WHITE PAPER

BEST PRACTICES FOR RUNNING ORACLE ON DELL EMC XTREMIO X2

Abstract

This White Paper describes the best practices and recommendations when deploying an Oracle® Database Management System 12c on Oracle Linux 7.x running on top of DELL EMC's XtremIO X2 enterprise all-flash storage array.

May, 2019

Contents

Executive Summary

Dell EMC XtremIO is a market-leading, purpose-built, all-flash array that offers consistently high performance with low latency, unmatched storage efficiency with inline, all-the-time data services, rich application, integrated copy services and unprecedented management simplicity. It is designed from the ground up to unlock flash technology's instant performance potential by uniquely leveraging the characteristics of SSDs and uses advanced inline data reduction methods to reduce the physical data that must be stored. XtremIO's storage system uses industry-standard components and custom-designed intelligent software to deliver unparalleled levels of performance, achieving consistent low latency for up to millions of IOPS.

XtremIO has always provided simple, easy-to-use management. The XtremIO Management Server (XMS) delivers an HTML5 user interface that is a simple and easy-to-use interface for storage administrators. XMS allows storage administrators the ability to provisions storage with very little setup and planning.

XtremIO is designed and optimized for databases and for DBAs, providing the following benefits.

Predictable Performance

XtremIO provides predictable and consistency low-latency performance. With XtremIO scale-up and scale-out architecture, year-to-year growth is easy. Initial investment is preserved, and application performance is improved. The performance is predictable and provides sub-millisecond response times regardless of the workload and environment – be it production, QA, test or development.

Incredible simplicity

With XtremIO, there is no need for planning and tuning the location and number of database files. XtremIO is a real scale-up / scale-out and N-active-active storage controller architecture in which all the array volumes are served by all array resources. All DBA tasks are fast and simple with 1 to 3 steps.

Agility

The typical enterprise applications require multiple copies such as test/development, reporting or online analytics. DBAs and test/dev engineers often have to spend hours managing the DB creation and refreshing the environments while often being limited by capacity, performance and number of copies. XtremIO's Integrated Copy Data Management (iCDM) allows for instant XtremIO Virtual Copies (XVCs) to be created from production with no performance impact. These copies can be repurposed for near real time analytics, test/dev and any other use case- all with complete space efficiency.

Protection

Protecting the database is easy with XtremIO. There is no need for any design covering RAID type, data file capacity, load balancing, and tuning. The data is protected with a proprietary flash-optimized algorithm called XtremIO Data Protection (XDP). XDP is very different from RAID in several ways. Since XDP is always working within an all-flash storage array, several criteria were important in the design of this protection scheme. XDP benefits include ultra-low capacity overhead, high levels of data protection in case of double SSD failure, rapid rebuild times, flash endurance, and of course extreme performance.

And with XtremIO virtual copies it is easy to protect and recover from any operational and logical corruption; XVC's allow the creation of frequent point-in-time copies (according to RPO intervals – seconds, minutes, hours) and use them to recover from any data corruption. An XVC can be kept in the system for as long as needed. Recovery using XtremIO virtual copy is instantaneous and does not impact system performance.

Introduction

Oracle's Database Management System (DBMS) operates at peak performance on the XtremIO Storage Array solution, regardless of the workload it encounters. This includes diverse workloads such as online transaction processing (OLTP), data warehousing, and hybrid workloads.

The XtremIO Storage Array delivers predictable high performance and consistent low latency. The recommendations and best practices described in this white paper are geared to assist Storage and Database administrators to maximize the performance and data capacity utilization of the XtremIO X2 Storage Array, when deploying an Oracle DBMS on Oracle Linux 7.x

Test Setup

All the tests performed on this white paper were conducted with the following equipment:

Component	**Properties**
XtremIO Array	XtremIO X2R, Single Brick, 18 x 1.92TB XtremApp 6.0.1-30
Server	Intel Based Server 755 GB RAM 2 x Intel(R) Xeon(R) CPU E5-2658 v4 @ 2.30GHz (28 cores)
Operating System	Operating System: Oracle Linux 7 (x86_64) UEK Release 4
Multipath	Device Mapper for Multi-Path Software + Oracle ASMLib 2.X
Volume Manager	Oracle ASM For Grid Infrastructure and Database
Oracle	Oracle Database 12c R2 Grid and Database Software

Test Performance Results

In this section, we take a deeper look at performance statistics from our XtremIO X2 array while running SLOB.

SLOB is an Oracle I/O workload generation tool kit which possesses the following characteristics:

- SLOB supports testing Oracle logical read (SGA buffer gets) scaling.
- SLOB supports testing physical random single-block reads (db file sequential read).
- SLOB supports testing random single block writes (DBWR flushing capacity).
- SLOB supports testing extreme REDO logging I/O.
- SLOB consists of simple PL/SQL.
- SLOB is entirely free of all application contention.

Software Configuration

- Red Hat Enterprise Linux Server release 6.7
- Oracle 12c 12.1.0.2
- Grid Control 12c 12.1.0.2
- SLOB 2.3

Database Storage Configuration

- ASM
- Diskgroup DATA of 4 volumes of 512 GB
- Diskgroup REDO of 4 volumes of 512 GB

Figure 1 shows SLOB performance test results on XtremIO X2. The storage array was prefilled up to 90% in order to simulate an environment as close as possible to a customer's production environment.

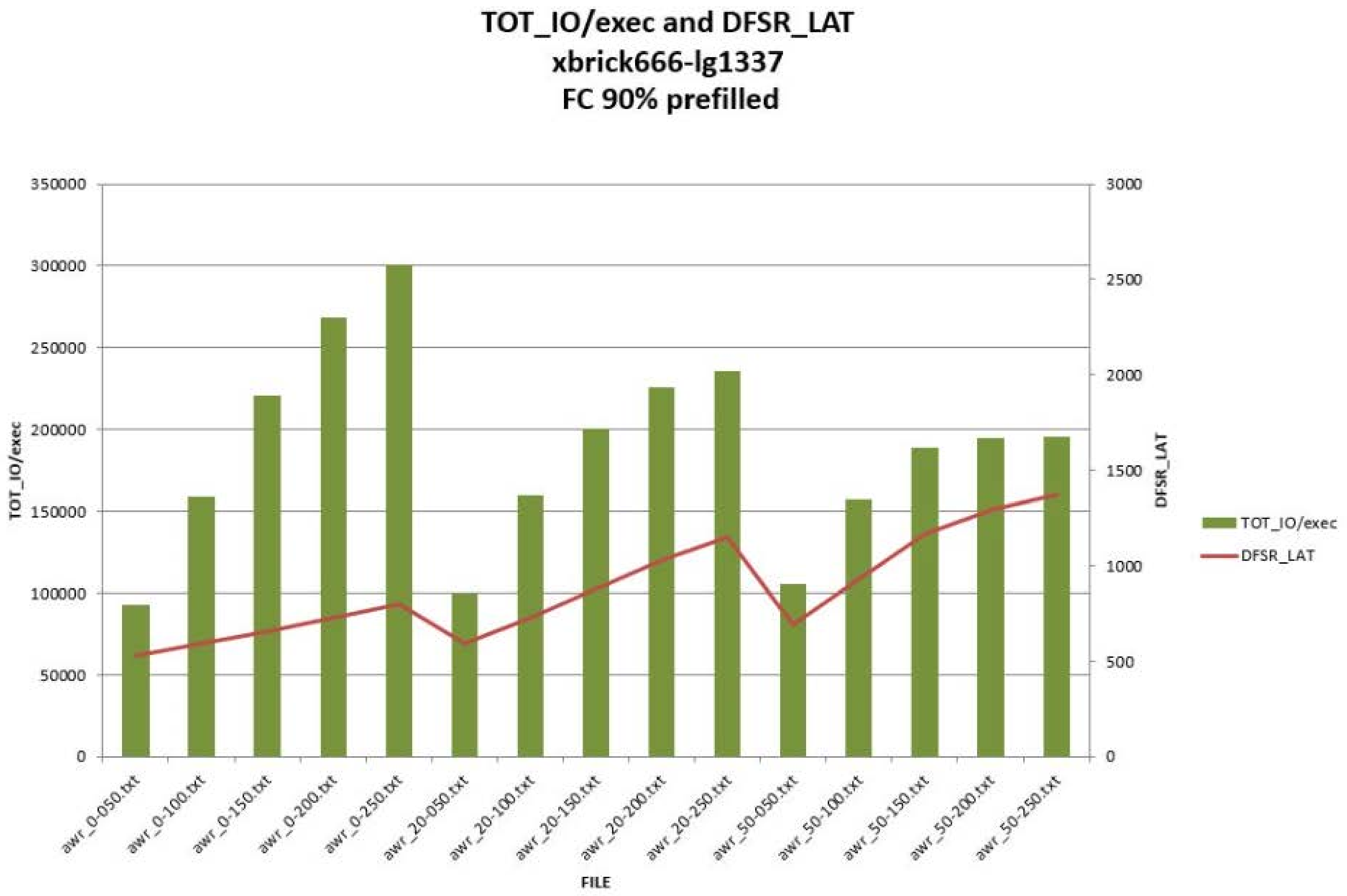
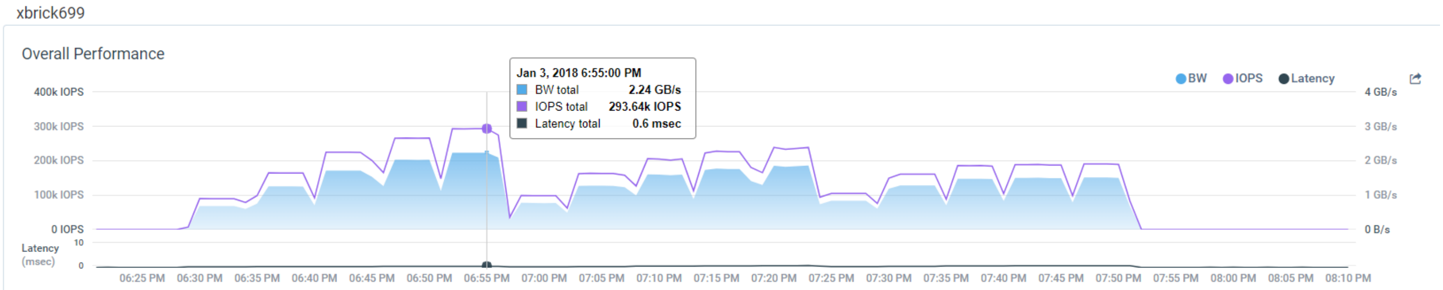


