IBM and SAS

Working Together to Promote Business Innovations with Grid Computing

The Power to Know.

A SAS White Paper

Table of Contents

Executive Summary

This paper provides an overview of how IBM and SAS can work together to create robust grid computing solutions that optimize the use of all available computing resources to solve data-intensive and computationally intensive applications. By combining IBM and SAS' leadership in the IT field, we can provide value-added solutions, state-of-the-art technologies, and extensive knowledge and experience to our mutual customers to solve their increasingly complex computing requirements. This paper is also a mid-level primer that will familiarize readers with grid concepts and terminology.

Grid Computing Overview

With grid computing, multiple heterogeneous systems are seamlessly integrated into a powerful single system.

The first grid(s) were called "compute grids" to reflect the fact that, by tying many machines together, tremendous computing power became available. Early grid projects, such as SETI@Home work by harnessing unused compute cycles from thousands of computers.

While this is an excellent application of grid technology, it only scratches the surface. The core concept of grid computing is to break a job into many smaller parts so that it can be processed more quickly. However, this raises a whole new set of issues. For example, how do we ensure that only authorized users can run jobs or that only authorized jobs can be run? How do we get the appropriate application binaries, licenses, and required data to the appropriate computers for execution? Grid computing attempts to answer these questions with well-defined standards that are based on an open, heterogeneous environment.

It's important to view grid computing from the top, down. In a traditional environment, a job runs on a machine. The faster the machine runs; the faster the results are received. Management is accomplished on the machine level. In a grid, jobs are automatically routed into the grid environment for execution, which enables a new level of functionality. Higher priority jobs can be given many computing nodes, and the jobs will be completed more quickly. Lower priority jobs will have fewer nodes and will take more time, but they will require fewer resources. In a grid environment, it's possible to size the run environment for a specific job to meet priority requirements and to move nodes in order to maximize the use of resources and to ensure that all jobs will be completed in the required time frame. In addition, grid computing increases the use of servers in a network by distributing the workload to servers that have a greater capacity to process the jobs in a shorter period of time. The computing resources have been virtualized, and job submission and run decisions are automated by an established set of rules.

Benefits of Grid Computing

The market for grid computing is growing rapidly. According to IDC, grid computing represents an over-12-billion-dollar opportunity in 2007 with a compound annual growth rate (CAGR) of 83%. In 2005, the grid computing market is estimated to be larger than 4 billion dollars.

Grid technology has gone from the planning and development stages to the academic field and is now being adopted by businesses worldwide. Grids are providing real value to research and development (R&D) and engineering fields, accelerating business analytics, and optimizing private enterprise and government applications. Economic trends indicate that technology budgets might not increase in tandem with enterprise goals, which makes technologies such as grid computing critical to business optimization.

The desire to analyze volumes of data in a shorter time frame is driving the attractiveness of grid computing. The grid can enable applications to process massive volumes of data efficiently in order to obtain timely business intelligence. The challenges that IT departments face today (budget cuts, server consolidation, hardware provisioning, and overall administration) are factors that are driving the interest in and implementation of grid computing. The convergence of recent hardware and software advances makes resource virtualization possible, which makes it easier to construct a grid. In hardware, recent advances include networked storage devices and low-cost, modular hardware components (such as blades). In software, recent advances include improved 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 the use of computing resources.

By implementing grid computing technology, organizations can optimize their return on investment (ROI), lower the costs of ownership, and they are able to do more with less. Grid computing provides the following benefits:

- **Cost savings** – leverages and exploits the unused or under-used power of all the computing resources in a network environment, including desktop PCs and servers.

- **Improved business agility** – decreases the time that it takes to process data and delivers results more quickly, which helps get new products to market faster. By delivering results more quickly, it also provides valuable insight into market requirements and the ability to adjust more quickly to changes.

- **Enhanced collaboration** – promotes collaboration so that IT resources can be shared and used collectively to solve computationally intensive problems efficiently and effectively.

