Paper 1330-2018

Creating a Multi-Tenant Environment using SAS® Enterprise Guide® and SAS Grid®

Tim Acton - General Dynamics Health Solutions

ABSTRACT

SAS® Enterprise Guide® is a powerful tool with an easy to use graphical interface. Combined with SAS Grid®, it becomes a scalable platform capable of serving a growing number of users. In our environment, SAS tools provide customers with access to large healthcare data through private work areas and custom data products.

In a shared environment, providing external users with access to secure data is challenging. Security is paramount, but there are also challenges with scale and complexity. This paper will outline methods to provide SAS Enterprise Guide users with access to large data sources in a secure, reliable, and scalable way.

INTRODUCTION

The Centers for Medicare & Medicaid (CMS) Chronic Condition Warehouse Virtual Research Data Center (CCW VRDC) is an alternative solution for accessing and analyzing CMS data for research purposes. Historically, CMS has provided data to researchers by preparing and shipping encrypted data files on external media. The CCW VRDC allows a researcher to access and perform analyses and manipulation of CMS data virtually from his or her own workstation. The CCW VRDC provides researchers with a secure mechanism to access timely data in a more efficient and cost-effective manner.

The CMS CCW VRDC is maintained by HealthAPT (a joint venture between NewWave Technologies and General Dynamics). As a trusted health solutions organization for more than 30 years, General Dynamics Health Solutions, part of General Dynamics Information Technology, provides end-to-end solutions and professional services to health organizations in the defense, federal civilian government, state and local government, commercial and international sectors.

The CCW SAS Grid utilizes SAS 9.4 M2 while providing clients with access to SAS Enterprise Guide 7.13 on Virtual Desktops. Current CCW SAS Grid infrastructure operates on Cisco B200 M3 and M4 Blades with 64-bit Linux. For back end storage IBM Spectrum Scale (GPFS) for a clustered file system is used. The SAN is a combination of EMC XtremeIO, VNX, and VMAX block storage.

OVERVIEW

This is a main topic in the body of the paper. This paragraph uses the PaperBody style.

Giving external customers access to a secure data warehouse presents several unique challenges. Users require the ability to run ad-hoc reports against hundreds of Terabytes of sensitive health care data. The level of SAS experience varies greatly (from total novice to power user) necessitating tools that are both powerful and intuitive. Users have the ability to access and create large datasets, sometimes hundreds of Gigabytes, making powerful processing tools and adequate data storage essential.

This paper will cover:

- Justification for the use of SAS Enterprise Guide and SAS Grid
- An overview of CCW infrastructure
- Mechanism for providing users secure access to SAS Enterprise Guide and the CCW data
- Tips and best practices for utilizing SAS Enterprise Guide with Grid in a large environment

WHY SAS ENTERPRISE GUIDE?

Ease of use is the primary reason for SAS Enterprise Guide within the CCW. New innovators and researchers are added to the CCW VRDC continually, including novices to SAS programming. These innovators and researchers pay to use the CCW VRDC resources, as a result efficiency and productivity are important. The ease of use and flexibility of SAS Enterprise Guide means users assimilate to the environment very quickly.

At the same time, SAS Grid provides users immediate access to powerful back-end CPU and storage. Enterprise Guide is configured to automatically log users onto the SAS Grid upon sign on to SAS Enterprise Guide, streamlining the process for the user. On the administrative side, setting up and maintaining Enterprise Guide on Grid is complex, but the benefits to the users and the ability to easily track and load-balance sessions for administrators is crucial. The incorporation of complex autoexec scripts and security measures within the environment has been reasonably straightforward.

INFRASTRUCTURE CONSIDERATIONS

This is a main topic in the paper. This paragraph uses the PaperBody style. This paragraph uses the PaperBody style.

Previously the CCW environment utilized Solaris on Sparc, with SAS 9.3 and SAS Grid, and EMC VMAX enabled storage for approximately 300T of data. An influx of new users in 2015 necessitated an update to the environment, including the following changes.