Figure 1. SLOB Performance Test Results on XtremIO X2

Figure 2 shows the performance metrics from the perspective of the array. As we can see, XtremIO X2 is handling storage bandwidths as high as ~2.24GB/s with over 290k IOPS during the SLOB performance test.

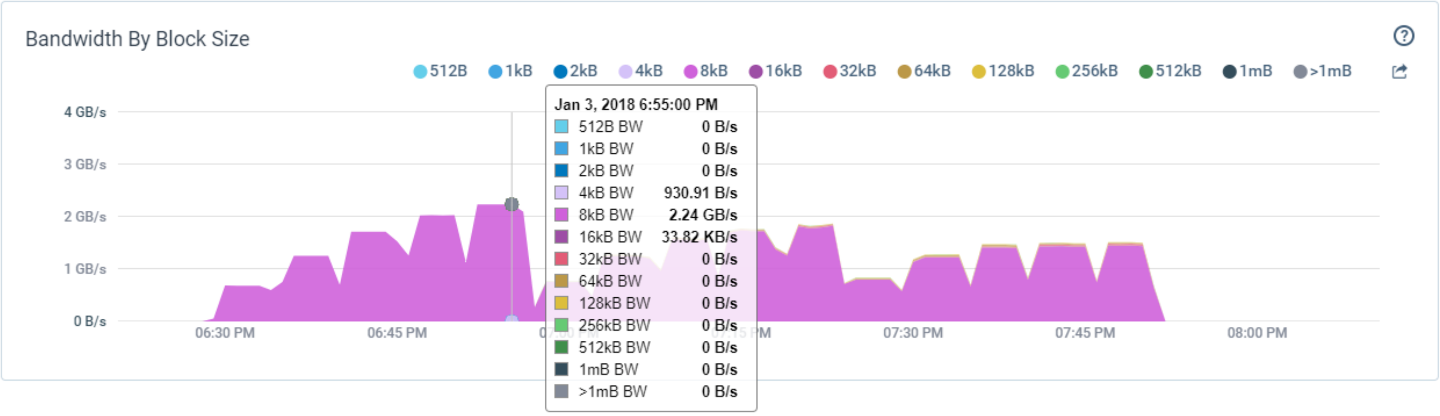


Jan 3, 2018 6:55:00 PM

BW total	2.24 GB/s
IOPS total	293.64k IOPS
Latency total	0.6 msec

Figure 2. XtremIO X2 IOPS and I/O Bandwidth During SLOB Performance Test

Figure 3 shows the block sizes distribution during the SLOB performance test. We can see that most of the bandwidth used is 8KB blocks, as this was the block size configured at the test level to use against our storage array.

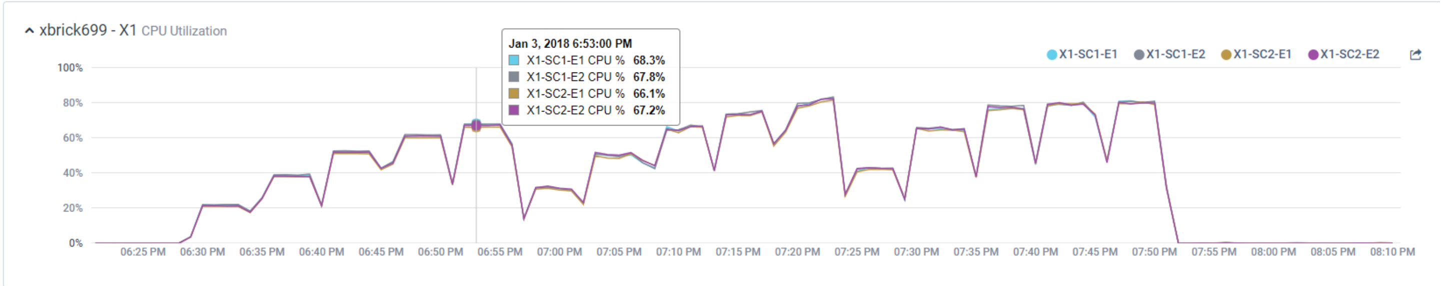


Jan 3, 2018 6:55:00 PM

512B BW	0 B/s
1kB BW	0 B/s
2kB BW	0 B/s
4kB BW	930.91 B/s
8kB BW	2.24 GB/s
16kB BW	33.82 KB/s
32kB BW	0 B/s
64kB BW	0 B/s
128kB BW	0 B/s
256kB BW	0 B/s
512kB BW	0 B/s
1mB BW	0 B/s
>1mB BW	0 B/s

Figure 3. XtremIO X2 IOPS and I/O Bandwidth During SLOB Performance Test

Figure 4 shows the CPU utilization of the Storage Controllers during the SLOB performance test. We can see that the CPUs are utilized well during this process with utilization close to 67%. We can also see the excellent synergy across the X2 cluster, with all the Active-Active Storage Controllers' CPUs sharing the load and effort, and the CPU utilization virtually equal for all the controllers for the entire process.



Jan 3, 2018 6:53:00 PM

X1-SC1-E1 CPU %	68.3%
X1-SC1-E2 CPU %	67.8%
X1-SC2-E1 CPU %	66.1%
X1-SC2-E2 CPU %	67.2%

Figure 4. XtremIO X2 IOPS and I/O Bandwidth During SLOB Performance Test

In Figure 5, we can see the IOPS and latency stats in the SLOB Performance Test. The graph shows again that IOPS are well over 290k but that the latency for all I/O operations remains less than 0.9msec, yielding the excellent performance of the Oracle Database.

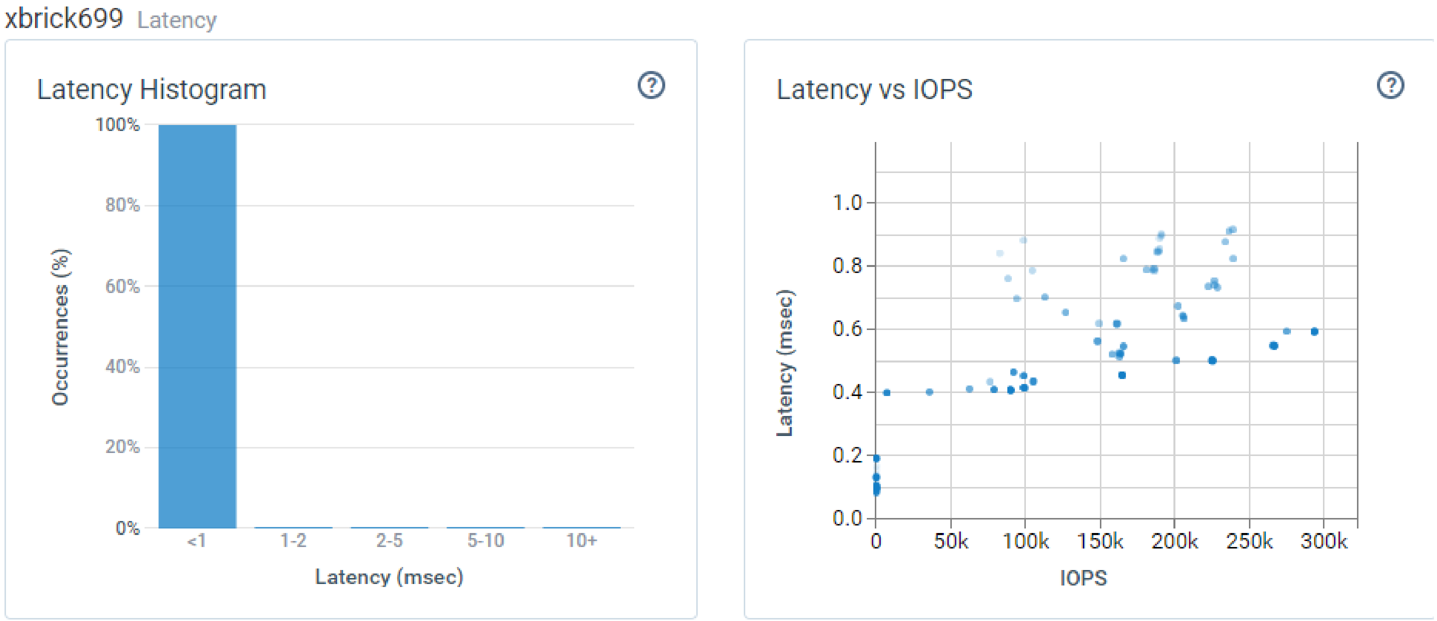


Figure 5. XtremIO X2 IOPS and I/O Bandwidth

Storage Array: Dell EMC XtremIO X2 All-Flash Array

Dell EMC XtremIO is an enterprise-class scalable all-flash storage array that provides rich data services with high performance. It is designed from the ground up to unlock Flash Technology's instant performance potential by uniquely leveraging the characteristics of SSDs. It also uses advanced inline data reduction methods to reduce the physical data that must be stored on the disks.

