Technical White Paper

Using Dell EMC PowerScale in an Equinix Cloud Connected Data Center

A detailed review

Abstract

This white paper discusses use cases, validates performance, and provides recommendations for running Dell EMC PowerScale in an Equinix Cloud Connected Data Center, attached to hyperscaler cloud providers.

August 2020

H18339

Revisions

Date	Description
July 2020	Initial release
August 2020	Updated with latest template

Acknowledgments

Author: Jeff Wiggins, Dell Technologies, Sr. Manager, Presales ANZ, India

Key contributors:

- Rajiv Juneja Dell Technologies
- Alex Seymour Dell Technologies
- Ryan Tassotti Dell Technologies
- Michael Yang Dell Technologies
- Brian Lepore Microsoft
- Jeff Tabor Microsoft
- Jer-Ming Chia Microsoft
- Gabriel Lageyre Equinix
- Lee Sharping Equinix
- Mischa Jampolsky Equinix
- William Lim Equinix

Table of contents

Executive summary

Public cloud infrastructure has changed the way IT organizations deploy, manage, and consume IT. For some, public cloud is something to explore as a potential location for segments of their IT workloads. However, having workloads in the public cloud can raise issues that are not applicable to an on-premises architecture, such as:

- Data sovereignty: Where is my data stored, and who has access to it?
- Cost predictability: What are my future public cloud charges?
- Performance predictability: How do I maintain consistent performance?
- Flexibility and choice: How do I use more than just one public cloud?

The public cloud conversation will explore taking a customer's on-premises workload (including compute and storage) and moving it in its entirety into a public cloud provider's infrastructure, which can raise the above issues for some customers. The purpose of this white paper is to validate an alternative approach, taking advantage of the compute benefits that public cloud provides while housing the customer's data in a managed service provider facility such as Equinix. In this case, the customer's data is housed on a Dell EMC PowerScale storage system within an Equinix facility with direct access to the compute instances in the public cloud addressing the issues and concerns listed above. The Dell EMC PowerScale family of storage solutions includes Isilon and PowerScale storage nodes and PowerScale OneFS operating system.

This white paper will outline the use cases of public cloud compute instances connected to a Dell EMC PowerScale storage array located in an Equinix cloud connected managed service provider facility. Furthermore, it will provide storage performance benchmarks for use cases hosted in the Microsoft Azure cloud and AWS cloud along with recommendations.

1 The case for high-performance file storage in a public cloud connected managed service provider facility

There are three commonly used architectures for cloud infrastructure:

- Private cloud (on premises): The customer's workloads run on private infrastructure.
- Public cloud: The customer's workloads are housed offsite on another party's infrastructure.
- Hybrid/Multi cloud: Some of the customer's workloads are in public cloud, and some workloads are in private cloud.

There are also mixed architectures where a customer will have their compute instances in public cloud, and their data (storage) housed privately in a public cloud connected facility of a managed service provider, such as Equinix.

1.1 Use cases

The demands of high performance or scale-out file storage are typically classified but not limited to the following use cases:

- **Data sovereignty**: A customer wants to take advantage of the economics of public cloud compute services and application, however, is mandated to keep data stored within a geographical location.
- **Disaster recovery**: A customer has an on-premises storage array and wants to enable disaster recovery by performing storage replication to an offsite storage array. If this offsite storage array is hosted in a managed service provider facility, such as Equinix, then public cloud compute instances can be spun up and attached to the offsite storage replicas in a disaster recovery scenario.
- **Predictable TCO**: Storage can be the highest cost component of public cloud. Hosting a storage array privately in a managed service provider facility, such as Equinix, provides a fixed cost for storage and reduced ingress and egress charges while leveraging public cloud compute services and applications.
- **Big data and Analytics**: Organizations are often looking to leverage new applications and tools available to analyse datasets by connecting to hyperscale compute resources from the Microsoft Azure cloud. Data generated both inside and outside of the cloud provider can be leveraged to create inputs for analytics workloads.

1.2 Why validate by testing?

Many customers using an on-premises private cloud architecture run workloads that are sensitive to performance. If a customer has a throughput-based workload that is important to their business or operation, the probability this customer will expect consistent predictable storage performance is high.

2 Microsoft Azure architecture and testing

This section will outline the Microsoft (MS) Azure architecture as tested, and detail the tests performed.

2.1 Architecture

The high-level architecture is shown in Figure 1. The architecture consists of:

- Dell EMC Isilon (now part of the Dell EMC PowerScale family) storage array, residing in a public cloud connected facility (Equinix);
- Dell EMC 10 Gb networking switch, also in a public cloud connected facility (Equinix);
- 2 Gb MS Azure ExpressRoute connection between Equinix facility and MS Azure public cloud;
- Compute instances in Azure that connect to the Dell EMC Isilon storage array over SMB and NFS.

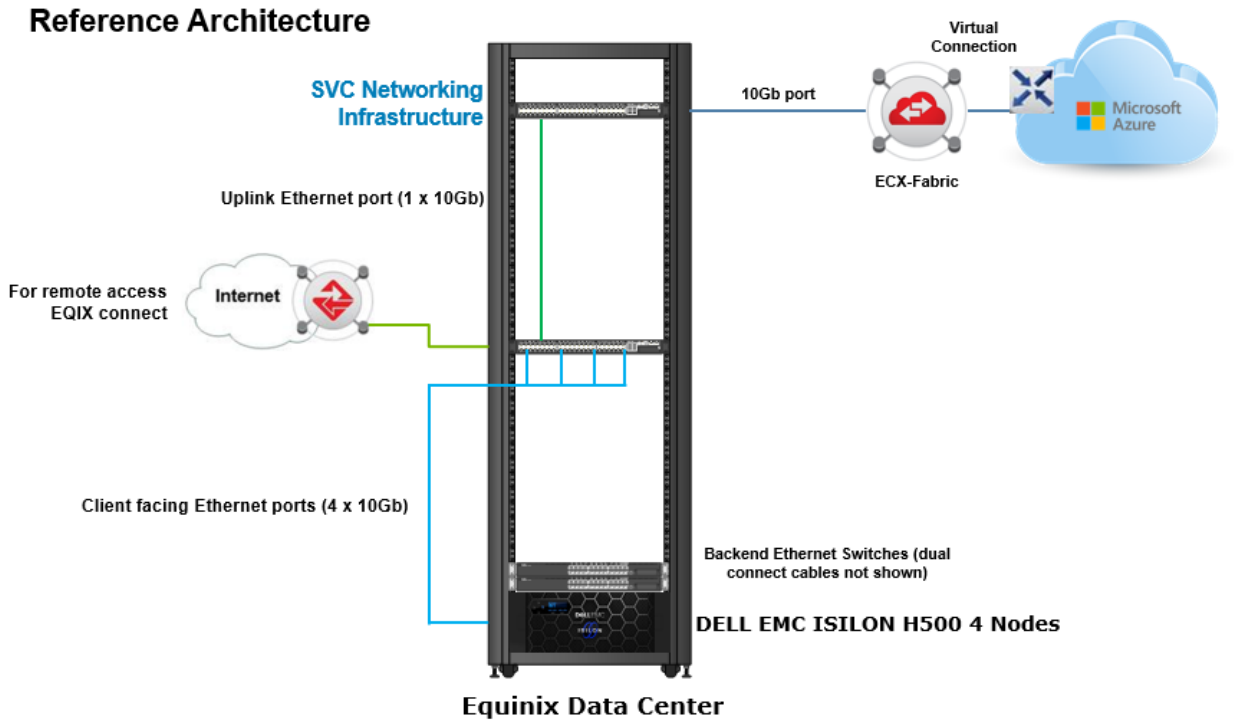


Figure 1 High-level network architecture

2.2 Azure testing configuration

This section will outline the as-tested configuration of the Azure compute instances, the Azure ExpressRoute connection, the Dell EMC PowerScale storage array, and the Dell EMC network switch.

Note: All Azure resources were deployed in the Australia East (Sydney) Region. The Dell EMC resources (Isilon storage array, network switch) were deployed at an Equinix facility in Sydney.

2.2.1 Azure Windows instance

The features of the chosen Microsoft Azure Windows instance are as follows:

- Standard D16s v3
- Windows Server 2019 Datacenter
- 64 GiB memory
- 16 vCPUs
- MTU = 1500

2.2.2 Azure Linux instance

The features of the chosen Azure Linux instance are as follows:

- Standard D16s v3
- CentOS (x86_64)
- lsb_release -d
- Description: CentOS Linux release 7.6.1810 (Core)
- 64 GiB memory
- 16 vCPUs
- uname -r
- 3.10.0-957.21.3.el7.x86_64
- MTU = 1500

2.2.3 Azure ExpressRoute

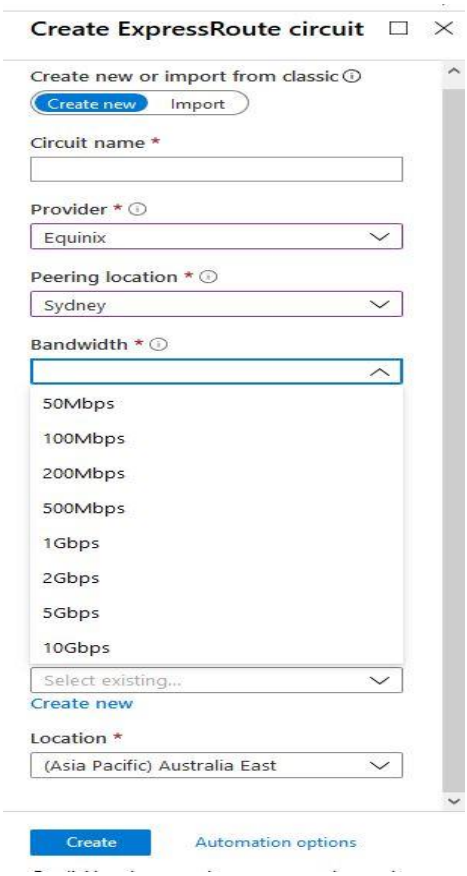
When located in an Equinix managed service provider facility, there are two ways to connect the Isilon storage array to the MS Azure compute instances:

- Equinix Cloud Exchange Fabric; or
- Azure ExpressRoute

Equinix Cloud Exchange Fabric

The major benefit of connecting into public cloud compute using Equinix Cloud Exchange Fabric™ (ECX Fabric™) is the availability of connections into all the major hyperscale providers including Amazon AWS, Microsoft Azure, Google Cloud Platform, Alibaba Cloud. ECX Fabric connects to the provider network over a single connection out of the customer's co-located network infrastructure. For example, the Isilon storage array connects to a network switch that directly connects to ECX Fabric. From there, virtual connections to the individual public cloud providers are securely established and dynamically managed using a self-service portal or API. The screen capture below shows the connection setup to ECX Fabric. More information can be found here.

> *ECX Fabric brings together cloud service providers and users, enabling them to establish affordable, private, high-performance connections within Platform Equinix®* (Equinix).



Azure ExpressRoute

ExpressRoute connects your private IT infrastructure with the Microsoft cloud over a private, high-speed connection providing access to such services as Microsoft Azure, Office 365, and Dynamics 365.

With ExpressRoute, establish connections at a cloud connected managed service provider facility, such as Equinix, or directly connect from your existing WAN network provider. ExpressRoute does not go over the public Internet offering more reliability, faster speeds, consistent latencies, and higher security than typical connections over the Internet. For information about how to connect your private IT infrastructure to the Microsoft cloud using ExpressRoute, see ExpressRoute connectivity models.

Pricing for ExpressRoute connections should be selected based on the expected bandwidth and data transfer requirements. The full pricing options are available online from the Azure ExpressRoute pricing page. Microsoft Azure offers plans which require charges per data transferred as well as fixed monthly options providing unlimited data transfers are made available for selection based on region.

2.2.4 Dell EMC PowerScale Storage

The Dell EMC Isilon H500, a part of the Dell EMC PowerScale storage family, is a versatile hybrid scale-out NAS platform that delivers high throughput and scalability. It is designed to provide organizations with a highly efficient and resilient scale-out storage platform which can support a wide range of enterprise file workloads and applications. The Isilon H500 is ideal for unstructured data storage use cases needing operational flexibility and the right balance of large capacity and performance. The Isilon H500 runs on the same powerful yet simple PowerScale OneFS operating system as other PowerScale storage systems and allows for simple integration with existing PowerScale clusters.

Available in several configurations, the Isilon H500 delivers 5 GB/s bandwidth per chassis and provides capacity options of 2 TB, 4 TB or 8 TB HDDs. The Isilon H500 houses 60 HDDs per 4U chassis ranging from 120 TB to 480 TB per chassis and scalable to 30 PB of capacity in a single Isilon cluster. Each H500 node supports 10 GbE or 40 GbE dedicated front-end client network options and InfiniBand QDR or 40 GbE internal backend networking options.

The configuration of the tested Dell EMC PowerScale storage array is as follows:

Dell EMC Isilon H500 4 node storage array (per node)
2.2 GHz 10-Core
128 GB memory
15 x 4 TB HDD | 1 x 1.6 TB SSD
2 x 10 GbE (Front end)
2 x QSFP+ 40 Gb Ethernet (internal backend network)
OneFS version 8.1.2
Patches

- 8.1.2.0_KGA-RUP_2019-06_250808
- 8.1.2.0_UGA-RUP_2019-06_250806

Protection Policy - +2d:1n
Pool Usable Capacity - 178 TB / 162 TiB @ 100% utilization
NFS v3 mount options (default OneFS values) –
`rw,noatime,vers=3,rsize=131072,wsize=524288,namlen=255,hard,proto=tcp,timeo=600,retrans=2`

Four separate NFS mount points created for mounting on each Linux VM

```
>  isilon-node1:/mount1     222T   23G  215T   1% /mnt/mount1
>  isilon-node2:/mount2     222T   23G  215T   1% /mnt/mount2
>  isilon-node3:/mount3     222T   23G  215T   1% /mnt/mount3
>  isilon-node4:/mount4     222T   23G  215T   1% /mnt/mount4
```

SMB v3
Four separate SMB shares created for mounting on each Windows VM

```
S:       \\172.16.0.5\smb1       Microsoft Windows Network
X:       \\172.16.0.6\smb2       Microsoft Windows Network
Y:       \\172.16.0.7\smb3       Microsoft Windows Network
W:       \\172.16.0.8\smb4       Microsoft Windows Network
```

Jumbo frames disabled
Six rack units (4U chassis + 2u backend network)

2.2.5 Dell EMC Network Switch

The S4148F-ON network switch is a one rack unit (1U), full-featured top-of-rack (ToR) 10/25/40/50/100GbE switch for 10G servers with small form-factor pluggable plus (SFP+), quad small form-factor pluggable plus (QSFP+), and quad small form-factor pluggable (QSFP28) ports.

Dell EMC Networking OS10 Enterprise Edition is a network operating system (OS) supporting multiple architectures and environments. As the networking world is moves away from a monolithic stack to a pick-your-own-world, the OS10 solution allows disaggregation of the network functionality. More information can be found here.

The configuration of the tested network switch is as follows:

- Model: S4148F-ON
- OS10 version: 10.4.2.0
- Connectivity to Unity storage array: 10 Gb TwinAx
- Connectivity to Direct Connect router: 10 Gb LR SFP+
- Jumbo frames disabled

2.3 Testing Methodology

The primary tool used for this benchmark testing is vdbench. Vdbench is a command-line utility created to generate disk I/O workloads for validating storage performance and storage data integrity. Iperf and ping tests were conducted before the benchmark testing to validate network bandwidth throughout the test infrastructure.

There are four main test categories used to validate the viability of the Dell EMC PowerScale solution in an Equinix managed datacenter.

- Iperf network bandwidth tests (bi-directional)
- Network latency (aka ping tests)
- Windows SMB benchmarking
- Linux NFSv3 benchmarking

2.3.1 Iperf network bandwidth (https://iperf.fr/)

iPerf is a tool for active measurements of the maximum achievable bandwidth on IP networks. iPerf2 is preinstalled on the OneFS operating system and should be leveraged to measure network performance before running any performance benchmark. Network bandwidth should be measured to set maximum performance expectations from the client and server network to the Isilon nodes. iPerf 2.0.9 was downloaded and used for the client systems on both Linux and Windows.

Iperf testing was performed from each Linux, Windows, and OneFS nodes to illustrate and test the read and write bandwidth over the ExpressRoute 10 Gbps network link. Initial testing provided some inconsistent results which lead to hardware changes of one network cable and one 10 Gbps SFP+ module.

Based on the Microsoft Azure VM definitions the maximum theoretical bandwidth per VM is 8 Gbps.

Start the iperf server on a OneFS environment.

isi_for_array iperf -s

Bidirectional testing with OneFS acting as the client pointing to a VM running the server instance of iPerf.

```
# iperf -c 10.10.0.205 -i 5 -t 60 -P 4
```

Start the iperf client on a Linux VM connecting to one of the Isilon nodes.

iperf -c 172.16.0.5

Start the iperf client on a Windows VM connecting to one of the Isilon nodes is the same command issued from the cli.

C:\Users\pocadmin\Downloads\iperf-2.0.9-win64\iperf-2.0.9-win64>iperf.exe -c 10.10.0.196

Summary of results:

Below is an example command from one Linux VM to one Isilon node. Testing was repeated from each VM to each Isilon node in the cluster to validate results and consistent network performance. Using the Isilon nodes as the server the bandwidth tested to ~ 7.21 Gbps per VM. (*VM limit is 8.0 Gbps)

Linux VM to Isilon node 1

```
# iperf -c isilon-node1 -i 5 -t 60 -P 4

------------------------------------------------------------

Client connecting to isilon-node1, TCP port 5001

TCP window size: 94.5 KByte (default)

------------------------------------------------------------

[  4] local 10.10.0.205 port 44506 connected with 172.16.0.5 port 5001

[SUM]  0.0-60.0 sec  50.3 GBytes  7.20 Gbits/sec
```

Two Linux VMs were also testing running iperf in parallel to maximize the ExpressRoute network link.

This test was dual writes from Linux VMs to separate Isilon nodes with iperf.

```
[pocadmin@Linux64GB16c-3 ~]$ iperf -c isilon-node3 -i 5 -t 40 -P 4
[SUM]  0.0-40.0 sec  22.5 GBytes  4.83 Gbits/sec

[pocadmin@linux-vm2 ~]$ iperf -c isilon-node2 -i 5 -t 40 -P 4
[SUM]  0.0-40.0 sec  22.1 GBytes  4.75 Gbits/sec
```

Looking at results of the iperf tests, writes appear to split evenly from the VMs to the Isilon nodes and saturate the bandwidth of the ExpressRoute.

