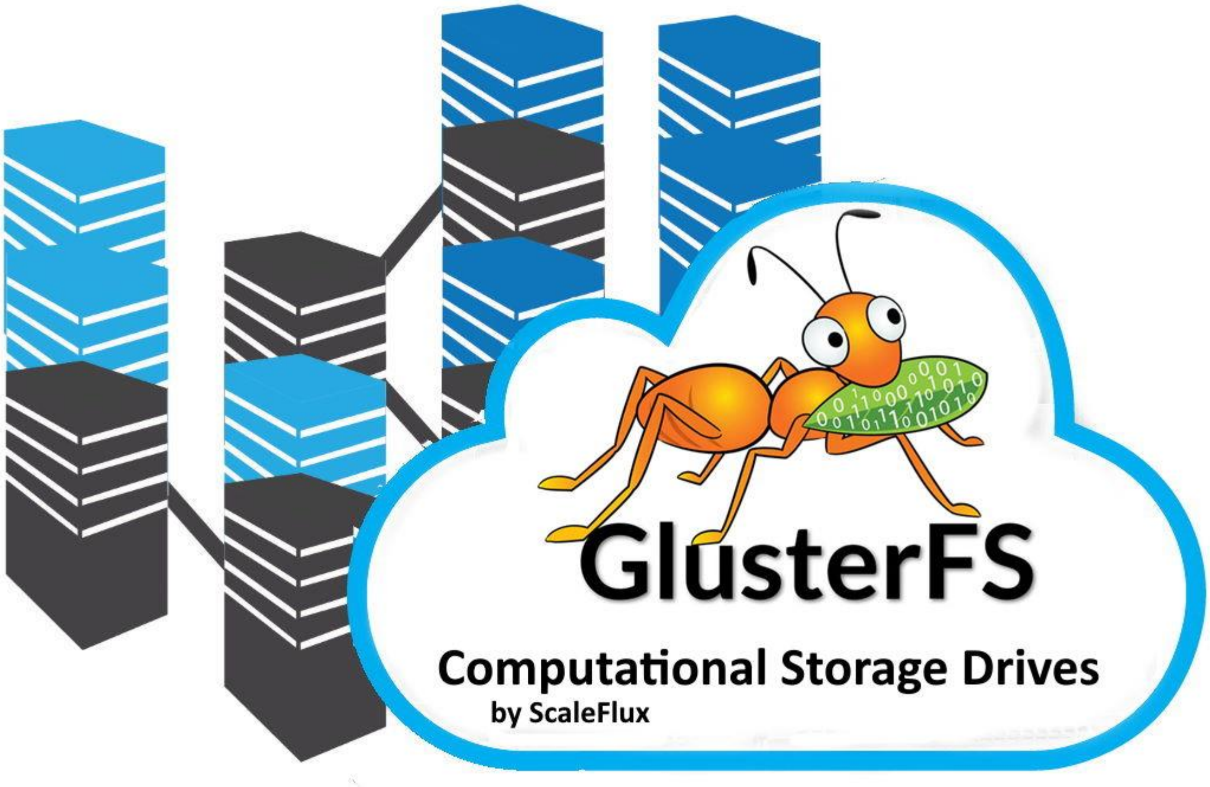
Gluster Scale Out Storage for Cloud using Computational Storage Drives CSD

By HYPER SCALERS



Tuesday, 19 July 2022

1 CONTENTS

2 List of Figures 3
3 Introduction 4
Audience and Purpose 4
Documents, Knowledge Base, and Technical Support 4
Digital IP Appliance Design Process 4
Appliance Optimizer Utility AOU 4
Important Considerations 6
BIOS is tuned to "High Performance" power and profile configuration. 6
CPUs are tuned to disable idling state 7
Proper ventilation and standard datacentre operating temperature 7
Native Sector Size 8
Memory requirements 8
Verify Supported Interfaces 9
Infrastructure Setup 9
Qualification Hardware with the CSD 9
Driver Install 10
Determine the Linux System Distribution 10
Determining the Kernel Version 10
Installing Binary Drivers 11
Installing Source Drivers 11
Validation 11
4 Base Product Deployment 13
Preinstallation Requirements 13
Modifying Drive Capacity 13
Logical Volume target for the Gluster File System 14
GlusterFS setup and configuration 16
5 Testing the Appliance 17
Extended Capacity Gluster exposed pool testing 17
6 Additional Setup and Deployment 18
Capacity Monitoring 18
Temperature Monitoring 19
Drive Management 20
Changing the Logical Sector Size 20
Enabling Atomic Writes 20
Data Wipe (Factory Reset) 21

7 Addendum........22

Test Suite on single drive........22

2:1 Compressed Individual drive test........22

2:1 Compressed drive LV test........24

Performance chart........26

Performance Comparison........26

8 Trademarks and Licensing........26

9 References........27

2 LIST OF FIGURES

Figure 1 Digital IP-Appliance Design Process........5
Figure 2 Socket performance configuration........6
Figure 3 Core frequency utilization........7
Figure 4 BMC sensor temperature informations........8
Figure 5 GlusterFS system architecture........9
Figure 6 ScaleFlux CSD status........12
Figure 7 ScaleFlux CSD status with increased drive capacity........14
Figure 8 KVPM graphical utility........15
Figure 9 Creating a volume group........15
Figure 10 Creating a logical volume with the grouped drives........16
Figure 11 IOMeter sequential read on GlusterFS target........17
Figure 12 IOMeter sequential write on GlusterFS targetss........18
Figure 13 ScaleFlux drive temperature status........19
Figure 14 IOPs and throughput with data compression and drive capacity........22
Figure 15 FIO sequential read test on individual drive........23
Figure 16 FIO sequential write test on individual drive........23
Figure 17 FIO random read test on individual drive........24
Figure 18 FIO random write test on individual drive........24
Figure 19 IOMeter Sequential read on two drive logical volume........25
Figure 20 IOMeter Sequential write on two drive logical volume........25
Figure 21 performance vs drive expansion........26
Figure 22 combined IOPs performance of 3:1 compressed drive comparison........26

3 INTRODUCTION

Hyperscalers has developed a storage solution based on Gluster FS and Computational Storage Drives by ScaleFlux. The result: The ability to expand capacity by 150%, double the read and write performance and halve the cost per TB.

Gluster provides an open-source platform for scale-out public and private cloud storage. When delivered as part an appliance by Hyperscalers, service providers are able to realise newfound benefits around capacity, performance and cost based on pre-engineered outcomes that can be delivered quickly and reliably.

Audience and Purpose

The document is for Engineers, Technical Sales personnel, IT Operations, ScaleFlux and HyperScalers staff.

The purpose of this technical document is to describe the deployment and testing of GlusterFS network filesystem on a 1U rack mounted server (S5B-D52B) with each ScaleFlux CSD 2000 expanded from 6.4TB to 9.6TB.

Documents, Knowledge Base, and Technical Support

Hyperscalers provides the reference architecture and a single point of demonstration to the high-performance, scalable, distributed, network filesystem using GlusterFS and ScaleFlux CSD2000 incorporating the drive expansion features.[1] https://www.hyperscalers.com/OCP-hyperscale-rack-solutions

The official documentation to access ScaleFlux CSD 2000 product brief and description is found in https://www.scaleflux.com/product/item/1002.[2]

Gluster can be used in different ways, it is supported by open community. The administrative guide and GitHub repository can server as the point-of-contact for understanding GlusterFS.[3]

Digital IP Appliance Design Process

Hyperscalers has developed a Digital- IP-Appliance Design Process and associated Appliance Optimizer Utility which can enable the productization of IT-appliances for Digital-IP owners needing to hyperscale their services very quickly, reliably and at a fraction of traditional costs.

Appliance Optimizer Utility AOU

The Appliance Optimizer Utility (AOU) automates the discovery of appliance bottlenecks by pinging all layers in the proposed solution stack. A live dashboard unifies all key performance characteristics to provide a head-to-head performance assessment between all data-path layers in the appliance, as well as a comparison between holistic appliances.

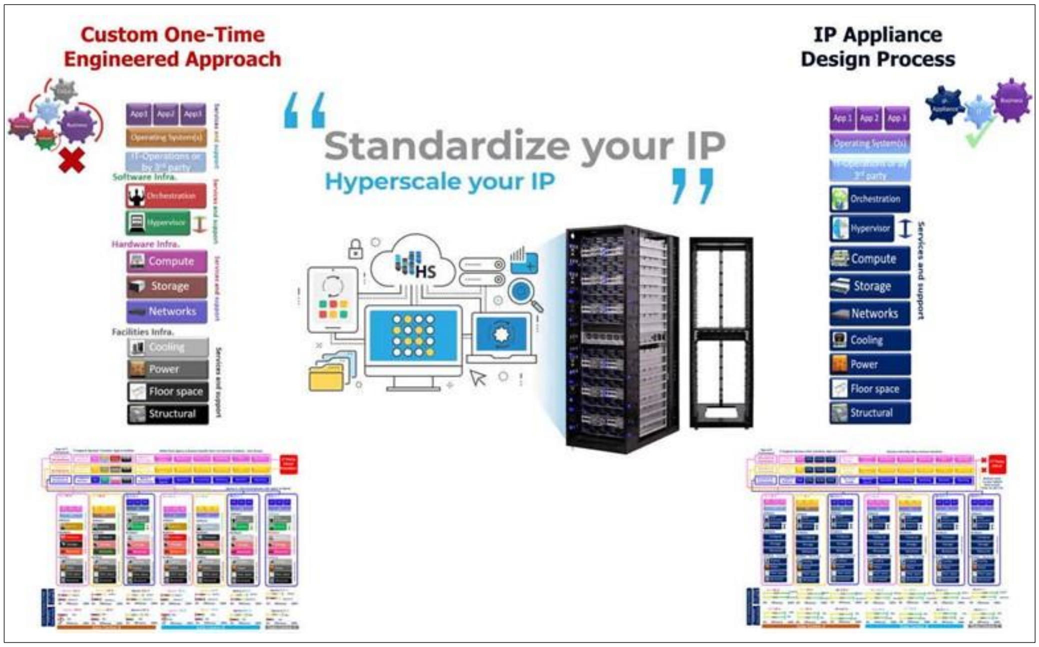


Figure 1 Digital IP-Appliance Design Process

Knowledge on basic Linux, distributed storage systems and networking are the pre-requisites to understand certain terminologies in this document.

The CSD 2000 drive features include:

Form Factors	Add in PCIe AIC & Hot-swappable 2.5" U.2
Interface	PCIe Gen3 x4 Low latency block storage device
NAND Media	3D TLC & 3D QLC
Drive Capacity	Up to 16TB Effective Capacity with data path compression (4/8TB raw)
Power Loss Data Protection	Yes
Compression Engine	Transparent GZIP Compression / Decompression Engine
Data Protection	End-to-End Protection ECC on all memories Full data-path CRC LDPC and die-level RAID protection
Power	18Watt Typical Active 25Watt Maximum 12Watt Idle (Zero Exit Latency)

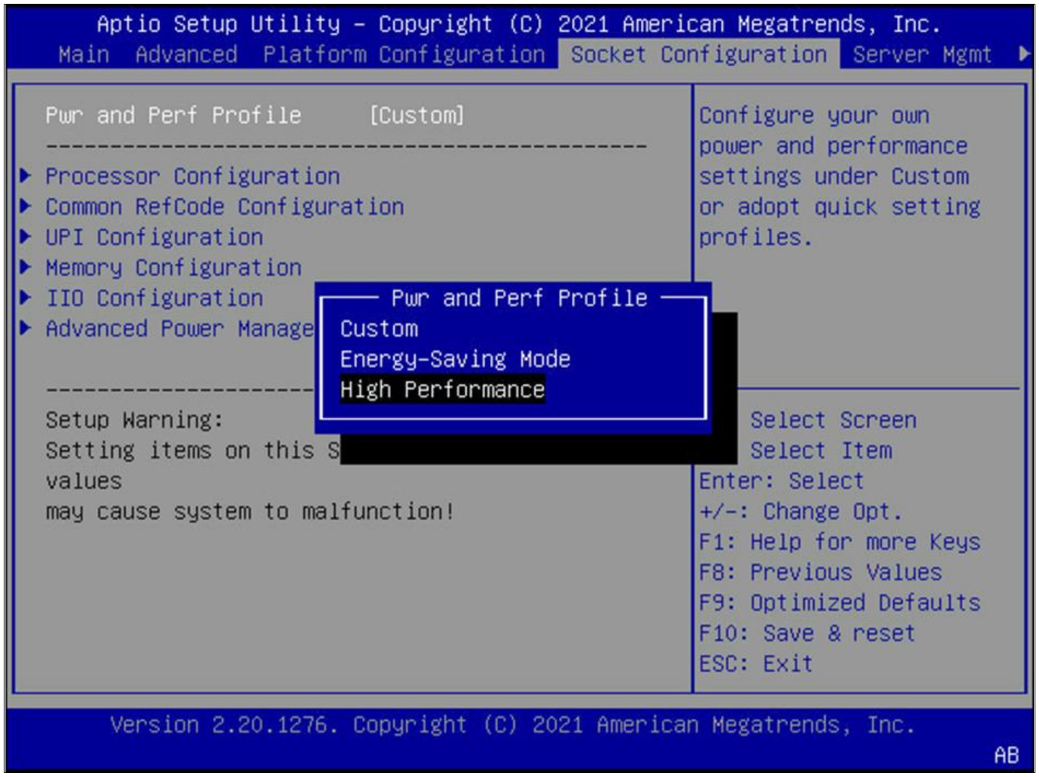
Operating Temperature	50°C @ 200LFM (AIC) 35°C @ 200LFM (U.2)
Temperature Protection	Thermal Throttling Enabled
MTTF	2 million Hours
Compute Capability	Transparent Datapath Compression Accelerated Performance Extended Capacity Adjustable drive settings
Software Compatibility	Linux OS 2.6 Kernel or later Only Repository Support: Ubuntu 16/18/20, RedHat/CentOS 6/7/8

Important Considerations

Hardware pre-requisite for optimal drive utilisation of the CSD 2000 drives to reach the vendor specified benchmark performance are listed below.