XtremIO's storage system uses industry-standard components and proprietary-intelligent software to deliver unparalleled levels of performance, achieving consistent low latency for up to millions of IOPS. The system comes with a simple, easy-to-use interface for storage administrators and requires very little planning to set up, before provisioning. XtremIO fits a wide variety of use cases for customers needing a fast and efficient storage system for their datacenters.

XtremIO X2 Overview

XtremIO X2 is the new generation of the Dell EMC's All-Flash Array storage system. X2 adds enhancements and flexibility to the previous generation storage array, which had already provided proficiency and high performance. By supporting the following features, the system provides the extra value and advancements required in the evolving world of computer infrastructure.

- Scale-Up for a more flexible system.
- Write boost for a more sensible and higher-performing storage array.
- NVRAM for improved data availability.
- New web-based UI for managing the storage array and monitoring its alerts and performance statistics.

The XtremIO X2 Storage Array uses building blocks called X-Bricks. Each X-Brick has its own compute, bandwidth and storage resources, and can be clustered together with additional X-Bricks to grow in both performance and capacity (Scale-Out). Each X-Brick can also grow individually in terms of capacity, with an option to add to up to 72 SSDs for each X-Brick.

XtremIO architecture is based on a metadata-centric, content-aware system. This helps streamline data operations efficiently without requiring any movement of data post-write for any maintenance reason (e.g. data protection and data reduction are done inline). The system lays out the data uniformly across all SSDs in all X-Bricks in the system, using unique fingerprints of the incoming data and controlling access with metadata tables. This provides an extremely balanced system across all X-Bricks, in terms of compute power, storage bandwidth and capacity.

Using the same unique fingerprints, XtremIO is equipped with exceptional always-on in-line data deduplication abilities, which highly benefits virtualized environments. Together with its data compression and thin provisioning capabilities (both also in-line and always-on), it achieves incomparable data reduction rates.

System operation is controlled by storage administrators via a stand-alone dedicated Linux-based server, called the XtremIO Management Server (XMS). An intuitive user interface is used to manage and monitor the storage cluster and its performance. The XMS can be either a physical or a virtual server and can manage multiple XtremIO clusters.

With its intelligent architecture, XtremIO provides a storage system that is easy to set-up, needs zero tuning by the client, and does not require complex capacity or data protection planning, which are handled autonomously by the system.

Architecture and Scalability

An XtremIO X2 Storage System is comprised of a set of X-Bricks that together, form a cluster. This is the basic building block of an XtremIO array. There are two types of X2 X-Bricks available: X2-S and X2-R.

X2-S is for environments with storage that are more I/O intensive than capacity intensive. Such applications would typically use smaller SSDs and less RAM. A suitable use of the X2-S is for environments that have high data reduction ratios (i.e. a high compression ratio or a great deal of duplicated data) which lower the capacity footprint of the data significantly.

X2-R X-Brick clusters are made for capacity-intensive environments, with bigger disks, more RAM and a higher potential for expansion in future releases. The two X-Brick types cannot be mixed together in a single system, so deciding which X-Brick type is suitable for your environment must be made in advance.

System Features

The XtremIO X2 Storage Array provides a wide range of built-in features that require no special license. The architecture and implementation of these features is unique to XtremIO and is designed with consideration in mind to the capabilities and limitations of flash media. The following sections list some key features included in the system.

Inline Data Reduction

XtremIO's unique Inline Data Reduction is achieved by two mechanisms: Inline Data Deduplication and Inline Data Compression.

Inline Data Deduplication

Inline Data Deduplication is the removal of duplicate I/O blocks from a stream of data prior to it being written to the flash media. XtremIO inline deduplication is always on, meaning no configuration is needed for this important feature. The deduplication is at a global level, meaning no duplicate blocks are written over the entire array. As an inline and global process, resource-consuming background processes or additional reads and writes are not necessary (in contrast to post-processing deduplication schemes which do require such background processes). This increases SSD endurance and eliminates performance degradation.

Inline Data Compression

Inline data compression is the compression of data prior to it being written to the flash media. XtremIO automatically compresses data after all duplications are removed, ensuring that the compression is performed only for unique data blocks. The compression is performed in real-time and not as a post-processing operation. This way, it does not overuse the SSDs or impact performance. Compressibility rates depend on the type of data written.

Thin Provisioning

XtremIO storage uses a small internal block size, and all volumes are natively thin provisioned. This means that the system consumes capacity only when it is needed. No storage space is ever pre-allocated before writing.

Integrated Copy Data Management

XtremIO pioneered the concept of integrated Copy Data Management (iCDM); the ability to consolidate both primary data and its associated copies on the same scale-out all-flash array for unprecedented agility and efficiency.

XtremIO is one-of-a-kind in its capabilities to consolidate multiple workloads and entire business processes safely and efficiently, providing organizations with a new level of agility and self-service for on-demand procedures. XtremIO provides consolidation, supporting on-demand copy operations at scale, and still maintains delivery of all performance SLAs in a consistent and predictable way.

XtremIO Virtual Copies

For all iCDM purposes, XtremIO uses its own implementation of snapshots, called XtremIO Virtual Copies (XVCs). XVCs are created by capturing the state of data in volumes at a specific point in time and allowing users to access that data when needed, no matter the state of the source volume, i.e. even if the source was deleted. XVCs allow any access type and can be taken either from a source volume or another Virtual Copy.

XtremIO's Virtual Copy technology is implemented by leveraging the content-aware capabilities of the system with a unique metadata tree structure that directs I/O to the data with the right timestamp. This allows efficient copy creation that can sustain high performance, while maximizing the media endurance.

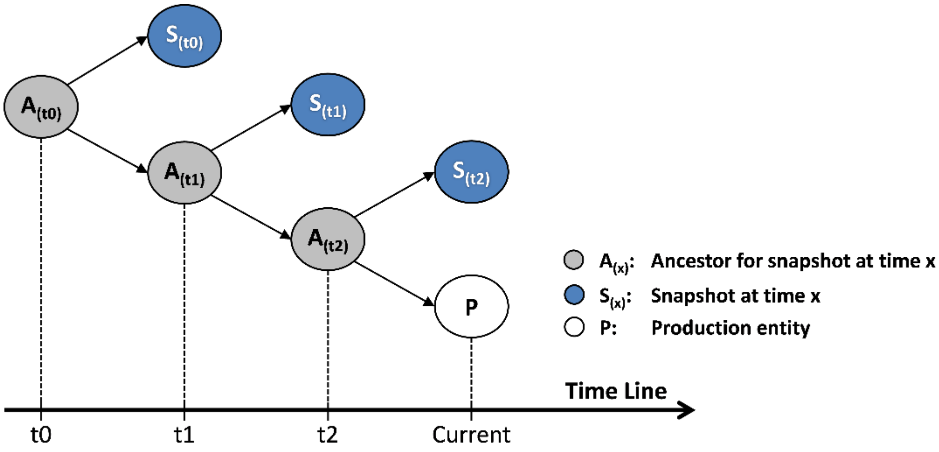


Figure 6. A Metadata Tree Structure Example of XVCs

When creating a Virtual Copy, the system only generates a pointer to the ancestor metadata of the actual data in the system, thus making the operation very quick. This operation does not have any impact on the system and does not consume any capacity at the point of creation, unlike traditional snapshots, which may need to reserve space or copy the metadata for each snapshot. Virtual Copy capacity consumption occurs only when changes are made to any copy of the data. Then the system updates the metadata of the changed volume to reflect the new write and stores its blocks in the system using the standard write flow process.

