Case Study

Dual-Core Intel® Xeon® Processor

Financial Services

High-Performance Computing

Scaling the Grid

Leading bank[1] proves application scalability on grid with Intel® Xeon® processor-based IBM BladeCenter* servers and GigaSpaces Technologies software

When it comes to financial service investments and their high-performance computing (HPC) platforms, speed is king. Analyze larger volumes of data faster and with greater subtlety, you can make smarter, faster trade and credit decisions, reduce investment and operational risk, and improve profitability.

A top-tier, US-based bank had adopted grid computing to scale its HPC resources. Now, it wanted to optimize application performance on the grid and use its grid resources for new classes of computing challenges, including transaction processing, databases and applications where memory latency and data persistence are crucial. It turned to Intel, IBM and GigaSpaces Technologies to prove the scalability of an innovative solution that would take its grid to the next level.

"We've gone from a few million to more than a billion calculations processed in a 24-hour period... and are saving tens of millions of dollars."

Business Unit Vice President
Leading Bank[1]

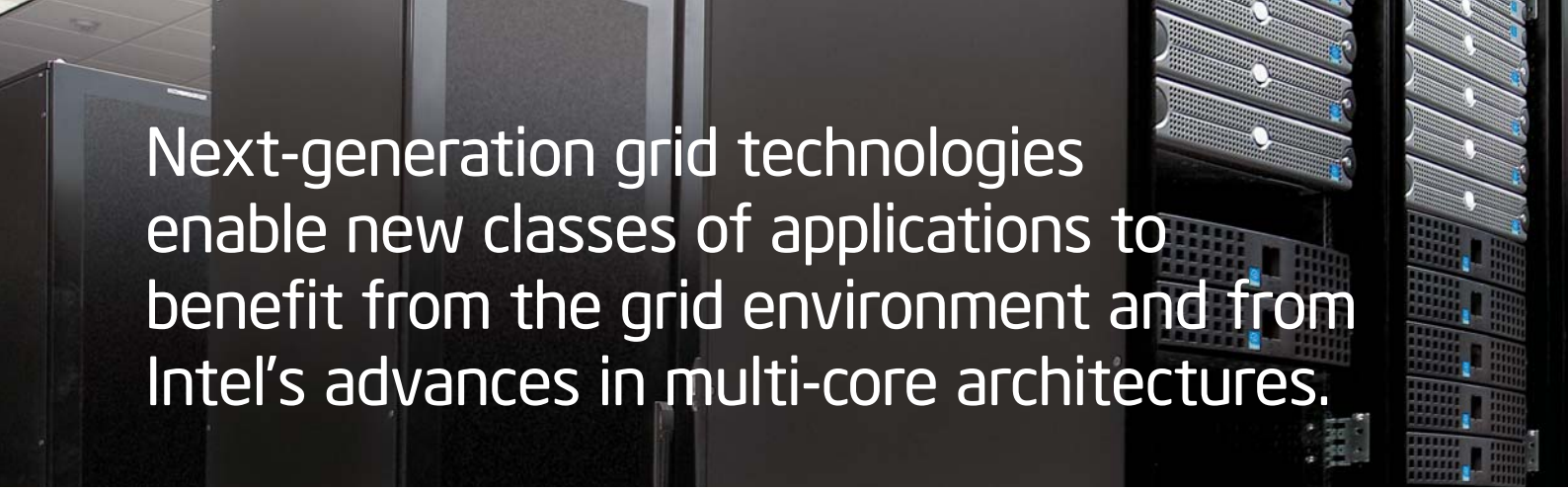
The Challenge

- **Accelerate business, broaden analysis.** To continue its growth, the bank needed to expand its already-extensive high-computing capabilities, in a quest to do more real-time business analysis and quoting and support a broader range of assets, products and locations.
- **Don't break the bank.** While the bank's demand for additional computing resources was virtually unlimited, its budget wasn't. The institution needed a cost-effective way to improve HPC speed and capacity.