BIOS is tuned to "High Performance" power and profile configuration.

Navigate to BIOS -> Socket Configuration -> Pwr and Perf Profile -> choose High Performance



```
        Aptio Setup Utility – Copyright (C) 2021 American Megatrends, Inc.
     Main  Advanced  Platform Configuration  Socket Configuration  Server Mgmt ▶

   Pwr and Perf Profile      [Custom]                    Configure your own
   ----------------------------------------------        power and performance
 ▶ Processor Configuration                               settings under Custom
 ▶ Common RefCode Configuration                          or adopt quick setting
 ▶ UPI Configuration                                     profiles.
 ▶ Memory Configuration
 ▶ IIO Configuration        ──── Pwr and Perf Profile ────
 ▶ Advanced Power Manage     Custom
                             Energy-Saving Mode
   ---------------------     High Performance
   Setup Warning:                                            Select Screen
   Setting items on this S                                   Select Item
   values                                                Enter: Select
   may cause system to malfunction!                      +/-: Change Opt.
                                                         F1: Help for more Keys
                                                         F8: Previous Values
                                                         F9: Optimized Defaults
                                                         F10: Save & reset
                                                         ESC: Exit

          Version 2.20.1276. Copyright (C) 2021 American Megatrends, Inc.
                                                                           AB
```

Figure 2 Socket performance configuration

CPUs are tuned to disable idling state.

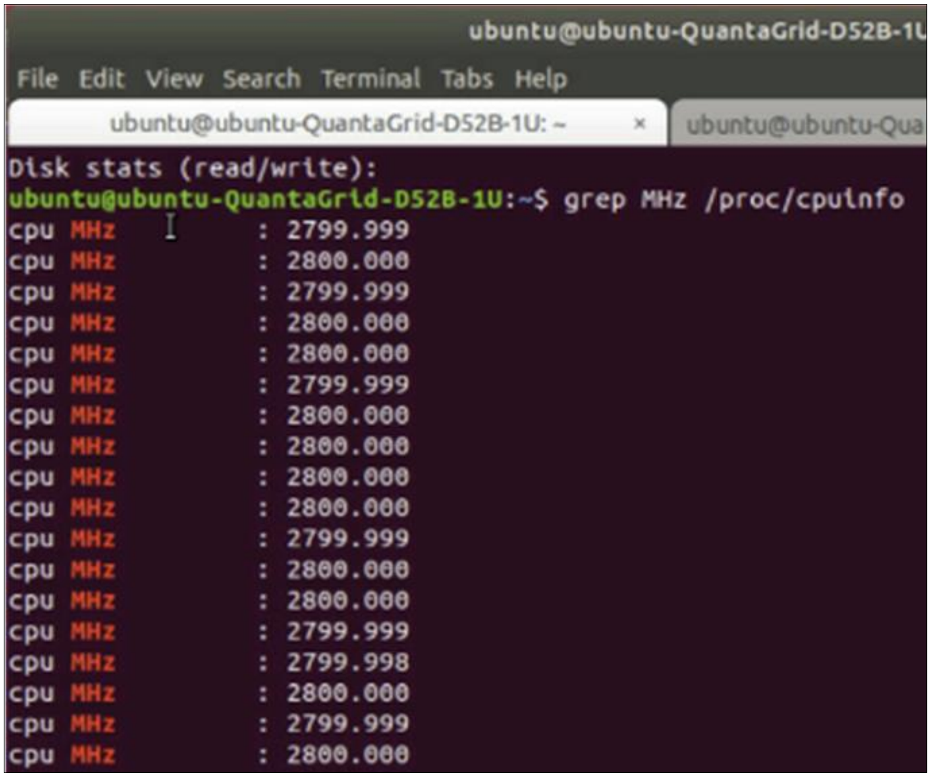
Install the SFX driver on the machine as mentioned in the section "Driver Install".

Open the OS terminal -> enter sudo session -> run the "opt/ScaleFlux/sfx-pincpu -u" script to refresh the driver. This will pin the process related to ScaleFlux to run across all the threads. Verify it with the command below.

```
$ grep MHz /proc/cpuinfo
```

To disable the CPU C-State or idling, run the below commands and update the grub.

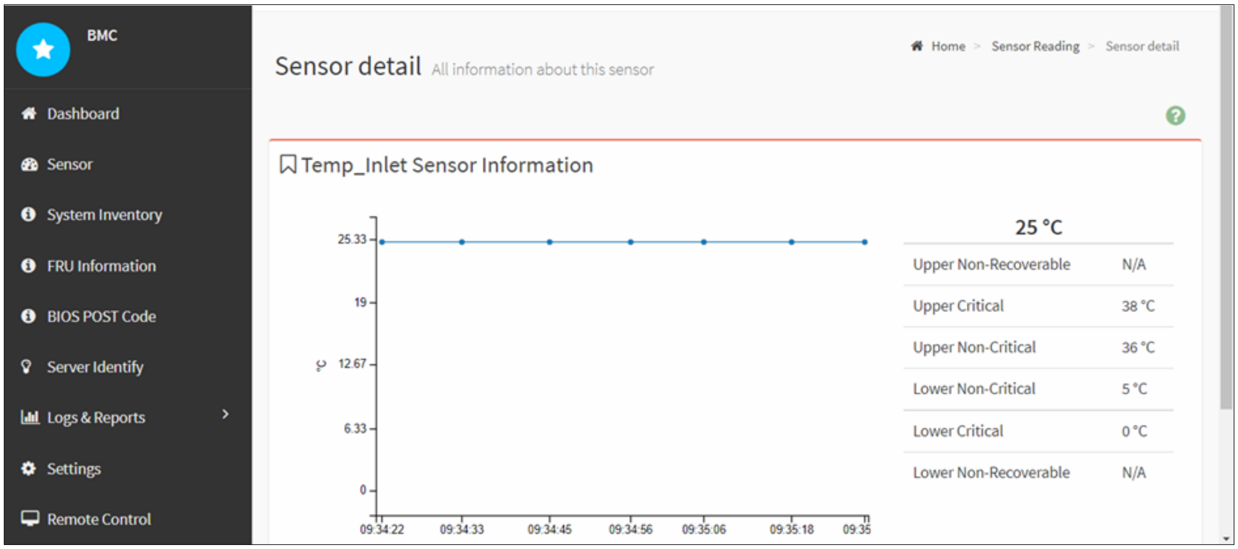
```
$>sudo nano /etc/default/grub
GRUB_CMDLINE_LINUX="crashkernel=auto rhgb quiet intel_idle.max_cstate=0 processor.max_cstate=0 idle=poll"
$>update-grub
```


ubuntu@ubuntu-QuantaGrid-D52B-1U
File Edit View Search Terminal Tabs Help
ubuntu@ubuntu-QuantaGrid-D52B-1U: ~
ubuntu@ubuntu-Qua

```
Disk stats (read/write):
ubuntu@ubuntu-QuantaGrid-D52B-1U:~$ grep MHz /proc/cpuinfo
cpu MHz         : 2799.999
cpu MHz         : 2800.000
cpu MHz         : 2799.999
cpu MHz         : 2800.000
cpu MHz         : 2800.000
cpu MHz         : 2799.999
cpu MHz         : 2800.000
cpu MHz         : 2800.000
cpu MHz         : 2800.000
cpu MHz         : 2800.000
cpu MHz         : 2799.999
cpu MHz         : 2800.000
cpu MHz         : 2800.000
cpu MHz         : 2799.999
cpu MHz         : 2799.998
cpu MHz         : 2800.000
cpu MHz         : 2799.999
cpu MHz         : 2800.000
```

Figure 3 Core frequency utilization

Proper ventilation and standard datacentre operating temperature



Sensor detail All information about this sensor

Temp_Inlet Sensor Information

25 °C

Upper Non-Recoverable	N/A
Upper Critical	38 °C
Upper Non-Critical	36 °C
Lower Non-Critical	5 °C
Lower Critical	0 °C
Lower Non-Recoverable	N/A

Figure 4 BMC sensor temperature informations

The CSD 2000 requires a 64-bit x86 server-class platform (e.g., Xeon Scalable or AMD Epyc processors). RISC based architectures such as ARM are not currently supported. A Linux based operating system is required. Driver software is provided in Debian and RPM formats with repository support for Ubuntu and RHEL/CentOS distributions.

Native Sector Size

The physical sector size of the CSD2000 drive is 4KB, and partitions that are not aligned to 4KB boundaries will impact performance. It is recommended to use the parted command to specify the partition size as a percentage. This way, the partition will be automatically aligned to the 4KB boundary when it is created. For example.

```
sudo parted /dev/sfdv0n1 -s "mklabel gpt mkpart primary 0% 30%"
```

Executing this command on a drive with a capacity of 6.4T will create a partition with a capacity of 1.8T and automatically align it to a 4KB boundary. Use the following command to see if the starting address is aligned to the 4KB boundary.

```
sudo parted /dev/sfdv0n1 -s "unit kiB print"
```

Memory requirements

The CSD 2000 uses an “open channel” style interface that requires the installation of a driver. The driver is based on the NVMe driver, with additional logic added for Flash management. Because the driver caches the Flash translation layer in host memory, it will occupy a portion of DRAM. The following formula calculates the amount of host memory needed to install the driver. If there is insufficient memory, the driver will not be loaded.

Logical Capacity (GB) x 0.2% + 3.5GB = Required System Memory per Drive

For example, for a 3.2TB CSD 2000:

3200 GB x 0.2% + 3.5 GB = 9.9 GB

When there are multiple drives installed in the system, multiply the number of drives by the memory required for a single drive to get the total amount of memory required. For example, if there are 12 3.2TB CSD 2000 drives installed in the system.

9.9 GB x 12 ~= 120 GB

Verify Supported Interfaces

CSD 2000 comes in add-in card and U.2 form factors. For add-in cards, a physical x8 CEM slot-connector is required. The slot must support PCIe Gen3. The 2.5" U.2 form factors support SAS/SATA or PCIe using the same SFF-8639 connector, but not at the same time. Because the connector is the same, a 2.5" U.2 drive will mechanically fit in the slot no matter which interface is present. Therefore, it is critical to verify that the U.2 drive bay is wired for PCIe and not SAS/SATA. Furthermore, the U.2 slots must not be attached to a storage controller (e.g., a Broadcom Mega RAID device) that prevents the host operating system from accessing PCIe devices directly. PCIe switches or re-timers do not pose any issues.

