ESG Lab Spotlight

EMC ScaleIO: Proven Performance and Scalability

Date: May 2015 **Author:** Aviv Kaufmann, ESG Lab Analyst

Abstract: *This ESG Lab Spotlight highlights the results of performance tests run on ScaleIO software on commodity hardware, demonstrating extremely high IOPS performance and near-linear scaling to 128 nodes.*

The Challenges

No matter how big or small your IT organization is, the challenge is the same: Do more with less. With budgets flattening and end-users continuing to demand services that require greater performance and agility from the supporting IT infrastructure, companies must find innovative ways to create scalable solutions. With capital and operational expenses constantly under scrutiny, IT admins must be conscious of the impact that their technology choices have on the bottom-line and be ready to defend their choices.

Traditional IT infrastructures consisting of silos of dedicated servers, networking, and storage are extremely complex and expensive to purchase and maintain. To limit capital and operational expenses, many organizations have turned to cloud-based solutions such as infrastructure-as-a-service (IaaS). ESG research indicates that 65% of enterprise organizations currently used or planned to use IaaS in 2014, with another 20% indicating that they were perhaps interested in using it later.[1] Many organizations leverage the cost effectiveness of these solutions temporarily as they struggle to design an in-house infrastructure that can deliver the same level of cost effectiveness, flexibility, and scalability to their end-users while also satisfying the demands of today's high-performance applications.

The Solution: EMC ScaleIO

EMC ScaleIO is a flexible software-defined solution that leverages host-based internal storage to create a scalable virtual SAN with performance that is comparable to or better than an external SAN, at a fraction of the cost and complexity.

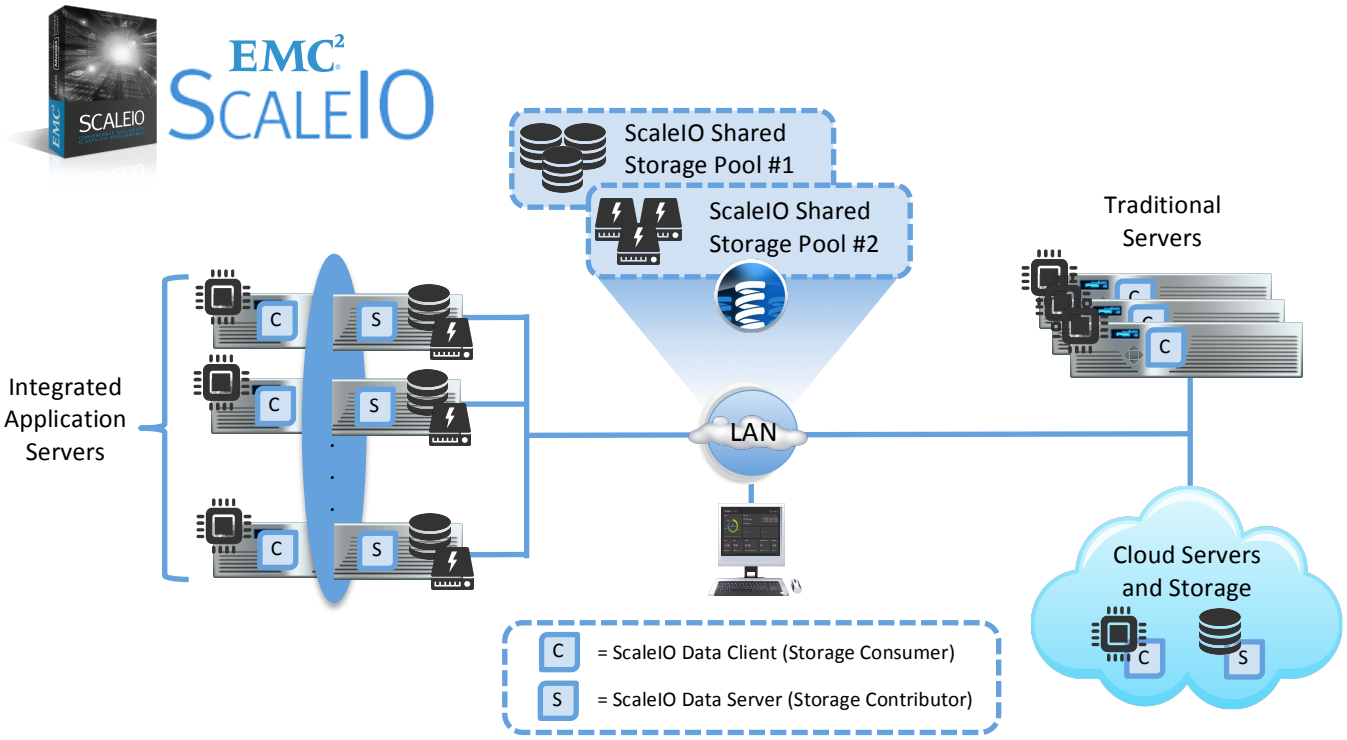
ScaleIO works by installing lightweight software components on application hosts. Application hosts contribute internal disks and any other direct attached storage (DAS) resources to the ScaleIO cluster by installing the ScaleIO Data Server software (SDS). Hosts can then be presented volumes from the ScaleIO cluster by leveraging the ScaleIO Data Client (SDC). These components can run alongside other applications on any server (physical, virtual, or cloud) and can leverage any type of storage media (disk drives, flash drives, PCIe flash cards, or cloud storage).

