SAS® GLOBAL FORUM 2021

Paper 1150-2021

# Running SAS Workloads Using Clustered File Systems on Amazon Web Services

Dilip Rajan, Amazon Web Services

## ABSTRACT

High-performing SAS applications in the cloud have different design patterns and storage options. Having the right architecture is critical in a customer's transition to the cloud, as setup and maintenance of poorly performing applications can be time-consuming and expensive to mitigate. With Amazon Web Services, customers have a variety of storage and compute options, but getting the right compute and storage configuration is vital to ensure success and a smooth transition. Amazon FSx and Amazon EFS provide simple, scalable, fully managed network file systems that are well-suited to provide the shared storage required for most customer applications. For customers that want to lift and shift their existing on-site SAN architecture to AWS without refactoring their cluster-aware file systems, such as Red Hat Global File System 2 (GFS2) or Oracle Cluster File System (OCFS2), another option is to use Amazon EBS volumes and the Multi-Attach feature. This feature can be used for highly available shared storage by using a cluster-aware file system (such as GFS2) that safely coordinates storage access between instances to prevent data inconsistencies. This paper is for customers who want to build highly available applications using clustered storage on Amazon EBS volumes. This paper will also walk through the process of setting up GFS2 using Multi-Attach enabled EBS volumes attached to multiple EC2 instances that are part of a Linux cluster and illustrate performance considerations along with best practices for the latest EBS volumes for SAS workloads.

## INTRODUCTION

Amazon Web Services (AWS) is the largest cloud provider with the most comprehensive depth of features and capabilities. In addition to unmatched operational excellence, AWS also has millions of customers who are constantly demanding more features and capabilities from AWS. To meet their demands, AWS innovates at a much faster clip and this is evident in the features that AWS rolls out frequently.

SAS Workloads including SAS Grid and SAS Viya demand the highest performance from their storage solutions and setting up one on-premises requires customers to pay for up front cost of a shared file system or a storage area network (SAN) along with support and maintenance even if the storage is not completely utilized. To solve this, AWS launched Amazon Elastic Block Store (EBS) IO2-Block Express, the first SAN in the cloud, that can be provisioned on-demand. In this paper, we shall outline the key performance characteristics of EBS IO2-Block Express that will exceed the performance requirements of the SAS workloads and will also outline how IO2-Block Express could be used in clustered file systems making your SAS workloads resilient to node failures.

## AWS SETUP AND BLOCK VOLUMES

Amazon EBS offers six different volume types to balance price and performance, all of them are optimized as block storage to meet the demands of throughput and transaction

intensive workload at any scale. Customers can run their mission critical systems with requirements from single digit millisecond latency to parallel distributed applications such as SAS Grid manager. In addition to high performance, EBS volumes are replicated within an Availability Zone (AZ) and with EBS snapshots can be backed up to Amazon S3 ensuring business continuity and data protection. It's because of this additional layer of durability that EBS volumes are preferred to NVMe ( Nonvolatile Memory) or ephemeral storage.

## SAS AND EBS

As mentioned above AWS offers 6 different types of the volumes

1. Throughput Optimized HDD (st1) – For Big data, Data warehouses, Log processing
2. Cold HDD (sc1) – Throughput-oriented storage that is infrequently accessed with lowest cost
3. General Purpose SSD (gp2, gp3) – Low latency interactive apps, development and test environments
4. Provisioned IOPS SSD (io2, io1) – Workloads that require sustained IOPS performance

For SAS 9.4 and SAS Grid, AWS and SAS have jointly recommended Throughput Optimized HDD (st1) volumes in a RAID 0 configuration. With a RAID 0 config, customers are able to achieve an aggregate throughput b/w 75 MB/s per physical core – 100 MB/s per physical core. We have also recommended that customers can provision IO1 type volumes with provisioned IOPS but would also need to have more than 2 volumes to achieve the aggregate throughput needed. In addition to the above EBS volumes, customers would be required to select instances with high EBS bandwidth and a min of 8 GB per physical core. As outlined in table 1, Amazon EC2 family of instance of type I3en, M5n, R5n are best suited for SAS workloads based on their throughput and memory specifications.

| Workload | Instances | Storage Recommendation |
|---|---|---|
| SAS 9.4 | i3e(n),r5(d)n,m5(d)n | EBS – ST1 12.5 TB, IO1 - 12 TB, 32KIOPS |
| SAS Grid | i3e(n),r5n(d),m5n(d) | FSx for Lustre – 100 TB(persistent) |
| SAS Viya | i3e(n),r5(n),m5(n) | EBS – ST1 12.5 TB IO1 - 12 TB, 32KIOPS |

