Redpaper

Joshua Blumert
Dave Weber
Stephen Smith

STAC-M3 Benchmarks on the IBM System x3750 M4

In financial markets, the ability to seize opportunities faster than the competition is everything. For high-frequency trading (HFT), smart order routing, crossing, and algorithmic trading, the difference between profit and loss can be measured in microseconds. Firms are in a constant race to cut the time that is taken to process market data, conduct risk analysis, plan trades, and place orders.

These financial systems are built on a highly tuned software architecture that uses industry-standard multi-core processors to deliver the highly deterministic performance that financial markets firms need. By running tick databases on compact IBM® System x® 3750 M4 servers with up to four 8-core Intel Xeon E5-4600 series processors, firms can reliably achieve ultra-low latency and predictable performance.

In this IBM Redpaper™ publication, we describe the baseline STAC-M3™ Benchmarks (the “Antuco suite”) that are performed by the Securities Technology Analysis Center (STAC®) on a stack involving Kx Systems kdb+ 2.8 that are hosted on an IBM System x3750 M4 server with four 8-core Intel Xeon E5-4650 processors and eight 800 GB Intel S3700 SATA MLC Enterprise solid-state drives (SSDs). The server was connected to the storage using direct SAS connections that were managed by an IBM ServeRAID M5110e SAS/SATA controller.

Here are the topics that are covered in this paper:

Note: The official record of these benchmark results is the STAC Report, which is available to the public at http://www.stacresearch.com/node/13956. The official record of the detailed system configuration is the STAC Configuration Disclosure at http://www.stacresearch.com/node/13957 (may be subject to STAC access rules). Portions of the STAC Report are reproduced here with the permission of STAC. STAC bears no responsibility for any errors or omissions in this document.

Overview of the STAC-M3 Benchmarks

In this section, we describe why this particular configuration was tested and provide a brief overview of the STAC-M3 Benchmarks methodology.

About STAC-M3

STAC-M3 is a set of benchmark specifications and accompanying tools that are provided by the Securities Technology Analysis Center (STAC) for testing time-series management solutions (tick databases). Such solutions take historical and real-time streaming data as input, and perform user-defined operations on the resulting time series.

STAC-M3 provides a common basis for quantifying the extent to which emerging hardware and software innovations improve the performance of tick storage, retrieval, and analysis.

The STAC-M3 benchmark specifications were developed by the STAC Benchmark™ Council, which consists of users from the Financial Services Market and the vendors who serve them. The STAC process is user-driven; while vendors can contribute to STAC Benchmark specifications, the requirements and priorities are set by the user firms, and only the user firms are able to vote on the specifications. This ensures that the benchmarks are focused on real business needs.

For more information about STAC, see the following website:

http://www.STACresearch.com

The reason for benchmarking

The ability to process data (move it, store it, and react to it) quickly in the fast-paced Financial Services Market is the key factor in detecting and responding to market events rapidly and with confidence and for staying ahead of your competition.

In the worldwide financial market, tick data streams run constantly from multiple exchanges and they are getting faster. In the realm of automated trading, data that arrives just a few hundred milliseconds slower than someone else's means slower trading decisions, which translates to lost opportunities. In microseconds, trades can be won or lost. In this battlefield, technology constantly strives to keep up (better storage, robust algorithm applications, reliable networking solutions, and so on), but there are always the inevitable bottlenecks. With all of the advances that have been made to keep pace, the constant hindrance is still *low data latency*.

Each new generation of processors increases the volume of data traffic a single server can handle, which in turn leads to the further elimination of latency in server-to-server hops. Using four-socket IBM System x3750 M4 servers with the latest Intel Xeon processors, financial markets firms can achieve low latency while using less rack space and power, and generate less heat. Particularly for co-located servers, this offers significant operational cost benefits.