ScaleIO can be deployed as storage only or can greatly simplify the infrastructure by converging the storage, compute, and networking resources into a single scalable building block that is easy to manage. Capacity and performance of all available resources are aggregated and made available to every participating ScaleIO server and application. Storage tiers can be created with media types and drive types that match the ideal performance or capacity characteristics to best suit the application needs.

Storage and compute resources can be added to or removed from the ScaleIO cluster as needed, with no downtime and minimal impact to application performance. The self-healing, auto-balancing capability of the ScaleIO cluster ensures that data is automatically rebuilt and rebalanced across resources when components are added, removed, or failed. Because every server and local storage device in the cluster is used in parallel to process I/O operations and protect data, system performance scales linearly as additional servers and storage devices are added to the configuration.

[1] Source: ESG Research Report, *2014 Public Cloud Computing Trends*, March 2014.

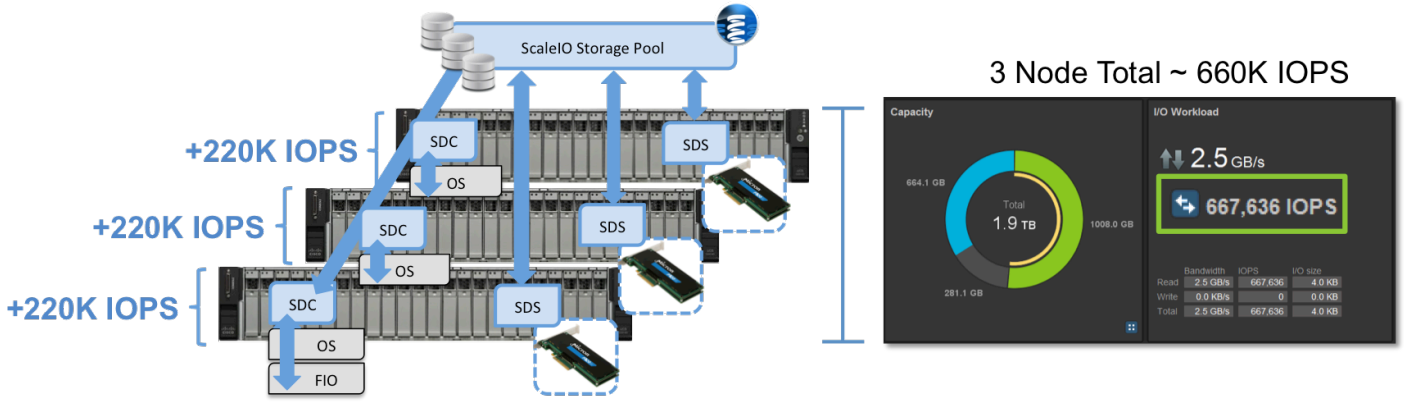
Figure 1. ScaleIO: Turning Commodity Hardware into a Converged, Elastic, Scalable Infrastructure