***Table 1 Current SAS on AWS Recommendations***

## AMAZON EBS IO2 – BLOCK EXPRESS

During reinvent 2020, AWS launched IO2-block express, the first SAN in the cloud. This new volume type was released to support the most demanding IO workloads such as SAS workloads. IO2-BlockExpress was designed to provide high throughput and minimize the need to RAID multiple EBS volumes together.
IO2-block express is built on a new architecture that takes advantage of advance communications protocols implemented by the AWS nitro system. In this modular storage system, SRD (scalable reliable datagrams) are implemented using custom-built, dedicated hardware, making communication between block express volumes and nitro-powered EC2

instances fast and efficient. IO2-block express can give you up to 256KIOPS with a maximum capacity of 64 TiB in a single volume.

## AMAZON EC2 – R5B

R5b family of EC2 instances is the latest generation of the compute instances under the R5 instance umbrella. R5 instances with EBS have been used by customers for large relational database workloads such as ERP systems and health care systems and rely on EBS to provide scalable durable and highly available block storage. While, the R5 instances provide sufficient storage performance for many use cases, some customer still require higher EBS performance on EC2.

With the new R5b instance, customers can now achieve upto 60 Gbps of bandwidth with 260K IOPS providing 3x higher EBS optimized performance compared to R5 instances. And since R5b and R5 instances have the same vCPU ( /2 physical cores) to memory ratio, SAS workloads as outlined in table 1 will still guarantee the memory required for in-memory processing.

| Instance Name | vCPUs | Memory | EBS Optimized Bandwidth (Mbps) | Max MB/s |
|---|---|---|---|---|
| r5b.large | 2 | 16 GiB | Up to 10,000 | 1,250 |
| r5b.xlarge | 4 | 32 GiB | Up to 10,000 | 1,250 |
| r5b.2xlarge | 8 | 64 GiB | Up to 10,000 | 1,250 |
| r5b.4xlarge | 16 | 128 GiB | 10,000 | 1,250 |
| r5b.8xlarge | 32 | 256 GiB | 20,000 | 2,500 |
| r5b.12xlarge | 48 | 384 GiB | 30,000 | 3,750 |
| r5b.16xlarge | 64 | 512 GiB | 40,000 | 5,000 |
| r5b.24xlarge | 96 | 768 GiB | 60,000 | 7,500 |
| r5b.metal | 96 | 768 GiB | 60,000 | 7,500 |

***Table 2 – R5b instance types***

## HOW TO SETUP IO2 BLOCK EXPRESS

Table 2 provides a list of all R5b instances with the vCPU, Memory and EBS Optimized Bandwidth. First let's go through the steps of setting up the IO2 block express EBS volume

1. Open the Amazon EBS Console
2. Select "**Create Volume**"
3. Under **Volume Type** – Select "***Provisioned IOPS SSD(io2)"*** ( By default, if you select IO2 a block express volume would be provisioned for you.