Infrastructure Setup

The system architecture for GlusterFS distributed network filesystem that is deployed using this document is as shown in the Figure 5 GlusterFS system architecture

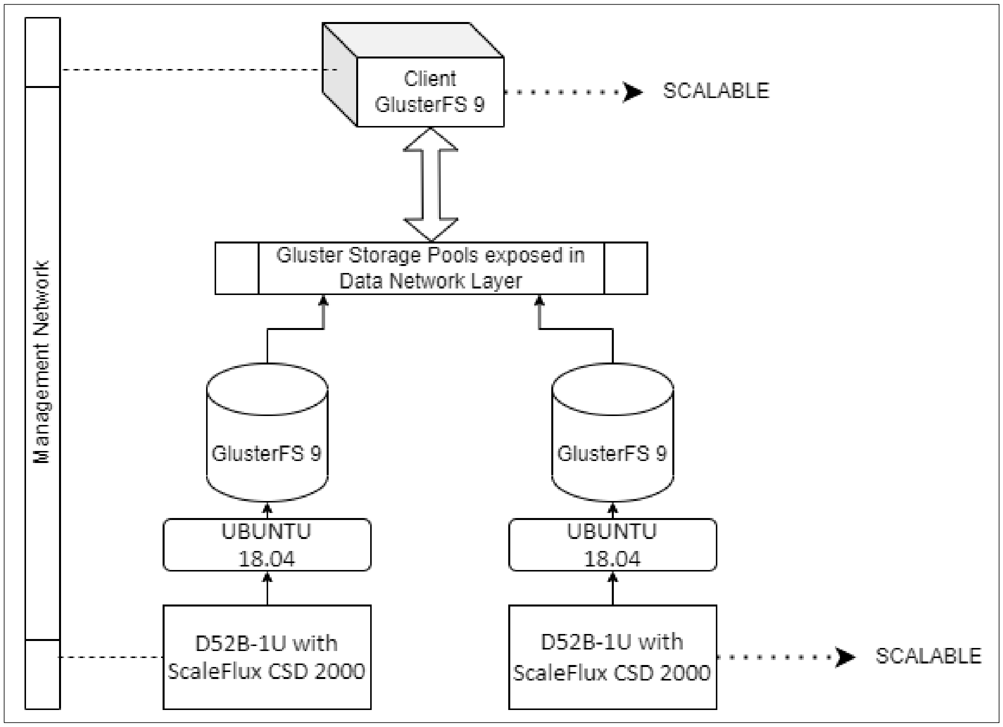


Figure 5 GlusterFS system architecture

Qualification Hardware with the CSD

The D52B-1U high availability 12x NVMe hot swappable server is used for the CSD 2000 qualification. The server configuration is as below:

BIOS Version	:	3B13
Firmware Revision	:	4.96

CPU	:	2x Intel(R) Xeon(R) Gold 6130 CPU @ 2.10GHz
RAM	:	4x 32 GB Samsung M393A4K40DB3-CWE 3200 MT/s
SAS Controller	:	1x LSI /Symbios SAS3216 PCI-Express Fusion-MPT SAS-3 (rev 01)
SATA drive	:	1x SAMSUNG MZ7LH240 404Q 240GB
OS	:	UBUNTU 18.04
Test drive	:	2x ScaleFlux CSD 2000
Test tools installed	:	FIO 3.1, IOMeter 1.1

Driver Install

Once the hardware installation is complete, confirm that the device is recognized properly by the operating system with the following command.

```
$ lspci -d cc53:
04:00.0 Non-Volatile memory controller: ScaleFlux Inc. Device 0002 (Rev 01)
```

Note: Systems running on older Linux kernel versions may not display the words "ScaleFlux Inc.".

The lspci tool is part of the “pciutils” package if it is not already installed with your distribution.

Determine the Linux System Distribution

The following command will show the running distribution [4]:

```
$ cat /etc/*rel*
```

For CentOS/RedHat distributions, run the following command to install the ScaleFlux repository:

```
$ curl -s https://packagecloud.io/install/repositories/scaleflux/sfx3x/script.rpm.sh | sudo bash
```

For Debian/Ubuntu distributions, run the following command to install the ScaleFlux repository:

```
$ curl -s https://packagecloud.io/install/repositories/scaleflux/sfx3x/script.deb.sh | sudo bash
```

For servers that do not have direct Internet access, please visit the following link, and enter the kernel version in the search box to locate and download the corresponding driver:

https://packagecloud.io/scaleflux/sfx3x

Determining the Kernel Version

The following command will return the running kernel version:

```
$ uname -r
```

Examples: 3.10.0-862.el7.x86_84 # CentOS

4.15.0-76-generic # Ubuntu

If using a binary driver, you will need to install a driver package that matches the running kernel.

Installing Binary Drivers

Install the corresponding precompiled package matching the running kernel version. For CentOS / RedHat distributions, execute the following command. Here "xxxx" is the kernel version number output by "uname -r", without the suffix (e.g. not including .el7.x86_64).

```
$ yum search sfx3xdriver-xxxx sudo yum install sfx3xdriver-xxxx
```

For Debian / Ubuntu distributions, execute the following command. Here "xxxx" is the kernel version number output by "uname -r", without the suffix (e.g. not including -generic).

```
$ apt search sfx3xdriver-xxxx sudo apt install sfx3xdriver-xxxx
```

Installing Source Drivers

The source package requires a compilation environment. Installing the source package will automatically install the package dependencies.

Note that the kernel headers that match the kernel need to be installed manually, such as kernel-headers-xxxx on centOS / RedHat. For CentOS / RedHat distributions, execute the following command.

```
$ yum search sfx3xdriver-src
$ sudo yum install sfx3xdriver-src
```

For Debian / Ubuntu distributions, execute the following command.

```
$ apt search sfx3xdriver-src
$ sudo apt install sfx3xdriver-src
```

Validation

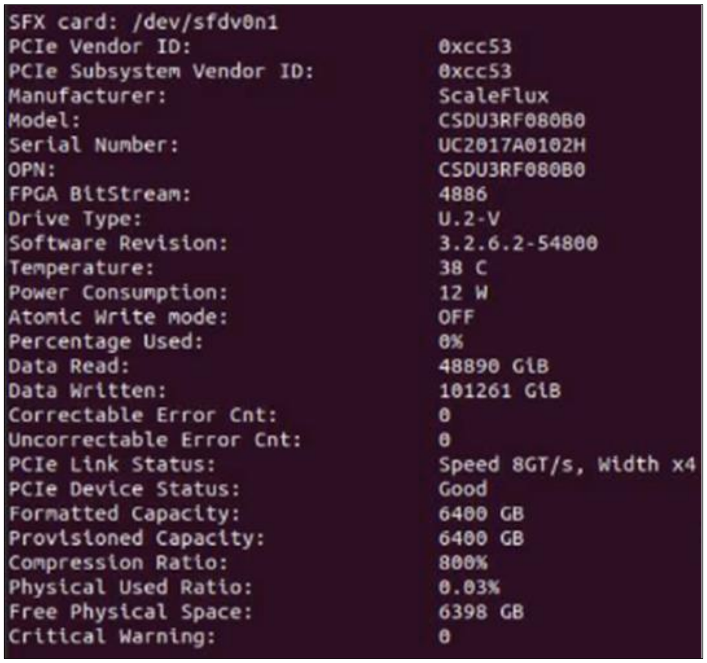
The sfx-status tool can be used to view the status of all drives or specified drives. The usage is as follows:

```
$ sudo sfx-status
```

To view the status of a specific drive, give the block device path, or represent it as a number. The number is an integer starting from 0.

```
$ sudo sfx-status /dev/sfdv[0-9]n1
```

Output descriptions are as follows.



```
SFX card: /dev/sfdv0n1
PCIe Vendor ID:                    0xcc53
PCIe Subsystem Vendor ID:          0xcc53
Manufacturer:                      ScaleFlux
Model:                             CSDU3RF080B0
Serial Number:                     UC2017A0102H
OPN:                               CSDU3RF080B0
FPGA BitStream:                    4886
Drive Type:                        U.2-V
Software Revision:                 3.2.6.2-54800
Temperature:                       38 C
Power Consumption:                 12 W
Atomic Write mode:                 OFF
Percentage Used:                   0%
Data Read:                         48890 GiB
Data Written:                      101261 GiB
Correctable Error Cnt:             0
Uncorrectable Error Cnt:           0
PCIe Link Status:                  Speed 8GT/s, Width x4
PCIe Device Status:                Good
Formatted Capacity:                6400 GB
Provisioned Capacity:              6400 GB
Compression Ratio:                 800%
Physical Used Ratio:               0.03%
Free Physical Space:               6398 GB
Critical Warning:                  0
```

Figure 6 ScaleFlux CSD status

• FPGA BitStream	- FPGA image version
• Drive Type	- Form factor
• Software Revision	- Driver version
• Temperature	- Current drive temperature (PCB)
• Power Consumption	- Current power consumption
• Percentage Used	- Used drive life
• Data Read	- Total data read by the host
• Data Written	- Total data written by the host
• Correctable Error Cnt	- Number of RAID-recovered reads
• Uncorrectable Error Cnt	- Number of un-correctable reads
• Formatted Capacity	- Host visible capacity
• Provisioned Capacity	- Physically available capacity
• Physical Used Ratio	- Used physical capacity
• Free Physical Space	- Available physical capacity

.

4 Base Product Deployment

This section is a guide to install and use the extended capacity of the ScaleFlux CSD 2000 with their inbuilt compression features exposed as GlusterFS target for clients to access.

Preinstallation Requirements

Modifying Drive Capacity

The following command can be used to expand the logical capacity of the drive, i.e., the capacity visible to the operating system.

```
$ sfx-nvme sfx change-cap <device> --cap=<NUM>|-c=<NUM>
```

Example:

```
$ sudo sfx-nvme sfx change-cap /dev/sfdv0n1 -c 9600
```

Expanding the logical capacity requires that the corresponding data compression ratio is achieved. For example, if the data is known to have a compression ratio of approximately 3:1 (i.e., 300G of data is compressed to 100G), then increasing the logical capacity of a 3.2T drive to 6.4T can be considered safe.

Regardless of the data compression ratio, after increasing the logical capacity, be sure to:

1. Use the "-o discard" parameter when mounting a drive/partition or run "fstrim" after deleting a file to notify the device that the space is no longer needed. Both "-o discard" and "fstrim" are used to mark blocks belonging to deleted files as free by sending trim commands to the device.

2. Implement monitoring of available physical capacity to avoid any unexpected out-of-space conditions.

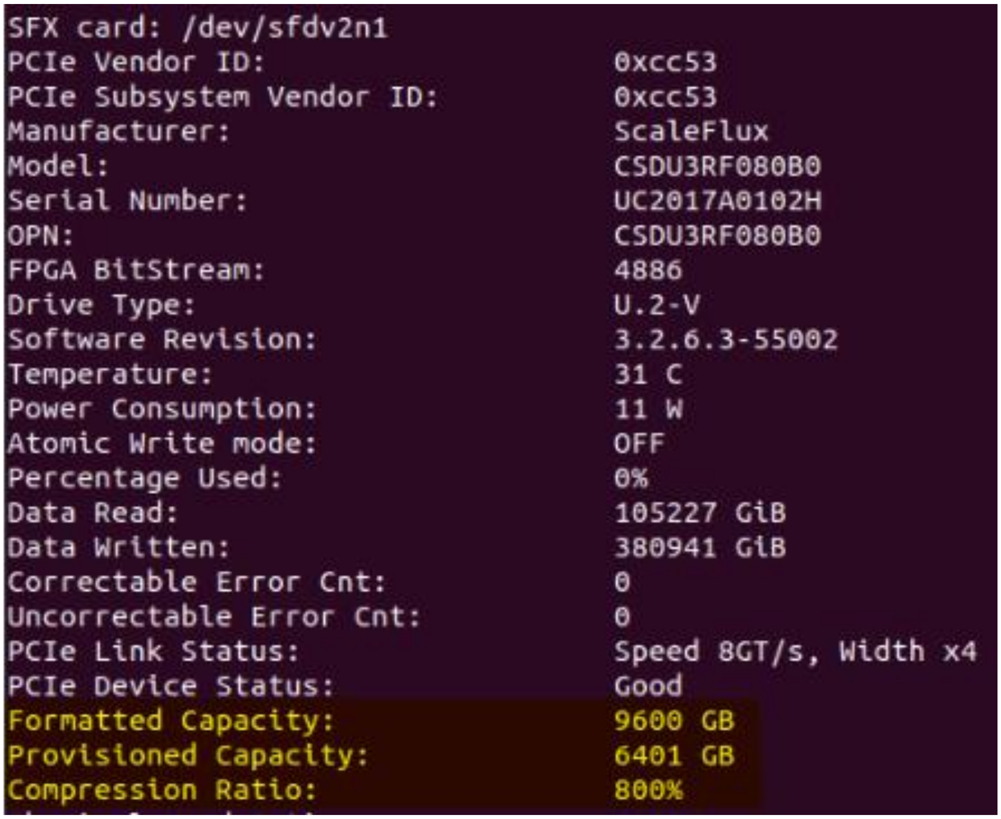
Expanding the logical capacity does not erase the data on the drive. User data is preserved. File system expansion needs to be performed manually by the user with the appropriate tool. Reducing the logical capacity, on the other hand, is destructive and needs to be done after first clearing the data. Data can be cleared using the following command:

```
$ sfx-nvme sfx set-feature -f 0xdc /dev/sfxv[0-9]+
```

