SAS® GLOBAL FORUM 2021

Stretch Your Mind And Your SAS® Grid

Diane Hatcher, Core Compete/Accenture

ABSTRACT

Deploying the SAS grid in a cloud provides the ability to use cloud capabilities to automatically scale your grid based on user demand. Grid nodes can be added as user volume increases and then scaled back down when the demand decreases. This paper will discuss how this can be done across all cloud vendors, including Azure, AWS, GCP and OpenShift. Deploying the SAS grid in the cloud can help you better manage TCO and storage costs without an upfront hardware investment. In addition, the SAS grid can also be complemented with SAS running in containers. There will also be a discussion on the future of parallel processing, where containers play a leading role.

INTRODUCTION

SAS Grid is a technology that has been around for many years, providing significant performance benefits for large-scale SAS usage. Essentially, SAS Grid is a set of SAS Foundation servers that are managed together as a single ecosystem. It is deployed across a number of servers or virtual machines (VMs) that sit on top of a shared file system. Each compute VM can be considered as a Grid node. The shared file system is typically a high-speed clustered filesystem, such as IBM Spectrum Scale or DDN EXAScaler (formerly Lustre). The shared file system allows the same files to be accessed by any node on the Grid using the same physical pathname.

SAS Grid is a great architecture for orchestrating and executing SAS workloads, and it is inevitably used to support many production-type workloads. These critical workloads deliver analytics, scores, and decisions to core operational processes across the organization, and they are relied upon to perform consistently when needed.

SAS GRID OVERVIEW

There are 2 main workload patterns for the SAS Grid.

First, there is the ability to break up a single SAS job into multiple tasks and run each task on separate nodes in parallel. This type of parallel-processing is helpful when users run the same set of code against different slices of data. For example, when forecasting demand for each product within a retail catalog, the same workload can be run for each individual product or SKU. Each execution can run on a different node in parallel, and the shared file system supports concurrent access to the data. Obviously, running the forecasts in parallel will complete much quicker than running them in sequence, one after the other.

The second pattern is what I call, "multi-tenant". You can configure the Grid to provide specific resources to different teams in the organization. For example, the Marketing team could access one set of servers, while the Finance team can access another set. This is especially useful when different teams have different resource demands. This is a very common configuration for SAS Grid deployments.

ON-PREMISE SAS GRID CHALLENGES

The challenge arises, then, when there is an increase in demand for resources. The SAS Grid is typically sized for an expected workload. If demand exceeds that capacity, performance of SAS jobs can be greatly reduced and effect other users on the Grid.

Especially in a multi-tenant setup, demand for SAS resources can fluctuate significantly – either because of regular analytic or reporting requirements, or because of special projects. We've seen Covid19 also impact many organizations by generating new analytical projects to understand and handle the changing business landscape. Coupled with overall usage increases, consistency of performance can become an issue.

SAS Grid, as I noted before, is architected to orchestrate across separate compute hardware and a shared file system. The upside is that performance is greatly enhanced. The downside is that this type of on-premise infrastructure cannot be adjusted quickly.

Organizations cannot dynamically respond to quick increases in the demand. Expanding the grid can take several months or weeks, as it means procuring and adding more hardware. This means

- additional capital expense and operating costs,
- delays waiting on hardware and configuring for your environment, and
- additional data center operating expenses

If the Grid is expanded to meet peak requirements, there is wasted capacity and costs during normal workload levels.

CLOUD INFRASTRUCTURE

So, let's talk a little bit about clouds and cloud infrastructure. As you know, the main cloud providers are Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform (GCP), and Redhat OpenShift. There are some nuances between what these vendors provide, but at a high level, the architectures are very similar.

The key elements for SAS Grid in the cloud are the same as an on-premise deployment. You need to setup Grid nodes using instance types with fast CPUs and network bandwidth. These instances need to be located in the same proximity zone, so they are physically close together to minimize network latencies. SAS Grid can also support different instance sizes, but it is important to use the same chipset across all Grid nodes.

All cloud vendors support automation for launching new instances and deploying software. This automation allows you to easily provision and configure instances as you need them.

You also need to configure a shared file system that is accessible to all Grid nodes. This shared file system needs to be configured to provide the I/O requirements needed for optimal performance. Good options are Lustre, Netapp, and GPFS IBM Spectrum Scale.

Cloud storage buckets are also important, as these represent low-cost storage options for large amounts of data. Cloud buckets don't provide the same high data transfer speeds that shared file systems provide. But, these are great options for storing data archives and user files.

SAS GRID IN CLOUD

All the major cloud vendors have the ability for you to deploy SAS Grid on VM instances with a shared filesystem that conform to the on-premise deployment requirements. SAS has provided excellent details on which instance types and storage options work great for each cloud vendor. Regardless of the cloud vendor, using cloud infrastructure has these benefits:

- No hardware purchases are required for servers and storage.
- No data center operating expenses for maintaining the hardware.
- If additional capacity is needed, you can easily provision more capacity with virtually no wait time.

With SAS 9.4m6 or later, SAS has made extending the SAS Grid much more seamless without impacting SAS users actively running jobs. You are able to add nodes to the Grid and register them with the Grid Controller without having to restart or pause the Grid.

Because of this capability, the cloud is a perfect place to deploy the SAS Grid.

SAS GRID IN CLOUD - MANAGED SCALABILITY

Remember that demand for SAS Grid resources can fluctuate significantly – either because of regular analytic or reporting requirements, or because of special projects. For example, Regular reporting requirements can use significant resources once a week, month, or year. Updates to pricing lists can require heavy short-term forecasting workloads.

