Technical Report

Oracle Databases on NetApp EF-Series

Mitch Blackburn, NetApp
October 2019 | TR-4794

Abstract

This guide is intended to help storage administrators and database administrators successfully deploy Oracle on NetApp® EF-Series storage.

NetApp®

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Introduction

In most OLTP systems, the processor, memory, and I/O subsystem in a server are well balanced and are not considered performance bottlenecks. The major source of performance issues in OLTP environments is typically related to storage I/O activity. The speed of existing HDD-based storage systems does not match the processing capabilities of the servers.

As a result, a powerful processor often sits idle, waiting for the storage I/O requests to complete. This situation negatively affects user and business productivity. The effect on productivity delays the return on investment (ROI) and increases overall TCO. Therefore, storage IOPS performance and latency become strategic considerations for business. It is critical to make sure that the response time goals are met, and performance optimization is realized for other system resources (processor and memory).

The NetApp EF-Series flash array is designed for performance-driven applications with microsecond-level latency requirements. The array is built on storage architecture developed by experts with more than 25 years of storage development experience; indeed, there are more than 1,000,000 systems in the field. Each EF-Series flash array can deliver extreme performance with microsecond-level response times, enabling business-critical applications to deliver faster results and improve the end-user experience. This combination of high IOPS and ultralow latency makes the EF-Series flash array an ideal choice for database-driven applications that require extreme performance.

The NetApp EF-Series flash array leads the market in delivering high performance and low latency. The EF570 is the $/SPC-2 MBps leader.[1] According to StorageReview.com, “To say the EF570 is quick is putting it mildly; it’s a monster.”[2]

NVMe has become the industry standard interface for Peripheral Component Interconnect Express (PCIe) SSDs. With a streamlined protocol command set and fewer clock cycles per I/O, NVMe supports up to 64K queues and up to 64K commands per queue. These attributes make it more efficient than SCSI-based protocols like SAS and SATA.

The introduction of NVMe over Fabrics (NVMe-oF) makes NVMe more scalable without affecting the low latency and small overhead that are characteristic of the interface. NetApp EF-Series EF570 and EF600 systems support NVMe over RoCE (NVMe/RoCE), NVMe over InfiniBand (NVMe/IB), and NVMe over Fibre Channel (NVMe/FC).

Both the EF570 and the EF600 support the SCSI-based fiber channel protocol (FCP) as well as NVMe/FC. Because little change is required in the standards to implement NVMe/FC, the introduction of NVMe/FC along with existing storage is easy, seamless, and noninvasive. And because NVMe/FC can use the same infrastructure components concurrently with other FC traffic, it is easy to migrate workloads at the pace that works for your organization. NVMe/FC also allows the efficient transfer of NVMe commands and structures end to end with no translations.

The leading-edge EF600, which is built on end-to-end NVMe technology, not only provides NVMe-oF to reduce latency between the server and the SAN. It also incorporates NVMe technology to accelerate access to your data.

The EF-Series array, available with up to 1.8PB of raw SSD capacity in the EF570 and up to 367TB in the EF600, provides the capacity and bullet-proof reliability to meet the requirements of the most demanding organizations. This technical report provides an overview of best practices for Oracle with NetApp EF-Series flash arrays.

---

[1] Storage Performance Council, Top 10 SPC-2 Version 1 by Price-Performance.
[2] StorageReview.com, NetApp EF570 All-Flash Array Review, by Adam Armstrong, October 2, 2018.

1.1 Intended Audience

This technical report is intended for NetApp customers, partners, employees, and field personnel who are responsible for deploying an Oracle database solution in a customer environment. It is assumed that the reader is familiar with the various components of the solution.

1.2 Caveats

This document covers the storage layout for an Oracle Automatic Storage Management (Oracle ASM) database that uses a NetApp EF-Series all-flash array as the underlying storage system. This document assumes that the database is either being relocated to an EF-Series storage system or being created on an EF-Series storage system to achieve high performance. This document also assumes that you want to improve performance of an OLTP application.

2 Introduction to EF-Series Storage Systems

2.1 EF-Series Hardware Overview

The EF-Series flash array continues NetApp's long-standing heritage of delivering powerful solutions to meet specific business needs. With high IOPS and sub millisecond response times, NetApp EF-Series arrays enable business-critical applications to deliver faster results and improve customer experience.

This combination of high IOPS and ultra-low latency makes an EF-Series flash array a great choice for database-driven applications that require a dedicated extreme performance solution.

Table 1 provides an overview of EF-Series flash array performance for RAID6 configurations plus Dynamic Disk Pool (DDP) configurations for the EF600.

Table 1) EF-Series performance, RAID6 and DDP as noted.

Product Comparison	EF280	EF570	EF600
SSD count	24	48	24
Max read IOPS	300K<210µs	1M<250µs	2M<250µs 2M<260µs DDP
Max write IOPS	45K<160µs	185K<250µs	340K<190µs 339K<190µs DDP
Read bandwidth	10GB/sec	21GB/sec	44GB/sec 43GB/sec DDP
Write bandwidth (cache mirroring enabled)	3.7GB/sec	9GB/sec	12.5GB/sec R6/DDP
Low latency (reads)	55K<140µs	100K<120µs	140K<110µs R6/DDP
Low latency (writes)	45K<160µs	150K<100µs	200K<80µs R6/DDP

Table 2 provides a comparison of the EF-Series product line.

Table 2) EF-Series product comparison.

Product Comparison	EF600	EF570	EF280
2U/24 expansion shelves	none	4	3

Product Comparison	EF600	EF570	EF280
Total drives	24	120	96
Raw capacity	367TB	1.8PB	1.45PB
Controller cache options	32GB or 128GB	16GB or 64GB	8GB or 32GB

Along with performance, the key to maximizing value is to maximize efficiency. Historically, companies have sacrificed efficiency to achieve extreme performance levels by overprovisioning their storage. But that is changing. The all-flash EF-Series storage array helps customers balance performance and efficiency by eliminating overprovisioning, resulting in dramatically reduced costs.

With the performance of more than 1,000 traditional drives, a single EF-Series flash array can meet extreme requirements with 95% less rack space, power, and cooling. This ability is a significant benefit to customers who are used to deploying partially filled disks to improve application performance.

In addition to cost efficiency, the EF-Series flash array provides application efficiency. By completing a higher volume of application operations, customers can become more efficient and obtain better results. The EF-Series flash array has a fully redundant I/O path that provides automated failover. Surprisingly, automated failover is not provided in many of the flash products available today, but it is an absolute requirement for enterprises that want to implement this type of technology.

All management tasks are performed while the EF-Series array remains online with complete read/write data access. This approach allows storage administrators to make configuration changes and conduct maintenance without disrupting application I/O.

The EF-Series flash array also offers advanced data protection common to enterprise storage to protect against data loss and downtime events. This protection is available locally with NetApp Snapshot™ technology and remotely with synchronous and asynchronous replication.

2.2 SANtricity OS

NetApp EF-Series flash array systems are managed by the SANtricity System Manager browser-based application. The SANtricity System Manager is embedded on the controller.

To create volume groups on the array, the first step is to assign a protection level during configuration. This assignment is then applied to the disks selected to form the volume group. EF-Series flash array storage systems support Dynamic Disk Pools (DDP) and RAID levels 0, 1, 5, 6, and 10. DDP was used for all configurations described in this document.

To simplify storage provisioning, SANtricity System Manager provides an automatic configuration feature. The configuration wizard analyzes the available disk capacity on the array. It then selects disks that maximize array performance and fault tolerance while meeting capacity requirements, hot spares, and any other criteria specified in the wizard.

For further information about NetApp SANtricity Unified Manager and SANtricity System Manager, see the E-Series Documentation Center.

