



Grid Computing and SAS®

SAS®9

Table of Contents

Abstract	1
Benefits of grid computing	1
Grid computing overview.....	2
Applications suited for the grid.....	3
Grid computing with SAS®.....	3
Customer grid computing success with SAS®	4
Texas Tech University.....	5
National Institutes for Environmental Health Statistics (NIEHS) National Toxicology Program.....	6
Future directions	7
Conclusion	8
For more information	8
References.....	8

Content providers for *Grid Computing and SAS®* were Merry Rabb, SAS Worldwide Marketing strategist, and Cheryl Doninger, SAS Research and Development Director.

Content providers for *Grid Computing and SAS®* were Merry Rabb, SAS Worldwide Marketing strategist, and Cheryl Doninger, SAS Research and Development Director.

Abstract

In today's economic climate, organizations are under pressure to speed up time-to-market and reduce costs. At the same time, constraints on processing power and the limitations on existing computing infrastructure often make it difficult for IT to implement effective systems. It becomes increasingly important to find ways to make the most of the resources you already have. In many industries including financial services, manufacturing, life sciences and the public sector, significant improvements to the bottom line have been realized through grid computing. Grid computing allows you to link together the processors, storage and/or memory of distributed computers to make more efficient use of all available computer resources to solve large problems more quickly. The benefits of this approach include cost savings, improved business agility by decreasing time to deliver results, and enhanced collaboration and sharing of resources. Grid computing is an innovative way to make the most of the computing resources that you already have, as well as speed up your time to intelligence. This paper will discuss grid computing and how SAS can work in a grid.

Benefits of grid computing

There are new economic as well as business factors that are contributing to the heightened interest in the development and implementation of grid computing. Because of the Internet and the way business is conducted today, we are inundated with data. As the data flood gates open wider, the window of opportunity for capturing and turning this data into information grows shorter and shorter. Computing applications in many industries involve processing large volumes of data and/or performing repetitive computations that exceed existing server platform capabilities. In order to use data analysis to achieve business intelligence and improve decision making, data must be analyzed in a timely manner. Today's business requirements often demand a much larger sample size for analysis or perhaps use of the entire data source for maximum accuracy.

The challenges that IT shops face today, including budget cuts, server consolidation, hardware provisioning and overall administration, are all factors driving interest in and implementation of grid computing. The convergence of recent hardware and software advances has made resource virtualization possible and made it easier to construct a grid. On the hardware side, these advances include networked storage devices and low-cost, modular hardware components (such as blades); on the software side, they include improvements in networking, Web services, databases, application servers and management frameworks.

Grid computing is an innovative solution to the explosion of data and IT challenges because it provides:

- **Scalability of applications** – long-running applications can be decomposed by either execution units, data subsets, or both, and executed in dramatically less time.
- **Scalability of number of users** – multiple users can access a virtualized pool of resources in order to obtain the best possible response time overall by maximizing utilization of the computing resources.

By implementing grid computing technology, organizations can optimize their return on investment, lower cost of ownership and are able to do more with less. Grid provides three main categories of benefits.

- **Cost savings:** leveraging and exploiting unutilized or underutilized power of all computing resources within a network environment – including desktops PCs and servers.
- **Improved business agility:** decreasing time to process data and deliver quicker results to bring new products to the market. By delivering quicker results, it provides insight and agility to adjust to changes in market requirements.
- **Enhanced collaboration:** promoting collaboration, so IT resources can be shared and utilized collectively to efficiently and effectively solve compute-intensive problems.

Grid computing overview

Grid computing began in the academic research community and the national defense industry, where researchers needed to process large amounts of data as quickly as possible for data-intensive projects. It is an innovative approach that leverages existing IT infrastructure to optimize compute resources and manage data and computing workloads. Using the original concept of grid computing, arrays of computational power are constructed from a network of many small and widespread computers and used to perform large calculations and operations that can be decomposed into independent units of work. This approach allows massive computational projects to achieve results that otherwise could not be completed even on today's largest computers.

As the concept has evolved, grid computing gained rapid acceptance in the commercial marketplace in a manner similar to the emergence of the Internet. Organizations with both large and small networks have been adopting grid techniques in order to reduce execution time as well as to enable resource sharing.

There are three kinds of grids that are often discussed in the market today:

1. Compute grid – multiple computers to solve one application problem.
2. Data grid – multiple storage systems to host one very large data set.
3. Utility grid – systems from multiple organizations for collaborating on a common issue.

This paper will focus on the use and benefits of SAS in a compute grid.