Example:

```
$ sudo sfx-nvme sfx set-feature -f 0xdc /dev/sfxv0
```

For more information on configuring drive capacity, please refer to the Extended Capacity User guide.



```
SFX card: /dev/sfdv2n1
PCIe Vendor ID:                     0xcc53
PCIe Subsystem Vendor ID:           0xcc53
Manufacturer:                       ScaleFlux
Model:                              CSDU3RF080B0
Serial Number:                      UC2017A0102H
OPN:                                CSDU3RF080B0
FPGA BitStream:                     4886
Drive Type:                         U.2-V
Software Revision:                  3.2.6.3-55002
Temperature:                        31 C
Power Consumption:                  11 W
Atomic Write mode:                  OFF
Percentage Used:                    0%
Data Read:                          105227 GiB
Data Written:                       380941 GiB
Correctable Error Cnt:              0
Uncorrectable Error Cnt:            0
PCIe Link Status:                   Speed 8GT/s, Width x4
PCIe Device Status:                 Good
Formatted Capacity:                 9600 GB
Provisioned Capacity:               6401 GB
Compression Ratio:                  800%
```

Figure 7 ScaleFlux CSD status with increased drive capacity

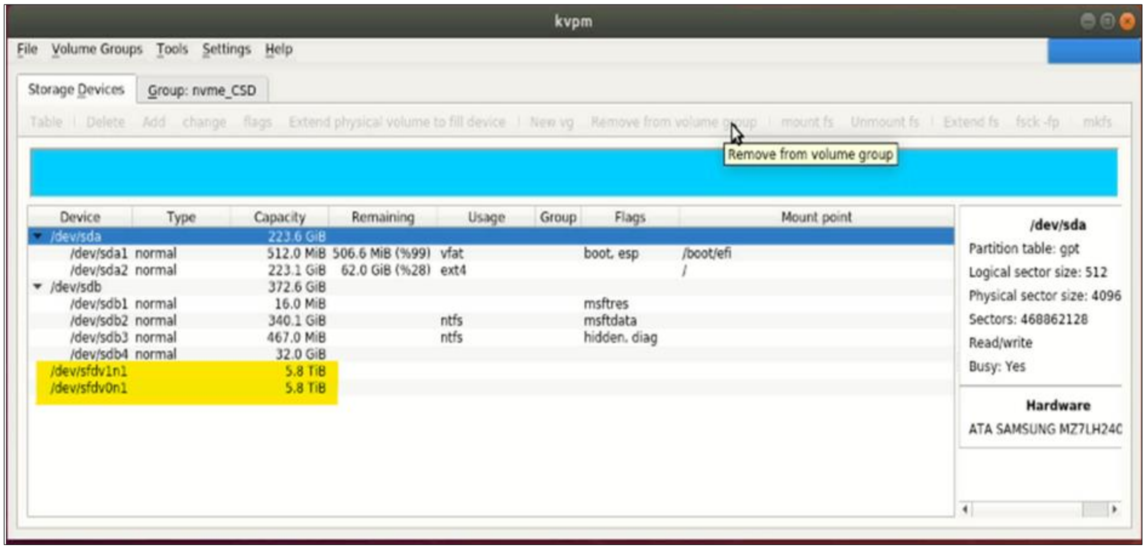
Logical Volume target for the Gluster File System

The CSDs on the server is pooled as a logical volume with the KVPM GUI utility on an UBUNTU 18.04. Install KVPM from the apt repository by

```
$ sudo apt install kvpm
$ sudo kvpm
```

Pre-requisite for the KVPM utility is the LVM2 package that can be installed using the command below:

```
$ sudo apt install lvm2
```



Device	Type	Capacity	Remaining	Usage	Group	Flags	Mount point
/dev/sda		223.6 GiB					
/dev/sda1	normal	512.0 MiB	506.6 MiB (%99)	vfat		boot, esp	/boot/efi
/dev/sda2	normal	223.1 GiB	62.0 GiB (%28)	ext4			/
/dev/sdb		372.6 GiB					
/dev/sdb1	normal	16.0 MiB				msftres	
/dev/sdb2	normal	340.1 GiB		ntfs		msftdata	
/dev/sdb3	normal	467.0 MiB		ntfs		hidden, diag	
/dev/sdb4	normal	32.0 GiB					
/dev/sfdv1n1		5.8 TiB					
/dev/sfdv0n1		5.8 TiB					

/dev/sda
Partition table: gpt
Logical sector size: 512
Physical sector size: 4096
Sectors: 468862128
Read/write
Busy: Yes

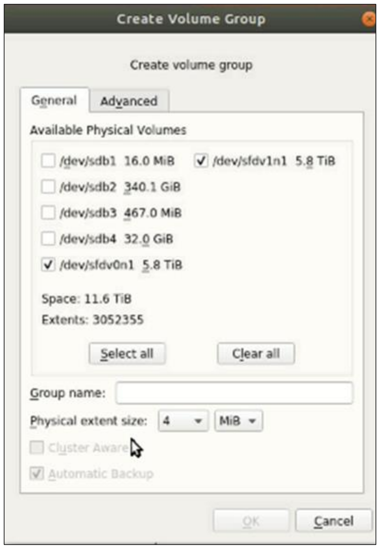
Hardware
ATA SAMSUNG MZ7LH240

Figure 8 KVPM graphical utility

Navigate to Volume Groups -> create a volume group -> select the devices involved in group -> provide a group name and click ok.

To create a Logical Volume -> Navigate to the Volume Group created in KVPM -> Select create a new volume -> Provide a name to the LV -> click on ok.

An example is as shown in the below pictures where two of the 5.8TiB ScaleFlux CSD NVMe drives are added to a logical volume "nvme_csd_volume" as a 11.6TiB LV.



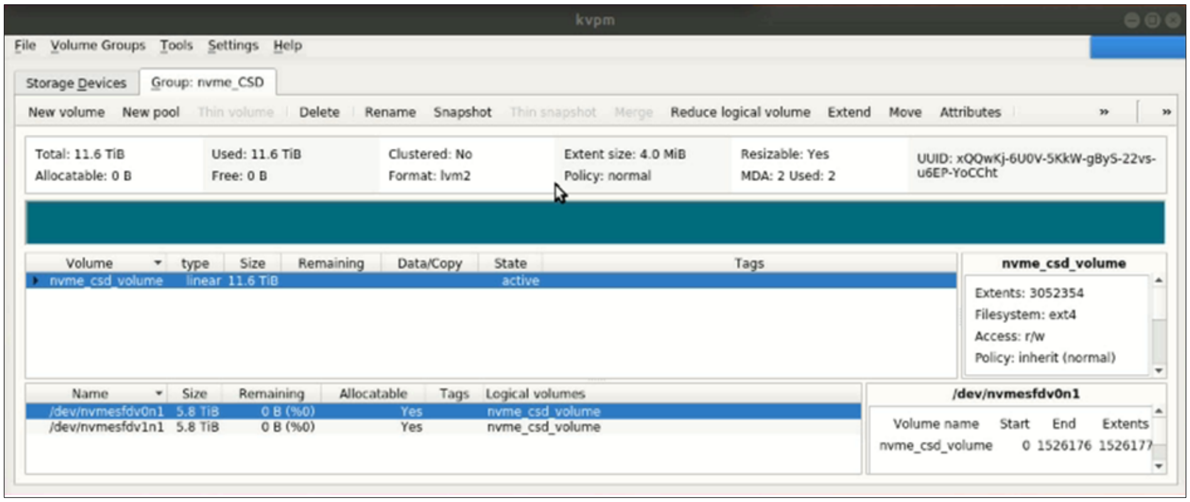
Available Physical Volumes

☐ /dev/sdb1 16.0 MiB
☐ /dev/sdb2 340.1 GiB
☐ /dev/sdb3 467.0 MiB
☐ /dev/sdb4 32.0 GiB
☑ /dev/sfdv0n1 5.8 TiB
☑ /dev/sfdv1n1 5.8 TiB

Space: 11.6 TiB
Extents: 3052355

Group name:
Physical extent size: 4 MiB
☐ Cluster Aware
☑ Automatic Backup

Figure 9 Creating a volume group



Total: 11.6 TiB
Allocatable: 0 B
Used: 11.6 TiB
Free: 0 B
Clustered: No
Format: lvm2
Extent size: 4.0 MiB
Policy: normal
Resizable: Yes
MDA: 2 Used: 2
UUID: xQQwKj-6U0V-5KkW-gByS-22vs-u6EP-YoCCht

Volume	type	Size	Remaining	Data/Copy	State	Tags
nvme_csd_volume	linear	11.6 TiB			active	

nvme_csd_volume
Extents: 3052354
Filesystem: ext4
Access: r/w
Policy: inherit (normal)

Name	Size	Remaining	Allocatable	Tags	Logical volumes
/dev/nvmesfdv0n1	5.8 TiB	0 B (%0)	Yes		nvme_csd_volume
/dev/nvmesfdv1n1	5.8 TiB	0 B (%0)	Yes		nvme_csd_volume

/dev/nvmesfdv0n1

Volume name	Start	End	Extents
nvme_csd_volume	0	1526176	1526177

Figure 10 Creating a logical volume with the grouped drives

GlusterFS setup and configuration

To Expose the logical block volume via the GlusterFS, perform the above-mentioned logical volume creation on two or more machines with CSD and follow the below steps [5]:

Step1: Install GlusterFS on the operating system using the command mentioned below:

```
$ sudo add-apt-repository ppa:gluster/glusterfs-9
$ sudo apt install glusterfs-server
```

Step2: Prepare the logical volume to be exposed via GlusterFS and mount it to a directory using the command:

```
$ sudo mkfs.ext4 /dev/mapper/nvme_CSD-nvme_csd_volume
$ sudo mkdir /mnt/HS-Gluster
$ sudo mount -t ext4 /dev/mapper/nvme_CSD-nvme_csd_volume /mnt/HS-Gluster/
```

Step3: Create a directory in the mounted location and initiate Gluster service to run as shown in the commands below:

```
$ sudo mkdir /mnt/HS-Gluster/gluster-volume
$ sudo gluster volume create <gluster volume name> <IP address of the first machine
involved>:/<directory path as created above> <IP address of the second machine
involved>:/<directory path as created above> …
```

Step4: Start the Gluster volume using the command below:

```
$ sudo gluster volume start <gluster volume name>
```

This pool is exposed via GlusterFS into the network and the clients can connect to this interface by installing GlusterFS.

Note: The version of the GlusterFS server and client needs to have consistent version to be connected.

