

A White Paper from Fiserv and Microsoft:

The
***Premier* com Scaling Tests**

**Scalable E-Commerce Solutions
for Internet Bankers**



Fiserv[®]

Microsoft[®]

Notice

The information in this document represents the views of Fiserv, Inc. regarding *Premier@com* scaling at the date of publication. The contents of this document should not be interpreted as a commitment, and are for informational purposes only. Fiserv makes no warranties, express or implied, in this document.

Microsoft, Windows NT and Windows are either registered trademarks or trademarks of Microsoft Corp. in the United States and other countries.

2000 Fiserv, Inc. All rights reserved.

2000 Microsoft Corp. All rights reserved.

Table of Contents

Notice	2
Table of Contents	3
Introduction.....	4
<i>Premier[®]com</i> and <i>Connect³</i>	4
Microsoft Architecture and Products	4
Scalability	4
Objectives	5
Considerations.....	5
Tiered Architecture.....	5
Data Tier Emulation	5
Telecommunications	5
Security	5
Connect ³ Application Server	5
Connect ³ Database Server	5
Connect ³ State Server.....	5
Standard Configurations	6
Standalone Application Server	6
Redundant Application Servers.....	7
Application Server Web Farm.....	8
Testing	9
Assumptions	9
Workload 1.....	9
Methodology	10
Tools	10
Scripting.....	10
Bandwidth Throttling	10
Script Item Delay.....	10
Terminology	10
Time To Last Byte (TTLB).....	10
Response Time	10
Active Server Pages: Requests/Sec	10
Data Tier Latency.....	10
Current Sessions.....	11
Concurrent Users	11
Results	11
Results Data Tables	11
Observations.....	12
Linear Scaling of the Connect ³ Application Server	12
Maximum Transaction per Hour Rates	13
Load Balancing	13
Fail-over and Redundancy.....	13
Conclusions	14
About the Participants	14
Microsoft Corp.	14
Fiserv, Inc.	14
For More Information.....	14

Introduction

Fiserv, Inc. has developed *PremierCom*, a new, integrated e-commerce solution capable of supporting heavy transaction volumes and large numbers of users. *PremierCom* is an ideal product for financial institutions requiring scalable, high-performance Internet banking software to serve the needs of their customers.

Recent testing conducted by Fiserv, Inc. and Microsoft Corp. at Microsoft's Bellevue, Washington, laboratories has conclusively verified the highly scalable capacity of the *PremierCom* application. The results of these tests, conducted July 12-21, 2000, provide clear evidence that even a modestly configured standalone application server can support a large enrolled user base, and that when deployed within a cluster of application servers (or "Web farm"), *PremierCom* can support extremely heavy transaction volumes.

This white paper details the test objectives, design, configurations, results and conclusions established during these recent tests of *PremierCom* scalability. The results published here demonstrate that this application, based on Microsoft® architecture and built with Microsoft tools, delivers unparalleled Internet banking scalability.

PremierCom and Connect³

PremierCom is an e-commerce software product designed to provide consumers with a secure, feature-rich, cross-platform Internet banking solution, while providing the licensing financial institution with smooth implementation, flexibility in branding, and control over areas like service charging. *PremierCom* provides banking customers with real-time access to their accounts and account activity, allowing funds transfer, loan and bill payments, service requests, viewing of check images, and retrieval of statements and notifications. *Connect³* middleware links an array of Fiserv electronic delivery systems, including browser-based consumer Internet banking and corporate cash management services. *Connect³* integrates with a variety of data tiers developed by Fiserv and other vendors.

Microsoft Architecture and Products

The Microsoft Windows NT® 4.0 operating system, SQL Server™ 7.0, Microsoft Transaction Server and Internet Information Server (IIS) make up Microsoft's platform for building and deploying interoperable Web solutions, and are part of a comprehensive family of server applications for building, deploying and managing scalable, integrated, Web-based services.

The Windows® platform architecture deployed by *PremierCom* consists of the following set of system services and component-based application services that support open technology standards:

- Presentation services (Active Server Pages [ASP], HTML, Dynamic HTML, scripting)
- Application services (IIS, Microsoft Transaction Server [MTS])
- Data services (SQL Server, Active Data Objects [ADO])
- System services (directory, security, management, networking, communications)

Scalability

"Linear scaling" refers to the ability of an application to efficiently accommodate increased workloads. A plan was developed to accurately simulate and measure real-world workloads, with test results demonstrating that *PremierCom* is a scalable solution. These results also provide a valuable tool to help estimate configurations that meet the needs of financial institutions of various sizes.