SAS defines grid computing as a means to apply the resources from a collection of computers in a network and to harness all the compute power into a single project. SAS additionally believes that grid computing needs to be a secure, coordinated sharing of heterogeneous computing resources across a networked environment that allows users to get their answers faster. The bottom line is that organizations need to obtain results faster and make more efficient use of the compute power they already have.

Applications suited for the grid

It is important to clearly define the types of applications that lend themselves to a compute grid implementation so that the right kind of project can be chosen, realistic expectations can be set and performance goals can be met. Typically, applications that are good candidates for a grid implementation take many hours and possibly even days or weeks to run. In some cases, the job is so big that it cannot be completed at all even given today's processor speeds. The reason for the long run time may be due to the application requiring many replicate runs of the same fundamental task, such as identical processing on many subgroups of a large data file, or certain types of optimizations or statistical simulations. Another example of a long-running job might be one where many independent tasks must run against the same large data source, as might happen in scoring or risk analysis. In general, an application would possess one or more of the following characteristics in order for a compute grid implementation to be considered:

- Takes a long time to execute.
- Involves many replicate runs of the same fundamental task.
- Processes large amounts of data.
- Decomposes into execution units, data subsets, or both.

Many applications involve repeating the same fundamental task many times against unique subsets of the data. While the execution of a single task against a single subset of the data may execute rather quickly, if you have to do this execution many times against hundreds, thousands, or even millions of subsets of the data, it can become extremely time intensive. These types of applications are often referred to as embarrassingly parallel and are very well suited to a grid implementation because the replicate tasks can be distributed across the grid to be executed in parallel and greatly reduce the total elapsed execution time.

Each of the fundamental tasks that get distributed across the grid must have access to all required input data. Sometimes the input data may be small (on the order of megabytes) and other times the data may be large (on the order of many gigabytes). In order to achieve the highest efficiency, the compute nodes should spend the majority of the time computing rather than communicating. Compute tasks that require substantial data movement generally do not perform well in a grid. Therefore the data must either be distributed to the nodes prior to running the application or, much more commonly, made available via shared network storage. There have been many recent advances in data storage hardware that provide fast read access to data and help contribute to the success of a grid.

Grid computing with SAS®

SAS provides scalability through parallel processing and the ability to manage, access and process data in a distributed environment. It also provides interoperability between different SAS releases running on any number of heterogeneous platforms. With its ability to work with scalable procedures and I/O engines, it gives applications unmatched potential to scale up in SMP environments and scale out on the network at the same time. Because SAS is so analytically powerful, many SAS applications tend to be very data and/or compute intensive.

As a result, the performance of these SAS applications can be improved dramatically by running in a grid environment.

SAS can be used to partition large jobs into independent tasks that can be performed in parallel. This type of parallelism is called independent parallelism. By distributing these tasks across a grid and executing them in parallel, a job can be performed in a fraction of the time required for executing the job sequentially on a single machine. In addition, SAS has support for pipeline parallelism. This allows dependent steps to overlap their execution by having the output of one process piped directly into the next process as input. This not only reduces overall elapsed execution time by allowing dependent steps to execute in parallel but also reduces disk space requirements by eliminating the need for the intermediate write to disk. Piping can be used to “chain” together any number of dependent processes.

Just as there are different uses of grids, there are also different choices for implementing a grid computing solution. For a SAS application, it is possible to implement your grid solution completely within SAS. The advantages are:

- A SAS grid solution embeds the SAS parallel distribution logic directly into the SAS application rather than requiring the SAS application to be split into multiple, possibly hundreds, of individual program files. These individual files would then have to be defined to some other grid middleware and individually maintained and managed, and some SAS applications may be very difficult to divide into individual files.
- In addition to the parallelization logic, SAS can handle any data transfer between grid nodes and data management that may be necessary before, during and after execution of the parallel processes. The SAS environment is always available for any aggregation or other post processing of results.
- A SAS centric grid solution is the only solution that can leverage all of the platforms supported by SAS, including several flavors of UNIX, Windows, Linux, Alpha/VMS and z/OS.
- A SAS grid solution balances the execution load between faster and slower machines in the grid such that the faster machines get more of the work in order to minimize overall execution time.
- A SAS grid solution provides the simplicity of a single vendor providing a complete solution.

Customer grid computing success with SAS®

Many customers are running SAS applications in a grid environment and realizing dramatic reductions in execution times. These applications have been run in a variety of environments including a grid of laptops, a Linux cluster and a grid of more than 200 heterogeneous Windows and UNIX platforms. In addition, these applications have been run using SAS Version 8, SAS®9, as well as a combination of the two. Two example customers are Texas Tech University and the National Institutes for Environmental Health Statistics (NIEHS) National Toxicology Program.