Step5: Connect to the exposed Gluster volume through the client machine after installing Gluster and creating a mount directory using the commands below:

```
$ sudo add-apt-repository ppa:gluster/glusterfs-9
$ sudo apt install glusterfs-server
$ sudo mkdir /mnt/gluster
$ sudo mount -t glusterfs <IP address of the Gluster server>:/<gluster volume name>
/mnt/gluster
```

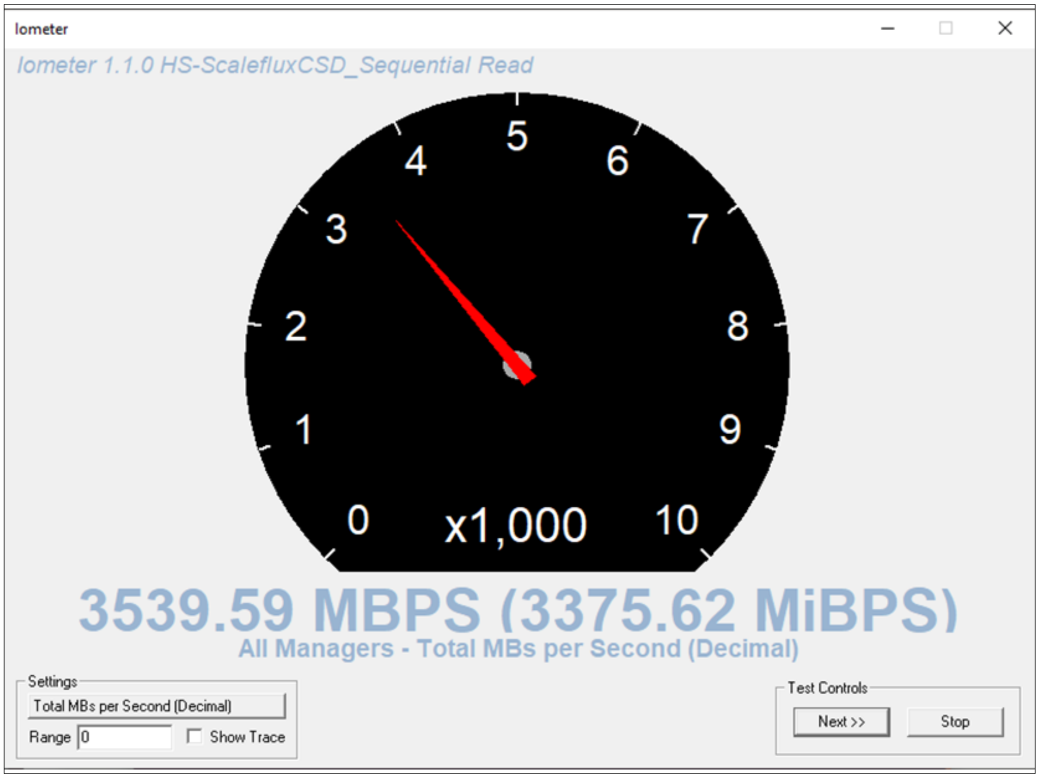
5 Testing the Appliance

The test is conducted to attach the client to the target exposing the ScaleFlux CSD 2000 drives as a logical volume. Sequential read and write tests are performed using the IOMeter tool running on a windows machine pinging the dynamo service running on the Linux target machines.

Extended Capacity Gluster exposed pool testing

The test setup is pre-conditioned with 2 of the CSD drives which are 1.5 times expanded (using procedure mentioned in section "Modifying Drive Capacity") and pooled as a logical volume using KVPM and LVM2.

The exposed pool is mounted in the client using GlusterFS Fuse. IOMeter test on the mounted file system hit a sequential read speed of 3.5GB/s and a sequential write speed of 2.7GB/s.



Iometer

Iometer 1.1.0 HS-ScalefluxCSD_Sequential Read

3539.59 MBPS (3375.62 MiBPS)

All Managers - Total MBs per Second (Decimal)

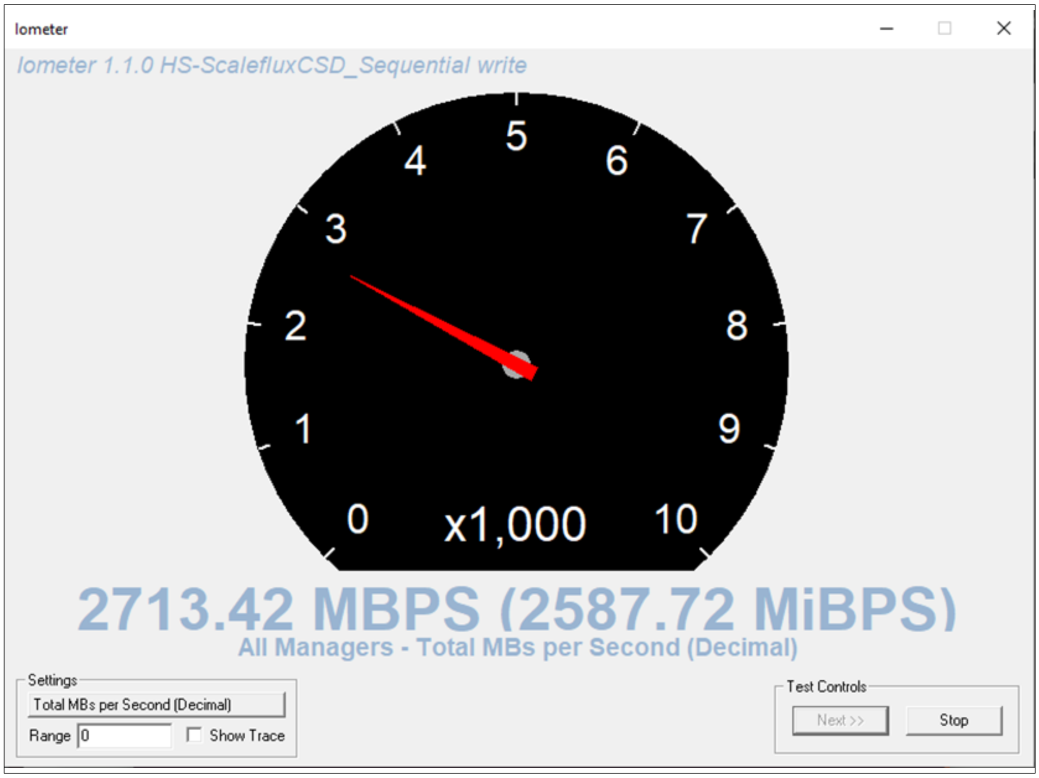
Settings

Total MBs per Second (Decimal)

Range 0

□ Show Trace

Figure 11 IOMeter sequential read on GlusterFS target



2713.42 MBPS (2587.72 MiBPS)
All Managers - Total MBs per Second (Decimal)

Settings: Total MBs per Second (Decimal)
Range 0

Figure 12 IOMeter sequential write on GlusterFS targetss

6 Additional Setup and Deployment

Capacity Monitoring

The CSD 2000 drive supports extending the advertised capacity to save cost when data is compressible (e.g., a 3.2TB drive can be set to report 6.4TB for storing data). Since the physically available storage is still fixed, if the compression ratio is not sufficient to support the extended capacity, a situation can occur where there is still host visible capacity, but the physical capacity is full. When using CSD 2000 with extended capacity, monitoring must be deployed to track the physical capacity and avoid out of space conditions.

The CSD 2000 provides a sysfs interface to obtain information about the current physical capacity utilization and overall compression ratio. The following sysfs file provides a consolidated human readable capacity report:

```
/sys/block/sfdv[0-9]n1/sfx_smart_features/sfx_capacity_stat
```

An example with field descriptions is as follows:

```
$ cat /sys/block/sfdv1n1/sfx_smart_features/sfx_capacity_stat
```

Output:

free_space	physical_size	logical_size	comp_ratio	provisioned_cap	space_flag
5303721064	947512904	3837963264	4.05	6251233968	0

- free_space Available physical capacity in sectors (512 bytes)

- physical_size Used physical capacity in sectors (512 bytes)
- logical_size Logical data size in sectors (512 bytes)
- comp_ratio Compression Ratio
- provisioned_cap Total logical capacity of the drive-in sectors (512 bytes)
- space_flag Out of capacity flag (a value of 1 indicates out of capacity)

The parameters included in this report are also available as individual sysfs parameters and are described as follows:

- sfx_freespace - Currently available physical capacity expressed in sectors (512 bytes) followed by the total physical capacity also expressed in sectors.
- sfx_logical_size - Amount of host data currently being stored expressed in sectors (512 bytes). Note that this may be higher than what is reported by a file system if trim is not used.
- sfx_physical_size - Amount of physical data currently being stored expressed in sectors (512 bytes).
- sfx_comp_ratio - Compression ratio of the current valid data (equal to sfx_logical_size / sfx_physical_size)
- space_flag - Out of space indicator flag (set to 1 when the drive is out of physical capacity)

In addition to the sysfs reporting mechanisms, an IOCTL interface is also available. Please contact ScaleFlux for details on the IOCTL interface.ss

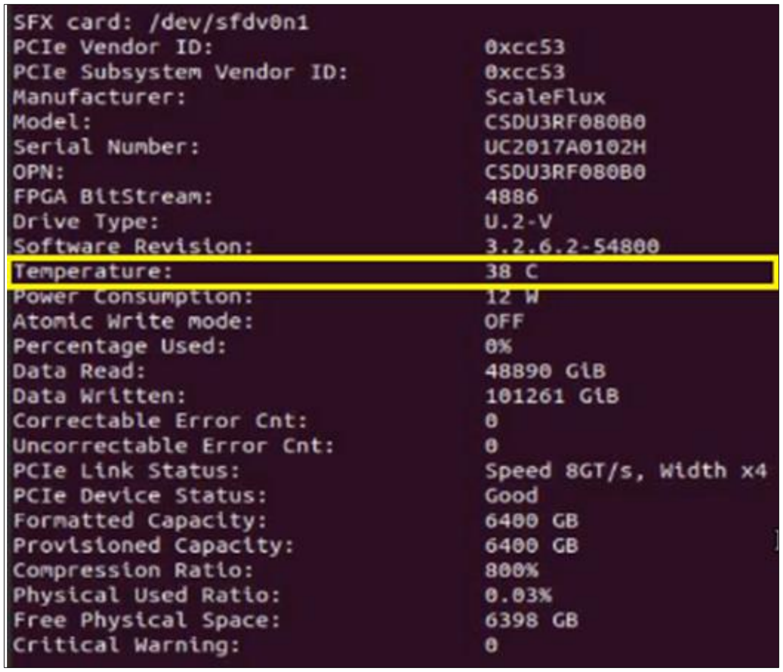
Temperature Monitoring

The nvme-cli tool can be used to conveniently monitor temperature and other key parameters.

```
$ sudo apt install nvme-cli
```

Upon installation, this tool will be installed with the name sfx-nvme, which is a version of nvme-cli that includes the ScaleFlux plugin and has been co-validated with installed driver.

Example output:



```
SFX card: /dev/sfdv0n1
PCIe Vendor ID:                     0xcc53
PCIe Subsystem Vendor ID:           0xcc53
Manufacturer:                       ScaleFlux
Model:                              CSDU3RF080B0
Serial Number:                      UC2017A0102H
OPN:                                CSDU3RF080B0
FPGA BitStream:                     4886
Drive Type:                         U.2-V
Software Revision:                  3.2.6.2-54800
Temperature:                        38 C
Power Consumption:                  12 W
Atomic Write mode:                  OFF
Percentage Used:                    0%
Data Read:                          48890 GiB
Data Written:                       101261 GiB
Correctable Error Cnt:              0
Uncorrectable Error Cnt:            0
PCIe Link Status:                   Speed 8GT/s, Width x4
PCIe Device Status:                 Good
Formatted Capacity:                 6400 GB
Provisioned Capacity:               6400 GB
Compression Ratio:                  800%
Physical Used Ratio:                0.03%
Free Physical Space:                6398 GB
Critical Warning:                   0
```

Figure 13 ScaleFlux drive temperature status

Drive Management

All device parameters can be modified using nvme-cli, but it is recommended to use the locally installed sfx-nvme version of nvme-cli to ensure that the ScaleFlux plug-in is up to date with the installed driver.

Changing the Logical Sector Size

The blockdev command can be used to get the current sector size. By default, this is 512 bytes.

```
$ sudo blockdev --getss /dev/sfdv0n1
```

The sector size can be set to 4kB as follows:

```
$ sudo sfx-nvme format /dev/sfdv0n1 -l 1
```

Changing back to 512-byte sectors:

```
$ sudo sfx-nvme format /dev/sfdv0n1 -l 0
```

Warning: Changing the logical sector size will erase all data on the drive.

Enabling Atomic Writes

Atomic write ensures that block IO requests are written as a unit such that all sectors are written, or no sectors are written. For example, a 16kB write IO request will never be broken into separate 4k writes.

The following command is used to enable or disable atomic writes:

```
$ sfx-nvme sfx set-feature <device> [--feature-id=<fid> | -f <fid>] [--value=<value> | - v
<value>]
```

[--feature-id=<fid> | -f <fid>] - feature id, atomic write feature id is 1

[--value=<value> | -v <value>] - Function value. Turn on is 1, turn off is 0

Examples:

```
$ sudo sfx-nvme sfx set-feature /dev/sfdv0n1 -f 1 -v 1 # Turn on atomic write
$ sudo sfx-nvme sfx set-feature /dev/sfdv0n1 -f 1 -v 0 # turn off atomic write
```

The use of atomic write has the following prerequisites.