Objectives

1. Simulate the workload of typical user sessions.
2. Simulate peak/extreme load conditions within various configurations.
3. Demonstrate that the application scales linearly.
4. Demonstrate that large numbers of concurrent users can be supported on standard configurations.
5. Demonstrate that a large enrolled user base can be supported on standard configurations.
6. Maintain an average transaction response time of five seconds or less.

Considerations

Tiered Architecture

Premier.com is structured using classic, three-tier application architecture consisting of the data, business and presentation tiers. (The term “business tier” describes the business rule functions within this structure.) This white paper focuses on tests and results related to scaling of the business and presentation tiers of the *Premier.com* application.

Data Tier Emulation

During scalability testing of the business tier, data tier response times were emulated to represent varying degrees of responsiveness delivered by different database systems. The parameters used to emulate the data tier were based on prior performance testing. This document refers to emulated data tier response time as “Data Tier Latency.”

Telecommunications

Physical network facilities affect achievable throughput, impacting both performance and scalability. This document does not attempt to quantify the amount of time telephone companies, Internet Service Providers (ISPs) or other routing variables add to real-world response time, herein defined as “Network Latency.”

Security

The standard configuration implements Secure Sockets Layer (SSL) on the Connect³ Application Server. The associated encryption and decryption processes occur in the IIS software, with all scalability tests conducted using 128-bit SSL.

Connect³ Application Server

The Connect³ Application Server provides business rule support for *Premier.com*, while IIS supports the HTML and ASP services required. Web Load Balancing Service (WLBS) is used to provide load balancing and fail-over capabilities across multiple Web servers.

Connect³ Database Server

The data tier resides on a Database Server or an Enterprise Server, and provides database services for Internet banking and related e-commerce transactions.

Connect³ State Server

In Web farm configurations, the state database requires a dedicated server (a State Server) that utilizes an SQL Server database. However, in non-Web farm configurations, the state database is an MSDE database hosted on the Application Server, and does not require a dedicated server.