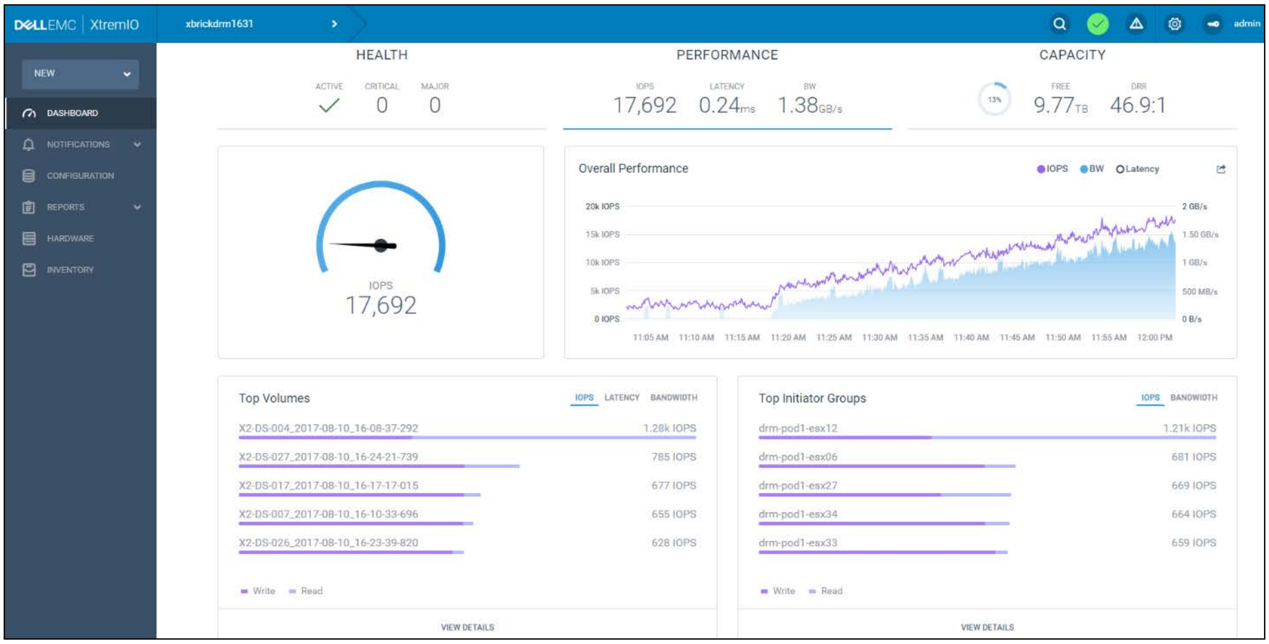
Dashboard

The Dashboard window presents an overview of the cluster. It has three panels:

1. Health: Provides an overview of the system's health status and alerts.

2. Performance (shown in Figure 7): Provides an overview of the system's overall performance and top used Volumes and Initiator Groups.

3. Capacity (shown in Figure 8): Provides an overview of the system's physical capacity and data savings. Note that these figures represent views available in the dashboard and not test results shown in earlier figures.



Top Volumes	
X2-DS-004_2017-08-10_16-08-37-292	1.28k IOPS
X2-DS-027_2017-08-10_16-24-21-739	785 IOPS
X2-DS-017_2017-08-10_16-17-17-015	677 IOPS
X2-DS-007_2017-08-10_16-10-33-696	655 IOPS
X2-DS-026_2017-08-10_16-23-39-820	628 IOPS

Top Initiator Groups	
drm-pod1-esx12	1.21k IOPS
drm-pod1-esx06	681 IOPS
drm-pod1-esx27	669 IOPS
drm-pod1-esx34	664 IOPS
drm-pod1-esx33	659 IOPS

Figure 7. XtremIO Web UI – Dashboard – Performance Panel

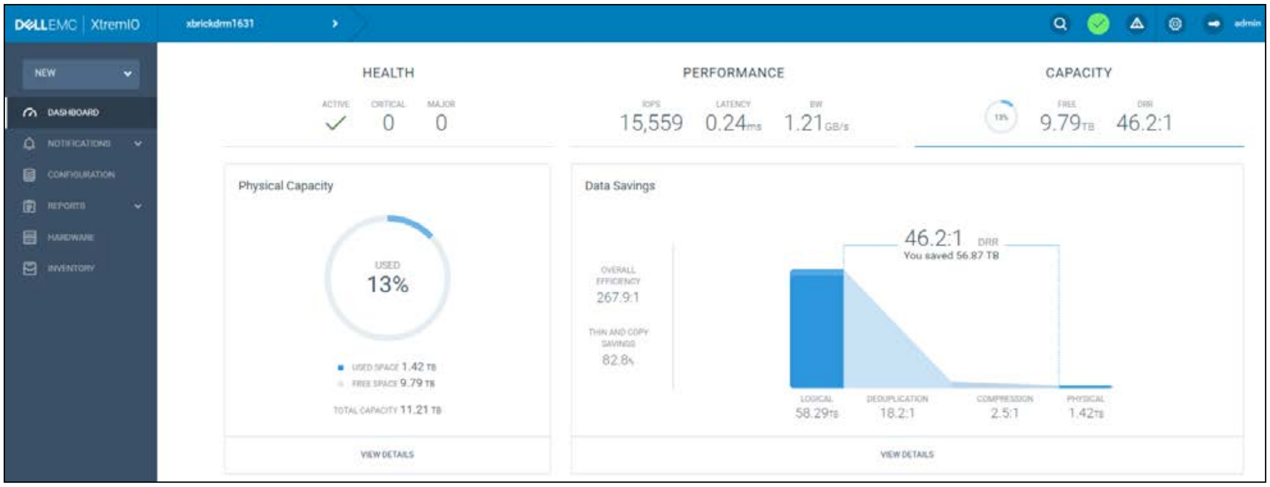


Figure 8. XtremIO Web UI – Dashboard – Capacity Panel

The main Navigation menu bar is located on the left side of the UI. Users can select one of the navigation menu options related to XtremIO's management actions. The main menus contain options for the Dashboard, Notifications, Configuration, Reports, Hardware and Inventory.

Solution's Software Layer

Oracle Physical Linux 7.4 Configuration

Note: Please refer to the latest Host Configuration guide for up to date procedures and best practices - https://support.emc.com/docu56210_XtremIO_Host_Configuration_Guide.pdf

General Guidelines

1. It is recommended to use 8 paths from each host server to the cluster.
2. Keep a consistent, duplex link speed on all paths between the host and the XtremIO cluster.
3. To ensure continuous access to XtremIO storage during cluster software upgrade, verify that a minimum I/O timeout of 30 seconds is set on the HBAs of all hosts connected to the affected XtremIO cluster. Similarly, verify that a minimum timeout of 30 seconds is set for all applications that are using storage from the XtremIO cluster.
4. The HBA queue depth (also referred to as execution throttle) controls the amount of outstanding I/O requests per HBA port. The HBA queue depth should be set to the maximum value.
5. The LUN queue depth controls the amount of outstanding I/O requests per a single path. These settings are controlled in the driver module for the card at the OS level. When connecting Linux host to XtremIO, LUN queue depth setting should retain its default values.
6. I/O scheduling controls how I/O operations are submitted to storage. Linux offers various I/O algorithms (also known as “Elevators”) to accommodate different workloads. When connecting a Linux host to XtremIO storage, set the I/O elevator to either noop or deadline. It is not recommended to use the cfq I/O elevator setting, as it is less optimal for XtremIO storage.

Configuring IO Elevator and Queue Depth using UDEV

1. For instructions on using CLI rather than UDEV, please refer to Host Configuration Guide.
2. As a general rule of thumb, do not change the default Queue Depth setting. If Oracle is the only application attached to XtremIO, consider increasing the Queue Depth between 128 and 256.
3. Create or edit the following file:

```
$ vim /etc/udev/rules.d/99-XtremIO.rules
```

4. Append the following rule to the file:

5. Save the changes made to the file.

```
#increase queue depth on the volume
ACTION=="add|change", SUBSYSTEM=="scsi", ATTR{vendor}=="XtremIO",
ATTR{model}=="XtremApp        ", ATTR{queue_depth}="32"
# Use noop scheduler for added performance
ACTION=="add|change", SUBSYSTEM=="block",
ENV{ID_VENDOR}=="XtremIO", ENV{ID_MODEL}=="XtremApp",ATTR{queue/scheduler}="noop"
# Use noop
ACTION=="add|change", SUBSYSTEM=="block", KERNEL=="dm*",
ENV{DM_NAME}=="??14f0c5*", ATTR{queue/Scheduler}="noop"
```

6. Run the following command to apply the changes:

```
$ udevadm trigger
```

Note: Some Linux operating systems may benefit from using deadline elevator configuration.

Note: In the first rule shown in step 4, there should be eight (8) spaces between 'XtremApp' and the closing quotation mark.

Installing and Configuring the DM-MPIO

The device mapper (also known as DM-MPIO) is a Linux multipathing software that is suitable for balancing I/O to XtremIO Storage Arrays. For more background on DM-MPIO refer to My Oracle Support (Doc ID 753050.1).

To install and configure the device mapper, the following steps should be followed:

4. Via YUM, install the device mapper package

```
[root@ucs3 ~]# yum install device-mapper
```

5. Use chkconfig to enable multipathd daemon:

```
[root@ucs3 ~]# chkconfig multipathd on
```

6. Configure the XtremIO disk device, modify the /etc/multipath.conf file with the following parameters:

```
devices {
      device {
            vendor XtremIO
            product XtremApp
            path_selector "queue-length 0"
            rr_min_io_rq 1
            path_grouping_policy multibus
            path_checker tur
            failback immediate
            fast_io_fail_tmo 15
            user_friendly_names no
            }
```

7. Restart multipathd daemon. As the root user, at the shell prompt, enter:

```
[root@ucs3 ~]# service multipathd restart
```

8. To ensure that the device mapper's nodes are updated and exposed, it is necessary to run the following commands:

```
[root@ucs3 ~]# multipath –F;multipath –v 2
```

Note: Setting the `user_friendly_names` parameter to `no`, sets the unique WWID as the multipath device name (`/dev/mapper/<NAA>`).