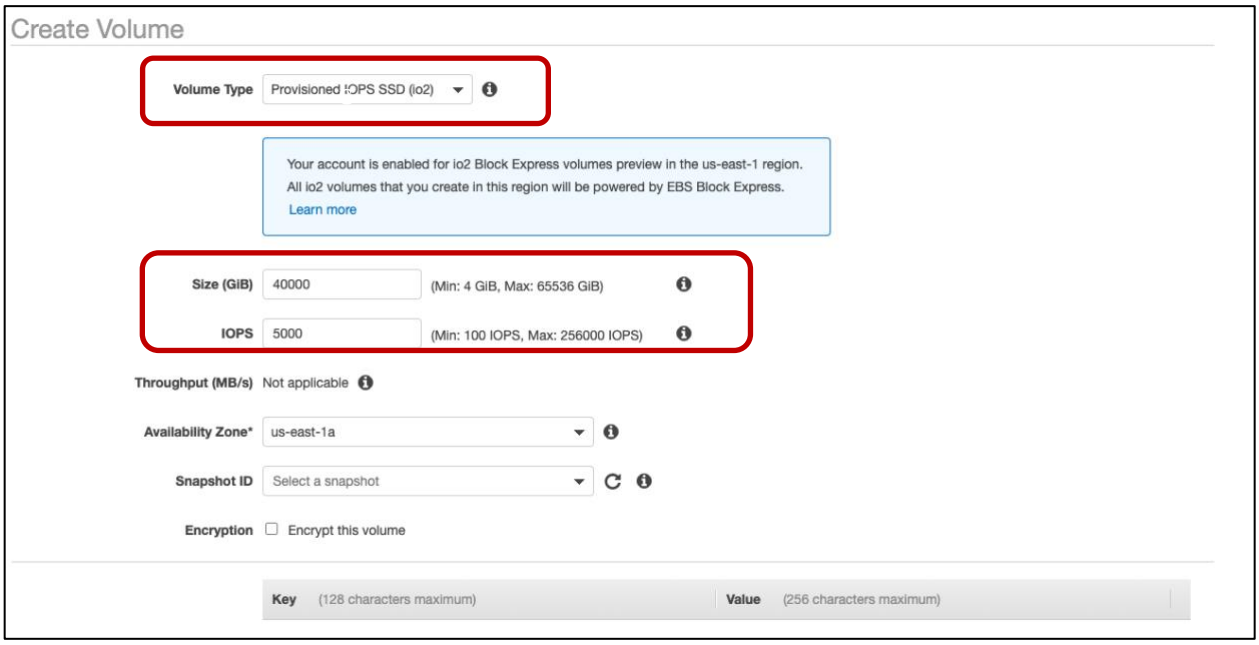


***Figure 1 – EBS IO2 Block Express***

4. For SAS IO test, we provision a volume of size xx TiB. Hence in the **Size (GiB)** text field enter "**xxxxx**" (depending upon your instance memory size and storage requirements)
5. Under **IOPS**, enter "**xxxx**" (enter a number based on a value from BEST PRACTICES section)
6. Select the appropriate **Availability Zone** as your SAS workload's placement group
7. Click on "Create Volume"

After creating the volume attach the volume to the appropriate EC2 R5b instance. To create an EC2 R5b instance, we select RHEL ( Red Hat Enterprise Linux) 8.0 as the operating system and leave the rest of the options as default.

## PERFORMANCE CONSIDERATIONS

We tested the performance against multiple R5b instance types with IOPS configurations between 2KIOPS to 64KIOPS to arrive at the best practices and recommendations. Before we get into the details, let's first take a look at the IO pattern of the tests mentioned above.

SAS Performance engineering team provides a RHEL IO script to measure the performance of the compute instance against the targe file system. The script launches multiple parallel IO processes, one for every physical core in the server. Each process performs read and write operations on a provisioned storage volume. IOs issued by this benchmark are of 64KiB in size, providing a clear IO pattern and has been used to validate IO2 block express.

The maximum IO size supported by IO2 block express is 256KiB and by HDD volumes is 1024KiB. Since the benchmark issues sequential IOs of 64KiB they can potentially be merged into larger ones. For example, 4 contiguous writes of 64KiB IOs can be merged into a single 256KiB IO. This means that this IO pattern is very likely to require less IOPs to achieve high throughput.

We also know that throughput is governed by the instance type. For ex – r5b.4xlarge can deliver a maximum EBS throughput of 1250 MB/s while r5b.12xlarge can deliver 3750 MB/s. Therefore, the minimum provisioned IOPS for maximum throughput is calculated by maximum throughput per instance divided by the max I/O size.

$$MIN\ IOPS\ for\ MAX\ THROUGHPUT = \frac{MAX\ THROUGHPUT\ PER\ INSTANCE}{MAX\ I/O\ SIZE\ OF\ EBS\ VOLUME}$$

For Example, $MIN\ IOPS\ for\ r5b\ 4x.large = {}^{1250\ MB\ per\ second}/_{256\ KiB} = 4882.8\ IOPS$

## PERFORMANCE VALIDATION

Now that we had a good understanding for IOPS configuration for each volume type, we can focus on some specific instances in the R5b family.

We knew that SAS also recommends to have at least 8 GB of RAM per physical core along with a 1: 4 VCPU to Memory ratio. Also, based on the past customer deployments, we concluded that r5b.4xlarge, r5b.8xlarge and r5b.12xlarge instance types would be most frequently used instances for a SAS compute / CAS deployment option.

