ESG Lab Review

EMC DSSD D5: Extreme Performance in a Shared Storage Environment

Date: April 2016  Author: Kerry Dolan, Senior Lab Analyst, and Tony Palmer, Senior Lab Analyst

Abstract

This ESG Lab Review documents performance test auditing of the EMC DSSD D5 “rack-scale flash” array, which is designed to enable extremely high performance with shared storage for the most latency- and performance-dependent applications of today and the future.

Background

Big data analytics are rapidly becoming mainstream IT functions for organizations of all sizes. The ability to capture and analyze massive data sets is leading to opportunities both profound and mundane: from truly life-saving medical breakthroughs and disaster forecasting to fraud detection and improved business agility. Businesses recognize the impact this type of analysis can make, particularly as they prepare for the huge amount of data that will be generated by the Internet of Things. This increase in the importance of analytics is supported by ESG research, in which business intelligence/data analytics initiatives have risen from the eighth most-cited response on the IT priority list in 2015 to the second most-cited in 2016, after cybersecurity.¹

Figure 1. Top 10 IT Priorities for 2016

Top 10 most important IT priorities over the next 12 months. (Percent of respondents, N=633, ten responses accepted)

- Cybersecurity initiatives: 37%
- Business intelligence/data analytics initiatives: 23%
- Managing data growth: 22%
- Data integration: 21%
- Improving data backup and recovery: 20%
- Major application deployments or upgrades: 20%
- Increasing use of server virtualization: 20%
- Desktop virtualization: 20%
- Improving collaboration capabilities: 19%
- Business continuity/disaster recovery programs: 18%

Source: Enterprise Strategy Group, 2016

The Challenges

Big data analytics are designed to get answers from large data sets, but that can be difficult. Imagine a complex query of a 10TB Oracle data set—for example, taking all store, catalog, and web sales, as well as returns from this year and last year to create a sales forecast. The data set size and query complexity make this a time-consuming operation. Further, workloads like this can involve not just a single process running on a single server, but many components running on hundreds of servers. Unfortunately, the longer it takes to process these workloads, the more stale and less useful the outcome becomes. To combat this, IT has developed workarounds such as complex partitioning, complicated indexing, and materialized views, to minimize I/O and shrink the data set to a more manageable size; these often mean lower performance—as indices must be ingested—more data to store, and even staler data. Other solutions include building out the infrastructure to provide sufficient compute power or ingest bandwidth, resulting in a huge waste of storage capacity.

The rise of flash storage—aided by the decrease in its price over the past few years—has helped, but only to a point. Replacing array-based HDDs with flash can reduce I/O latency, but does nothing to reduce the latency of fibre channel and InfiniBand connectivity. Even though shared storage solutions using these technologies can support much higher throughput and lower latency than HDD-based arrays, the size and complexity of the data sets in use today present a serious challenge and even these solutions can struggle to provide enough bandwidth. Server-attached flash can boost throughput and reduce latency further, but it keeps the storage isolated, i.e., not shareable, and doesn’t scale easily.

The Solution: DSSD D5

To meet today’s requirements, a new architecture is needed, one that offers the performance benefits of direct-attached or internal storage, but that is sharable with no single point of failure like networked storage. DSSD D5 provides that new architecture, leveraging the fastest flash on the market, NVM Express (NVMe), which connects via PCIe bus. D5 delivers the kind of performance that can eliminate the need for size-reducing workarounds, and can handle high-performance computing and analysis on massive data sets. It can make data management simpler and more responsive, and keep the analysis on full sets of live data instead of subsets of stale data. According to EMC, DSSD D5 can generate 100 GB/sec, more than 10 Million IOPS, and latency as low as 100 microseconds, many times the performance of other flash solutions; this enables multi-step analytics workloads to run on the same platform instead of separately on multiple platforms.

Figure 2. DSSD D5

DSSD D5: Rack-scale Flash

- Performance-focused architecture
- Dense, shared flash
- Native and flexible data access
- Enterprise Reliability

Source: Enterprise Strategy Group, 2016

© 2016 by The Enterprise Strategy Group, Inc. All Rights Reserved.
Hardware

Each DSSD D5 comes in a 5U form factor and provides 144 TB of flash (100 TB useable, thin-provisioned) which can be accessed by up to 48 Linux hosts. Since PCIe is the fastest storage connector, it is used for both the NVMe flash modules and the server interconnects. Separate control and data planes ensure that applications can leverage all the performance of the NVMe drives. All major components of the D5 are field replaceable and redundant, providing no single point of failure.

