

# Benchmarking & Performance Optimization for XceedIOPS SSD

---

Application Note AN001

January 2011



## Table of Contents

1.	Introduction	2
2.	Performance States of Solid State Drives	3
3.	Key attributes that influence SSD performance	4
3.1	Chipset & SATA Driver	4
3.2	Partition Alignment & I/O Alignment	5
3.3	Transaction History	5
3.4	Transfer Type and Transfer Size	6
3.5	Data Compression Level and Data Entropy	7
4.	XceedIOPS SSD Benchmarking Test Procedure	8
4.1	System/OS Setup	8
4.2	Partition Alignment	8
4.3	IOMETER Benchmark Software Setup	9
4.4	Test Sequence	9
4.5	Secure Erase/Purge	9
5.	Preconditioning	10
5.1	Preconditioning for Sequential Workload	10
5.2	Preconditioning for Random Workload	10
5.3	Preconditioning for Mixed Random/Sequential Workload	10
5.4	Preconditioning Time	11
6.	Steady State Testing	11
6.1	Definition of Steady State	11
6.2	Random Workload Benchmark	12
6.3	Sequential Workload Benchmark	12
7.	Reporting	13

## 1. Introduction

Designers of enterprise computing applications are starting to take advantage of solid state drives (SSD) in order to increase performance, save space, improve reliability, and reduce power consumption. When compared to traditional 10K or 15K hard disk drives, SSDs are being used to develop systems and applications capable of delivering previously unheard of levels of performance.

But unlike hard disk drive technology, measuring the performance of solid state drives is more complicated due to the internal architecture and storage media used. For example, the performance of an SSD may change as the device is being used and achieving stable and repeatable results requires different test methodologies than those used for traditional HDD devices. Furthermore, performance can be a variable of the data pattern that is written to the drive. Understanding the entropy level (i.e. "randomness") of the data and the related write amplification on the drive is key to predicting the performance for a specific application.

This application note discusses the key parameters that are relevant to the performance of SMART's XceedIOPS SSDs. It provides guidelines and methods for a benchmark test setup required to obtain repeatable, stable and optimal performance results. SMART Modular is an active member in the Solid State Storage Initiative (SSSI) of SNIA, and follows the benchmarking guidelines set forward in the Solid State Storage (SSS) Performance Test Specification (PTS)<sup>1</sup>.

Note: This application note uses an Intel chipset-based motherboard that runs IOMETER software. It should be noted that the performance of the XceedIOPS SSD may vary based on benchmark platform setup, I/O alignment, workload, and data pattern. It is furthermore assumed that remainder of the benchmark infrastructure (servers, networks, etc) is sufficiently robust to force the XceedIOPS SSD to be the bottleneck or performance limiting resource.

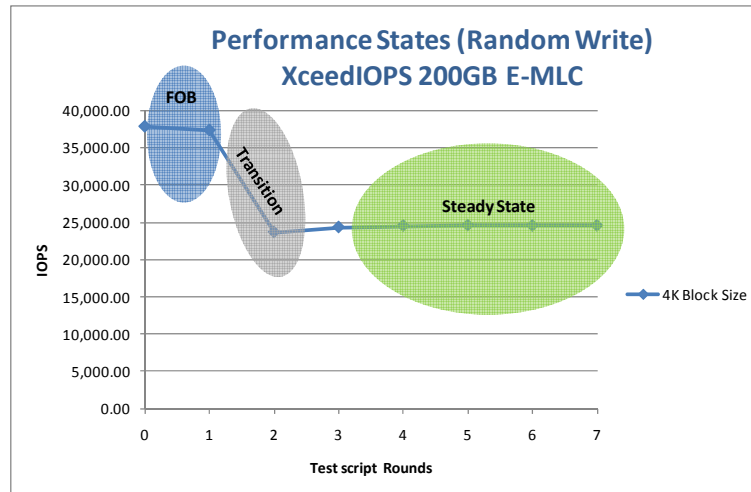


<sup>1</sup> For more information on SSSI, visit: <http://www.snia.org/forums/sssi>

## 2. Performance States of Solid State Drives

SMART's XceedIOPS SSD, similar to any other SSD on the market, will show different performance numbers when it has never been used vs. when it has been used for a while. The transition from "Fresh Out of the Box" (FOB) to steady state is shown in Figure 1 below; in this case the test script writes a 4K random write workload in a repeat loop.

Figure 1: Performance States XceedIOPS SSD



**FOB or "Fresh Out of Box":** This is the condition in which the drive contains no user data. This happens right after the drive is shipped from SMART's manufacturing line, or after a Security Erase command. The NAND flash cells are all in an erased state and are ready to accept new data. As the drive is written and the NAND blocks fills up, the performance of the drive decreases as more and more blocks contain old and invalid data and need to be erased before they can be used again.

**Transition:** Immediately following the FOB state, the drive enters a transition state marked by steadily decreasing performance as more data gets written and the flash management algorithm start to perform background operations. A process called "Garbage Collection" is invoked to accumulate flash blocks that contain old data and erase them to make room for new data. Once the drive has reached steady state performance, the algorithms will ensure that background operations will not show an impact on drive performance.

