

PARADIGM AND DDN: ACHIEVING THE ULTIMATE EFFICIENCY FOR SEISMIC DATA ANALYSIS

THE NEED FOR I/O IMPROVEMENTS IN SEISMIC PROCESSING ENVIRONMENTS

The pressure to reduce both operating and capital costs in seismic data analysis drives an on-going demand for efficiency improvements in computational processing facilities. This is against a background of dramatically increasing seismic data volumes. Wide/Multi/Rich-azimuth methods using multi-sensor arrays and sophisticated acquisition techniques are producing higher-fidelity subsurface images, and modern analytics techniques are enabling continued advancement in the interpretation of seismic data for both newly acquired data and historical oil field data. As a result of the volume and scale of the seismic data required for modern HPC-based seismic processing and imaging, the performance of the associated storage subsystem can be a source of the greatest overall efficiency improvements. For this report, DDN performed a number of experimental benchmarks to attain optimal IO rates for Paradigm[®] Echos[®] application workloads. We present results from IO intensive Echos micro-benchmarks to illustrate the DDN GRIDScaler[®] performance benefits and provide some detail to aid optimal job packing in 40G Ethernet clusters.

DDN GRIDScaler appliances combine a storage system and a parallel file system into one appliance. The GS7K[®], GS12K[®] and GS14K[®] systems comprise a set of appliances capable of more than 10, 20 and 35 GB/s respectively. At the lower levels of software, a highly optimized real-time I/O engine, SFAOS, moves data between file system services and storage (SSD, SAS, NL-SAS). SFAOS leverages Storage Fusion Xcelerator[®] (SFX) for hybrid deployments of traditional SAS drives with SSDs. SFX[®] integrates application-centric intelligence with flash media to accelerate awkward read intensive workloads. The storage is then presented to clients using the IBM[®] Spectrum Scale[™] file system, a true parallel file system with a strong enterprise-class feature set. All this runs embedded within a fully redundant, active-active controller that supports backend data volumes up to 10PB. Furthermore, the GRIDScaler systems scale out to hundreds of GB/s throughput and hundreds of PB of capacity within a single namespace.

RESULTS

The following results summarize a set of experiments performed to identify optimal client configurations as supported by a DDN GS12K appliance. Customer workload-based Echos read-and-write benchmarks were performed on 1- and 2--socket Intel systems with 10G and 40G network connections. The load per server was increased from 1 to 18 concurrent jobs per node.

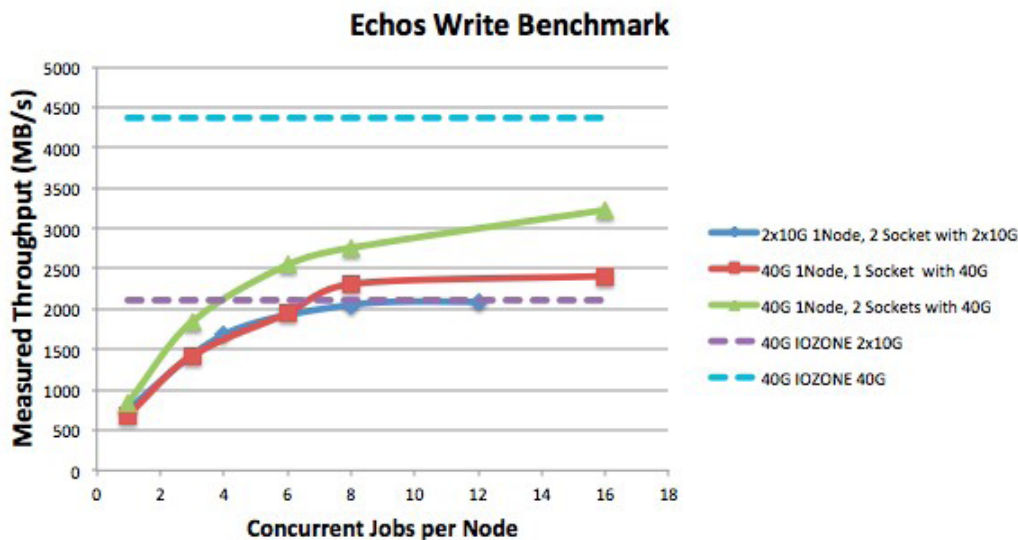


Figure 1: Paradigm Echos Write Benchmark showing the measured write performance when loading a single client with multiple copies of the IO benchmark.

“The Paradigm Seismic Processing and Imaging solutions are fully supported for use on DDN high performance storage systems.”

