February 2004

Prudential Financial IT Operations
Data Warehouse with IBM® DB2® for Linux And Topspin® InfiniBand Server Switch

Case Study

1. Executive Summary

In late-2003, IBM and Topspin worked with Prudential Financial to implement a new switched computing architecture based on IBM and Topspin products that dramatically improves the system economics for data warehouses. The solution involves using the IBM DB2 Integrated Cluster Environment database across a set of IBM eServer xSeries servers with Intel Xeon processors connected through an InfiniBand-based Topspin Server Switch with Topspin InfiniBand host side drivers.

The Prudential team, lead by Don Canning, chose this architecture for the following benefits.

1. Ease and economics_- Leverage standard OS and volume systems—including x86, Linux and InfiniBand—for lowest up-front cost.
2. "Pay-as-you-grow" – Achieve granular scalability by adding additional computing power over time rather than paying for extra capacity up-front.
3. Adapt with speed of business – Flexible switching architecture scales up and down quickly based on business requirements.

Early Prudential results show the IBM xSeries and Topspin InfiniBand cluster delivering a 20X system price/performance improvement (2.5X speedup at 1/8th the cost) relative to its traditional UNIX clusters. The following paper outlines the basis for this improvement, along with the details of the configuration and test. Prudential is now moving the Topspin cluster from R&D phase into a productive mode on a current data warehousing project.

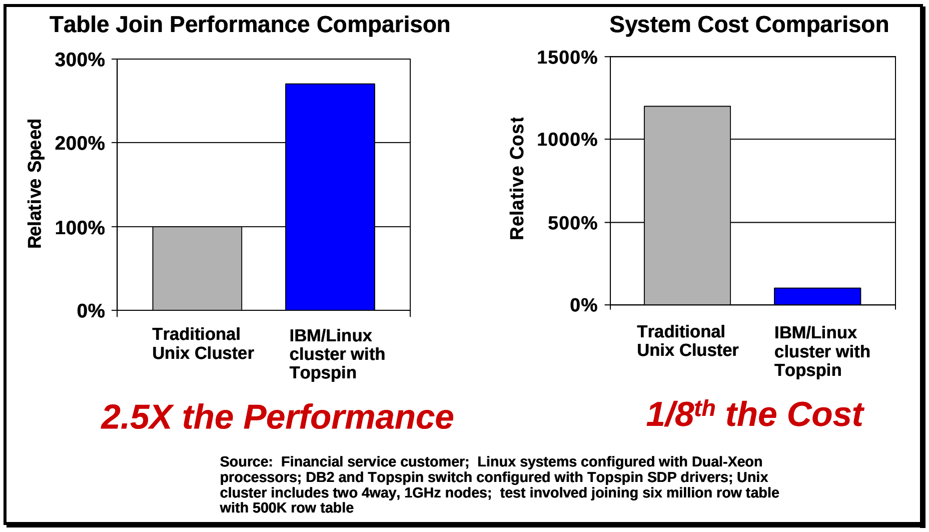


Table Join Performance Comparison

System Cost Comparison

2.5X the Performance

1/8th the Cost

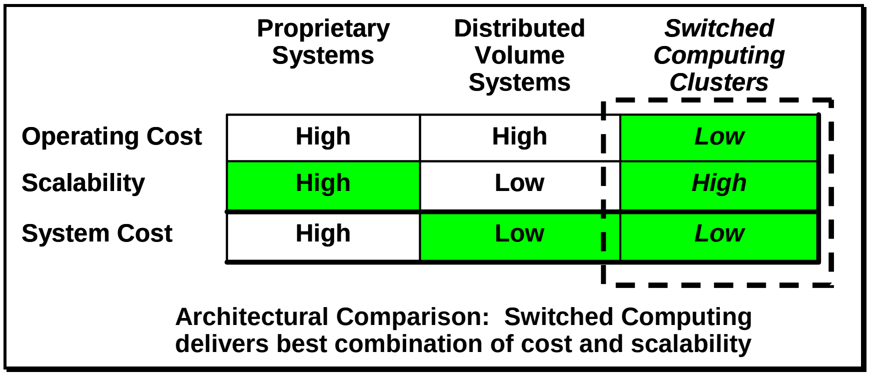
Source: Financial service customer; Linux systems configured with Dual-Xeon processors; DB2 and Topspin switch configured with Topspin SDP drivers; Unix cluster includes two 4way, 1GHz nodes; test involved joining six million row table with 500K row table

2. Benefits of Switched Computing

Faster ROI with Scale-Out Architecture

Much faster return on investment can be achieved by clustering groups of entry-level Linux servers over InfiniBand versus using smaller numbers of mid-range, or even mainframe, servers. Prudential currently runs multiple versions of IBM DB2 across several hundred enterprise-class servers clustered using server-class specific interconnects—including clusters of traditional UNIX systems. This traditional approach delivers outstanding performance, but at a high cost, making it impractical for wide adoption.

