

# WHITE PAPER.

## **SAS Parallel Scoring Optimization.**

Once data mining models are created, that's when the real work begins. As data proliferates, organizations are taking advantage of data mining techniques to develop tactical and strategic insight into these vast data stores. Algorithms, software tools and systems to mine this data can help customers in business, science, national defense, engineering and health care industries overcome their most pressing challenges. This white paper discusses the business benefits of data scoring as well as the mechanics of implementing an optimized data center to support an enterprise-class data mining operation.

- > **Consulting.**
- > **Systems Integration.**
- > **Outsourcing.**
- > **Infrastructure.**
- > **Server Technology.**

**UNISYS**

Imagine it • Done •

# CONTENTS.

<b>The Big Picture.</b>	<b>1</b>
▶ <b>Parallel scoring optimization.</b>	<b>1</b>
▶ <b>The mechanics of scoring.</b>	<b>2</b>
▶ <b>A real-life scoring case study.</b>	<b>2</b>
<b>Developing a solution.</b>	<b>3</b>
▶ <b>Test development.</b>	<b>3</b>
▶ <b>Parallel scoring benchmark description.</b>	<b>3</b>
▶ <b>Test environment configuration.</b>	<b>4</b>
▶ <b>The test results.</b>	<b>5</b>
▶ <b>Conclusions.</b>	<b>7</b>
▶ <b>About the Unisys Center of Excellence for SAS.</b>	<b>7</b>

## Parallel scoring optimization.

As data proliferates, organizations are taking advantage of data mining techniques to develop tactical and strategic insight into these vast data stores. Algorithms, software tools and systems to mine this data are critical to overcoming a wide variety of challenges in business, science, national defense, engineering and health care. For example, data mining has been used successfully in detecting credit card fraud, in criminal investigations, insurance fraud, credit card acquisition and risk management and for product recommendations.

The payoff of applying data mining results can be enormous. Financial services use data mining to optimize marketing, commercial and risk-related goals. By calculating propensity scores to use in direct marketing, one large bank has doubled its customer profitability. It regularly updates its scores so current information can be used to define selling recommendations for sales. Use of data mining results in credit scoring enables financial institutions to offer more credit broadly (e.g. Web, marketing offers, directly from branch), increasing profitability and reducing the risk of loan default.

The business impact of data mining is driving an urgent need for efficient deployment of current scoring information for optimizing day-to-day operations. The data mining process involves the following key steps:

- ▶ Formulate a well-defined business/research problem insuring you have representative data and an interdisciplinary team of business managers, data managers and data analysts to address the problem.
- ▶ Clean and prepare a data set for model development.
- ▶ Develop competing data mining models.
- ▶ Select a champion model that generalizes well on hold out data and also satisfies business requirements.
- ▶ Deploy the model into operational systems.
- ▶ Monitor the model performance.

Model deployment is the end result of data mining—the final phase in which the ROI from the mining process is realized. Deployment typically consists of three different activities. First, data is scored using the statistical or data mining model. Scoring is the critical step in applying the intelligence of the model to the underlying target data. The scoring step, although far less glamorous than model development, is data mining's sole reason for existing. It calculates the model results for subsequent use in strategy execution.

Second, these scores are deployed into operational systems and also used as the basis for various reports. Scoring is typically conducted on a periodic basis—either daily, weekly, monthly or perhaps in real time or on an event-driven basis.

Third, on a periodic basis, a new model should be built and compared to the existing model. If required, the old model can be replaced by a new model, and the deployment process starts over.

Accuracy requires fresh models on fresh data, which means updating the model as frequently as the data demands. Maximum return on model investment requires efficient, timely model deployment so results can be used to yield optimized business results.

## **The mechanics of scoring.**

The process of using the model is distinct from the process of creating the model. A model is typically used multiple times after it is created to score different data. The output of the model is the score. The most common scores are numbers that represent the probability of a particular outcome (for example, the probability of a customer changing service providers in the next 90 days). The scoring process is critical to realizing the benefit of the model output. For example, timely information about customers at risk of defecting and the ability to drive intervention strategies (direct mail retention offers, customer satisfaction service calls, etc.) can determine whether or not hundreds of millions of dollars will leave or stay in your revenue pipeline. Successfully applying data mining techniques can create a new reliance and urgency on your IT organization's ability to provide fresh, up-to-date results.