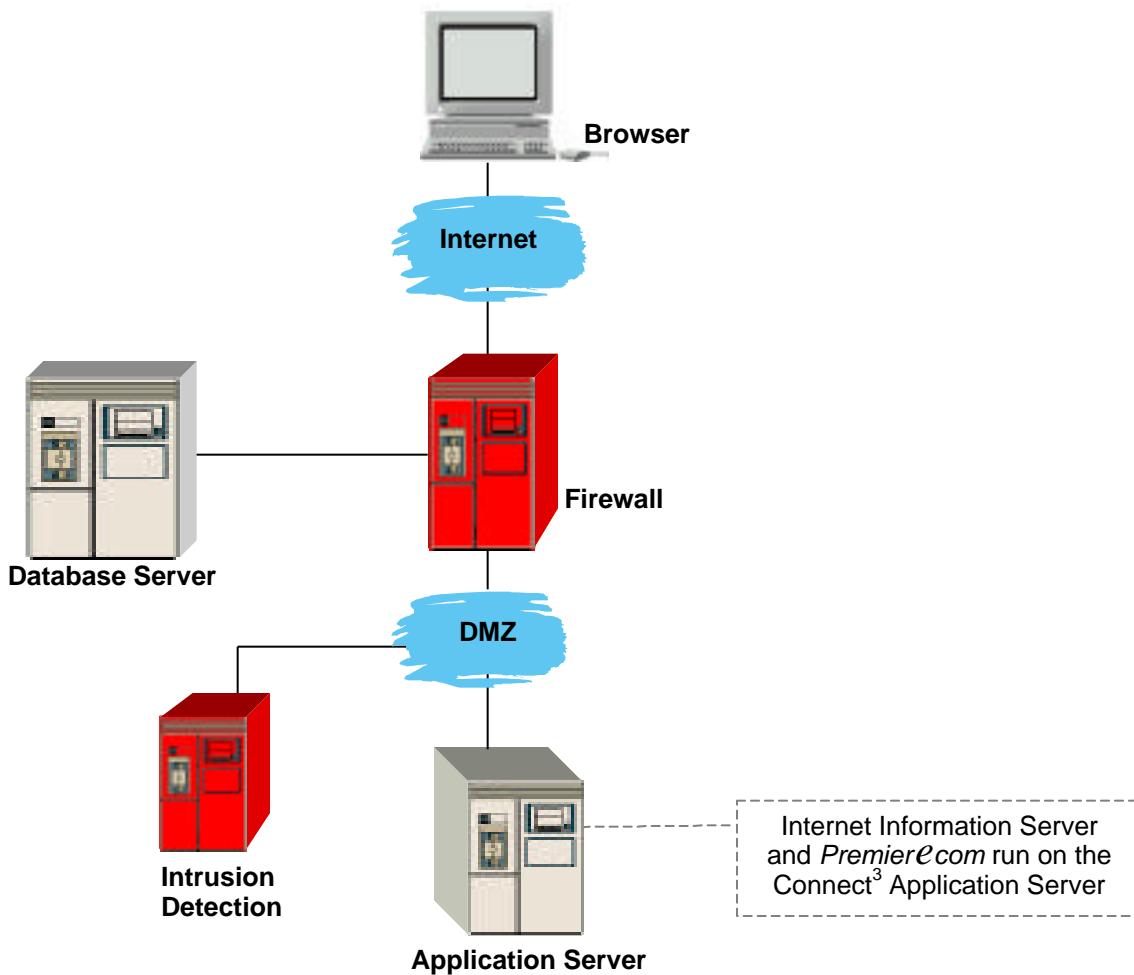
Standard Configurations

Continuous increases in processing capacity are being driven by rapid advances in computing technology. While obviously of benefit to the financial industry, this evolution necessarily limits the scope of hardware specification to the present and the immediate future. The following should therefore be seen as appropriate configurations for typical institutional needs at the time the testing was completed.

From among the configurations tested, the following basic application server configurations are most likely to be deployed: 1) a single, or *standalone application server*; 2) a pair of *redundant application servers*; and 3) from three to eight linked application servers, also known as an *application server Web farm*.

