EMC POWERPATH
Optimized IO Multipathing for All Flash Arrays

JANUARY 2015

All-flash arrays are changing the datacenter for the better. No longer do we worry about IOPS bottlenecks from the array: all-flash arrays (AFA) can deliver a staggering amount of IOPs. AFAs with the ability to deliver hundreds of thousands of IOPs are not uncommon. The problem now, however, is how to get the IOPS from the array to the servers. We recently had a chance to see how well an AFA using EMC PowerPath driver works to eliminate this bottleneck—and we were blown away. Most comparisons with datacenter infrastructure show a 10-30% improvement in performance; but, the performance improvement that we saw with PowerPath was extraordinary.

Getting bits from an array to server is easy —very easy, in fact. The trick is getting the bits from a server to an array in an efficient manner when you have many virtual machines (VM) on multiple physical hosts that are transmitting the bits over a physical network with a virtual fabric overlay; this is much more difficult. Errors can get introduced and must be dealt with, the most efficient path must be obtained and established, re-evaluated and reestablished continually, and any misconfiguration can produce less than optimal performance. In some cases, this can cause outages or even data loss. In order to deal with the "pathing," or how the I/O travels from the VM to storage, the OS running on the host needs a driver, or in the case where multiple paths can be taken from the server to the array, a multipathing driver needs to be used to direct the traffic.

Windows, Linux, VMware and most other modern operating systems include a basic multipath driver; however, these drivers tend to be generic and not code optimized to extract the maximum performance from an array and come with only rudimentary traffic optimization and management functions. In some cases these generic drivers are fine, but in the majority of datacenters the infrastructure is overtaxed and its equipment needs to be used in the most efficient manner possible. Fortunately, storage companies such as EMC are committed to making their arrays work as performant as possible and spend a considerable amount of time and research to develop multipathing drivers optimized for their arrays. EMC invited us to take a look at how PowerPath, their optimized "intelligent" multipath driver, performed on an XtremIO flash array connected to a Dell PowerEdge R710 server running ESX 5.5 while simulating an Oracle workload. We looked at the results of the various tests EMC ran comparing PowerPath/VE multipath driver against VMware's ESXi Native Multipath driver and we were impressed—very impressed—by the difference that an optimized, multipath driver like PowerPath can make in a high IO traffic scenario

Multipathing, Native Multipathing and PowerPath and PowerPath/VE

Datacenters, in order to insure reliability and increase performance, have best practices in place to minimize "single point of failures." For example, and as shown in Figure 1, we see that multiple hosts are connected to multiple switches via various paths to multiple storage processors (SP) to a storage

87 Elm Street, Suite 900 · Hopkinton, MA 01748 · T: 508.435.2556 · F: 508.435.2557

array. This fabric, along with appropriate software drivers, ensures that if a component or path goes down, another path will be available to enable data continuity in a datacenter.

Multipathing is a generic term and is simply a technique used to connect a server to its storage over multiple paths. Because it is aware of all the paths connecting the server to storage, multipathing allows increased performance and reliability of those connections. In its simplest form, it sets up the connections in an active/passive configuration where all the traffic is passed over one connection; if that connection goes down, it switches over to the other path. As active/passive configurations are not very efficient, engineers have come up with more sophisticated techniques to utilize the multiple paths between the storage and server.