**Steady State:** This is the condition in which there is relatively small change in performance over a relatively large timeframe. The SNIA Performance Test Specification (PTS) considers a drive to be in steady state when the performance stays within a 20% range of the average performance for at least 5 iterations (or rounds) of the test script, and the slope of the curve is within 10% of the average for the last 5 iterations.

When benchmarking the XceedIOPS SSD, it is important to only take performance measurements once the drive has reached steady state. Section 5 explains in further detail the preconditioning steps required that result in repeatable and consistent results.

### 3. Key attributes that influence SSD performance

Figure 1 below indicates many key attributes that a system designer needs to consider when benchmarking an Enterprise Grade SSD. Let’s examine closely what each key attribute describes and how it can impact SSD performance.

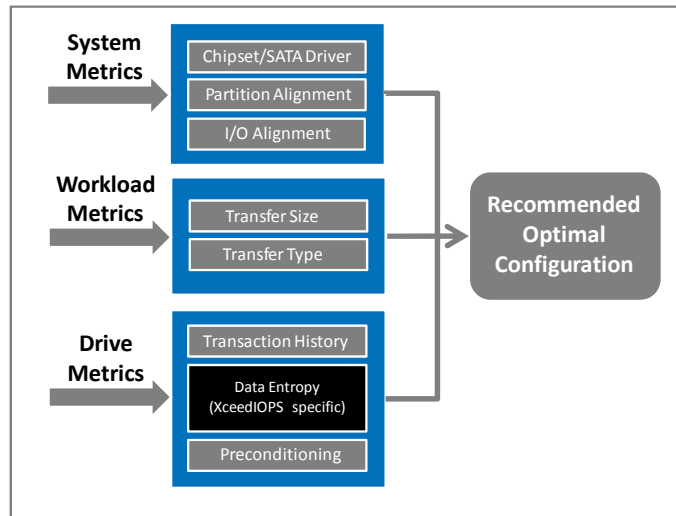


Figure 2: Key attributes that are relevant to SSD performance

#### 3.1 Chipset & SATA Driver

The Intel Southbridge chipset controls all system I/O functions, including the SATA controller. The Intel ICH10R SATA controller, combined with the Intel Matrix SATA Storage driver, delivers higher read/write throughput than many other chipsets, and is used for the benchmark testing described in this application note. The Intel Southbridge chipset ICH10R supports the AHCI (Advanced Host Controller Interface) mode, which enables advanced SATA features, such as NCQ (Native Command Queuing). In multi-threaded software enterprise environments or in virtualized storage architectures, access patterns are highly random. Without NCQ support, read/write commands are executed in a non-optimized order, causing the drive to potentially conduct more random access operations that could be avoided by re-ordering the sequence of pending drive operations. The harder the SSD has to work to shuffle data to make room for new data, the longer it will take for the drive to accept new commands from the application/host.

The NCQ feature helps the SSD build a performance-optimized queue of drive operations in the most efficient order to reduce random access operations, thus improving the drive’s performance and endurance. The XceedIOPS SSD supports up to 32 NCQ commands.

Note: The benchmark test platform, OS, and drivers can have a significant effect on the performance results. It is therefore advised to run performance testing on the same platform(s) when comparing results between different drives, firmware revisions, competitive products, etc.

### 3.2 Partition Alignment & I/O Alignment

NAND flash devices are divided into erasable blocks composed of multiple pages (4KB per page, 256 pages per block for E-MLC flash that XceedIOPS SSD uses). A flash block that contains data must be fully erased prior to writing new data to the block. The erase process for a single block can take up to several milliseconds. Partition alignment between the Operating System (OS) and the NAND flash blocks is critical in terms of obtaining maximum write performance from an SSD. When the partitions are misaligned, write performance typically suffers a great deal because the XceedIOPS SSD controller has to perform unnecessary block erase operations which are referred to as read-modify-write operations. An example of a read-modify-write operation is when the host wants to write 4KB of data (one flash page) to an SSD. The SSD must read an entire block (1MB), update the single page and then write the data for the entire block back.

Windows XP and Windows Server 2000/2003 start partition offset at 31.5KB (32,256 bytes). Due to this misalignment, clusters of data are spread across physical memory block boundaries, incurring a read- modify-write penalty. As a result, the XceedIOPS SSD controller must write up to 200% more data to the flash than is sent from the host to the drive.