IBM and SAS' Grid Strategy

IBM and SAS consider grid computing to be a strategic initiative. Grid is a key element of IBM's on-demand strategy. Using open standards and building on existing service paradigms, grid computing seeks to abstract computing and data resources so that IT can focus on completing jobs in the most efficient manner.

Grid computing is a fundamental facet of SAS' enterprise-class strategy and a strategic direction for SAS technologies and solutions. The SAS grid roadmap includes promoting the existing capabilities as well as enhancing the SAS grid offering throughout 2005 and beyond. We are currently grid-enabling SAS Solutions that process high volumes of data and require computationally intensive resources. SAS strongly believes that grid computing increases computing capability and maximizes the value of existing resources, thereby creating extraordinary value for our customers.

IBM and SAS are active members and participants in the Global Grid Forum (GGF). Both companies have taken a leadership role in developing the grid market by helping to create grid standards with others in the GGF and implementing the standards with the Globus Alliance. In addition, IBM and SAS are networking with thought leaders and early adopters of grid computing to establish best practices for an emerging technology.

IBM and SAS each have customers in the academic, scientific, and commercial industries who are implementing grid computing. Any industry that collects volumes of data and requires computationally intensive resources to process that data is an ideal market for grid computing. Grid computing has migrated from R&D in academic institutions to the commercial sector. Our companies have experienced an uptake of grid computing in the financial, health and life sciences, government and public sectors, and manufacturing. IBM and SAS are promoting grid computing as a means to achieve the optimization of enterprise resources.

IBM and SAS – Working Together to Promote Business Innovations with Grid Computing

IBM and SAS have the expertise and skills to deliver industry-specific solutions. Both companies have implemented grid computing solutions that have resulted in performance gains as high as 95% to 99% for a variety of customers. By working together, the best capabilities of both companies can be harnessed to create more effective solutions.

SAS applications are analytically powerful and, often, very data and computationally intensive. The performance of these SAS applications can be improved dramatically by running them in a grid environment. Recognizing the relevance of grid computing to these types of applications, SAS uses parallel processing to distribute applications across a grid, manage the parallel execution of the applications, and aggregate the results. Grid enablement can be accomplished entirely by the SAS System or, for some types of SAS programs, be accomplished in conjunction with third-party grid middleware. The grid environment can be augmented with IBM technologies for additional functionality.

IBM provides leading storage and provisioning capabilities that are critical to the success of the more complex grid implementations. As applications become more distributed and flexible, it becomes increasingly important to have seamless access to storage resources, regardless of location. Because multiple, heterogeneous computing systems are used, it's necessary to be able to seamlessly access storage in many formats such as CIFS, NFS, or FTP, or in multiple databases such as DB2 from IBM or Oracle. IBM Tivoli Provisioning products can easily and automatically build and configure grid resources such as computing nodes and storage and network resources.

Recently, IBM and SAS worked together to develop a new method for accelerating customer-insight business processes. It was shown that the current SAS credit-scoring application can be enabled to run in grid environments, and that it reduces the cycle time for executing statistical models and complex data analyses. Shorter cycle times provide additional opportunities to enhance the scoring models by enabling a greater number of modeling iterations in a shorter time frame. This can directly benefit customer acquisition and retention, cross-selling, and up-selling.

Using IBM and SAS technologies, such as the SAS credit-scoring application in an IBM grid environment, provides a cost-effective model for acquiring, deploying, and managing system resources. Because the existing insight-application infrastructure doesn't need to be replaced, enterprises can leverage their existing capital investments and optimize the efficiency of their customer-insight business applications by using grid technology. All this is achieved while migrating to a high-performance, cost-effective, standards-based infrastructure.

IBM and SAS' Grid Computing Model

The IBM and SAS grid computing model enables businesses to run existing analytics software in a more timely and efficient manner. Figure 1 shows how this computing model works.