The switched computing architecture delivers the performance benefits associated with mid-range and mainframe-class clusters with the up-front equipment costs associated with standard Linux/x86 building blocks. Overall operating costs are reduced by taking advantage of the dynamic resource mapping capabilities provided in the Topspin Server Switch.



	Proprietary Systems	Distributed Volume Systems	*Switched Computing Clusters*
Operating Cost	**High**	**High**	***Low***
Scalability	**High**	**Low**	***High***
System Cost	**High**	**Low**	***Low***

Architectural Comparison: Switched Computing delivers best combination of cost and scalability

Performance Enablers

InfiniBand delivers a two-prong performance boost for data warehouse clusters by achieving better performance from each server in the cluster and by enabling larger clusters with standard components. The core enablers are CPU-offload—which frees a large fraction of CPU cycles to focus on application computing, and remote direct memory access (RDMA)—which speeds up inter-process communication between nodes. Latency between two nodes in a cluster becomes comparable to latency between two processors within an SMP server.

RDMA makes inter-process communication faster by enabling application instances to transfer data directly between each other, bypassing the traditional TCP stack and lowering the number of times each packet of data must traverse the server's internal memory bus and OS kernel. This lowers network latency from several hundred microseconds for TCP/IP to tens of microseconds for InfiniBand.

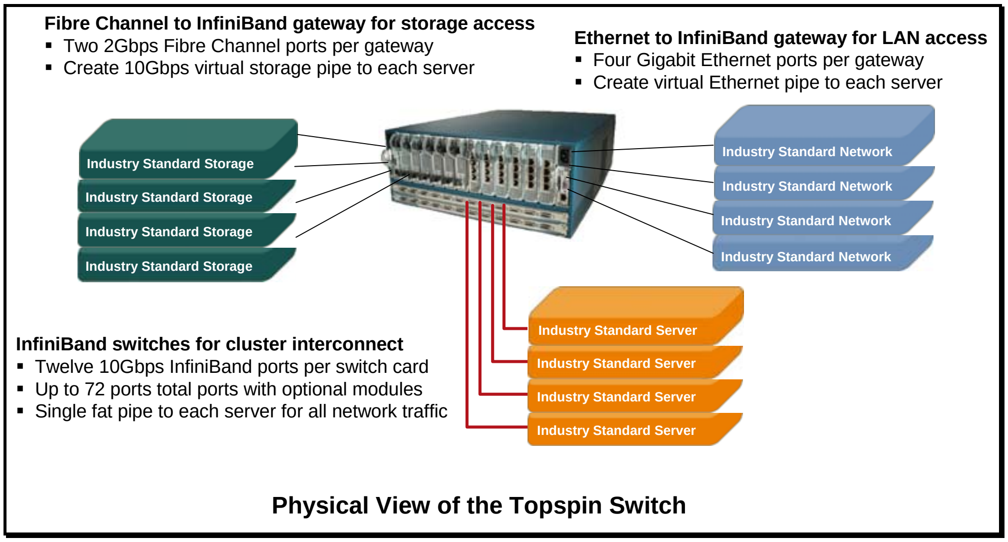
RDMA also provides CPU-offload, freeing a large fraction of the host server CPU cycles for running the database application. It does this by shifting the communications stack into the InfiniBand network hardware, lowering overall CPU utilization for communications from more than 50 percent for TCPIP to 1-3 percent for InfiniBand.[1]

DB2 UDB for Linux uses sockets direct protocol (SDP) to implement InfiniBand RDMA for communication between the cluster nodes and between the cluster and the application tier. A second protocol, SCSI remote protocol (SRP) is used by Topspin's gateways to encapsulate Fibre Channel traffic over InfiniBand for server and storage communication. Both SDP and SRP are industry standard protocols defined by the InfiniBand Trade Association (see http://www.infinibandta.org for details).