Dynamic Disk Pools

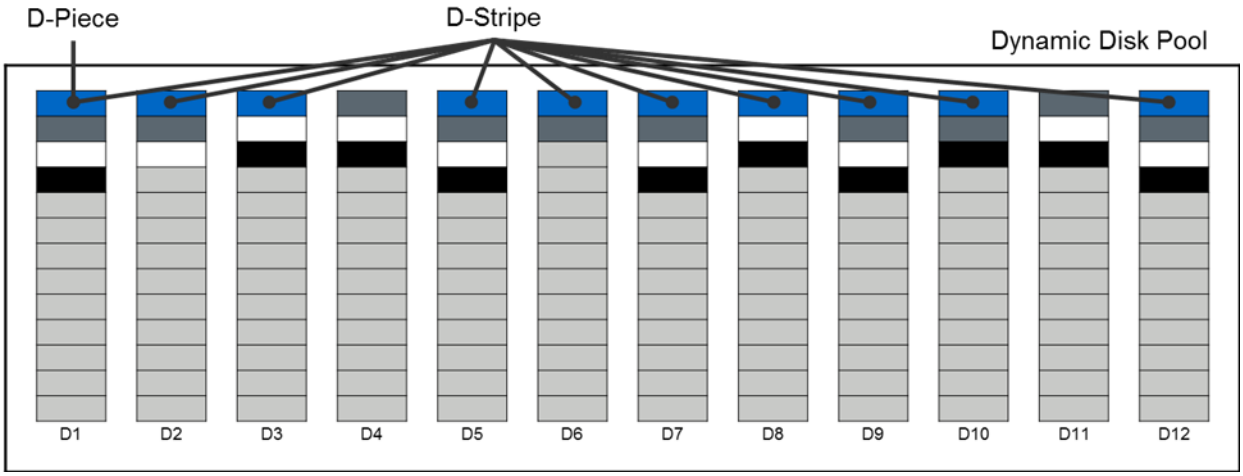
With seven patents pending, the DDP feature dynamically distributes data, spare capacity, and protection information across a pool of drives. These pools can range from a minimum of 11 drives to all the supported drives in a system. In addition to creating a single pool, storage administrators can mix traditional volume groups and DDP or even multiple pools, offering an unprecedented level of flexibility.

A pool is composed of several lower-level elements. The first of these elements is a D-piece. A D-piece consists of a contiguous 512MB section from a physical disk that contains 4,096 128KB segments. Within a pool, the system chooses 10 D-pieces by using an intelligent optimization algorithm from selected

drives in the pool. Together, the 10 associated D-pieces are considered a D-stripe, which is 4GB of usable capacity. Within the D-stripe, the contents are similar to a RAID 6 8+2 scenario. There, eight of the underlying segments potentially contain user data, one segment contains parity (P) information calculated from the user data segments, and one segment contains the Q value as defined by RAID 6.

Volumes are then created from an aggregation of multiple 4GB D-stripes as required to satisfy the defined volume size up to the maximum allowable volume size in a pool. Figure 1 shows the relationship between these data structures.

Figure 1) DDP components.

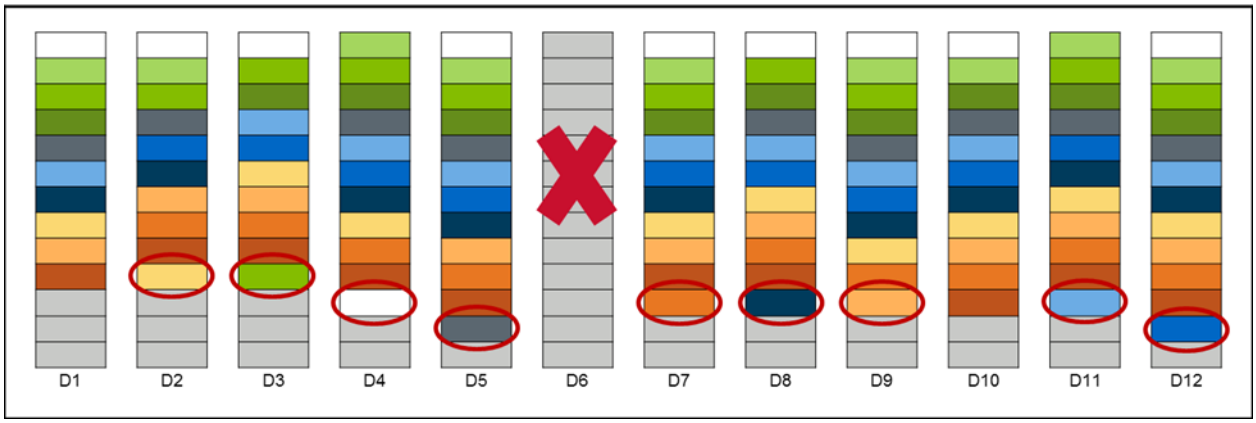


Another major benefit of a pool is that, rather than using dedicated stranded hot spares, the pool contains integrated preservation capacity to provide rebuild locations for potential drive failures. This approach simplifies management, because individual hot spares no longer need to be planned or managed. The approach also greatly improves the time for rebuilds, if necessary, and enhances volume performance during a rebuild, as opposed to traditional hot spares.

When a drive in a pool fails, the D-pieces from the failed drive are reconstructed to potentially all other drives in the pool by using the same mechanism normally used by RAID 6. During this process, an algorithm internal to the controller framework verifies that no single drive contains two D-pieces from the same D-stripe. The individual D-pieces are reconstructed at the lowest available Logical Block Access (LBA) range on the selected drive.

In Figure 2, drive 6 (D6) is shown to have failed. Next, the D-pieces that previously resided on that disk are recreated simultaneously across several other drives in the pool. Because there are multiple disks participating in the effort, the overall performance effect of this situation is lessened, and the length of time needed to complete the operation is dramatically reduced.

Figure 2) DDP drive failure.



When multiple disk failures occur in a pool, priority for reconstruction is given to any D-stripes missing two D-pieces to minimize data availability risk. After those critically affected D-stripes are reconstructed, the remainder of the necessary data is reconstructed.

From a controller resource allocation perspective, there are two user-modifiable reconstruction priorities in DDP:

- Degraded reconstruction priority is assigned to instances in which only a single D-piece must be rebuilt for the affected D-stripes; the default for this value is high.
- Critical reconstruction priority is assigned to instances in which a D-stripe has two missing D-pieces that need to be rebuilt; the default for this value is highest.

For large disk pools with two simultaneous disk failures, only a relatively small number of D-stripes are likely to encounter the critical situation in which two D-pieces must be reconstructed. As discussed previously, these critical D-pieces are identified and reconstructed initially at the highest priority. Doing so returns the pool to a degraded state quickly so that further drive failures can be tolerated.

In addition to improving rebuild times and providing superior data protection, DDP can also greatly improve the performance of the base volume under a failure condition compared with the performance of traditional volume groups.

For more information about DDP, see TR-4652: SANtricity Dynamic Disk Pools—Feature Description and Best Practices.

2.3 Provisioning the EF-Series Flash Array

SANtricity DDP technology allows storage administrators to simplify RAID management, improve data protection, and maintain predictable performance under all conditions. DDP evenly distributes data, protection information, and spare capacity across the entire EF-Series pool of drives, simplifying setup and maximizing use. Its next-generation technology minimizes the performance effect of a drive failure and can return the system to optimal condition up to eight times more quickly than traditional RAID. With shorter rebuild times and patented technology to prioritize reconstruction, DDP significantly reduces exposure to multiple disk failures, offering a level of data protection that simply can't be achieved with traditional RAID.

With SANtricity software, all management tasks can be performed while the storage remains online with complete read/write data access. Storage administrators can make configuration changes, conduct maintenance, or expand the storage capacity without disrupting I/O to attached hosts. SANtricity software's online capabilities include the following:

- Dynamic volume expansion allows administrators to expand the capacity of an existing volume.

- Dynamic segment size migration enables administrators to change the segment size of a given volume.
- Dynamic RAID-level migration changes the RAID level of a RAID group on the existing drives without requiring the relocation of data. Supported RAID levels are 0, 1, 5, 6, and 10.
- Nondisruptive controller firmware upgrades are supported with no interruption to data access.

For the detailed information on provisioning EF-Series flash array, see the E-Series Documentation Center. If there is repeated volume creation and deployment, REST API commands can be used to automate this task.

Best Practice
NetApp recommends using Dynamic Disk Pools for all storage configuration.

3 Oracle Automatic Storage Management

Oracle Automatic Storage Management (Oracle ASM) provides integrated cluster file system and volume-management features, removing the need for third-party volume management tools and reducing the complexity of the overall architecture.