1. Linux kernel version needs to be 2.6 or higher.

2. Maximum request is 256KB.

3. The sector size must be 4KB. 512B atomic writes are not supported.

4. Atomic writes need to be used in conjunction with file system Direct I/O to achieve the desired effect.

5. The file system block size must be adjusted to the desired atomic unit. For example, if MySQL uses 16K pages with the ext4 file system, 16K allocation units must be configured when using atomic writes. This can be achieved as follows:

```
$ sudo mkfs.ext4 -O extent,bigalloc -C 16384 /dev/sfdv0n1
```

Data Wipe (Factory Reset)

The CSD 2000 driver provides a data wipe feature that can be used to wipe the data on the drive and reset all statistics except life usage:

```
$ sudo sfx-nvme sfx set-feature -f 0xdc /dev/sfxv0
```

Warning: This operation will erase all data on the drive.

7 Addendum

Test Suite on single drive

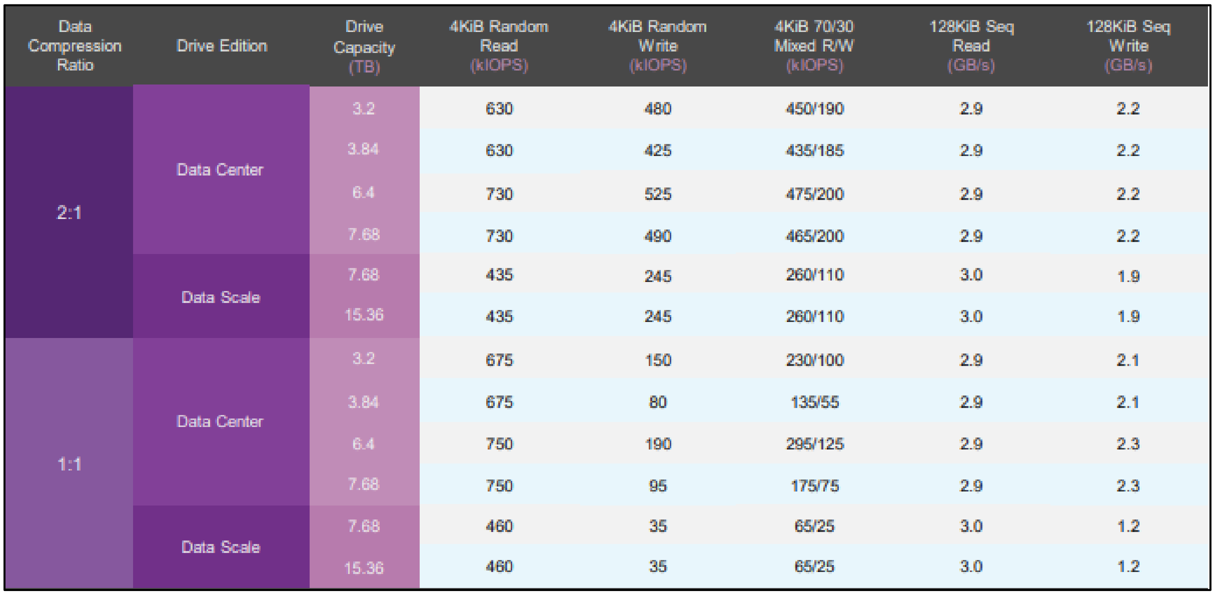
Baseline FIO test https://github.com/jinqiangwang/fio-baseline

Run fio-baseline.sh from the github repository by modifying the "comp_ratio" and "disks" variable corresponding to the environment.

Example: comp_ratio= 80 for 3:1 compression and disks= (sfdv0n1)

buffer compress percentage	compression ratio
0-39	1:1
40-69	2:1
70-89	3:1
90	5:1
95	8:1
100	28:1

ScaleFlux speed test specification can be inferred from the table below and the actual test results for various test scenarios are provided below.

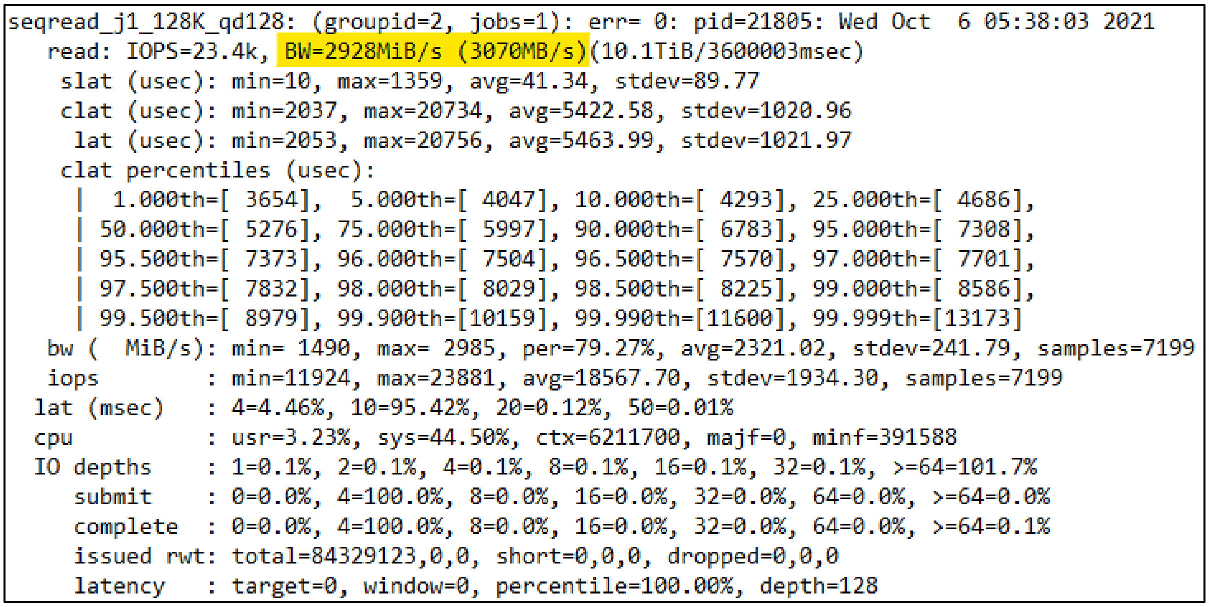


Data Compression Ratio	Drive Edition	Drive Capacity (TB)	4KiB Random Read (kIOPS)	4KiB Random Write (kIOPS)	4KiB 70/30 Mixed R/W (kIOPS)	128KiB Seq Read (GB/s)	128KiB Seq Write (GB/s)
2:1	Data Center	3.2	630	480	450/190	2.9	2.2
		3.84	630	425	435/185	2.9	2.2
		6.4	730	525	475/200	2.9	2.2
		7.68	730	490	465/200	2.9	2.2
	Data Scale	7.68	435	245	260/110	3.0	1.9
		15.36	435	245	260/110	3.0	1.9
1:1	Data Center	3.2	675	150	230/100	2.9	2.1
		3.84	675	80	135/55	2.9	2.1
		6.4	750	190	295/125	2.9	2.3
		7.68	750	95	175/75	2.9	2.3
	Data Scale	7.68	460	35	65/25	3.0	1.2
		15.36	460	35	65/25	3.0	1.2

Figure 14 IOPs and throughput with data compression and drive capacity

2:1 Compressed Individual drive test

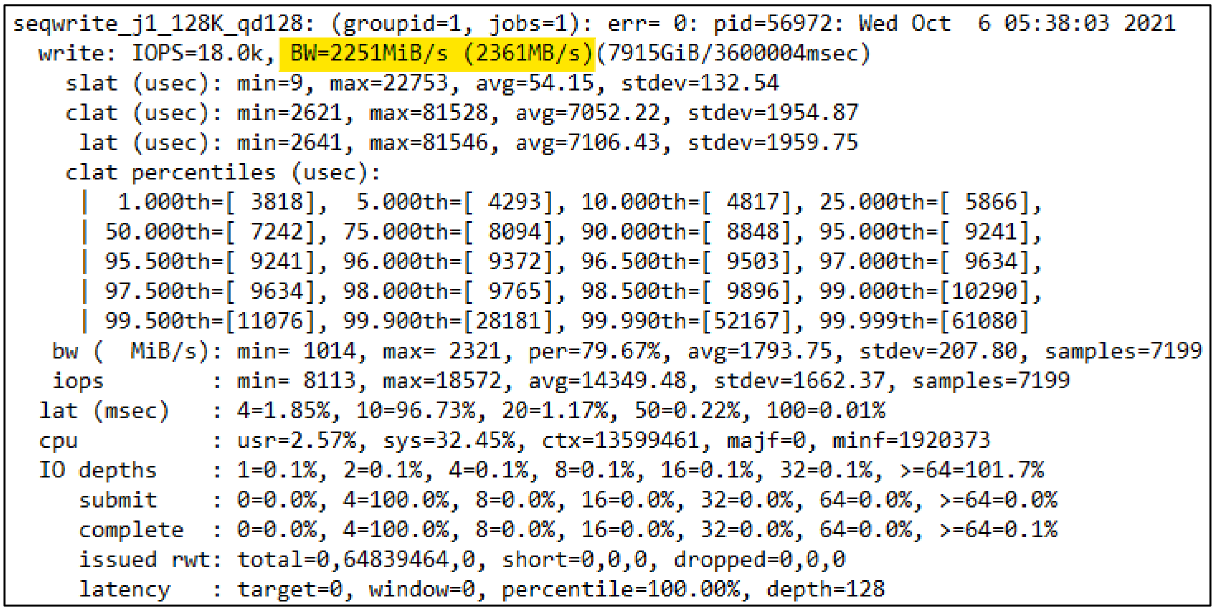
The Individual drive is tested with its actual drive capacity and the sequential read test shows a bandwidth of 2.9 GB/s on a 2:1 compression type data. These results make the drive efficient for seamless application with large multimedia storage or backups.



```
seqread_j1_128K_qd128: (groupid=2, jobs=1): err= 0: pid=21805: Wed Oct  6 05:38:03 2021
   read: IOPS=23.4k, BW=2928MiB/s (3070MB/s)(10.1TiB/3600003msec)
    slat (usec): min=10, max=1359, avg=41.34, stdev=89.77
    clat (usec): min=2037, max=20734, avg=5422.58, stdev=1020.96
     lat (usec): min=2053, max=20756, avg=5463.99, stdev=1021.97
    clat percentiles (usec):
     |  1.000th=[ 3654],  5.000th=[ 4047], 10.000th=[ 4293], 25.000th=[ 4686],
     | 50.000th=[ 5276], 75.000th=[ 5997], 90.000th=[ 6783], 95.000th=[ 7308],
     | 95.500th=[ 7373], 96.000th=[ 7504], 96.500th=[ 7570], 97.000th=[ 7701],
     | 97.500th=[ 7832], 98.000th=[ 8029], 98.500th=[ 8225], 99.000th=[ 8586],
     | 99.500th=[ 8979], 99.900th=[10159], 99.990th=[11600], 99.999th=[13173]
   bw (  MiB/s): min= 1490, max= 2985, per=79.27%, avg=2321.02, stdev=241.79, samples=7199
   iops        : min=11924, max=23881, avg=18567.70, stdev=1934.30, samples=7199
  lat (msec)   : 4=4.46%, 10=95.42%, 20=0.12%, 50=0.01%
  cpu          : usr=3.23%, sys=44.50%, ctx=6211700, majf=0, minf=391588
  IO depths    : 1=0.1%, 2=0.1%, 4=0.1%, 8=0.1%, 16=0.1%, 32=0.1%, >=64=101.7%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     complete  : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.1%
     issued rwt: total=84329123,0,0, short=0,0,0, dropped=0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=128
```