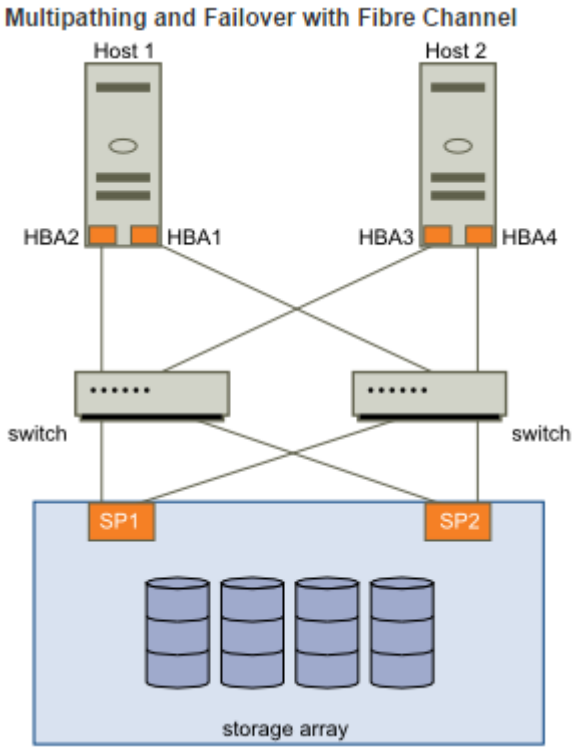

Figure 1 (Source: VMware)

In order to connect storage to an ESX server in a more efficient manner, VMware, with the cooperation of its storage partners, developed a Native Multipathing Plug-in (NMP). Generally, the VMware NMP supports all storage arrays listed on the VMware storage HCL and provides the default path selection algorithm based on the array type. However, each array has its own idiosyncrasies, and each has its own algorithms and ways of handling path failover, as well as determining which physical path is used to issue an I/O request to storage. VMware NMP was designed to distribute the load over all the available paths and provide failover protection in the case of path, port or HBA failure, but it has not been fully optimized to work with the controllers in a storage systems. In such systems, LUNs are fully owned or preferentially managed by one controller. With the latest vSphere integrations, the vendor can influence path selection; however, less than optimal path selection may result in paths moving between controllers. This movement can cause unnecessary cache repopulation or other internal array gyrations in order to transfer active volume ownership across controllers, or "LUN trespassing." VMware's NMP Round Robin policy does not have the intelligence that PowerPath has as PowerPath uses testing and diagnostics to continually monitor an environment to determine the optimal path for queuing requests and will adapt to current conditions.

Noting the inefficiencies in VMware's NMP driver, EMC developed a set of drivers specifically designed to overcome these limitations and improve the performance and reliability of the data passing between an array and a server. EMC developed the PowerPath family of products optimized specifically for Linux, Microsoft Windows, and UNIX Operating Systems as well as PowerPath/VE for VMware vSphere and Microsoft Hyper-V hypervisors.

To fully use all the pathways from the server to the array, PowerPath has patented "intelligent" algorithms to optimize and load-balance the paths to and from an array. This "intelligence" guarantees that some paths will not be overloaded while others are underutilized, a condition we have witnessed in the past using other multipath drivers. PowerPath is constantly monitoring the loads on the path and will shift traffic as needed to balance the load.

Cables fail and adapters go bad; this can lead to path failure. PowerPath monitors path health and tests for path failure. In the case of a path failure, PowerPath will redirect all outstanding and any subsequent I/O requests to alternative paths. PowerPath is continually probing paths and once a path is restored it will, without any user intervention, begin to use the path again. This ensures data continuity and integrity.

Proliferation of all-flash arrays and bottlenecks

In the past five years the price of flash has fallen considerably and the technology has matured to the point where AFAs are no longer considered "exotic" niche products; companies of all sizes have been acquiring them and they have been acquiring a lot of them. All-flash array sales have exceeded even the most optimistic expectations. Even small companies are embracing AFA's for applications that are IOPS intensive or departments such as test and development that appreciate them due to the data efficiencies that some AFAs have. There have been reports of large companies that are starting to implement "all flash datacenters." The growth of AFAs has been explosive and they are no longer reserved for fringe applications.

POWERPATH – NOT JUST FOR VMWARE AND EMC ARRAYS

The PowerPath family of products isn't just for EMC Arrays and vSphere. The PowerPath family of products is supported on Linux, Microsoft Windows, and UNIX as well as VMware vSphere and Microsoft Hyper-V. Supported storage includes all EMC block arrays as well as qualified arrays from all the major vendors.