ESG Lab Validation of Performance and Testing Methodology

ESG Lab began by validating the results of performance tests on a minimal three-node ScaleIO cluster made up of commodity Cisco UCS C240 M3S servers, each populated with a single Micron P320 700GB PCIe flash storage device. The popular and public benchmarking tool FIO was used to generate a random, 4KB read workload. The three-node cluster was able to achieve a total of 660K read IOPS, for an impressive average of about 220K IOPS per node. Figure 2 shows the results of the three-node performance tests with PCIe flash devices.

Figure 2. Results of 4KB Random Read Testing on Three-node ScaleIo Cluster with P320 PCIe Flash Devices

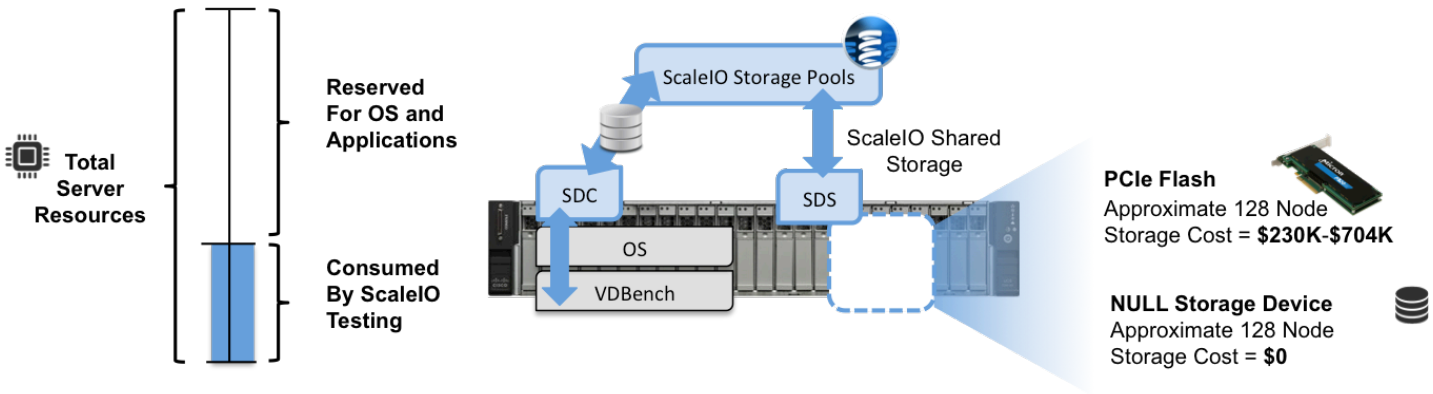


Next, ESG reviewed the test methodology proposed by EMC performance engineers to demonstrate performance scalability to a large number of ScaleIO nodes in a single cluster. One hundred twenty eight identically configured Cisco UCS C240 M3S servers were procured for performance testing. However, populating this many servers with enough SSDs or PCIe flash cards to drive the nodes to saturation was cost prohibitive for a single test. In addition to the cost of the

servers, ESG Lab estimates that, at 128 nodes, this test would require well over half a million dollars for storage costs alone.

To limit the cost of the test, EMC Performance Engineers proposed that, instead of the SDS providing storage from an expensive PCIe flash card, storage would be provided by the SDS to the ScaleIO Storage Pool from /dev/NULL. This operation would require that no actual data be read or written from a physical storage device, but would otherwise be identical in operation to tests done with PCIe flash storage devices and include all software and network overhead. This methodology would simulate an extremely fast flash-like device capable of stressing all ScaleIO components. As illustrated in Figure 3, the test would consume nearly all of the cores allocated to SDS (storage related) functions, while still reserving plenty of CPU resources for the OS and applications.

