REVIONICS
REVENUE BIONICS

WHITE PAPER

Grid Computing in a SaaS Environment

By Keith Busloff, Jim Sills, and Jeff Moore

Table of Contents

Keith Busloff has over 22 years of systems architecture, design and development experience. As Revionics' Senior Software Engineer, Busloff is responsible for designing, engineering and implementing the Revionics Grid. Prior to joining Revionics, Keith held several systems architect, project management, data management and software engineering roles for tax, retirement and cost accounting projects with the State or California, State of Hawaii and National Semiconductor. Keith holds a B.S. Computer Science from National University at San Diego.

Jim Sills is an experienced software development executive with over 20 years of experience. He is currently vice president of Engineering at Revionics. Sills has also served as vice president of Trade Development at SAP, where he managed a staff of over 70 employees working on planning applications for retail environments. He is widely credited with pioneering scientific retailing at Khimetrics, where he served as Vice President of Research. A former faculty member at the University of Texas, San Antonio, Sills received his Doctorate and Masters in electrical engineering from Georgia Institute of Technology. He has a Bachelors in lectrical engineering from the University of Dayton. He holds six patents and has published over 20 papers.

Jeff Moore comes from an electrical engineering background with 13 years of experience. He joined Revionics early in 2008 as Director of Research. He was the lead architect of Demand Management Science at SAP where he led a core team of 10 scientists and engineers in the integration of Demand Forecasting with numerous retail applications. He was a researcher and systems engineer at pricing pioneer Khimetrics, designing solutions for markdown, promotions, and replenishment as well as developing core demand modeling and forecasting science. He holds a Masters in electrical engineering with a minor in mathematics from the Georgia Institute of Technology. He holds a Bachelors in electrical engineering from the University of Arizona.

Introduction

Software-as-a-Service (SaaS) is revolutionizing how retailers leverage their IT dollars to more efficiently access powerful planning applications. In the past, retailers purchased planning systems and maintained on-premise hardware. Besides requiring a large initial investment, these systems were expensive to maintain and upgrade (see the Revionics white paper: "Software-as-a-Service—A Better Approach"). In contrast, SaaS requires a minimal up-front investment and eliminates maintenance and upgrade costs.

> *"SaaS requires a minimal up-front investment and eliminates maintenance and upgrade costs."*

Retail planning systems, such as price and promotion optimization, are computationally intensive applications that require the latest hardware to push on-demand performance. Analysis of sales history involves applying advanced mathematical algorithms against millions (for the largest retailers, billions) of rows of historical data. Due to the sheer volume of data, use of Transactional Log (TLog) data for market-based and customer analysis makes even greater demands on the processing power of retail systems. These applications run extremely fast on the Revionics SaaS Computing Grid because computations are performed in parallel on a large pool of hardware resources.

The Revionics Computing Grid provides:

- Supercomputing power for intensive analytics for every retail customer
- Ability to scale processing capacity to any size using low-cost Intel hardware
- Sharing of resources to maximize hardware use
- No costly up-front hardware and IT expense to retailers
- High-availability processing resources (no single point of failure) in a secure SaaS environment

The whitepaper provides an overview of the Revionics SaaS Computing Grid. It includes the following sections:

- SaaS Grid Computing – Describes how SaaS grid computing surpasses on-premise applications in meeting retailers' planning and data processing needs
- The Revionics SaaS Computing Grid – Provides an overview of the Revionics Grid
- Performance Benchmarks – Presents performance benchmarks that show the advantages of parallel computing on the Revionics Grid

SaaS Grid Computing

Retail planning applications such as pricing, promotions, and markdown require weekly, daily, and on-demand processing to leverage the most up-to-date data and enable retailers to optimize profit and revenue. These applications are computationally intensive in that they involve processing large volumes of data. For example, a mid-size retailer with 30 stores, each with 30 categories, must process over 3 gigabytes of sales data. Larger retailers with thousands of stores typically occupy a terabyte or more of data storage and require huge processing capacity. The hardware costs of dedicated on-premise or hosted planning applications make even weekly processing prohibitive. As a result, retailers don't benefit from the most up-to-date information, such as cost changes, and miss opportunities to adjust prices. Worse still, on-premise applications use only a fraction of their hardware capacity because retail workflow requirements leave the hardware idle a large percent of the time.