2.3.2 Network Ping Testing

Using the ping command to test overall network latency between Linux and Windows VMs to the Isilon nodes.

Linux VM1 (Azure Cloud) to Isilon node 1 (over ExpressRoute network)

```
[pocadmin@Linux64GB16c ~]$ ping isilon-node1
PING isilon-node1 (172.16.0.5) 56(84) bytes of data.
64 bytes from isilon-node1 (172.16.0.5): icmp_seq=1 ttl=61 time=2.33 ms
64 bytes from isilon-node1 (172.16.0.5): icmp_seq=2 ttl=61 time=1.97 ms
64 bytes from isilon-node1 (172.16.0.5): icmp_seq=3 ttl=61 time=2.10 ms
64 bytes from isilon-node1 (172.16.0.5): icmp_seq=4 ttl=61 time=2.16 ms
```

Windows VM1 to Linux VM1 hosted in Azure cloud

```
C:\Users\pocadmin>ping linux-vm1
Pinging linux-vm1 [10.10.0.196] with 32 bytes of data:
Reply from 10.10.0.196: bytes=32 time<1ms TTL=64
Reply from 10.10.0.196: bytes=32 time<1ms TTL=64
Reply from 10.10.0.196: bytes=32 time<1ms TTL=64
Reply from 10.10.0.196: bytes=32 time<1ms TTL=64
```

2.3.3 Windows SMB3 Benchmarking

To test SMB throughput performance, vdbench is used to illustrate the impact of reads and writes to a 1 GB file. Vdbench parameters were specified to generate multiple threads and multiple IO transfer sizes.

The four tests were performed for each of the following IO transfer sizes: 64 k, 128 k, 256 k, 512 k, and 1024 k. The parameters below show an example of the Windows disk benchmarking commands using the 1024 k (1 M) transfer size.

```
fsd=throughput_$host_dir1,anchor=Z:\\$host\throughput,depth=2,width=10,files=10,
size=1G

fwd_tpr_$host,fsd=throughput_$host_*,operation=read,threads=20,xfersize=1M

fwd=fwd=fwd_tpw_$host,fsd=throughput_$host_*,operation=write,threads=10,xfersize
=1M,openflags=directio
```

This test will run the following benchmark, the `xfersize` variable **determines the IO transfer size**.

- Create folder structure with 2 subdirectories, 10 parent directories, each containing ten 1 GB files
- Read the files with 20 threads
- Writes to the files with 10 threads, `directio` flag enabled to eliminate cache

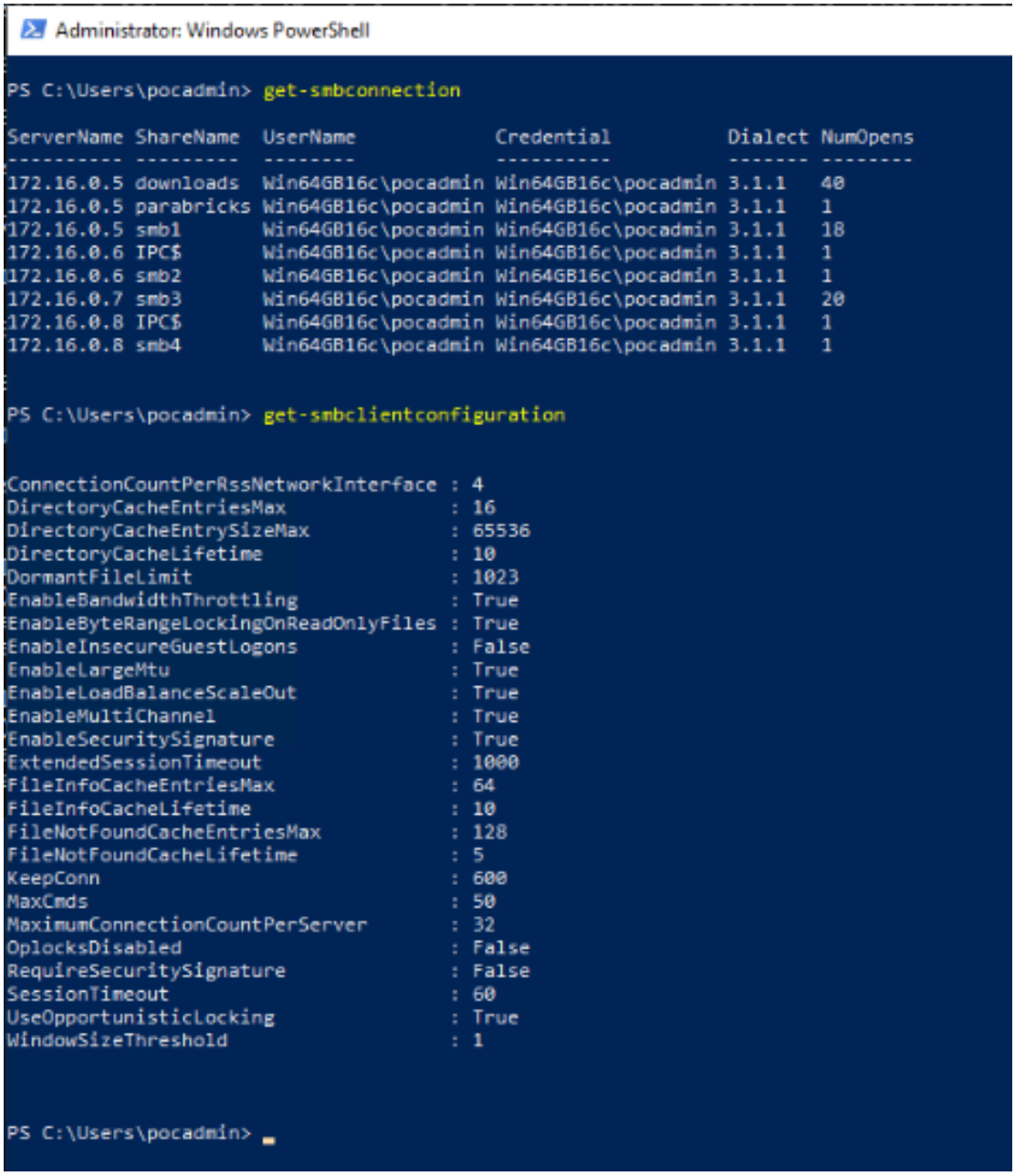
Note: There are five iterations of each test run for reads and writes. The value selected is the average for each iteration.

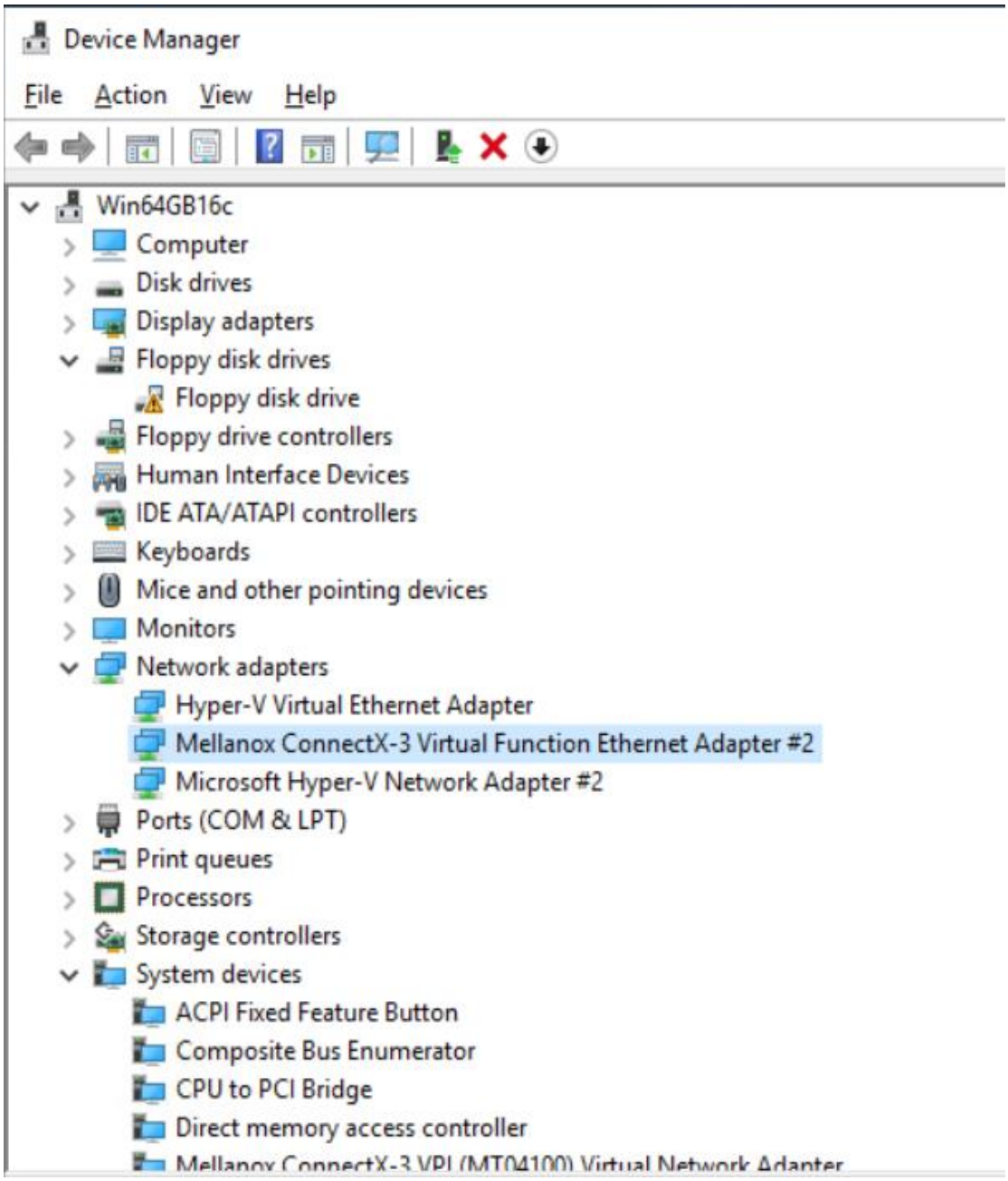
Additional notes about these tests and results:

- Each test iteration is run five times at random times throughout a 24-hour period.
- Each test iteration reads and writes the file based on vdbench parameters.
- Testing alternates between read test and write testing for each run.
- The five overall results presented are averaged and graphed in the results section

Windows SMB settings

Settings for SMB version and multi-channel connections are shown below by using Windows PowerShell.






```
PS C:\Users\pocadmin> get-smbmultichannelconnection

Server Name Selected Client IP   Server IP  Client Interface Index Server Interface Index Client RSS Capable Client RDMA Capable
----------- -------- ---------   ---------  ---------------------- ---------------------- ------------------ ----------
172.16.0.5  True     10.10.0.197 172.16.0.5 10                     1                      True               False
172.16.0.7  True     10.10.0.197 172.16.0.7 10                     1                      True               False
172.16.0.6  True     10.10.0.197 172.16.0.6 10                     1                      True               False
172.16.0.8  True     10.10.0.197 172.16.0.8 10                     1                      True               False
```

2.3.4 Linux NFSv3 Benchmarking

To test NFS throughput performance, vdbench is used to illustrate the impact of reads and writes to a 1 GB file. Vdbench parameters were specified to generate multiple threads and multiple IO transfer sizes.

The four tests were performed for each of the following IO transfer sizes: 64 k, 128 k, 256 k, 512 k, and 1024 k. The parameters below show an example of the Windows disk benchmarking commands using the 1024 k (1 M) transfer size.

fsd=throughput_$host_dir1,anchor=/mnt/mount1/throughput/$host,depth=2,width=10,files=10,size=1G

fwd=fwd_tpr_$host,fsd=throughput_$host_*,operation=read,threads=20,xfersize=1M

fwd=fwd_tpw_$host,fsd=throughput_$host_*,operation=write,threads=20,xfersize=64k,openflags=directio

This test will run the following benchmark, the `xfersize` variable determines the IO transfer size.

- Create folder structure with 2 subdirectories, 10 parent directories, each containing ten 1 GB files
- Read the files with 20 threads
- Writes to the files with 10 threads, `directio` flag enabled to eliminate cache

---

Note: There are five iterations of each test run for reads and writes. The value selected is the average of for each iteration.

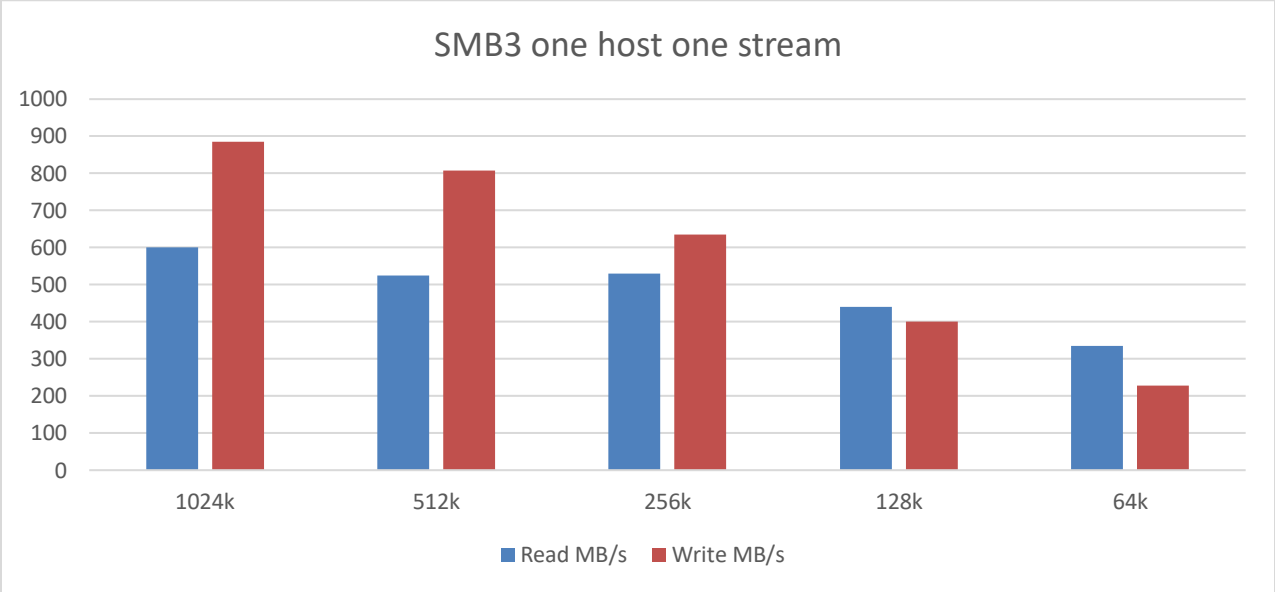
Additional notes about these tests and results:

- Each test iteration is run five times at random times throughout a 24-hour period.
- Each test iteration reads and writes the file based on vdbench parameters.
- Testing alternates between read test and write testing for each run.
- The five overall results presented are averaged and graphed in the results section

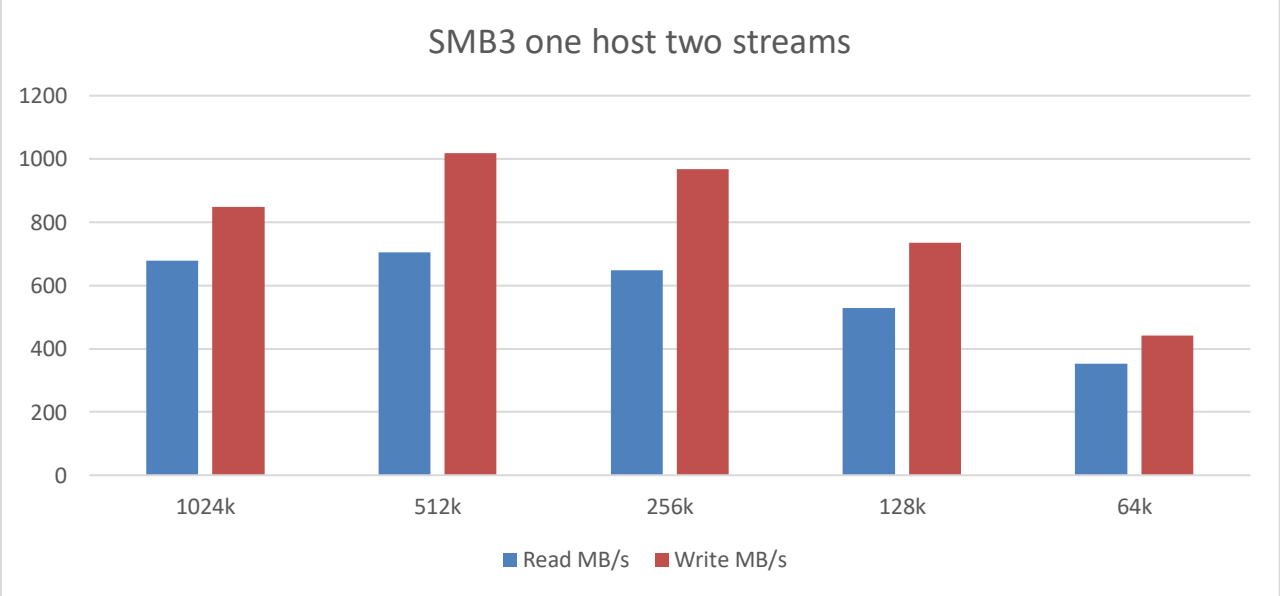
3 Microsoft Azure testing results

This section provides the results for running the various vdbench outputs for both Windows and Linux VMs connecting to Dell EMC PowerScale/Isilon storage.

