Lab Spotlight

Apeiron Data Systems ADS1000: Native NVMe Networking Accelerates Real-time Workload Performance

Date: April 2016 **Authors:** Kerry Dolan, Senior Lab Analyst, and Mike Leone, Senior Lab Analyst

Abstract: ESG Lab previewed Apeiron's ADS1000 Non Volatile Memory express (NVMe) storage solution (general availability: June 2016) and audited performance testing; this paper provides an overview of the architecture and impressive performance results.

The Challenges

Organizations are collecting and leveraging massive (and growing) data sets that they must process with extreme speed, often in real time, for fraud detection, web personalization, video streaming, genomic sequencing, and other high-performance computing and database activities. These jobs demand speed at any cost. Traditional scale-out architectures cannot deliver the required performance without expensive overprovisioning and wasted CPU cycles. Even the non-volatile memory solutions available today, which can deliver real-time performance and low latency, are extremely expensive silos of infrastructure. Sharing these data sets via InfiniBand or PCIe storage networks is expensive, complex, and introduces significant latency; in addition, those protocols typically cannot leverage the full capabilities of NVMe.

Apeiron ADS1000

NVMe is a new storage protocol designed to use the internal PCIe Gen3 bus to take advantage of the bandwidth, low latency, and parallel capabilities of flash technology. Apeiron has created a Direct Scale-out Flash (DSF) solution that leverages "NVMe over Ethernet" (NoE) to create a storage network with the high performance and simplicity of internal or captive storage. The system can use any commercially available NVMe drive, enabling organizations to choose the proper drive profile and supplier for its applications.

The ADS1000 delivers extreme performance in a shared infrastructure. This integrated storage and networking solution leverages the ease of use and cost-efficiencies of the Ethernet ecosystem to provide extremely high performance and the ability to independently scale compute and storage. To the client, the ADS1000 looks and acts like DAS or internal storage, but with all the operational and economic advantages of a storage network—such as the ability to access thousands of NVMe drives—without the typical network latency. The efficient NoE switching moves the bottleneck to the NAND drive architecture itself. The solution is currently demonstrating a total latency of ~100 microseconds, most of which (~95 microseconds, depending on SSD manufacturer) is in the SSD itself. Because the ADS1000 can leverage any commercial NVMe drive, the solution supports future drive innovations such as Intel's 3D XPoint technology, which has the potential to reduce latency to less than 10 microseconds.

The solution is a combination of driver-level software and hardware. The ADS1000 array is a 2U drive and network chassis capable of presenting 38 TB to 192 TBs; total capacity will increase as higher capacity NVMe SSDs come to market. The device includes dual 16-port I/O modules with redundant power and cooling. The server component includes dual 40GbE PCIe HBAs, which can be purchased separately or included in Apeiron's x86 1U compute node. A key feature of this design is a fully integrated 40Gb Ethernet switch. In addition to the performance benefits, the elimination of all external switching devices delivers significant consolidation benefits over traditional storage arrays and server-based scale-out architectures. The company's intellectual property is found in the driver and firmware built into the Intel (Altera) FPGAs, which reside in two places: in the ADS1000, and on the Apeiron HBAs. The driver and HBAs virtualize each storage command by inserting a fixed, four-byte identifier and sending it over hardened Layer-2 Ethernet to the integrated switch and FPGA component; the packet is identified and either sent directly to the appropriate drive via native NVMe, or to the dedicated server-to-server communication channel. This out-of-band server network eliminates server "chatter" from the data path. To ensure network robustness, Apeiron has integrated all of the error checking and correction capabilities of Layer 3, but without the overhead. By moving switching and storage functions (such as the physical-to-virtual mapping) to the FPGAs, and eliminating server-to-server traffic from the switching infrastructure, Apeiron has developed ultra-high performance NVMe networking at petabyte scale.

ESG Lab Testing

The Apeiron storage architecture creates a high-performance storage network that can replace traditional scale-out architectures without the need for a storage controller. Similar to scale-out solutions, the ADS1000 relies on the application or OS for functions such as RAID protection and snapshots. The system was built for workloads that demand speed first and foremost, and compute-intensive storage operations dilute the advantages of NVMe flash. However, the ADS1000 does include Tier-1 protection capabilities. It has the ability to execute and manage "hidden replication" by leveraging the FPGA infrastructure and dedicated network channels. This replication is outside of the data path and has no impact on the server CPU. Offloading storage management functions to the HBAs and I/O modules improves CPU utilization, reducing the number of servers needed for a given workload. According to Apeiron, deployment of the ADS1000 can result in 70%-150% improvement in server performance. In addition, Apeiron can achieve levels of performance as high as 72 GB/s sustained bandwidth, and 17.8M IOPS for random 4k reads in a 2U enclosure with six x86 servers and 24 NVMe drives.