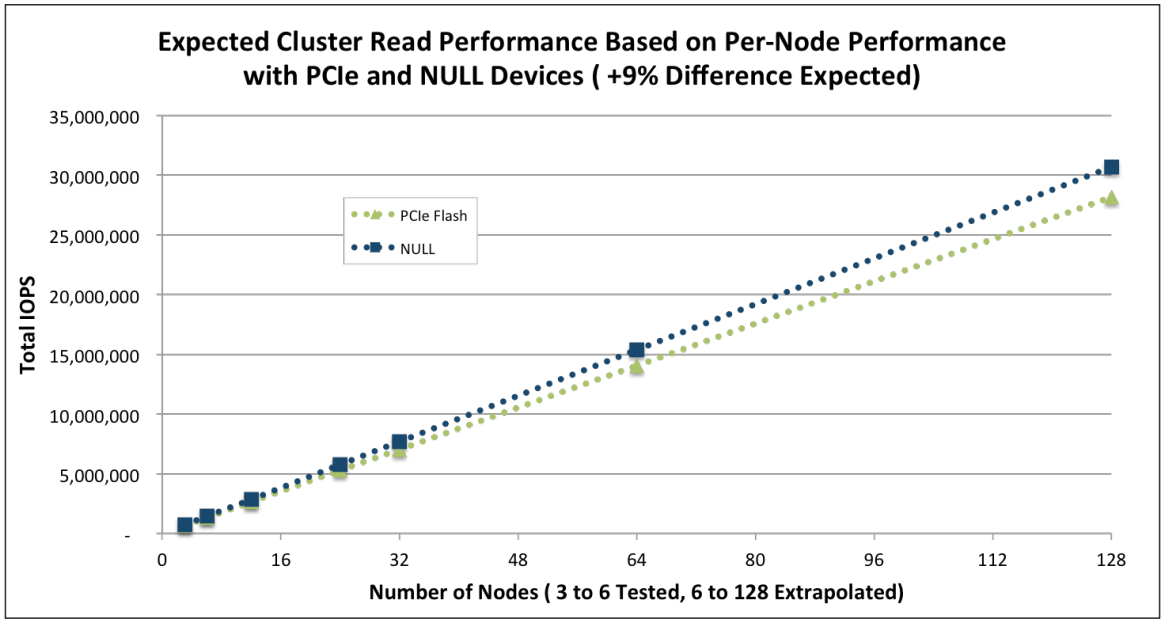
Figure 3. Using NULL Devices to Cost-effectively Measure Performance and Scalability



ESG Lab audited the results of performance tests that compared per-node cluster performance when storage was provided from a single Micron P320 PCIe storage device and when storage was provided from a NULL device. The tests were performed on three-node and six-node clusters to ensure that results were consistent and scalable. Two 4KB workloads were tested, consisting of 100% reads and a mixed workload of 70% reads and 30% writes. The *irqbalance* service was disabled using standard Red Hat Enterprise Linux commands to ensure an optimal balance of CPU interrupts between the SDC and SDS processes. This is a common option available to any consumer of the RHEL operating system.

Results of the read-only tests showed that there was little difference in the maximum IOPS achievable through a single node when storage is supplied from PCIe flash (~220K IOPS per node) compared with when storage is supplied by NULL devices (~240K IOPS per node). This small 9% difference proved consistent between the three-node and six-node results. The mixed workload results showed a slightly larger differential of about 15%. ESG Lab firmly believes that while we cannot rely on the results of NULL devices to accurately represent the response time of a physical storage device, it is reasonable to assume that commodity storage devices such as a PCIe flash card or SSDs could deliver similar levels of IOPS scalability. The results of the three-node and six-node tests were then used to extrapolate node scaling results to 128 nodes. The results are shown in Figure 4.