The key Oracle ASM features include the following:

- Automatic file and volume management
- Database file system with performance of raw I/O
- Automatic distribution and striping of data
- A choice of external (array-based) data protection and two-way and three-way mirror protection
- Control over which copy of mirrored data should be used preferentially

With these capabilities, Oracle ASM provides an alternative to the third-party file system and volume-management solutions for database storage management tasks such as creating or laying out databases and managing the use of disk space. Volume-management tasks on the Oracle server host can be performed by using familiar create, alter, and drop SQL statements, simplifying the job of database administrators regarding database storage provisioning. Load balancing avoids performance bottlenecks by enabling the I/O workload to use all the available disk-drive resources.

NetApp EF-Series storage systems automatically load balance I/O among all the solid-state drives (SSD) for the underlying volume group. All LUNs placed within the single volume group can use all the volume group's SSD in a balanced manner. Oracle ASM provides further load balancing of I/O across all LUNs or files in an Oracle ASM disk group by distributing the contents of each data file evenly across the entire pool of storage in the disk group based on a 64MB stripe size. This provides even performance through the available SCSI devices at the host and network layer.

When used with Oracle ASM, NetApp load balancing allows multiple LUNs and file system data to share common disk drives. This functionality reduces the number of LUNs required for each Oracle ASM disk group, which improves manageability without compromising performance. This provides optimal read and write performance in high-transaction database environments.

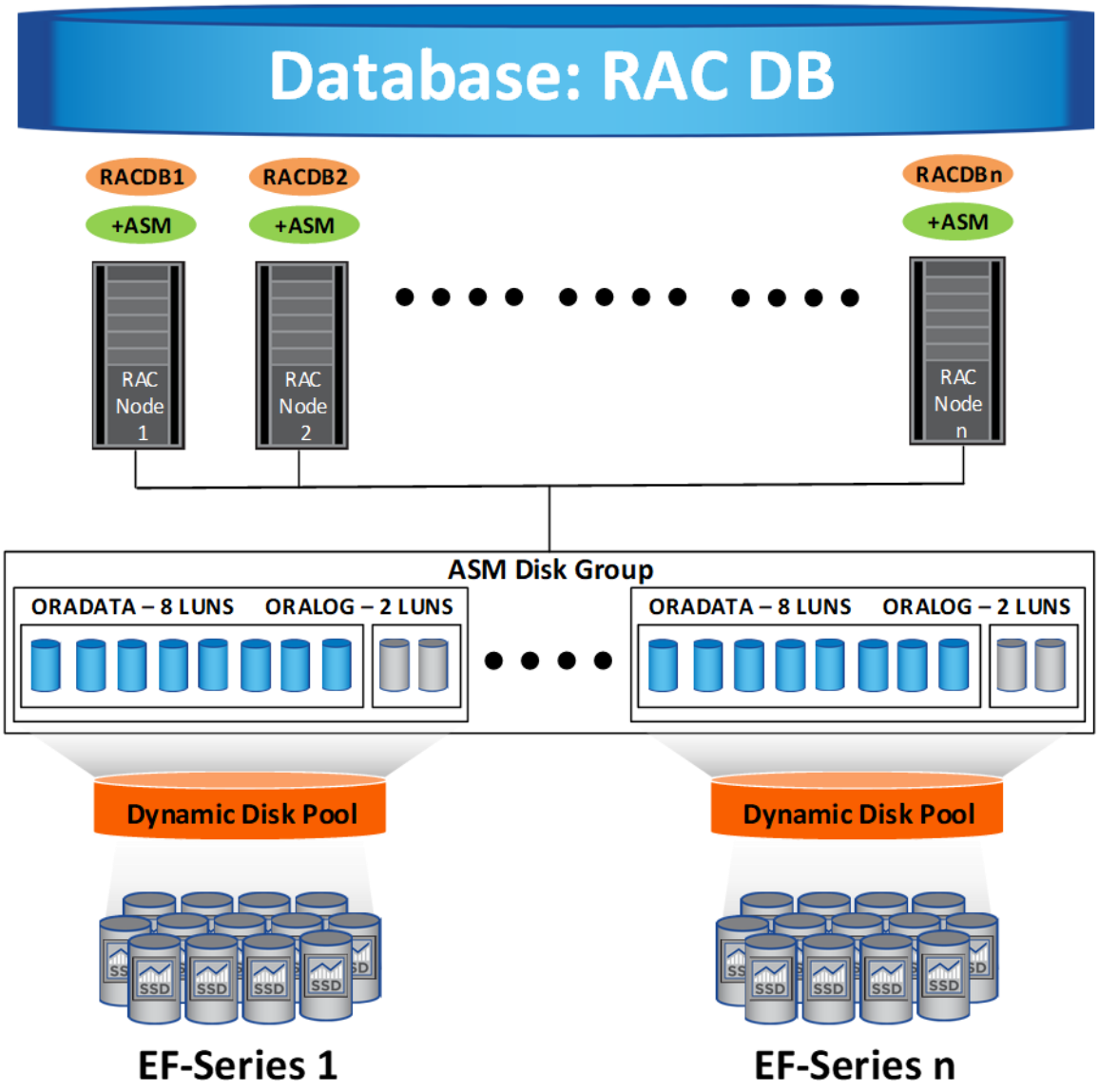
Best Practice

When you create Oracle ASM disk groups on NetApp EF-Series storage, NetApp recommends configuring at least eight LUNs per DDP for the data disk group. Configuring multiple LUNs can maximize I/O concurrency to the disk group, overcoming any per-LUN limits of the host driver stack (operating system, host bus adapter driver, or multipath driver). Even if initially unnecessary, this provisioning can avoid unneeded data movement that would occur if LUNs within the same volume were later added to the disk group.

Figure 3 illustrates how an Oracle RAC database can be configured with multiple EF-Series all-flash arrays for a multi-node RAC database. A single DDP is created using all available drives on each EF-Series, and the data and log volumes are created from it. The ORADATA and ORALOG LUNs are provisioned on the EF-Series arrays by using SANtricity System Manager and then presented to the Oracle RAC nodes. A single ASM disk group can be created to span all the LUNs to present a single file system from a database and host perspective.

The result of this configuration is that the workload is evenly distributed across all the storage LUNs and volumes on all arrays. External redundancy is chosen for the ASM disk group because the NetApp EF-Series already provides protection at the array level. Optionally, customers can choose ASM mirroring to increase data protection, provided the storage arrays are provisioned with enough capacity.

Figure 3) Oracle ASM disk group layout per database.



4 High Availability

4.1 EF-Series Systems and SANtricity OS

The NetApp EF-Series storage system has been architected for high reliability and high availability with features such as:

- A dual active controller with automated I/O path failover
- RAID levels 0, 1, 5, 6, and 10 or DDP
- Redundant, hot-swappable controllers, disks, power supplies, and fans
- Automatic drive failover detection and rebuild using global hot spares
- Mirrored data cache with battery backup and destage to memory
- Nondisruptive controller firmware upgrades

- Proactive drive health monitoring
- Background media scan with autoparity check and correction

All components are fully redundant, and you can swap them without powering off the system or even halting operation. Redundant components include controllers, disks, power supplies, and fans. The EF-Series power supplies offer an 80-plus efficiency rating. The EF-Series flash array features several functions designed to protect data in every circumstance. Multiple RAID levels are available for use with varying levels of redundancy. If a connection is lost, failover from one path to another is also automatically included with the system. Within the shelf, each drive has a connection to each controller so that even internal connection issues can be quickly overcome. Volumes on the system are available for host I/O from the moment they are created and can even have significant properties altered without stopping I/O.

Other features of the EF-Series flash array that protect data include mirroring and backing up the controller cache. If power is lost while the system is operating, onboard batteries destage the data from cache memory to internal controller flash so that it is available when power is restored. The RAID algorithms allow the system to recreate any lost data in case of drive failure. You can also confirm data with RAID parity and even continue a rebuild if you hit an unreadable sector.

Behind the scenes, the system performs other tasks that protect data at all times. The optional media scan feature looks for inconsistencies, even on sectors not currently accessed by any host. The EF-Series array proactively tracks SSD wear and flags drives that are approaching the end of their expected life. All types of diagnostic data are routinely collected for use later by NetApp Support, if necessary.

