CLABBY ANALYTICS

Research Report

Fast Flash Memory-to-Memory Systems: The Microsoft/Mellanox/Violin Memory Relationship

Executive Summary

Microsoft and Violin Memory have announced a jointly developed Windows-based solution (Window Flash Array – WFA) that speeds data delivery between systems (between clients and servers, as well as between servers and servers) – essentially performing ***high-speed memory-to-memory transfer without burning CPU cycles***. This is accomplished by using a protocol known as Remote Direct Memory Access (RDMA).

> ***Using this solution, enterprises can transfer data memory-to-memory between Windows-based systems at close to line speed.***

The types of workloads that benefit from this environment are those that suffer from bandwidth-related issues and I/O bottlenecks – for instance, large scale simulations, rendering, large scale software compilation and trading decisions. Additionally, branch offices that need fast access to remote files and data benefit from this solution (remote files and data feel "local" using this solution).

Using this solution, information technology (IT) managers and administrators can expect to see:

- Improved database performance. For instance:
  - SQL Server database write speed can be increased by 2X;
  - SQL Server reads can take place up to 1.5x faster.
- Hyper-V virtual machine (VM) writes can take place up to 1.4x faster;
- An increase in virtual machine density (meaning more work can be derived from existing server investments);
- CPU (central processing unit) performance improvements can be improved by up to 30%
  - Leading to improved server return-on-investment and lower software costs; and,

Note: *these measurements are against industry standard all flash arrays. Hence, if compared against hybrid, SSD, or HDD arrays, the performance improvements would be dramatically higher*.

From our perspective, we see this solution as a way to work around having to invest heavily in primary dynamic random access memory -based (DRAM) memory. We observe that DRAM (main memory) is always faster than NAND-based flash – but we also observe that the performance characteristics are similar in that both are much faster than traditional mechanical disk drives. In the specific case of DDR3 DRAM vs MLC NAND Flash (the FLASH used in the WFA), DDR3 ***DRAM is 17X more expensive than MLC NAND Flash*** in terms of $/GB. Accordingly, if the performance of MLC NAND is about 1/6 (or better) than that of DDR3, ***FLASH can cost about 1/3 as much (on a performance adjusted basis)***.

In this *Research Report*, *Clabby Analytics* takes a closer look at this fast "memory-to-memory" solution. And we conclude that IT executives looking for significantly improved Windows database performance, increased VM density, and better CPU utilization and performance should closely examine this tri-vendor integrated file/data transfer solution.

Background

The goal in building a balanced information systems design is to maximize efficiency. With more efficient systems, IT executives can achieve a better return-on-investment while lowering systems/software acquisition costs (because fewer servers are needed to execute workloads – and fewer servers means fewer software licenses are required).

In order to maximize efficiency, a balance between processing speed and data delivery must be achieved. Data needs to be fed to the processor at a rate that doesn't swamp the processor, or that doesn't leave the processor thirsting for data. To date, much of this balancing act has been accomplished by finding the right amount of primary system memory to dedicate to a given processor. Providing too much memory can be expensive (main memory is costly and should not be wasted); providing too little means that the processor will waste cycles with nothing to process.

> ***To feed data to CPUs at high speed, Microsoft and Violin Memory have jointly developed a balanced file/data delivery environment that consists of file/data transfer software that can exploit RDMA, All Flash Arrays and RDMA adapters/switches. The file/data transfer software that exploits RDMA is known as Microsoft's SMB Direct – it is an implementation of SMB 3 (simple message block file transfer software) that was specifically designed to exploit RDMA (it was released as part of Windows Server 2012 and the follow-on R2 version). Violin Memory delivers an All Flash Array known as Windows Flash Array that feeds data to SMB Direct a very high speed. Mellanox builds Infiniband and Ethernet adapters and switches capable of handling RDMA requests.***

Is This Solution Really Memory-to-Memory?

