Paper 2752-2018

Frequently Asked Questions about Getting Started with SAS® Grid

Andy Peredery, Nick Welke, and Ivan Gomez
Zencos Consulting

ABSTRACT

Scalability. Fault tolerance. Load balancing. High performance. High Availability. SAS® Grid Manager. These are all phrases that commonly show up during analytics infrastructure conversations. However, maybe you are still confused about how all this relates to SAS Grid and what it would mean for your organization.

When considering SAS Grid, many customers have similar points of confusion about how and if their SAS workflow would change, how do they connect to the Grid, how do you manage it, what do they need to successfully implement it. In this paper, we cover questions we are asked most often about implementation, administration, and usage of a SAS Grid environment.

INTRODUCTION

This paper answers some of the common questions we receive when we work with customers who want to implement SAS Grid. It is intended for users, managers or SAS administrators considering modernizing their SAS platform by adding SAS Grid.

WHAT IS SAS GRID?

Simply put: SAS Grid is a distributed computing environment based on Platform Computing's Load Share Facility (LSF) job scheduling technology (owned by IBM as of 2012) that has been combined with SAS software.

SAS Grid allows a user to batch submit a SAS job to a job queue associated with a pool of computers, and the LSF loading balancing algorithm will schedule which computer will run the SAS job, based on a variety of factors, including computer resource availability (RAM, CPU, Disk usage), time of day, or other dependencies and conditions. This process is illustrated in Figure 1 below.

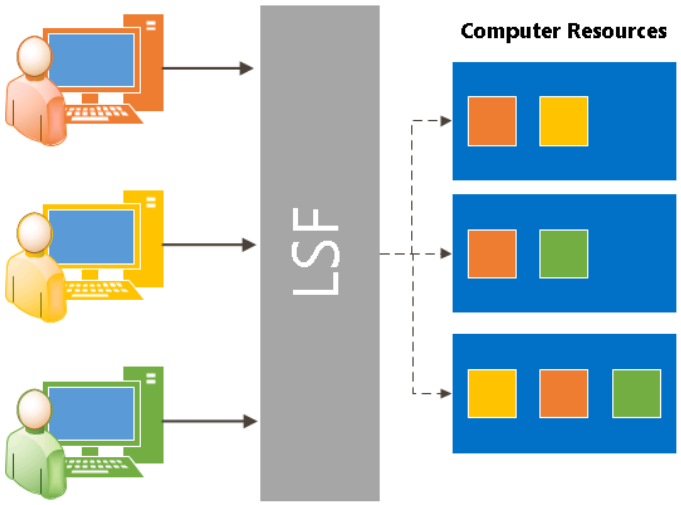

Figure 1. Flow of Job Submissions in a SAS Grid

SAS Grid is much more powerful than locally installed PC SAS since you can leverage the larger computing power of server-class machines, as compared to PC SAS which simply utilizes your local PC or laptop resources. SAS Grid also differs from a typical SAS BI Server installation which usually only has a single compute server. SAS Grid LSF job scheduler on the other hand, has access to a set of compute nodes.

I'VE HEARD ABOUT HADOOP. HOW DOES IT COMPARE TO SAS GRID?

HADOOP is another form of distributed computing. The key difference between HADOOP and SAS Grid is how the parallel processing of job gets distributed and run across the compute nodes. HADOOP typically processes a job's data in parallel across multiple worker nodes and consolidates the "results" at the very end on a client node, whereas a SAS Grid job only executes on a single (best) compute node, although the data can be located across multiple nodes. Both HADOOP and SAS Grid allow for multiple concurrent/parallel jobs to execute at the same time.

The two technologies fundamentally differ in how data is managed and attempt to solve dissimilar computing issues using contrasting approaches.

HADOOP attempts to speed access to data by loading it equally across multiple nodes. HADOOP is used to store massive amounts of data, such as data sets or tables in the terabyte or larger range. Organizations typically try to use HADOOP as a "data lake", and in the process end up expending large resources on data governance. SAS Grid, on the other hand, is traditionally used to improve computing power for an existing organization by consolidation or re-purposing computing resources. Data is normally left alone on the databases or file servers where it currently resides – it is left as-is, or with only a cursory folder re-organization on the existing servers; unlike HADOOP, which requires the extra step of a larger effort of re-distributing and loading the data into a HADOOP Distributed File System (HDFS). HDFS is installed on-top of an existing (UNIX-based) filesystem. Depending on data access requirements, some SAS Grid implementations may require upgrading the file server's filesystems to a clustered filesystem such as IBM's GPFS.

