SOLUTION BRIEF

# Plexistor Uses Mellanox Technology to Show SDM Can Handle Millions of Remotely Mirrored Writes at Low Latency

## Software Defined Memory overcomes fundamental limitation of traditional storage technology

In the last decade, flash-based solid state drives (SSDs) have been widely adopted for accelerating latency-sensitive applications. Recently, new device types were introduced that are orders of magnitude faster and run at memory or at near-memory speeds.

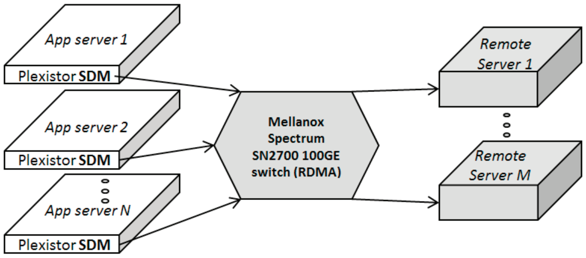
NVDIMM devices reside on the memory interconnect which is designed for low latency at high throughput. NVDIMM-N are NVDIMM devices that are fully mapped to the memory address space and can be accessed at cache-line granularity, unlike SSD devices which require their I/O events to be organized into larger blocks. Also, NVDIMM-N devices respond at near-memory speed, which makes the traditional operating system practice of caching data in main memory redundant.

NVDIMMs are faster, but also more expensive, than flash. Building a cost-effective solution requires mixing storage media types. Plexistor's Software Defined Memory (SDM) is designed from the ground up to marry the performance of persistent memory (NVDIMM-N) with the capacity of flash devices. SDM consumes fast memory resources alongside more-affordable SSD resources and presents them in an abstracted way through conventional APIs.

Plexistor's SDM eliminates the redundant block abstraction layer and collapses the multiple storage software layers into a single layer, which is tailored for modern multi-core processors. It further cuts latency and saves resources by allowing users to directly access the storage media without creating an additional cached copy. Plexistor's SDM eliminates redundant software caching and enables byte-level addressing. The net result provides near-memory performance levels at cost levels associated with Flash media.

Resolving the local storage performance issues moved the performance bottleneck to the next weakest link — to the network of highly available configurations. While SDM is proven to provide exceedingly low latency for local file-level access, in order for the benefits to also be achieved in high-availability configurations, the solution must be able to write locally to persistent storage media while also mirroring the written data and metadata to non-local destinations at various levels of throughput and to maintain the data even in the face of full node failures.

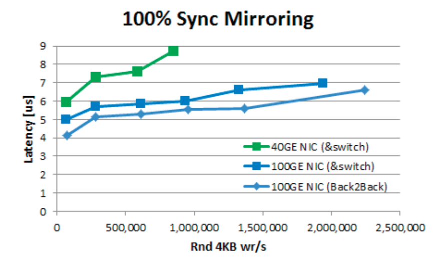
Working with Mellanox, the market leader for RDMA and 100GbE networking infrastructure, Plexistor proved that SDM can easily handle millions of remotely mirrored Writes per second at latencies as low as a few microseconds.

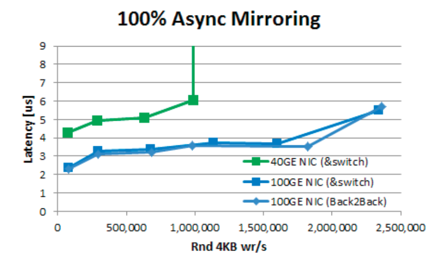


In recent testing, N single-socket E5-2667v3 application servers, running CentOS 7.2, were each installed with Plexistor SDM[1] software and a

[1] Plexistor's SDM is available at plexistor.com/download. The mirroring feature specifically discussed in this paper is not, however, publicly available yet.

high-speed RDMA-enabled ConnectX®-3 (40 GbE) or ConnectX®-4 (100 GbE) NIC from Mellanox. These "initiators" mirrored all data and metadata via a Mellanox Spectrum™ 100GbE SN2700 switch to M persistent Bricks, each comprised of persistent memory and flash, in order to reduce the solution cost to a minimum.





Using the synthetic FIO benchmark[2], latency was measured as a function of throughput for the most demanding workload: 100% random Write accesses. Each Write access was measured end-to-end, i.e., from before the user space application triggered the Write system call, via the context switches, until after it completed, resulting in a persistent local copy and a volatile confirmed/unconfirmed second copy of the data and metadata on the remote node.

In actual production environments, of course, users rarely if ever exceed one million Writes per second. However, the results of this testing reveal that even for such a stressful workload as a sustained throughput of 1M Writes per second with full mirroring, completion time latencies were measured at less than 4µs for asynchronous Writes, and less than 6µs for synchronous Writes. Such low latencies only become possible with the introduction of the Plexistor SDM cluster architecture's ability to harness the RoCE capabilities provided by the Mellanox switch and network adapters.

Also, to isolate the effect of the switch, the 100GbE tests were repeated "back-to-back" — directly connected with no switch between the local and remote servers. It is impressive to note that even for Sync Writes (with more bi-directional network traffic), the switch was responsible for less than 10% additional latency. Given that it is impractical in most production environments to pair all systems with back-to-back cabling, it is reassuring to know that leveraging the many benefits of a Mellanox switch — simplifying deployment, administration, scaling, and so on — will not significantly impact performance, if at all, for even the most demanding workloads.

Finally, these results also illustrate the obvious dramatic benefit for upgrading the network from 40GbE to 100GbE. Obviously, 100GbE offers 2.5X more throughput than 40GbE, but it is also important to note, as shown here, that even before the 40GbE results had saturated the network with 1M Writes, 100GbE latencies per Write were consistently lower by at least 1.5µs.

Often, application environments must cope with high rates of incoming data (input) that must be logged as well as processed, and all results (output) must also be saved. All Writes must be made safely without errors or excuses, while performance must be optimal. Maintaining performance while mirroring all Writes is a common goal of strategic approaches, but until now there has been no cost-effective way to avoid a painful tradeoff.

This dynamic tension can be found in the high-stakes worlds of financial trading, fraud detection systems for transaction processing, military command & control, and many others. In a wider variety of database environments, similar requirements for low latency and data safety are also common, but with lower transaction rates.

The Plexistor Cluster Architecture, in which SDM leverages Mellanox's implementation of RDMA over 100GbE, is an ideal solution for exactly these industries and environments.

## About Mellanox

Mellanox Technologies is a leading supplier of end-to-end Ethernet interconnect solutions and services for enterprise data centers, Web 2.0, cloud, storage and financial services. More information is available at www.mellanox.com.

## About Plexistor

Plexistor is the first to implement the new and innovative Software Defined Memory architecture, enabling enterprises to process huge amounts of data at near-memory speed and to respond to customer needs with faster response time. Learn more at www.plexistor.com.

[2] *fio --invalidate=1 --sync=1 --bs=4k --size=1g --numjobs=16 --rw=randwrite --allrandrepeat=1 --ioengine=psync --thinktime=<TT> --thinktime_spin=<TT> --directory=/mnt/ --name=test*