Ensuring LUN Accessibility

1. To ensure that XtremIO devices are properly exposed and remain readily accessible via the host without requiring a host reboot, it is recommended to install the `sg3_utils` package.

```
[root@ucs3 ~]# yum install sg3_utils
```

2. To probe the SCSI bus for new LUNs on channels, execute the following as root:

```
[root@ucs3 ~]# rescan-scsi-bus.sh
```

3. At the shell prompt, enter the following command:

```
[root@ucs3 ~]# multipath -ll
3514f0c5c83a001b1 dm-44 XtremIO ,XtremApp
size=2.0T features='0' hwhandler='0' wp=rw
`-+- policy='queue-length 0' prio=1 status=active
  |- 0:0:4:39 sddb 70:144  active ready running
  |- 0:0:5:39 sddv 71:208  active ready running
  |- 1:0:4:39 sdep 129:16  active ready running
  `- 1:0:5:39 sdfj 130:80  active ready running
```

4. Use the following command to correlate between the volume NAA and volume name on the attached XtremIO cluster:

```
[root@lgsup22 ~]# for i in `multipath -ll |grep XtremIO |awk '{print
"/dev/mapper/"$1}'`; do lu_name=$(sg_inq --page=0x83 $i |grep "vendor specific:"|
sed -n 1p |awk '{print $NF}'); br_name=$(sg_inq --page=0x83 $i |grep "vendor
specific:"| sed -n 2p |awk '{print $NF}'); echo $i $lu_name $br_name; done

/dev/mapper/3514f0c553b800001 oracle-db1 xbrick10
```

Oracle ASM

Oracle Automatic Storage Management (ASM) is Oracle's recommended software for supporting Oracle database files.

ASM Features in Oracle 12

- 511 Disk Groups
- 10,000 Oracle ASM disks
- 1 million files for each disk group
- Without any Oracle Exadata Storage, Oracle ASM has the following storage limits if the COMPATIBLE.ASM or COMPATIBLE.RDBMS disk group attribute is set to less than 12.1:
    - 2 TB maximum storage for each Oracle ASM disk
    - 20 PB maximum for the storage system
- Without any Oracle Exadata Storage, Oracle ASM has the following storage limits if the COMPATIBLE.ASM and COMPATIBLE.RDBMS disk group attributes are set to 12.1 or greater:
    - 4 PB maximum storage for each Oracle ASM disk with the AU size equal to 1MB
    - 8 PB maximum storage for each Oracle ASM disk with the AU size equal to 2MB
    - 16 PB maximum storage for each Oracle ASM disk with the AU size equal to 4MB
    - 32 PB maximum storage for each Oracle ASM disk with the AU size equal to 8MB
    - 320 exabytes (EB) maximum for the storage system

For more information:
https://docs.oracle.com/database/121/OSTMG/GUID-BC6544D7-6D59-42B3-AE1F-4201D3459ADD.htm#OSTMG10042

ASM General Recommendations

- External redundancy is generally recommended for XtremIO.
- The XtremIO Storage Array natively provides flash-optimized data protection.

Database Files Location in ASM Disk Groups

The best practices for Storing Oracle DBMS file types in ASM disk groups are outlined in following table:

Database Type	**Grid DG**	**Data DG**	**Redo 1xDG**	**Redo 2xDG**	**FRA DG**
Single-Instance	N/A	Control File SPFILE Data Files Temp Undo	Redo Logs	Multiplexed Redo Logs (if applicable)	Archive Logs Flashback Logs Backup Components
RAC	OCR Voting File SPFILE	Control File Data Files Temp Undo	Redo Logs	Multiplexed Redo Logs (if applicable)	Archive Logs Flashback Logs Backup Components

Note: The second redo data group (DG) is applicable if redo logs are multiplexed.

Number of LUNS per Disk Group

Excellent cluster performance is achieved using an XtremIO Storage Array with just a single LUN in a single disk group. However, in order to maximize performance from a single host, parallelism and adequate utilization of device queues are required.

The best practice to achieve this is using a minimum of four LUNs for the data disk group per array. Doing so enables the hosts, or applications, to use parallelism at various queuing points. This method ensures optimal performance from the XtremIO Storage Array.

The best practices for Disk group configuration and data placement are outlined in following table:

Database Type	**Grid DG**	**Data DG**	**Redo 1xDG**	**Redo 2xDG**	**FRA DG**
Single-Instance	N/A	4 x LUNs per Array	1 x LUN	1 x LUN	1 x LUN per component: Archive, Flashback, Backup, etc.
RAC	1 x LUN	4 x LUNs per Array	1 x LUN	1 x LUN	1 x LUN per component: Archive, Flashback, Backup, etc.

For Oracle 12.x GRID installations with ASM, do not set the physical-block-size, neither for X1 nor for X2 arrays (leave the setting at the default of 512). Example for configuring ssd.conf on Solaris SPARC is shown below.

```
ssd-config-list = "XtremIO XtremApp","throttle-max:64,delay-busy:30000000000,retries-
busy:90,retries-timeout:30,retries-notready:30,cache-nonvolatile:true,disksort:false";
```

Creating a Linux Partition as Required by ASMLib

- Oracle ASMLib is an optional host software that offers another method for handling persistent device naming and other features generally included in later releases of Linux.
- Although many DBAs prefer Linux UDEV (8) for device naming, some may still prefer using ASMLib. The below URL covers the differences between Oracle ASMLIB and Oracle ASM Filter Driver (ASMFD): http://docs.oracle.com/database/121/OSTMG/GUID-9C4245E0-279B-4832-A2FA-00E57B34D604.htm#OSTMG95908
- For DBAs who wishing to transition away from ASMLib, My Oracle Support note 1461321.1 provides a step-by-step guide for converting from ASMLib to UDEV (8).
- If ASMLib is required for specific business needs, the following information should be considered.
    - When working with ASMLib, some customers may create partitions. In such a case, the system administrator must decide which utility to use for partitioning. An example with FDISK (8) is provided below.
    - The first addressable sector for each device is sector 0, and each sector is 512 bytes in size. As a general rule, the best practice when partitioning the device is to explicitly assign the starting offset, such as one megabyte. This one megabyte of extra room is reserved by defining the partition to start at sector 2048. The extra room is available for storing the ASMLib header, which serves to minimize the occurrence of ASMLib header corruption.

Note: As recommended, partitioning drives also guarantees that I/O requests will be aligned properly for XtremIO.

Example for using fdisk utility:

1. At the shell prompt, enter:

```
# fdisk –u /dev/mapper/<NAA>
```

2. Enter the following values:

- n – for new
- p – for partition
- 1 – for partition 1
- 2048 – for the starting sector
- Enter – to accept the last sector
- w – to save

3. To Access the recently created partition on the block device:

```
# kpartx -av /dev/mapper/<NAA>
```

The addressable block device partition becomes: `/dev/mapper/<NAA>p1`

4. If /dev/mapper/<NAA>p1 is not displayed, it is necessary to restart multipathd via the service(8):

```
# service multipathd restart
```

5. The following describes an Example for Initializing a LUN for Oracle ASMLib:

```
# oracleasm createdisk DATAOFF4 /dev/mapper/3514f0c5c83a001b9p1
```

In Linux clustering, it is common for hosts to assign different "friendly names" (e.g. mpathX) to share LUNs when the hosts boot up. This is often referred to as "device slip". Device slips can be prevented with UDEV (8). However, since the topic at hand is ASMLib, it should be noted that the `oracleasm-support` package labels disks with cluster-wide unique headers on each device.

Enabling Load Balancing when Using ASMLib

To ensure that DM-MPIO nodes are suitably utilized for load balancing, it is recommended to explicitly modify the ASMLib configuration file. The best practice is to perform the modifications while the existing ASM disk groups are unmounted.

1. Modify the /etc/sysconfig/oracleasm file as below:

```
ORACLEASM_ENABLED=true
# ORACLEASM_UID: The default user owning the /dev/oracleasm mount point
ORACLEASM_UID=oracle
# ORACLEASM_GID: The default group owning the /dev/oracleasm mount point
ORACLEASM_GID=dba
# ORACLEASM_SCANBOOT: When set as “true”, the system scans for ASM disks upon
boot
ORACLEASM_SCANBOOT=true
# ORACLEASM_SCANORDER: Matching patterns to order disk scanning
ORACLEASM_SCANORDER="dm"
# ORACLEASM_SCANEXCLUDE: Matching patterns to exclude disks from scan
ORACLEASM_SCANEXCLUDE="sd"
```

2. Restart ASM Daemon for Changes to take effect.

```
# /etc/init.d/oracleasm stop
# /etc/init.d/oracleasm start
```

512 versus 4K Advanced Format Considerations

The default setting for XtremIO volumes is 512 bytes. It is recommended keep the default setting and not use 4K the Advanced Format for Oracle Database deployments. There are no performance ramifications when using 512B volumes in conjunction with an Oracle database. On the contrary, 4K Advanced Format is rejected by many elements of the Oracle and Linux operating system stack.