Unified Data Warehouse Networks

Building the data warehouse from a cluster of commodity servers instead of large SMP systems creates a potential challenge by introducing a large number of cables between each server and its external storage and LAN networks. Topspin solves this problem by linking each server in the cluster to a central server switch through a single 10 gigabit InfiniBand port (two connections can be used for redundancy). The switch encapsulates all inter-process communication (IPC), storage and LAN traffic over the InfiniBand connection to each server, eliminating the need for separate networks within the cluster for each of these three types of data traffic. Deployment of the Topspin Server Switch dramatically reduces the cost and complexity of building and managing the cluster.

[1] A normal TCP/IP stack needs about 75000 processing steps from the user level application through the stack. A shortcut RDMA code path needs about 250 steps.



Fibre Channel to InfiniBand gateway for storage access

- Two 2Gbps Fibre Channel ports per gateway
- Create 10Gbps virtual storage pipe to each server

Ethernet to InfiniBand gateway for LAN access

- Four Gigabit Ethernet ports per gateway
- Create virtual Ethernet pipe to each server

InfiniBand switches for cluster interconnect

- Twelve 10Gbps InfiniBand ports per switch card
- Up to 72 ports total ports with optional modules
- Single fat pipe to each server for all network traffic

Industry Standard Storage
Industry Standard Storage
Industry Standard Storage
Industry Standard Storage
Industry Standard Network
Industry Standard Network
Industry Standard Network
Industry Standard Network
Industry Standard Server
Industry Standard Server
Industry Standard Server
Industry Standard Server

Physical View of the Topspin Switch

Prudential takes advantage of this capability to consolidate its server-to-storage connections though the switch, eliminating the need for expensive Fibre Channel host bus adapters (HBAs) and switch ports dedicated to each xSeries server in the cluster.[2] This capability is provided by Topspin's Fibre Channel-to-InfiniBand gateways which plug into the server switch. In addition to basic connectivity, the gateways provide dynamic port aggregation, load balancing, and auto-rerouting for up to ten gigabits of Fibre Channel or Ethernet to each server within the server cluster.

Dynamic Resource Allocation

With a single switch interconnecting servers, storage, and LAN systems together, the Topspin system enables dynamic system resource mapping to applications based on externally defined policies and triggers. Prudential can wire its systems once into the switch and then use the switch to implement resource policies based on the needs of the business. For example, additional data warehouse nodes can be moved into the cluster in anticipation of demand spikes or if a system fails. This drives high resource-utilization and lower overall operating costs in line with Prudential's data warehousing strategy.

Customer Configuration

Prudential required integration with existing system and management components, including the IBM storage area network, as a prerequisite for the deployment of the cluster.

System Configuration

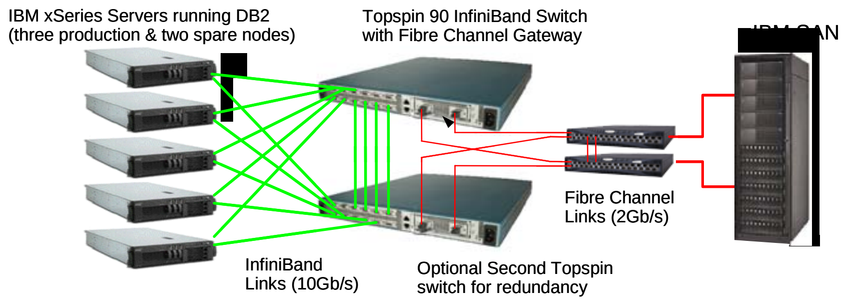
- IBM eServer xSeries x345 server with Intel Xeon processors and Topspin InfiniBand Host Channel Adapters
- IBM FastT 700 Storage Servers with IBM Exp700 disk cabinets
- IBM Fibre Channel switches
- Topspin 90 InfiniBand Switch with Fibre Channel Gateway

Major software components:

- RedHat Enterprise Linux 2.1 Advanced Server
- IBM DB2 Universal Database (UDB) Enterprise Server Edition DPF, Version 8 for Linux

[2] 10 gigabit InfiniBand switch ports are now available below $500/port, vs. roughly $1600/port for 2 gigabit Fibre Channel switch ports.

The system was designed for full redundancy on the switching infrastructure, and includes the option for adding a second switch and storage gateway for redundant paths across all components. The following diagram shows the overall system setup. The cluster includes five xSeries servers, three of which were used for DB2 UDB testing. The Topspin 90 Server Switch can interconnect up to twelve servers, allowing for expansion of the data warehouse cluster as more performance is required. Additional switches can also be added transparently, without changing any of the application or storage components.