The benefits that are realized by this powerful architecture include the following ones:

- Average execution latency as low as 3.0 microseconds provides first-mover advantage for financial markets firms.
- The usage of advanced processor features that are found in the Intel Xeon E4-4600 processors, such as Intel Streaming SIMD Extensions (SSE) 4.1 and 4.2, and Intel Advanced Vector Extensions 2 (AVX2), enable faster vector calculations.
- IBM servers include factory-integrated low-latency network cards, for consistent performance.
- IBM has low-latency tuning expertise and presence in global financial centers, which combines to minimize support issues.

What the benchmark tests

Analyzing time-series data, such as tick-by-tick quote and trade histories, is crucial to many trading functions, from algorithm development to risk management. But the domination of liquid markets by automated trading, especially high-frequency trading, has made such analysis both more urgent and more challenging.

As trading robots try to outwit each other on a microsecond scale, they dish out quotes and trades in ever more impressive volumes. This places a premium on technology that can store and analyze that activity efficiently. For example, the faster an algorithm developer can back-test and discard a haystack of unprofitable ideas, the faster he finds the needle of a winning algorithm, leaving more time to use it in the market.

STAC-M3 tests the ability of a solution stack, such as columnar database software, servers, and storage, to perform various operations on a large store of market data. The STAC-M3 Working Group designed these test specs to enable useful comparisons of entire solution stacks (that is, to gauge the state of the art) as well as comparisons of specific stack layers while holding other layers constant. Comparisons can include (but are not limited to):

- Different storage systems, including SSD, DRAM, interconnects, and file systems
- Different server products, processors, chipsets, and memory
- Different tick-database products

As shown in Figure 1, the test setup for STAC-M3 consists of the “stack under test” (SUT) and client applications. No restrictions are placed on the architecture of the SUT or clients (though members of the STAC-M3 Working Group frequently provide input on architectures they want to see tested). Threads within the clients take in Randomized Reference Data (RRD), such as dates and symbols, submit requests for the required operations, receive responses, and store the timings and results from these queries. Vendor-supplied code for the operations and latency calculations are subjected to a combination of source-code inspection and empirical validation.