3.1 Windows single host single stream benchmarking results

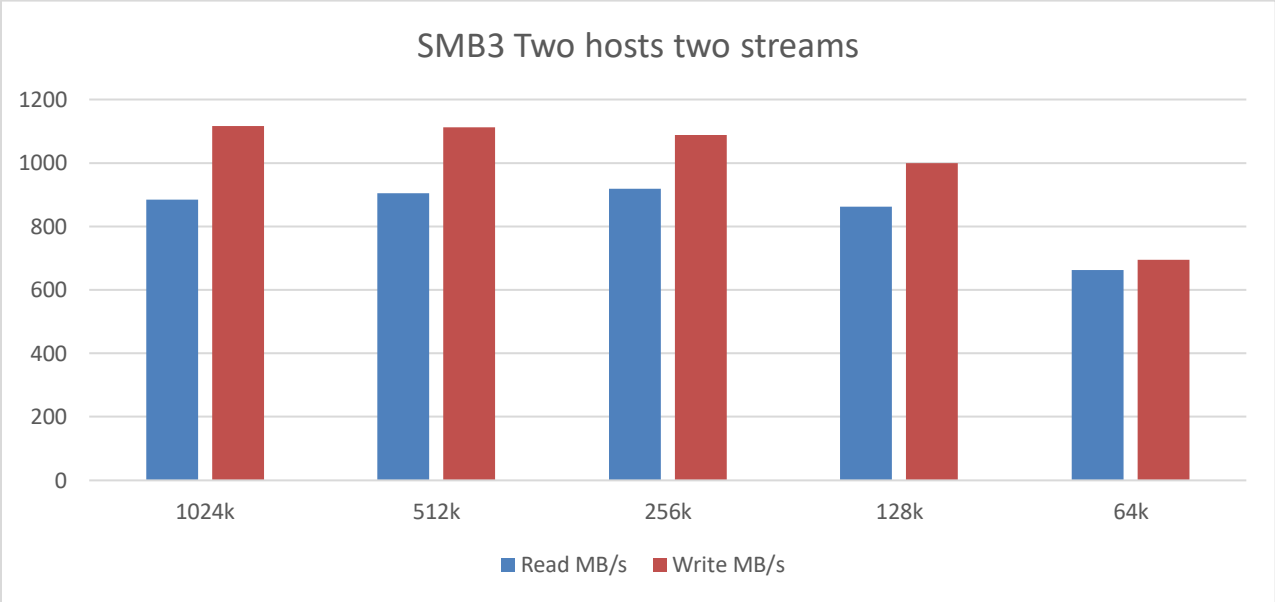


3.2 Windows single host two streams benchmarking results



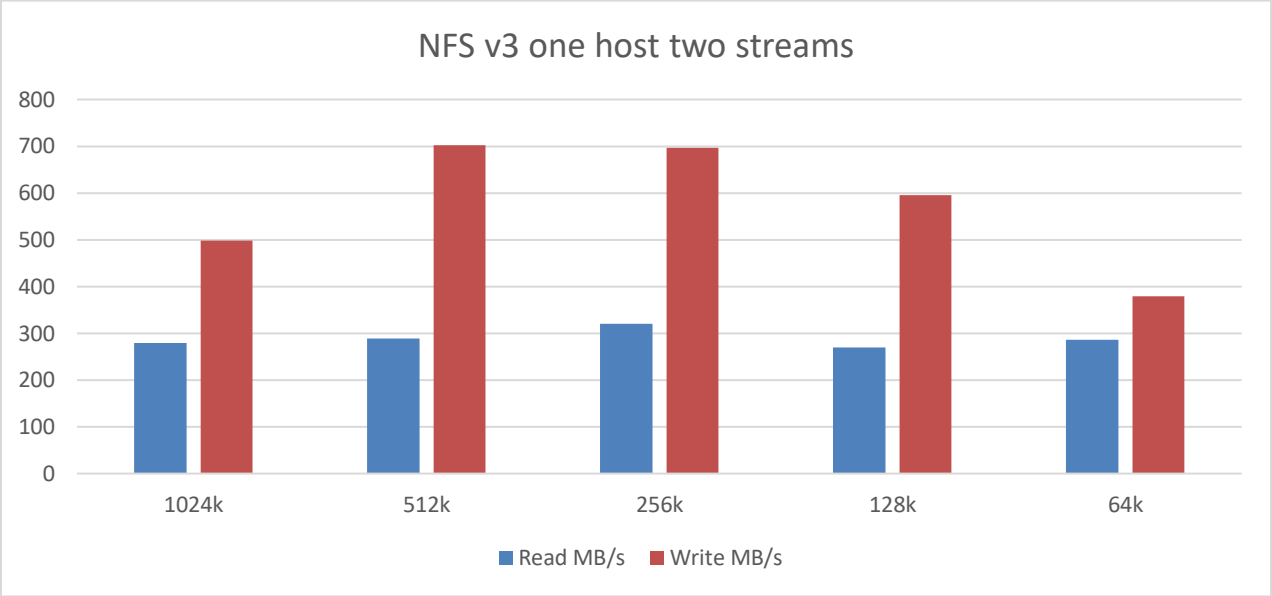
The two write streams tests reached max bandwidth of the ExpressRoute link. Two multi-channel SMB3 connections were able to exceed the single VM specification of 8Gbps.

3.3 Windows – Two hosts two total streams benchmarking results



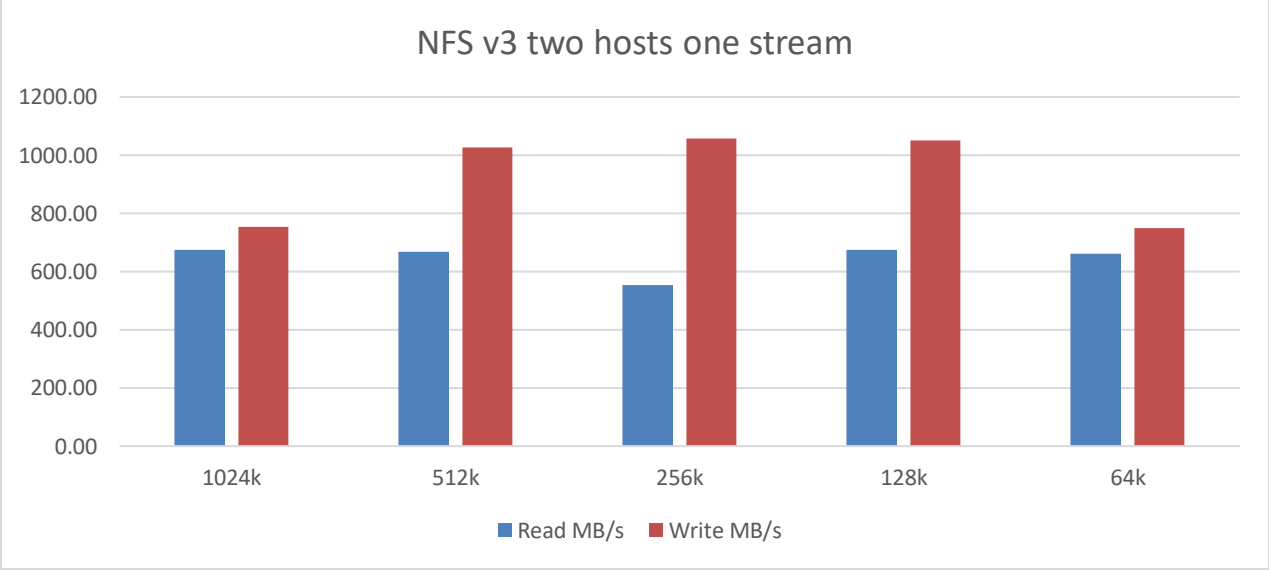
This table shows two Windows VMs could reach the ExpressRoute bandwidth limit with multiple IO sizes.

3.4 Linux – One host two streams benchmarking results



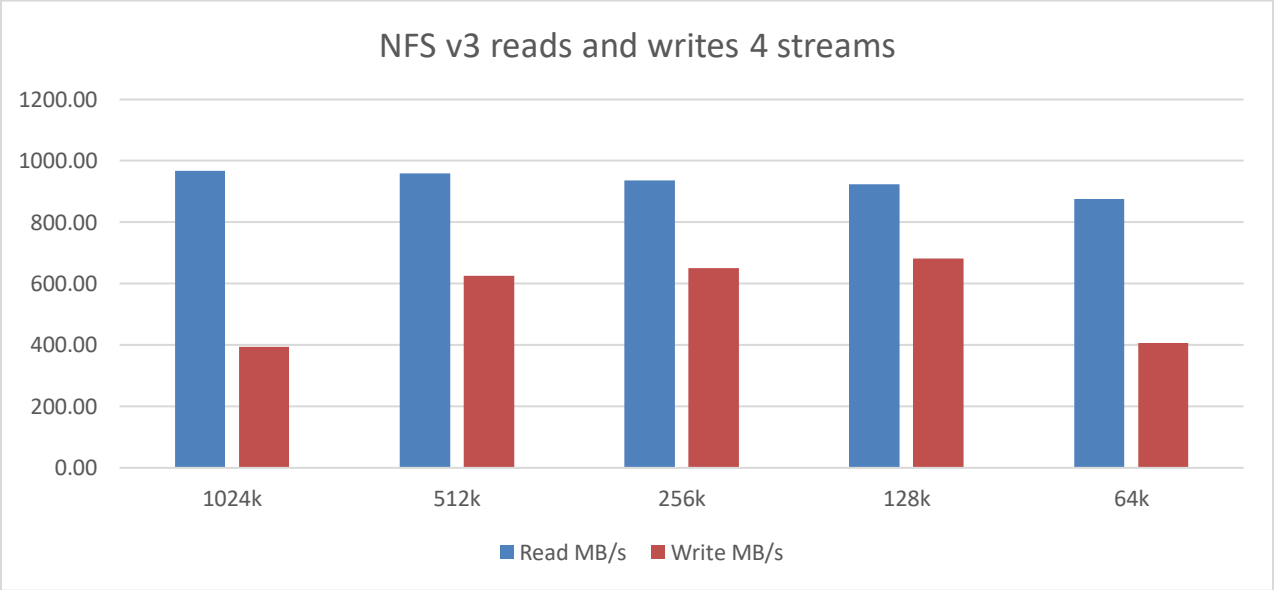
A single Linux VM with NFSv3 produced higher write performance than read performance when testing.

3.5 Linux – Two hosts one stream each benchmarking results



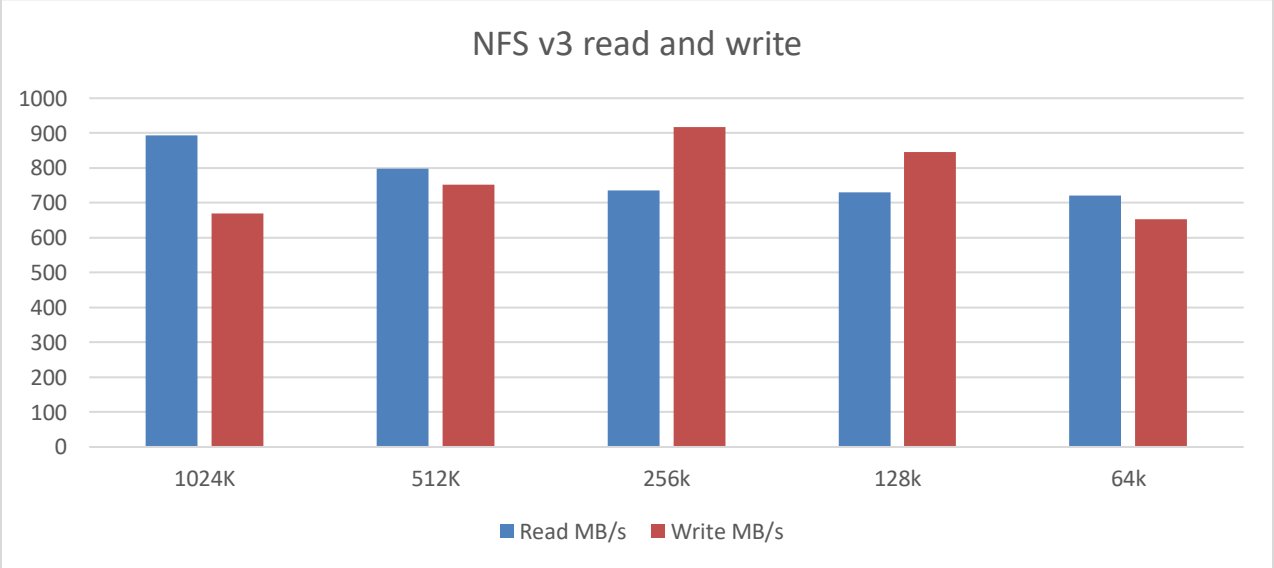
Two Linux VMs writing data to two separate Isilon nodes could saturate the ExpressRoute link with multiple IO transfer sizes from 512k to 128k.

3.6 Linux one host 4 streams benchmarking results



A single Linux VM reading and writing data to four separate Isilon nodes is limited by the virtual machine's network bandwidth with multiple IO transfer sizes from 512k to 128k.

3.7 Linux three hosts one stream benchmarking results



Three Linux VMs provided aggregate bandwidth utilization for both read and write workloads. Overall, IO sizes at 256 K provided the highest write performance while read performance benefited from larger IO sizes to achieve maximum bandwidth.

4 Vertical industry testing

Dell EMC PowerScale storage, which includes PowerScale and Isilon nodes, is the leading NAS solution across a large group of vertical industries collectively known as Commercial HPC. Two such industries are Media and Entertainment (M&E) and Healthcare Life Sciences (HLS). These industries demand scalable and high-throughput file storage that the Isilon solution provides, they also demand flexible and high-performance compute farms that can be built in the Microsoft Azure cloud. The purpose of this section is to focus on demanding applications within these industries and prove they can be delivered on an architecture with Isilon in a managed service provider facility, such as Equinix, which is nearby an Azure datacenter.

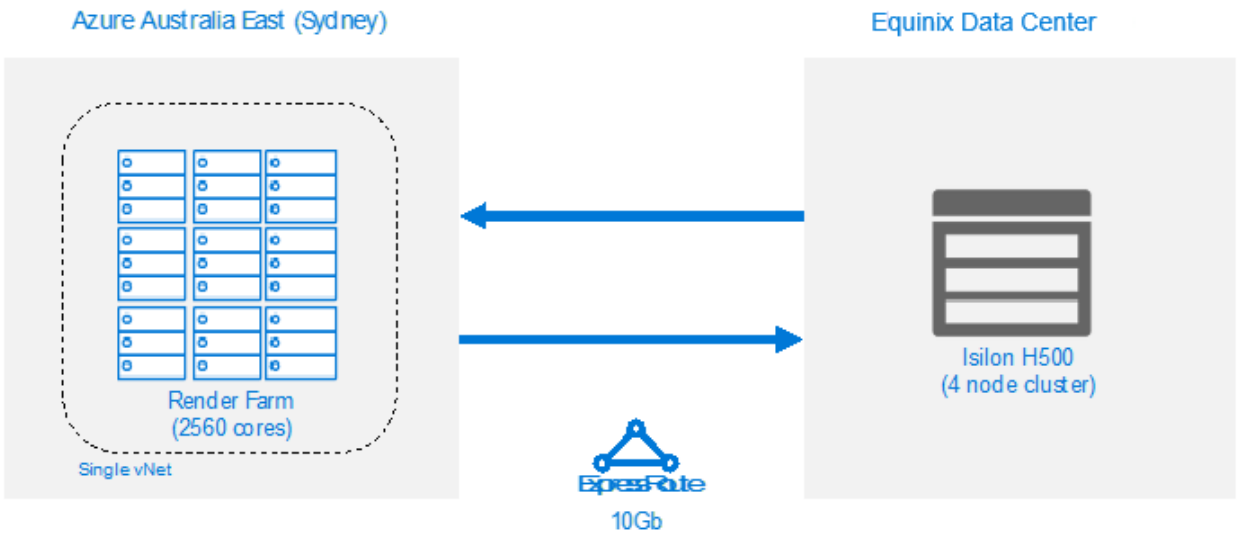
4.1 Rendering in the M&E industry

Making modern movies is a technical process that includes high demand for both storage and compute servers. Rendering is one of the most taxing applications in the post-production workflow. Rendering is the process to create scenes that cannot be filmed (such as a dragon) or the director has chosen not to film for cost savings (for example a sand storm). Over the past ten plus years, nearly all the top blockbuster movies have used rendering to create special effects.

Rendering requires scalable and high-throughput file storage that Isilon systems can deliver. Rendering also requires scaling a compute farm (a.k.a. render farm) used for the rendering process which is composed of thousands of CPU cores that Azure compute cloud can deliver. The system architecture tested in this whitepaper uses an ExpressRoute 10 Gb link to connect the render farm in Azure to the Isilon storage located in an Equinix cloud connected facility. The goal of this testing is to prove these two pieces of infrastructure separated by an ExpressRoute 10 Gb link can deliver successful rendering.

4.2 Rendering Test Plan and Results

The rendering test infrastructure is shown in the following diagram:



The rendering test plan consisted of the following steps:

1. Download Movie Clip source files to the Isilon H500 storage system located in the Equinix cloud connected managed service provider facility
2. Create a 2560-core render farm in Azure compute using qty = 160 DS16 Linux Virtual Machines (VMs)
3. Connect the 2560-core render farm to the Isilon H500 using an Ultra Performance VNet Gateway and an ExpressRoute 10 Gb link
4. Run rendering steps a-c in parallel across the cores used and output frames rendered*
    a. Read sources files from Isilon H500 to cores
    b. Render output frames
    c. Write output frames back to Isilon H500
5. Measure and record the Wall Clock Time from the first read to the last write
6. Repeat steps 4 and 5 multiple times looking for optimal number of cores and output frames for the ExpressRoute 10 Gb link

The rendering test results are below. The number of tasks in column 3 is derived from cores used and output frames rendered.