Once we provisioned the volumes based on the minimum required IOPS, volumes can be mounted as per the docs here and RHEL IO tests were executed against those volumes.

```
Bash>> ./rhel_io.sh -t /nvme1n1
```

Based on initial tests, following were our observations as shown in Table 2

| Instance Type | Memory (GB) | EBS Volume | Storage (TB) | IOPS | RAID | Read Throughput (MBps/core) | Write Throughput (MBps/core) |
|---|---|---|---|---|---|---|---|
| r5b.4xlarge | 128 | io2-bx | 40 | 5K | None | 150.29 | 150.30 |
| r5b.8xlarge | 256 | io2-bx | 40 | 10k | None | 152.88 | 152.59 |
| r5b.12xlarge | 384 | io2-bx | 40 | 15K | None | 152.90 | 152.64 |

***Table 3 Performance Test Results***

As observed from the results above, for both read throughput and write throughput IO2 block express is able to provide above 150 MBps/ per physical core, far exceeding the performance benchmark requirement of 100-125 MBps/per physical core. This was a significant improvement over the previous recommendations where up to 8 volumes had to stripped together to get this type of performance on AWS.

We also observed that for higher instances of size r5b.16xlarge and r5b.24xlarge, we did need to stripe 2 volumes in RAID 0 config to achieve the same performance, as the EBS throughput delivered by these instances is higher than the maximum throughput a single IO2 Block Express volume currently supports, 4000 MB/s

## BEST PRACTICES

Based on our observations following best practice would yield the maximum performance:

For r5b.4xlarge, r5b.8xlarge and r5b.12xlarge with a single IO2 block express volume:

1. Provision the minimum number of IOPs required to achieve the maximum throughput of the instance: PIOPs = MAX_TPUT_PER_INSTANCE / 256KiB . For example, with r5.12xl the minimum io2-bx PIOPs = 3750MBps/256KiB = 14,305
2. Set the IO2-block express volume's block device read_ahead to 256 KiB in order to amplify block sizes on reads and maximize throughput:

   ```
   echo 256 > /sys/block/nvme1n1/queue/read_ahead_kb
   ```

   assuming nvme1n1 is the IO2-block express volume's block

For r5b.16xlarge and r5b.24xlarge with IO2-Block Express:

1. Provision two volumes, each with half the minimum IOPs required to achieve maximum instance throughput: PIOPs/volume = (MAX_TPUT_PER_INSTANCE / 256KiB) / 2.
2. Aggregate the volumes with MDADM striping using a stripe size of 1MB

   ```
   sudo mdadm --create --verbose /dev/md0 --level=0 --name=md0 --
   chunk=1024 --raid-devices=2 /dev/nvme2n1 /dev/nvme1n1
   ```

   ```
   sudo mkfs.xfs /dev/md0
   ```

Based on your initial test results, we recommend retesting after increasing/decreasing the number of IOPs by 10% to find out what the optimum provisioned IOPs that satisfy your throughput *requirements*

## COST CONSIDERATIONS

Setting up on a SAN on the cloud can be expensive, and this can be further elevated with additional maintenance and support. With the on-demand nature of pricing on AWS, IO2-block express can be provisioned as needed thereby switching the high cost upfront investment to an amortized as needed pricing model.

For IO2-block express, customers would need to pay a flat amount by storage volume along with a tiered amount provisioned IOPS. Following are the exact numbers:

Storage costs -- $0.125/GB-month

Provisioned IOPS -- $0.065/provisioned IOPS-month up to 32,000 IOPS

Provisioned IOPS -- $0.046/provisioned IOPS-month from 32,001 to 64,000 IOPS

Provisioned IOPS -- $0.032/provisioned IOPS-month for greater than 64,000 IOPS

## CLUSTERED FILE SYSTEMS

With IO2-block express, it is also possible now for customers to setup their favorite clustered file systems like GFS2 on AWS. To enable this multi-attach capability is being tested and will release soon and this blog will help customers set up their favorite file

systems like GFS2 on IO2-block express. Below architecture highlights the setup on multi-attach enabled volumes.