As already described, the EF-Series array offers many reliability and availability features. In addition, NetApp SANtricity software makes it possible to maximize availability. For example, SANtricity performs the following tasks:

- Enables high-speed, high-efficiency NetApp Snapshot technology
- Protects data in seconds
- Reduces flash consumption by storing only changed blocks
- Provides robust disaster recovery protection
- Supports synchronous mirroring for no-data-loss protection of content
- Supports asynchronous mirroring for long-distance protection and compliance
- Maximizes ROI with flexible protection
- Supports flash, near-line SAS (NL-SAS), or a mix of recovery targets based on cost and performance needs
- Delivers speed without breaking budgets

4.2 Oracle HA Options

Oracle provides many high availability options for different database configurations. Descriptions of the various options can be here.

5 Performance Optimization and Benchmarking

Accurate testing of database storage performance is a complicated subject. It requires an understanding of IOPS and throughput. This subject also requires an understanding of the difference between foreground and background I/O operations, the impact of latency upon the database, and numerous OS and network settings that also affect storage performance. In addition, there are non-storage database tasks to consider. There is a point where optimizing storage performance yields no useful benefits because storage performance is no longer a limiting factor for performance.

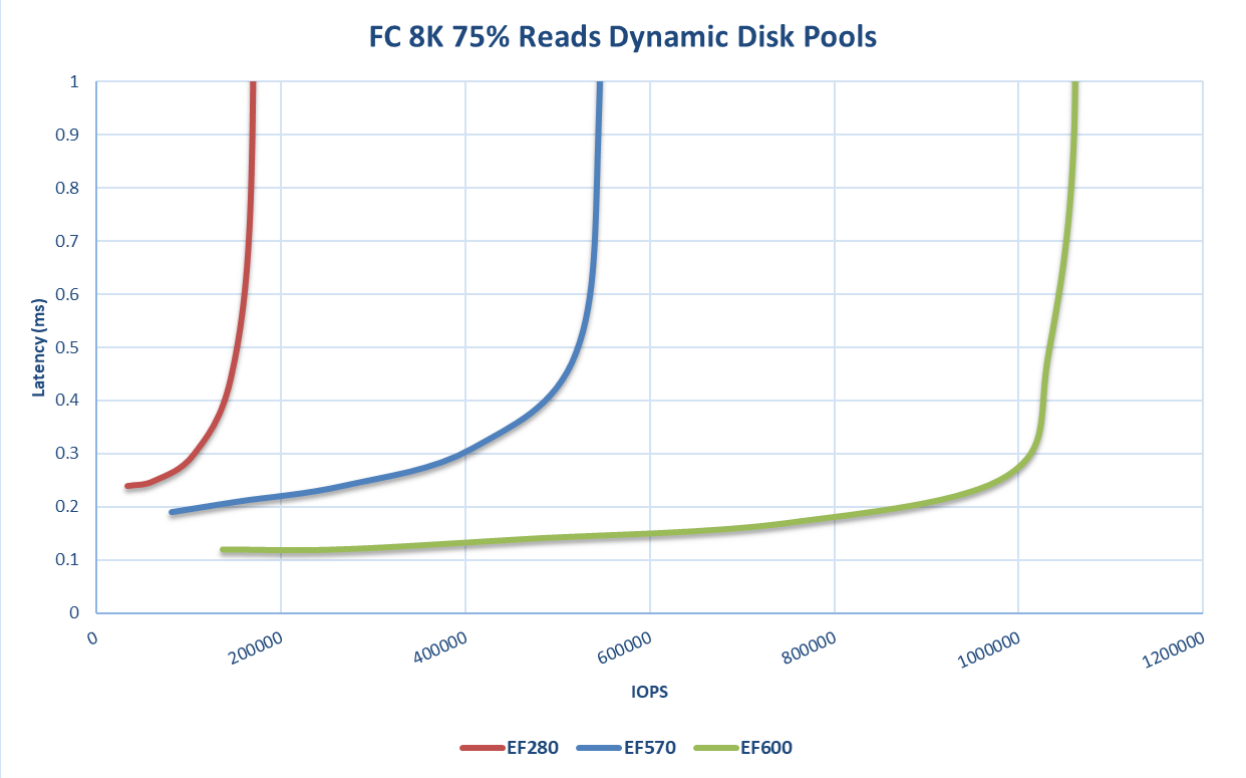
There are some additional considerations when choosing an all-flash array like the EF-Series.

- With a 75/25 read/write ratio, an EF600 can deliver approximately 800K random database IOPS at a latency of 300µs. This is so far beyond the current performance demands of most databases that it is difficult to predict the expected improvement. Storage would be largely erased as a bottleneck.
- Network bandwidth is an increasingly common source of performance limitations. For example, spinning disk solutions are often bottlenecks for database performance because the I/O latency is very high. When latency limitations are removed by an all-flash array, the barrier frequently shifts to the network. This is especially notable with virtualized environments and blade systems where the true network connectivity is difficult to visualize. This can complicate performance testing if the storage system itself cannot be fully utilized due to bandwidth limitations.
- Comparing the performance of an all-flash array with an array containing spinning disks is generally not possible because of the dramatically improved latency of all-flash arrays. Test results are typically not meaningful.
- Comparing peak IOPS performance with an all-flash array is frequently not a useful test because databases are not limited by storage I/O. For example, assume one array can sustain 500K random IOPS, whereas another can sustain 300K. The difference is irrelevant in the real world if a database is spending 99% of its time on CPU processing. The workloads never utilize the full capabilities of the storage array. In contrast, peak IOPS capabilities might be critical in a consolidation platform in which the storage array is expected to be loaded to its peak capabilities.
- Always consider latency as well as IOPS in any storage test. Many storage arrays in the market make claims of extreme levels of IOPS, but the latency renders those IOPS useless at such levels. The typical target with all-flash arrays is the 1ms mark. A better approach to testing is not to measure the maximum possible IOPS, but to determine how many IOPS a storage array can sustain before average latency is greater than 1ms.

5.1 EF-Series OLTP Performance

Figure 4 shows EF-Series performance results for the entry-level EF280, the midrange EF570, and the end-to-end NVMe EF600 configured with DDP for a typical OLTP workload. Testing was done with the array configured to use 24 SSDs in a pool, the FC host protocol or NVMe/FC for the EF600, and a workload of 75% read and 25% write. At approximately 300µs of latency, the EF280 produces 100,000 IOPS, the EF570 produces 400,000 IOPS, and the EF600 delivers over 1,000,000 IOPS.

Figure 4) Performance results comparison for OLTP configuration.



5.2 Oracle Automatic Workload Repository and Benchmarking

The gold standard for Oracle performance comparison is an Oracle Automatic Workload Repository (AWR) report.

There are multiple types of AWR reports. From a storage point of view, a report generated by running the `awrrpt.sql` command is the most comprehensive and valuable because it targets a specific database instance and includes some detailed histograms that break down storage I/O events based on latency.

Comparing two performance arrays ideally involves running the same workload on each array and producing an AWR report that precisely targets the workload. In the case of a very long-running workload, a single AWR report with an elapsed time that encompasses the start and stop time can be used, but it is preferable to break out the AWR data as multiple reports. For example, if a batch job ran from midnight to 6 am, create a series of one-hour AWR reports from midnight to 1 am, from 1 am to 2 am, and so on.

In other cases, a very short query should be optimized. The best option is an AWR report based on an AWR snapshot created when the query begins and a second AWR snapshot created when the query ends. The database server should be otherwise quiet to minimize the background activity that would obscure the activity of the query under analysis.

Note: Where AWR reports are not available, Oracle statspack reports are a good alternative. They contain most of the same I/O statistics as an AWR report.

5.3 Oracle AWR and Troubleshooting

An AWR report is also the most important tool for analyzing a performance problem.

As with benchmarking, performance troubleshooting requires that you precisely measure a workload. When possible, provide AWR data when reporting a performance problem to the NetApp support center or when working with a NetApp or partner account team about a new solution.

When providing AWR data, consider the following requirements:

- Run the `awrrpt.sql` command to generate the report. The output can be either text or HTML.
- If Oracle RAC is used, generate AWR reports for each instance in the cluster.
- Target the specific time the problem existed. The maximum acceptable elapsed time of an AWR report is generally one hour. If a problem persists for multiple hours or involves a multihour operation such as a batch job, provide multiple one-hour AWR reports that cover the entire period to be analyzed.
- If possible, adjust the AWR snapshot interval to 15 minutes. This setting allows a more detailed analysis to be performed. This also requires additional executions of `awrrpt.sql` to provide a report for each 15-minute interval.
- If the problem is a very short-running query, provide an AWR report based on an AWR snapshot created when the operation begins and a second AWR snapshot created when the operation ends. The database server should be otherwise quiet to minimize the background activity that would obscure the activity of the operation under analysis.
- If a performance problem is reported at certain times but not others, provide additional AWR data that demonstrates good performance for comparison.