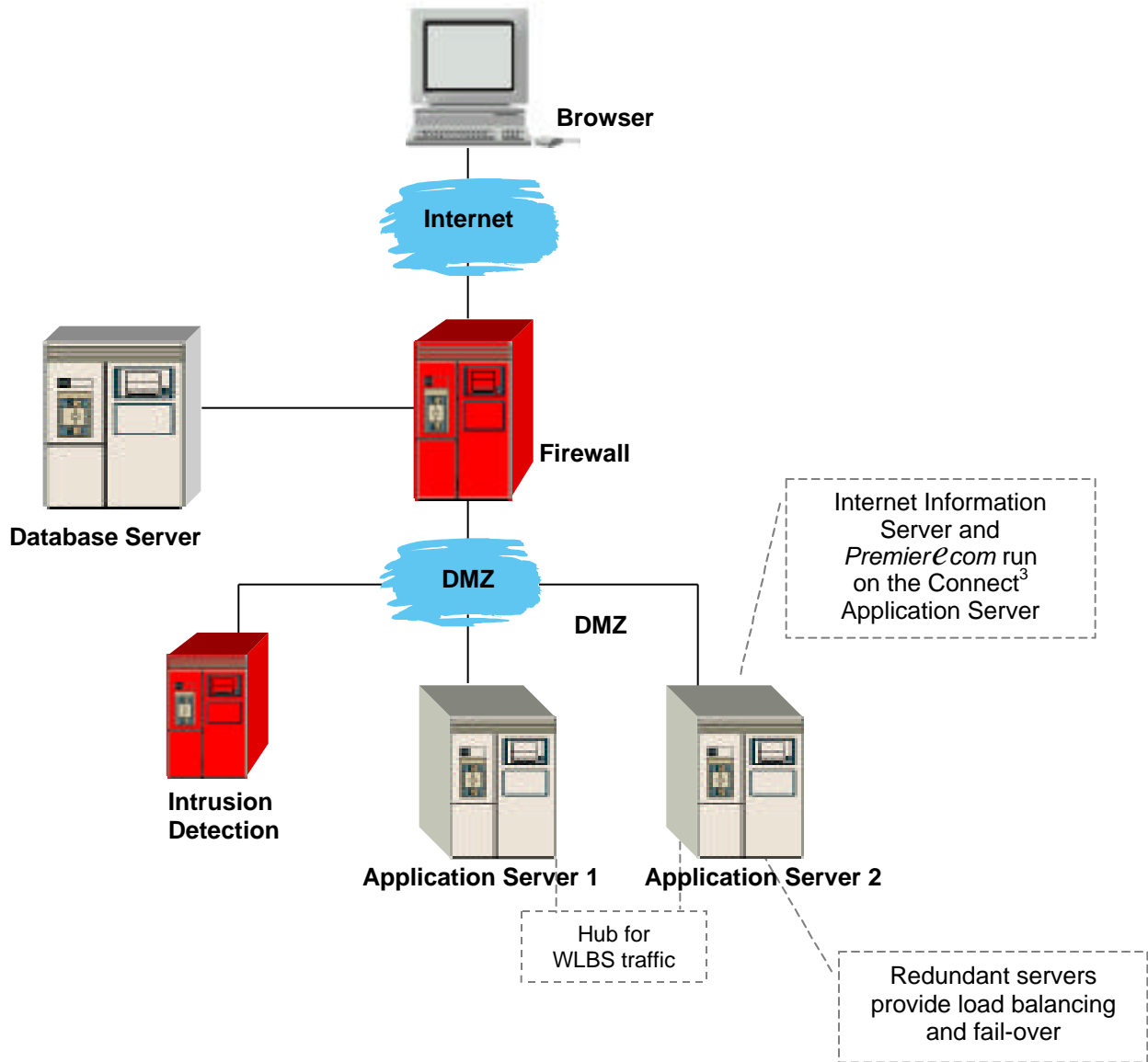
Standalone Application Server

- 1 Server
- Dual Processors (two Pentium III, 550 Mhz), 512 MB of memory
- Windows NT Server



Redundant Application Servers

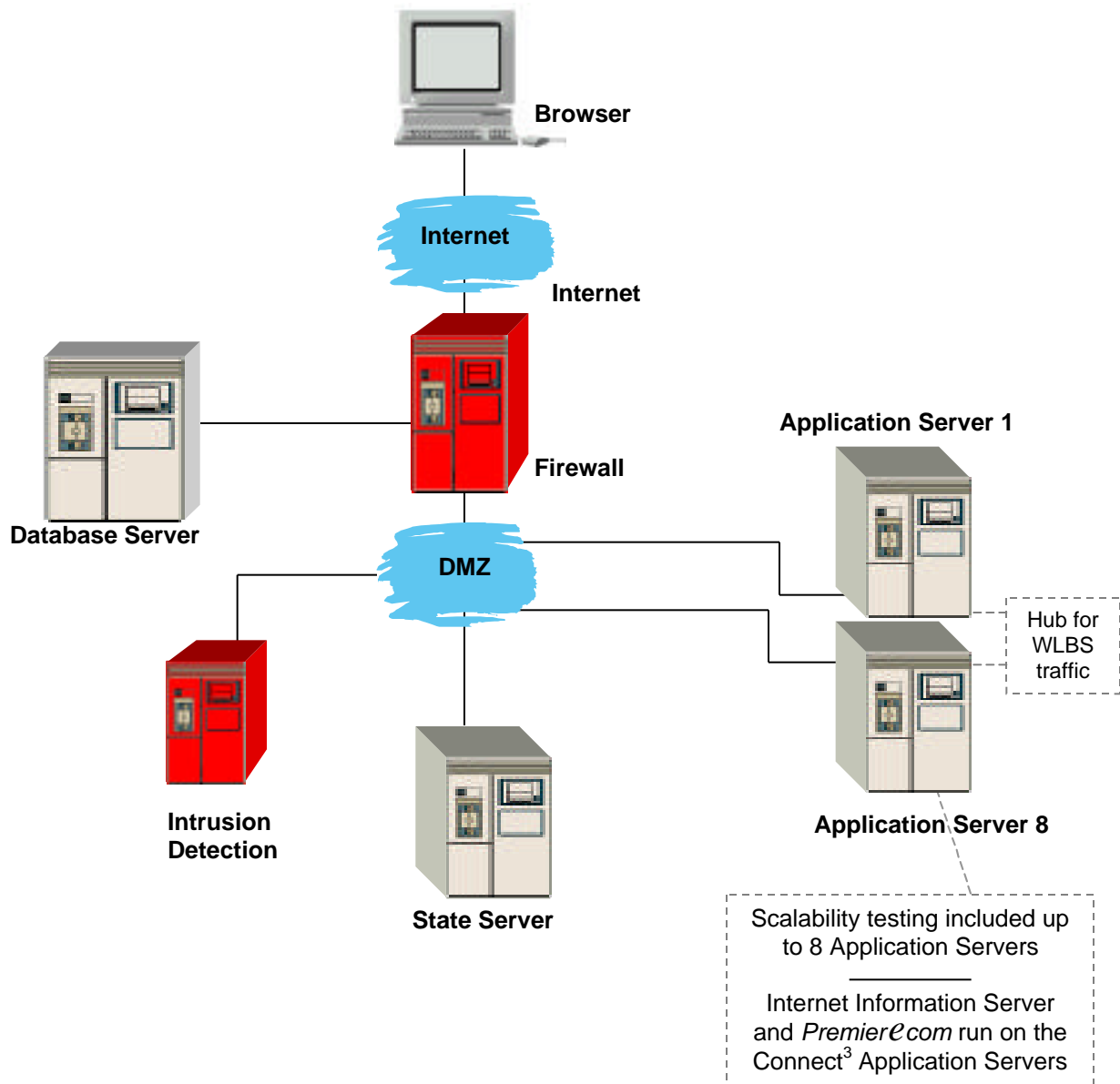
- 2 Servers
- Dual Processors (two Pentium III, 550 Mhz), 512 MB of memory
- Windows NT Enterprise Edition (with Web Load Balancing Service)