Many software components in both Oracle and Linux operating system layers do not function properly with 4K logical sector sizes. One example of Linux operating system functionality which does not work with 4K Advanced Format is the direct I/O (O_DIRECT) support on both EXT4 and XFS file systems.

My Oracle Support Doc ID 1630790.1 and Doc ID 1681266.1 provide more information on the 4K Advanced Format.

Multiblock I/O Request Sizes

Oracle Database performs I/O on data files in multiples of the database block size (db_block_size), which by default is 8KB. The default Oracle Database block size is optimal on XtremIO. XtremIO supports larger block sizes as well. In the case of multiblock I/O (e.g., table/index scans with access method full), one should tune the Oracle Database initialization parameter `db_file_multiblock_read_count` to limit the requests to 128KB. This is derived with the following formula:

`db_file_multiblock_read_count` is: `db_file_multiblock_read_count = 128KB / db_block_size`

Historically, Oracle Database was optimized to perform very large transfers to mitigate the seek cost due to multiblock reads on mechanical drives. In a seek-free environment, such as XtremIO, there is no need for such mitigation. Also, most modern Fiber Channel host bus adapters require Linux to segment large requests into multiple requests. For example, an application I/O request of one megabyte is fragmented by the Linux block I/O layer into two 512KB transfers in order to suit the HBA maximum transfer size.

Redo Log Block Size

The default block size for REDO LOG is 512 bytes. I/O requests sent to the redo log files are in increments of the redo block size. This is the blocking factor Oracle uses within REDO LOG files and has nothing to do with the on-disk format of the XtremIO LUN.

Our recommendation for XtremIO X1 and X2 arrays is to create REDO LOG files with 4K block size. For more details check Oracle Support notes 1681266.1.

Notes: On Oracle versions prior to 12.2.0.1.0, you should set the parameter `_disk_sector_size_override` to `TRUE` when creating a redo log with 4K-block size in the database instance. This issue is fixed in later Oracle versions. For more information see Oracle Doc ID# 1918508.1.

Do not set the parameter `_disk_sector_size_override` in the ASM instance. Once the instance is running, simply add more redo logs with the BLOCKSIZE option set to 4KB and then drop any redo logs that have the default 512B block size.

Grid Infrastructure Files – OCR/Voting

The block size for both Oracle Cluster Registry (OCR) and Cluster Synchronization Services (CSS) voting files are 512 bytes. I/O operations to these file objects are therefore sized as a multiple of 512 bytes. This is consistent with best practice for XtremIO volume creation which also uses 512 byte formatting.

Implementing Oracle Quality of Service (QoS)

The XtremIO Storage Array is an "equal-opportunity" array, servicing all I/O requests from all hosts with simple first-in-first-out fairness. Potentially non-mission-critical applications may utilize a larger share of the array's performance capacity than desired by the administrator. However, host I/O on Linux platforms can be easily managed with the Linux Control Groups.

The following references offer more information regarding implementing QoS at the host level:

http://www.oracle.com/technetwork/articles/servers-storage-admin/resourcecontrollers-linux-1506602.html

The following procedure is an example of implementing Host QoS to Limit Performance of DEV Server.

1. At the shell prompt, enter:

```
mkdir /cgroup/blkio
mount -t cgroup -o blkio none /cgroup/blkio cgcreate -t oracle:dba -a oracle:dba
-g blkio:/iothrottle
```

2. Identify the Device “Major:Minor” Number:

```
[root@ucs3 Scripts]# ls -l /dev/oracleasm/disks/DATA*
brw-rw----. 1 oracle oinstall 253, 6 Nov 20 05:57 /dev/oracleasm/disks/DATA1
brw-rw----. 1 oracle oinstall 253, 7 Nov 20 05:57 /dev/oracleasm/disks/DATA2
brw-rw----. 1 oracle oinstall 253, 8 Nov 20 05:57 /dev/oracleasm/disks/DATA3
brw-rw----. 1 oracle oinstall 253, 9 Nov 20 05:57 /dev/oracleasm/disks/DATA4
```

3. Limit DEV Server Maximum Read IOPS to 10KB:

```
[root@OEL63-1 ~]# echo “253:6 10000” >
/cgroup/blkio/blkio.throttle.read_iops_device
```

4. Optionally limit Source Database Server Maximum Read IOPS to 100KB:

```
[root@OEL63-1 ~]# echo “253:6 100000” >
/cgroup/blkio/blkio.throttle.read_iops_device
```

Simplicity of Operation – Provision Capacity Without Complexity

Capacity is the main concern for provisioning XtremIO storage for Oracle Databases.

XtremIO LUN provisioning and presentation is very simple and can be performed via the XtremIO WebUI or the XMCLI. To provisioning host storage from XtremIO, follow the procedure below:

1. On the XtremIO Storage Array:
   a. Create Volumes.
   b. Create an Initiator Group.\
   c. Map the Volumes to the Initiator Group.
2. On the host:
   a. Perform a "host LUN discovery".

Utilities for Thin Provisioning Space Reclamation

Oracle Automated Storage Management does not trim the space potentially made available by files that were deleted in the disk group. Instead of trimming space, ASM marks the free space as "available for overwrite", causing inaccurate reporting of logical space used by XtremIO.

The ASM Storage Reclamation Utility® (ASRU)[1] inserts deleted files with zero-byte blocks in released space in an ASM disk group.

Executing ASRU in an ASM disk group that has had many deleted files serves to adjust accounting of logical used capacity on XtremIO. If deleted files are not referenced anywhere else, ASRU corrects the reported physical capacity used.

Using an XFS and ext4[2] file system, deleted files are automatically trimmed by specifying the "Discard" mount option. This then propagates to the array. Alternatively, one can forego the "Discard" mount option and perform trim operations out-of-band with the `fstrim(8)` command.

Snapshots Used for Backup-to-Disk

XtremIO Storage Array snapshots are precise point-in-time copies of source volumes which essentially are a collection of pointers referencing the source volume blocks. Therefore, snapshots consume no physical capacity.

Executing snapshots is an extremely rapid and efficient backup-to-disk methodology. This is because snapshots are based completely on metadata operations. Snapshots employ the same benefits that are attributed to the source volumes, including high-level performance, XtremIO Data Protection (XDP), automatic data distribution, global deduplication and thin provisioning.

The space that is saved by creating snapshots is not reflected in the deduplication ratio (as is the case with RMAN Image copies). This is because snapshots are pointer based, so they are not actual duplicated blocks.

The space saved by snapshots is tremendous, especially at the time the snapshot is created. Over time, as source volumes are updated, and snapshots mounted and accessed for writes as needed, the net physical capacity consumed by both source volumes and snapshots grows. For backup purposes, it is imperative to invoke snapshots while the database is in "Backup" mode. This is done in order to create valid image copies on snapshots and to enable rolling-forward the database utilizing logs (such as offline logs and/or online or redo logs) up to the desired System Change Number (SCN). This establishes a consistent point in time or "latest SCN".

Invoking snapshots to roll-forward to a latest SCN establishes the most recent consistent state (from a database perspective). As a precaution, the recommended best practice is to create a backup control file prior to initiating a backup-to-disk process.

[1] ASRU is a trademark or registered trademark of the Oracle Corporation.

[2] The ext4 or "fourth extended filesystem" is a journaling file system for Linux, developed as the successor to ext3.

For recovery purposes, the recommended best practice is to separate data files and logs (both offline and online), hence enabling a recovery from various points-in-time. XtremIO backup-to-disk image creation (snapshots) is a seamless and fast process, and results in no perceived degradation in terms of performance of the source volumes. Freeze and thaw of the source volumes are implicitly performed internally on XtremIO, via SCST during snapshot operations.

The Snapshot Groups[3] feature is supported to ensure that headers within the database files (such as control files, data files, log files and optional application volumes) remain consistent. Multiple snapshots or Snapshot Groups of the source volumes, as well as snapshots of snapshots, are fully supported.

This support enables best practice precautionary steps be taken before attempting actual restores and recoveries, such as performing a mock restore and recovery. Unlike the RMAN "Restore" process, the snapshot process for restoring is very fast.

The best practice is to backup image copies on snapshots to a separate storage or tape.

The following URL link provides comprehensive details for RMAN backup concepts.

http://docs.oracle.com/cd/E11882_01/backup.112/e10642/toc.htm

The following URL link provides comprehensive steps for backing up existing image copy backups with RMAN.

http://docs.oracle.com/cd/E11882_01/backup.112/e10642/rcmbckba.htm#BRADV 89561

Snapshots Used for Manual Continuous Data Protection (CDP)

As the implementation of snapshots is so efficient on the XtremIO Storage Array, the snapshots feature may be used as part of a business continuance strategy or for continuous data protection (CDP). Two options can be used for this strategy:

- A crash-consistent, or "restartable" image

  OR

- A recoverable image

[3] Snapshot Group refers to any snapshot action that is performed on a folder, or on a manually-selected list of volumes.

Crash-Consistent Image

A crash-consistent or restartable image is a point-in-time image of the primary database on disk; i.e. a snapshot.

This option entails taking snapshots and/or Snapshot Groups of the primary database while it is up and operational. The image that is captured is similar to the state of the primary database, once the `shutdown abort` command is issued against it.