Alistair Downie
Product Manager, Echos, at Paradigm

 **Paradigm[®]**

“Paradigm software solutions, including Echos, provide a flexible system for implementing seismic processing and imaging workflows, which scales to the largest HPC clusters. DDN is delighted to support HPC clusters using DDN HPSS and Paradigm Echos to create highly parallel framework for high-definition geoscience workflows”

James Coomer
 EMEA Technical Director,
 DDN Storage

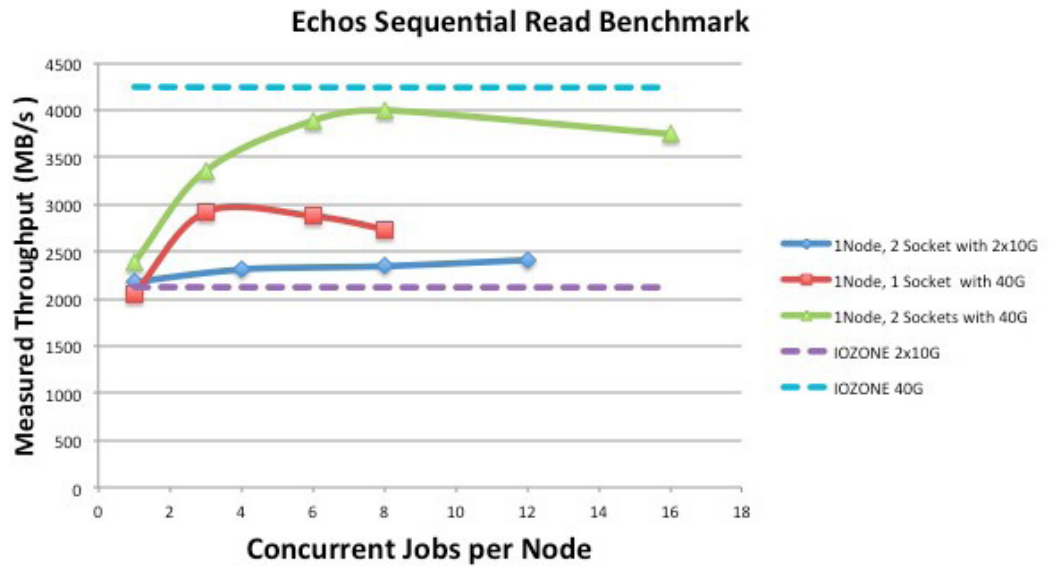


Figure 2: Paradigm Echos Read Benchmark showing the measured read performance when loading a single client with multiple copies of the IO benchmark.

The sequential read benchmark shows a similar overall pattern, with the 2-socket node clearly delivering benefits over a single socket node. The maximum read performance is achieved with just 8 concurrent jobs per node and reaches very close to 4GB/s – within 80% of the IOZONE measurements.

The above throughput tests demonstrate strong scaling to around 8 concurrent jobs before the IO and network bottleneck significantly constrains the workload. In fact for other reasons including CPU/memory bandwidth considerations, the chosen optimal value was 3. Thus with 3 concurrent jobs per node, we further investigated the number of nodes supported efficiently via a single GS12K appliance.

	Measured throughput for 45 concurrent jobs across 15 nodes
SEQUENTIAL WRITE	13,155 MB/s
SEQUENTIAL READ	18,153 MB/s

It was found that a healthy compromise involved 15 nodes each running 3 jobs per node. In this case the measured throughput on the storage system indicated that the IO benchmarks attained near the maximum throughput of the GS12K of ~20GB/s for Sequential Reads.

One more preliminary test was performed to establish the impact of SSDs: on the “Strided Read” benchmark, where the IO pattern is based on reading the seismic data back in a different order than the order it was written essentially resulting in somewhat random access into the seismic data set. A smaller test system was used and compared the performance of 100 spindles alone, with 100 spindles accelerated with 12 SSDs. The smaller test system used 4 clients, with each running the 3 instances of the Strided Read benchmark.

NO SFX	SFX ACCELERATED
1639 MB/s	5110 MB/s

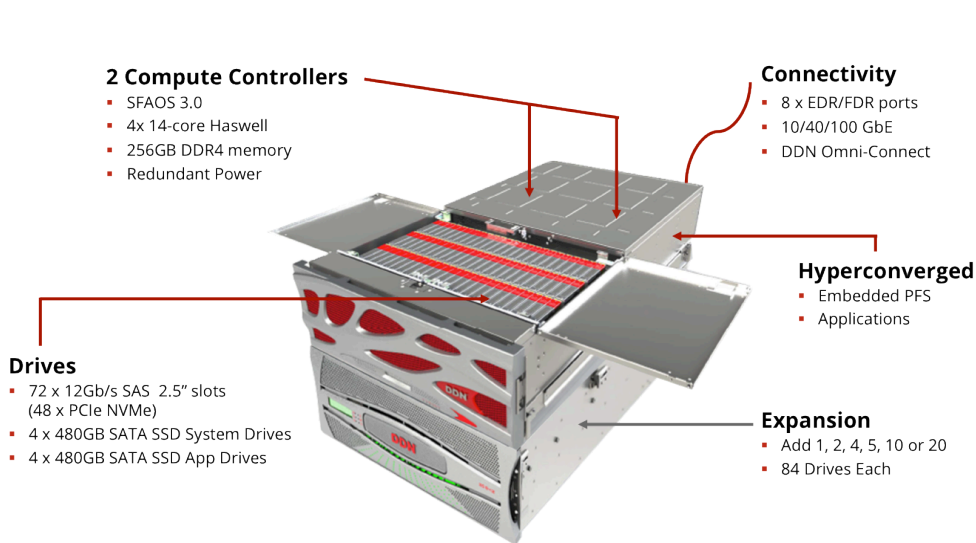
Clearly SFX-read caching on the SSDs eliminates the disk contention and corresponding performance bottleneck for this benchmark on HDDs. This result is important as SFX allows a relatively small expenditure on a handful of SSDs to have a large impact on runtimes.