Job Name	Wall Clock Time	Nodes, Cores, Tasks	Throughput	Storage Source	Render Job notes
Job 64	5 minutes	160 D16 nodes, 2560 tasks	9.4 Gbps	Direct	Movie Clip, separate output directory
Job 65	4.5 minutes	160 D16 nodes, 2560 tasks	9.8 Gbps	Direct	Movie Clip, separate output directory
Job 66	4.5 minutes	160 D16 nodes, 2560 tasks	6.1 Gbps	Direct	Movie Clip, separate output directory
Job 67	3.5 minutes	160 D16 nodes, 1280 tasks	4.5 Gbps	Direct	Movie Clip, separate output directory
Job 68	3.5 minutes	160 D16 nodes, 1280 tasks	7.1 Gbps	Direct	Movie Clip, separate output directory
Job 69	5.5 minutes	160 D16 nodes, 1280 tasks	7.1 Gbps	Direct	Movie Clip, separate output directory
Job 70	3.5 minutes	160 D16 nodes, 1280 tasks	5 Gbps	Direct	Movie Clip, separate output directory
Job 71	4.5 minutes	160 D16 nodes, 1920 tasks	5.9 Gbps	Direct	Movie Clip, separate output directory
Job 72	4.5 minutes	160 D16 nodes, 1920 tasks	9.3 Gbps	Direct	Movie Clip, separate output directory
Job 73	3.5 minutes	160 D16 nodes, 1920 tasks	6.8 Gbps	Direct	Movie Clip, separate output directory
Job 74	4.5 minutes	160 D16 nodes, 1600 tasks	6.8 Gbps	Direct	Movie Clip, separate output directory
Job 75	4.5 minutes	160 D16 nodes, 1600 tasks	7.4 Gbps	Direct	Movie Clip, separate output directory
Job 76	4 minutes	160 D16 nodes, 1600 tasks	6.3 Gbps	Direct	Movie Clip, separate output directory
Job 77	5 minutes	160 D16 nodes, 2240 tasks	7.9 Gbps	Direct	Movie Clip, separate output directory
Job 78	4.5 minutes	160 D16 nodes, 2240 tasks	9.8 Gbps	Direct	Movie Clip, separate output directory
Job 80	4.5 minutes	160 D16 nodes, 2240 tasks	9 Gbps	Direct	Movie Clip, separate output directory

Takeaways from rendering results:

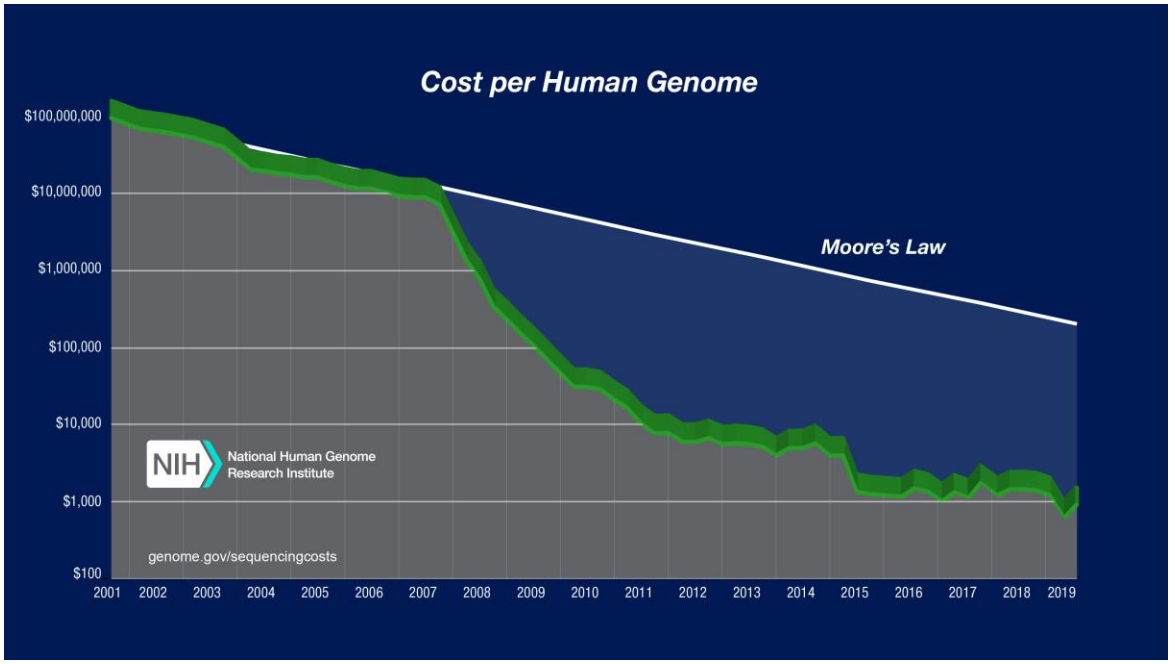
- Scaling rendering to 2560 cores and 2560 output frames is achievable
- None of the rendering jobs fully saturated the ExpressRoute 10 Gb link
- The Wall Clock Time required was consistent (3.5 to 5.5 minutes range)
    a. Average duration of 4.5 minutes

4.3 Rendering conclusions

Consistently good performance for rendering can be delivered when running on a render farm built of Azure compute cores over an ExpressRoute 10 Gb link to an Isilon H500 storage system in an Equinix facility. At the maximum tested configuration of 2560 cores and 2560 output frames, rendering is successful, and the ExpressRoute 10 Gb link was mostly saturated. Since the ExpressRoute 10 Gb link was not fully saturated, the quantity of cores and output framed can likely increase (10-20%). If the goal is to substantially increase (30-100%) the quantity of cores and output frames, then we recommend adding a second ExpressRoute 10 Gb link.

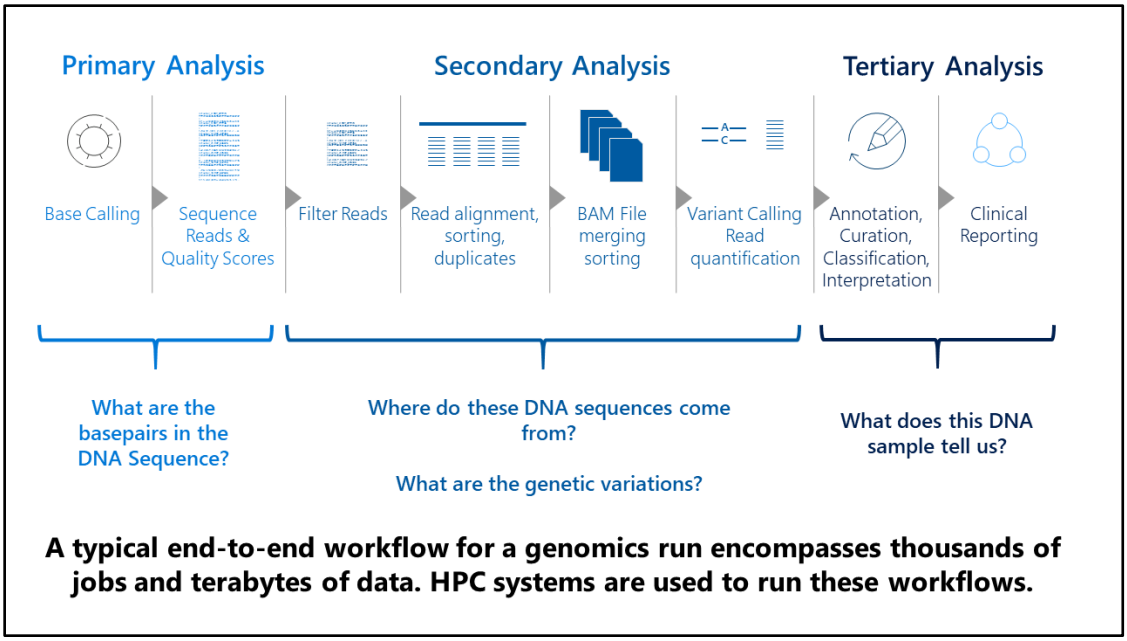
4.4 Genomic analysis in the HLS industry

The field of Genomics, which is the study of DNA/RNA at the level of whole genomes has seen a generational change in DNA sequencing technologies in the past decade. Technological innovation is exponentially increasing the rate and decreasing the cost at which we can sequence DNA accelerating the field of Genomics. DNA (and by extension RNA) sequencing is now used routinely in the medical domain, driving both diagnostics and therapeutics applications.



Source: NHGRI: https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

The rate of advance of DNA sequencing technologies has outpaced Moore's Law, with the current generation of instruments capable of generating up to 100 TB of data per week each. As a result, organizations face capacity constraints in their compute and storage infrastructure for analysing and storing these Genomics data.



A typical end-to-end workflow for a genomics run encompasses thousands of jobs and terabytes of data. HPC systems are used to run these workflows.

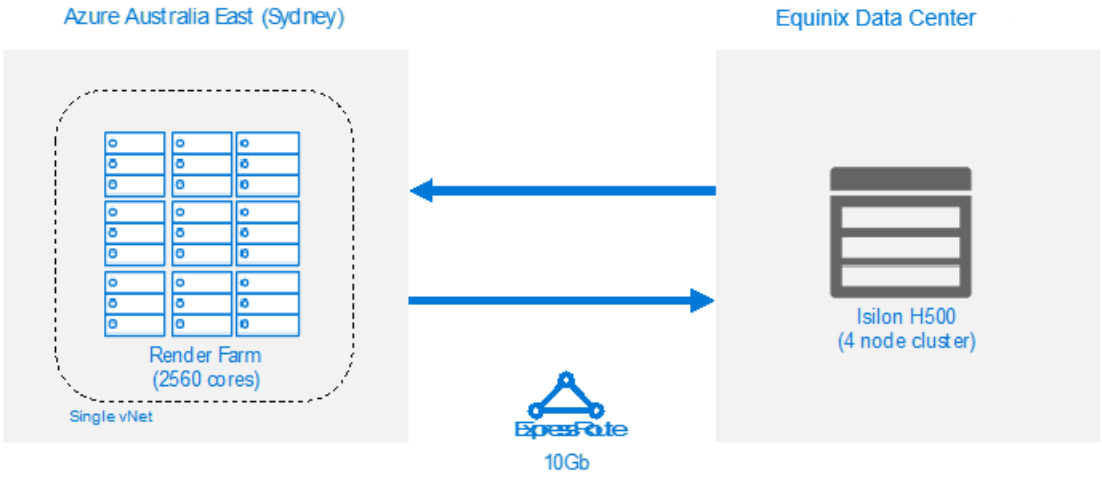
A typical genomics workflow consists of three stages: Primary, Secondary, and Tertiary. The Primary stage processes the raw output from the sequencing instruments, performs quality control and filtering, and converts the raw output into an ASCII format of the data. The ASCII output represents the four bases of DNA (A, T, C, G -- Adenine, Thymine, Cytosine, and Guanine). The Secondary stage takes the DNA data from each sample and identifies the contents at the organism, chromosome, or gene level. This process is done by running a mapping stage also known as the DNA alignment process. The alignment results are merged and collated, and then used to derive biological significance. The biological significance may be genetic variants unique to the sample, or gene expression values which indicate the activity of genes and proteins in the sample. Finally, the Tertiary stage of the workflow encompasses the data annotation and statistical analyses, which drives biological context to the DNA data. This information is then used for diagnostics and increasingly for therapeutic decisions.

The secondary analysis, specifically the alignment process, is the most computationally intensive stage of the entire workflow. This process runs parallel threads consuming multiple CPU cores and can rapidly become IO bound. The speed at which the DNA data can be aligned depends on the performance of the storage server. The storage server contains the output from the Primary analysis stage and stores the results from the Secondary analysis stage.

The use case in this study is designed to stress test this scenario – demonstrating the Isilon performance scales out linearly to match the IO demands of an increasing number of Azure CPUs doing DNA alignment. During the test, the bandwidth of the ExpressRoute connection between the Isilon and the Azure network will be monitored for saturation. The DNA alignment tool used in this benchmark is Bowtie2, a standard DNA alignment application used in this field. The DNA sample used is the NA12787 sample from the 1000 Genomes Project, specifically the ERR194147 DNA library that is sequenced at a depth of 50x. In other words, this DNA sample covers the human genome 50 times over. This sample is the standard used for genomics benchmarking.

4.5 Genomic Analysis Test Plan and Results

The genomic analysis test infrastructure using Bowtie2 is shown in the following diagram:



The Bowtie2 test plan for alignment and sorting consisted of the following steps:

1. Download the following genome to the Isilon H500 system in the Equinix Data Center: https://www.ncbi.nlm.nih.gov/sra/?term=ERR194147
2. Create qty = 5 H16 Linux VMs
3. Download Bowtie2 code to each VM
4. Connect above VMs to the Isilon H500 using an Ultra Performance VNet Gateway and an ExpressRoute 10 Gb link
5. Run Bowtie2 alignment – read FASTQ file from Isilon H500, process on VMs, write SAM file back to Isilon.
6. Measure Wall Clock Time, bandwidth on ExpressRoute 10 Gb link, and SAM file size
    a. For Wall Clock Time and ExpressRoute throughput, see table below. SAM file size is 468 GB.
7. Check that egress charges are not too high: Cost of ~8 hours of 1Gb/s data output at $0.05/GB is: 8hr/test * 60min/hr * 60s/min * 1Byte/8bits * $0.05/GB = $180/test is not too high!
8. Egress charges are not too high ($180/test), scale to 20 VMs and run Bowtie2 alignment again
9. If scaling to 20 VMs is successful, then scale to 40 VMs and run Bowtie2 alignment again
10. Start from the beginning and this time run Bowtie2 alignment **and sorting** on 2 VMs
11. Run Bowtie2 alignment **and sorting** on 40 VMs

The Bowtie2 alignment and sorting test results are as follows:

Job	# jobs	processes	smpslots	VM type	# DNA molecules from the file	Output written: Local or Shared Storage	Avere in-line	# tasks	Max ER Tput - In (Mbps) (during peak total throughput)	Max ER Tput - Out (Mbps) (during peak total throughput)	Total ER Tput (Mbps) (peak)	Wall Clock Time (Max)	Task ID	Client Cached	Sorted
481-485	5	80	16	H16	all of ERR194147	shared, direct to isilon	no	1	1370	827	2197	7hr 40m	0	no	no
491-510	20	320	16	H16	all of ERR194147	shared, direct to isilon	no	1	6890	4320	11210	7hr 42m	0	no	no
511-555	40	640	16	H16	all of ERR194147	shared, direct to isilon	no	1	6420	3720	10140	8hr 20m	0	no	no
934	1	16	16	H16	all of ERR194147	shared, direct to isilon	no	1				13h 7m	1	no	yes
934	1	16	16	H16	all of ERR194147	shared, direct to isilon	no	1				13h 15m	2	no	yes
944	40	640	16	H16	all of ERR194147	shared, direct to isilon	no	40	5110	4760	9870	13h 13m	1-40	no	yes

Other test results and observations for alignment only (a-c below) and alignment and sorting (d-e below) tests:

a. With 5 VMs, the test ran for 7 hr 40 min. They peaked at ~2 Gb/s total at the beginning of the test which was mostly read throughput over the ER link. The rest of the test ran with a steady mostly (~90%) write throughput of ~1 Gb/s over the ER link.
b. Ran test with 20 VMs. They finished in 7 hr 45 min and peaked at over 10 Gb/s in the beginning which was mostly read throughput over the ER link. The rest of the test ran with a steady ~4.5 Gb/s throughput which was 90% write over the ER link.
c. Ran with 40 VMs and finished in 8 hr 20 min. Peaked at 10 Gb/s in the beginning which was mostly read throughput over the ER link. The rest of the test ran with a steady ~9.5 Gb/s throughput which was 90% write over the ER link.
d. Ran alignment and sorting on 2 VMs and finished in 13 hr 15 min with <1 Gbps on ER link, 7 hr45 min for alignment and 5 hr 30 min for sorting. Variant calling finished in 15 min. NOTE: Job 934 spans 2 rows in the table above and these jobs were run simultaneously.
e. Ran 40 alignment and sorting jobs on 40 VMs and the jobs completed in 13 hr 13 min, demonstrating scaling to 40 VMs and jobs. The ER link was mostly saturated at the beginning and end of the test.

4.6 Genomic analysis conclusions

Bowtie2 alignment and sorting jobs scale well running on Azure compute VMs reading source files and writing output files over an ExpressRoute 10 Gb link to an Isilon H500 storage system in an Equinix facility. At the maximum tested configuration of 40 VMs, alignment and sorting is successful and the ExpressRoute 10 Gb link was mostly saturated. Since the ExpressRoute 10 Gb link was not fully saturated, the quantity of VMs and Bowtie2 jobs can likely increase (10-20%). If the goal is to substantially increase (30-100%) the quantity of VMs and Bowtie2 jobs, then we recommend adding a second ExpressRoute 10 Gb link.

5 Amazon Web Services architecture and testing

This section will outline the Amazon Web Services (AWS) architecture as tested, and detail the tests performed.

5.1 Architecture

The high-level architecture is presented in Figure 2. The architecture consists of:

- Dell EMC Isilon storage array, residing in a public cloud connected managed service provider facility (Equinix)
- Dell EMC 10 Gb networking switch, also in a public cloud-connected co-located facility (Equinix)
- 10 Gb AWS DirectConnect connection between the Equinix facility and AWS cloud
- Compute instances in AWS that connect to the Dell EMC Isilon storage array over SMB and NFS

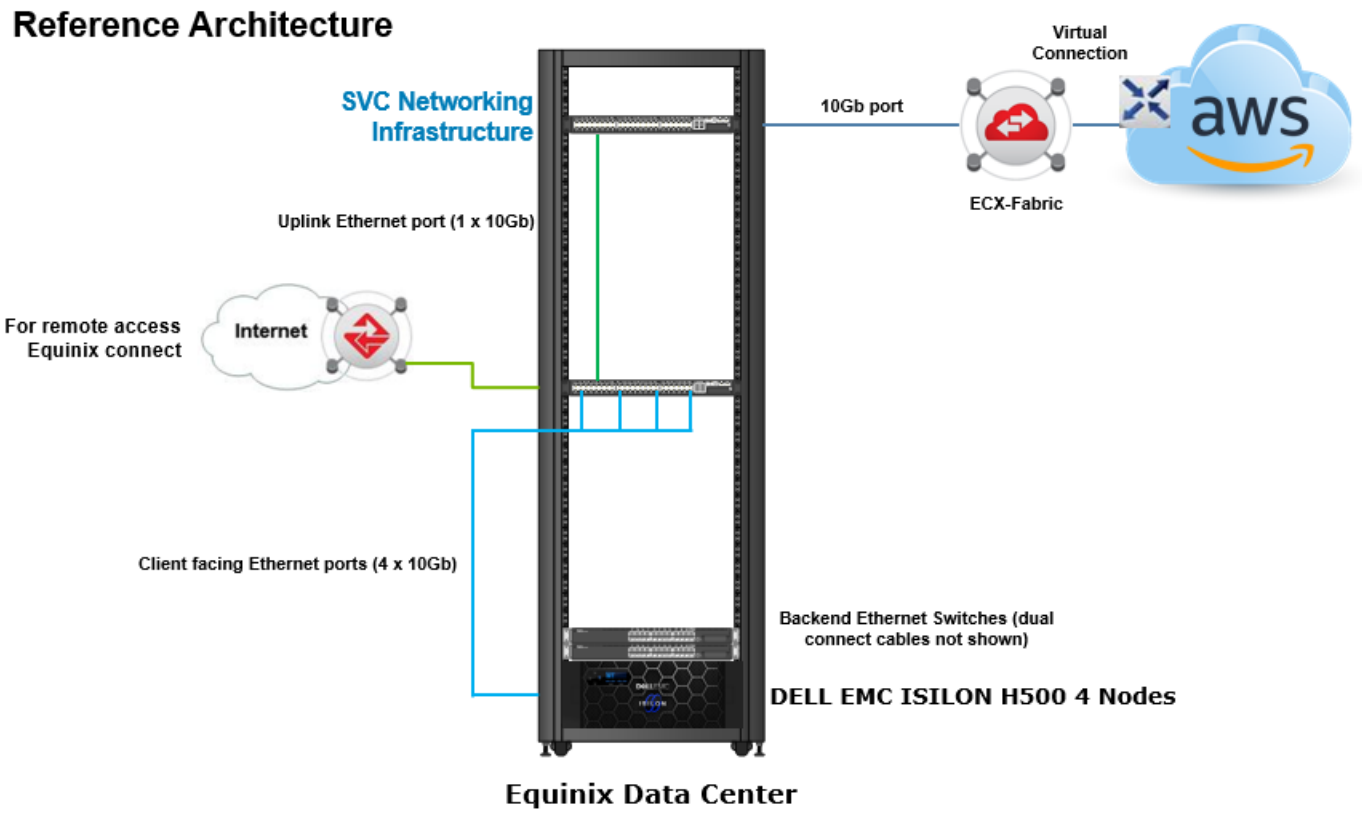


Figure 2 High-level network architecture

5.2 AWS testing configuration

This section will outline the as-tested configuration of the AWS compute instances, the AWS DirectConnect connection, the Dell EMC PowerScale storage array, and the Dell EMC network switch.