5.4 calibrate_io

The `calibrate_io` command should never be used to test, compare, or benchmark storage systems. As stated in the Oracle documentation, this procedure calibrates the I/O capabilities of storage.

Calibration is not the same as benchmarking. The purpose of this command is to issue I/O to help calibrate database operations and improve their efficiency by optimizing the level of I/O issued to the host. Because the type of I/O performed by the `calibrate_io` operation does not represent actual database user I/O, the results are not predictable and are frequently not even reproducible.

5.5 SLOB2

SLOB2, the Silly Little Oracle Benchmark, has become the preferred tool for evaluating database performance. It was developed by Kevin Closson and is available here. It takes minutes to install and configure, and it uses an actual Oracle database to generate I/O patterns on a user-definable tablespace. It is one of the few testing options available that can saturate an all-flash array with I/O. It is also useful for generating much lower levels of I/O to simulate storage workloads that are low IOPS but latency sensitive.

5.6 Swingbench

Swingbench can be useful for testing database performance, but it is extremely difficult to use Swingbench in a way that stresses storage. NetApp has not seen any tests from Swingbench that yielded enough I/O to be a significant load on any EF array. In limited cases, the Order Entry Test (OET) can be used to evaluate storage from a latency point of view. This could be useful in situations where a database has a known latency dependency for specific queries. Care must be taken to make sure that the host and network are properly configured to realize the latency potentials of an all-flash array.

5.7 HammerDB

HammerDB is a database testing tool that simulates TPC-C and TPC-H benchmarks, among others. It can take a lot of time to construct a sufficiently large data set to properly execute a test, but it can be an effective tool for evaluating performance for OLTP and data warehouse applications.

5.8 Orion

The Oracle Orion tool was commonly used with Oracle 9, but it has not been maintained to ensure compatibility with changes in various host operation systems. It is rarely used with Oracle 10 or Oracle 11 due to incompatibilities with OS and storage configuration.

Oracle rewrote the tool, and it is installed by default with Oracle 12c. Although this product has been improved and uses many of the same calls that a real Oracle database uses, it does not use precisely the same code path or I/O behavior used by Oracle. For example, most Oracle I/Os are performed synchronously, meaning the database halts until the I/O is complete as the I/O operation completes in the foreground. Simply flooding a storage system with random I/Os is not a reproduction of real Oracle I/O and does not offer a direct method of comparing storage arrays or measuring the effect of configuration changes.

That said, there are some use cases for Orion, such as general measurement of the maximum possible performance of a specific host-network-storage configuration, or to gauge the health of a storage system. With careful testing, usable Orion tests could be devised to compare storage arrays or evaluate the effect of a configuration change so long as the parameters include consideration of IOPS, throughput, and latency and attempt to faithfully replicate a realistic workload.

6 General Oracle Configuration

The following parameters are generally applicable to all configurations.

6.1 filesystemio_options

The Oracle initialization parameter `filesystemio_options` controls the use of asynchronous and direct I/O. Contrary to common belief, asynchronous and direct I/O are not mutually exclusive. NetApp has observed that this parameter is frequently misconfigured in customer environments, and this misconfiguration is directly responsible for many performance problems.

Asynchronous I/O means that Oracle I/O operations can be parallelized. Before the availability of asynchronous I/O on various OSs, users configured numerous dbwriter processes and changed the server process configuration. With asynchronous I/O, the OS itself performs I/O on behalf of the database software in a highly efficient and parallel manner. This process does not place data at risk, and critical operations, such as Oracle redo logging, are still performed synchronously.

Direct I/O bypasses the OS buffer cache. I/O on a UNIX system ordinarily flows through the OS buffer cache. This is useful for applications that do not maintain an internal cache, but Oracle has its own buffer cache within the SGA. In almost all cases, it is better to enable direct I/O and allocate server RAM to the SGA rather than to rely on the OS buffer cache. The Oracle SGA uses memory more efficiently. In addition, when I/O flows through the OS buffer, it is subject to additional processing, which increases latencies. The increased latencies are especially noticeable with heavy write I/O when low latency is a critical requirement.

The options for `filesystemio_options` are:

- **async**. Oracle submits I/O requests to the OS for processing. This process allows Oracle to perform other work rather than waiting for I/O completion and thus increases I/O parallelization.
- **directio**. Oracle performs I/O directly against physical files rather than routing I/O through the host OS cache.
- **none**. Oracle uses synchronous and buffered I/O. In this configuration, the choice between shared and dedicated server processes and the number of dbwriters are more important.
- **setall**. Oracle uses both asynchronous and direct I/O.

In almost all cases, the use of `setall` is optimal, but consider the following issues:

- Some customers have encountered asynchronous I/O problems in the past, especially with previous Red Hat Enterprise Linux 4 (RHEL4) releases. These problems are no longer reported, however, and asynchronous I/O is stable on all current OSs.
- If a database has been using buffered I/O, a switch to direct I/O might also warrant a change in the SGA size. Disabling buffered I/O eliminates the performance benefit that the host OS cache provides for the database. Adding RAM back to the SGA repairs this problem. The net result should be an improvement in I/O performance.
- Although it is almost always better to use RAM for the Oracle SGA than for OS buffer caching, it might be impossible to determine the best value. For example, it might be preferable to use buffered I/O with very small SGA sizes on a database server with many intermittently active Oracle instances. This arrangement allows the flexible use of the remaining free RAM on the OS by all running database instances. This is a highly unusual situation, but it has been observed at some customer sites.

Note: The `filesystemio_options` parameter has no effect in DNFS and ASM environments. The use of DNFS or ASM automatically results in the use of both asynchronous and direct I/O.

Best Practice

NetApp recommends setting `filesystemio_options` to `setall`, but be aware that under some circumstances the loss of the host buffer cache might require an increase in the Oracle SGA.

6.2 db_file_multiblock_read_count

The `db_file_multiblock_read_count` parameter controls the maximum number of Oracle database blocks that Oracle reads as a single operation during sequential I/O. This parameter does not, however, affect the number of blocks that Oracle reads during any and all read operations, nor does it affect random I/O. Only sequential I/O is affected.

Oracle recommends that the user leave this parameter unset. Doing so allows the database software to automatically set the optimum value. This generally means that this parameter is set to a value that yields an I/O size of 1MB. For example, a 1MB read of 8KB blocks would require 128 blocks to be read, and the default value for this parameter would therefore be 128.

Most database performance problems observed by NetApp at customer sites involve an incorrect setting for this parameter. There were valid reasons to change this value with Oracle versions 8 and 9. As a result, the parameter might be unknowingly present in `init.ora` files because the database was upgraded in place to Oracle 10 and later. A legacy setting of 8 or 16, compared to a default value of 128, significantly damages sequential I/O performance.

Best Practice

NetApp recommends the `db_file_multiblock_read_count` parameter should not be present in the `init.ora` file. NetApp has never encountered a situation in which changing this parameter improved performance, but there are many cases in which it caused clear damage to sequential I/O throughput.

6.3 Redo Block Size

Oracle supports either a 512-byte or 4KB redo block size. The default is 512 bytes. The best option is expected to be 512 bytes because this size minimizes the amount of data written during redo operations. However, it is possible that the 4KB size could offer a performance benefit at very high logging rates. For example, a single database with 50MBps of redo logging might be more efficient if the redo block size is larger. A storage system supporting many databases with a large total amount of redo logging might

benefit from a 4KB redo block size. This is because this setting would eliminate inefficient partial I/O processing when only a part of a 4KB block must be updated.

It is not correct that all I/O operations are performed in single units of the redo log block size. At very high logging rates, the database generally performs very large I/O operations composed of multiple redo blocks. The actual size of those redo blocks does not generally affect the efficiency of logging.

Best Practice

NetApp recommends only changing the default block size for cause, such as a documented requirement for an application or because of a recommendation made by NetApp or Oracle customer support.