Scoring models against the operational data store can be a time-consuming activity, and it is critical to ensure that the process is properly managed and performed in a timely fashion. As models become an integral optimization strategy for many aspects of a business, the practicalities of model deployment are becoming more challenging.

## **A real-life scoring case study.**

One of Fortune's "Top 10" financial services organization has made a significant investment in SAS data mining technologies to support a variety of business processes including marketing, risk management and anti-fraud. Today, they have hundreds of models that need to be deployed against massive amounts of data. As they have grown to rely on SAS models, so has the IT challenge of deploying those models and scoring the data.

In the case of this financial services client, its IT organization faced a challenging task: Scoring 150 million rows of data against 100 different models in an overnight update window. The scoring process had previously taken weeks – even months – to accomplish because the SAS models were being recoded into another language rather than being deployed in native SAS. The business demands require regular scoring model updates. The cost of deployment was high and introduced an element of risk for each model update. Delays in their deployment process were compromising the effectiveness of their modeling efforts, reducing the optimization yields and the return on investment. Their primary goal was to complete the scoring process in a single 12-hour window.

# DEVELOPING A SOLUTION.

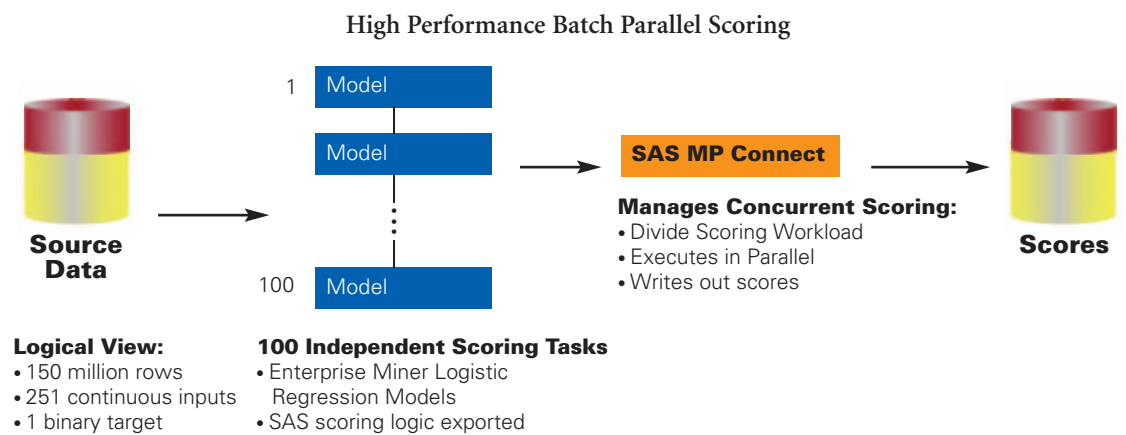
## Test development.

Feedback from several customers indicated that scoring extremely large amounts of data with a large number of models was an issue. As a result, the SAS Enterprise Miner team approached the Unisys Center of Excellence for SAS, located at SAS world headquarters, to help address these parallel scoring requirements. The Unisys Center of Excellence provides expertise in SAS performance optimization for large-scale, enterprise implementations on both Microsoft Windows and Linux operating environments. The Center is equipped with high performance computing and storage subsystems, which are critical to addressing these types of data warehousing, analytics and parallel scoring demands.

## Parallel scoring benchmark description.

A benchmark parallel scoring test application was developed using specific client requirements. One hundred SAS Enterprise Miner™ logistic regression models were developed using 252 variables. SAS Enterprise Miner™ supplies analytical models in SAS, C, Java and PMML. It not only captures the code for analytic models, but it also captures the code for preprocessing activities.

In this test, the SAS Enterprise Miner™ scoring code was deployed in parallel using SAS MP Connect (Multi-process) on a Unisys ES7000 server. MP CONNECT allows SAS jobs to be divided into multiple, independent units of work that execute in parallel, so that the jobs can be performed in less time than if they were performed sequentially.



The goal of the testing was to reach a sustained parallel updating rate of between 4,000 to 5,000 rows per second, which would equate to over 150 million rows within ten hours. Achieving this sustained rate of processing would enable the entire database to be updated against the 100 client models in the fixed overnight batch window available to the client.