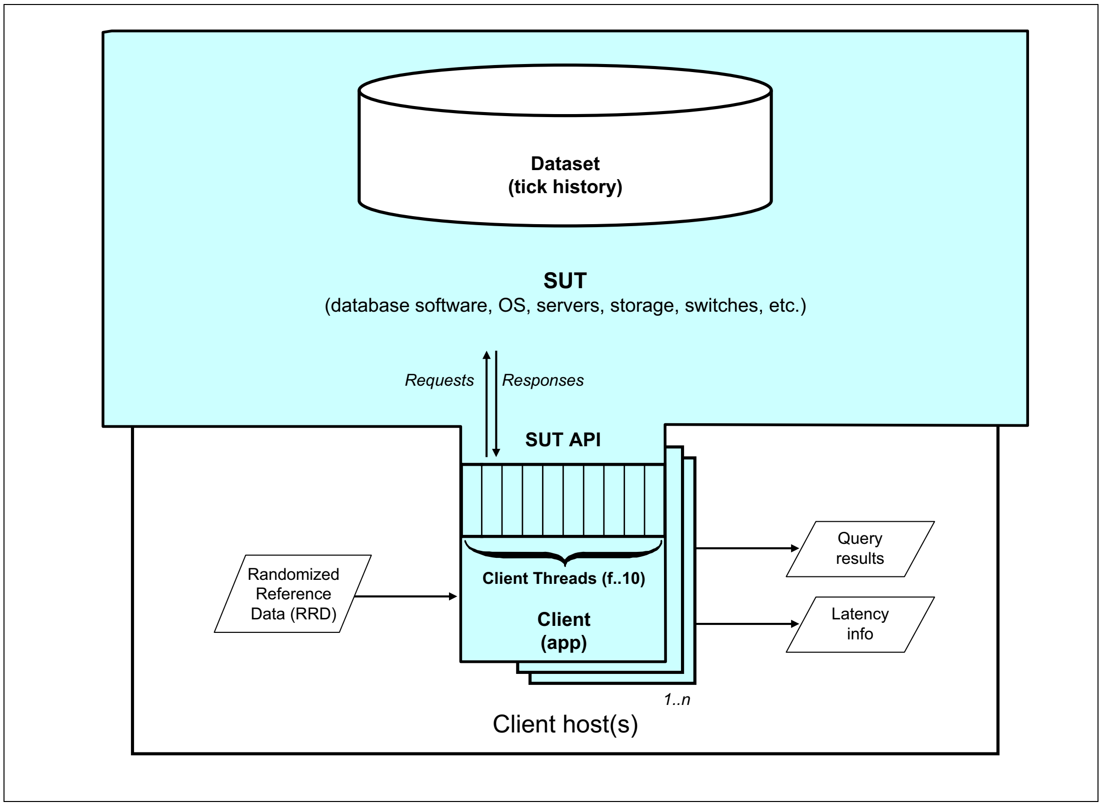

Figure 1   STAC-M3 Benchmark Stack Under Test (SUT) overview

Synthetic data set

STAC-M3 draws from client experience with equities and FX use cases. The database is synthetic, and is modeled on NSYE TAQ data (US equities). Although it is also desirable to test with real data, synthetic data has three advantages that make it compelling for this STAC-M3 suite:

- Synthetic data allows you to control the database properties exactly, which in turn allows us to randomize elements of queries from project to project while keeping the resulting workload the same (for example, you control how much volume is associated with each symbol).
- Synthetic data does not incur fee liability from a third party, such as an exchange.
- Synthesizing the data makes it easy to scale the database to an arbitrarily large size and run benchmarks against projected future data volumes.

The data set consists of high-volume symbols and low-volume symbols in proportions that are based on observed NYSE data. The data volume per symbol was based on doubling the typical volume in NYSE TAQ in 1Q10. The resulting database occupies approximately 4 TB of storage and is smaller than comparable databases at customer sites, as they typically contain many years of data. This was a deliberate choice by the STAC-M3 Working Group to minimize the cost of running benchmarks while still yielding valuable results.

Benchmarks that scale the database to the size of existing customer footprints and beyond are contained in the Kanaga suite of STAC-M3 Benchmark specifications.

Test metrics

The key metric in STAC-M3 is the latency of query responses (response times). Latency measurements are performed in the clients. A client thread gets a local time stamp (`tsubmit`) just before submitting a query. When the first results arrive, the client gets another time stamp (`tfirst`). When it receives the complete results (sorted appropriately), the client immediately gets a third time stamp (`tlast`). For systems that return all results in one chunk, the first-result and last-result time stamp are identical.

The algorithms in all benchmarks are defined to keep the result sets small. This ensures that network I/O between the test clients and servers is negligible compared to back-end processing times.

Some of the I/O-focused benchmarks also measure the bytes read per second from persistent storage (for example, excluding server cache), which is computed from the output of appropriate system utilities.

The STAC-M3 Benchmark Audit that was conducted by STAC included the following actions:

- Validating the database
- Inspecting any source-code revisions to the STAC Pack
- Validating the operation results
- Running the tests
- Checking the system configuration
- Documenting the results

Table 1 shows a brief overview of each test in this STAC-M3 suite.