The Solution

- **Virtualize and conquer.** The bank virtualized computing resources at several of its data centers around the world, creating a 10,000-CPU grid that included thousands of Intel® Xeon® based IBM servers.
- **Optimize scalability and performance on the grid.** To enhance application performance and data liquidity on its grid, the bank undertook a proof-of-concept (PoC) in the IBM Deep Computing Capacity on Demand (DCCoD) center in New York of compute and data grid scalability with GigaSpaces Technologies software running on Dual-Core Intel Xeon processor-based IBM xSeries* servers. With clear evidence that the solution could handle even terabyte-sized reads and writes to a distributed cache, the bank is moving confidently toward production deployment — and is well-positioned to take full advantage of Intel's advances in dual-core computing and hardware-assisted virtualization.

[1] Customer wishes to remain anonymous.



Next-generation grid technologies enable new classes of applications to benefit from the grid environment and from Intel's advances in multi-core architectures.

Assessing the Situation

The benefits of grid computing and other service oriented approaches are widely recognized, particularly in industries such as financial services, energy and manufacturing, where increased computing capacity can have a direct bearing on the bottom line.

A leading, full-service financial services institution[1] was an early adopter of grid solutions. Beginning with a pilot operation in one of its major US facilities, the bank created a risk management solution using Intel® Xeon® processor-based IBM servers and a grid management software.

Soon, the pilot grid was running 23 hours a day and processing hundreds of millions of calculations involving terabytes of data. Analysts got results faster and were able to perform many more "what-if" scenarios. The bank then extended its grid to include more than 10,000 processors at three global locations, including over 3,700 IBM servers or blades powered by the Intel Xeon processor.

Efficient Use of Resources

"Resources that we were paying for but using only 8 to 12 hours a day can be used around the clock," says the vice president who oversaw the launch of the initial grid deployment. "We've gone from a few million to more than a billion calculations processed in a 24-hour period. Jobs that used to take 90 minutes can be completed in 20. Computations for a single trade have been reduced from 4 hours to 40 minutes. Scenarios that would have taken 24 hours to run on our old infrastructure now run in five. It used to be that an analyst running a scenario would have his data the next day. Now, a user who asks for data at 9 a.m. can have it 20 minutes later, and we're running 600 scenarios in a day."

More efficient use of computing resources not only helps the bank process calculations faster, but also enables it to defer investments in additional IT infrastructure. Along with making better use of existing production resources, the bank can leverage costly contingency resources that normally sit idle, using them to increase capacity in its production environments.

"Thanks to the increased utilization, grid computing will save the bank tens of millions of dollars in IT costs over the next three years with better application convergence," says the vice president. "I am a big fan of grid technology. It's a wonderful solution to make your organization more efficient – increasing agility and flexibility while ensuring full utilization of IT resources throughout the enterprise."

Taking Grid to the Next Level

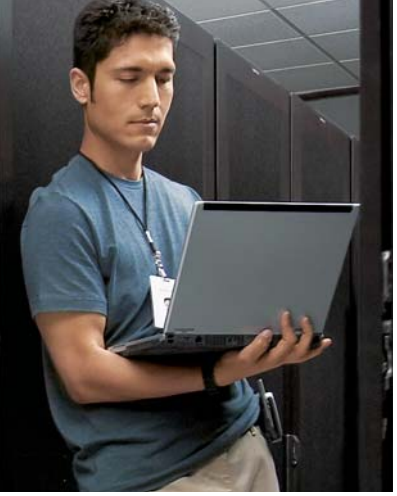
The bank wanted to optimize application scalability on the grid and run a wider range of applications on its distributed infrastructure, including real-time transactional applications. Data latency and data liquidity were particular issues. Driving down data latency – the time an application or user takes to access cached data – is important for delivering the rapid quoting and execution that today's competitive climate demands. Improving data liquidity – the ability to access data anywhere and at any time, regardless of where the data resides or how it is structured – is critical to achieving the bank's goal of expanding its in position in electronic trading, as well as delivering real-time insights into risk and P&L capabilities across businesses.

The bank turned to Intel, IBM and GigaSpaces to help address these requirements and take its grid to the next level.