Texas Tech University

Texas Tech University, located in Lubbock, Texas, is a state-supported institution consisting of seven colleges with a total student population of more than 24,000. They recently launched a high-performance computing initiative in order to improve the performance of data-intensive projects that require a great deal of time and resources. In addition to the need to process huge volumes of data in a timely manner and having limited IT capacity, they needed to enable collaboration on projects across campus and minimize expenditures.

One area of research involved developing statistical resampling methods to determine whether announcements and other historical events affect stock prices. Resampling is a compute-intensive method where the data are sampled repeatedly (say 10,000 times) with or without replacement. In addition, each resampled data set required costly matrix inversions. Adding to this computational complexity, the resampling procedure itself was studied using 10,000 simulations for a total of 100,000,000 data sets to be processed. The problem grew even larger when 10 parameters were attached to each simulation which resulted in 1 billion data sets.

Previous versions of this research relied on a "sneaker grid," where parcels of code reflecting portions of the billion data sets were given to graduate students to run overnight on their machines. The so-called "sneaker grid" is thus named because the process can be viewed as a person running from office to office "in his sneakers," handing out parcels of code. The results were then collated (essentially manually) from output files, and the "sneaker grid" process was repeated over multiple nights until the 1 billion data sets had been processed.

To grid enable this project TTU combined the distribution capabilities of SAS with heavy-duty SAS analytics to implement their financial application on a grid. They used more than 200 high-powered Windows machines in the computer labs of the Rawls College of Business Administration. Only 100 of these machines are available for use at any given time because of the number of SAS licenses purchased, with available licenses managed through a keyserver application. Thus, the computing environment can be conceptually viewed as a virtual 100 node (2.66 GHz per node) super computer with 100 gigs of combined RAM. These computers are used during the day by students to complete their daily assignments. SAS grid jobs are run while students are using them without affecting performance. However, the prime opportunity to leverage these resources for grid computing is during off peak hours and nights when students have no need for these machines.

The grid computing capabilities of SAS offer a fantastic advantage over the sneaker grid in that the jobs to process the 1 billion data sets are all sent at the same time and all data are sent back to the client machine for automatic summarizing using SAS analytics. In addition, the SAS grid enabled TTU to reduce their execution time from 25 hours on a single machine to just 40 minutes on the grid, more than a 95 percent reduction in time.

As a result of TTU's initial success with their grid, they are currently implementing their next SAS grid application, which is a portfolio selection and analysis project. The study involves randomly forming 300 portfolios, each comprised of 50 securities taken from the CRSP daily database, and then randomly choosing a one-year sequence of daily stock prices. There are more than 20,000 securities in the CRSP database; a subsetted SAS data set with essential variables requires 1.362 gigabytes. Each portfolio requires 127,500 models using PROC AUTOREG of SAS/ETS. On an 866 Megahertz PC the computations for each portfolio take approximately 40 hours and the entire analysis would require around 500 days of continuous compute time on a dedicated machine. The only feasible solution to this computation problem is to use SAS grid computing.

“Texas Tech University recently embarked on a high-performance computing initiative to use grid computing to leverage resources campus-wide,” said Peter Westfall, director of the Center for Advanced Analytics and Business Intelligence at Texas Tech. “SAS’ advanced multiprocessing capabilities are critical in driving the success of this initiative and enabling us to be innovative, such as in the creation of our advanced analytics and business intelligence center. With SAS, we are able to significantly improve the performance of particular projects that would normally require a great deal of time and resources.”

National Institutes for Environmental Health Statistics (NIEHS) National Toxicology Program

The research of the National Toxicology Program (NTP) of NIEHS has helped eliminate, reduce or control many hazards: lead, mercury, asbestos and many industrial and agricultural chemicals. NIEHS research has also begun to unravel the causes of disease at a cellular level. Part of their mission is to improve the statistical computations delivering critical results that impact our environmental health. As always, there are challenges such as silos of data, the time required to collect and analyze data and the need to make better more efficient use of skills and resources.

The specific project undertaken at NIEHS involved the data analysis of a toxicogenomics microarray study. Microarrays provide a snapshot of all of the genes in a given biological sample. This allows gene expression profiling across thousands of genes simultaneously. Gene samples were taken in order to determine the interaction of all possible gene pairs as the result of injections. Using a sample of 500 genes resulted in the need to process 124,000 possible combinations. Grid computing is ideally suited to this type of computationally-intensive problem that involves repeating the same analysis over thousands of pairs of genes.