Figure 15 FIO sequential read test on individual drive

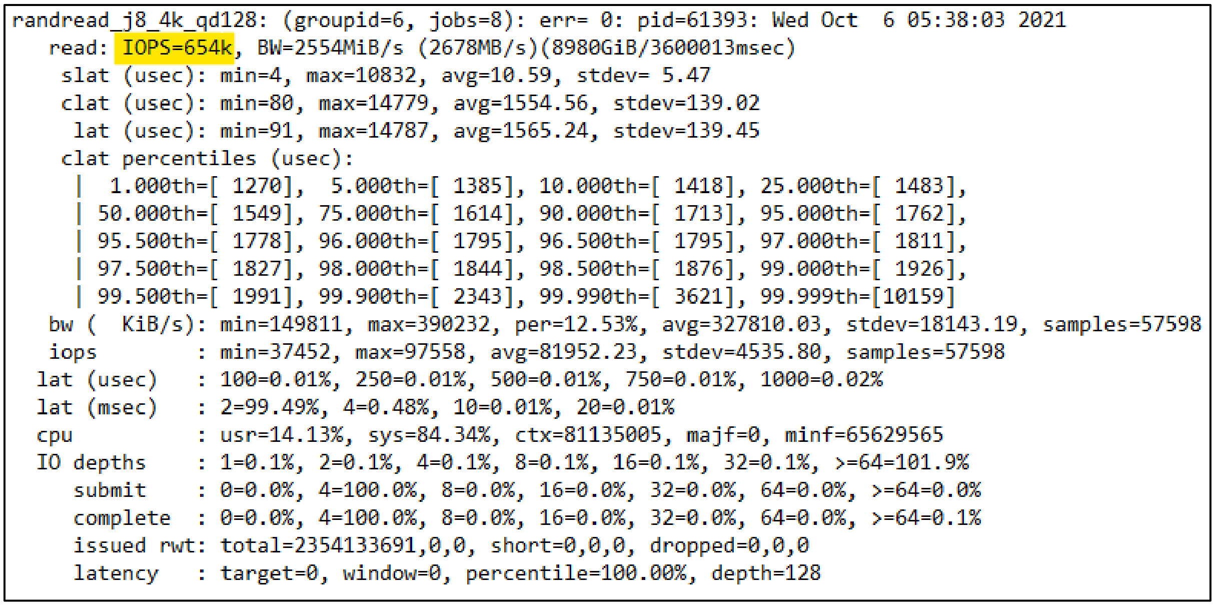
The sequential write test shows a bandwidth of 2.3GB/s to write the data that can be compressed as 2:1 ratio.



```
seqwrite_j1_128K_qd128: (groupid=1, jobs=1): err= 0: pid=56972: Wed Oct  6 05:38:03 2021
  write: IOPS=18.0k, BW=2251MiB/s (2361MB/s)(7915GiB/3600004msec)
    slat (usec): min=9, max=22753, avg=54.15, stdev=132.54
    clat (usec): min=2621, max=81528, avg=7052.22, stdev=1954.87
     lat (usec): min=2641, max=81546, avg=7106.43, stdev=1959.75
    clat percentiles (usec):
     |  1.000th=[ 3818],  5.000th=[ 4293], 10.000th=[ 4817], 25.000th=[ 5866],
     | 50.000th=[ 7242], 75.000th=[ 8094], 90.000th=[ 8848], 95.000th=[ 9241],
     | 95.500th=[ 9241], 96.000th=[ 9372], 96.500th=[ 9503], 97.000th=[ 9634],
     | 97.500th=[ 9634], 98.000th=[ 9765], 98.500th=[ 9896], 99.000th=[10290],
     | 99.500th=[11076], 99.900th=[28181], 99.990th=[52167], 99.999th=[61080]
   bw (  MiB/s): min= 1014, max= 2321, per=79.67%, avg=1793.75, stdev=207.80, samples=7199
   iops        : min= 8113, max=18572, avg=14349.48, stdev=1662.37, samples=7199
  lat (msec)   : 4=1.85%, 10=96.73%, 20=1.17%, 50=0.22%, 100=0.01%
  cpu          : usr=2.57%, sys=32.45%, ctx=13599461, majf=0, minf=1920373
  IO depths    : 1=0.1%, 2=0.1%, 4=0.1%, 8=0.1%, 16=0.1%, 32=0.1%, >=64=101.7%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     complete  : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.1%
     issued rwt: total=0,64839464,0, short=0,0,0, dropped=0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=128
```

Figure 16 FIO sequential write test on individual drive

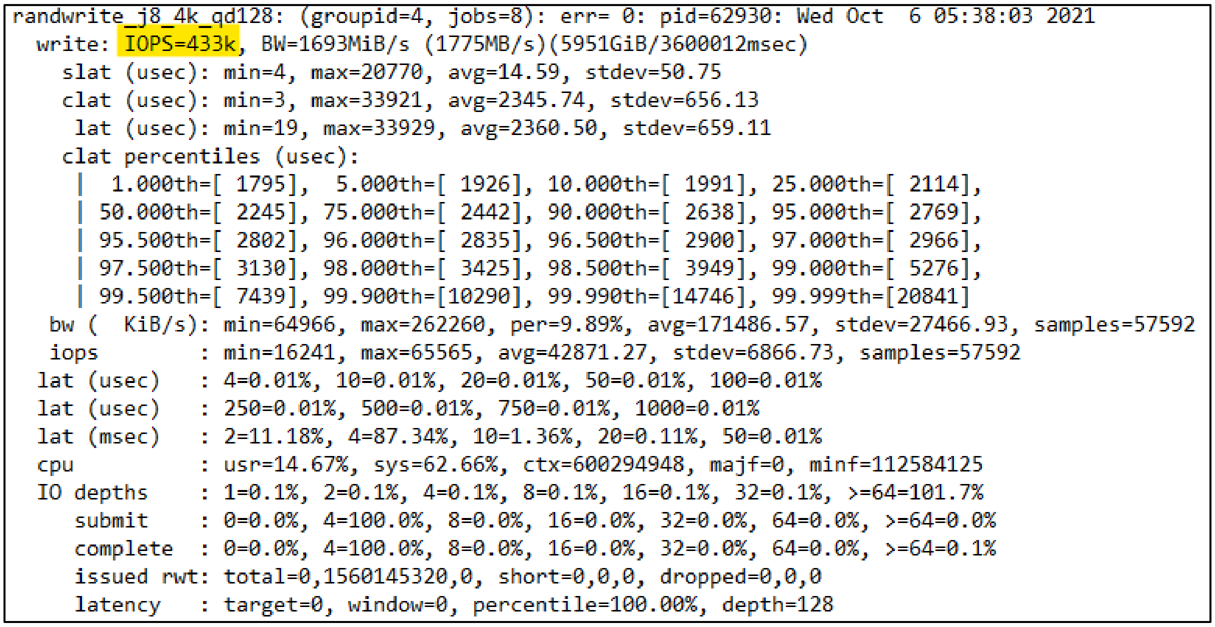
The random read performance of the drive can reach 654k IOPs on a 2:1 compressed data created by FIO utility.



```
randread_j8_4k_qd128: (groupid=6, jobs=8): err= 0: pid=61393: Wed Oct  6 05:38:03 2021
   read: IOPS=654k, BW=2554MiB/s (2678MB/s)(8980GiB/3600013msec)
    slat (usec): min=4, max=10832, avg=10.59, stdev= 5.47
    clat (usec): min=80, max=14779, avg=1554.56, stdev=139.02
     lat (usec): min=91, max=14787, avg=1565.24, stdev=139.45
    clat percentiles (usec):
     |  1.000th=[ 1270],  5.000th=[ 1385], 10.000th=[ 1418], 25.000th=[ 1483],
     | 50.000th=[ 1549], 75.000th=[ 1614], 90.000th=[ 1713], 95.000th=[ 1762],
     | 95.500th=[ 1778], 96.000th=[ 1795], 96.500th=[ 1795], 97.000th=[ 1811],
     | 97.500th=[ 1827], 98.000th=[ 1844], 98.500th=[ 1876], 99.000th=[ 1926],
     | 99.500th=[ 1991], 99.900th=[ 2343], 99.990th=[ 3621], 99.999th=[10159]
   bw (  KiB/s): min=149811, max=390232, per=12.53%, avg=327810.03, stdev=18143.19, samples=57598
   iops        : min=37452, max=97558, avg=81952.23, stdev=4535.80, samples=57598
  lat (usec)   : 100=0.01%, 250=0.01%, 500=0.01%, 750=0.01%, 1000=0.02%
  lat (msec)   : 2=99.49%, 4=0.48%, 10=0.01%, 20=0.01%
  cpu          : usr=14.13%, sys=84.34%, ctx=81135005, majf=0, minf=65629565
  IO depths    : 1=0.1%, 2=0.1%, 4=0.1%, 8=0.1%, 16=0.1%, 32=0.1%, >=64=101.9%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     complete  : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.1%
     issued rwt: total=2354133691,0,0, short=0,0,0, dropped=0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=128
```

Figure 17 FIO random read test on individual drive

For applications that require higher IOPs in write performance, below are the results that show the ScaleFlux 2000 series SSD can randomly write up to 433k IOPs under 2:1 compressed data load.

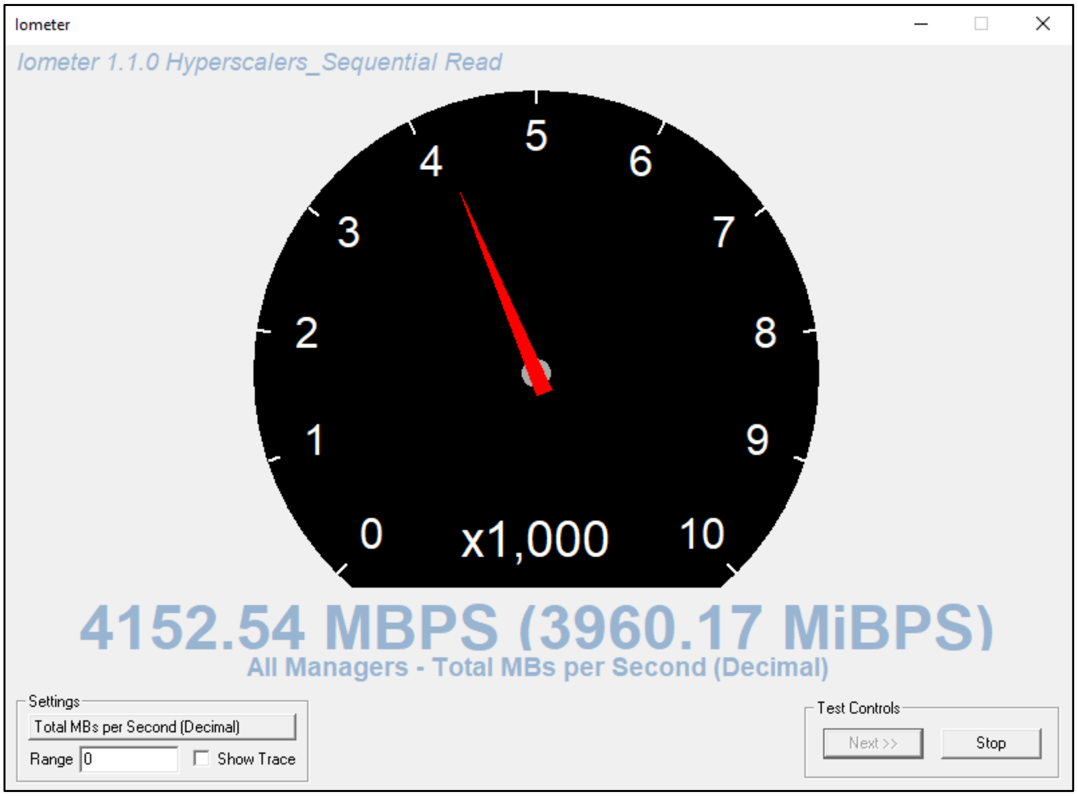


```
randwrite_j8_4k_qd128: (groupid=4, jobs=8): err= 0: pid=62930: Wed Oct  6 05:38:03 2021
  write: IOPS=433k, BW=1693MiB/s (1775MB/s)(5951GiB/3600012msec)
    slat (usec): min=4, max=20770, avg=14.59, stdev=50.75
    clat (usec): min=3, max=33921, avg=2345.74, stdev=656.13
     lat (usec): min=19, max=33929, avg=2360.50, stdev=659.11
    clat percentiles (usec):
     |  1.000th=[ 1795],  5.000th=[ 1926], 10.000th=[ 1991], 25.000th=[ 2114],
     | 50.000th=[ 2245], 75.000th=[ 2442], 90.000th=[ 2638], 95.000th=[ 2769],
     | 95.500th=[ 2802], 96.000th=[ 2835], 96.500th=[ 2900], 97.000th=[ 2966],
     | 97.500th=[ 3130], 98.000th=[ 3425], 98.500th=[ 3949], 99.000th=[ 5276],
     | 99.500th=[ 7439], 99.900th=[10290], 99.990th=[14746], 99.999th=[20841]
   bw (  KiB/s): min=64966, max=262260, per=9.89%, avg=171486.57, stdev=27466.93, samples=57592
   iops        : min=16241, max=65565, avg=42871.27, stdev=6866.73, samples=57592
  lat (usec)   : 4=0.01%, 10=0.01%, 20=0.01%, 50=0.01%, 100=0.01%
  lat (usec)   : 250=0.01%, 500=0.01%, 750=0.01%, 1000=0.01%
  lat (msec)   : 2=11.18%, 4=87.34%, 10=1.36%, 20=0.11%, 50=0.01%
  cpu          : usr=14.67%, sys=62.66%, ctx=600294948, majf=0, minf=112584125
  IO depths    : 1=0.1%, 2=0.1%, 4=0.1%, 8=0.1%, 16=0.1%, 32=0.1%, >=64=101.7%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     complete  : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.1%
     issued rwt: total=0,1560145320,0, short=0,0,0, dropped=0,0,0
     latency   : target=0, window=0, percentile=100.00%, depth=128
```

Figure 18 FIO random write test on individual drive

2:1 Compressed drive LV test

When the CSD is tested after creating a logical volume with two drives of its actual provisioned capacity, the sequential read speed reached to about 4.2GB/s and sequential write speed reached to about 3.2GB/s.