Note: All AWS resources were deployed in the ap-southeast-1c *(Singapore) Availability Zone*. The Dell EMC resources (Isilon storage array, network switch) were deployed at an Equinix cloud connect managed service provider facility in Singapore.

5.2.1 AWS Windows instance

The features of the chosen Windows instance are as follows:

Standard m5a.4xlarge
Windows Server 2019
64 GiB memory
16 vCPUs
MTU = 1500

5.2.2 AWS Linux instance

The features of the chosen Linux instance are as follows:

Standard m5a.4xlarge
Centos (x86_64)
lsb_release -d
Description: CentOS Linux release 7.6.1810 (Core)
64 GiB memory
16 vCPUs
uname –r
3.10.0-957.1.3.el7.x86_64
MTU = 1500

5.2.3 AWS Direct Connect

When located in an Equinix facility, there are two ways to connect the Isilon storage array to the AWS compute instances:

Equinix Cloud Exchange Fabric; or
AWS DirectConnect

Equinix Cloud Exchange Fabric

The major benefit of connecting into public cloud compute using Equinix Cloud Exchange Fabric™ (ECX Fabric™) is the availability of connections into all the major hyperscale providers including Amazon AWS, Microsoft Azure, Google Cloud Platform, Alibaba Cloud. ECX Fabric connects to the provider network over a single connection out of the customer's co-located network infrastructure. For example, the Isilon storage array connects to a network switch that directly connects to ECX Fabric. From there, virtual connections to the individual public cloud providers are securely established and dynamically managed using a self-service portal or API. The screen capture below shows the connection setup to ECX Fabric. More information can be found here.

> *ECX Fabric brings together cloud service providers and users, enabling them to establish affordable, private, high-performance connections within Platform Equinix®* (Equinix).

Amazon Web Services Direct Connect

AWS Direct Connect is a cloud service solution that makes it possible to connect your private IT infrastructure to AWS over a private, high-speed connection.

With AWS Direct Connect, establish connections at a cloud-connected co-location facility, such as Equinix, or directly connect from your existing WAN network provider. AWS Direct Connect does not go over the public Internet and allows Direct Connect connections to offer more reliability, faster speeds, consistent latencies, and higher security than typical connections over the Internet.

AWS Direct Connect has two billing elements: port hours and outbound data transfer. Port hour pricing varies by connection type - Dedicated Connection or Hosted Connection - and capacity. Data transfer out over AWS Direct Connect is charged per GB. https://aws.amazon.com/directconnect/pricing/

5.2.4 Dell EMC PowerScale Scale-out NAS

A part of the Dell EMC PowerScale storage family, the Isilon H500 is described in section 2.2.4. The configuration of the Dell EMC Isilon storage array tested for the AWS testing is as follows:

Dell EMC Isilon H500 4 node storage array (per node)
2.2 GHz 10-Core
128 GiB memory
15 x 4 TB HDD | 1 x 1.6 TB SSD
2 x 10 GE (Front end)
2 x QSFP+ 40 Gb Ethernet (internal backend network)
OneFS version 8.1.2
Protection Policy - +2d:1n
Pool Usable Capacity - 178 TB / 162 TiB @ 100% utilization
NFS v3 mount options (default OneFS values) -
`rw,noatime,vers=3,rsize=131072,wsize=524288,namlen=255,hard,proto=tcp,timeo=600,retrans=2`
Four separate NFS mount points created for mounting on each Linux VM

```
- isilon-node1:/mount1     222T   23G  215T   1% /mnt/mount1
- isilon-node2:/mount2     222T   23G  215T   1% /mnt/mount2
- isilon-node3:/mount3     222T   23G  215T   1% /mnt/mount3
- isilon-node4:/mount4     222T   23G  215T   1% /mnt/mount4
```

SMB v3
Four separate SMB shares created for mounting on each Windows VM

```
S:      \\172.16.0.5\smb1       Microsoft Windows Network
X:      \\172.16.0.6\smb2       Microsoft Windows Network
Y:      \\172.16.0.7\smb3       Microsoft Windows Network
W:      \\172.16.0.8\smb4       Microsoft Windows Network
```

Jumbo frames disabled
Six rack units (4U chassis + 2u backend network)

5.2.5 Dell EMC Network Switch

The S4148F-ON network switch used here is similar to the switch described in section 2.2. The configuration of the tested network switch in the AWS testing is as follows:

Model: S4148F-ON
OS version: 10.4.2.0
Connectivity to Unity storage array: 10 Gb TwinAx
Connectivity to Direct Connect router: 10 Gb LR SFP+
One rack unit
Jumbo frames disabled

5.3 Testing Methodology

The primary tool used for this benchmark testing is vdbench. Vdbench is a command-line utility created to generate disk I/O workloads used for validating storage performance and storage data integrity. Iperf and ping tests were conducted before the benchmark testing to validate network bandwidth throughout the test infrastructure.

There are four main test categories used to validate the viability of the Dell EMC PowerScale solution in an Equinix managed datacenter.

- Iperf network bandwidth tests (bi-directional)
- Network latency (aka ping tests)
- Windows SMB benchmarking
- Linux NFSv3 benchmarking

5.3.1 Iperf network bandwidth (https://iperf.fr/)

iPerf is a tool for active measurements of the maximum achievable bandwidth on IP networks. iPerf2 is preinstalled on the OneFS operating system and should be leveraged to measure network performance before running any performance benchmark. Network bandwidth should be measured to set maximum performance expectations from the client and server network to the Isilon nodes. iPerf 2.0.9 was downloaded and used for the client systems on both Linux and Windows.

Testing was performed using both Linux and Windows compute servers connected to the OneFS nodes. The iperf tests measured the read and write bandwidth over the DirectConnect 10 Gbps network link. AWS testing provided some inconsistent results similar to the results with MS Azure ExpressRoute testing. Storage benchmarks proceeded with this noted.

Based on the AWS VM definitions, the maximum theoretical bandwidth per VM is 2120 Mbps operating in standard networking mode.
https://aws.amazon.com/blogs/aws/m5-the-next-generation-of-general-purpose-ec2-instances/

To start the iperf server on a OneFS environment.

```
# isi_for_array iperf -s
```

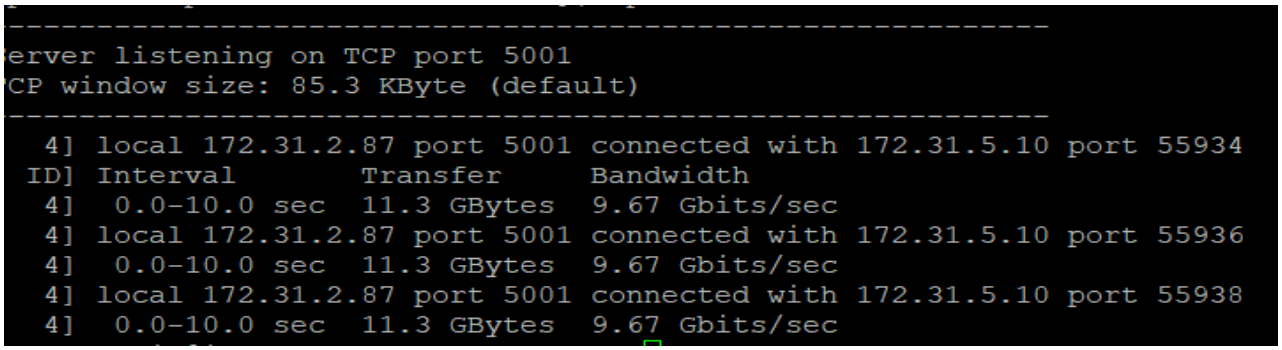
To start the iperf client on a Linux VM connecting to one of the Isilon nodes, it is the same command issued from the cli

Summary of results:

Below is an example command from one Linux VM to one Isilon node. Test was repeated from each VM to each Isilon node in the cluster to validate results and consistent network performance.

```
sgdch500-1: [  4]  0.0-10.0 sec  1.75 GBytes  1.50 Gbits/sec
sgdch500-2: [  4] local 172.16.1.12 port 5001 connected with 172.31.5.10 port 33718
sgdch500-2: [ ID] Interval       Transfer     Bandwidth
sgdch500-2: [  4]  0.0-10.0 sec  1.76 GBytes  1.51 Gbits/sec
sgdch500-4: [  4] local 172.16.1.14 port 5001 connected with 172.31.5.10 port 42362
sgdch500-4: [ ID] Interval       Transfer     Bandwidth
sgdch500-4: [  4]  0.0-10.0 sec  1.58 GBytes  1.36 Gbits/sec
```

Also shown below are the results for iperf connections between two compute VMs.



5.3.2 Network ping testing

Using the ping command to test overall network latency between Linux and Windows VMs to the Isilon nodes.

Linux VM1 (AWS Cloud) to Isilon node 1 (over AWS Direct Connect network)

- `ping 172.16.1.14`
- `PING 172.16.1.14 (172.16.1.14) 56(84) bytes of data.`
- `64 bytes from 172.16.1.14: icmp_seq=1 ttl=62 time=1.82 ms`
- `64 bytes from 172.16.1.14: icmp_seq=2 ttl=62 time=1.82 ms`

Linux VM1 to Linux VM1 hosted in AWS cloud

- `ping 172.31.5.10`
- `PING 172.31.5.10 (172.31.5.10) 56(84) bytes of data.`
- `64 bytes from 172.31.5.10: icmp_seq=1 ttl=64 time=0.303 ms`
- `64 bytes from 172.31.5.10: icmp_seq=2 ttl=64 time=0.114 ms`
- `64 bytes from 172.31.5.10: icmp_seq=3 ttl=64 time=0.098 ms`
- `64 bytes from 172.31.5.10: icmp_seq=4 ttl=64 time=0.090 ms`

5.3.3 Windows SMB3 Benchmarking

To test SMB throughput performance, vdbench is used to illustrate the impact of reads and writes to a 1 GB file. Vdbench parameters were specified to generate multiple threads and multiple IO transfer sizes.

The four tests were performed for each of the following IO transfer sizes: 64 k, 128 k, 256 k, 512 k, and 1024 k. The parameters below show an example of the Windows disk benchmarking commands using the 1024 k (1 M) transfer size.

```
fsd=throughput_$host_dir1,anchor=Z:\\$host\throughput,depth=2,width=10,files=10,
size=1G

fwd_tpr_$host,fsd=throughput_$host_*,operation=read,threads=20,xfersize=1M

fwd=fwd=fwd_tpw_$host,fsd=throughput_$host_*,operation=write,threads=10,xfersize
=1M,openflags=directio
```

This test will run the following benchmark, the `xfersize` variable **determines the IO transfer size**.

- Create folder structure with 2 subdirectories, 10 parent directories, each containing ten 1 GB files
- Read the files with 20 threads
- Writes to the files with 10 threads, `directio` flag enabled to eliminate cache

---

Note: There are five iterations of each test run for reads and writes. The value selected is the average for each iteration.

Additional notes about these tests and results:

- Each test iteration is run five times at random times throughout a 24-hour period.
- Each test iteration reads and writes the file based on vdbench parameters.
- Testing alternates between read test and write testing for each run.
- The five overall results presented are averaged and graphed in the results section

Windows SMB settings

Settings for SMB version and multi-channel connections are shown below by using Windows PowerShell.

```
PS C:\Users\Administrator> get-smbconnection

ServerName  ShareName UserName                    Credential              Dialect NumOpens
----------  --------- --------                    ----------              ------- --------
172.16.1.11 downloads EC2AMAZ-E9U1NER\Administrator EC2AMAZ-E9U1NER\pocadmin 3.1.1   4
172.16.1.11 IPC$      EC2AMAZ-E9U1NER\Administrator EC2AMAZ-E9U1NER\pocadmin 3.1.1   1
172.16.1.11 smb1      EC2AMAZ-E9U1NER\Administrator EC2AMAZ-E9U1NER\pocadmin 3.1.1   29


PS C:\Users\Administrator> get-smbmultichannelconnection

Server Name Selected Client IP    Server IP   Client Interface Index Server Interface Index Client RSS Capable Client
                                                                                                               RDMA
                                                                                                               Capable
----------- -------- ---------    ---------   ---------------------- ---------------------- ------------------ --------
172.16.1.11 True     172.31.10.16 172.16.1.11 8                      2                      True               False


PS C:\Users\Administrator>
```

```
PS C:\Users\Administrator> get-smbclientconfiguration


ConnectionCountPerRssNetworkInterface : 4
DirectoryCacheEntriesMax              : 16
DirectoryCacheEntrySizeMax            : 65536
DirectoryCacheLifetime                : 10
DormantFileLimit                      : 1023
EnableBandwidthThrottling             : True
EnableByteRangeLockingOnReadOnlyFiles : True
EnableInsecureGuestLogons             : False
EnableLargeMtu                        : True
EnableLoadBalanceScaleOut             : True
EnableMultiChannel                    : True
EnableSecuritySignature               : True
ExtendedSessionTimeout                : 1000
FileInfoCacheEntriesMax               : 64
FileInfoCacheLifetime                 : 10
FileNotFoundCacheEntriesMax           : 128
FileNotFoundCacheLifetime             : 5
KeepConn                              : 600
MaxCmds                               : 50
MaximumConnectionCountPerServer       : 32
OplocksDisabled                       : False
RequireSecuritySignature              : False
SessionTimeout                        : 60
UseOpportunisticLocking               : True
WindowSizeThreshold                   : 1
```

```
PS C:\Users\pocadmin> get-smbmultichannelconnection

Server Name Selected Client IP   Server IP  Client Interface Index Server Interface Index Client RSS Capable Client
                                                                                                             RDMA
                                                                                                             Capable
----------- -------- ---------   ---------  ---------------------- ---------------------- ------------------ ----------
172.16.0.5  True     10.10.0.197 172.16.0.5 10                     1                      True               False
172.16.0.7  True     10.10.0.197 172.16.0.7 10                     1                      True               False
172.16.0.6  True     10.10.0.197 172.16.0.6 10                     1                      True               False
172.16.0.8  True     10.10.0.197 172.16.0.8 10                     1                      True               False
```

5.3.4 Linux NFSv3 Benchmarking

To test NFS throughput performance, vdbench is used to illustrate the impact of reads and writes to a 1 GB file. Vdbench parameters were specified to generate multiple threads and multiple IO transfer sizes.

The four tests were performed for each of the following IO transfer sizes: 64 k, 128 k, 256 k, 512 k, and 1024 k. The parameters below show an example of the Windows disk benchmarking commands using the 1024 k (1 M) transfer size.

```
fsd=throughput_$host_dir1,anchor=/mnt/mount1/throughput/$host,depth=2,width=10,f
iles=10,size=1G

fwd=fwd_tpr_$host,fsd=throughput_$host_*,operation=read,threads=20,xfersize=1M

fwd=fwd_tpw_$host,fsd=throughput_$host_*,operation=write,threads=20,xfersize=64k
,openflags=directio
```

This test will run the following benchmark, the `xfersize` variable determines the IO transfer size.

- Create folder structure with 2 subdirectories, 10 parent directories, each containing ten 1 GB files
- Read the files with 20 threads
- Writes to the files with 10 threads, `directio` flag enabled to eliminate cache

Note: There are five iterations of each test run for reads and writes. The value selected is the average of for each iteration.