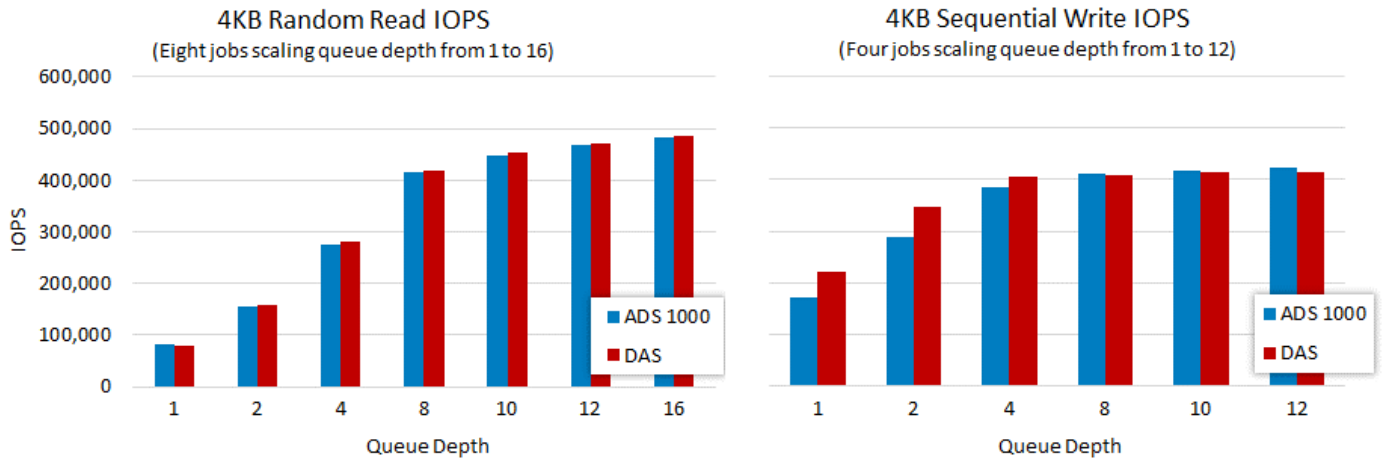
ESG Lab audited performance tests run by Apeiron with a goal of baselining the sustainable performance capabilities of the technology. Tests were run against both an Apeiron configuration and a DAS configuration for performance comparison. The same x86 server was used in both tests, and contained two 14-core E5-2695 V3 2.3GHz CPUs with 128 GB of DDR4-2133MHz RAM. An 800GB Intel P3700 NVMe SSD was used in both tests and was accessed via a Gen3 8-lane PCIe slot. For the DAS configuration, the NVMe drive was placed directly in the PCIe slot, while the Apeiron configuration leveraged an Apeiron dual 40Gb port HBA in the PCIe slot. The HBA was networked to an ADS1000 containing the NVMe drive.

Performance-sensitive applications that benefit from flash storage often require high performance for relatively small I/O requests (e.g., an OLTP database application with 4KB I/Os, or a financial application writing logs at 1KB block sizes). Others have high throughput requirements (e.g., an HPC application processing a large, machine-generated data set with 512KB I/Os). For this testing, ESG Lab focused on the small block, 4KB I/Os and measured IOPS, throughput, and response times across a series of tests that scaled both the queue depth (number of outstanding I/Os) and the number of jobs (number of simultaneous threads). The commonly used FIO utility served as the workload generation tool to produce random 4KB reads and writes.

IOPS

The first series of validated results demonstrate the pure, random read and write IOPS performance of both test scenarios (Figure 1). This represents the amount of I/O that can be processed by the underlying storage. We tested small block reads to evaluate the maximum IOPS supported by the drive, as well as small block writes to evaluate any limitations such as those that OLTP workloads might present. The first key takeaway in both the read and write IOPS performance charts is that when comparing Apeiron to DAS, the results are nearly indistinguishable as the peak performance levels are attained. The secondary takeaway is that from a pure IOPS standpoint, achieving nearly 500,000 read IOPS and 425,000 write IOPS from a single external NVMe drive validates the high level of efficiency of the Apeiron data fabric, which leverages an interconnect and an integrated network and storage appliance to service the I/O. Essentially, the ADS1000 storage network presents no distinguishable latency from an internally installed SSD.