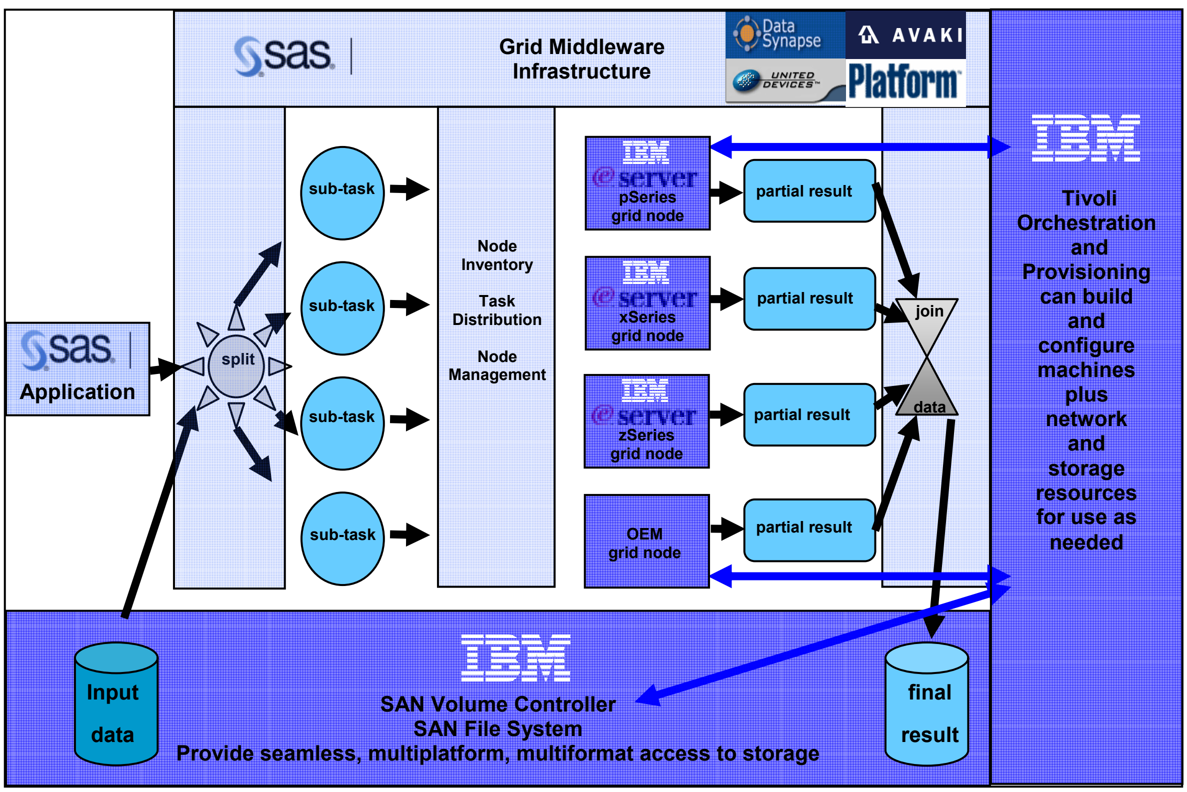

Figure 1. The IBM and SAS Grid Computing Model

Following is a description of the main components of the IBM and SAS grid computing model:

- **SAS Application** – a SAS program that is made up of one or more SAS DATA steps and SAS procedures and involves one or more SAS products. Any SAS application that is very computationally intensive, data intensive, or both and can be decomposed into sub-units of execution and data will probably benefit from a grid implementation. For example, a SAS application that requires many repeated runs of the same fundamental task (such as identical processing on many sub-groups of a large data file) or various types of optimization or statistical simulation. Another example is a SAS application that runs many independent tasks against a single, large data source such as you would find in scoring or in risk analysis.

- **Grid Middleware** – provides the ability to split the SAS application into sub-units of execution; takes care of the node inventory, task distribution, and node management; and aggregates the partial result sets into the final results. It is the job of the grid middleware to determine which machines are available and to dispatch the pieces of execution to the SAS nodes. This ensures maximum use of the nodes and minimal execution time of the application. Some examples of grid middleware are: SAS/CONNECT from SAS Institute, GridServer from DataSynapse, Platform LSF from Platform Computing, GridMP Enterprise from United Devices, and Avaki EII (enterprise information integration) from Avaki.

