flash dies through Flash Fabric hardware to Violin Operating System software – we solve, avoid, or dramatically improve every shortcoming of the basic flash medium itself and of less-thoroughly engineered solid state storage solutions. The benefits to enterprise customers are clear – much higher system performance, ultra-low latency, decades-deep reliability, high density, low power/cooling needs as well as easy installation and management. Perhaps most important, thanks to the efficiencies derived from our custom-engineered FFA and OSes, Violin customers can get storage in a flash for costs equivalent to enterprise hard disk drive arrays that soon will be obsolete – at any price.

Find out more about Violin technology, products and solutions at www.violin-memory.com

About Violin Memory

Business in a Flash. Violin Memory transforms the speed of business with high performance, always available, low cost management of critical business information and applications. Violin's All Flash optimized solutions accelerate breakthrough CAPEX and OPEX savings for building the next generation data center. Violin's Flash Fabric Architecture (FFA) speeds data delivery with chip-to-chassis performance optimization that achieves lower consistent latency and cost per transaction for Cloud, Enterprise and Virtualized mission-critical applications. Violin's All Flash Arrays and Appliances, and enterprise data management software solutions enhance agility and mobility while revolutionizing data center economics. Founded in 2005, Violin Memory is headquartered in Santa Clara, California.

For more information, visit www.violin-memory.com. Follow us on Twitter at twitter.com/violinmemory

Violin Memory, Inc. | 4555 Great America Parkway, Santa Clara, CA 95054 USA | Tel: 1-650-396-1500 • Fax: 1-650-396-1543 | www.violin-memory.com

02-0207 rev 1.1r 03/16