SaaS Grid Computing shares resources among many retailers and provides much more efficient use of hardware. For example, a retailer may require 10 computing resources to complete a daily planning application task in one hour. The other 23 hours of the day, the computing resources are idle. In the SaaS Grid, there may be 100 computing resources available, which through parallization can be used to completed the task in 6 minutes instead of one hour. Rather than sitting idle, these resources are constantly being used. Bottom line is that retailers benefit because more computational resources are available when they are required.

The Revionics SaaS Computing Grid

The Revionics SaaS Computing Grid is enabled by proprietary Revionics IP that separates compute-intensive business logic into small, manageable, atomic units of work and distributes them across the network where they can be executed in parallel. The end result is that an entire business process that uses a large volume of data can be computed in a fraction of the time it would take using traditional sequential methods. For the retailer, this means a large reduction in overall time and cost.

The Revionics Grid is implemented using a Service-Oriented Architecture (SOA) to ensure:

- Computing faults are recoverable without service outage (High Availability)
- Software systems grow inward and outward without downtime as processing demand is increased or reduced (Scalability)
- Computing capacity is spread evenly across the enterprise during peak demand periods (load balancing)
- Data security. Customer data is managed at Revionics' Class A Data Center. The Center ensures there is no single point of failure in our environment. Customer data is never co-mingled, divulged to third parties, or aggregated in any fashion.

> *“ …an entire business process that uses a large volume of data can be computed in a fraction of the time it would take using traditional sequential methods. ”*

Revionics implements a specialized Grid designed for the specific needs of our retail customers and their processing requirements. The Grid provides our retailers:

- Near-perfect systems availability for our online customers and our batch processing needs
- Optimal scalability without service interruption as our market share grows and our customers' data volume increases
- Online and batch processing activities are evenly distributed and load balanced across the enterprise network and data centers
- Ability to procure off-the-shelf hardware and software systems so that cost efficiency is maximized
- A secure processing environment located inside of the Revionics firewall, ensuring that sensitive data is not compromised

The next section describes the Grid architecture.

Architecture

The Revionics Grid is implemented using a Service-Oriented Architecture (SOA). The architecture ensures high availability, scalability, and load balancing, while limiting overall cost for Revionics customers now and in the future.

The architecture includes Master Services that are responsible for managing the Grid's command and control, load balancing, and failover, as well as Work Services that are responsible for executing on demand and batch processing atomic tasks. The architecture is designed to ensure that both sets of services can be scaled outward within the Revionics infrastructure, without service interruption, so that processing bandwidth is increased at any point in time. The Grid's SOA design is built using Microsoft technologies that are proven to be highly efficient, mature, and very cost effective for large-scale information technology implementations. Figure 1 illustrates a conceptual view for the Revionics Grid. The Revionics Portal (and any other client), connects through the firewall to the Master Services, and submits job requests such as Price Optimization, Markdown, etc. The Master Services transfer job requests to the Work Services, which execute the jobs using data stored within the customer data repositories.

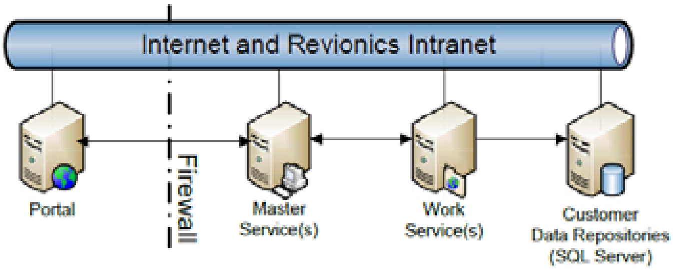
Internet and Revionics Intranet
Portal
Firewall
Master Service(s)
Work Service(s)
Customer Data Repositories (SQL Server)

Figure 1. Grid Architecture Conceptual View

Functional Components

Figure 2 illustrates the component view of the Revionics Grid. Clients connect and submit job requests to the Grid through Web Services that are hosted within Internet Information Server (IIS). The job requests are submitted using SOAP messaging over the HTTP protocol. The Web Services connect and forward job requests to the Master Service or one of the Backup Master Services using Windows Communication Foundation (WCF) over TCP. The Master Services store job requests in a shared Master Queue (SQL Server). Master Services then determine which Work Services have capacity to execute jobs using optimized load balancing algorithms. Then, using WCF over TCP, Master Services forward the job requests to the Work Services.