Application Server Web Farm

- 3-8 Application Servers
- Dual Processors (two Pentium III, 550 Mhz), 512 MB of memory
- Windows NT Enterprise Edition (with Web Load Balancing Service)

- Dedicated State Management Server
- Quad-Processor (four Pentium III, 550 Mhz), 1,024 MB of memory
- Windows NT Enterprise Edition (with Web Load Balancing Service)



Testing

Assumptions

The following assumptions were made during benchmark testing and in deriving conclusions. Wherever possible, assumptions are supported by field data from institutions using *Premier@com* in a production environment, results from Fiserv's internal testing, and/or accepted industry standards and publications.

1. Workload 1 represents the workload of a typical user. Workload 1 design was based on transaction history reports from actual customers. See Workload 1 table for transaction details.
2. Field data indicates that two minutes is the approximate length of a session. In Workload 1, Script Item Delay (the total time taken by the user to view, scroll and think) is estimated to total 80 seconds, with the remaining 40 seconds attributed to System Response Time.
3. Explicit log-outs by users do not always occur. Seventy-five percent of all log-out transactions were eliminated from Workload 1 to simulate those users who do not explicitly log out. (These users tie up session resources on the Web server, thereby reflecting real-world use.)
4. One out of every 750 users will be online at any given time. This is inferred from field data showing the percentage of log-ins occurring during peak hours, average session length, the number of users enrolled for Internet banking services, the percentage of the user base that logs in daily, and the percentage of traffic during peak periods. Note that during a recent study, it was determined that one out of every 769 enrolled users is online during peak periods. During these tests, a more conservative assumption was made that one out of every 750 users is online during peak periods.

Workload 1

Session Transactions (all times in seconds)	Script Item Delay	System Response Time	Total Time Elapsed
Log-in (authentication, followed by a welcome page)	5	5	10
Account List (list of accounts and current position)	5	5	10
Account Detail (includes many balances, dates and amounts)	20	5	25
Transaction List (most likely for a current statement)	30	5	35
Payment List (used to find a specific payment)	5	5	10
Payment Inquiry (payment made, followed by confirmation)	5	5	10
Payment Change (change date or amount)	5	5	10
Payment List (list of payments with updated payment)	5	5	10
Log-out (only 25% of users explicitly log out, accounted for in script)	0	0	0
Session Totals The average Total Time Elapsed per transaction is 15 seconds (a 10-second average Script Item Delay plus a 5-second average System Response Time).	80	40	120

Methodology

Measurements are made using a script that simulates a workload. The typical workload (Workload 1) is based on transaction histories of existing electronic banking users and adjusted to reflect transaction patterns that have emerged since the release of *Premier.com*.

Tools

The Microsoft Web Application Stress Test tool, currently code-named "Homer," is designed to realistically simulate multiple browsers requesting pages from a Web application. The workload is defined through the Web Application Stress Tool. This tool is used because it can emulate multiple users, reflect user pauses, apply stress factors (numbers of users), automatically capture performance monitor statistics, and manage the multiple client PCs used to generate large loads against a Website.