### Test environment configuration.

The performance tests were run on a Unisys ES7000 Aries 420 server with 16 Intel® Itanium® 2 processors, 64GB RAM and 2 EMC CX600 disk subsystems. The tests were performed using Microsoft Windows Server 2003 Datacenter Edition for Intel® Itanium® based systems. The configuration is detailed in Figure 1 (below).

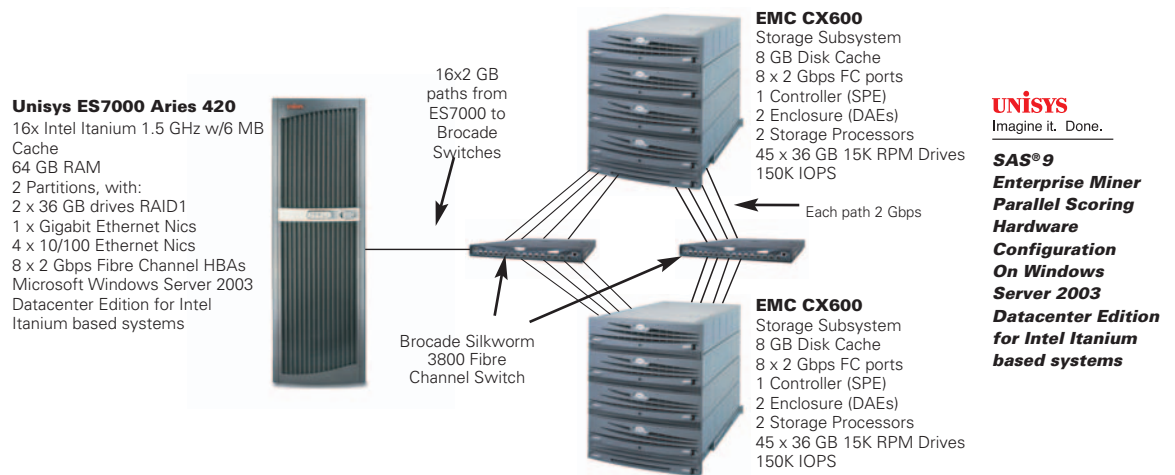


Figure 1: The parallel scoring test environment configuration for achieving 4,000 to 5,000 rows per second, which would equate to over 150 million rows within ten hours.

The EMC disk subsystems used hardware RAID within the disk subsystem. A combination of RAID 1+0 and RAID 0 volumes were used for this test. The configuration details are represented in Figure 2 (page 5).

Once the volumes were configured in the disk subsystem and presented to be mounted in the Linux partition. Software RAID was used in the operating system to stripe the smaller volumes from the disk subsystems into large volumes and then mounted as three mounted volumes in Linux.

This methodology was designed to maximize performance and distribute the workload across both disk subsystems and storage processors within the subsystems. Cascading the software across multiple storage processors and disk subsystems allowed for load balancing the IO across eight Emulex LP 9802 2Gbps Fibre channel Host Bus Adapters.

The Input volume where the source dataset resides used RAID 1+0. This type of RAID was chosen for better-read performance. The volumes where SASWORK and OUTPUT resided used RAID 0.