Microsoft, Mellanox and Violin Memory position their combined solution as "tier-0" or "tier-1 storage solutions. ("Tier-0" and "tier-1) are storage classifications that refer to fast/expensive storage – typically solid state drives with Flash memory as compared with mechanical storage). We, however, we view this solution differently. We see the Violin Memory's WFA as extended memory. And we think *this distinction is important because it means that IT buyers may not have to purchase as much primary systems memory as they have had to in the past in order to speed data transfer between systems*. Instead, large amounts of less expensive Flash-as-memory All Flash Arrays can be used as a substitute for memory – *enabling enterprises to build high performance systems that behave as very large memory environments*.

Here are the basic arguments in favor of classifying the Windows Flash Array as storage versus classifying it as extended memory:

- *Windows Flash Array as Storage* – It can certainly be argued that the Violin Memory's involved in this solution is storage because the array offers typical storage array features that include data deduplication, storage live migration, thin provisioning, compression, replication, scale-out file services and transport level encryption.
- *Windows Flash Array as Memory* – Primary systems memory can interface with a processor over a dedicated, high-speed memory channel.
  - The speed at which primary systems memory interfaces with the processor is variable depending on memory type and memory speed. So, for instance, a DR3-1333 memory module can achieve a bandwidth rate of 10.6 GB/s; while a DR3 – 1866 can achieve a bandwidth rate of 14.9 GB/s when interacting with an x86 processor.

- Violin Memory’s Windows Flash Array can interface directly with primary systems memory over a high-speed *PCIe channel* using a facility known as direct memory access (DMA). PCIe3 bandwidth is 15.754 GB/s.

> ***To us, this means that Flash data can be placed in primary systems memory at a rate that is equivalent to the speed at which primary memory can deliver it to the processor across a memory channel. We see primary systems memory as a pass-through for extended Flash memory.***

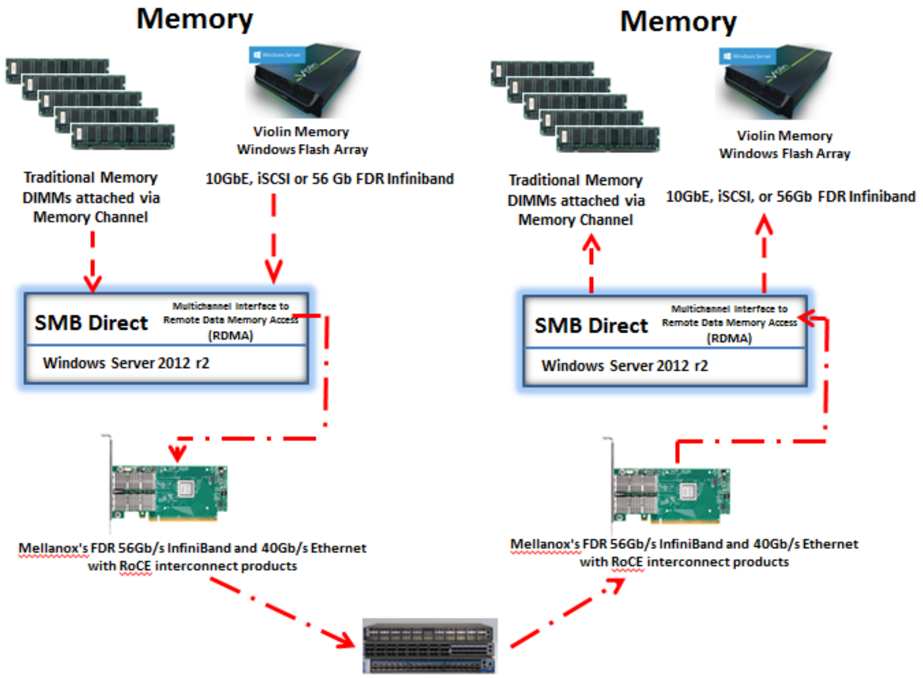
A Closer Look at This File Transfer Solution

To understand this data flow, it is necessary to better understand the product offerings of each vendor – as well as what Remote Direct Memory is:

- Data contained in Violin Memory’s Windows Flash Array is sent to Microsoft’s SMB Direct for file transfer;
- SMB Direct (server message block direct) and NFS (network file system) 3.0/4.1 enable data on one computer to read/write, as well as to request services, from another computer. SMB Direct makes use of RDMA, whereas plain NFS does not use RDMA. The WFA still delivers fast NFS, but it does not exploit RDMA as SMB Direct does.;
- SMB Direct can exploit a communications protocol known as Remote Direct Memory Access to streamline communications between the processor and the underlying PCIe3-based I/O subsystem. (RDMA is known for its zero copy facilities);
- RDMA makes use Infiniband and Ethernet network interface cards (NICs) provided by Mellanox.