The SAS parallelization functionality was utilized to distribute SAS analytics across 32 nodes in a Linux cluster. The total elapsed time was 14.5 hours to complete the 124,000 groups. If the entire job had run on a single node in the cluster, it would have taken nearly 448 hours to execute. This is nearly a 97 percent decrease in elapsed time. The same application was also run on a grid made up of 100 heterogeneous nodes. These nodes have a variety of processor speeds, a combination of various Windows and UNIX operating systems, and ran a combination of SAS Version 8 as well as SAS®9. The fact that they could move this same application to a heterogeneous grid shows the flexibility of SAS software. The total elapsed time in this scenario was only 5.25 hours, which would represent a 99 percent decrease in elapsed time if the entire job had run on one average node out of this grid.

The scientists at NIEHS are very pleased with the 97 percent speedup that they were able to achieve. As Roy Reter, IT security officer and systems administrator for NIEHS' Division of Intramural Research, said, "Grid computing has enabled NIEHS to see an impressive breakthrough in statistical analysis computations, which aids our scientists in quickly examining their data."

Future directions

SAS has many customers using our software across a grid, putting them in the forefront of the grid computing wave and giving them a competitive advantage over other organizations. However, we are constantly looking for innovation by adding features and capabilities that will enhance SAS' grid computing offering. One of the ways we intend to accomplish this is to enhance our existing integration with some of the industry-leading grid middleware vendors to leverage the additional capabilities that they provide and surface these capabilities through a SAS application. The solution will offer a user-friendly interface to configure the machines within the grid, provide additional monitoring and management capabilities and support the execution of SAS and non-SAS applications across a grid environment.

We are also actively working to grid enable our SAS solutions and vertical products. An article in *ComputerWorld*, Oct. 13, 2003, emphasized the amount of time it will take most solution providers to parallelize their applications. "Parallelizing these applications is a major rewrite. It will be five years before applications such as supply chain systems become suitable for grid computing."

However, in many cases SAS will be able to very quickly grid enable its solutions and vertical products that can benefit from a grid implementation. This is because many of these solutions, including those for data mining and risk management, have already incorporated the parallelization features of SAS in order to scale up and take advantage of multiple processors on SMP machines. At SAS, we are working on changes that would allow these applications to quickly and easily run on a grid and be grid automated.

In addition, SAS is also working to provide scalability for multiple users in a grid environment. This will enable an organization to create and optimize a virtual pool of resources allowing concurrent SAS users to submit jobs more efficiently and effectively. This type of grid would provide better turn around time for each individual job, sharing of resources, improved service levels and more efficient use of these resources.

In 2003 SAS became the first enterprise business intelligence vendor to join the Global Grid Forum (GGF), a community-initiated forum of more than 5,000 individual researchers and practitioners working on distributed computing, or "grid," technologies. GGF's primary objective is to promote and support the development, deployment and implementation of grid technologies and applications via the creation and documentation of best practices, including technical specifications, user experiences and implementation guidelines. SAS' participation in GGF will allow us to stay abreast of emerging standards and to influence the direction of grid computing as it becomes a critical component to the way we do business.

Conclusion

Grid computing will have a major impact on productivity and cost improvements at the enterprise level. It exploits unutilized or underutilized power of all computing resources within a network environment, including desktop PCs. It helps organizations make the most of the computing and time resources they already have. SAS defines grid computing as a means to apply the resources from a collection of computers in a network and to harness all the compute power into a single project. While the development and implementation of grid computing is still emerging, we expect the need and adoption of the grid computing market to continue to increase rapidly over the next several years.

For those applications that lend themselves to a grid implementation, SAS can be used in a grid to tap unused capacity and drastically reduce total elapsed execution time. This is proven by the customer results described above. SAS can leverage a grid environment today and we are in a great position to advance our capabilities as the grid computing industry matures.

For more information

To learn more about SAS and grid computing visit us on the Web at www.sas.com or visit the SAS Scalability Community at <http://support.sas.com/rnd/scalability/grid>.

For more information on the Global Grid Forum visit their Web site at www.gridforum.org/.

References

Bremer R., Perez J., Smith P. and Westfall, P.H. (2004). *Grid Computing at Texas Tech University Using SAS*. South Central SAS User Group Conference, 2004.

Anthes, Gary A., "Grids Extend Reach," *ComputerWorld*, October 13, 2003.



World Headquarters
and SAS Americas
SAS Campus Drive
Cary, NC 27513 USA
Tel: (1) 919 677 8000
Fax: (1) 919 677 4444
U.S. & Canada sales:
(1) 800 727 0025

SAS International
PO Box 10 53 40
Neuenheimer Landsr. 28-30
D-69043 Heidelberg, Germany
Tel: (49) 6221 4160
Fax: (49) 6221 474850
www.sas.com