## Parallel Scoring Disk Layout

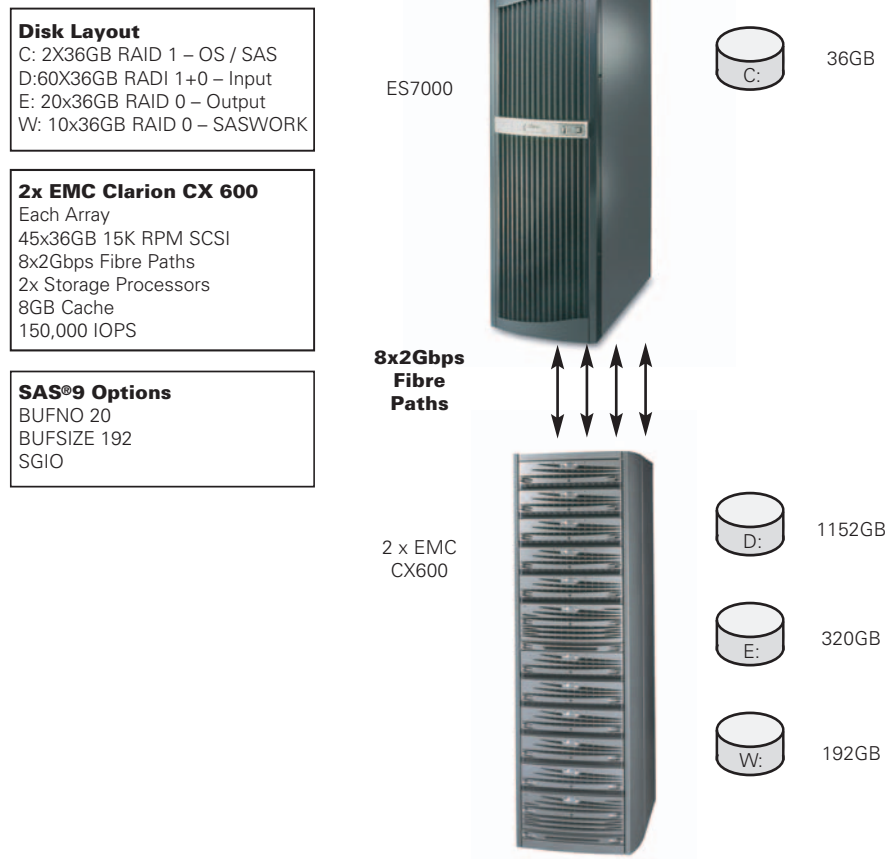


Figure 2. Parallel scoring disk layout.

### The test results.

A variety of test configurations were attempted before isolating the ideal test application tuning and configuration characteristics. This included determining the correct balance between IO usage, CPU usage and the number of models processed per CPU. When the parallel scoring run was executed, the tests used 15 out of 16 available processors. Unisys Server Sentinel Process Control was used to affinitize one CPU for the operating system and 15 CPUs were isolated for exclusive access by SAS. Each processor ran four models simultaneously. Therefore, two passes were needed to process all 100 models.

The first pass scored 60 models (15 CPUs x 4 models per CPU). The second pass scored 40 models (10 CPUs x 4 models per CPU). To process 150 million rows of data, it took 5 hours to complete the first pass and an additional 5 hours for the second pass. These results demonstrated combined capability of SAS MP Connect and a Unisys ES7000 system to meet the client's requirement for scoring a large number of models against a large volume of data within a specific window of time.