WHAT ABOUT USING SAS MP CONNECT?

SAS MP Connect is used to break a single SAS program up into multiple tasks that can be executed in parallel on one or more compute servers. MP Connect is the most granular level of optimizing the execution of a SAS program, and is typically done by the programmer themselves, although SAS Enterprise Guide (EG) has some capabilities of automating this parallelization process.

Programs are typically modified by including RSUMBIT…ENDRSUBMIT blocks of code (that are remote submitted to another server), with WAIT and WAITFOR statements interspersed to act as blocking semaphores. The primary goal when using MP Connect, is to ensure parallel access to the data is optimized without creating dead-lock or live-lock situations. From a distributed computing perspective SAS MP Connect, behaves like a poor-man's SAS Grid. SAS MP Connect enabled programs can be run by SAS Grid. SAS MP Connect enabled programs cannot run on HADOOP, but they can be used in SAS programs that retrieve data from HDFS but continue to execute on a SAS compute node.

WHAT IS THE SAS GRID ARCHITECTURE?

SAS Grid consists of the typical 3-tier SAS implementation, but includes additional compute nodes and a special Grid Controller node in the Compute Tier which runs the LSF processes. Specifically, the SAS Grid requires the following logical tiers:

- Metadata Server
- Mid-tier Server
- Grid Controller (GC) Server
- Grid Compute Node(s)

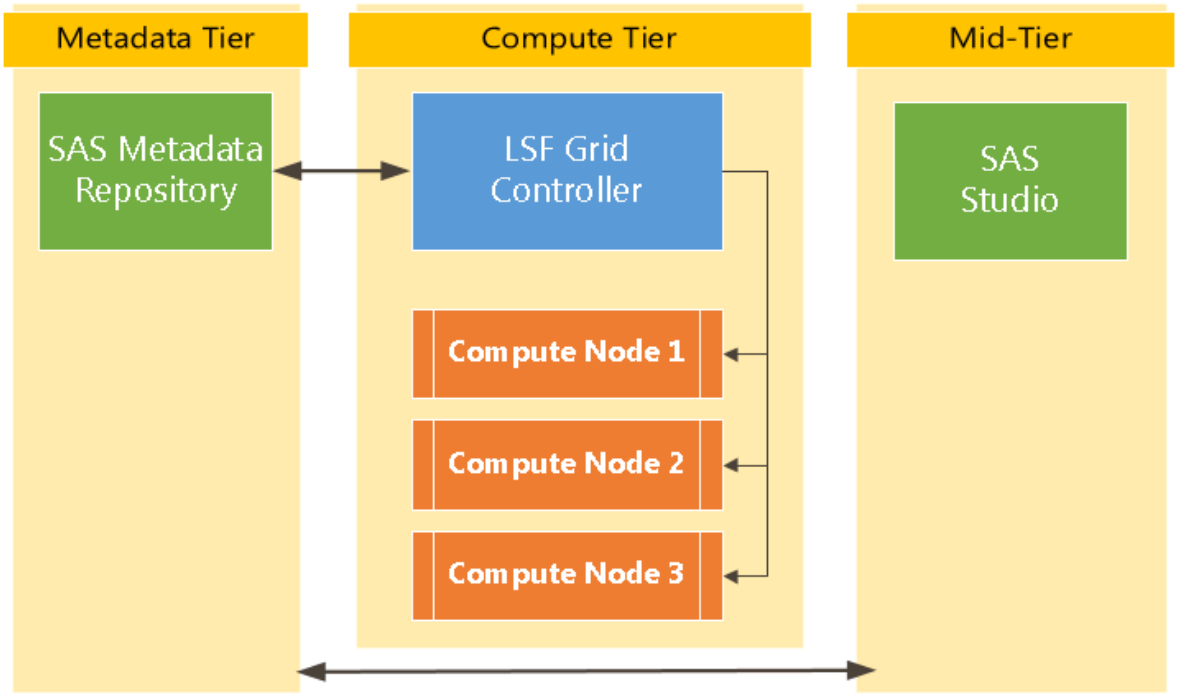

Figure 2. SAS Grid Architecture

Depending on the complexity of the load-balancing and fail-over requirements, the Meta, Mid and GC servers can each be clustered. Grid compute nodes are considered a cluster already.