- SAS 9.3 to SAS 9.4 M2 update
- Transition from 5 Solaris SPARC servers to 14 Cisco B200 M3 and M4 blades running 64-bit Linux
- Change from legacy VMAX block storage system (and Veritas CFS) to an Isilon NFS system*
- Converted SAS data from Solaris to Intel format
- Migrated 300 TB of SAS data
- *Further migration from Isilon to VNX and VMAX block storage using IBM Spectrum Scale (aka GPFS)

Cisco servers are used with Intel processors (Dual socket E5-2697 and E5-2697 Xeon processors). Commonly SAS environments use CPU’s with the highest available clock frequency ensuring a lower number of total cores per socket. In the CCW environment, the number of cores per socket was prioritized due to an extensive   real time user base executing ad-hoc type work. That naturally leads in to the next consideration, I/O throughput.

In a sizable system, providing the SAS recommended 100-150MB per second of I/O per job can be extremely challenging. For example, 300 jobs would require 45GB per second of throughput. A SAN to meet that level of need is prohibitively expensive. To ease the SAN requirements, servers with one 1.5TB PCIE flash card each were implemented. The flash memory card is used exclusively for WORK and UTIL allowing the utilization of a more “modest” SAN. While this solution is effective, occasionally users will fill the work directory, executing a large join, for example.

Coordinating with the move from Solaris to Linux x86_64, was the transition from EMC VMAX to EMC Isilon. Isilon is an NFS storage solution which offers several advantages over traditional block storage. Reduced cost is the primary advantage, and it is easier to provision. Testing revealed that, for long running SAS jobs, better performance was observed on Isilon than on the older VMAX system. However, Isilon proved to be incompatible with other SAS use cases. Issues with file locking were the most common error. EMC was able to resolve the issue eventually, but the issue reappeared after an Isilon OneFS update. In addition to locked files, overall performance of Isilon was poor; small jobs that had taken minutes to run were now processing for hours. Run time for jobs that accessed many files increased significantly. The assignment of large volumes of libraries (sometimes up to 400) means that SAS Enterprise Guide accessed every library before loading. With the Isilon, login times greatly increased and the workspace load balancing timeout was reset to 5 minutes. Performance in the SAS Enterprise Guide GUI was also an obstacle, since Enterprise Guide was executing multiple metadata calls to the file system. (Some filesystem metadata issues with Enterprise Guide were alleviated by a 2017 update. Updates to the current Enterprise Guide client are recommended).

Following the transition to Isilon, all SAS binaries were moved to block storage on EMC XtremeIO. Eventually migration of all data occurred from EMC Isilon to an EMC VNX and later, an EMC VMAX. A brief attempt was made using GFS2 for a clustered file system, however SAS recommends using IBM Spectrum Scale (AKA GPFS) and after implementation of GPFS no issues were observed allowing for the painless mixing of a variety of storage devices.

In 2017, major changes to the backup and recovery strategy commenced. The system currently houses approximately 1.5 Petabytes of data and projected to be in the 2 Petabyte range in 2018. Prior to 2017, CCW had been using a large tape library and Veritas NetBackup. Data was moved offsite to cold storage, and there was no secondary datacenter for replication. At 1 Petabyte of data, the time needed to achieve full backup was prohibitive, and concerns about backups not completing in time for normal business hours were well-founded. Disaster recovery exercises indicated that up to a month may be needed to restore all data.

To mitigate these risks, a two pronged strategy was implemented. First, an IBM Spectrum Protect (AKA Tivoli or TSM) was rolled out. The primary advantage of Spectrum Protect is the avoidance of full backups. With the increasing size and scale of the CCW VRDC warehouse, it became necessary to avoid full backups whenever possible. The large tape library was replaced with a disk backup system engaging large amounts of flash storage and allowing for significantly faster backups and restores.

Replication over WAN to a ‘warm’ offsite datacenter became the offsite storage solution. A caveat with this arrangement is the large amount of data to replicate daily. Since multiple TB of data are updated daily, monitoring the WAN connection is critical.

System performance monitoring utilizes Foglight (see Figure 1.0 below) and allows for the easy custom creation of daily dashboards. For example, system administrators have two dashboards up at all times. One displaying the amount of space being used in all WORK directories, and another containing a breakdown of all CPU usage across our SAS Grid servers. Custom monitoring scripts for the observation of WORK usage, disk capacity, and SAS Metadata cluster status.

Figure 1 is a sample Foglight Dashboard.