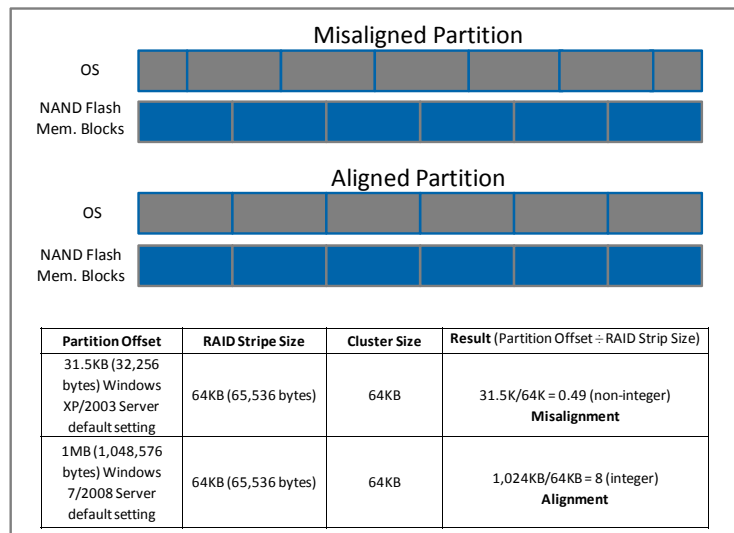


Figure 3: Misaligned Partition vs. Aligned Partition

When choosing a partition starting offset, SMART Modular recommends that system integrators correlate the partition offset with the RAID stripe size and cluster size to achieve optimal SSD I/O performance. Figure 2 above shows an example of a misaligned partition offset and an example of an aligned partition offset for Windows Server.

### 3.3 Transaction History

Beyond the initial transition from FOB to steady state as described in section2, XceedIOPS SSD will optimize its performance to a specific workload, which may take up to 10-12 hours. The transition time can vary, but is typically shorter when changing the workload from random to sequential than vice versa.

Figure 4 below shows that it takes approximately 1.5 hours to transition from a steady state random workload to a steady state sequential write workload. In contrast, Figure 5 shows that it can take up to 10 hours to transition from a steady state sequential workload to a steady state random workload. In order to reduce the transition time to a minimum, SMART Modular recommends starting each benchmark test with a Security Erase/Purge command and a preconditioning stage.

Figure 4: Transition time from 4KB random write to 128KB sequential write workload (IOMeter 2008)

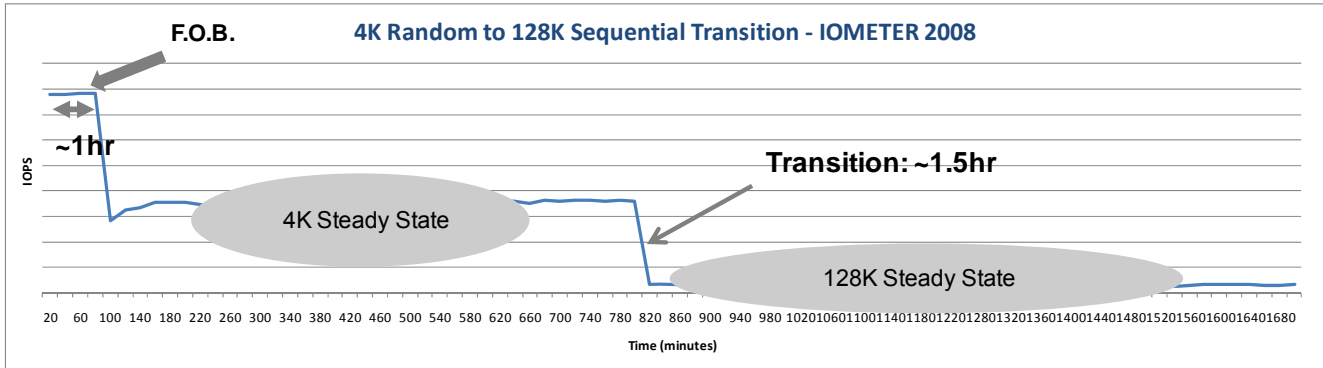
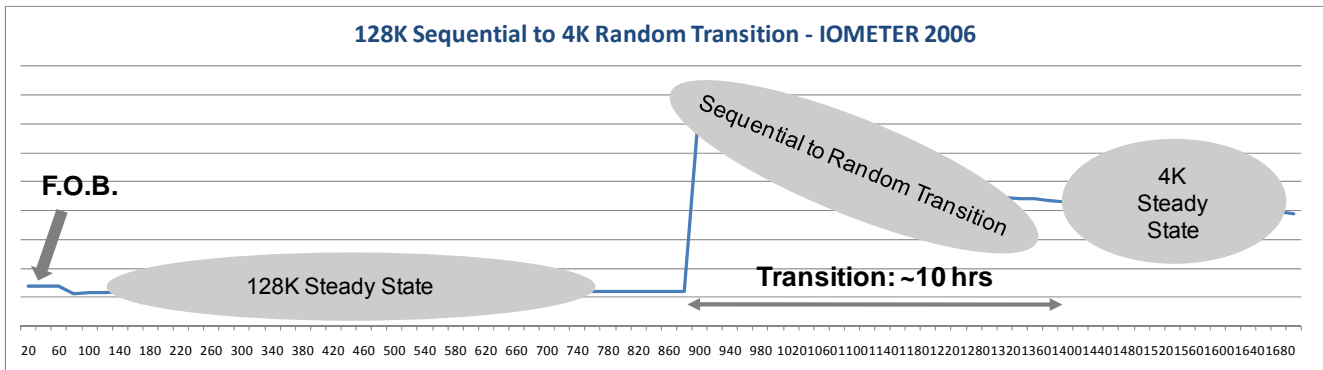


Figure 5: Transition time from 128KB sequential write to 4KB random write workload (IOMeter 2006)



### 3.4 Transfer Type and Transfer Size

The size of the data transfer is another critical performance attribute. SSD performance is optimized when the amount of data transferred to the flash is equal to the NAND Flash page size (4KB for 32nm E-MLC).