There are several architectural innovations that make this all possible, in particular the D5’s Flash Modules, Control Modules, and I/O Modules.

- **Flash Modules.** The D5 includes 36 custom, hot-pluggable, 2TB or 4TB NVMe flash modules, providing parallel access to more than 18,000 NAND dies. Each flash module has dual 4GB/s PCIe interfaces that connect to Gen 3 x4 lane connections in the I/O modules, for a total of 8 GB/s of throughput. Built-in vaulting circuitry protects data from power failures. Each flash module is connected to dual Control Modules for enterprise redundancy and availability. The flash modules are ready to support future advancements such as capacity increases through 3D NAND and NVMe technologies.

- **Control Modules.** Dual, active-active Control Modules deliver intelligence with high availability, tracking what and where data is, but remaining separate from the data path. They manage a single, logical pool of flash (instead of individual SSDs), enabling multiple servers to share bits of data for parallel processing.

- **I/O Modules.** This PCIe mesh consists of redundant, active-active I/O modules, each containing 48 PCIe Gen 3 x4 lane ports, for a total of 96 ports. I/O flows directly between the flash modules and the applications through the I/O modules, for direct memory access (DMA). Up to two client cards per server are connected by dual, hot-pluggable PCIe Gen 3 x4 cables. NVMe MPIO is always on, enabling transparent path failover and the ability to add bandwidth or servers while the D5 is serving I/O to other hosts.

In addition, the power profile enables D5 to deliver each flash module 50 watts of power; this keeps all 18,000+ NAND dies reading and writing simultaneously, something other flash solutions don’t have enough power to do. And to keep things cool, not only are there redundant fans, but even the fans have dual rotors in case one set fails.

Software

Software innovations are also important to the D5; the change in the architecture required software to make it simpler and more efficient. An essential component is the Flood software that runs on the client (installed via the client card) and on the D5. Flood provides multiple functions: the client interface, the DMA engine, a high performance object store, data management and protection, and the appliance CLI.

The legacy I/O stack was designed for HDDs and requires multiple steps that add unpredictable latency. With the D5, Flood enables applications to issue I/O requests directly to the PCIe fabric without calls to the OS, buffer copies, volume managers, or file systems. Data moves directly between the application and the NVMe drives through the PCIe fabric.

Because it is an object store, D5 can provide high performance to many types of modern applications and data types. The block interface allows block applications (unmodified) to access virtual LUNs in the D5; the Flood API supports multiple object types including key value collections, and a plug-in supports Hadoop nodes. All access models can run simultaneously from various processes, whether within a server or among multiple servers.

Other Flood capabilities include:

- Global wear leveling and improved garbage collection, to prevent hot spots on flash, ensuring the maximum lifetime of the flash media and the best performance.
- Flash physics control, to optimize NAND dies according to age and system temperature, and extend the lifetime of the flash.
Cubic RAID, for multi-dimensional data protection. Rack-scale flash needs a different way to protect data than traditional RAID. When a standard hard disk fails, the whole thing fails, and there is a relatively small number of them in a managed unit (such as 24 HDDs in an array). With flash, each NAND die can fail on its own, taking with it the stored bits; in a D5, each Flash Module contains 512 flash cells, with more than 18,000 of them in a fully populated D5. With Cubic RAID, all NAND dies are protected at the chip layer within and across flash modules, with greater resiliency than previous RAID algorithms. The grid-like management includes row and column parity bits as well as intersection parity bits that can repair both row and column errors, enabling improved data recovery.

ESG Lab Tested
ESG Lab audited testing of the EMC DSSD D5 in EMC’s Menlo Park facility. Testing was designed to validate that the D5 provides sufficient performance to deliver faster query execution, more reliable runtimes, lower management overhead, and reduced data duplication as compared to traditional SAN-attached shared flash implementations.

The test environment included eight Dell R630 servers, each with dual 18-core Intel E5 processors and 256 GB of memory, running Oracle Real Application Clusters (Oracle RAC) 12c. Each server was connected to a single D5 using the DSSD client card, in a single card configuration, utilizing two ports per server. (The standard configuration utilizes dual cards with four ports per server.) The testing used a 5TB data set and a schema designed to emulate a decision support system for a modern retail organization. Performance was tested using a complex sales forecasting query incorporating multiple UNION and JOIN operations.