6.4 Checksums and Data Integrity

One question commonly directed to NetApp is how to secure the data integrity of a database. This question is particularly common when a customer who is accustomed to using Oracle RMAN streaming backups migrates to snapshot-based backups. Notably, RMAN performs integrity checks during backup operations. Although this feature has some value, its primary benefit is for a database that is not used on a modern storage array. When physical disks are used for an Oracle database, it is nearly certain that corruption eventually occurs as the disks age, a problem that is addressed by array-based checksums in true storage arrays.

With a real storage array, data integrity is protected by using checksums at multiple levels. If data is corrupted in an IP-based network, the Transmission Control Protocol (TCP) layer rejects the packet data and requests retransmission. The FC protocol includes checksums, as does encapsulated SCSI data. After it is on the array, SANtricity OS has RAID and checksum protection. Corruption can occur, but, as in most enterprise arrays, it is detected and corrected. Typically, an entire drive fails, prompting a RAID rebuild, and database integrity is unaffected. Less often, SANtricity OS detects a checksum error, meaning that data on the disk is damaged. The disk is then failed out and a RAID rebuild begins. Once again, data integrity is unaffected.

The Oracle datafile and redo log architecture is also designed to deliver the highest possible level of data integrity, even under extreme circumstances. At the most basic level, Oracle blocks include checksum and basic logical checks with almost every I/O. If Oracle has not crashed or taken a tablespace offline, then the data is intact. The degree of data integrity checking is adjustable, and Oracle can also be configured to confirm writes. As a result, almost all crash and failure scenarios can be recovered, and in the extremely rare event of an unrecoverable situation, corruption is promptly detected.

Most NetApp customers using Oracle databases discontinue the use of RMAN and other backup products after migrating to snapshot-based backups. There are still options in which RMAN can be used to perform block-level recovery with SMO. However, on a day-to-day basis, RMAN, NetBackup, and other products are only used occasionally to create monthly or quarterly archival copies.

Some customers choose to run `dbv` periodically to perform integrity checks on their existing databases. NetApp discourages this practice because it creates unnecessary I/O load. As discussed above, if the database was not previously experiencing problems, the chance of `dbv` detecting a problem is close to zero, and this utility creates a very high sequential I/O load on the network and storage system. Unless there is reason to believe corruption exists, such as exposure to a known Oracle bug, there is no reason to run `dbv`.

7 Sizing

Oracle performance has been centered on I/O. Traditionally, users improved this performance by either increasing the number of spindles or making the spindles go more quickly. With the advent of the NetApp EF-Series flash array, you can improve performance by using SSDs.

7.1 EF-Series I/O Overview

There are several factors that can affect the overall performance of an EF-Series storage system, including physical components, such as networking infrastructure, and the configuration of the underlying storage itself. Generically, storage system performance tuning can be defined as following a 40/30/30 rule: 40% of the tuning and configuration is at the storage system level, 30% is at the file system level, and the final 30% is at the application level. The following sections describe the 40% related to storage system specifics. For the file system and application level, some of the general considerations include the following:

- **I/O size.** EF-Series storage systems are largely responsive systems. To complete an I/O operation, they require a host to request that operation. The I/O size of the individual requests from the host can have a significant effect on either the number of IOPS or throughput, described in megabytes per second (MBps) or gigabytes per second (GBps). Larger I/Os typically lead to lower numbers of IOPS and larger MBps, and the opposite is true as well. This relationship is defined with the equation Throughput = IOPS × I/O size.
- **Read versus write requests.** In addition to the I/O size, the percentage of read versus write I/O requests processed at the storage system level also has a potential effect on the storage system. You should consider this percentage when designing a solution.
- **Sequential versus random data streams.** Host requests to the underlying disk media logical block addresses can be sequential or random, which has a significant effect on performance at the storage system level. The sequence influences the ability of physical media to respond effectively to the request with minimal latency; the sequence also influences the effectiveness of the storage system's caching algorithms. An exception to increased latency of random requests is for SSDs, which do not have mechanically invoked latency.
- **Number of concurrent I/O operations.** The number of outstanding I/O operations applied to a given volume can vary according to several factors, including whether the file system uses raw, buffered, or direct I/O. Generally, most volumes in an EF-Series storage system are striped across several drives. Providing a minimal amount of outstanding I/O to each individual disk can cause underutilization of the resources in the storage system, resulting in less than desired performance characteristics.

If you are new to NetApp EF-Series technology, it might be helpful to review the differences between RAID 10, RAID 5, RAID 6, and DDP technology. Table 3 compares the usable capacity for different RAID levels. For completeness, all RAID levels supported by EF-Series systems are shown.

Table 3) Comparison of usable capacity for different RAID levels.

Desired Feature	RAID 0	RAID 1 RAID 10	RAID 5	RAID 6	DDP
Usable capacity	100%	50%	(N-1) ÷ N where N is the selected drive count in the volume group	(N-2) ÷ N where N is the selected drive count in the volume group	80% minus selected preservation capacity

7.2 Estimating I/O

Estimating the number of I/O operations required for a system is crucial when you size a database. This exercise helps you understand how to keep the database instance performing within acceptable limits.

You must estimate I/O when you are unable to get the actual physical I/O numbers for the system. This typically happens for new systems that are in the process of being constructed. The following sections provide formulas for estimating I/O.

New OLTP Database System

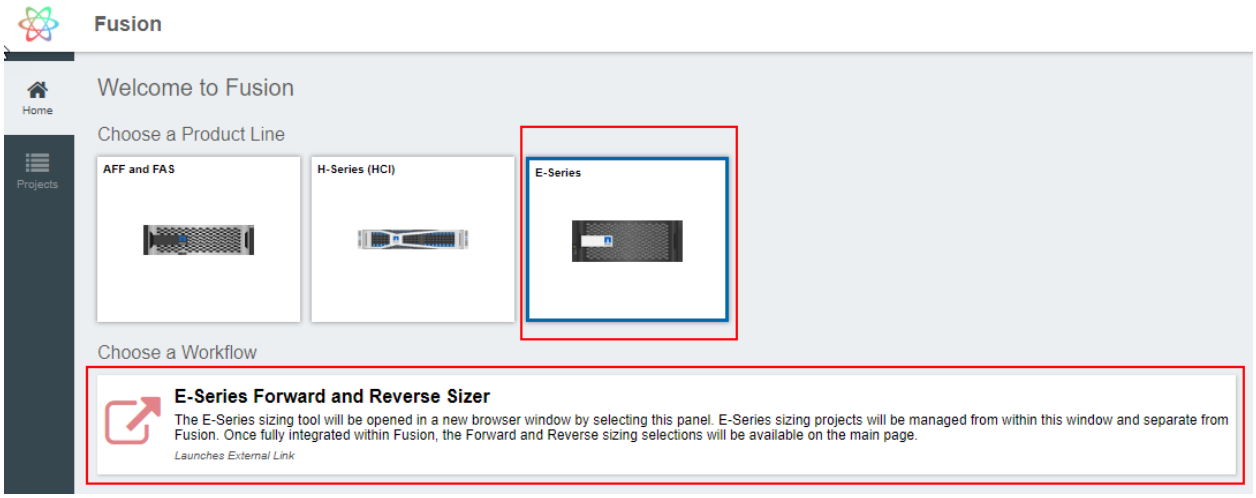
The following items must be considered:

- Business transaction
- Duration of business transaction
- Acceptable latency for the application
- Approximate ratio of transaction and system transactions (I/O)

The easiest way to estimate sizing for Oracle with EF-Series systems is by following this example:

1. Estimate the number of business transactions. For example, estimate the number of business transactions as 600,000,000 per day. The business runs 24/7.
2. Assume that one business transaction creates 25 I/O operations. Therefore, for our example, the I/O operations per day are 600,000,000 x 25 = 15,000,000,000 I/O per day.
3. Using standard storage sizing tools requires IOPS information. From the example, this is (15,000,000,000) ÷ 86,400 = 173,612 IOPS.
4. Acceptable latency for the application should be considered. For example, the application might want less than 1 millisecond of latency.
5. Input this number to the NetApp Fusion tool and use the E-Series Forward and Reverse Sizer. See Figure 5.

Figure 5) Selecting the E-Series Forward and Reverse Sizer.



6. To use the sizer, you must also estimate the percentage of reads and writes. 80% reads and 20% writes are typical of an OLTP database.

Existing OLTP Database System

When sizing for an existing database environment, understanding the type of workload and interpreting the statistical data is helpful. It is important to gather statistics during periods of peak stress on the system. The Oracle AWR report described in Oracle Automatic Workload Repository and Benchmarking can provide the peaks for the time frame in which you monitor the system.