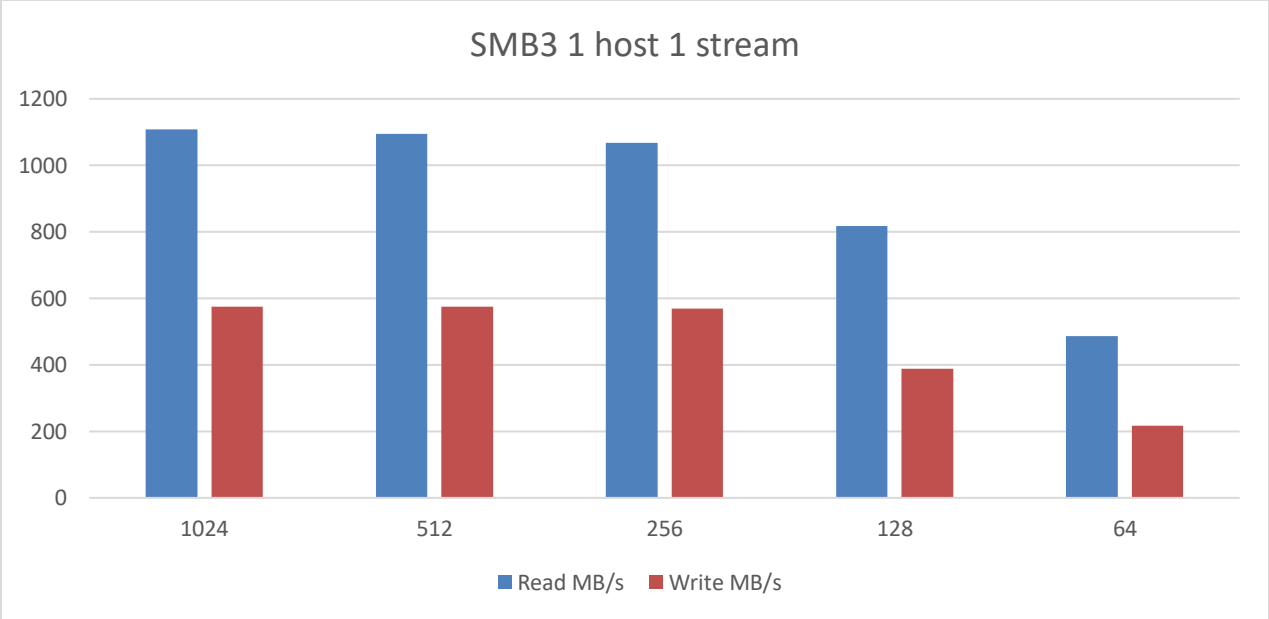
Additional notes about these tests and results:

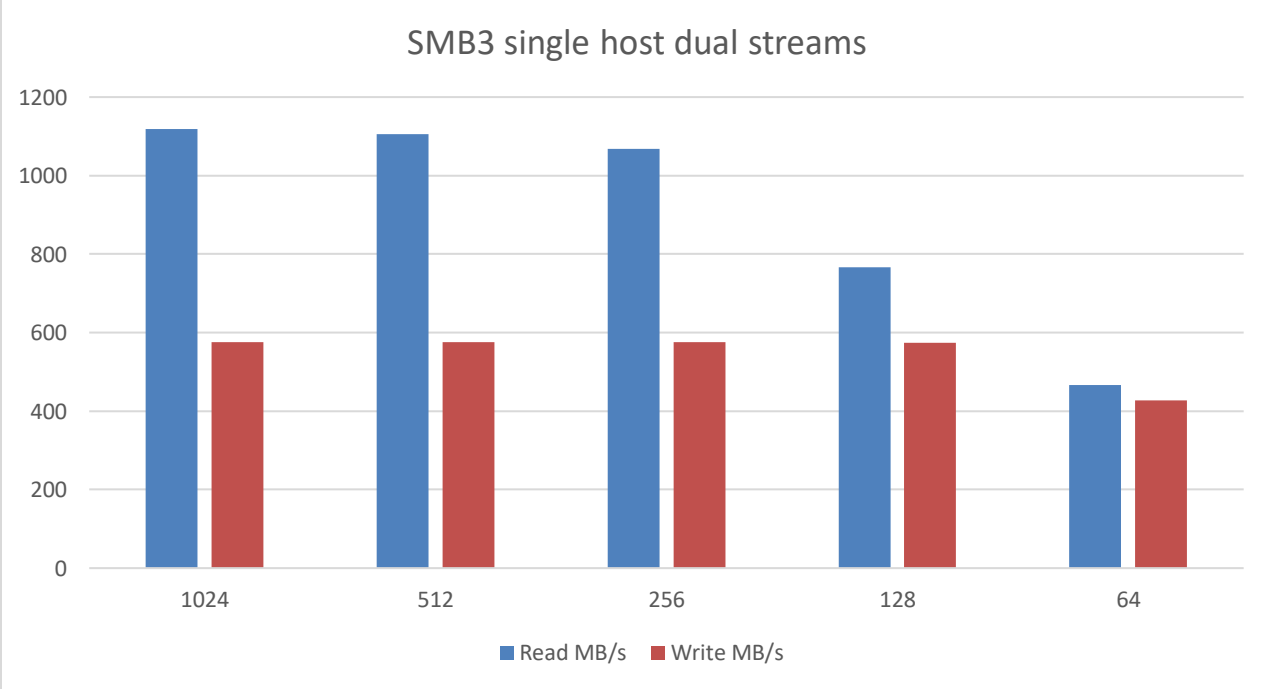
- Each test iteration is run five times at random times throughout a 24-hour period.
- Each test iteration reads and writes the file based on vdbench parameters.
- Testing alternates between read test and write testing for each run.
- The five overall results presented are averaged and graphed in the results section

6 Amazon Web Services testing results

This section provides the results of running various vdbench outputs for both Windows and Linux VMs connecting to Dell EMC PowerScale storage located in the Equinix managed service provider facility.

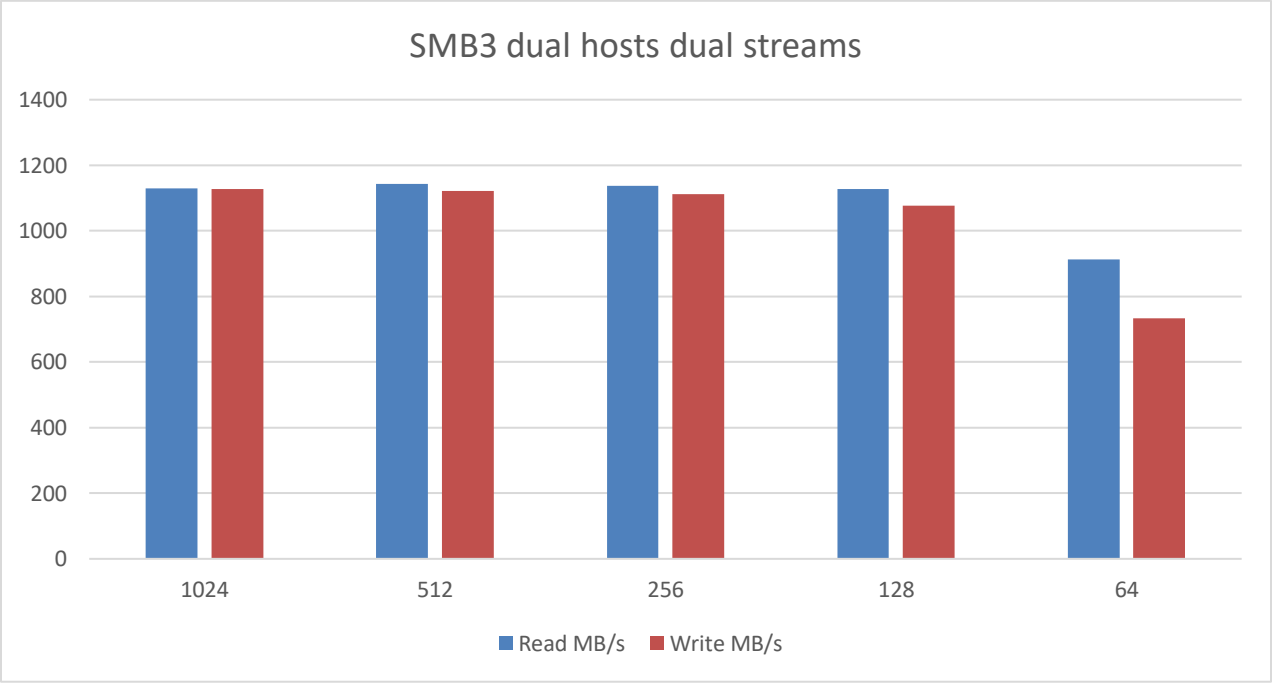
6.1 Windows – Single host benchmarking results



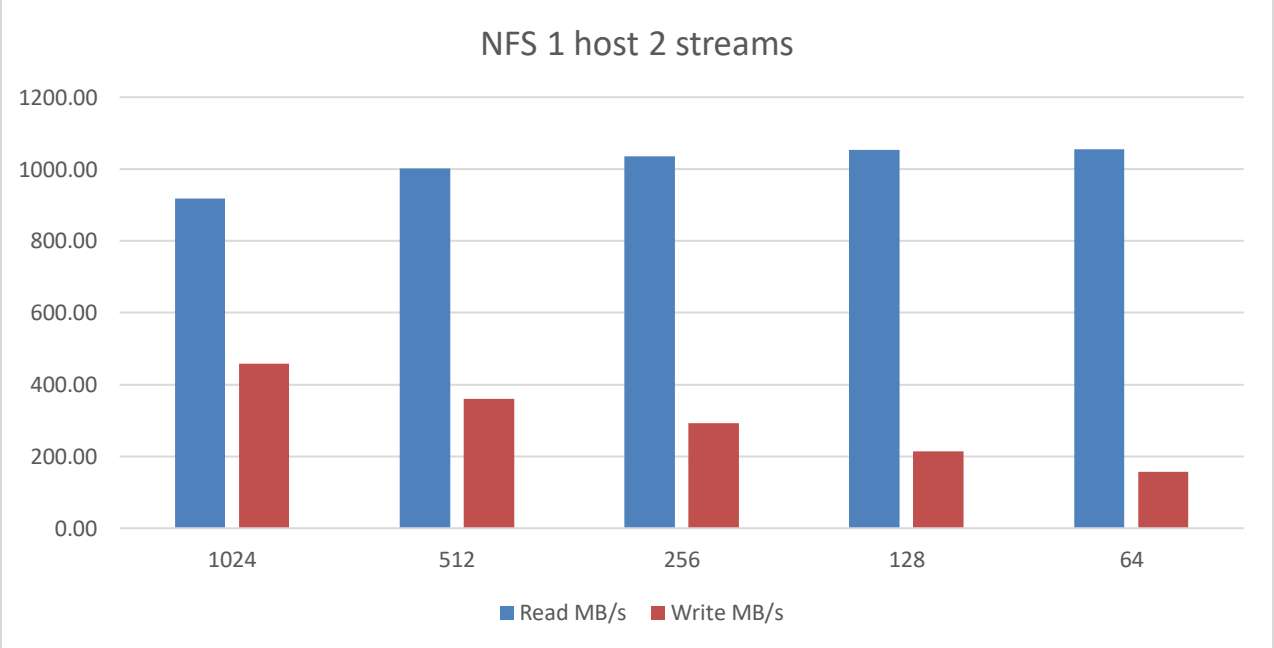


6.2 Windows – Two hosts two total streams benchmarking results

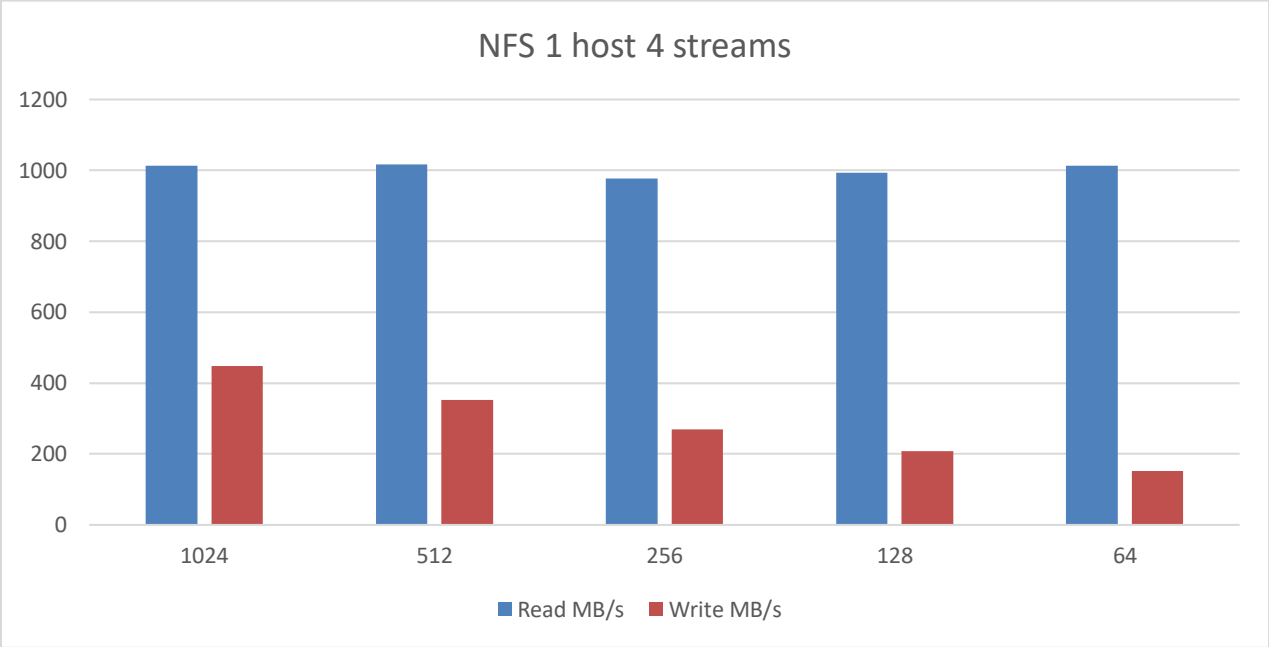
Two streams writes reached max bandwidth of a single Direct Connect link SMB3 using multi-channel connections.



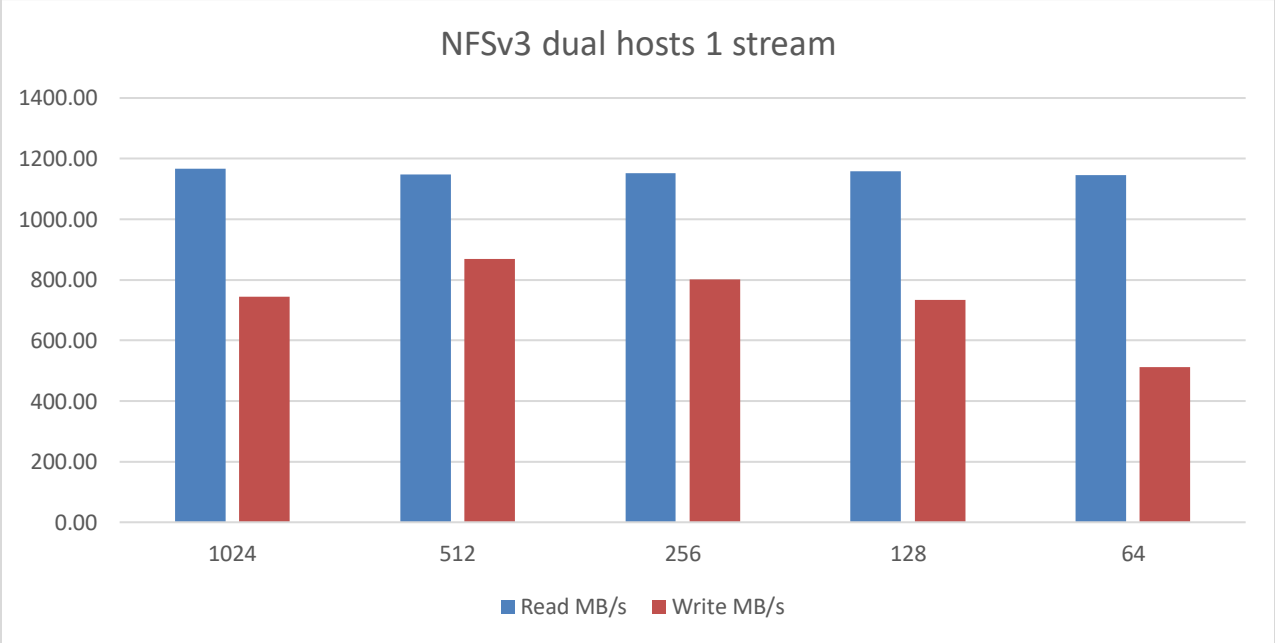
6.3 Linux – One host two streams benchmarking results



6.4 Linux – One host four streams benchmarking results



6.5 Linux – Two hosts single stream benchmarking results



7 Amazon Web Services testing comparison

This section provides results of running various vdbench outputs for Linux VMs connecting to a Dell EMC Isilon cluster located within an Equinix facility. The results compare to native EFS and Lustre file services from AWS.

The same Linux VMs were used to provide a comparison using 1 MB and 10 MB files.

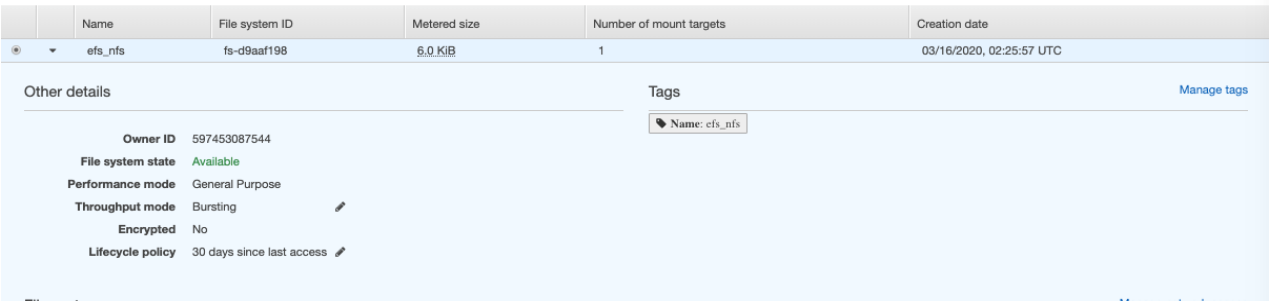
The benchmark tool vdbench was used to create 280,000 1 MB files spread over 840 directories with a depth of two and an IO size of 64 k. This dataset size was used to ensure VM memory was surpassed in sizing.

For the second test, the vdbench was used to create 30,000 10 MB files spread over 220 directories with a depth of two. This dataset size was used to ensure VM memory was surpassed in sizing.

Note: Larger dataset sizes can be used to extend the overall length of time.

The native file systems inside AWS offer different configuration options for capacity and performance at different price points. FSX file systems allow customers to select fixed sizes and specific performance profiles at an increased cost. EFS uses scalable EBS storage and provides NFS and SMB file access where Lustre offers higher performance scratch space scalable for bursting workloads.

Figure 3: EFS standard settings



Two different EFS file systems were tested in both General Purpose Burst and Provisioned.

https://docs.aws.amazon.com/efs/latest/ug/performance.html

The following table provides examples of bursting behavior.