Table 1 Summary of STAC-M3 Benchmark tests

Operation	Number of requesting client threads	Algorithm that is performed on behalf of each requesting client thread	Algorithm I/O intensity	Algorithm compute intensity	Input date range[a]
VWAB-Day	1	4-hour volume-weighted bid over one day for 1% of symbols (such as VWAP but operating on quote data, so much higher input volume).	Heavy read	Light	Last 30 days
VWAB-12DaysNoOverlap	100	4-hour volume-weighted bid over 12 days for 1% of symbols. No overlap in symbols among client threads.	Heavy read	Light	Full year
Year High Bid	1	Maximum bid over the year for 1% of symbols.	Heavy read	Light	Full year
Year High Bid Rerun	1	Rerun of Year High Bid (same symbols) without clearing the cache.	Heavy read[b]	Light	Full year
Quarter HighBid	1	Maximum bid over the quarter for 1% of symbols.	Heavy read	Light	Most recent quarter
Month High Bid	1	Maximum bid over the month for 1% of symbols.	Heavy read	Light	Most recent month
Week High Bid	1	Maximum bid over the week for 1% of symbols.	Heavy read	Light	Most recent week
Aggregate Stats	10	One set of basic statistics over 100 minutes for all symbols on one exchange. Each 100-minute range crosses a date boundary.	Heavy read	Heavy	Full year
Stats - Unpredictable Intervals	1, 10, 50, or 100 (more optional)	Per-minute[c] basic statistics over 100 minutes for all high-volume symbols on one exchange. Each 100-minute range crosses a date boundary.	Heavy read	Heavy	Full year
Market Snapshot	10	Most recent trade and quote information for 1% of symbols as of a random time.	Heavy read	Heavy	Full year
Volume Curves	10	Create an average volume curve (using minute intervals that are aligned on minute boundaries) for 10% of symbols over 20 days selected at random.	Light read	Heavy	Full year
Theoretical P&L	10	For a basket of 100 trades on random dates, find the future times at which 2X, 4X, and 20X the trade size traded in each symbol. Trade sizes cause up to 5 days of forward searching. Calculate the corresponding VWAP and total volume traded over those periods.	Light read	Heavy	Full year

Operation	Number of requesting client threads	Algorithm that is performed on behalf of each requesting client thread	Algorithm I/O intensity	Algorithm compute intensity	Input date range[a]
NBBO	1	Create the NBBO across all 10 exchanges for all symbols on the most recent day. Write to persistent storage.	Heavy read and write	Heavy	Most recent day
Write	1	Write one day's quote data to persistent storage, following the same algorithm that is used to generate the randomized data set used in the other operations.	Heavy write	Light	N/A
Storage efficiency	n/a	Reference Size of the data set divided by size of the data set in the SUT format that is used for the performance benchmarks. Expressed as a percentage.	N/A	N/A	N/A

[a] In some cases, one or more dates at the end of the year were excluded from eligibility to prevent an algorithm that crosses days from running out of input data.

[b] Typically, this is reads from DRAM cache.

[c] In this case, interval start times are offset from minute boundaries by a consistent random amount per test run so that the SUT cannot rely on pre-calculated minute statistics.

IBM System x3750 M4

The IBM System x3750 M4, shown in Figure 2, is a 4-socket server that features a streamlined design, optimized for price and performance, with best-in-class flexibility and expandability. Models of the x3750 M4 are powered with Intel Xeon E5-4600 processors, up to 8 cores each, for an entry-level 4-socket solution. The x3750 M4 provides maximum storage density, with flexible PCI and 10 Gb Ethernet networking options in a 2U form factor.

Figure 2   IBM System x3750 M4 server

The x3750 M4 has outstanding memory performance that is achieved by supporting three-RDIMM-per-channel configurations at speeds up to 25% faster than the Intel specification, while still maintaining world-class IBM reliability. LR-DIMM speeds are also 25% beyond the Intel specification for 1.35 V DIMMs, and this speed improves not only performance, but reduces overall system power at the same time.

The x3750 M4 offers a flexible, scalable design and simple upgrade path to 16 hard disk drives (HDDs) or 32 IBM eXFlash solid-state drives (SSDs), with up to eight PCIe Gen 3 slots and up to 1.5 TB of memory. The flexible embedded Ethernet solution provides two standard Gigabit Ethernet ports onboard, along with a dedicated 10 GbE slot that allows for a choice of either two copper or two fiber optic connections. Comprehensive systems management tools with the next-generation Integrated Management Module II (IMM2) make it easy to deploy, integrate, service, and manage.