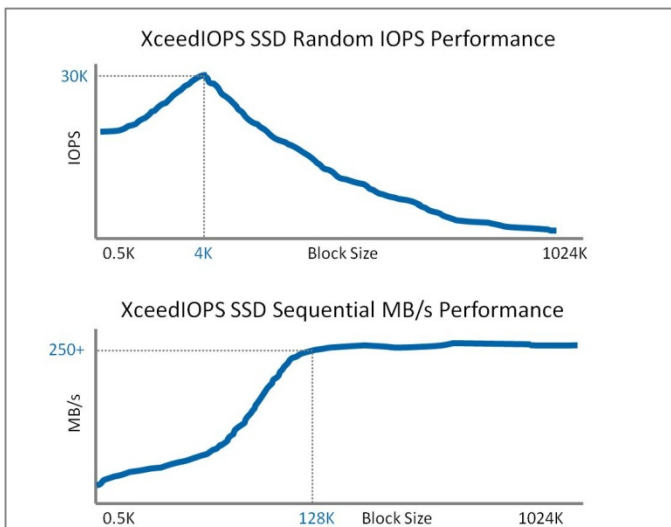


Figure 6: XceedIOPS SSD Random Performance

As shown in Figure 6, XceedIOPS SSDs are optimized for high random write IOPS in 4K block size and high sequential write in 128K block size or larger.

Typical benchmark settings for the transfer size are 4KB for random read/write tests and 128KB for sequential read/write tests.

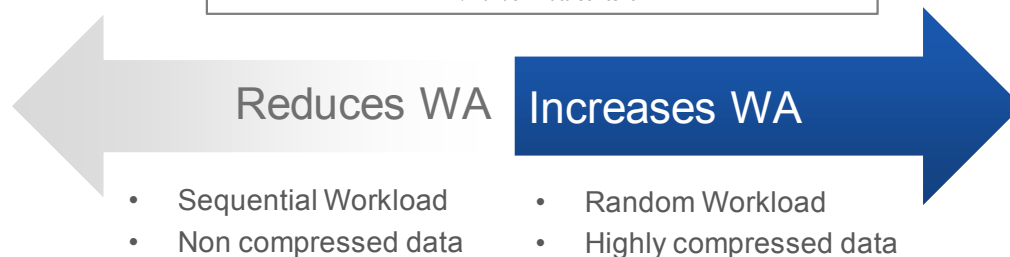
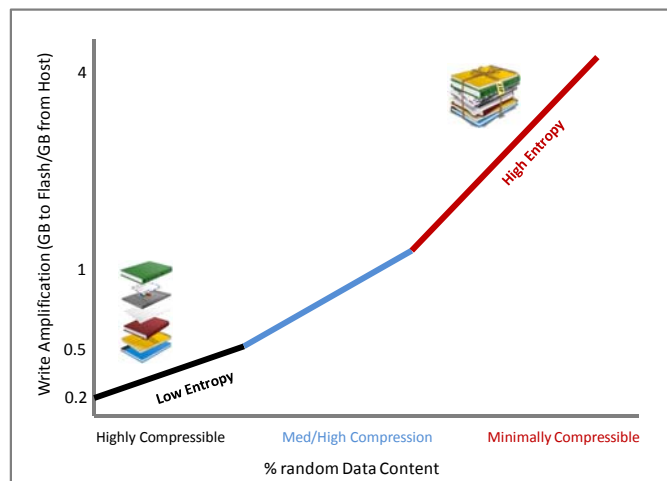
### 3.5 Data Compression Level and Data Entropy

XceedIOPS SSDs employ a data "awareness" hardware engine to both increase performance and decrease write amplification, resulting in better overall drive endurance. Write Amplification is defined as follows:

$$\text{Write Amplification} = \frac{\text{Amount of Data Written to Flash}}{\text{Data Written By Host}}$$

Figure 7 below shows how the data content, or entropy (a measure for data randomness), affects the Write Amplification of XceedIOPS SSD. Typical database log files and text files with no or low efficiency encoding are considered to have low entropy. Write Amplification of the XceedIOPS SSD for low entropy data content ranges from 0.2 - 0.75, depending on the workload (sequential vs. random). Highly efficient encoded file formats such as MPEG-4 video files that are highly compressed are considered to have high entropy. Write Amplification for high entropy data content is higher since not much further compression can be achieved and all data needs to be written to the flash. Write Amplification of the XceedIOPS SSD for high entropy data content ranges from 1.1 to 4, depending on the workload (sequential vs. random).

Figure 7: Write Amplification vs. Data Content (Entropy)



IOMETER 2008 generates lower entropy data patterns than IOMETER 2006. As a result, IOMETER 2008 write performance metrics are higher due to a lower Write Amplification factor. Data patterns generated by IOMETER 2008 are more representative of database applications, whereas IOMETER 2006 represents more video streaming applications.