IBM xSeries Servers running DB2
(three production & two spare nodes)
Topspin 90 InfiniBand Switch
with Fibre Channel Gateway
[redacted]
Fibre Channel
Links (2Gb/s)
InfiniBand
Links (10Gb/s)
Optional Second Topspin
switch for redundancy

Data Warehouse Setup

The Topspin Server Switch binds the five Linux nodes into a server cluster, with the IBM DB2 software operating as one single database instance, transparent to all applications utilizing the database. Regardless of which database server acts as the connection point, applications see the complete database as one single entity. This configuration mimic's Prudential's traditional data warehousing approach using small clusters of mid-range UNIX servers bound together across a server-class-specific interconnect. Prudential expects that a cluster of multiple volume-server nodes will provide equal or greater power than their traditional UNIX clusters, and at a much lower cost.

Topspin's Server Switch enables this performance by consolidating all micro nodes which push I/O over a single InfiniBand switching network, thereby mimicking the traditional UNIX interconnect fabric. This new architecture removes bottlenecks at the cluster, storage, and application tiers by providing each micro server with an extremely fast IPC interconnect within the database cluster, and up to 10 gigabits of I/O to storage and LAN networks outside the cluster.

4. Price/Performance Calculation

Prudential's data warehouse architecture team compared the performance of the InfiniBand micro cluster with their traditional UNIX cluster configurations, using three tests:

- Server to storage I/O throughput—UNIX vs. SRP over InfiniBand
- Data warehouse table join—UNIX vs. IP over InfiniBand
- Data warehouse table join—UNIX vs. SDP over InfiniBand

In the first test, millions of rows of data were extracted from the UNIX cluster hosting two large UNIX systems (4-way, 1GHz CPU, 6 Gig memory) onto IBM xSeries three node cluster, interconnected via InfiniBand. The database load operation with xSeries cluster was configured using the InfiniBand SRP stack, and it passed without any problems[3].

In the second test, a fully functional DB2 UDB on Linux environment executed queries from the operating systems command line. The test further compared the query execution time for the database on the UNIX

[3] The load time was not measured in detail. However the technician's comment was that it seemed comparable to the UNIX load time. Experience shows that the limit for actual load operations is more defined through the bandwidth in the actual backend of the storage systems than in the performance of the database servers in front of them.

cluster versus the xSeries cluster over InfiniBand TCP-IP. The test involved joining a six-million row transaction table with a five-hundred-thousand-row product table. The same database configuration, design, and data content were used on both clusters. In this case, the results were close to equal, measured by elapsed query-execution time. This was significant, since the xSeries-class architecture delivered similar results at a fraction of the cost.

In the third test, the xSeries InfiniBand configuration was changed from the classic TCP-IP communications stack to SDP, the RDMA communications stack designed largely around database clustering. In this case the xSeries query results were approximately two-and-one-half times faster than the UNIX cluster, completing the table-join in sixteen seconds, versus forty-three seconds.

Although specific cost and UNIX platform data are not available at this time, the IBM xSeries cluster with InfiniBand is as a little as one-eighth the cost of a comparable UNIX cluster[4]. This cost savings translates into as much as 20 times better price/performance for the InfiniBand cluster running this data warehouse operation in this environment.

The same database configuration, schema, and design—as well as data content—were used on each cluster, including the multi-million line table-join. This is a realistic starting point for Prudential, whose actuarial data warehouses typically scale into the hundreds of millions of rows and federated databases into the multi-terabytes.

These results were not performed with scientific rigor, but are documented here as observations of behavior. Prudential has not yet tested the environment to determine how well it scales with additional xSeries nodes, however, the observations are initially favorable. The base cluster is currently being transitioned from an R&D phase into production on a current data warehousing project.

Topspin, IBM and Horizon Data Systems, who are familiar with parallel MPP systems, worked closely with Prudential to ensure that the InfiniBand switched computing cluster works well with key applications, servers, and storage systems.

[4] Thee-node IBM xSeries server cluster with Topspin server adapters, Topspin InfiniBand server switch and InfiniBand to Fibre Channel gateway costs range between $40K to $70K for well configured system.

IBM Canada
8200 Warden Avenue
Markham, ON
L6G 1C7
Canada

Printed in United States of America
10-03

References in this publication to IBM products or services do not imply that IBM intends to make them available in all countries in which IBM operates.
The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law:
INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.
Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurement may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

The information in this white paper is provided AS IS without warranty. Such information was obtained from publicly available sources, is current as of 01/30/2003, and is subject to change. Any performance data included in the paper was obtained in the specific operating environment and is provided as an illustration. Performance in other operating environments may vary. More specific information about the capabilities of products described should be obtained from the suppliers of those products.