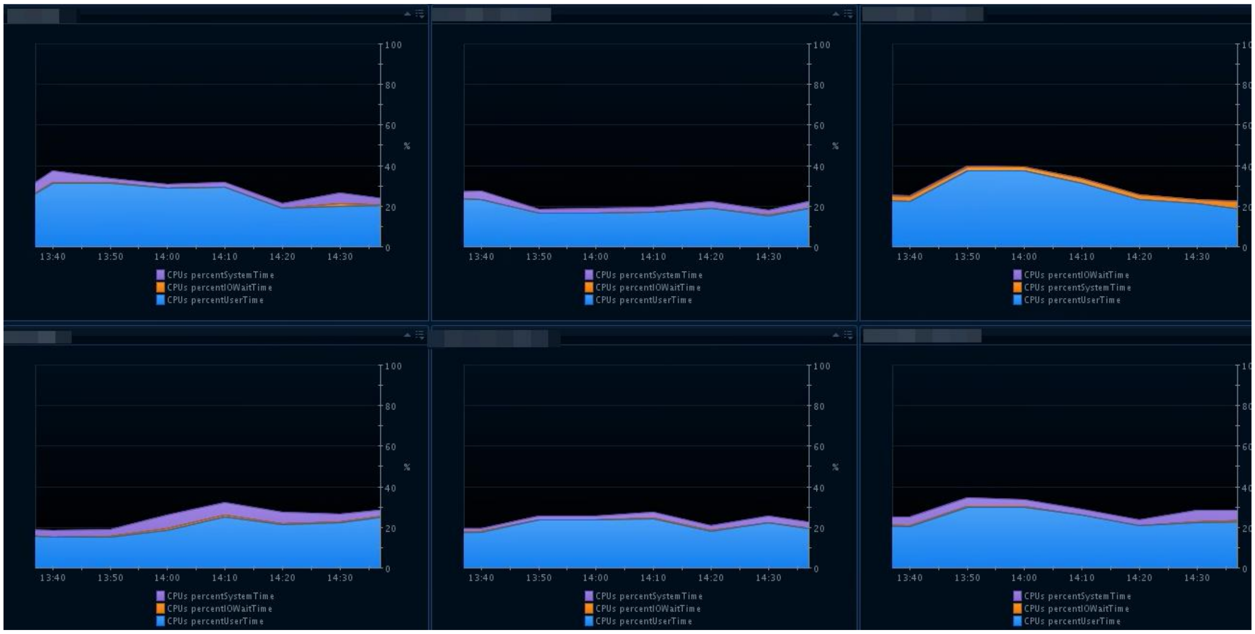


Figure 1 Sample Foglight Dashboard

PRESENTING SAS ENTERPRISE GUIDE AS A SERVICE

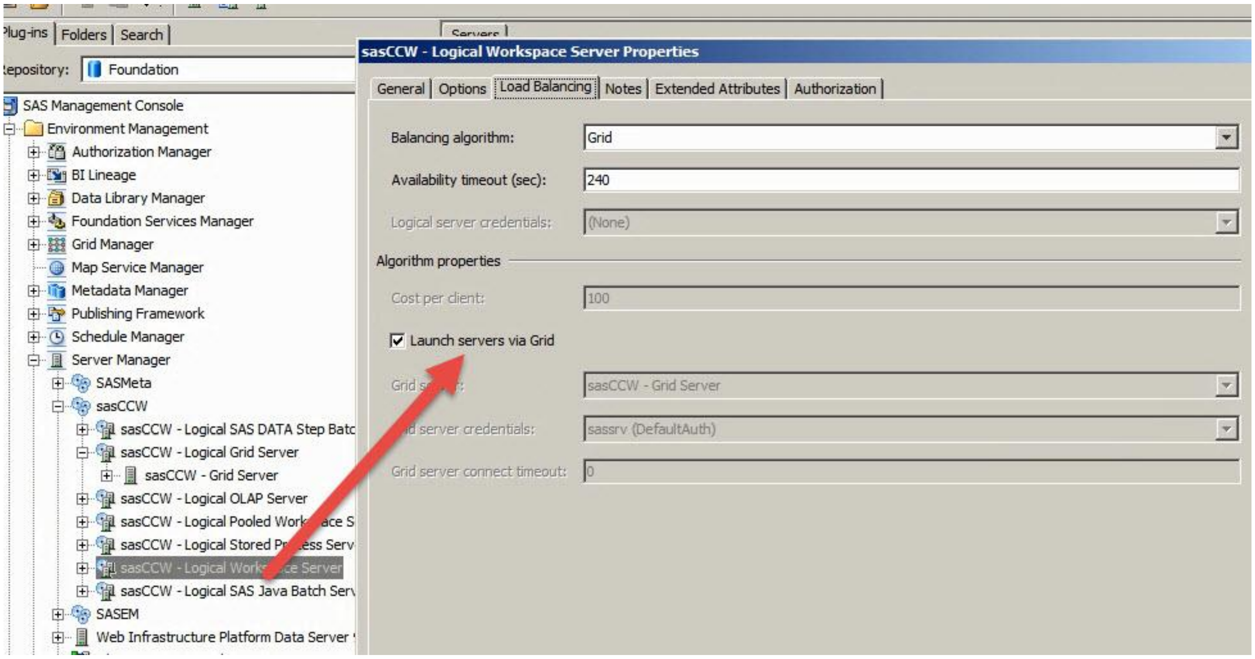
In order to grant users CCW data access as quickly and securely as possible, a combination of VMware Horizon and SAS Enterprise Guide is employed. Users are presented with a secure virtual desktop over the internet from which they access the full SAS 9.4 application suite, focused on SAS Enterprise Guide. SAS Enterprise Guide simplifies access to the SAS compute nodes for both administrators and users. All file copies to and from the virtual desktop are disabled, and all data migrating in and out of the CCW VRDC is highly restricted.