The data flow of these technologies is illustrated in Figure 1 (follow the red line to trace data flow).

<u>Figure 1: Data Flow Using the Microsoft/Mellanox/Violin Memory File Transfer Solution</u>


Memory
Violin Memory
Windows Flash Array
Traditional Memory
DIMMs attached via
Memory Channel
10GbE, iSCSI or 56 Gb FDR Infiniband
SMB Direct
Multichannel Interface to
Remote Data Memory Access
(RDMA)
Windows Server 2012 r2
Mellanox's FDR 56Gb/s InfiniBand and 40Gb/s Ethernet
with RoCE interconnect products
Memory
Violin Memory
Windows Flash Array
Traditional Memory
DIMMs attached via
Memory Channel
10GbE, iSCSI, or 56Gb FDR Infiniband
SMB Direct
Multichannel Interface to
Remote Data Memory Access
(RDMA)
Windows Server 2012 r2
Mellanox's FDR 56Gb/s InfiniBand and 40Gb/s Ethernet
with RoCE interconnect products

Source: Clabby Analytics, January, 2015

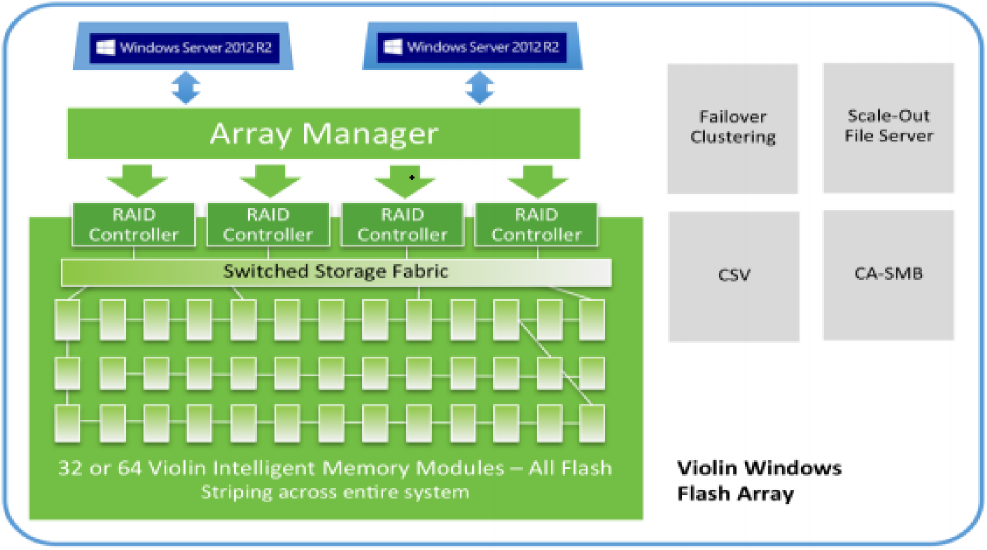
Violin Memory's Windows Flash Array

Violin Memory's WFA is an all-Flash array jointly developed by Microsoft and Violin memory to serve Microsoft's Windows Storage Server 2012 R2 environment. This array has been integrated with Windows at the kernel level – meaning that deep integration has taken place to improve WFA integration and increase array performance.

We do not support mirroring on the WFA, we do support DFS replication, but not mirroring in the strictest sense. Fault tolerance is built into the Flash Fabric Architecture, RAID configuration as well as full redundancy on all devices/components.

The array takes advantage of Violin Memory's Flash Fabric Architecture, an all silicon approach that uses patented Flash optimization algorithms implemented in hardware that operate at line speed (see Figure 2).