Note: An open source utility to measure the entropy of a file is available at <http://www.fourmilab.ch/random/>.

## 4. XceedIOPS SSD Benchmarking Test Procedure

SMART Modular, a founding member of the SSSI (Solid State Storage Initiative) of SNIA (Storage Networking Industry Association), has been working with other SSSI members to develop a comprehensive SSS (Solid State Storage) Performance Test Suite Specification<sup>2</sup>. The benchmarking test procedure discussed in this section follows the general guidelines of the SNIA SSS Performance Test Suite Specification. SMART Modular recommends that storage system designers follow the procedures below to benchmark the XceedIOPS SSD.

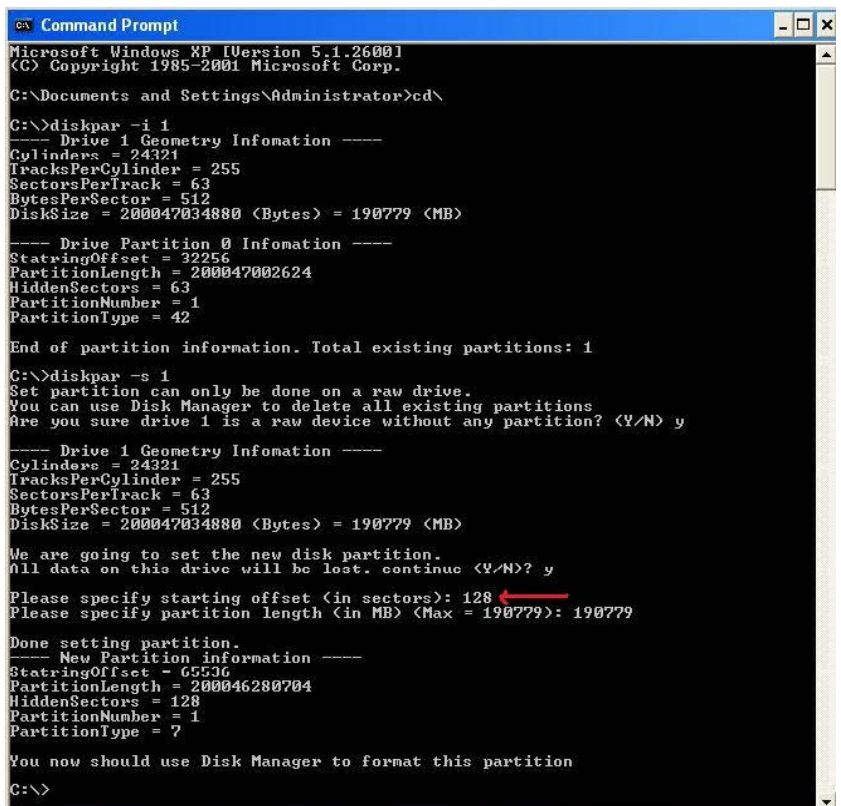
### 4.1 System/OS Setup

SMART Modular recommends that a high performance host system (CPU, memory, I/O, OS) is used to benchmark the XceedIOPS SSD to achieve consistent and maximum results. It is recommended using the following system configuration or a configuration with an equivalent level of system resources:

- CPU: Intel Core i7 920 2.67GHz
- Southbridge Chipset: Intel ICH10R (or similar with AHCI BIOS support)
- Memory: 8GB DDR3 SMART Modular
- OS: Windows 64-bit 7 Home Edition
- Intel AHCI Driver: Intel matrix Storage Manager ver. V9.6.0.1014 (<http://downloadcenter.intel.com/>)
- BIOS: Award Modular v6.00PG
- BIOS Setting: AHCI mode – NCQ enable

### 4.2 Partition Alignment

Partitions created on earlier versions of Windows XP are misaligned. New partitions on Windows Vista, Windows 7, and Windows Server 2008 are properly aligned. Figure 8 shows how to set the partition offset using Microsoft's Diskpar program. Diskpar can be downloaded at (<https://kb.wisc.edu/images/group14/4556/diskpar.exe>). SMART Modular recommends setting the offset to 128 on Windows XP.



```

C:\Documents and Settings\Administrator>cd\

C:\>diskpar -i 1
----- Drive 1 Geometry Information -----
Cylinders = 24321
TracksPerCylinder = 255
SectorsPerTrack = 63
BytesPerSector = 512
DiskSize = 200047034880 (Bytes) = 190779 (MB)

----- Drive Partition 0 Information -----
StartingOffset = 32256
PartitionLength = 200047002624
HiddenSectors = 63
PartitionNumber = 1
PartitionType = 42

End of partition information. Total existing partitions: 1

C:\>diskpar -s 1
Set partition can only be done on a raw drive.
You can use Disk Manager to delete all existing partitions
Are you sure drive 1 is a raw device without any partition? (Y/N) y

----- Drive 1 Geometry Information -----
Cylinders = 24321
TracksPerCylinder = 255
SectorsPerTrack = 63
BytesPerSector = 512
DiskSize = 200047034880 (Bytes) = 190779 (MB)

We are going to set the new disk partition.
All data on this drive will be lost. continue (Y/N)? y

Please specify starting offset (in sectors): 128
Please specify partition length (in MB) (Max = 190779): 190779

Done setting partition.
----- New Partition information -----
StartingOffset = 65536
PartitionLength = 200046280704
HiddenSectors = 128
PartitionNumber = 1
PartitionType = 7

You now should use Disk Manager to format this partition

C:\>
  