Another challenge faced within the environment is securing file level access. Many users have access to XCMD functionality in SAS, some even have SSH and SFTP access. At the same time, over 400 libraries are assigned out, along with over 140,000 SAS datasets. Since SAS Metadata doesn’t handle setting file level access, a custom access tool was developed for this purpose. The tool pulls from an access (note: not MS Access) database and assigns libraries to users. It then goes through all libraries and sets file level ACL permissions, allowing for the establishment of table level access for users, without using SAS Metadata server. SAS Metadata server is still used for authentication.

USE GRID LAUNCHED WORKSPACE SERVERS

Launching SAS Enterprise Guide via Grid is a critical function of the system. By launching and balancing workspace servers via Grid, load balancing and tracking Enterprise Guide sessions is possible. While real time sessions present issues to load balancing, SAS Grid adequately determines general load and directs sessions to servers that have availability.

Figure 2 shows the screen for configuring grid launched workspace servers.


Plug-ins | Folders | Search
Repository: Foundation
SAS Management Console
Environment Management
Authorization Manager
BI Lineage
Data Library Manager
Foundation Services Manager
Grid Manager
Map Service Manager
Metadata Manager
Publishing Framework
Schedule Manager
Server Manager
SASMeta
sasCCW
sasCCW - Logical SAS DATA Step Batc
sasCCW - Logical Grid Server
sasCCW - Grid Server
sasCCW - Logical OLAP Server
sasCCW - Logical Pooled Workspace S
sasCCW - Logical Stored Process Serv
sasCCW - Logical Workspace Server
sasCCW - Logical SAS Java Batch Serv
SASEM
Web Infrastructure Platform Data Server
Servers
General | Options | Load Balancing | Notes | Extended Attributes | Authorization

sasCCW - Logical Workspace Server Properties

Balancing algorithm: Grid

Availability timeout (sec): 240

Logical server credentials: (None)