Figure 2: Violin Memory's Flash Array Architectural Design



Source: Violin Memory, January, 2015

The array core offers a resilient, highly available mesh of thousands of Flash dies that to continuously optimize performance and decrease latency. "Violin Intelligent Memory Modules" (VIMMs) organize these dies into intelligent Flash management units. A Flash Translation Layer helps manage these VIMMs, performing garbage collection as well as wear leveling and error/fault management services. Violin Switched Memory (vXM) manages array power efficiency and performance. Together VIMMs and the Flash Fabric Architecture work with Violin's vRAID (redundant array of independent disks) algorithm to increase reliability and reduce latency. All of these modules work in conjunction with Violin Memory's own Flash operating environment.

Violin Memory positions this array as a storage array that can serve a failover cluster; that can act as a scale-out file server; that can serve cluster shared volumes (CSVs) in a Windows Server 2012 R2 environment; and that can provide continuous availability services for a Microsoft SMB environment (CA-SMB).

From our perspective, however, this Microsoft SMB/Mellanox/Violin Memory configuration acts more like memory than storage. As we stated in the Executive Summary, DRAM (main memory) is always faster than NAND-based flash – but we also observe that the performance characteristics are similar in that both are

> ***much faster than traditional mechanical disk drives. In the specific case of DDR3 DRAM vs MLC NAND Flash (the FLASH used in the WFA), DDR3 DRAM is 17X more expensive than MLC NAND Flash in terms of $/GB. Accordingly, if the performance of MLC NAND is about 1/6 (or better) than that of DDR3, FLASH costs about 1/3 as much (on a performance adjusted basis).***

Microsoft's SMB Direct Environment

Simple Message Block (SMB) architecture first made its appearance in the 1980s as a means to transfer files (first released in Microsoft LAN Manager in 1987 – after which it was worked into Windows operating environments under different names including CIFS for "common Internet file system)). Successive generations of this architecture continually expanded functionality while improving performance. (For a complete history of SMB, see this Web page).

In 2013, when Microsoft introduced the second version of its Windows Server 2012 (known as Windows Server 2012 R2), several improvements were made to SMB including the ability of SMB Direct to support RDMA. To detect RDMA capabilities in the underlying network, SMB Direct uses its multi-channel facility – and if RDMA connections are found, SMB Direct creates multiple RDMA connections (two per interface) – enabling large volumes of data to be transferred.

What is special about RDMA is that it features "zero copy" – which means that data does not need to be copied between memory and data buffers. By not having to copy and buffer data, the impact of processing communications on a central processing unit (CPU) is greatly reduced because the CPU doesn't have to spend processing cycles handling communications overhead. As a result of not having to perform as much communications handling, processors are free to perform more useful work processing applications and data (a much better use of processor investment).

RDMA greatly improves communications flow by removing the need to process thousands to tens-of-thousands of commands and calls. For instance, using the tradition I/O flow model, device drivers are called, copy or pin source data is derived, memory mapped I/O needs to take place, acceleration needs to take place, polls and interrupts need to complete, copy or unpin results need to be correlated – and then, finally, data from the device driver can be delivered. This process represents a lot of processor management/communications overhead just to manage device interaction with a processor. RDMA speeds I/O to CPU communications by bypassing many of these obstacles – thus greatly improving I/O to CPU data passing speed – and allowing data to be directly placed into the memory of another computer.

In more practical terms, SMB Direct makes it possible to get more work out of processors – so fewer processors are required to execute a given workload. (One vendor claims a 30% improvement in CPU utilization due to the use of SMB Direct). Fewer processors means that fewer application and database licenses need to be purchased – saving enterprises REALLY BIG MONEY in terms of systems and applications/database license costs (remember: most software licenses are priced per CPU. Drive down the number of CPUs needed and the software license costs decrease accordingly).