SUMMARY

The DDN systems deliver very strong, scalable performance along with a small footprint and enterprise features. Each GS12K appliance delivers 20GB/s to the clients, and each single client achieves more than 4GB/s.

Newly released systems now improve upon these figures with the launch of the GS14K representing the latest SFA technology. The GS14K controller embeds the GPFS file system and SFAOS real-time IO engine into a pair of active-active controllers housed in a 4U hyperconverged unit with up to 72 SSDs. Additional storage – up to 1680 drives – of SSD SAS or NL-SAS is housed in up to 20 SS8462 Expansion Enclosures. The GS14K exceeds 35GB/s throughput and supports Omnipath, Infiniband, and Ethernet networks.

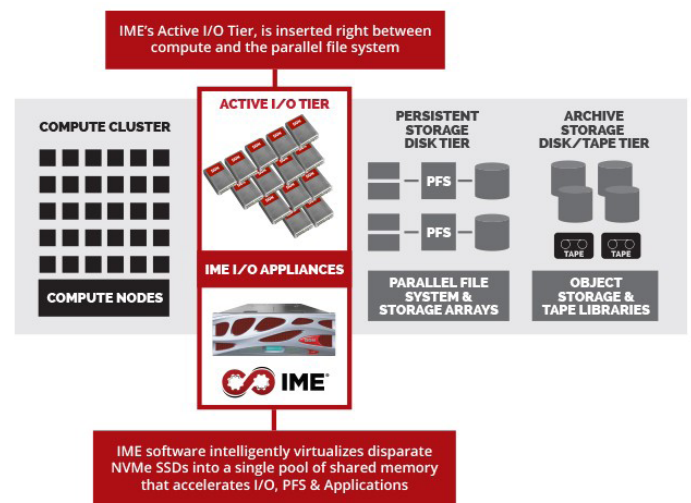
The key benefit of a true parallel file system (e.g., over scale-out NAS) is the very high single-client performance that can be delivered and sustained even when many hundreds of clients are working concurrently. A good portion of the theoretical bandwidth was attainable by Paradigm Echos IO intensive regions – particularly alongside a strong 40GE interconnect. We also have demonstrated DDN's SFX as a cost-effective way to gain application acceleration with SSDs with a minimal additional investment.



“Developing workload benchmarks that effectively simulate a customer usage of high performance storage systems requires a combination of customer focus and technical enablement. DDN appreciate the cooperation and support of Paradigm in this benchmarking effort.”

Francesco Torricelli
 Director International Business Development, DDN Storage

The next step-change in improving IO efficiencies, particularly for complex seismic workloads, is delivered by DDN's Infinite Memory Engine® (IME). IME® introduces a new tier into the IO hierarchy. Seismic analysis is particularly challenging for today's file systems due to a tendency towards large random IO and share-file IO. IME radically changes how information is saved and accessed by compute and allows data to reside next to compute in a very fast, shared pool of non-volatile memory (NVM) bypassing remaining file system bottlenecks. IME utilizes common protocols, making it transparent to both applications and the parallel file system, requiring no code modifications.



ABOUT DDN®

DataDirect Networks (DDN) is the world's leading big data storage supplier to data-intensive, global organizations. For more than 15 years, DDN has designed, developed, deployed, and optimized systems, software, and solutions that enable enterprises, service providers, universities, and government agencies to generate more value and to accelerate time to insight from their data and information, on premise and in the cloud. Organizations leverage the power of DDN technology and the deep technical expertise of its team to capture, store, process, analyze, collaborate, and distribute data, information, and content at largest scale in the most efficient, reliable, and cost effective manner. DDN customers include many of the world's leading financial services firms and banks, healthcare and life science organizations, manufacturing and energy companies, government and research facilities, and web and cloud service providers. For more information, visit our website www.ddn.com or call 1-800-837-2298.