With this widespread acceptance of AFAs, and the increase in performance they have enabled, datacenters are beginning to show a weakness in dataflow. Companies that had demanding applications (such as Oracle databases) in the past were bound by storage or CPU, but with the advent of AFAs (and servers that support 32 cores), they are now being bound by a new constraint: the connection between the server and the array, and specifically, its inability to pass data between the two so they can be utilized to their full potential.

XtremIO Flash Array

XtremIO was an early entrant into the all-flash array market and was acquired by EMC in May of 2012. This acquisition let EMC jump into the AFA market with a well-accepted and highly regarded product that has helped make EMC a leader in the AFA market. XtremIO architecture is based on an "X-Brick" scheme. Each X-Brick has either 5 TB, 10 TB or 20 TB of raw flash capacity and clusters of X-Bricks scale linearly in capacity and performance; each X-Brick adds both capacity and performance. This allows XtremIO to scale out in a linear fashion and provide a consistent and predictable I/O throughput rate. XtremIO differentiates itself from other all-flash arrays by embedding capacity efficiency technologies that can reduce data storage requirements while still providing key features like full VMware VAAI integration and consistent performance.

An XtremIO array can be composed of up to 6 bricks for up to 120TB of raw flash capacity that, when deduplication, compression and other space savings techniques are taken into account combined with the XtremIO scale-out architecture, can equate (based on EMC calculations) to petabytes of effective logical capacity. The capacity of the cluster is impressive, but what is more impressive is its ability to deliver over one million IOPS on a pure read scenario or more than six hundred thousand IOPs in a 50/50 split of random reads and writes. The average latency of these reads and writes is delivered with sub-millisecond response times. This kind of performance can greatly enhance an application's performance but only if the IOPS can reach the server in an efficient manner. Most

87 Elm Street, Suite 900 · Hopkinton, MA 01748 · T: 508.435.2556 · F: 508.435.2557

multipath drivers in the market today were simply not optimized for this type of activity but PowerPath was engineered explicitly to support this type of load.

Test configuration

Although EMC ran the tests presented below the Taneja Group spent a considerable amount of time doing a deep review of the tests results presented below We discussed the test methodology and process with the EMC engineers that did the actual testing and feel comfortable with the results that they obtained.

All hardware used for the testing can commonly be found in a datacenter. All of the test equipment was set up using publicly available best practices.

Server: Dell PowerEdge R710, 2 Quad (8 CPU) Xeon E5530 2.4GHz, 12 GB of RAM, 4 HBA ports (Qlogic, 8Gbps)

Hypervisor: ESX 5.5 GA (build 1331820)

Multipath software: NMP (comes with ESX 5.5), PowerPath/VE = 59 SP1 – No special or additional tuning was performed on either software package

Array: XtremIO array (single X-Brick) with 2.02 microcode, 4 Fibre Channel storage ports (8Gbps per port), 4 LUNs (5 GB per LUN), 25 SSD drives

Switch: Brocade DS 300

Performance testing: VDbench running on Red Hat Linux 6.2 VM, number of threads: 60 per LUN (240 total)

Test results

Workloads such as VDI, databases and HPC can be IOPS intensive, and as such demonstrate weakness in the data flow between a server and array. VDI tends to be "bursty" during boot storms and AV scans, whereas high performance databases such as Oracle and Sybase can continually push a datacenter infrastructure to its limits and making its weakness more apparent. To test PowerPath, a load was needed to push it to its limit. Accordingly, a VDbench benchmarking test was run which replicated the load of an active database and Exchange server.

Different tests were run to see how PowerPath compares to VMware NMP in high IO load situations. The tests replicated five different environments that are commonly encountered in a datacenter: Exchange, an Oracle OLTP, a Decision Support System (DSS), a data warehouse, and a Sybase OLTP. Each of these loads has slightly different characteristics, and in each case, data is moved in a slightly different manner. The one thing that they did have in common was that they all performed much better with PowerPath than with NMP. At this point we need to note that only specific workloads were tested during the benchmarking and, of course, results will differ depending on your workload and hardware configuration.