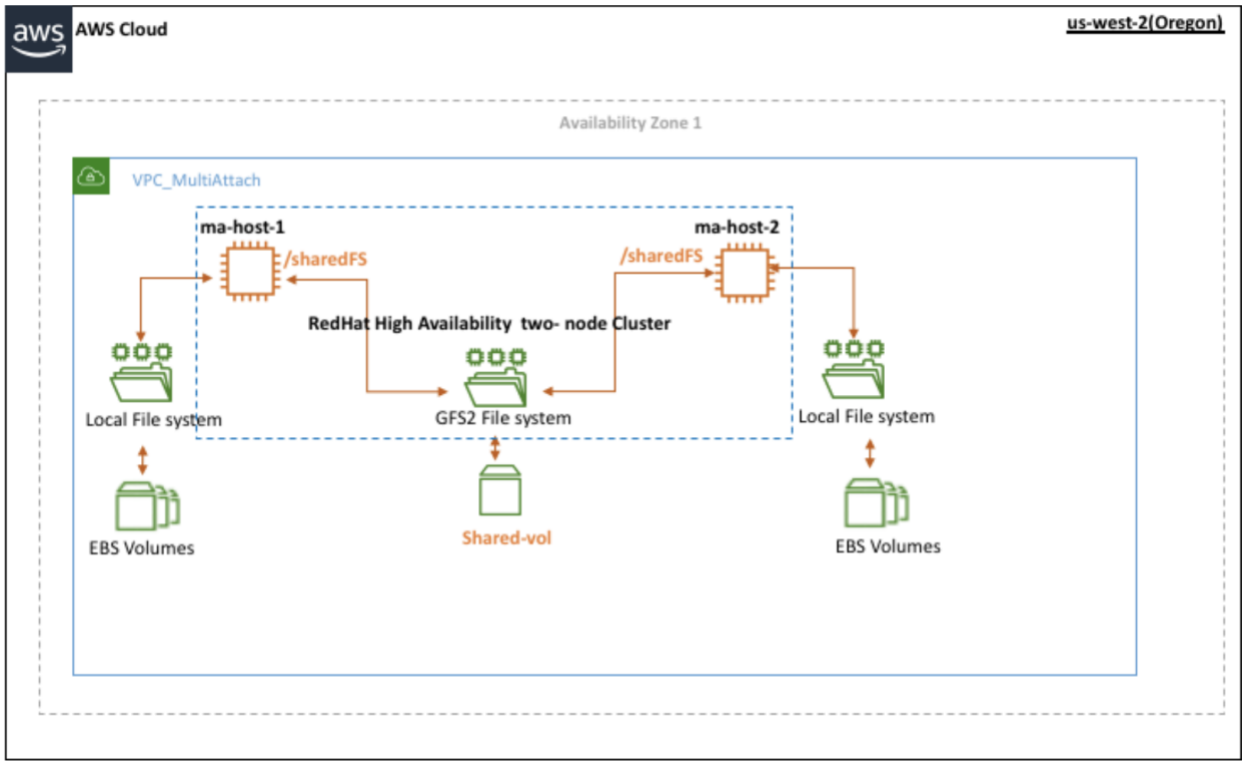


***Figure 2: GFS2 Architecture***

## CONCLUSION

In this paper, we were able to outline how the latest EBS volumes from AWS such as IO2-block express can exceed the performance requirements for SAS workloads. Not only is IO2-block express an excellent choice for a SAN from a performance perspective but also provides customers with strong cost benefits. Additionally, IO2-block express can be also configured to build a clustered file system where multiple nodes might need to communicate with a shared persistent storage to reduce the impact of node failures. Although, customers have many options for persistence storage such as Amazon S3, Amazon EFS and Amazon FSx/Lustre, AWS believes in customer choice and provides another very valuable feature in terms of EBS – IO2-block express to meet the demands SAS workload requirements.

## REFERENCES


https://aws.amazon.com/blogs/storage/clustered-storage-simplified-gfs2-on-amazon-ebs-multi-attach-enabled-volumes/

https://support.sas.com/kb/59/680.html

https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ebs-using-volumes.html

https://docs.aws.amazon.com/prescriptive-guidance/latest/migration-sas-grid/welcome.html

https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/volume_constraints.html

https://aws.amazon.com/blogs/aws/new-amazon-ec2-r5b-instances-providing-3x-higher-ebs-performance/


## ACKNOWLEDGMENTS


Sammy Frish, Sr Systems Engineer, Amazon Web Services

Nishit Nagar, Sr Product Manager, Amazon Web Services

Venky Nagapudi, Principal Product Manager, Amazon Web Services

Suney Sharma, Solution Architecture Manager, Amazon Web Services

Jim Kuell, Sr Software Engineer, SAS Institute

Margaret Crevar, Sr Performance Engineering Manager, SAS Institute


## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:


Dilip Rajan
Amazon Web Services
rajand@amazon.com