It’s important to note that testing was designed to compare the utility of using materialized views to optimize complex and lengthy queries versus simply leveraging the bandwidth of the D5, and not to validate the upper limits of D5 performance.

Figure 3 shows the explain plan for the complex query as presented by Oracle Enterprise Manager 13c. The highlighted section shows the portion of the query replaced by the materialized view. The materialized view was engineered to reduce the I/O scan volume by one third, while leaving in three JOIN and two UNION operations.

Figure 3. Oracle Enterprise Manager 13c Explain Plan
First, the full query was run against the database, with no optimization. Figure 4 shows the D5 user interface while the query was running. As seen here, the single D5 was servicing nearly 35GB/sec of throughput. The query completed in 4.5 minutes.

**Figure 4. EMC DSSD D5 User Interface – Query Running**

Next, the materialized view-optimized query was run. Optimization resulted in reduction of the query runtime to 3.4 minutes, or about 24%. It’s important to note that the reason for this is that the non-I/O-intensive portions of the query—specifically the BUFFER SORT phase—become dominant when the I/O requirement is reduced. This is best illustrated by comparing the **Metrics** tab statistics in Oracle Enterprise Manager. As seen in Figure 5, the same spike occurs in the **CPU Used** charts for both the full and optimized queries, while the **I/O Throughput** charts show a valley in both queries at the same spot.

**Figure 5. CPU utilization and I/O Bandwidth During Queries**
The implication here is that while materialized views might reduce I/O bandwidth requirements substantially, significant CPU requirements may remain, depending on how much of the query can be materialized. With the D5 there is sufficient raw bandwidth so that the I/O against the raw tables becomes less significant than the residual processing. In addition, on all-flash arrays, CPU is required to serve data. On DSSD D5, hosts serve themselves with data via Direct Memory Access (DMA), allowing them to leverage full parallel access to all 18,000 NAND chips in the appliance, dramatically reducing contention between sessions.

To see what effect this could have in the real world, ESG Lab compared the performance of the D5 with that of a typical all-flash array, based on a single array configuration with a “datasheet” throughput specification of 3 GB/sec, which is a bit higher than the average ESG Lab has observed for single all-flash arrays running database workloads. The query runtimes for the all-flash array were modeled by taking the total amount of data scanned during both the full query and materialized view, calculating how long it would take the all-flash array to scan the data, and then adding that time to the non-I/O portion of the query.

**Figure 6. D5 versus All-Flash Array—Full Query and Materialized View**

As Figure 6 shows, while the materialized view delivers a reduction of 64% in query time for the standard all-flash array by reducing I/O, there is a much smaller optimization effect for the D5 (24%). When the significant amounts of complex development and operations time to create and maintain the views—along with data staleness and extra capacity requirement issues inherent in materialized views—are taken into consideration, the value and usefulness of materialized views drops further.

It’s important to note that this lab test simulates just one business analyst running one query. The impact is intensified when multiple business analysts are running multiple queries simultaneously. Even if they are using materialized views on all-flash arrays, the organization will get its answers faster using DSSD.
Figure 7 compares the D5 running the full query to the all-flash array running the optimized query using the materialized view. With no optimizations the D5 reduced query time by more than 55%.

**Figure 7. D5 Full Query versus All-Flash Array Materialized View**

The data used to model the performance comparisons is detailed in Table 1.

**Table 1. Complex Query Runtime Results**

<table>
<thead>
<tr>
<th>Platform</th>
<th>Bandwidth</th>
<th>Query</th>
<th>Data Scanned (GB)</th>
<th>Time to Scan (Seconds)</th>
<th>Non-I/O Portion of Query (Seconds)</th>
<th>Total Time (Seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMC DSSD D5</td>
<td>32 GB/sec (Tested)</td>
<td>Full</td>
<td>4,710.4</td>
<td>147.2</td>
<td>122.8</td>
<td>270.0</td>
</tr>
<tr>
<td>All Flash Array</td>
<td>3 GB/sec (Reported)</td>
<td>Full</td>
<td>4,710.4</td>
<td>1,570.1 (Modeled)</td>
<td>122.8</td>
<td>1,692.9 (Modeled)</td>
</tr>
<tr>
<td>EMC DSSD D5</td>
<td>32 GB/sec (Tested)</td>
<td>Materialized View Optimized</td>
<td>1,331.2</td>
<td>41.6</td>
<td>162.4</td>
<td>204.0</td>
</tr>
<tr>
<td>All Flash Array</td>
<td>3 GB/sec (Reported)</td>
<td>Materialized View Optimized</td>
<td>1,331.2</td>
<td>443.7 (Modeled)</td>
<td>162.4</td>
<td>606.1 (Modeled)</td>
</tr>
</tbody>
</table>