Algorithm properties

Cost per client: 100

☑ Launch servers via Grid

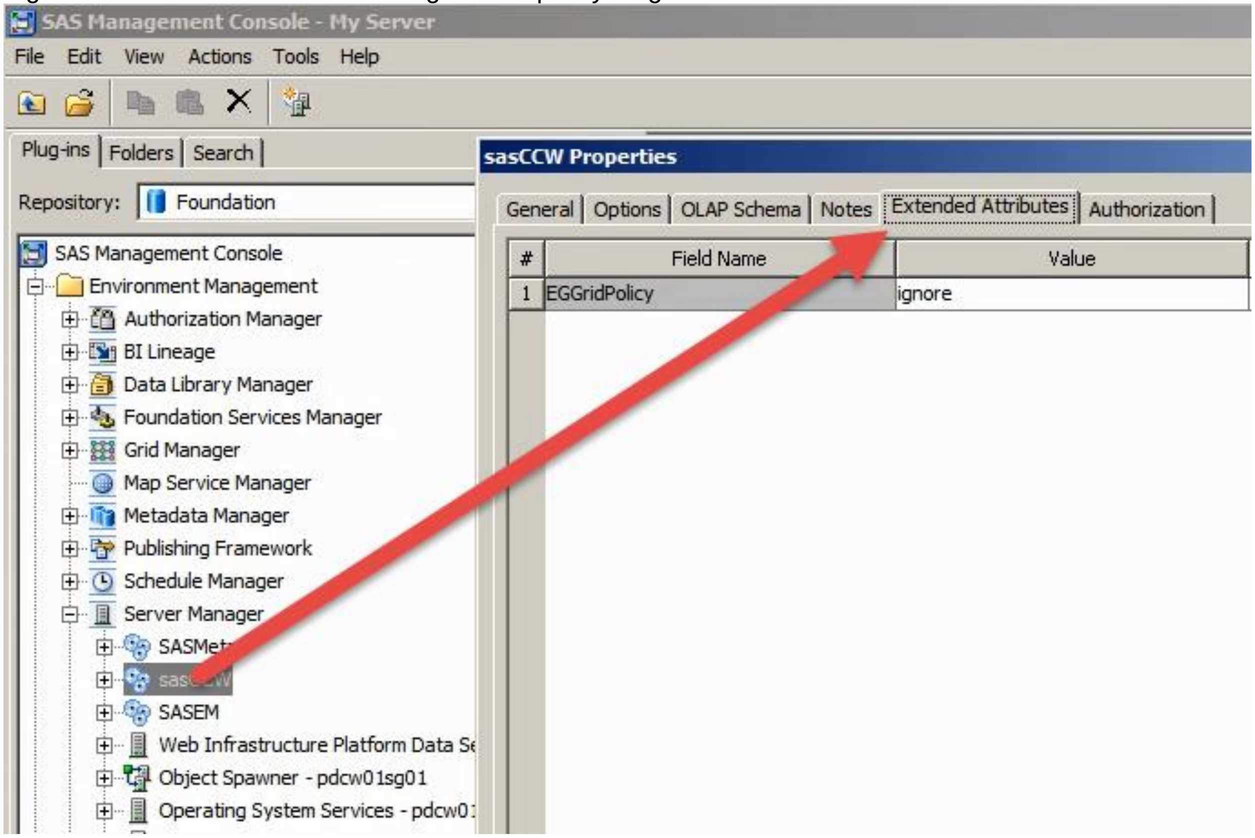
Grid server: sasCCW - Grid Server

Grid server credentials: sassrv (DefaultAuth)

Grid server connect timeout: 0

Figure 2. Enable grid launched workspace servers

To ensure that SAS Enterprise Guide Session remain in grid sessions, set ‘EGGridPolicy’ to ignore. Figure 3 shows the screen for setting EGGridpolicy to ignore.


SAS Management Console - My Server
File Edit View Actions Tools Help
Plug-ins | Folders | Search
Repository: Foundation
SAS Management Console
Environment Management
Authorization Manager
BI Lineage
Data Library Manager
Foundation Services Manager
Grid Manager
Map Service Manager
Metadata Manager
Publishing Framework
Schedule Manager
Server Manager
SASMeta
sasCCW
SASEM
Web Infrastructure Platform Data Se
Object Spawner - pdcw01sg01
Operating System Services - pdcw0
General | Options | OLAP Schema | Notes | Extended Attributes | Authorization