```

Figure 8: Using Microsoft's DiskPar to align partition on Windows XP

<sup>2</sup> SSSI is expected to publish the SSS Performance Test Suite in early 2011

### 4.3 IOMETER Benchmark Software Setup

- IOMETER 2008 Windows version 06-22-rc2 (<http://sourceforge.net/projects/iometer/>)
- IOMETER 2006 Windows version 06.27 (<http://www.iometer.org/doc/downloads.html>)
- # of Outstanding I/Os per Target: 32
- # of Workers: 1
- Align I/Os on: data transfer size (i.e. 4K, 128KB)

Note: Running a single worker with outstanding I/O setting at 32 yields the same results as running multiple workers

### 4.4 Test Sequence

In order to create consistent and repeatable performance test results, it is recommended to follow the below test sequence:

1. Secure Erase/Purge the XceedIOPS SSD to bring it back to the initial FOB state (see section 4.5)
2. Properly align the partition offset (see section 4.2)
3. Align I/O on flash page boundary (see section 3.2)
4. Precondition the drive until it reaches a Steady State (see section 5)
5. Measure the performance and create a report (see section 7)

### 4.5 Secure Erase/Purge

Prior to preconditioning, SMART Modular recommends that the benchmark tests start with an XceedIOPS SSD that is in a state that is as close as possible to the initial fresh-out-of-box (FOB) state. The security erase/purge procedure can be done through the *ATA Security Erase* commands for XceedIOPS SATA or through the *FORMAT Unit* command for XceedIOPS SAS SSD. Please refer to the XceedIOPS SATA/SAS SSD product specifications for more details on these commands.

Note: It is also possible to use the following freeware security erase utility, available from <http://cmrr.ucsd.edu/people/Hughes/SecureErase.shtml>. This is a DOS utility.

## 5. Preconditioning

In order to reach a steady state performance as fast as possible, preconditioning for a random write workload will be different than preconditioning for a sequential write workload. The following sections describe the steps that are required for preconditioning for different types of workloads.

Note: SMART can provide IOMeter preconditioning and workload test scripts upon request.

### 5.1 Preconditioning for Sequential Workload

If the performance benchmark test will focus solely on sequential read/write performance, steady state will be achieved as quick as possible by preconditioning the drive with a sequential 100% write pattern, 128KB block size. The following IOMETER configuration is recommended for this type of preconditioning:

- # of Outstanding I/Os per Target: 32
- # of Workers: 1
- Select Transfer Type: 128KB, 100% sequential write
- Set I/O alignment on 128KB
- Run the IOMETER preconditioning test script until Steady State is reached

### 5.2 Preconditioning for Random Workload

When the performance benchmark test will focus on random read/write workload, it is recommended that the drive will first be preconditioned with a sequential workload (as described in section 5.1), followed by a preconditioning step with a random 100% write pattern, 4KB block size. The following IOMETER configuration is recommended for this type of preconditioning:

- # of Outstanding I/Os per Target: 32
- # of Workers: 1
- Select Transfer Type: 4KB, 100% random write
- Set I/O alignment on 4KB
- Run the IOMETER preconditioning test script until Steady State is reached

### 5.3 Preconditioning for Mixed Random/Sequential Workload

Preconditioning for a test workload that will include a mix of random and sequential workloads would have to follow the same preconditioning steps as for a random workload, described in section 5.2.

Once the preconditioning step is finished and the test will start, it is highly recommended to start with the sequential workload before transitioning to the random workload. This, because it takes time for the drive to transition from one workload to the next (see section 3.3). Additionally, it would be recommended to determine whether the random write performance has reached a steady state. It might be required to do a longer preconditioning or “warm up” period in between the two workloads, in order to ensure that the performance is not measured during a transition period.

## 5.4 Preconditioning Time

The time it takes to reach steady state depends on the capacity of the drive, as well as the benchmark test that is used. Table 1 below shows some recommended preconditioning times for the various capacities XceedIOPS SSDs.

Table 1: Time required for Preconditioning XceedIOPS SSD

Drive Capacity	IOMeter 2006				IOMeter 2008			
	50GB	100GB	200GB	400GB	50GB	100GB	200GB	400GB
Sequential Write Preconditioning 128KB	10 min	20 min	40 min	80 min	45 min	90 min	180 min	360 min
Random Write Preconditioning 4KB	30 min	60 min	120 min	240 min	45 min	90 min	180 min	360 min
Full Preconditioning Duration	40 min	80 min	160 min	320 min	90 min	180 min	360 min	720 min

Note: For a benchmark that only measures sequential write performance, no random write preconditioning is required.