Scripting

Bandwidth Throttling

In order to emulate actual working conditions, "Homer" allows simulation of bandwidth. For this test, a bandwidth 56 Kbps was used for the simulated users.

Script Item Delay

Script Item Delay is user view/scroll/think time. We assume a total of 80 seconds of Script Item Delay (see Workload 1) during a typical user session.

Terminology

Time To Last Byte (TTLB)

This measurement reflects elapsed time from transaction start to the time the Web server (IIS) has transmitted the final byte of a response back to the browser. It does not reflect when the browser received it, and thus does not account for Network Latency. TTLB is reported in milliseconds.

Response Time

The metric used to derive Response Time is TTLB. In order to determine average response time, the TTLB figures reported by "Homer" for each request/response are totaled and divided by the number of Internet banking transactions.

Active Server Pages: Requests/Sec

Requests per second helps measure application performance, but results can be skewed by the performance of the Data Tier used. By simulating (controlling) Data Tier response times, the Requests/Sec metric provides a stable measurement of *Premier.com*'s performance.

Data Tier Latency

Multiple Data Tier Latencies were simulated during these benchmark tests. Several scenarios are represented in the results. These scenarios represent how quickly the various data tiers might respond.

Current Sessions

Current Sessions is the number of sessions currently being serviced by the Web server; it is made up of active sessions currently executing transactions, and inactive sessions that are not currently executing transactions but have not explicitly logged out. After 10 minutes of inactivity, these sessions are terminated by IIS.

Concurrent Users

Concurrent Users is the term used to describe the "active" user load (i.e., the total number of users currently and simultaneously conducting transactions on the system).

Results

The following Results Tables reflect the performance of optimal configurations most likely to be deployed by Fiserv at this time.

Table 1 reflects the standard configuration for most financial institutions, comprising a single standalone Web server with an integrated MSDE-based state server database. **Table 2** is a configuration that might be recommended for an institution that desires redundant servers and/or slightly higher volumes than those of the first tier. This second configuration also has an integrated MSDE-based state server database. **Table 3** represents a high-volume Web farm requiring a dedicated SQL Server-based state server.

All tests were conducted in a secure environment using SSL. Although the 128-bit encryption provided by SSL necessarily induces additional processor burden, this requirement reflects the high security levels Fiserv recommends to financial institutions.

Table 1: Standalone Application Server

(A single Web server with an integrated MSDE-based state server database)

Web Servers	CPUs per Web Server	Memory (MB) per Web Server	Con-current Users	Online User Base	Data Tier Latency	Hits per Hour	Trans per Hour	Total RPS	Response Time
1	1	128	50	37,500	1	24,750	15,342	6.88	2.351
1	1	128	50	37,500	2	22,956	13,956	6.38	3.644
1	1	128	50	37,500	4	20,082	12,456	5.58	5.335
1	2	256	50	37,500	1	26,544	16,404	7.37	1.447
1	2	256	50	37,500	2	26,394	15,966	7.33	2.186
1	2	256	50	37,500	4	22,656	13,698	6.29	4.138
1	2	512	100	75,000	1	41,598	25,746	11.56	5.024
1	2	512	100	75,000	2	41,010	25,434	11.39	5.183
1	2	512	100	75,000	4	35,970	22,584	9.99	6.994

Table 2: Redundant Application Servers

(Two Web servers with an integrated MSDE-based state server database)

Web Servers	CPUs per Web Server	Memory (MB) per Web Server	Con-current Users	Online User Base	Data Tier Latency	Hits per Hour	Trans per Hour	Total RPS	Response Time
2	2	512	200	150,000	1	79,872	49,902	22.18	5.51
2	2	512	200	150,000	2	77,166	48,552	21.44	5.96
2	2	512	200	150,000	4	71,676	44,808	19.9	6.99

Table 3: Application Server Web Farm

(Three to eight Web servers with a dedicated SQL Server-based state server)

Servers		CPUs per Server		Memory (MB) per Server		Con-current Users	Online User Base	Data Tier Latency	Hits per Hour	Trans per Hour	Total RPS	Response Time
Web	State	Web	State	Web	State							
4	1	2	4	512	1024	400	300,000	1	187,494	116,334	52.08	2.79
4	1	2	4	512	1024	400	300,000	2	180,294	110,754	50.08	3.64
4	1	2	4	512	1024	400	300,000	4	163,356	100,470	45.48	5.2
8	1	2	4	512	1024	800	600,000	1	338,280	210,030	93.92	3.488
8	1	2	4	512	1024	800	600,000	2	337,812	208,458	93.76	3.65
8	1	2	4	512	1024	800	600,000	4	309,018	190,980	85.84	5.09