> ***SMB Direct, and the underlying RDMA protocol, helps enterprises save REALLY BIG MONEY by reducing communications processing overhead on the processor. This in turn means fewer processors are needed to execute a given workload – and this means fewer servers and fewer software licenses are needed to execute that workload. From a performance perspective, SMB Direct and RDMA also leverage the full throughput of a network (thanks to reduced overhead and copying); and they streamline WAN file sharing (making the files respond at speeds that seem "local").***

The Mellanox Contribution

According to the Mellanox Web site, the company is a "leading supplier of end-to-end InfiniBand and Ethernet interconnect solutions and services for servers and storage. Mellanox interconnect solutions increase data center efficiency by providing the highest throughput and lowest latency, delivering data faster to applications and unlocking system performance capability. Mellanox offers a choice of fast interconnect products: adapters, switches, software and silicon that accelerate application runtime and maximize business results for a wide range of markets including high performance computing, enterprise data centers, Web 2.0, cloud, storage and financial services".

> ***Mellanox is a networking company, most noted for its high speed networking products***

Mellanox network solutions can support server and storage connections to 40Gb/s Ethernet with RoCE (RDMA over Converged Ethernet); and/or FDR 56Gb/s InfiniBand.

High-Value Solutions

As mentioned in the *Executive Summary*, the combined Microsoft/Mellanox/Violin Memory solution is geared to best serve application/data environments that are "I/O challenged", including large scale simulations, rendering, large scale software compilation and trading decisions.

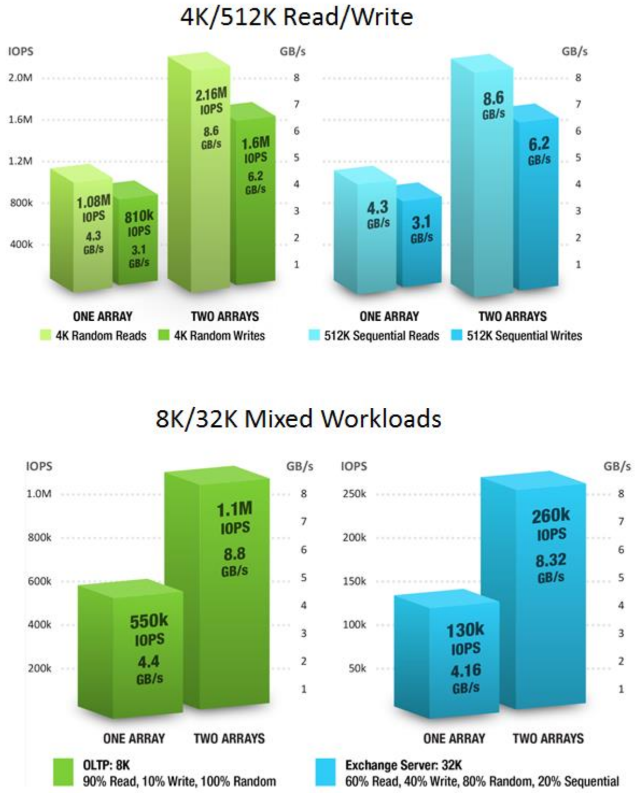
A recent Microsoft white paper describes some of the virtualization and scale-out file services advantages that can be found when using the combination of SMB Direct and Violin Memory's WFAs. According to this white paper, one particularly good high-value scenario involves using this platform to create virtualized Hyper-V (Microsoft's hypervisor virtualization solution) instances on scale-out file servers in order to store virtual disk files. Under this scenario, SMB Direct and WFA combine to support a new virtual hard disk (VHD) format known as VHDX. This format enables much larger file sizes than its VHD predecessor – and also provides data corruption protection, and helps optimize performance on large sector physical disks. SMB Direct helps administrators create scale-out, clustered, continuously available Hyper-V virtual server environments quickly.

From a performance perspective, Microsoft's white paper shows that "synthetic virtualized IO workloads running in Hyper-V VMs can linearly scale to over two million random read IOPS and over 8.6 GB/s sequential read bandwidth with two Violin WFA-64 arrays in a Scale-Out File Server Cluster. In this platform, 99th percentile latencies of 4.5ms can be achieved at a load of 2 million random read IOPS. For simulated OLTP IO traffic, 99th percentile latencies of 3.7-4ms can be achieved at a load of 1.15 million IOPS". This report concludes that "*the Violin WFA with its high performance, availability and scalability can easily keep up with customer's most demanding application SLAs while providing increased density and efficiency in a virtualized environment*".