## 6. Steady State Testing

### 6.1 Definition of Steady State

Steady state is reached when a relatively stable performance state for the workload being applied is reached. According to the SNIA PTP, the last 5 rounds of the test script (i.e. measurement window, see Figure 9) need to show similar test results before it can be concluded that the drive is in steady state. To be more specific, Steady State is achieved if the following conditions are met:

- Variation of y in measurement window is within 20% of average (see Figure 10)
- Trending of y within measurement window is within 10% of average (see Figure 10)

Figure 9: Measurement Window

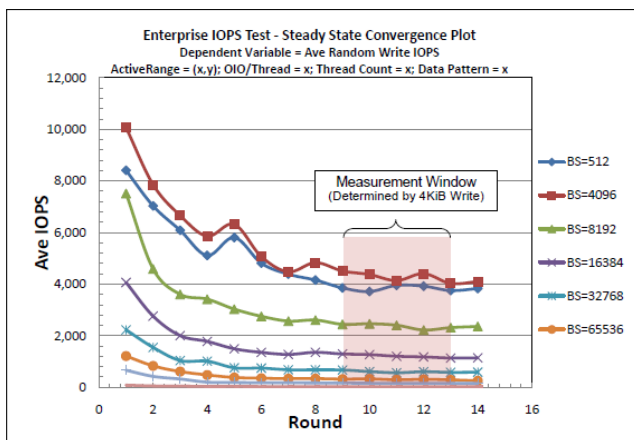
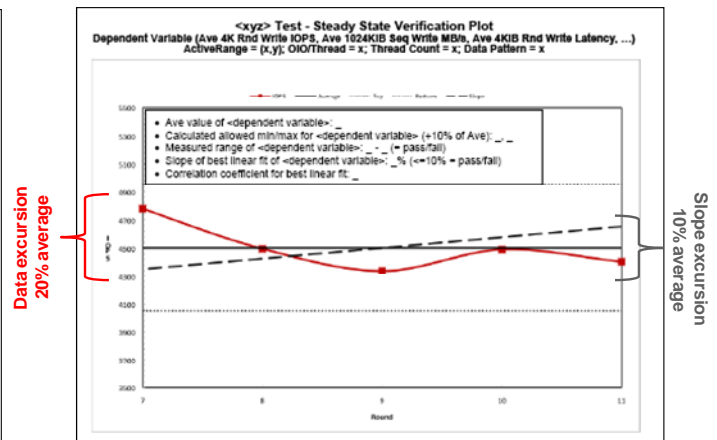


Figure 10: Trending and variation of y



## 6.2 Random Workload Benchmark

In order to understand the performance of an XceedIOPS SSD across all block sizes and read/write mixes, the following steps are recommended to run a random I/O test benchmark:

1. Secure Erase XceedIOPS SSD (see section 4.5)
2. Preconditioning for sequential workload (see section 5.1)
3. Precondition for random workload (see section 5.2)
4. Run a script that creates the following workloads:

Transfer Size (KB)	Read/Write Mix %						
	0/100	20/80	35/65	50/50	65/35	80/20	100/0
0.5	•	•	•	•	•	•	•
1	•	•	•	•	•	•	•
2	•	•	•	•	•	•	•
4	•	•	•	•	•	•	•
8	•	•	•	•	•	•	•
16	•	•	•	•	•	•	•
32	•	•	•	•	•	•	•
64	•	•	•	•	•	•	•
128	•	•	•	•	•	•	•
256	•	•	•	•	•	•	•
512	•	•	•	•	•	•	•
1,024	•	•	•	•	•	•	•

5. Set Run Time: 30 minutes for each transfer size
6. Run the script up until drive is in steady state (see section 6.1)
7. Take measurements

Note: When constrained for time, SMART Modular recommends a minimum of the three highlighted workloads above (0/100, 65/35, and 100/0) to be tested. For a script that runs longer than 5hrs, it is expected that steady state will be achieved around the 5hr point.

Note: The order in which these tests are run can change the results (start left or right; start top or bottom). To delete the dependency of the order of tests, a purge would be required between each test run.

Note: If both IOMETER 2006 and IOMETER 2008 will be used to benchmark the SSD, SMART Modular recommends benchmarking with IOMETER 2008 prior to IOMETER 2006, since IOMETER 2008 generates lower entropy data patterns.

## 6.3 Sequential Workload Benchmark

In order to understand the performance of an XceedIOPS SSD across all block sizes and read/write mixes, the following steps are recommended to run a random I/O test benchmark:

1. Secure Erase XceedIOPS SSD (see section 4.5)
2. Preconditioning for sequential workload (see section 5.1)
3. Run a script that creates the following workloads:

Transfer Size (KB)	Read/Write Mix %						
	0/100	20/80	35/65	50/50	65/35	80/20	100/0
0.5	●	●	●	●	●	●	●
1	●	●	●	●	●	●	●
2	●	●	●	●	●	●	●
4	●	●	●	●	●	●	●
8	●	●	●	●	●	●	●
16	●	●	●	●	●	●	●
32	●	●	●	●	●	●	●
64	●	●	●	●	●	●	●
128	●	●	●	●	●	●	●
256	●	●	●	●	●	●	●
512	●	●	●	●	●	●	●
1,024	●	●	●	●	●	●	●

8. Set Run Time: 30 minutes for each transfer size
9. Run the script up until drive is in steady state (see section 6.1)
10. Take measurements

Note: When constrained for time, SMART Modular recommends a minimum of the three highlighted workloads above (0/100, 65/35, and 100/0) to be tested. For a script that runs longer than 5hrs, it is expected that steady state will be achieved around the 5hr point.