Observations

Linear Scaling of the Connect³ Application Server

The Results Tables demonstrate that two processors in a single Connect³ Application Server, while not twice as fast, do significantly reduce the response time (e.g., from 2.351 seconds to 1.447 seconds) and significantly increase the number of supportable users (e.g., from 50 to 100 concurrent users) on a single server.

Fiserv configurations use two-processor Web servers as the optimal and therefore standard configuration.

The Transactions per Hour results support the objective of linear scalability through the addition of Web servers: One server processed 25,746 transactions per hour (see Table 1), two servers processed 49,902 (Table 2), four servers processed 116,334 and eight servers processed 210,030 (Table 3).

Maximum Transaction per Hour Rates

Eight servers with quad-processors processed in excess of 225,000 transactions per hour. This was the maximum processed during these tests (see the following table). However, the most practical and cost-efficient means of achieving 225,000 and higher rates is a configuration that clusters multiple state servers. Though not tested during this phase, we believe this configuration would allow a financial institution to nearly double the number of concurrent users it can serve, without increasing the number of Web servers required.

Servers		CPUs per Server		Memory (MB) per Server		Con-current Users	Online User Base	Data Tier Latency	Hits per Hour	Trans per Hour	Total RPS	Response Time
Web	State	Web	State	Web	State							
8	1	4	4	512	1024	800	600,000	1	362,912	225,832	100.8	3.45

At one point during testing, it was noted that as many as 8,107 Current Sessions were being successfully managed by the Web farm.

Load Balancing

The Standalone Application Server is designed for scalability and is capable of supporting all business tier functions, including Web services, SSL cryptography, and business rules. However, in a Web farm, the application server must take on the additional function of load-balancing. The Windows Load Balancing Service (WLBS) was the load-balancing method tested.

There are configurations that would remove the load-balancing requirement from the application server. However, we were unable to detect any significant WLBS-induced degradation on the Web server during the tests, concluding that WLBS performed very well. Note that all Web farm servers were configured with dual Network Interface Cards (NICs). One NIC was used for WLBS traffic to the rest of the cluster.

Fail-over and Redundancy

Note that all Web farms use clustering technology (WLBS) that provides for fail-over; in the event of Web server failure, subsequent traffic is directed to the other Web server(s). The state server should be configured as a fault tolerant system, because in this configuration it becomes a single point of failure for the entire Website.

Premier.com eliminates server affinity (i.e., each time a user hits the site, the transaction can be processed by a different server). This enables a session to be serviced by another Web server if the Web server servicing a user's session fails, enhances the ability of WLBS to balance load, and avoids problems common to specific ISPs when many users make requests from the same IP address.

Conclusions

Premier.com supported loads of up to 800 concurrent users with response times well under the five-second objective. This was achieved without specialized hardware or tuning.

Premier.com exceeded expectations in its ability to process more than 225,000 transactions per hour. (Note that real-world results will vary for each financial institution.)

Application server configurations can be scaled up simply by adding hardware (either more processors or more servers). To realize the benefits of this scalability, the data tier must also scale to provide reasonable Data Tier Latencies.

There is good reason to believe that the application will scale well beyond the levels currently testable with these laboratory configurations.

About the Participants

Microsoft Corp.

Founded in 1975, Microsoft (NASDAQ:MSFT) is the worldwide leader in software, services and Internet technologies for personal and business computing. The company offers a wide range of products and services designed to empower people through great software — any time, any place, and on any device.

Fiserv, Inc.

Fiserv, Inc. (NASDAQ:FISV) is an independent, full-service provider of integrated data processing and information management systems to the financial industry. As a leading technology resource, Fiserv serves more than 10,000 financial services providers worldwide, including banks, broker-dealers, credit unions, financial planners/investment advisers, insurance companies, mortgage banks and savings institutions. Headquartered in Brookfield, Wisconsin, Fiserv can also be found on the Internet at www.fiserv.com.

For More Information

For more information about these tests or products, contact Fiserv by telephone at 402.421.4251, or e-mail esolutions@fiserv.com.