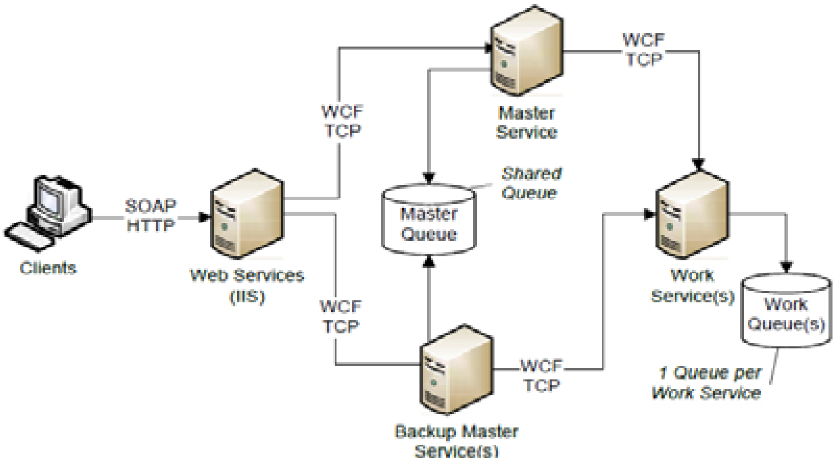

Figure 2. Revionics Grid Component View

"The Grid's SOA design is built using Microsoft technologies that are proven to be highly efficient, mature, and very cost effective for large-scale information technology implementations."

Table 1 identifies each Grid component along with its function, objective, and/or benefit.

Table 1. Grid Components

Component	Function, Objective, and/or Benefit
Master and Backup Master Services	• Provides command and control services for clients and Grid Work Services • Provides job submission services for clients and Grid Work Services • Provides job status services for clients and Grid Work Services • The Backup Master Services ensures high availability by picking up the load in the event that the primary Master Services fail • Additional Backup Master Services can be deployed at any time without service interruption to ensure that Master Services can scale outward as Revionics processing needs arise • Master and Backup Master Services are responsible for managing each Work Services processing load, ensuring optimal load balancing across the Grid enterprise
Master Queue	• Primary Master Services and Backup Master Services share the same service queue to ensure that all Service Requests are handled
Work Service(s)	• Executes atomic units of work such as Price Optimization for Zone and Categories • Each Work Service operates independently of other Work Services, ensuring that any job failure does not affect another unit of work • Each Work Service maintains its own work queue to ensure that all job requests are handled • Additional Work Services can be deployed at any time without service interruption to ensure that Revionics processing capabilities scale outward as Revionics processing needs arise
Internet Information Server (IIS)	• Host Web Services
Web Services	• Provides public Grid interface
Windows Communication Foundation (WCF)	• SOA integration and TCP messaging

Performance Benchmarks

Revionics combines the Grid capability with advanced algorithmic efficiencies to achieve unparalleled performance in retail planning. While computational tasks in other planning applications takes weeks to complete, Revionics can perform these tasks in minutes or hours. A typical planning task involves modeling demand for each item in each store. Revionics' Item-Store demand modeling benchmark test analyzes 2 years of sales history for every item in every store for select customer scenarios. These models are used for downstream forecasting and optimization processes in planning applications. Item-store demand modeling is CPU and database intensive because of the very large volume of data, complex calculations, and database read/write operations.

> *“ Revionics combines the Grid capability with advanced algorithmic efficiencies to achieve unparalleled performance in retail planning ”*

The Revionics Grid was used to perform the following processes, yielding benchmark results that demonstrate its ability to scale to very large data sets:

- Read the complete set of Items and Stores from a SQL server database and decompose them into atomic units of work
- Submit the atomic units of work to the Grid for Item-Store demand modeling
- Distribute the atomic units of work across the Grid's Work Services for execution
- Process the atomic units of work and store the results in a SQL server database

The Grid test environment configuration is listed in Table 2, and the results from 2 of the test executions are listed in Table 3 and Table 4. The results clearly show that computations that require days or weeks using on premise or hosted solutions, are completed in just minutes or hours using the Revionics Computing Grid.