# of Rows	Times in Hours	Rows per Second
50 Million	3:24:00	4084
100 Million	6:47:44	4087
150 Million	10:04:54	4126

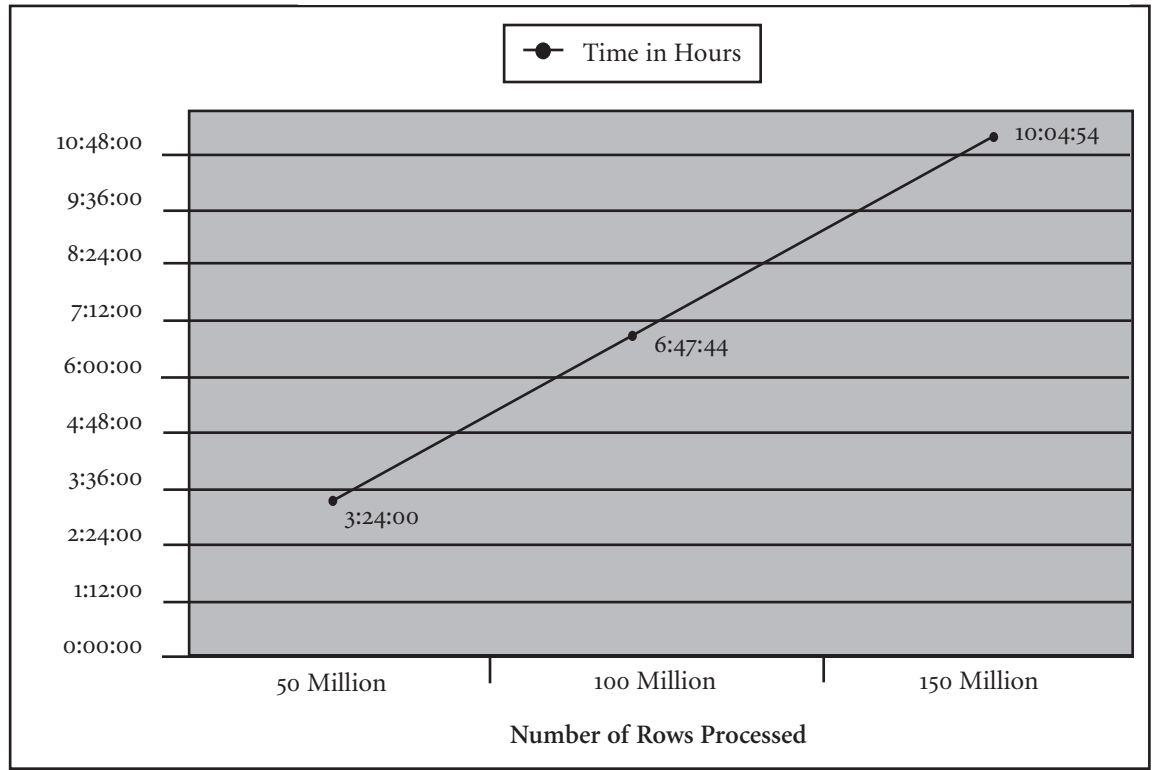


Figure 3: The linear scalability measurement of the Unisys ES7000 server as it processed the 150-million-row workload.

As illustrated in figure 3, the parallel scoring processing scaled linearly on the Unisys ES7000 server as additional rows of data were updated. The system scaled to 150 million rows. It completed the parallel scoring effort in 10 hours for the 150-million-row workload. Based on these results, we can project with a high degree of confidence that an ES7000 configured with more processors (in this case a 28-32 CPU system) would effectively score 100 models with 25 scoring processes (CPUs) with 4 models per process in one pass at the data within five hours.

These results prove that serious large-scale parallel scoring capabilities can be addressed with a Unisys ES7000 server running Microsoft Windows Server 2003 Datacenter Edition. Businesses can deploy more predictive models into production for a given time period, enabling more accurate decisions related to customer retention, fraud detection, credit applications and marketing campaigns. Because SAS model development techniques are leveraged directly in the production environment, the financial services organization profiled in this white paper can experience productivity gains, decreased risk exposure and improved workflow.

## **Conclusions.**

As data mining moves into the mainstream for tackling the growing volumes of data organizations face, the need to deploy models efficiently and the need to reduce the processing time for scoring grow in importance. The challenge of scoring will continue to grow as data volumes grow, processing windows shrink and updates are needed more frequently. IT organizations will need to respond to the business demands and meet the daily, weekly, monthly or perhaps real-time and event-driven needs. Efficient parallel scoring is key to deploying models. Demand will increase as model scores are integrated into other business processes such as credit scoring, campaign management and other business processes that require fresh and up-to-date results.

The test results documented in this paper demonstrate that Unisys ES7000 servers configured with Intel® Itanium® 2 processors and running Microsoft Windows can successfully handle the massive parallel scoring challenge of one of the world's largest financial services providers. ES7000 servers provide compelling economic alternative to proprietary mainframe and UNIX options and can help you solve your large-scale SAS data warehousing and data management challenges.

## **About the Unisys Center of Excellence for SAS.**

The SAS Enterprise Miner Research and Development team collaborated with the Unisys SAS Center of Excellence, located at SAS world headquarters in Cary, North Carolina, to address the needs of the client's specific parallel scoring requirements discussed in this white paper. The Unisys Center of Excellence (COE) for Business Intelligence at SAS is equipped with state-of-the-art ES7000 server technology and an expert staff and global support who collaborate with SAS Research and Development, Technical Architects and pre-sales support organizations to reduce client risk associated with enterprise-wide SAS deployments. The Unisys COE for Business Intelligence at SAS specializes in performance and scalability testing for SAS data warehousing, analytics and business intelligence applications and databases. The Center provides the skills and resources to help plan, deploy and optimize BI solutions using Unisys ES7000 servers and leading best-of-breed partner technologies in a no-risk environment, using your own data. Additional Unisys Business Intelligence Centers of Excellence are located in Amsterdam, Netherlands and Redmond, Washington.

# NOTES.

# NOTES.

© 2005 Unisys Corporation. All rights reserved.

Unisys is a registered trademark of Unisys Corporation. Intel is a registered trademark of Intel Corporation. Microsoft and Windows are registered trademarks of Microsoft Corporation. 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. UNIX, used under an exclusive license, is a registered trademark of The Open Group. All other brands and products referenced herein are acknowledged to be trademarks or registered trademarks of their respective holders.

Printed in US America 2/05



4126 5489-100