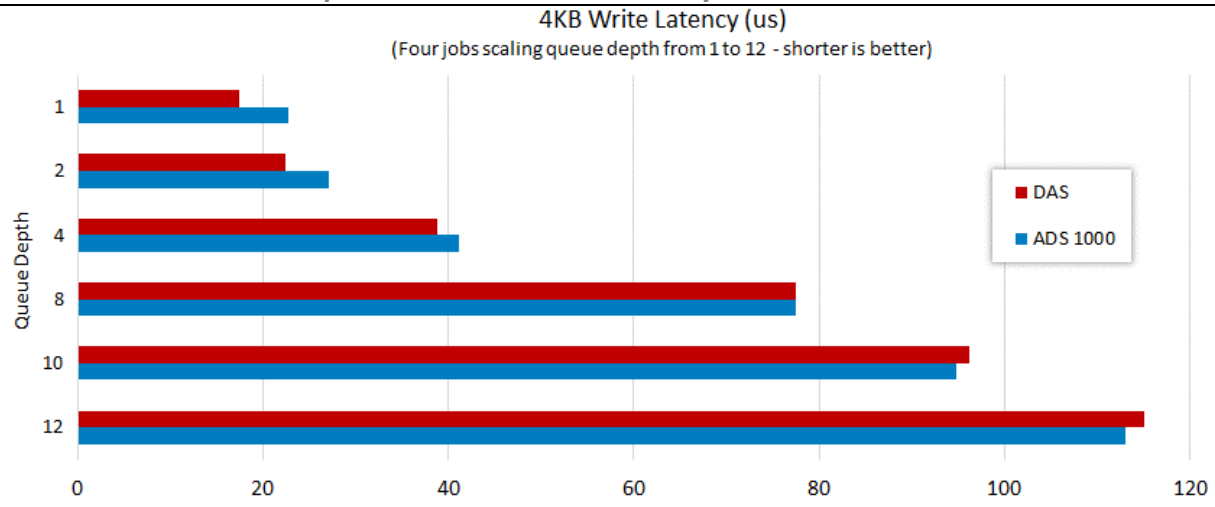
FIGURE 1. ADS1000 to DAS Comparison of NVMe SSD IOPS Performance



Latency

After comparing the IOPS performance between a single NVMe SSD installed in an ADS1000 and a direct-attached NVMe SSD, ESG Lab shifted to latency, or the time it takes to service each I/O. Latency ties directly to the end-user experience– the longer an I/O takes to complete, the longer end-users are waiting. Particularly important to the end-user experience, especially for mission-critical database workloads such as OLTP, is the amount of time it takes to complete small block write I/Os. Figure 2 displays the measured latency of 4KB writes across both test scenarios when running four simultaneous jobs and scaling the queue depth from one to twelve. Again, the difference between DAS and the ADS1000 is nearly indistinguishable, a truly impressive feat considering Apeiron requires additional I/O travel time across its fabric.