sasCCW Properties

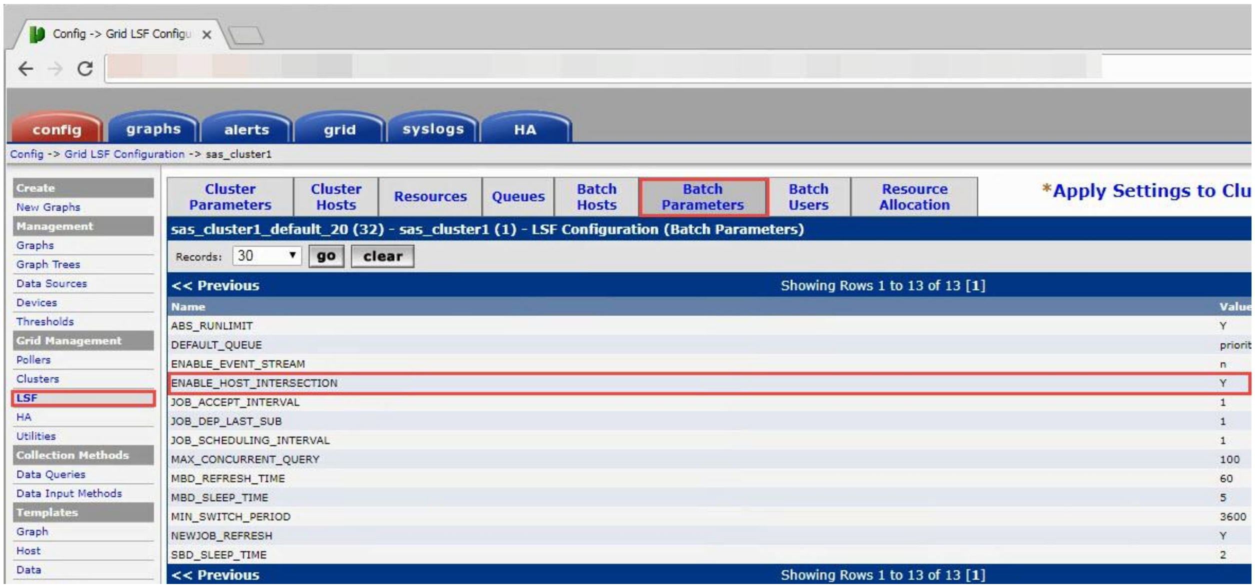
#	Field Name	Value
1	EGGridPolicy	ignore

Figure 3. Set EGGridpolicy to ignore

MORE FLEXIBLE GRID QUEUES

An issue arose when some queues did not have all grid hosts in them. For example, some servers have special software installed for encrypting files. By default, attempting to limit queues to only certain hosts causes issues with SAS Enterprise Guide. By setting ENABLE_HOST_INTERSECTION=Y, the issue is circumvented.

Figure 4 shows the screen for setting ENABLE_HOST_INTERSECTION in RTM.



sas_cluster1_default_20 (32) - sas_cluster1 (1) - LSF Configuration (Batch Parameters)

Name	Value
ABS_RUNLIMIT	Y
DEFAULT_QUEUE	priorit
ENABLE_EVENT_STREAM	n
ENABLE_HOST_INTERSECTION	Y
JOB_ACCEPT_INTERVAL	1
JOB_DEP_LAST_SUB	1
JOB_SCHEDULING_INTERVAL	1
MAX_CONCURRENT_QUERY	100
MBD_REFRESH_TIME	60
MBD_SLEEP_TIME	5
MIN_SWITCH_PERIOD	3600
NEWJOB_REFRESH	Y
SBD_SLEEP_TIME	2

Figure 4. ENABLE_HOST_INTERSECTION

TUNABLES AND SYSTEM CONFIGURATIONS

For Linux systems SAS has guidelines on tuning and system configurations available, and in general those were implemented in the CCW system. However resource issues arose, especially on the server that maintains the grid master, primary Object Spawner, and Web Infrastructure Data Server. Here are some of the most important settings to adjust:

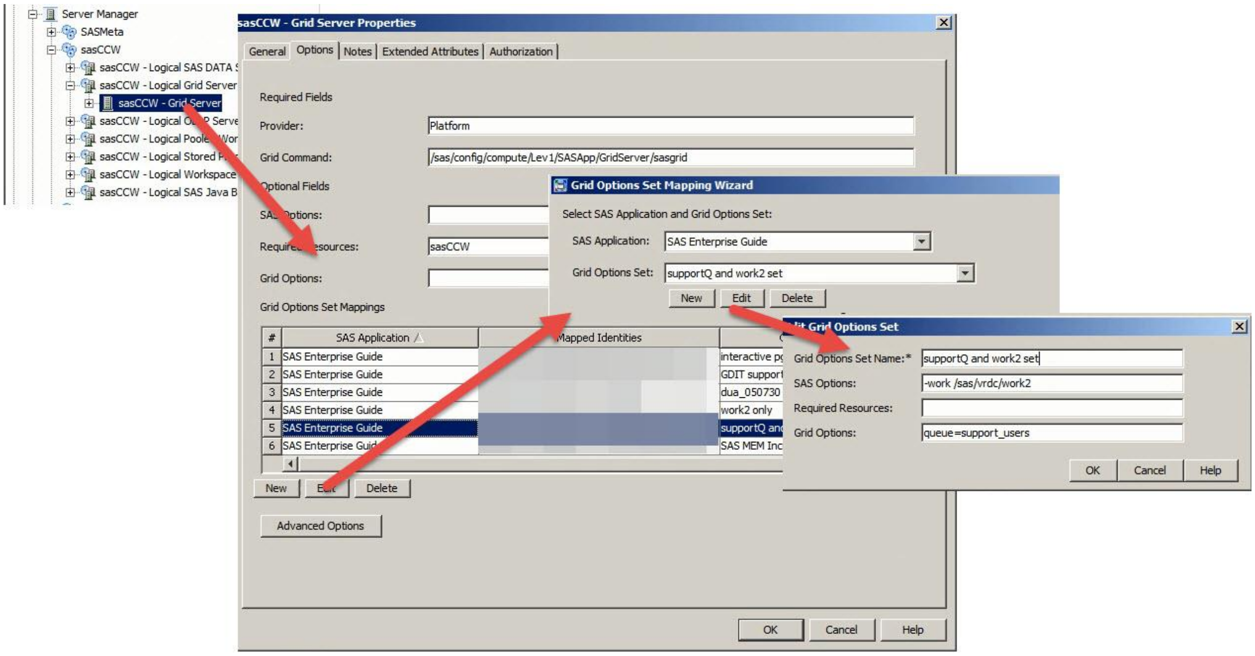
```
/etc/security/limits.conf
sasowner        hard    nofile          65535
sasowner        soft    nofile          20480


/etc/sysctl.conf
net.core.rmem_default = 16777216
net.core.rmem_max = 16777216
net.core.wmem_default = 16777216
net.core.wmem_max = 16777216
net.core.optmem_max = 16777216
```

GRID OPTION SETS

In the CCW environment, there is heavy use of highly customized autoexec scripts. These scripts set environment variables (such as WORK location and default LSF queue) and assign SAS libraries. SAS Metadata is mostly used for authentication and system configuration. However the Grid Option Sets have been a tremendous resource in the system, allowing specific groups and users “one off” customizations. Individual users and groups are given their own custom settings for LSF queues, WORK location, and even custom memory allocations.

Figure 5 shows an example grid option set.


Server Manager
SASMeta
sasCCW
sasCCW - Logical SAS DATA S
sasCCW - Logical Grid Server
sasCCW - Grid Server
sasCCW - Logical OLAP Serve
sasCCW - Logical Pooled Wor
sasCCW - Logical Stored P
sasCCW - Logical Workspace
sasCCW - Logical SAS Java B
sasCCW - Grid Server Properties
General
Options
Notes
Extended Attributes
Authorization
New
Edit
Delete
Advanced Options
OK
Cancel
Help
Grid Options Set Mapping Wizard
New
Edit
Delete
Edit Grid Options Set
OK
Cancel
Help

Required Fields

Provider: Platform

Grid Command: /sas/config/compute/Lev1/SASApp/GridServer/sasgrid

Optional Fields

SAS Options:

Required Resources: sasCCW

Grid Options:

Grid Options Set Mappings

#	SAS Application	Mapped Identities	
1	SAS Enterprise Guide	[redacted]	interactive p
2	SAS Enterprise Guide	[redacted]	GDIT support
3	SAS Enterprise Guide	[redacted]	dua_050730
4	SAS Enterprise Guide	[redacted]	work2 only
5	SAS Enterprise Guide	[redacted]	supportQ and
6	SAS Enterprise Guide	[redacted]	SAS MEM Inc

Select SAS Application and Grid Options Set:

SAS Application: SAS Enterprise Guide

Grid Options Set: supportQ and work2 set

Grid Options Set Name:* supportQ and work2 set

SAS Options: -work /sas/vrdc/work2

Required Resources:

Grid Options: queue=support_users

Figure 5. Using grid option sets

REFERENCES

Website Riva, Edoardo. “Effective Usage of SAS® Enterprise Guide® in a SAS® 9.4 Grid Manager Environment.” 2014. Available at http://support.sas.com/resources/papers/proceedings14/SAS375-2014.pdf

Website Configuration Guide for SAS 9.4 Foundation for UNIX Environments. 2017. “Economic Research.” Accessed March 1, 2018. http://support.sas.com/documentation/installcenter/en/ikfdtnunxcg/66380/PDF/default/config.pdf

CONTACT INFORMATION

Contact the author at:

Tim Acton
General Dynamic Health Solutions
tim.acton@gdit.com