File System Size	Aggregate Read/Write Throughput
A 100-GiB file system can...	• Burst to 100 MiB/s for up to 72 minutes each day, or • Drive up to 5 MiB/s continuously
A 1-TiB file system can...	• Burst to 100 MiB/s for 12 hours each day, or • Drive 50 MiB/s continuously
A 10-TiB file system can...	• Burst to 1 GiB/s for 12 hours each day, or • Drive 500 MiB/s continuously
Generally, a larger file system can...	• Burst to 100MiB/s per TiB of storage for 12 hours each day, or • Drive 50 MiB/s per TiB of storage continuously

Figure 4: Lustre settings for FSx1

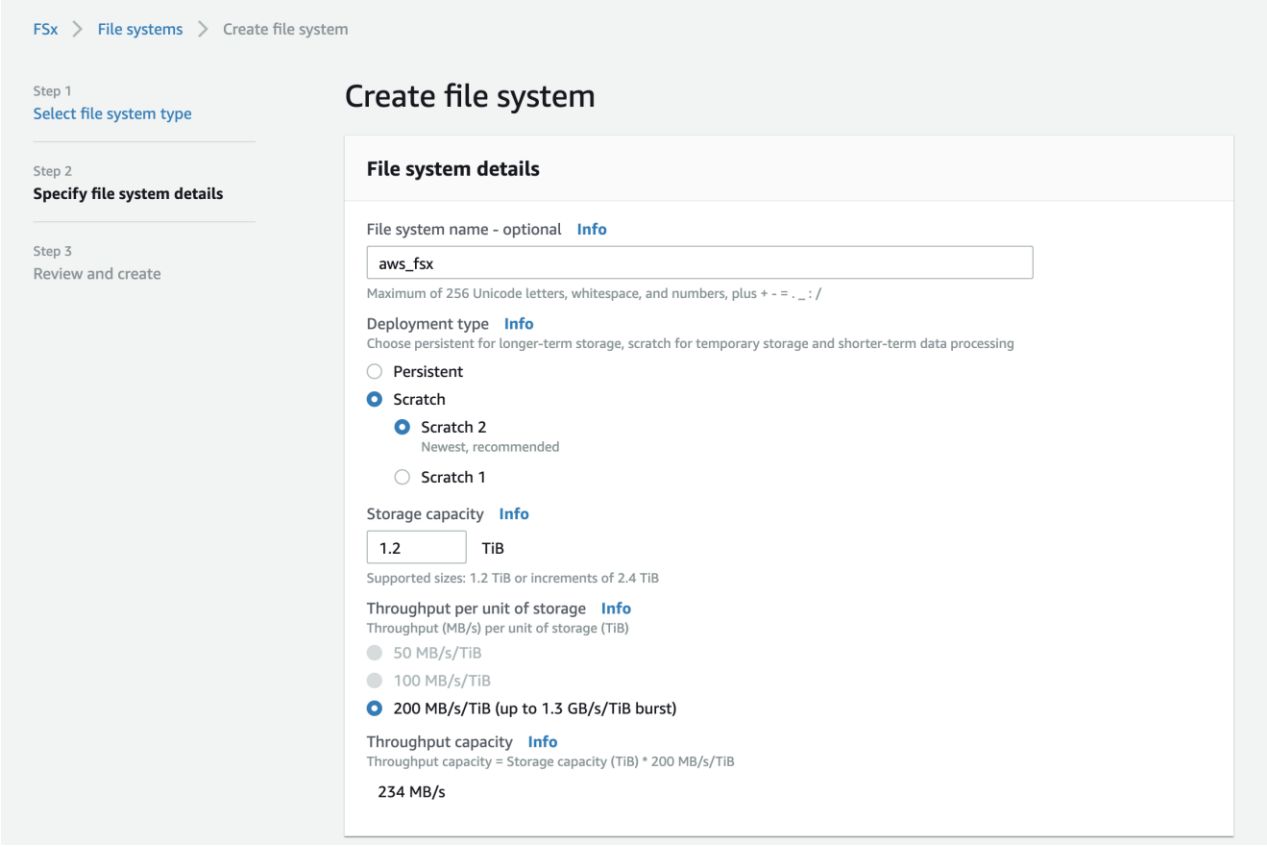


Figure 5: Lustre settings for FSx2

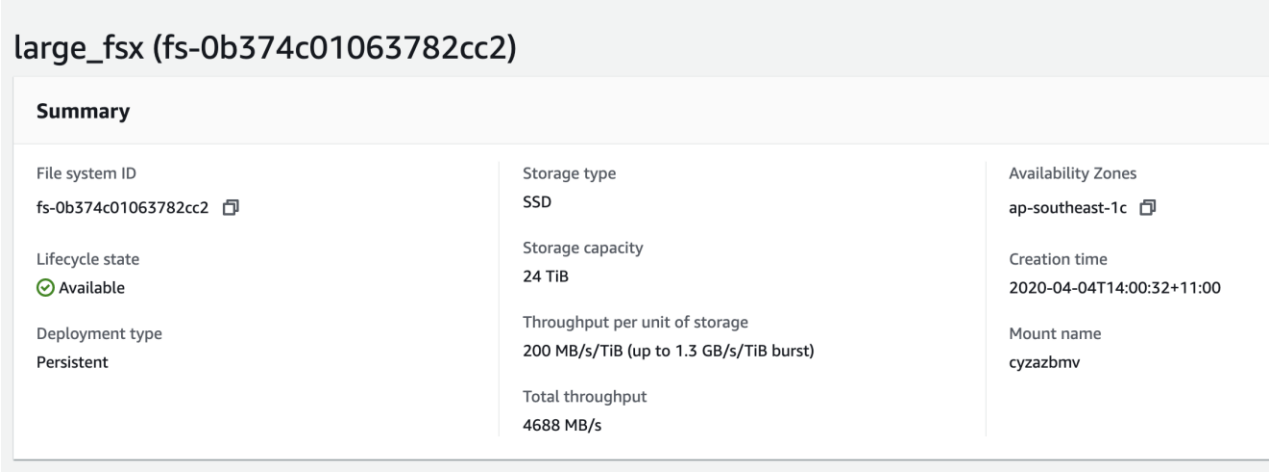
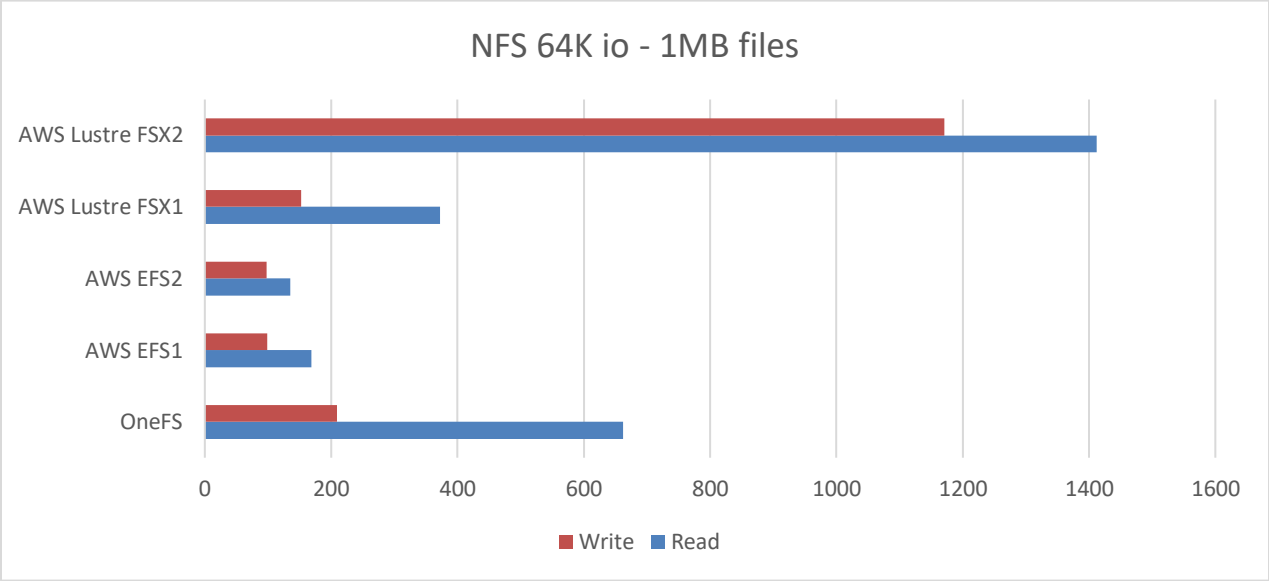


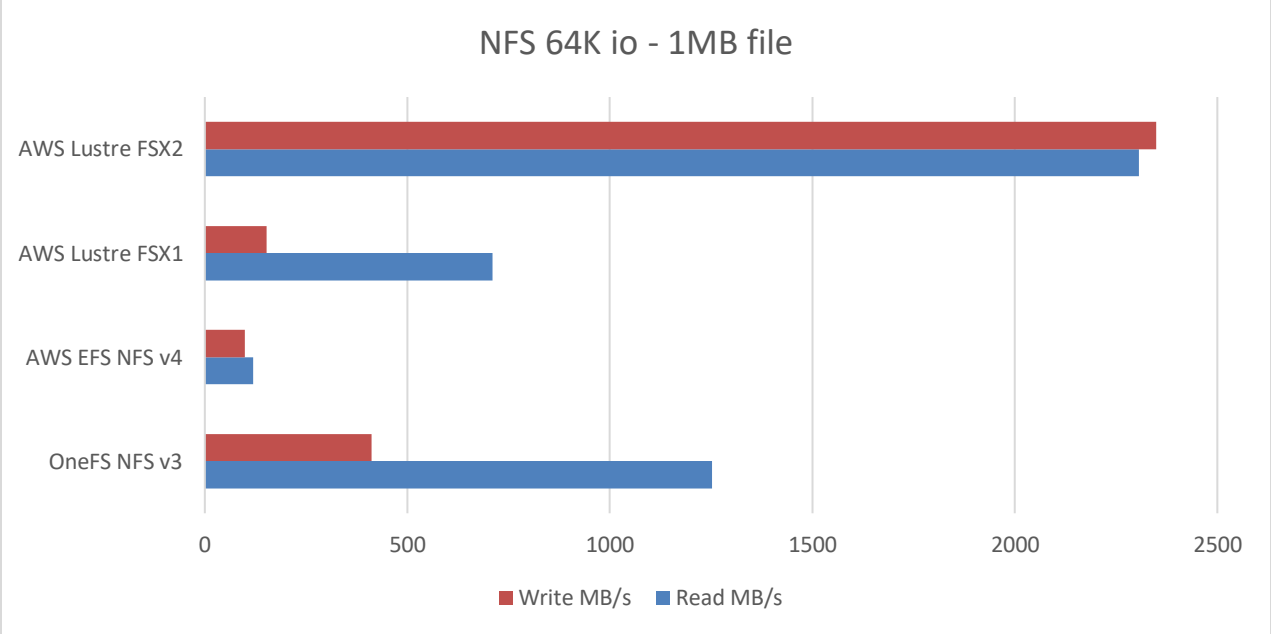
Figure 6: USD pricing to create FSx including $/GB rates

FSx		**$1,491.56**
Asia Pacific (Singapore)		**$1,491.56**
Amazon FSx CreateFileSystem:Lustre		$1,491.56
$0.168 per GB-Month of provisioned Lustre storage - Asia Pacific (Singapore)	336.667 GB-Mo	$56.56
$0.339 - per GB-Month of provisioned persistent Lustre storage with 200 MB/s per TiB - Asia Pacific (Singapore)	4,233.025 GB-Mo	$1,435.00

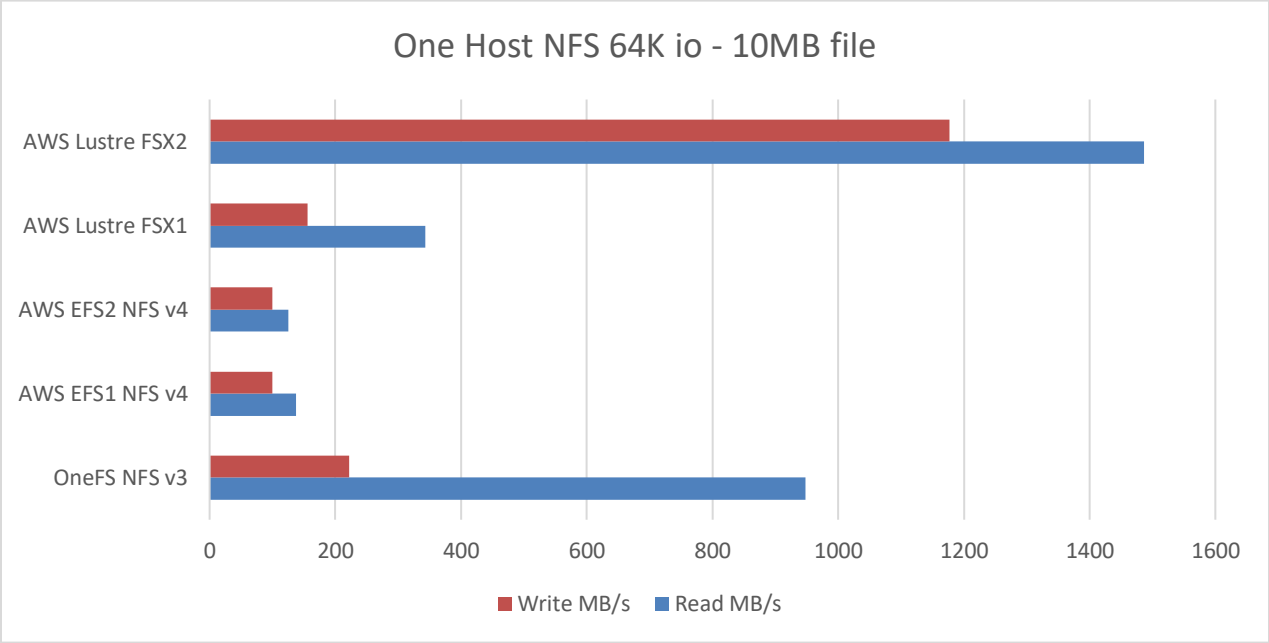
7.1 One Linux host – Single stream 280,000 files of 1 MB



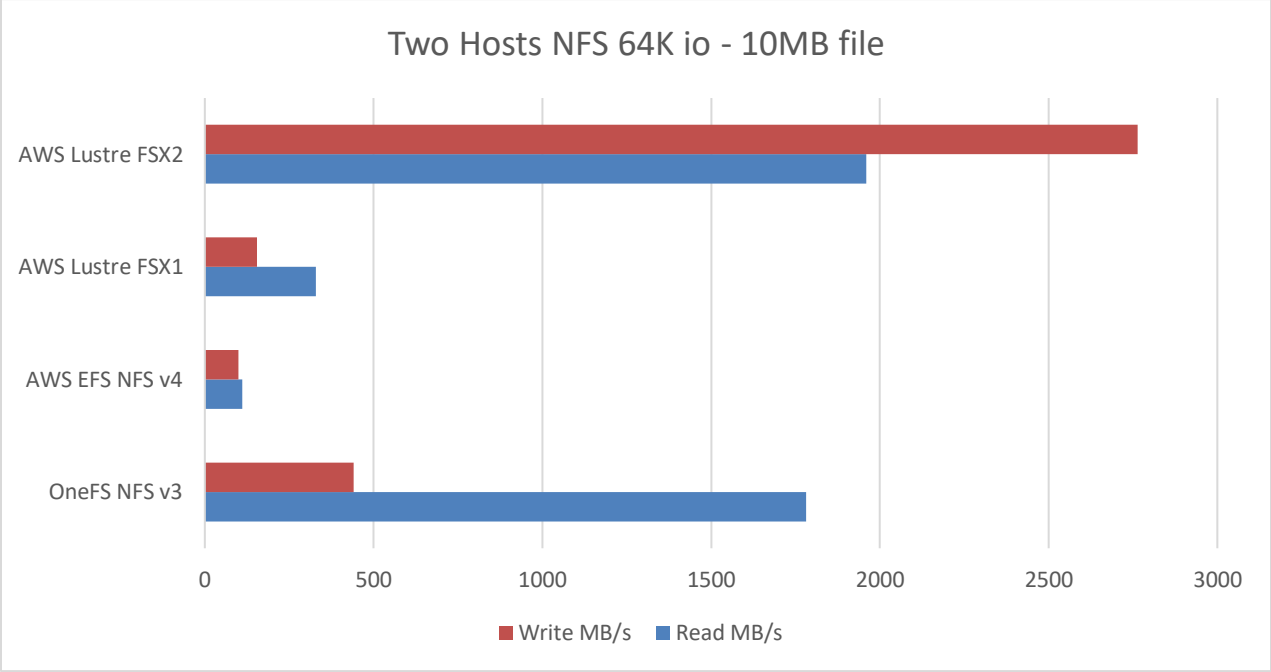
7.2 Two Linux hosts – Single stream 280,000 files of 1 MB



7.3 One Linux host – Single stream 30,000 files of 10 MB



7.4 Two Linux hosts – Single stream 30,000 files of 10 MB



8 SyncIQ between cloud locations

In multi-cloud environments data replication is a key requirement for customers. OneFS provides for SyncIQ replication between cluster instances over distance using asynchronous replication. The details below outline bi-directional replication between H500 systems located in Singapore and Sydney, Australia. Network replication was over a dedicated 100 Mb using the Equinix cloud fabric.

Network latency showed 87 ms as the ping time between locations over the allocated network.