- **Grid Nodes** – used for executing applications. A fundamental tenant of grid computing is heterogeneity, which makes possible a wide variety of choices of hardware and operating environments. Any operating environment that is supported by SAS is available to the grid environment. Grid middleware will ensure that the appropriate applications are routed to the correct grid node.

- **Provisioning** – prepares and adds additional resources to the operating environment if there are insufficient resources available for processing. The grid controller can communicate with **TIO/TPM** (Tivoli Intelligent Orchestrator /Tivoli Provisioning Manager) and, automatically, prepare and add resources to the operating environment. This process can be as simple as changing configuration files or installing a required application, or it can be more complex such as changing the operating environment from Linux to Windows. TIO/TPM is not specific to IBM platforms, therefore, a wide variety of hardware, operating environments, and software solutions are supported.

- **Storage Pool** – consists of the SAN Volume Controller and a SAN File System, which can seamlessly present the data from a common pool in a variety of formats such as database, NFS, or CIFS/SMB. A grid unifies and virtualizes heterogeneous storage and data sources and presents them in a seamless, easily accessed format. The storage source can be from IBM or from other vendors. Each grid node has access to the storage pool.

IBM Services personnel are available and trained in the use of all these grid components. IBM has partner agreements with a wide variety of middleware vendors to facilitate the architecture and assembly of a complete solution.

SAS Consulting Services and the SAS Enterprise Excellence Center are available to provide SAS expertise and to assist with SAS grid implementations and best practices.

Conclusion

This document describes how to use grid technologies from both IBM and SAS to parallelize the execution of SAS applications across heterogeneous computer systems and to seamlessly provide access to the required data. The result is an application that runs faster. Faster execution enables more computations and better quality results. The grid middleware manages the environment and provides excellent use of all resources on the grid by more effectively distributing work.

Grid computing is an emerging technology that will have a major impact on productivity and value at the enterprise level. IT departments are faced with limited budgets and need to fully use the servers and workstations in which they have invested. Organizations are inundated with data, but still need information for on-time decision making. The workload to process the data is overwhelming and continues to increase at an incredible rate. Grid computing addresses this dilemma by making it possible to get results that are more accurate in less time. Computationally intensive projects that would take weeks or months to process can now be completed in days or hours if you use grid computing.

Grid computing is a strategic initiative for both IBM and SAS. As our collaboration evolves, there will be a corresponding evolution of our joint collateral, using our offerings to deliver complete grid computing solutions to our customers. From accelerated analytics to enterprise resource optimization, IBM and SAS will deliver more white papers and information about grid computing, and document the solutions that will power the computing infrastructures of the future.

Resources

Global Grid Forum. http://www.ggf.org.

IBM Grid Computing. http://www.ibm.com/grid.

SAS Grid Computing. http://support.sas.com/rnd/scalability/grid/index.html.
The Globus Alliance. http://www.globus.org.

Vrablik, R., Strysniewicz, R., Reech, c, and Spexet, D. *"Analytics Acceleration Grid Environment".* IBM developerWorks, November 2004. Available http://www-128.ibm.com/developerworks/grid/library/gr-aage1/.

Contact Information

Greg Kettmann
Business Development Executive – Partner Relations
IBM, Americas
(603) 472-4327
gkettman@us.ibm.com

Tamara Crawford
Program Director, Global Grid Alliances
IBM, World Wide
(312) 245-5582
tamarac@us.ibm.com

Merry Rabb
Product Manager, Grid Computing
SAS Institute Inc., Worldwide Strategy
(919) 531-7042
Merry.Rabb@sas.com

Helen Morin
Manager, IBM Alliance
SAS Institute Inc., Customer Solutions and Alliance
(602) 265-0410 ext. 230
Helen.Morin@sas.com

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

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

www.sas.com

 3326877US.505