Table 2. Test Environment Configuration

Service Type	Make/Model	OS	CPU	Mem
Master Service	Dell PowerEdge 1955	Win 2003 Server x64 SP2	4 Core / 1.60ghz	8GB
Work Service 1	Dell PowerEdge 1955	Win 2003 Server x64 SP2	4 Core / 1.60ghz	8GB
Work Service 2	Dell PowerEdge 1955	Win 2003 Server x64 SP2	4 Core / 3.00ghz	32GB

> *“ Results clearly show that computations that require days or weeks using on-premise or hosted solutions, are completed in just minutes or hours using the Revionics Computing Grid. ”*

Table 3. Benchmark Test 1

Description	Value / Result
Number of Products	139,931
Number of Stores	148
Number of Sales History records	96,780,732
Number of Product/Store combinations in Sales History	4,273,796
Records inserted into Table 1 (mod_parm)	21,368,980
Records inserted into Table 2 (mod_metric)	25,333,674
Average number of records inserted per minute	367,154
Number of jobs (Atomic Units of Work)	10277
Approximate elapsed processing time per job	.06 (Seconds)
Overall elapsed processing time	1 Hour, 9 Minutes

Table 4. Benchmark Test 2

Description	Value / Result
Number of Products	50,106
Number of Stores	35
Number of Sales History records	7,834,301
Number of Product/Store combinations in Sales History	340,200
Records inserted into Table 1 (mod_parm)	1,701,000
Records inserted into Table 2 (mod_metric)	2,041,200
Average number of records inserted per minute	255,150
Number of jobs (Atomic Units of Work)	281
Approximate elapsed processing time per job	.02 (Seconds)
Overall elapsed processing time	8 Minutes

The benchmark test was performed with a small hardware landscape consisting of two processing servers and a single Master Server in a controlled environment (a total of 8 dedicated processing CPUs). In a production environment, the Grid Computing environment scales out with the addition of processing nodes. The figure below shows the scalability profile of the Grid based on Benchmark #1 to illustrate how throughput is impacted by number of processing nodes inthe Grid landscape. The measurement is time per million product-stores processed using the modeling benchmark versus the number of CPUs inthe Grid landscape.

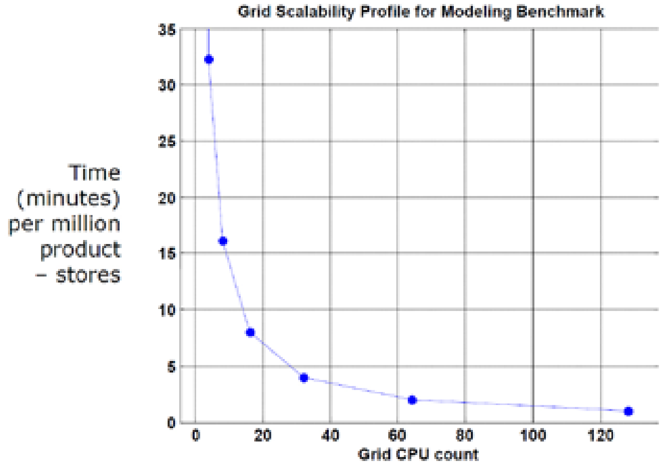

Figure 3. Revionics Grid Processing Scalability Profile

Scaling the system for size and volume of customers in the Revionics Grid environment requires only assessment of the processing load for scheduled tasks and dedication of processing nodes to match the peak processing requirements dictated by processing schedules. Addition of new nodes to the Grid is seamless and requires no downtime for the existing grid. On top of scheduled processing requirements, additional Grid capacity is budgeted to allow for on-demand processing tasks (user-initiated forecasting and optimization operations).

Conclusion

The Revionics Grid Computing Infrastructure gives users supercomputing performance in an affordable manner. Revionics' SaaS environment provides every user with cutting-edge processing capacity, and without the need for service contracts or expensive hardware. The key to Revionics Grid Computing is the highly secure network of shared hardware resources. In an era where technology costs must be justified quickly, Grid Computing delivers intelligent business software to retailers in an affordable manner.