8.1 SyncIQ testing

Average transfer speed is 9930981023 / (13 * 60 + 38) = 92.6 Mb/s

On the source cluster:

```
sdch500-1# cd parabricks
sdch500-1# ls -lsap
total 12956480
     32 drwxr-xr-x     2 root  wheel          77 Jun 21  2019 ./
     32 drwxr-xr-x    12 root  wheel         291 Aug 20  2019 ../
     72 -rwxr-xr-x +   1 root  wheel       15105 Jun 20  2019 parabricks.tar.gz
12956344 -rwxr-xr-x +   1 root  wheel  9924454379 Feb 22  2019 parabricks_sample.tar.gz
sdch500-1# du -sh
 12G    .
sdch500-1# du -sh -m
12653   .
sdch500-1#
```

On the target cluster:

```
sgdch500-2# cd parabricks_DR
sgdch500-2# ls -lsap
total 12960776
     32 drwxr-xr-x     2 root  wheel          77 Jun 21  2019 ./
     32 drwxrwxrwx    10 root  wheel         198 Jun 19 20:46 ../
     72 -rwxr-xr-x +   1 root  wheel       15105 Jun 20  2019 parabricks.tar.gz
12960640 -rwxr-xr-x +   1 root  wheel  9924454379 Feb 22  2019 parabricks_sample.tar.gz
sgdch500-2# du -sh -m
12657   .
sgdch500-2#
```

Subreport summary

Job ID:	1
Policy name:	Alex3
Status:	Finished
Started:	2020-06-20 06:59:49
Ended:	2020-06-20 07:13:27
Duration:	13 m 38 s
Errors:	No errors reported

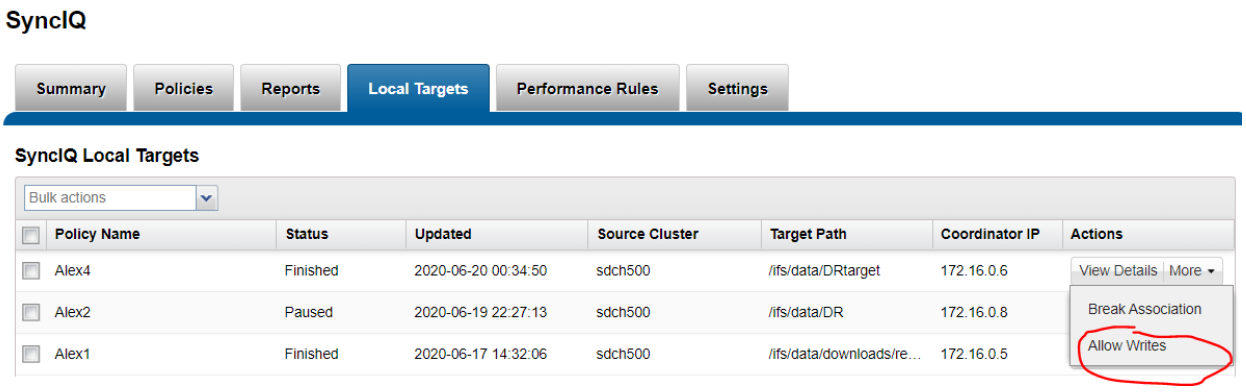
Subreport details

Sync type:	Initial
Action:	Run
Directories	
Source directories visited:	1 directories
Target directories deleted:	0 directories
Files	
Total files:	2 files
Actually transferred:	0 files
New files:	2 files
Updated files:	0 files
Automatically retransmitted files:	0 files
Target files deleted:	0 files
Number of committed files with conflicts:	0 conflicts
Skipped files	
Up-to-date (already replicated):	0 files
Modified while being replicated:	0 files
I/O errors occurred:	0 files
Network errors occurred:	0 files
Integrity errors occurred:	0 files
Data transfer	
Total network traffic:	9930981023 bytes
Total data:	9924469504 bytes
File data:	9924469504 bytes
Sparse data:	0 bytes
Snapshots	
Target:	SIQ-Failover-Alex3-2020-06-19_21-00-24

Policy summary

Policy name:	Alex3
Action:	Sync
Source root directory:	/ifs/data/azdata/parabricks
Included directories:	No value
Excluded directories:	No value
File matching criteria:	No value
Target host:	172.16.1.11
Target directory:	/ifs/data/parabricks_DR

The policy engine on OneFS can be configured to run on a scheduled basis or when new data has been written to the source file system. Target locations for SyncIQ datasets are held in read-only mode until promoted to the active file system to allow new writes. Testing of failover and failback process was performed in the environment as per the SyncIQ best practices.



Change the file by adding a new line.

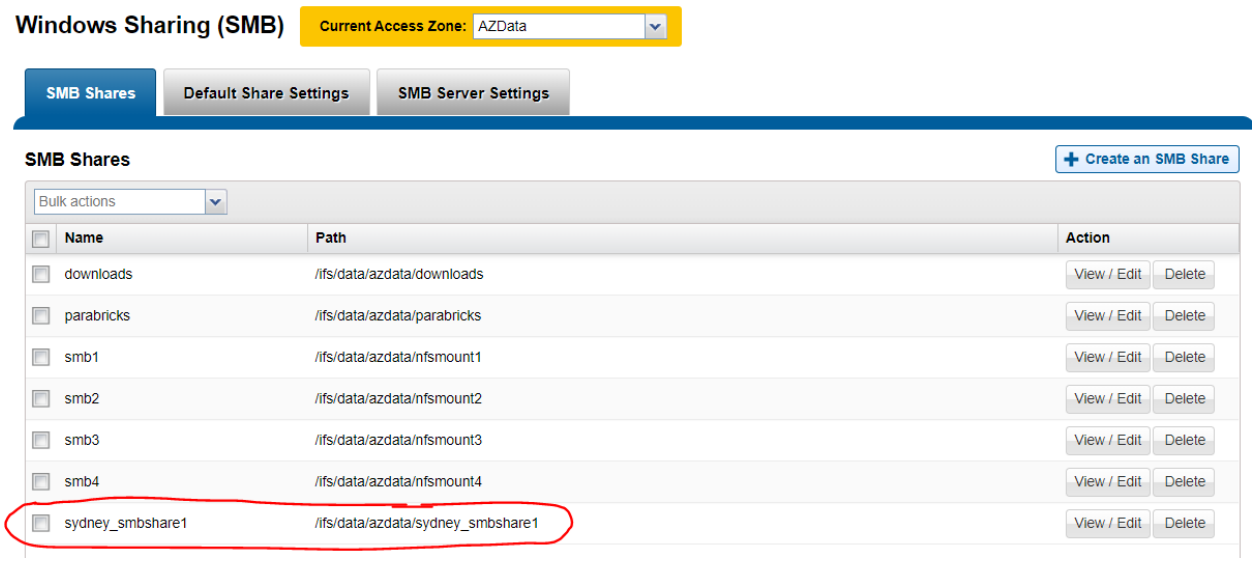
```
sgdch500-2# cat newfile_syd
this is the line from sydney.
this is the line from Singapore.
```

Also, create a new file from Singapore.

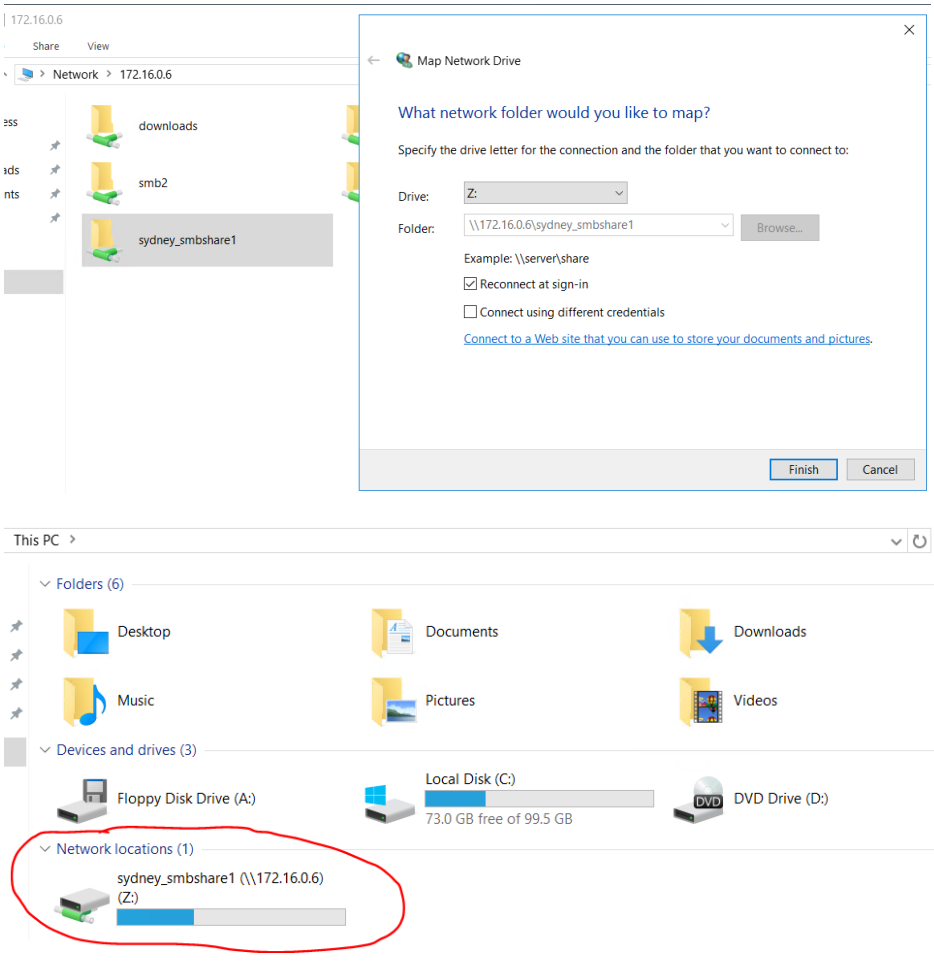
```
sgdch500-2#
sgdch500-2#
sgdch500-2# cat newfile_syd
this is the line from sydney.
this is the line from Singapore.
sgdch500-2# ls -lsap
total 112
32 drwxr-xr-x     2 root  wheel   29 Jun 19 22:41 ./
32 drwxrwxrwx    11 root  wheel  224 Jun 19 22:34 ../
48 -rw-r--r--     1 root  wheel   63 Jun 19 22:41 newfile_syd
sgdch500-2# touch newfile_sgp
sgdch500-2# ls -lsap
total 136
32 drwxr-xr-x     2 root  wheel   58 Jun 19 22:42 ./
32 drwxrwxrwx    11 root  wheel  224 Jun 19 22:34 ../
24 -rw-r--r--     1 root  wheel    0 Jun 19 22:42 newfile_sgp
48 -rw-r--r--     1 root  wheel   63 Jun 19 22:41 newfile_syd
```

Host access was tested from both Windows (SMB) and Linux (NFS) servers. The Windows server access is shown below and was used to validate access. These VM's simulate the ability for the dataset to be accessible over the local hyperscale cloud providers network connection.

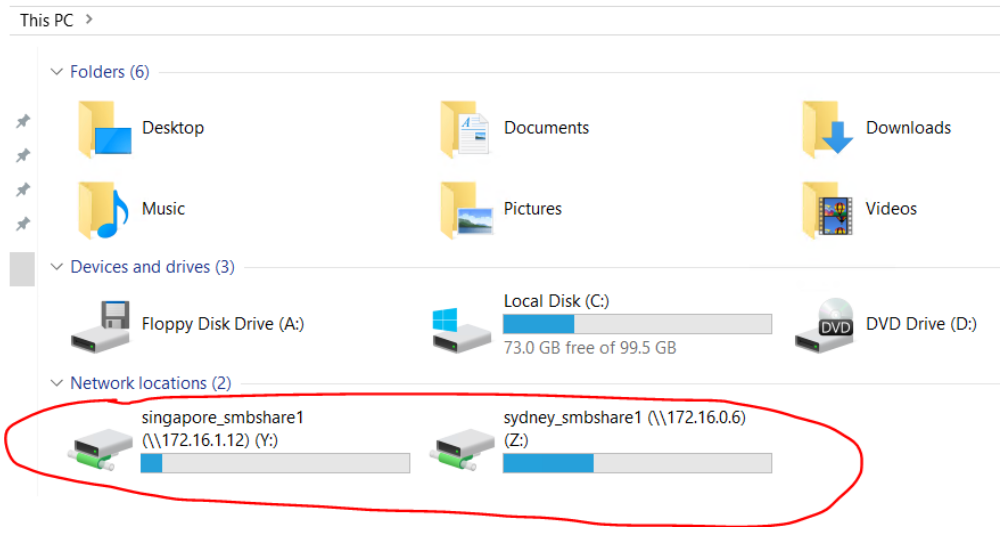
Create a new SMB share at Sydney Isilon.



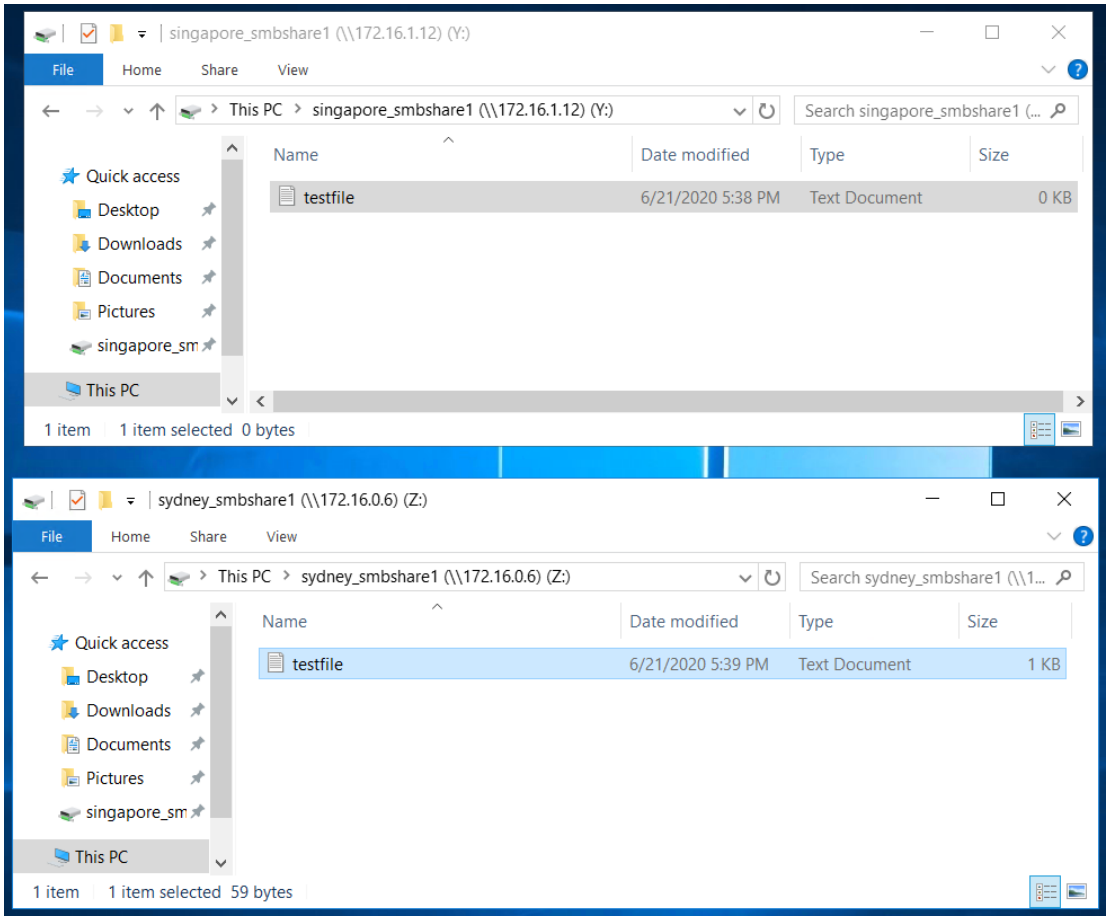
Present to the Windows client and create a local mapped drive against it.



From the single Windows client, you can see two SMB shares from both Isilon cross the regions.



Now you can easily copy the files between those two SMB shares.



Testing showed correct data access to the active file system in each location from any assigned VM located in either Singapore or Sydney.

9 Conclusion

Dell EMC PowerScale storage located within a cloud connected managed service provider facility, such as Equinix, provides off-prem multi-cloud solutions for permanent and short-term high-performance unstructured data applications. The testing and results prove high-performance connectivity into all the major hyperscale providers is simple and effective using Equinix Cloud Exchange Fabric™ (ECX Fabric™).

9.1 Microsoft Azure results summary and observations

The combined Dell EMC PowerScale, Microsoft Azure, and Equinix solution provides customers with options to leverage virtual compute resources and value-added applications within the Azure cloud environment. Connecting to a high-performance filesystem based on PowerScale OneFS hosted within an Equinix facility, customers now have additional choices on how they obtain value from their datasets proven by the vertical industry test results in section 4. Performance, data sovereignty, data gravity, data security, and data ownership are now managed by the service provider when using a hyper-scale cloud provider solution for their IT needs.

9.2 AWS comparison results summary and observations

AWS offers filesystem options to customers with native file services. PowerScale OneFS looks to provide customers with high-performance file storage connected to cloud resources. OneFS located within an Equinix facility provided consistent performance and scaling over the Direct Connect network link as shown in section 6. As the workloads scaled starting with a single host running a single stream to one Isilon node, moving to two hosts accessing two Isilon nodes, the Direct Connect network link eventually becomes saturated.

The comparison charts in section 7, provide outputs to compare both EFS and Lustre filesystem options with OneFS running on a PowerScale/Isilon H500 four node cluster. The OneFS cluster located within Equinix provided approximately 2x MB/s read performance and between 3x and 4x MB/s write performance when compared to native AWS EFS filesystems for 1 MB and 10 MB file sizes. The Lustre filesystem tests show better performance than the 4-node, single chassis Isilon H500. OneFS can scale up to 63 chassis (252 nodes per cluster) each offering up to 5 GB/s throughput per chassis offering extreme scale for high demanding workloads.

The tested FSX1 filesystem consisted of a 1.2 TiB volume which has a 240 MBps throughput limit. In order to maximize performance, AWS recommends striping files over multiple OST creating additional management and cost overheads for deployments. The FSX2 filesystem was created as a 24 TiB volume with Figure 5, section 7 showing the maximum performance capability.

9.3 Recommendations

Choose a cloud connected managed service provider, such as Equinix, located as close to the public cloud resources as possible. The further the distance, the greater the latency observed at the host.

A Technical support and resources

- Dell EMC PowerScale family of scale-out NAS storage including Isilon
  - https://www.dellemc.com/en-au/storage/isilon/index.htm
- Dell EMC S4148F-ON Network Switch
  - https://www.dell.com/support/home/us/en/04/product-support/product/networking-s4148f-on/

A.1 Related resources

- Equinix Cloud Exchange Fabric
  - use the global url located at: equinix.com/interconnection-services/cloud-exchange-fabric/
- Equinix Data Centers and Co-location
  - https://www.equinix.com/services/data-centers-colocation/
- Microsoft Azure ExpressRoute
  - https://azure.microsoft.com/en-us/services/expressroute/
- Microsoft Azure Disk Storage
  - https://docs.microsoft.com/en-us/azure/virtual-machines/windows/disks-types#premium-ssd
- Microsoft Azure Virtual Machines
  - https://docs.microsoft.com/en-us/azure/virtual-machines/windows/sizes
- Oracle vdbench
  - https://www.oracle.com/downloads/server-storage/vdbench-downloads.html