Figure 2 and Figure 3 show the throughput measured during the IO jamming tests of the NMP and PowerPath drivers. Apples to apples, the throughput of the PowerPath driver was 1.5-6 times better than that of NMP. EMC attributes this dramatic increase in bandwidth to PowerPath's efficient use of all the available paths to move the data from the server to the array. Before AFAs became commonplace, a multipath driver's inability to move data as efficiently as possible would have been masked by an array's inability to supply the data. Now we have all-flash arrays that unmask these weaknesses and can overwhelm a multipath driver, resulting in sub-optimal performance, which can severely hobble business critical applications.

87 Elm Street, Suite 900 · Hopkinton, MA 01748 · T: 508.435.2556 · F: 508.435.2557

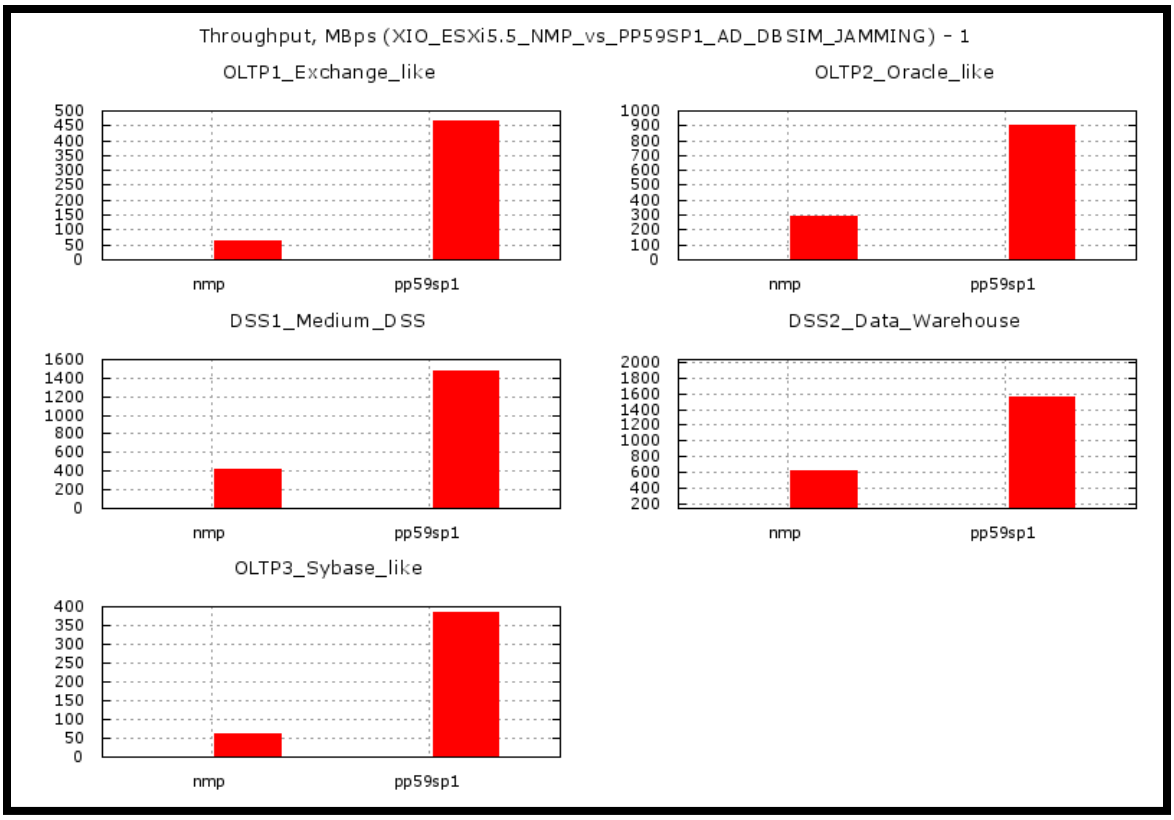

Figure 2: Throughput, in MBps, of PowerPath/VE compared to NMP driver

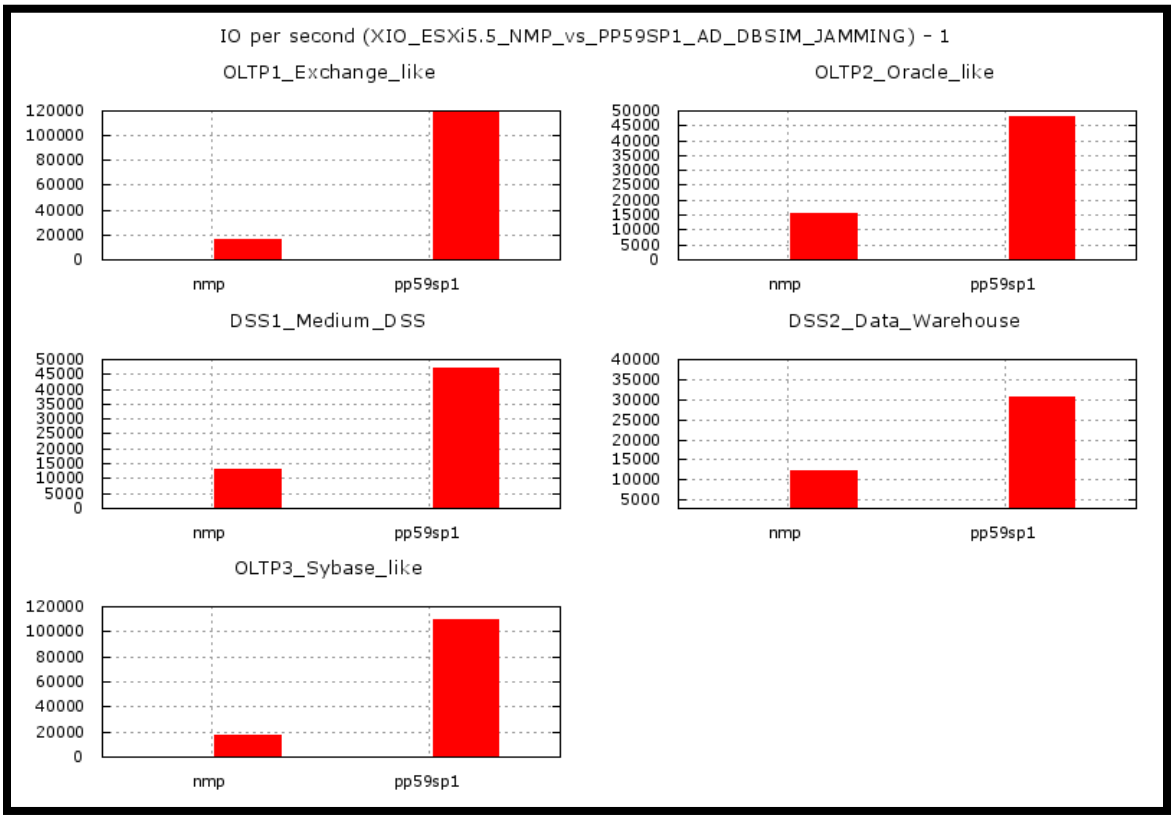

Figure 3: Throughput, in IOPS, of PowerPath/VE compared to NMP driver

87 Elm Street, Suite 900 · Hopkinton, MA 01748 · T: 508.435.2556 · F: 508.435.2557

The next metric that we looked at Figure 4) was the latency. The difference between the two drivers surprised us; we thought that the latency experienced by the application would be near identical and would be heavily dependent on the capabilities of the array, but this was not the case. To understand this better we talked to several EMC engineers. What we learned was that the overall latency (the time difference between a VM asking for something and getting it) can be poor even though the array is performing at a much lower latency. Clearly, a high latency array will deliver high overall latency but a low latency array is not guaranteed to deliver a low overall latency. This has to do with the way an I/O is handled as it is passed through the queues and/or buffers on the virtual machine, hypervisor, server's HBA, switch (ingress and egress), target's HBA, array controller and the storage media. If any of these queues or buffers becomes backed up, they become a choke point and the latency will increase. Traditional non-flash storage arrays had higher latency and often did not expose the weaknesses in a multipath driver's inability to pass data in an efficient manner. All-flash arrays can, and do, push the pathing software to its limit and now we are finding that the array is not the weakest link in the data path chain.

Jamming Tests

The jamming tests which resulted in the performance benchmarks seen in this paper were created by using IOMeter to simulate various scenarios. During the jamming tests IO paths were overwhelmed with IO requests. These tests were designed to simulate real world database and data warehousing applications. EMC PowerPath/VE displayed performance improvements in this specific scenario but please note that performance will vary depending on the specific environment and IO load conditions.