During the database restart on the snapshots and/or Snapshot Groups, the database automatically performs a recovery, using the online logs. All committed transactions are included, and all uncommitted transactions are rolled back. The Recovery Point Objective (RPO) is defined per interval. The interval is the scheduled time for snapshot or Snapshot Group creation, which can be set as daily, hourly, or defined in minutes (for example, every 30 minutes).

To perform a restore operation, unmount the disk groups or file systems (if applicable), and unmap all of the source volumes comprising the database (data files plus control file, online logs, archived log destination). Once these actions have been successfully performed, map the corresponding snapshot or Snapshot Group.

To perform a recover operation using SQLPLUS, enter `startup` at the prompt for the primary database snapshot.

Recoverable Image

A recoverable image is an image of the primary database on disk; i.e. a snapshot. This option entails taking snapshots and/or Snapshot Groups of the primary database while the database is in "Backup" mode.

The image should be captured after the `alter database begin backup` command is issued. To avoid excessive logging, the `alter database end backup` command should be executed shortly thereafter.

It is also highly recommended to have a backup file of the control file prior to commencing the backup process, and after completion of the backup process.

The recovery point objective (RPO) is defined per interval, once snapshots and/or Snapshot Groups are created. The interval can be set as daily, hourly, or defined in minutes.

Unlike with the crash-consistent image iteration, data files on the replica can be rolled forward through time. This is performed by using logs up to a consistent point in time, either to the desired SCN or up to the latest SCN (captured in the control file). This means that RPO has a much higher time resolution than is available with scheduled intervals. In this way, not only can an image be recovered via the scheduled intervals, but points in time in-between intervals can also be recovered. This works in conjunction with an Oracle media recovery.

Snapshots for Cloning Primary Databases

Clones of the primary database may be deployed using the methodologies described above.

Oracle provides a utility called "NID" (or "NEWDBID®")[4] which is used to facilitate renaming of the `database ID` and database Name properties automatically, as opposed to having to recreate the control file.

[4] NID and NEWDBID are trademarks or registered trademarks of the Oracle Corporation.

Recovery Manager Image Copies for Backup to Disk

The Oracle Recovery Manager (RMAN) is an Oracle-native tool for backing up, restoring and recovering an Oracle database. The tool is an integral part of Maximum Availability Architecture (MAA) employed by Oracle for making robust database deployments.

RMAN creates backups on disk and backup sets by default, rather than creating image copies. A backup set consists of physical files which can be written to either disk or tape, but the format is native to RMAN only (as opposed to image copies).

Image copies created via the "`BACKUP AS COPY`" command are bit-by-bit copies of database files. Image copies may be directed to disk just like a backup set. The copies are then recorded in the RMAN repository, either via an RMAN catalog or via a control file of the target database (in cases where a catalog does not exist or was not used).

Image copy is the recommended format for backup-to-disk, as the format provides the highest level of space savings on the XtremIO Storage Array. This is done by providing up to a 2:1 deduplication ratio between the source database and the RMAN backup to disk. Using RMAN to clone the primary database also benefits from XtremIO's deduplication feature. In an environment where the primary database, RMAN backup-to-disk (image copies), and primary database clone all reside on the XtremIO Storage Array, the DRR (data reduction ratio) can reach 3:1.

Use Cases for RMAN Image Copies:

- Image copies may be used to restore control files, data files and logs, when primary files are corrupted or are inadvertently deleted.
- Image copies on disk may also be used as point-in-time copies of the actual database files. Thus, the time-consuming restore from the backup location to the actual primary volumes can be avoided. Regardless of whether the image copies reside on ASM or on the file system, RMAN automatically re-directs the pointers to the image copies, updating the control files accordingly.
- Image copies on disk may also be used to create a clone of the primary database on the same host or on another host.
- Image copies may be used to create secondary, backup copies, either to tape media or to another storage device.

References

1. Dell EMC XtremIO Main Page:
   http://www.dellemc.com/en-us/storage/xtremio-all-flash.htm

2. Introduction to EMC XtremIO 6.0

3. Dell EMC XtremIO X2 Specifications:
   http://www.emc.com/collateral/specification-sheet/h16094-xtremio-x2-specification-sheet-ss.pdf

4. Dell EMC XtremIO X2 Datasheet:
   http://www.emc.com/collateral/data-sheet/h16095-xtremio-x2-next-generation-all-flash-array-ds.pdf

5. XtremIO CTO Blog (with product announcements and technology deep dives):
   https://xtremio.me/

6. EMC Host Connectivity:
   https://www.emc.com/collateral/TechnicalDocument/docu5265.pdf

7. Oracle Database Online Documentation 12c Release 1 (12.1):
   https://docs.oracle.com/database/121/index.htm

8. SLOB Overview: https://kevinclosson.net/slob/

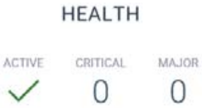
Appendix A – XtremIO Monitoring

XtremIO X2 contains various options to verify proper configuration of the array and troubleshoot performance issues.

The sections below describe several methods for troubleshooting performance related issues.

WebUI

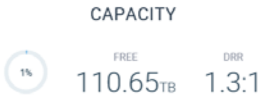
1. Open the WebUI management page: http://<XMS IP>
2. Verify the health of the array.



HEALTH

ACTIVE	CRITICAL	MAJOR
✓	0	0

3. Verify the array capacity is not exceeding 90%.



CAPACITY

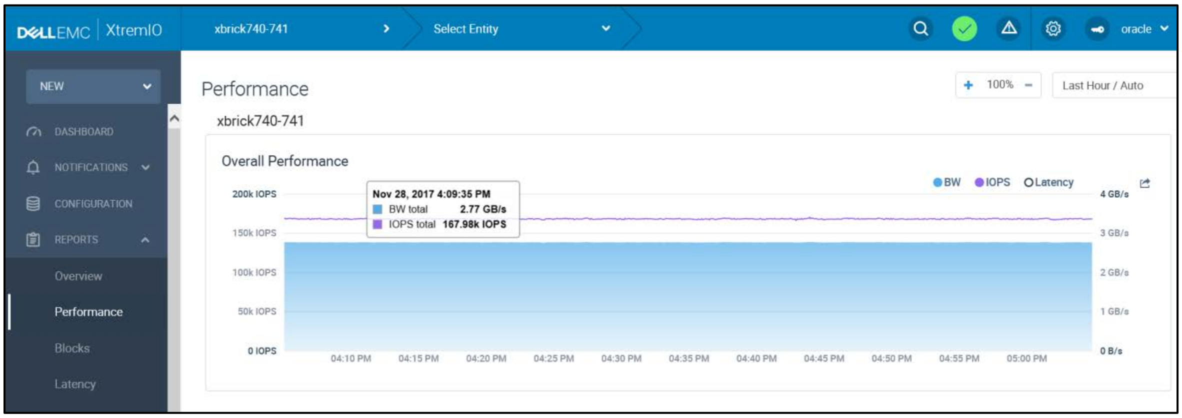
1%

FREE 110.65TB

DRR 1.3:1

4. Access the reports tab and check for the following:

- Performance



Performance

xbrick740-741

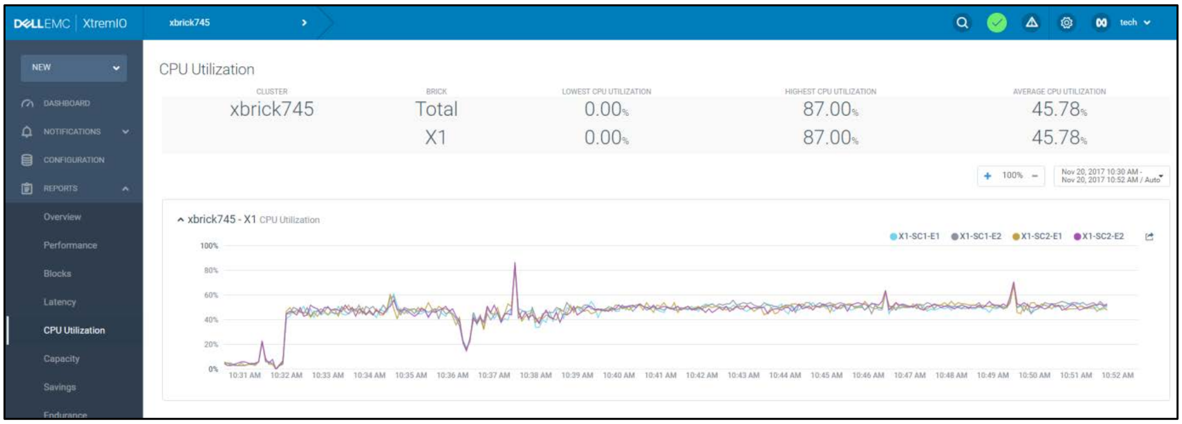
Overall Performance

Nov 28, 2017 4:09:35 PM

BW total 2.77 GB/s

IOPS total 167.98k IOPS

- CPU utilization: If CPU utilization is constantly above 70%, the environment should be fully analyzed.