Grid compute nodes can be heterogeneous; In other words, each node does not have to have the same computing resources (CPU, RAM, storage, etc.), although at minimum it is a best practice to at least have the same operating system installed on each node. This flexibility allows companies to repurpose existing hardware and incorporate it into the Grid. Conversely, HADOOP nodes are typically implemented with stricter restrictions for identical node hardware, especially for worker nodes, although this restriction has been somewhat lifted with more recent releases of HADOOP.

WHY ADD SAS GRID TO YOUR ENVIRONMENT?

Most organizations that are considering a SAS Grid environment, are looking to reduce costs in three key areas.

1. Laptop/Desktop PC SAS licensing costs
2. Support Costs
    a. How much time is spent per user to keep their environment (PC SAS) up
    b. Simplified SAS license renewals
3. Operational Costs
    - capacity planning (charts) – centralized computing resources
    - faster running jobs
    - applying hotfixes/patches

WHAT SHOULD I CONSIDER WHEN PLANNING FOR SAS GRID?

Gathering business requirements for a typical SAS Grid implementation includes collecting information such as:

- Total number of users
- Number of concurrent users (+ expected usage growth)

- Total Data volume – from all sources (datasets, DBMS, CSV files, Web Services, data scrapers, etc. – this is a great auditing exercise, as it reveals a lot of archived sources)
- Existing Jobs (reporting, operational, statistical, interactive, etc. that will be switched over to the Grid – also a great audit exercise)
- Typical Job Runtimes (provides base-line comparisons when jobs eventually run on the grid)
- User training needs (It cannot be over-stated how much transition time and training is required when a significant infrastructure change is made at an organization; rule of thumb is to conservatively estimate the time and resources, then at minimum, double it; tripled estimates are the typical result for a successful implementation)

As always, the more information that is gathered, the better the planning goes. Keep in mind special-use cases when doing your planning.

WHAT DO I CONSIDER WHEN INTRODUCING SAS GRID TO USERS?

Following the installation of the various tiers making up the SAS Grid environment, and completing PAT (stress), UAT, and BAT testing, the next step is to transition users to start using the SAS Grid.

Most environments that transition to a SAS Grid have users that are familiar with Windows, given the predominant Windows-based OS is installed on desktops and laptops. Transitioning to a Linux or other Unix flavored SAS Grid implementation usually presents a massive challenge to many users.

Most IT shops prefer to run SAS Grid on the more efficient and cost-effective Linux platform, since the OS overhead is much less than the equivalent Windows Server, but a SAS Grid can be implemented on Windows-based hardware.

Windows-based Grids have some key advantages over Linux, primarily around ease of data access which ultimately creates nightmares for users not familiar with Linux/Unix environments. The additional time spent by a user to simply retrieve results from a batch submission in Linux--without CIFS or Samba support for the networked storage, can lead to a plethora of ongoing headaches for the support team: From a Grid implementation perspective, the data-access problem is simply being deferred to another team to solve. The simple advantage of mouse-click job submission, continued DDE support for Microsoft Excel spreadsheets or navigating to a network drive/shared folder to access data and results, directly from a users' Windows laptop may out-weight the costs saved by introducing a Linux-based Grid, which does not natively support these functions.

If a Linux Grid implementation is the choice for an organization, then batch submission tools can be custom built to assist users with job submission to the SAS Grid. The Linux advantage is the support of a vibrant open source community that provides a diverse variety of Grid-enabled solutions and capabilities that are not readily found or supported in the Windows world

CONCLUSION

A SAS grid computing environment can bring a lot of power to your organization. This product is good for organizations looking to save operational, license, and support costs. A knowledgeable source can assist with the planning and rollout for the SAS Grid Platform.

REFERENCES

- IBM Platform LSF, https://www.ibm.com/support/knowledgecenter/en/SSETD4/product_welcome_platform_lsf.html

RECOMMENDED READING

- The Top Four User-Requested Grid Features Delivered with SAS® Grid Manager 9.4 Available at http://support.sas.com/resources/papers/proceedings13/470-2013.pdf
- RTM and SASGSUB, the Power to Know®… what your Grid is doing Available at: http://support.sas.com/resources/papers/proceedings12/370-2012.pdf

CONTACT INFORMATION

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

Andy Peredery
Zencos Consulting
aperedery@zencos.com

Ivan Gomez
Zencos Consulting
igomez@zencos.com

Nick Welke
Zencos Consulting
nwelke@zencos.com