Figure 4. Expected Results of Node Scaling Tests Comparing PCIe Flash with NULL Devices



Why This Matters

Performance testing of node scalability for extremely large clusters requires an extremely expensive infrastructure. ScaleIO, like its target customers, was looking to leverage an intelligent alternative to get the most for its investment dollar while demonstrating extreme levels of scale.

ESG Lab validated that the results obtained using ScaleIO volumes configured on NULL devices could be used to reasonably estimate the IOPS performance and scalability of ScaleIO nodes populated with a PCIe flash card or SSDs, avoiding the prohibitive cost and allowing more of the test budget to be spent on additional nodes rather than flash storage devices.

ESG Lab Performance Testing

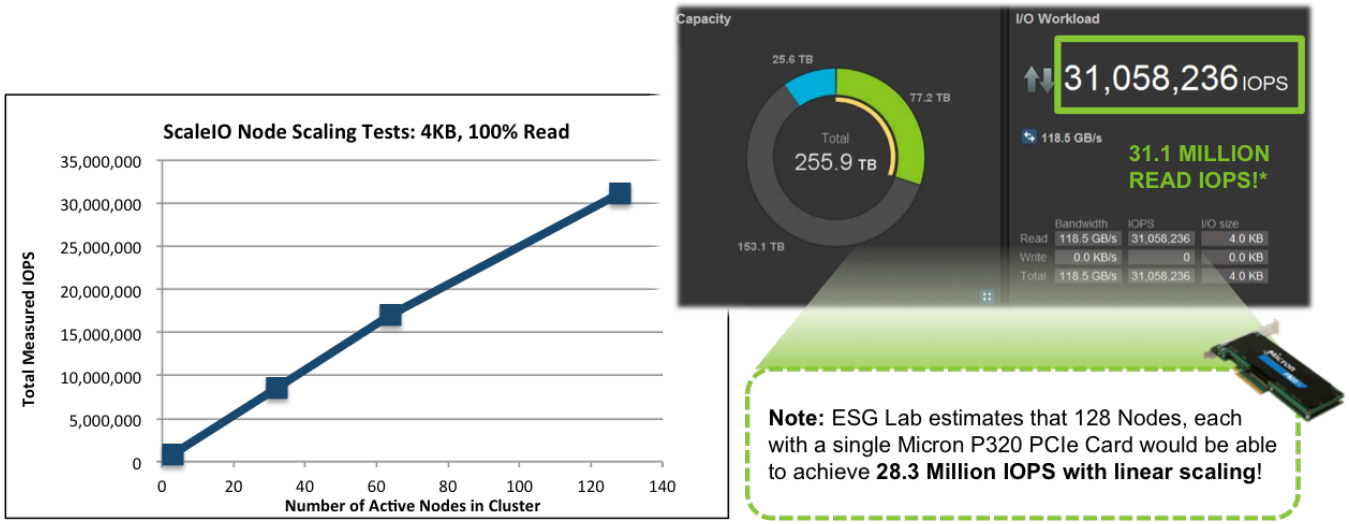
ESG Lab validated the results of performance testing on up to 128 identically configured ScaleIO nodes. Each Cisco UCS C240 M3S node used in the test setup contained two ten-core 2.8Ghz Intel Xeon E5-2680v2 CPUs with hyper-threading enabled and 64GB of RAM. To better balance the handling of interrupts, for this test, the RHEL IRQbalance service was disabled using parameters available to any end-user. ESG Lab also audited out-of-the-box performance results with no IRQbalance tuning and estimates that the OS-dependent IRQbalance tuning accounted for up to a 25-30% performance increase per node by assigning network interrupts to a single socket.

FIO was used to generate and measure the performance of both a 4KB random reads workload and a mixed 4KB random workload consisting of 70% reads and 30% writes to ScaleIO volumes. The ScaleIO volumes were configured from a ScaleIO pool built by presenting two devices from each SDS. As ESG Lab validated previously, a node that is configured with a single Micron P320 PCIe card or SSDs would be expected to achieve 91% of the random read IOPS performance, and 85% of the 70/30 mixed workload IOPS performance that was achieved by testing to NULL devices.