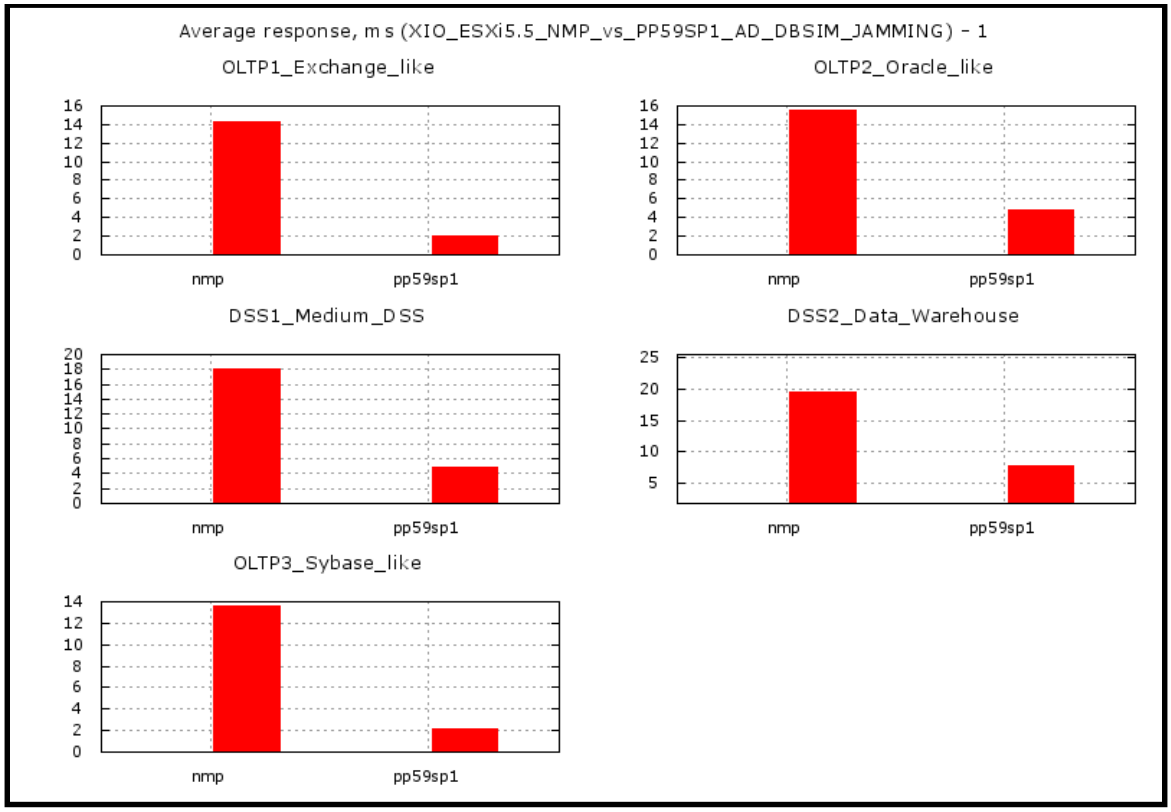

Figure 4: Latency, in ms, of PowerPath/VE compared to NMP driver

87 Elm Street, Suite 900 · Hopkinton, MA 01748 · T: 508.435.2556 · F: 508.435.2557

In the benchmarking tests with PowerPath/VE we saw a reduction in latency of up to 86% over NMP. This is substantial. As shown above we can see, with all other components being equal, that an optimized multipath driver can make a big difference in the performance of a demanding application such as an Exchange server or an Oracle database. Assuming the results of these tests are broadly applicable to a range of IO intensive applications with PowerPath, not only will transactions be handled faster, but also more data can be delivered to and from the server.

Indications that your array is being bottlenecked by your multipath software

Identifying bottlenecks in your existing environment caused by multipath performance can be tricky. Below are a few indications that your array may be suffering from less than optimal performance due to your multipath driver.

- Your array stats are considerably less than their stated capacity. Make sure that you are supplying the array with enough server I/O workload to push throughput to capacity
- Queuing delays. Native Multipathing (NMP) uses round robin, which is unable to determine congestion on backed-up I/O queues. NMP keeps sending the IOs down the same path regardless of the queue depth.
- Latency increases. The same I/O queuing can affect response time which is another telltale sign
- Some paths can be ok for certain I/O sizes but not others. Round Robin does not distinguish this, which can result in an increase in I/O errors, retries, and latency (see above).
- Sometimes storage front-end processors can be overloaded if the IOs are not balanced properly throughout the system. General system imbalances are due to application workload increases and random I/O access. All can be exacerbated by poor multipathing.
- Check the ESX stats – VMware has written many excellent articles on using esxtop to gather and interpret statistics that can be used to diagnose multipath issues. It takes time and commitment to master esxtop but the payoff can be a deeper understanding of your data path.