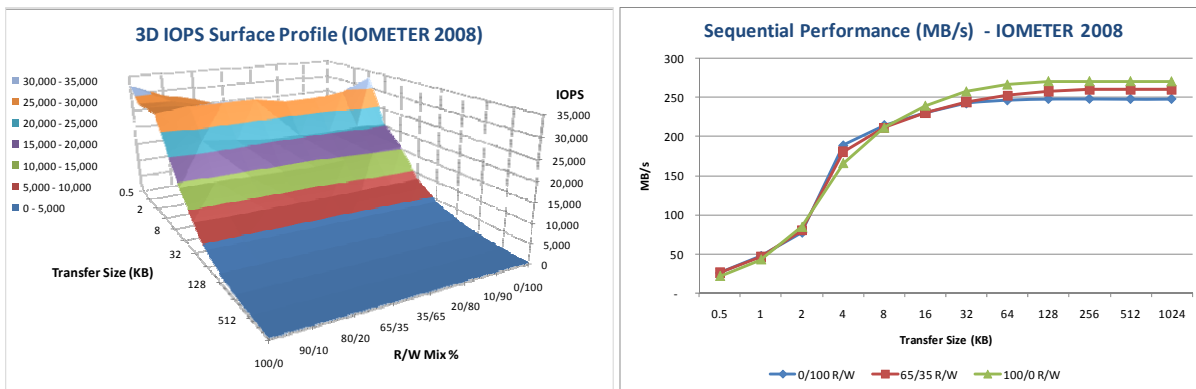
Note: The order in which these tests are run can change the results (start left or right; start top or bottom). To delete the dependency of the order of tests, a purge would be required between each test run.

Note: If both IOMETER 2006 and IOMETER 2008 will be used to benchmark the SSD, SMART Modular recommends benchmarking with IOMETER 2008 prior to IOMETER 2006, since IOMETER 2008 generates lower entropy data patterns.

## 7. Reporting

Once benchmark testing is finalized, it is recommended to present the results in graphic form. It is common practice to display random performance in a 3D IOPS Profile and sequential performance in a 2D format, with only 2 or 3 read/write mixes as data points (see Figure 11).

Figure 11: Performance Graphs



Note: Performance test results for XceedIOPS SATA and SAS SSDs are available upon request.

**Disclaimer:**

No part of this document may be copied or reproduced in any form or by any means, or transferred to any third party, without the prior written consent of an authorized representative of SMART Modular Technologies, Inc. ("SMART"). The information in this document is subject to change without notice. SMART assumes no responsibility for any errors or omissions that may appear in this document, and disclaims responsibility for any consequences resulting from the use of the information set forth herein. SMART makes no commitments to update or to keep current information contained in this document. The products listed in this document are not suitable for use in applications such as, but not limited to, aircraft control systems, aerospace equipment, submarine cables, nuclear reactor control systems and life support systems. Moreover, SMART does not recommend or approve the use of any of its products in life support devices or systems or in any application where failure could result in injury or death. If a customer wishes to use SMART products in applications not intended by SMART, said customer must contact an authorized SMART representative to determine SMART's willingness to support a given application. The information set forth in this document does not convey any license under the copyrights, patent rights, trademarks or other intellectual property rights claimed and owned by SMART. The information set forth in this document is considered to be "Proprietary" and "Confidential" property owned by SMART.

ALL PRODUCTS SOLD BY SMART ARE COVERED BY THE PROVISIONS APPEARING IN SMART'S TERMS AND CONDITIONS OF SALE ONLY, INCLUDING THE LIMITATIONS OF LIABILITY, WARRANTY AND INFRINGEMENT PROVISIONS. SMART MAKES NO WARRANTIES OF ANY KIND, EXPRESS, STATUTORY, IMPLIED OR OTHERWISE, REGARDING INFORMATION SET FORTH HEREIN OR REGARDING THE FREEDOM OF THE DESCRIBED PRODUCTS FROM INTELLECTUAL PROPERTY INFRINGEMENT, AND EXPRESSLY DISCLAIMS ANY SUCH WARRANTIES INCLUDING WITHOUT LIMITATION ANY EXPRESS, STATUTORY OR IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

©2010 SMART Modular Technologies, Inc. All rights reserved.

**Corporate Headquarters:** 39870 Eureka Dr., Newark, CA 94560, USA ♦ Tel: (510) 623-1231 ♦ Fax: (510) 623-1434 ♦ E-mail: info@smartm.com

**Flash Design Center:** Three Highwood Dr., Ste. 103E, Tewksbury, MA 08176, USA ♦ Tel: (978) 805-2100 ♦ Fax: (978) 805-2357

**Asia:** Plot 18, Lrg Jelawat 4, Kawasan Perindustrian Seberang Jaya 13700, Prai, Penang, Malaysia ♦ Tel: +604-3992909 ♦ Fax: +604-3992903