> ***In other words, the Microsoft/Violin Memory combo can help enterprises create very dense virtualized server environments that offer outstanding read/write speed (especially when compared to mechanical disks), very low latency – and that offer support for larger file sizes and better reliability/availability features.***

Violin Memory provided us with two graphics that show the read/write performance characteristics under two scenarios: 1) a 4K random reads/writes scenario with 512K sequential reads; and, 2) an 8K/32K read/write scenario running an on-line transaction processing application – and an 8K/32K read/write scenario running an Exchange Server environment. These scenarios are illustrated in Figure 3 (next page).

Figure 3 – Read/Write Scenarios



Source: Violin Memory, January, 2015

How To Determine How Much Primary System Memory vs. Flash Memory to Use

IT buyers need to consider that primary systems memory (dual in-line memory modules – DIMMs) can be very expensive. Flash memory, by comparison, can be significantly less expensive (perhaps about a third the cost of primary memory). Another difference is that Violin Memory All Flash Arrays are persistent storage (no power backup required), whereas DIMMs used in system memory are non-persistent. Both types of memory can deliver data expeditiously to a processor. So the big question becomes: "is the speed advantage that primary memory has over Flash memory worth three times the cost?"

> ***One way to determine how much primary systems memory should be used to support a given workload involves a trial-and-error approach. Typically, a workload is assigned a block of memory. The amount of memory in this block can be overprovisioned – and then notched back until input/output (I/O) bottlenecks start to occur. Try augmenting primary systems memory with Flash-as-memory. The point at which the workload has enough memory to execute is considered the balance point.***

Summary Observations

In days long gone, systems makers used to improve server performance simply by increasing processor speed. But several years ago microprocessors hit physical limitations in the 5-6 GHz range. To overcome this performance limitation, microprocessor vendors started to build multi-core designs as a means to increase processing power. And, as a result, it is now common to find servers offered in dual-, quad-, octa- and even greater-core configurations.

But improving performance using multi-core processors is not the only way to accelerate systems performance. Further performance improvements can be achieved by:

- Speeding up the path between the input/output (I/O) subsystem and the processor;
- Improving memory and memory channel speed and performance; and,
- Using hardware accelerators (such as field programmable gate arrays [FPGAs], graphical processing units [GPUs], and other specialized processors).

Microsoft, Mellanox and Violin memory are using some of these approaches to speed file transfer from memory in one system to memory in another:

- With SMB Direct, more SQL Servers can be fed more data more quickly – meaning that results can be obtained more quickly;
- RDMA decreases communications overhead processing on CPUs.
    - More efficient processing means fewer servers are needed to execute a given workload (resulting in server acquisition cost savings for the enterprise).
    - Fewer servers (and more specifically, fewer CPUs) help lower software costs (because software is often priced by the number of CPUs in a given system).
        - Therefore, fewer CPUs = lower software license costs (in some cases *huge software licensing savings are possible*);
- Better file transfer facilities combined with faster networking (thanks to the use of Mellanox RDMA-enabled adapters and switches) yields better throughput and faster response times.
    - In many cases, response time for remote data feels "local".

Note that Violin Memory markets this solution as a Flash storage solution. From our perspective, the Flash array used in this architecture behaves more like extended memory – and ***represents a way to undercut expensive primary system memory – thus driving down the cost for application/database workloads that can benefit from large memory configurations***.

> ***Information technology (IT) buyers looking for an extremely high speed, scalable, low latency, highly available clustered Windows-based file server solution that can rapidly service virtualized Hyper-V and SQL Server environments need to evaluate this integrated Windows file serving solution. Further, enterprises that are experiencing I/O latency problems would be well served to evaluate this jointly developed solution.***

Clabby Analytics
http://www.clabbyanalytics.com
Telephone: 001 (207) 846-6662

January, 2015

Clabby Analytics is an independent technology research and analysis organization. Unlike many other research firms, we advocate certain positions — and encourage our readers to find counter opinions —then balance both points-of-view in order to decide on a course of action. Other research and analysis conducted by Clabby Analytics can be found at: www.ClabbyAnalytics.com.