After either IOPS or throughput (MBps) of the system is captured, it can be entered into the E-Series performance sizing tool. Go to NetApp Fusion and select the E-Series product line for an E-Series or EF-Series sizing. Figure 6 shows the input fields for a reverse sizing. From our example, 173,612 IOPS were required. As shown in Figure 7, one EF280 with 24 SSDs can provide over 184,000 IOPS with under 1ms of latency.

Figure 6) E-Series reverse sizer input fields.

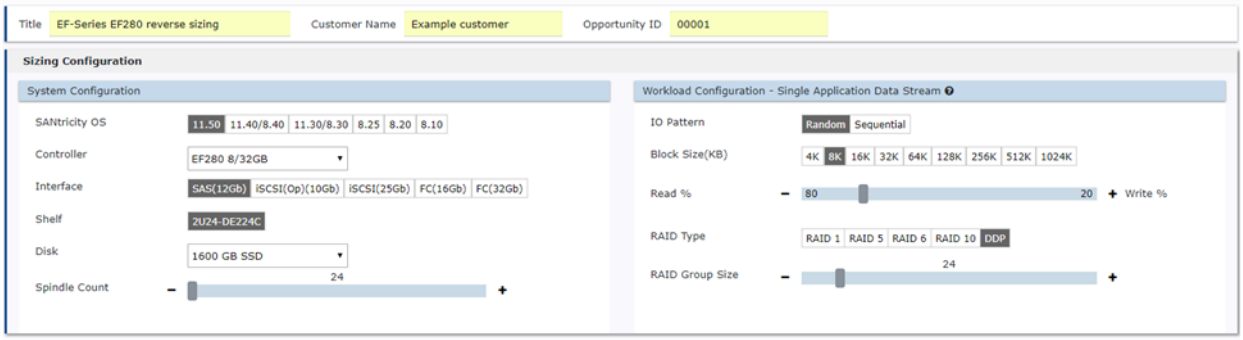
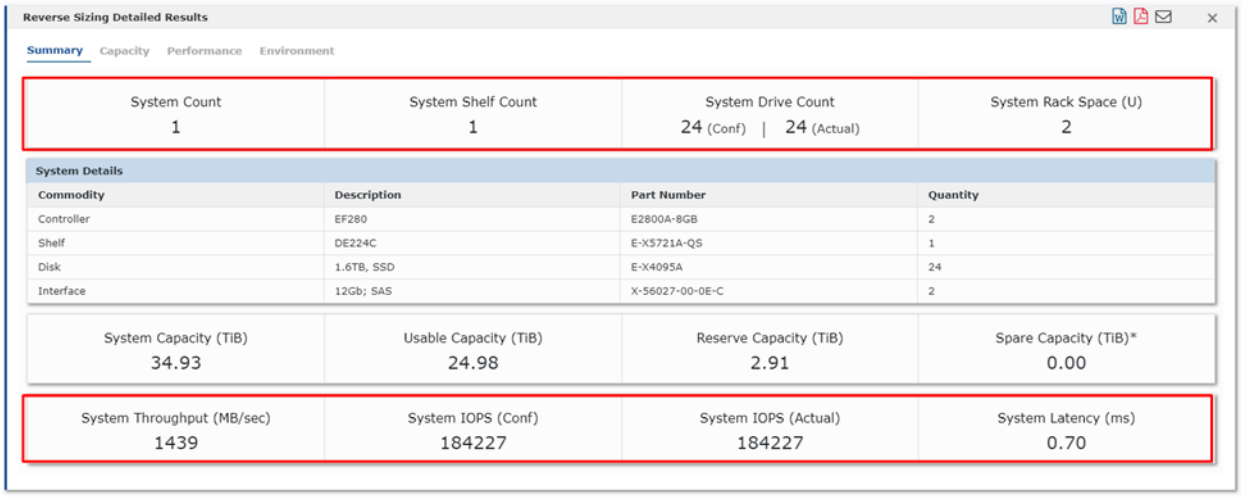


Figure 7) Reverse sizing output.



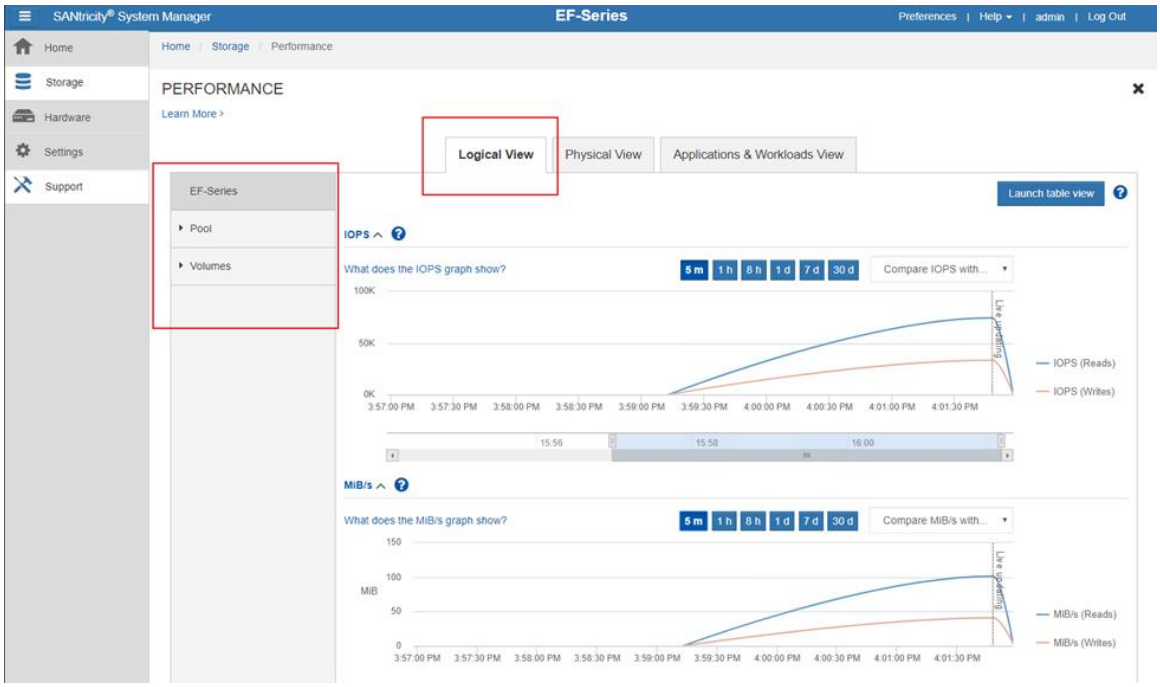
8 EF-Series Performance Monitoring Using SANtricity System Manager

During storage system operations, it can be useful to monitor storage system performance. Using NetApp SANtricity System Manager, you can view EF-Series performance data in both textual and graphical dashboard formats.

SANtricity System Manager provides exceptional performance monitoring that enables you to capture logical, physical, and application-level workloads. It adds new functionality, including application and workload tagging, enhanced performance data, an embedded monitor, and a graphical view of the volume usage. You can observe the performance monitor by selecting View Performance Details in SANtricity System Manager. You can see the performance of three main categories:

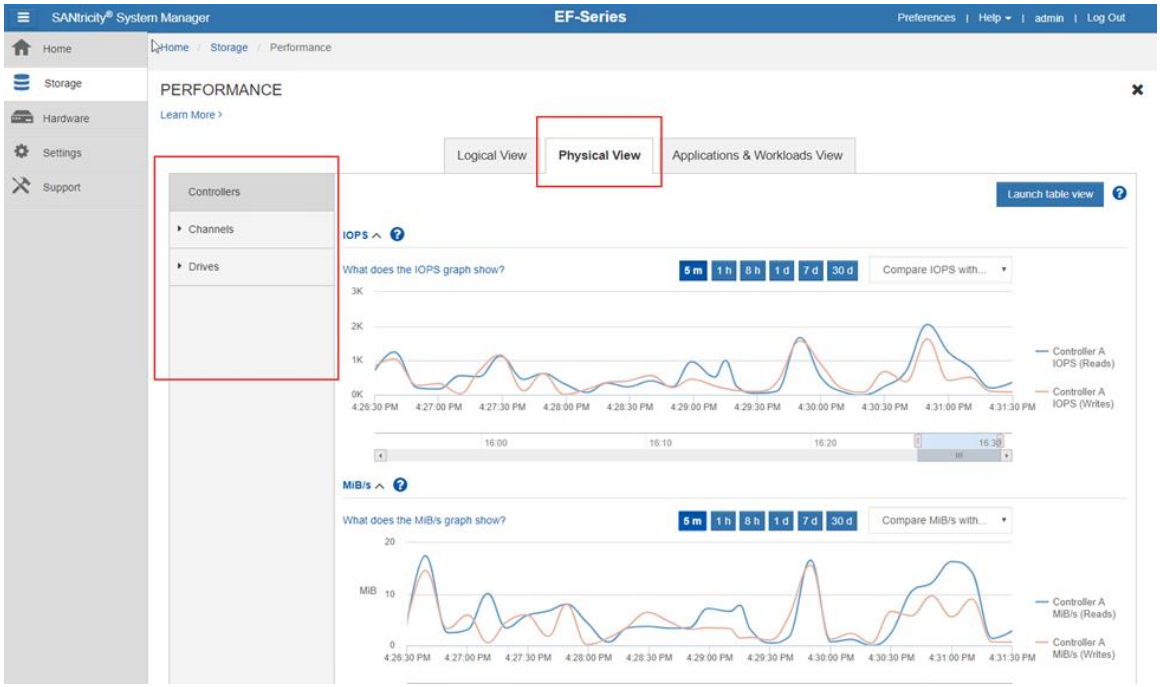
- The logical view enables you to filter information by pools and volume groups or volumes. See Figure 8.

Figure 8) Example of performance monitor with logical view.



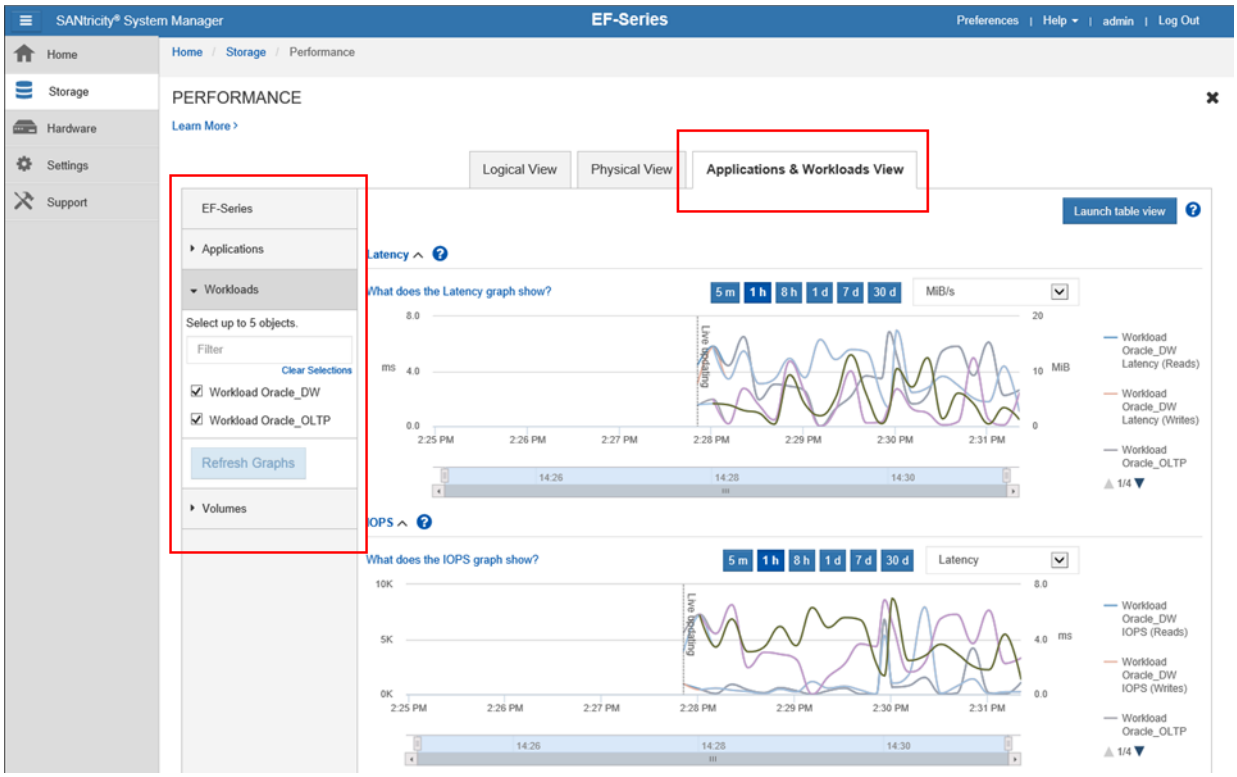
- The physical view enables you to filter information by host channels and drives. See Figure 9.

Figure 9) Example of performance monitor with physical view.



- The applications and workloads view enables you to filter information by applications, workloads, and volumes within the workloads. See Figure 10.

Figure 10) Example of performance monitor with applications and workloads view.



For additional detail on the performance monitor, see the video EF-Series Performance Monitor.

9 Recommendations

EF-Series requires very few changes to the default settings for maximum performance.

9.1 Storage

Table 4 shows the basic tuning parameters, default settings, and recommendations for the EF-Series storage array.

Table 4) Storage tuning parameters.

Name	Context	Default	Available Options	Recommendations
Cache block size	Array	32K	4K, 8K, 16K, and 32K	32K
Automatic cache flushing	Array	80%	0% to 100%	80%
Segment size	Volume	128K	32K, 64K, 128K, 256K, 512K	64K or 128K
Read caching	Volume	Enable	Enable/disable	Enable
Dynamic cache read prefetch	Volume	Enable	Enable/disable	Disable
Write cache	Volume	Enable	Enable/disable	Enable

Name	Context	Default	Available Options	Recommendations
Write cache mirroring	Volume	Enable	Enable/disable	Enable

9.2 Oracle ASM

Table 5 shows the basic tuning parameters for Oracle ASM and database instance.

Table 5) Oracle ASM and instance settings.

Name	Default	Available Options	Recommendations
ASM AU	1MB	1MB to 64MB	64MB
File system I/O options	None	None	Not used
db_file_multiblock_read_count	128	8 to 128	Default

9.3 Linux and NVMe-oF

Table 6 shows the basic tuning parameters for Linux. For complete setup of the Linux OS and NVMe-oF see the SANtricity Software Express Configuration for Linux.

Table 6) Linux tuning parameters.

Name	Default	Available Options	Recommendations
I/O scheduler	cfq	cfq, anticipatory, deadline, or noop	noop

10 Conclusion

In this report, we present the NetApp EF-Series solution to high-performance Oracle databases. This solution allows you to leverage best-in-class, end-to-end, modern SAN and NVMe technologies to deliver business-critical IT services today while also preparing for the future.

With the EF-Series, NetApp has created a SAN array that is future-ready, usable today; and easily implemented within your current operational processes and procedures.

The NetApp EF-Series flash array is a market leader in delivering high performance, consistent low latency, and advanced HA features.

The array is easy to provision with the embedded System Manager. If you have many systems to prepare, you can also provision the array by using the REST API.

The EF-Series flash array also has robust built-in monitoring capabilities that enable you to troubleshoot performance issues at the logical level, physical level, and application level, as shown in the video EF-Series Performance Monitor.

In addition to solving the extreme latency requirements, the EF-Series flash array also resolves Oracle database challenges in the following ways:

- It dramatically boosts the performance of existing applications and lowers the cost per IOPS ratio without requiring that you rearchitect the application.
- It increases Oracle performance with RAID10 and DDP.
- Better response times increase user productivity, improving business efficiency.

- Oracle ASM provides an alternative solution for high availability and disaster recovery.

For more technical information, see the EF-Series Flash Storage Arrays page.

Where to Find Additional Information

To learn more about the information that is described in this document, review the following websites:

- Oracle Product Documentation
  https://docs.oracle.com/
- NetApp Product Documentation
  https://docs.netapp.com

Version History

Version	Date	Document Version History
Version 1.0	August 2019	Initial release

Refer to the Interoperability Matrix Tool (IMT) on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer's installation in accordance with published specifications.

Copyright Information

Trademark Information