CPU Utilization

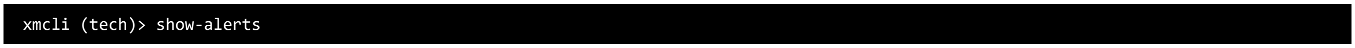
CLUSTER	BRICK	LOWEST CPU UTILIZATION	HIGHEST CPU UTILIZATION	AVERAGE CPU UTILIZATION
xbrick745	Total	0.00%	87.00%	45.78%
	X1	0.00%	87.00%	45.78%

Nov 20, 2017 10:30 AM - Nov 20, 2017 10:52 AM / Auto

xbrick745 - X1 CPU Utilization

XMCLI

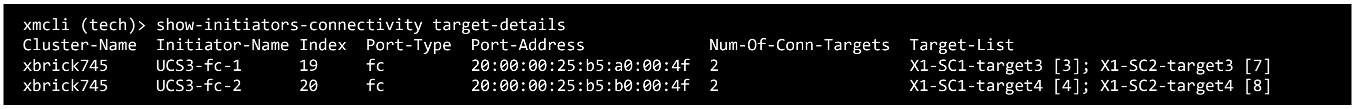
1. Use the following command to check the overall health of the array:



```
xmcli (tech)> show-alerts
```

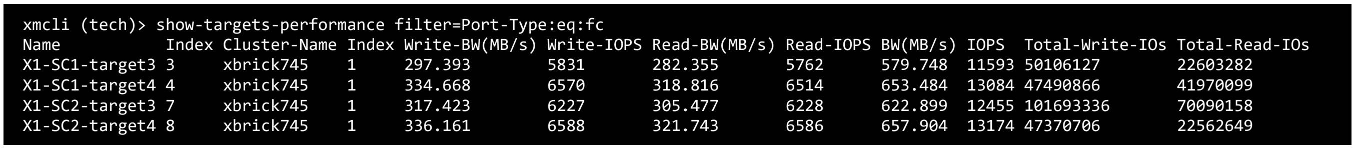
2. The following command shows the initiator's connectivity to FC-target ports.

Note: Verify each initiator is zoned according to XtremIO best practices.



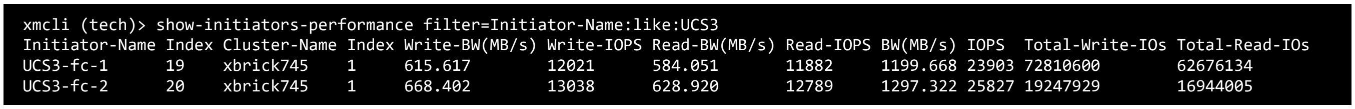
```
xmcli (tech)> show-initiators-connectivity target-details
Cluster-Name  Initiator-Name Index  Port-Type  Port-Address             Num-Of-Conn-Targets  Target-List
xbrick745     UCS3-fc-1      19     fc         20:00:00:25:b5:a0:00:4f  2                    X1-SC1-target3 [3]; X1-SC2-target3 [7]
xbrick745     UCS3-fc-2      20     fc         20:00:00:25:b5:b0:00:4f  2                    X1-SC1-target4 [4]; X1-SC2-target4 [8]
```

3. Use the following command to verify the IOPS/Bandwidth is balanced across the target ports:



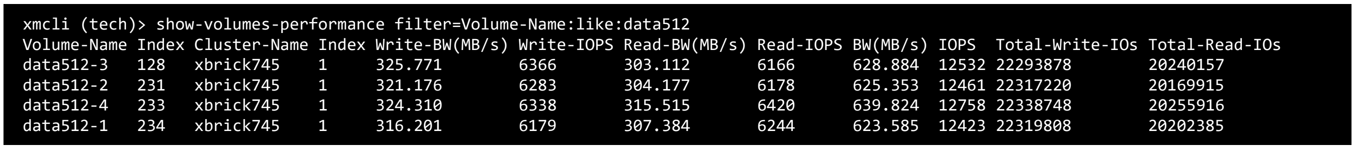
```
xmcli (tech)> show-targets-performance filter=Port-Type:eq:fc
Name           Index Cluster-Name Index Write-BW(MB/s) Write-IOPS Read-BW(MB/s) Read-IOPS BW(MB/s) IOPS  Total-Write-IOs Total-Read-IOs
X1-SC1-target3 3     xbrick745    1     297.393        5831       282.355       5762      579.748  11593 50106127        22603282
X1-SC1-target4 4     xbrick745    1     334.668        6570       318.816       6514      653.484  13084 47490866        41970099
X1-SC2-target3 7     xbrick745    1     317.423        6227       305.477       6228      622.899  12455 101693336       70090158
X1-SC2-target4 8     xbrick745    1     336.161        6588       321.743       6586      657.904  13174 47370706        22562649
```

4. Use the following commands to verify the IOPS/Bandwidth is balanced between the various initiators of the host:



```
xmcli (tech)> show-initiators-performance filter=Initiator-Name:like:UCS3
Initiator-Name Index Cluster-Name Index Write-BW(MB/s) Write-IOPS Read-BW(MB/s) Read-IOPS BW(MB/s) IOPS  Total-Write-IOs Total-Read-IOs
UCS3-fc-1      19    xbrick745    1     615.617        12021      584.051       11882     1199.668 23903 72810600        62676134
UCS3-fc-2      20    xbrick745    1     668.402        13038      628.920       12789     1297.322 25827 19247929        16944005
```

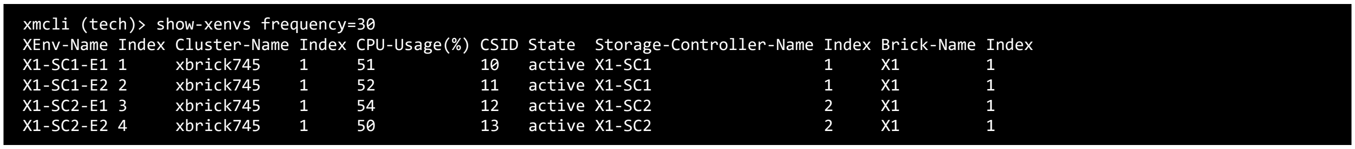
5. The following command shows the comparative performance utilization of a group of application volumes on the XtremIO Storage Array:



```
xmcli (tech)> show-volumes-performance filter=Volume-Name:like:data512
Volume-Name Index Cluster-Name Index Write-BW(MB/s) Write-IOPS Read-BW(MB/s) Read-IOPS BW(MB/s) IOPS  Total-Write-IOs Total-Read-IOs
data512-3   128   xbrick745    1     325.771        6366       303.112       6166      628.884  12532 22293878        20240157
data512-2   231   xbrick745    1     321.176        6283       304.177       6178      625.353  12461 22317220        20169915
data512-4   233   xbrick745    1     324.310        6338       315.515       6420      639.824  12758 22338748        20255916
data512-1   234   xbrick745    1     316.201        6179       307.384       6244      623.585  12423 22319808        20202385
```

6. Use the following command to check internal XtremIO module utilization.

Note: if XENV utilization is constantly above 70%, the environment should be fully analyzed.



```
xmcli (tech)> show-xenvs frequency=30
XEnv-Name Index Cluster-Name Index CPU-Usage(%) CSID State  Storage-Controller-Name Index Brick-Name Index
X1-SC1-E1 1     xbrick745    1     51           10   active X1-SC1                  1     X1         1
X1-SC1-E2 2     xbrick745    1     52           11   active X1-SC1                  1     X1         1
X1-SC2-E1 3     xbrick745    1     54           12   active X1-SC2                  2     X1         1
X1-SC2-E2 4     xbrick745    1     50           13   active X1-SC2                  2     X1         1
```

Appendix B – ASM Disk Group Sector Size – ASMLib Ramifications

512B is the best practice for XtremIO used with Oracle Database.

This section is provided for informational purposes only as it pertains to the use of the 4K Advanced Format which is not recommended for XtremIO used with Oracle Database.

The minimum I/O-transfer size for files in an ASM disk group is determined by the sector size of the underlying physical drive.

Oracle ASM queries devices for the logical sector size of the drive and assigns this value to the `sector_size` disk group attribute (see My Oracle Support note 1938112.1). This is the expected behavior for ASM disks that are not accessed with ASMLib. However, an exception to this behavior was exhibited in early versions of Linux 6.x, with native multi-pathing software (e.g. Device Mapper). In these older Linux versions, the physical sector size was adopted by ASMLib for the ASM disk group instead of the logical sector size.

When using EMC PowerPath, instead of Device-Mapper, Oracle queries the device to verify that the logical-sector size of the LUN is the same as the physical-sector size. Therefore, no work-around is required with ASMLib.

If your business requirements specifically demand the combination of 4K Advanced Format, ASMLib with DM-MPIO, and neither udev(8) control nor EMC Powerpath on XIOS 6.x , refer to My Oracle Support note 1526096.1 for more detailed information.

How to Learn More

For a detailed presentation explaining XtremIO X2 Storage Array's capabilities and how XtremIO X2 substantially improves performance, operational efficiency, ease-of-use and total cost of ownership, please contact XtremIO X2 at XtremIO@emc.com. We will schedule a private briefing in person or via a web meeting. XtremIO X2 provides benefits in many environments and mixed workload consolidations, including virtual server, cloud, virtual desktop, database, analytics and business applications.

Learn more about Dell EMC XtremIO

Contact a Dell EMC Expert

View more resources

Join the conversation @DellEMCStorage and #XtremIO