With a cloud-based deployment, you can make your SAS Grid elastic on a managed basis by scheduling additional capacity to address known peaks in demand, then scaling back down when the peaks are over.

By leveraging automatic deployments, you can add additional capacity on a scheduled basis. Then, you can scale the grid back down by quiescing the recently added a Grid node, allowing existing jobs to finish, then shut it off completely. This allows you to add capacity without having to purchase hardware or pay for capacity when it's not needed.

You can also take advantage of "bursting" SAS jobs out of the SAS Grid. For individual projects that require short-term capacity, you can schedule these jobs to run completely outside of the Grid by leveraging containers to run project workloads as batch jobs. This is a great option for "ephemeral" type, or temporary, jobs versus extending the Grid for longer-term additional capacity. The objective is to remove these workloads off the grid altogether – allowing them to run in temporary containers that shut down as soon as the job is finished. You only pay for the container for the time it takes the job to run. This also helps to maintain a consistent load on the Grid by eliminating short-term, resource-intensive jobs.

IAAS BENEFITS

With cloud providers, you can take advantage of:

- Only paying for resources when you use them.
- Scripting to launch new server instances.
- Scheduling tools for running the scripts to launch new server instances.

The latest SAS Grid deployed on cloud, then, you can

- Add capacity on a dynamic basis without impacting other SAS users or SAS jobs already running on the Grid.
- Gracefully remove extra capacity by allowing active jobs to finish on a node and then shutting it down.
- Take advantage of additional cloud options to provide extra capacity when needed.

SAS GRID IN CLOUD - STORAGE MANAGEMENT

Another nice feature of cloud providers is the ability to easily manage storage requirements and minimize costs associated with storing data. SAS Grid users tend to create a lot of permanent data, so managing storage costs are top of mind for IT administrators.

Just as you do on-premise, you also need to configure a shared filesystem in the cloud. Lustre, for example, is available across AWS, Azure, and GCP, and is a solid choice for SAS Grid performance.

Shared file systems are needed for performance and concurrent acces, but they are not elastic like cloud storage and represent a more costly option.

However, in the cloud, you only need to configure a large enough shared storage to support your "hot" data for active workloads. Data that is not needed right now does not have to be stored in the shared file system.

Cloud vendors also provide low-cost storage options, such as AWS S3, Azure Cloud Storage, and Google Cloud Storage. Using cloud buckets with your shared file system allows you to better manage overall storage costs. Warm and cold data can be maintained in the cloud bucket and typically represents the bulk of your data assets. You can:

- Automatically synch data between shared storage and cloud storage.
- Enable users to seamlessly retrieve data from cloud storage into shared storage.
- Automatically move data out of shared storage into cloud storage as it cools due to inactivity.

MIGRATE SAS GRID TO CLOUD

What is the best way to move your Grid to the cloud? One option is to completely replicate your current environment in the cloud – exactly mimicking your existing infrastructure and storage, then migrating all SAS code and data from on-premise shared storage to cloud shared storage.

However, this is not necessarily the most cost effective option, nor is it necessarily the fastest way to move. Shared storage, as noted earlier, can become expensive for large storage clusters. If you have lots of production-type jobs, the validation of all the jobs could take a lot of time to ensure there is no disruption to your business. The net result, then, is that you are running essentially 2 identical production systems until validation is complete and keeping everything in synch will be complicated and risky.

A more common approach, with less risk, is to move data and workloads in batches or phases, such that you can encapsulate specific processes and/or departments and move them together. The validation of smaller batches can be done more quickly, and these workloads can move off the on-premise Grid in phases. While the cloud Grid can be expanded as migration progresses, the on-premise Grid can be downsized to remove excess capacity.

You can also build up the data pipelines to migrate just the needed data in each phase to the cloud Grid environment. Data pipelines can be used to help keep this subset of data in synch until the cloud validation is completed. Data pipelines can also be used to hydrate cloud data stores from enterprise data lakes, regardless of where the data lake is on-premise or in the cloud. Data pipelines can help to provide data continuity, even if the data lake moves or changes.

To determine how to phase your migration, it is best to perform a detailed assessment of your existing environment. This is more than taking inventory, although that is an important piece of information. You also want to understand the nature of your workloads and the data/execution patterns they represent. You should:

1. Understand what your workloads are doing – such as data prep, reporting, analytics

2. Analyze your data pipelines to understand what data is being accessed, written, and shared

3. Categorize and identify logical groupings of workloads by application and/or department.

These details you gather are used as input into developing your migration roadmap. Your roadmap should identify the best way forward to migrate quickly and to minimize impact on day-to-day operations.

- Only move the data that you need and where you need it

- Ensure production jobs continue to execute when they are needed

- Provide a consistent user experience to quickly drive adoption and migration of ad-hoc workloads.

CONCLUSION

If you are running SAS Grid today on premise, migrating it to the cloud should be considered as a next step. As your infrastructure begins to age, you should weigh the benefits of refreshing your on-premise infrastructure against a cloud-based deployment. Migrating your Grid to the cloud allows you to:

- Retain all the SAS assets your organization has already created
- Quickly configure a baseline Grid environment without having to purchase or wait on hardware
- Allow you to expand quickly as cloud adoption increases within your organization.

RECOMMENDED READING

Crevar, Margaret. 2020. Important Considerations When Moving SAS9® To A Public Cloud. *Proceedings of the SAS Global 2020 Conference.* Cary, NC : SAS Institute. Available at https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2020/4312-2020.pdf

CONTACT INFORMATION

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

Diane Hatcher
Core Compete/Accenture
diane.hatcher@corecompete.com
https://www.corecompete.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.