Performance was tested for clusters consisting of three nodes, 32 nodes, 64 nodes, and 128 nodes. For the 100% random read tests, ESG Lab witnessed almost completely linear scaling up to 128 nodes. Each node provided roughly 240K IOPS for a total cluster throughput of over 31 million IOPS. Although this impressive performance was achieved to NULL devices, ESG Lab estimates with high confidence that had these nodes been populated with a single Micron P320 PCIe flash storage device or SSDs, the 128-node system would achieve roughly 28.3 million IOPS. This performance,

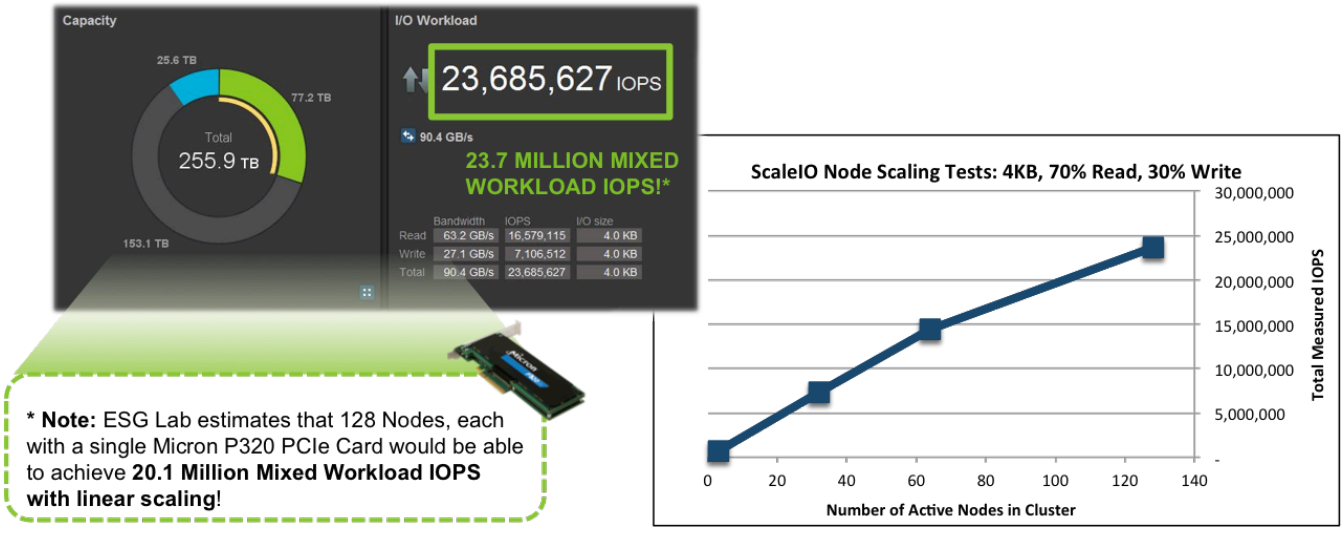
achievable with commodity hardware, is magnitudes higher than the best IOPS performance demonstrated by storage vendors to date. The results of the 100% read node scaling test are shown in Figure 5.

Figure 5. ScaleIO: Results of 128-node Scaling Tests with 100% Random Reads



Next, ESG Lab validated the results of node scaling performance tests with the mixed 70% read, 30% write workload. Once again, the results scaled near-linearly up to 128 nodes with each node averaging around 220K IOPS until dropping off slightly to around 185K IOPS per node during the the 128-node test. The 128-node system achieved roughly 23.7 million mixed workload IOPS and could be expected to provide over 20 million mixed workload IOPS when configured with physical flash storage devices. Figure 6 shows the results of the mixed workload node scaling tests.