Just because you aren't seeing or experiencing any indications of bottlenecks at this instant in time doesn't mean you haven't experienced multipath bottlenecks in the past and won't see them in the future. When you do run up against data path bottlenecks it may be at the most inopportune time—when you need the capacity to support business-critical application performance.

Taneja Group Opinion

We knew through anecdotal evidence that PowerPath delivered better performance under load than NMP; we hadn't seen it quantified before, but we didn't expect to see such a dramatic difference between PowerPath and NMP. As mentioned at the beginning of the paper, we consider a 10-30% performance gain commendable and worthy of mention; to see a 150-600% improvement is outstanding. EMC has done a brilliant job of identifying and understanding a problem and then creating a high-performance solution for it.

All-flash arrays are not inexpensive but they do offer good value for the money if they are fully utilized. Many datacenters focus on the performance of their servers and arrays but forget about the connection between them. If you are running applications like Oracle, Exchange or other demanding workloads, and your application performance is suffering, there is a good chance that your existing multipath driver has starved your array in the past and will continue to do so in the future. Technologies like PowerPath can allow you to fully utilize your resource and get the full benefit from your datacenter assets.

87 Elm Street, Suite 900 · Hopkinton, MA 01748 · T: 508.435.2556 · F: 508.435.2557

So, all in all, we found that PowerPath delivers the superior I/O path management and performance that is needed in today's flash environment. Its automated data path failover and recovery scheme ensures business critical applications are always available. PowerPath shouldn't be thought of as just a high performance conduit between ESXi and an XtremIO array, as most popular OSes and a wide range of arrays from various vendors could benefit from it as well.

If you are using an all-flash array in your datacenter and feel like you are not getting the maximum performance from your applications, you owe it to yourself to take a close look at your existing multi-path approach. If you find your existing multi-path performance is falling short, consider an optimized solution such as EMC PowerPath in order to extract the maximum value from this valuable resource.

87 Elm Street, Suite 900 · Hopkinton, MA 01748 · T: 508.435.2556 · F: 508.435.2557