Finally, ESG Lab examined a 100% read workload generated using the SLOB2 utility. As seen in Figure 8, the D5 under test was able to drive 43.7 GB/Sec of throughput, and 5.35 million 8KB Oracle IOPS at 300 microsecond average response time, confirming that the DSSD D5 has the headroom to support additional workloads on top of the complex query detailed above.
To validate the real world application of the concepts explored in this report, ESG Lab spoke with one of EMC's DSSD D5 customers. CMA Consulting builds software product suites focused on database optimization, financial management, and human resources, and provides technology and management consulting services for public sector and commercial clients. CMA has observed more than 54 GB/sec of throughput from a single run of their real-life query workloads. According to Brian Dougherty, CMA’s Chief Technical Architect, “The CMA workload hammers the storage system and the D5 has brought Oracle to life again.” This example confirms that a single DSSD D5 can handle the most complex Oracle queries with additional bandwidth for additional concurrent queries.

**Why This Matters**

Big data analytics and high performance database processing are not esoteric requirements for a niche market any more. While many of its applications are exotic—genomics, high performance computing, even the search for extraterrestrial life—big data analytics are needed for tasks such as keeping up with customers’ online transactions, handling seasonal workload increases, and simply enabling organizations to outmaneuver the competition, which is only a click away.

ESG Lab validated the extreme performance of the DSSD D5 running complex queries against a 5TB data set in an Oracle RAC environment and compared those results to what can be accomplished using a standard all-flash array. ESG Lab found that the D5 could complete a complex, long-running query against the full data set with no optimizations in less than half the time it would take with a typical all-flash array.

DSSD D5 eliminates the need for complex workarounds to reduce data set sizes so that analysis can be done in a timely fashion with fresher data. DBAs can spend less time trying to optimize queries and tune the system, and more time on productive business efforts. Companies get faster time to information—the actual answers to their questions—with a vastly simpler application environment and a storage solution that packs a huge punch in a small footprint.
The Bigger Truth

Legacy infrastructures were not built for today’s applications, data sets, or even storage media, and they are simply not able to keep up with the performance requirements. Today, organizations are using in-memory computing and applications like NoSQL, Hadoop, and Splunk to query petabytes of data that are stored on SSDs. Adding flash to traditional arrays or servers delivered some incremental improvements—but only with a redesign of both hardware and software can you make full use of the performance capabilities of today’s NVMe flash.

While “application latency” may sound like a boring IT metric, in fact latency is often the key to the kingdom. Speeding up processing of large data sets can mean saving billions of dollars, by reducing fraud detection processing from 60 ms to 1 ms. It can mean saving lives by reducing the time to sequence a genome from years down to minutes. The processes that can benefit are too numerous to count—for example, the average airline flight generates 500GB of data, so major airports that handle upwards of 1500 flights per day, each with a two-hour maintenance window, need to process massive amounts of data incredibly fast to maintain flight safety and keep things moving.

DSSD D5 offers the benefits of both server-side latency and shared storage. It is designed for workloads that require extremely fast performance and leverage large data sets. It delivers the performance of 18,000+ NAND dies, all working in parallel, running over the PCIe fabric to PCIe connected clients, with advanced flash data protection. This enables organizations to use the largest data sets and to do more analysis with less tuning. It means organizations can get to the answers they are seeking much faster, with fresher data.

ESG Lab validated that the DSSD D5 can leverage its massive bandwidth to run complex queries against large data sets without labor-intensive optimizations or workarounds to reduce the amount of data to scan. D5 demonstrated that it can perform non-optimized analysis of a 5TB data set in less than half the time required by a standard all-flash array using materialized views.

ESG Lab was extremely impressed with the DSSD D5. The hardware and software architectures are built specifically for NVMe flash, which acts differently from HDDs; as a result, the D5 enables levels of performance and data protection that other all-flash or hybrid solutions simply cannot match. DSSD D5 provides greater problem solving capabilities than have been possible by orders or magnitude—as a result, customers may have difficulty at first realizing what they can do with a solution like this. But won’t it be exciting to watch as they figure it out!