Figure 6. ScaleIO: Results of 128-node Scaling Tests with Mixed Workload



The impressive scalability of the ScaleIO cluster validated by ESG Lab with commodity hardware certainly places ScaleIO at or near the top of the price/performance lists of competitive offerings. EMC can quickly demonstrate impressive competitive advantage and total cost of ownership (TCO) savings of around 60% versus traditional SAN using internally generated tools. ESG Lab looks forward to validating these tools and perhaps performing our own TCO analysis sometime in the near future.

The Bigger Truth

Today's cloud-scale data centers require a simple, agile, cost-effective, and, most importantly, scalable storage infrastructure. Scale-out storage systems have proven very successful in healthcare, financial, media, and communications data centers, and more IT organizations are looking to deliver the improved agility gained through building their own cloud-provider-like infrastructure-as-a-service (IaaS) IT model. High-performance, scale-out storage systems can be costly, and ESG research has found that purchasing technologies with improved return on investment (ROI) is clearly one of the top considerations in justifying IT investments as reported by survey respondents over the past two years.[2] With ScaleIO, EMC allows organizations to create a high-performing and highly available scale-out infrastructure using the commodity hardware components that best fit their requirements and budget.

ESG Lab validated the convergence, performance, elasticity, and reliability of a ScaleIO solution in 2014 and was impressed with the simple, scalable performance that was achievable using commodity hardware components. In this report, ESG Lab demonstrated that an eight-node ScaleIO cluster in a high-performance Oracle RAC configuration was able to produce over 830K query SLOB IOPS and over 21.2 GB/sec when performing table scans. This performance was on par with or better than some Oracle RAC converged solutions costing twice as much. ESG Lab also audited the results of some performance testing at scale in a QA environment where the 53-node cluster scaled well and produced some eye-popping numbers.

By testing to NULL devices rather than high-performance PCIe flash devices, EMC was impressively able to demonstrate near-linear performance scaling up to 128 nodes in a single cluster. The 128-node cluster was able to achieve over 31 million random read IOPS, and 26.7 million mixed workload IOPS (70% read, 30% write). To put this level of performance into perspective, feel free to look at the highest number of IOPS that has been demonstrated and published by storage vendors to date. What's perhaps more impressive is that the number of nodes under test in the cluster was limited only by the availability of hardware and budget, and not by the lack of ability to scale beyond this count.

Competitors will inevitably be tempted to dismiss the performance scalability testing results as irrelevant because they were achieved by testing to NULL devices. However, it is imperative that intelligent customers realize that the methodology was not meant to deceive or propagate claims of unachievable performance. ESG Lab feels that these test results are analogous to testing an air tank to levels of pressure that it would not normally be subjected to in normal operating conditions to show that it is more than capable of holding up in the most extreme conditions. It must be realized that it would have been far easier to demonstrate scalability with lesser performing devices (like spinning media, SSDs, or even PCIe flash devices) than it was to demonstrate linear scalability with NULL devices.

The performance results clearly show that ScaleIO lives up to its name and continues to scale, even when nodes are populated with extremely fast storage devices. ESG Lab has no doubt that a ScaleIO cluster will continue to scale in both capacity and performance as nodes are added to hundreds or even thousands of nodes. To help organizations better get a feel for ScaleIO technologies, EMC has launched a ScaleIO community and plans to offer a free and frictionless trial version of ScaleIO. The trial version will be made available to potential customers simply looking to evaluate the technology and will allow for unlimited capacity without time restriction for non-production use. This generous offer shows that EMC stands behind the scalability of its product and is not afraid to prove it. If your organization is struggling to deploy technologies that are flexible enough to meet unpredictable application demands while also minimizing costs, then ESG Lab strongly suggests downloading and trying ScaleIO in your data center today.[3]

[2] Source: ESG Research Report, *2015 IT Spending Intentions Survey*, February 2015.

[3] Join the ScaleIO Community at http://emc.com/scaleiocommunity and download a free trial version after May 29th from http://www.emc.com/products-solutions/trial-software-download/scaleio.htm