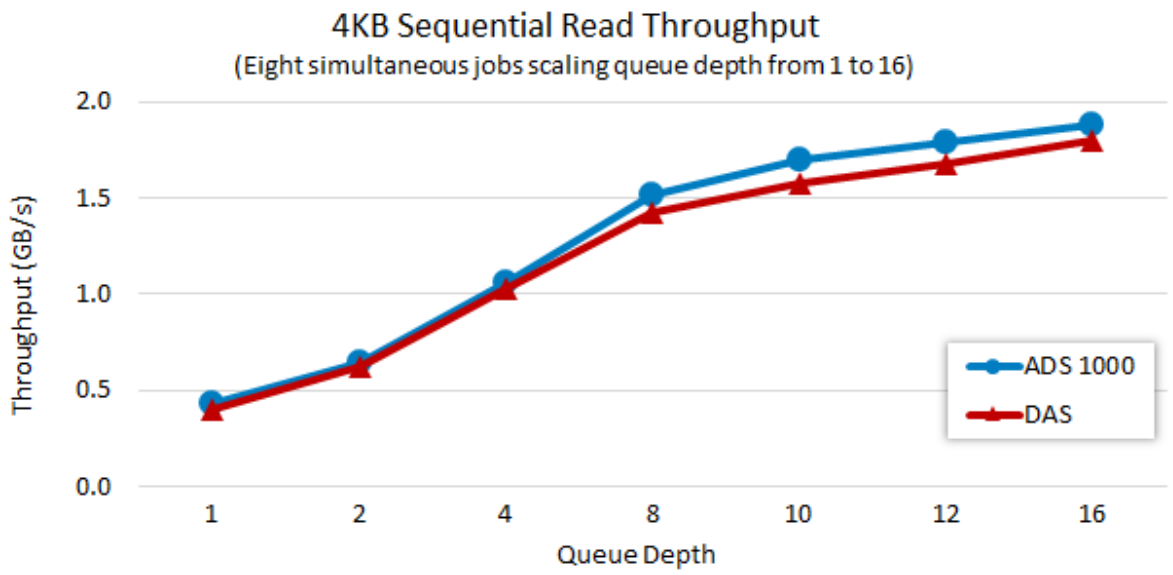
FIGURE 2. ADS1000 to DAS Comparison of NVMe SSD Latency



Throughput

The third performance metric we tested is throughput, which represents the speed at which data is transferred. Common applications associated with throughput are video streaming (large block reads) or backup (large block writes). For video streaming, many services break large video files into easily transferrable chunks of data; these become small block I/Os that minimize buffer times and get pieced back together sequentially for continuous playback. Figure 3 shows the throughput results of 4KB sequential reads. At the low queue depth scaling points, the performance of both configurations yielded similar results. As the queue depth increased to 16, ESG Lab witnessed a slight improvement of the ADS1000 configuration over internal storage. At the peak performance level of 16 outstanding I/Os, nearly 2 GB/s of throughput was achieved from a single NVMe SSD.

FIGURE 3. ADS1000 to DAS Comparison of NVMe SSD Latency



The Bigger Truth

For many organizations, essential processes need to be executed on massive data sets in real time. This is true for genomic sequencing, real-time fraud detection, online video streaming, and many other advanced analytics and HPC applications. Traditional storage solutions can't handle the amount of data or the low latency required without performing "unnatural acts" such as massive overprovisioning of storage, or splitting data into manageable subsets and then aggregating the results. Both scenarios result in large, complex environments and waste significant CPU cycles as data sets grow.

The move from fibre channel HDDs to IOPS-dense flash storage enabled a big leap in performance; NVMe drives offer the next major performance leap, with much lower latency and greater density. Apeiron brings NVMe out of the rigid silos typical of scale-out architectures and into a networked environment. The ADS1000 provides a single pool of high performance NVM that IT can allocate to an unlimited number of servers. With petabyte-scale NVMe storage, no external switching, and significant server CPU benefits, the consolidation of hardware and IT functions provides a compelling ROI/TCO justification for the system.

NVMe was designed to get the most out of flash drives by optimizing the data path to the drive. Other solutions that use Infiniband or PCIe to access NVMe drives require external switches that add to the acquisition and management costs and drag down performance. The Apeiron architecture was designed to get the most performance from NVMe, and leverages an Ethernet fabric whose standardization and consolidation capabilities drive down costs and risk.

With our preview of the solution, ESG Lab validated that the ADS1000 delivers performance equal to or better than direct-attached storage, while providing a robust storage network for NVMe. IOPS, latency, and throughput comparisons yielded indistinguishable differences between the ADS1000 with an external NVMe SSD and an NVMe SSD internal to the server. These performance capabilities offer the ability to leverage VLUNs for server and drive optimization, increasing server efficiency without the storage overhead. Apeiron enables organizations to create performance-focused NVMe networks that can truly leverage the next generation of Hadoop, NoSQL, Splunk and other big data applications.

The ADS1000 delivers multiple racks of commercial NVMe drives with an integrated data fabric, all in a single 2U building block, capable of scaling to thousands of drives and hundreds of servers. First and foremost, this enables extreme performance for real-time and HPC workloads; it also enables lower capital and operational costs through the massive consolidation of servers and external switching. This all means more performance in a smaller footprint—and as NVMe drive densities and performance increase in the future, the Apeiron advantages will increase in kind.