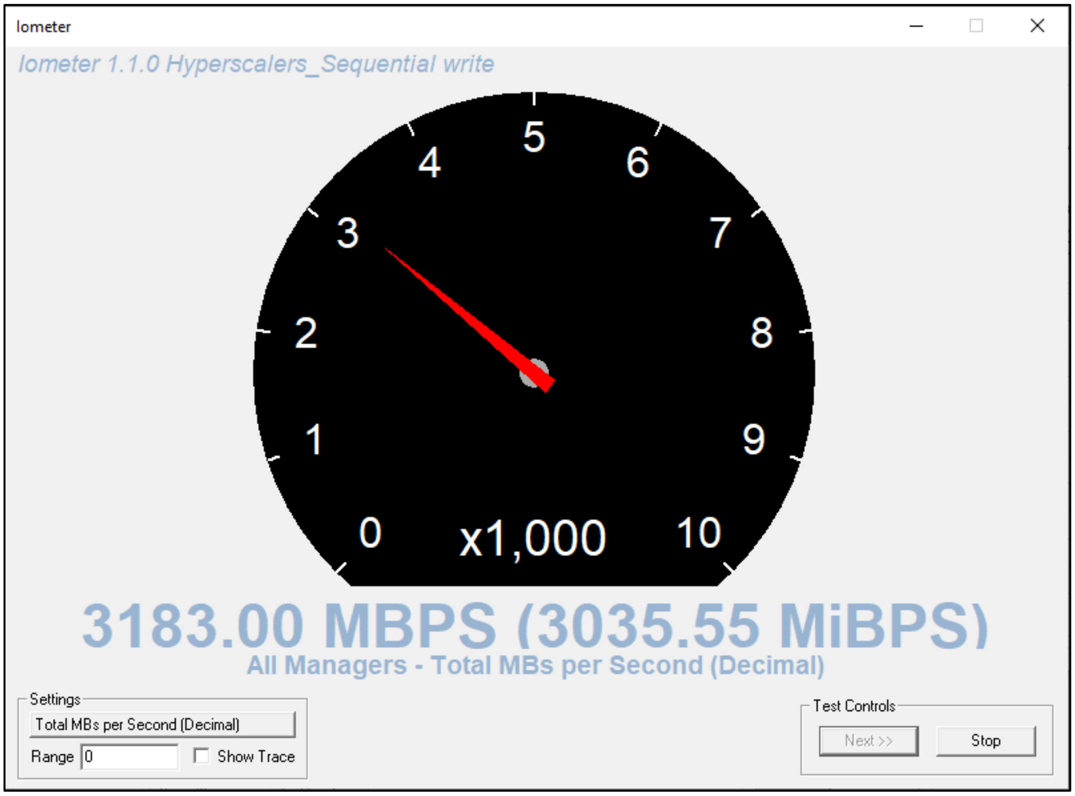
Iometer 1.1.0 Hyperscalers_Sequential Read

4152.54 MBPS (3960.17 MiBPS)
All Managers - Total MBs per Second (Decimal)

Settings
Total MBs per Second (Decimal)
Range 0 □ Show Trace

Test Controls

Figure 19 IOMeter Sequential read on two drive logical volume



Iometer 1.1.0 Hyperscalers_Sequential write

3183.00 MBPS (3035.55 MiBPS)
All Managers - Total MBs per Second (Decimal)

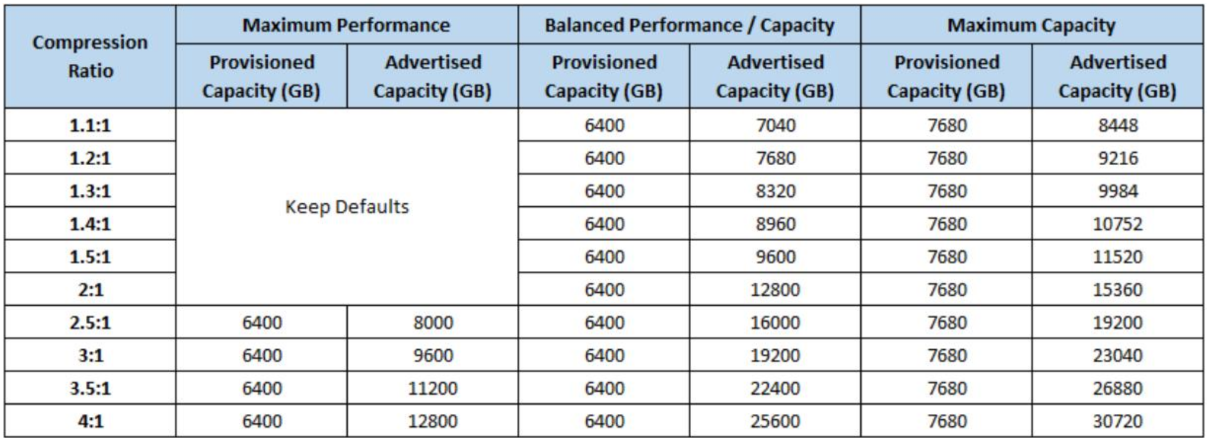
Settings
Total MBs per Second (Decimal)
Range 0 □ Show Trace

Test Controls

Figure 20 IOMeter Sequential write on two drive logical volume

Performance chart

Greater range of configuration vs performance in terms of drive capacity expansion along with data types of different compression capacity can be referenced from the below table.[2]

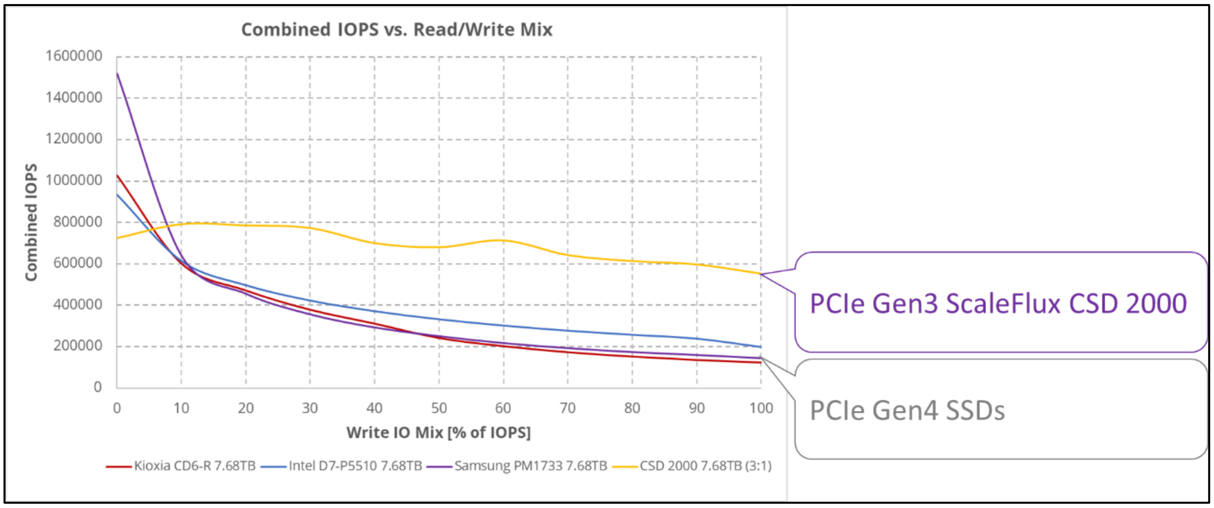


Compression Ratio	Maximum Performance		Balanced Performance / Capacity		Maximum Capacity	
	Provisioned Capacity (GB)	Advertised Capacity (GB)	Provisioned Capacity (GB)	Advertised Capacity (GB)	Provisioned Capacity (GB)	Advertised Capacity (GB)
1.1:1	Keep Defaults		6400	7040	7680	8448
1.2:1			6400	7680	7680	9216
1.3:1			6400	8320	7680	9984
1.4:1			6400	8960	7680	10752
1.5:1			6400	9600	7680	11520
2:1			6400	12800	7680	15360
2.5:1	6400	8000	6400	16000	7680	19200
3:1	6400	9600	6400	19200	7680	23040
3.5:1	6400	11200	6400	22400	7680	26880
4:1	6400	12800	6400	25600	7680	30720

Figure 21 performance vs drive expansion

Performance Comparison

ScaleFlux CSD 2000 with 3:1 compression is compared with other NVMe drives as shown in the graph below.[2]



PCIe Gen3 ScaleFlux CSD 2000

PCIe Gen4 SSDs

Figure 22 combined IOPs performance of 3:1 compressed drive comparison

8 COPYRIGHT AND LICENSING

9 REFERENCES

[1] "S5B TC | D52B-1U", *Hyperscalers.com*, 2022. [Online]. Available: https://www.hyperscalers.com/quanta-qct-server-1u/s5b-qantagrid-d52b-1u-all-nvme-ssd-hdd-storage-server-qct-buy-distributor-hp-proliant-dl360-dell-poweredge-r640-cisco-c220-m5-compare-usa. [Accessed: 02- May- 2022].

[2] "CSD 2000 - Product - ScaleFlux Computational Storage", *Scaleflux.com*, 2022. [Online]. Available: https://www.scaleflux.com/product/item/1002. [Accessed: 02- May- 2022].

[3] "Gluster", Gluster.org, 2022. [Online]. Available: https://www.gluster.org/. [Accessed: 02- May- 2022].

[4] Youtube.com, 2022. [Online]. Available: https://www.youtube.com/watch?v=sCtN7qIdsbo&ab_channel=ScaleFlux. [Accessed: 02- May- 2022].

[5] "Install - Gluster Docs", *Docs.gluster.org*, 2022. [Online]. Available: https://docs.gluster.org/en/latest/Install-Guide/Install/. [Accessed: 02- May- 2022].

Index

A

Access and Default Credentials..13
Addendum 23
Additional Setup and Deployment...................................19
Appliance Optimizer Utility AOU 4
Audience and Purpose 4

B

Base Product Deployment 14

C

Capacity Monitoring 19
Changing the Logical Sector Size......................................21
Compressed drive LV test 25

D

Data Wipe (Factory Reset) 22
Determine the Linux System Distribution........................11
Determining the Kernel Version 11
Digital IP Appliance Design Process 4
disable idling state 7
Documents, Knowledge Base, and Technical Support 4
Drive Management 21
Driver Install..10

E

Enabling Atomic Writes 21
Extended Capacity Gluster exposed pool testing.............18
External Stakeholders 6

I

Important Considerations..6
Individual drive test 23
Infrastructure Setup..9
Installing Binary Drivers 11
Installing Source Drivers 11
Internal Staff 6
Introduction 4

L

Logical Volume target for the Gluster File System 15

M

Memory requirements 9
Modifying Drive Capacity 14

N

Native Sector Size.. 8

O

operating temperature 8

P

Performance chart 27
Performance Comparison 27
power and profile configuration 6

Q

Qualification Hardware with the CSD.............................. 10

R

References... 28

T

Temperature Monitoring 20
Test Suite on single drive.. 23
Testing the Appliance 18
Trademarks and Licensing 27

V

Validation 12
Verify Supported Interfaces 9