Delivering the Solution

With Intel Xeon processor-based IBM servers throughout its infrastructure, the bank was well acquainted with the performance, scalability and reliability of Intel and IBM technologies, and the companies frequently joined forces to apply new and emerging technologies to the bank's critical architectural and business issues. GigaSpaces offers award-winning infrastructure software solutions built for the most demanding distributed systems and service-oriented architectures. The bank wanted to understand whether GigaSpaces software could improve application scalability on the grid, even where data latency and liquidity are important considerations.

Typical grid solutions provide scalability at the data center level – for example, assigning an application to run from 6 pm to 6 am on a given set of computing resources. GigaSpaces



"The methodology was to create data grids with distributed caches capable of loading extremely large amounts of data in memory, and then have a lot of applications and users consuming that data."

Vice President of Architecture
Leading Bank[1]

software goes further, to optimize the application's performance, capacity and throughput on those assigned resources.

Using a three-dimensional, space-based architecture, GigaSpaces Technologies creates a large, in-memory virtual cache or data fabric – an in-memory data grid layer that takes advantage of the large memory and cache capacity of the latest-generation Dual-Core Intel Xeon processor. By permitting more efficient use of memory, GigaSpaces software enables high-speed data access for distributed applications and allows even data-sensitive applications to be distributed effectively across a grid. And, by providing end-to-end, on-demand scalability at the application level, GigaSpaces software addresses bottlenecks at all levels of application performance: data access, data processing and messaging.

Proof-of-Concept Testing

The four companies conducted a large-scale proof-of-concept using GigaSpaces software on a grid of Intel Xeon processor-based IBM server blades to validate both data fabric and parallel computing scalability. Benchmark tests were conducted at the IBM DCCoD Center, one of four world class centers which provide customers access to Supercomputers on a pay-for-use basis.

Intel and IBM created a grid at the test facility containing 512 two-way IBM xSeries* 355 servers powered by the dual Intel® Pentium® 4 processors with 3 GB RAM on a 1 GB Ethernet network. The processor combines Intel's dual–core technology and innovative power management capabilities, helping to boost energy–efficient price/performance with up to two to four times the performance–per–watt of previous Intel Xeon processors and platforms.

Performance was compared using a variety of Java* development kits, which ran on the Java run-time environment in a Java virtual machine. "The methodology was to create data grids with distributed caches capable of loading extremely large amounts of data in memory, and then have a lot of applications and users consuming that data," explains a vice president of architecture at the bank, who headed the PoC effort. The benchmark tests proved that the Dual-Core Intel Xeon processor-based IBM xSeries server is a high-performance platform for running the bank's financial applications in a large-scale GigaSpaces grid computing environment. The tests also validated the scalability enabled by GigaSpaces Technology and its underlying approach to managing large environments. Using partitioned, remote caches with up to 15 million objects in the cache, the suite of benchmark tests showed that the bank could distribute data-sensitive applications across the grid without degradations in scalability, and could increase processing speed, capacity and throughput proportionally as it scaled its applications across more servers.

Among the specific results:

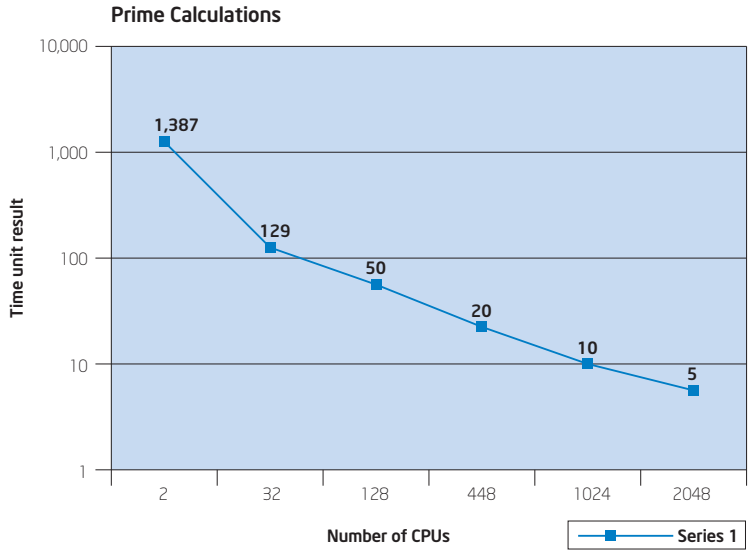
- Parallel processing, reading data from the cache, writing data to the cache, and the ability to load extremely large caches were all proven scalable. As many as 200,000 objects were written and read to the data grid concurrently, without glitches or errors.
- Even with terabyte-sized caches, the grid platform scaled to process large amounts of data quickly and efficiently. The system processed object payloads as large as 100 KB at rates of up to 100,000 transactions per second and higher, with cache sizes ranging from 15 GB to 1 TB, in configurations ranging from one to 500 client nodes.
- The speed of parallel processing in the grid scaled linearly.
- Results were validated up to the available 1,024 CPUs and extrapolated to a target initial deployment of 2,048 CPUs, demonstrating infrastructure capability to provide low latency for large-scale, end-to-end, real-time transactions.

Key Technologies: Proof-of-Concept

- 512 two-way IBM xSeries 355 servers, each running the Low-Voltage Dual-Core Intel Xeon processor with 3 GB of RAM on a 1 GB Ethernet network
- Linux* operating system
- GigaSpaces Technologies Software, Enterprise Edition & Caching Edition
- Java* development kits from IBM, BEA and Sun

Integral Answers

- There were no inhibitors to smooth linear scaling of up to 2,048 CPUs, with the expectation of smooth scaling well beyond that figure.
- Using distributed data caches, the solution handled terabytes of data in memory, with many users and applications reading and writing to large data caches.
- The bank can distribute data-sensitive applications across the grid without degradations in scalability, and can increase processing speed, capacity and throughput proportionally as it scales its applications across more servers.



The more CPUs, the faster the processing: The data grid proof-of-concept showed no inhibitors to smooth linear scaling of up to 2,048 CPUs, with the expectation of continued smooth scaling well beyond that figure.

Moving Forward

Following the success of the large-scale proof-of-concept, numerous groups in the bank's capital markets organization – including equities, derivatives, mortgage-backed securities, global foreign exchange, electronic trading and the rates business – are moving toward production deployment. In doing so, they are taking advantage of a new generation of data distribution solutions exemplified by GigaSpaces software, as well as power-efficient, dual-core Intel® technologies with hardware-assisted virtualization capabilities built in.

Already experiencing great success in using its grid to conquer CPU bottlenecks, the bank is poised to enhance the performance and scalability and drive down the costs of its data-sensitive computing applications. The tested grid solution handled massive amounts of data in memory, with many users and applications and reading and writing to those large caches.

Find a business solution that is right for your company. Contact your Intel representative, visit the Intel® Business/Enterprise Web site at: intel.com/business or visit the industry solutions-specific sites at: intel.com/business/bss/industry/ To learn more about digital office platforms visit: intel.com/it/digital-office/

Return on Investment

- With its virtualized, service-oriented environment, the bank can meet rising demands for computing resources while avoiding costly purchases of servers, storage and software licenses. The bank's IT leaders say the solution is saving them tens of millions of dollars over a three-year period.
- By expanding its ability to scale and accelerate its analytical, simulation and transaction processing applications, the bank can reduce risk exposure, improve profitability and adjust dynamically and efficiently to market shifts and customer demands. It is also positioned to execute its strategy of moving more aggressively into electronic trading.
- For companies who don't have the computing resources to perform similar tests, the proof-of-concept provided clear evidence that IBM xSeries servers with the Dual-Core Intel® Xeon® processor and GigaSpaces Technologies software provided smooth linear scaling of up to 2,048 CPUs, and used distributed data caches to deliver low latency for large-scale, data-intensive computing.

Solution provided by: IBM GIGASPACES

0906/KRS/RL/XX/PDF 315254-001US