The value of the x3750 M4, especially in relation to the STAC M3 workloads, is two fold:

- With the 48 DIMM sockets, the amount of memory that the x3750 M4 can have is significant. In such workloads, an increase in the amount of available memory means better performance.
- The x3750 M4 is a dense 2U system that offers four processors or up to 32 cores. This level of compute power density leads to a significant efficiency/consolidation advantage for customers, who can save on footprint space, which is an excellent value proposition, especially with expensive data center or collocation costs.

Key features

The IBM System x3750 M4 blends outstanding flexibility and expandability. The x3750 M4 2+2 socket design enables pay-as-you-grow processing with the new Intel Xeon E5-4600 series processors and memory scalability to help lower cost and manage growth. The 5+3 PCIe socket design allows you to pay for PCIe capabilities as needed.

With the capability to support up to 48 DIMMs, four sockets, mix and match internal storage with up to 16 HDDs or 32 eXFlash SSD drives, 6 hot-swap dual rotor fans, two power supplies, and integrated 10 GbE networking with options for fiber or copper, the x3750 M4 provides unmatched features and capabilities in a dense 2U design.

Scalability and performance

The x3750 M4 offers numerous features to boost performance, improve scalability, and reduce costs:

- The Intel Xeon processor E5-4600 product family improves productivity by offering superior system performance with 8-core processors and up to 2.9 GHz core speeds, up to 20 MB of L3 cache, and up to two 8 GTps QPI interconnect links.
- The x3750 M4 2+2 processor socket design enables pay-as-you-grow processing with the Intel new Xeon E5-4600 series processors and memory scalability to help lower cost and manage growth.
- Up to four processors, 32 cores, and 64 threads maximize the concurrent execution of multithreaded applications.
- Intelligent and adaptive system performance with Intel Turbo Boost Technology 2.0 allows processor cores to run at maximum speeds during peak workloads by temporarily going beyond processor thermal design power (TDP).
- Intel Hyper-Threading Technology boosts performance for multithreaded applications by enabling simultaneous multithreading within each processor core, up to two threads per core.
- Intel Virtualization Technology integrates hardware-level virtualization hooks that allow operating system vendors to better use the hardware for virtualization workloads.
- Intel Advanced Vector Extensions (AVX) improve floating-point performance for compute-intensive technical and scientific applications compared to Intel Xeon 5600 series processors.

- The outstanding RDIMM memory performance of the x3750 M4 is achieved by supporting three DIMMs per channel configurations at speeds up to 25% faster than the Intel specification.
- 48 Load Reduced DIMMs (LR-DIMMs) of 1333 MHz DDR3 ECC memory provide speed, high availability, and a memory capacity of up to 1.5 TB.
- LR-DIMM speeds implemented in the x3750 M4 are also 25% beyond the Intel specification at 1.35 V for one, two, and three DIMM per channel configurations. This configuration improves performance and reduces overall system power at the same time, all while maintaining reliability.
- The use of IBM eXFlash solid-state drives (SSDs) instead of, or along with, traditional hard disk drives (HDDs), can improve I/O performance. An SSD can support up to 100 times more I/O operations per second (IOPS) than a typical HDD.
- Up to 16 HDDs or 32 eXFlash SSDs, together with an optical drive at the same time, provide a flexible and scalable all-in-one platform to meet your increasing demands.
- The server offers a SAS switch backplane option (88Y7421) to allow up to 16 SFF devices to attach to a single controller.
- The server has two integrated Gigabit Ethernet ports and two optional 10 Gb Ethernet ports that do not consume PCIe slots.
- The 5+3 PCI Express socket design of the server allows you to pay for PCIe capabilities as needed.
- The server offers PCI Express 3.0 I/O expansion capabilities that improve the theoretical maximum bandwidth by almost 100% (8 GTps per link using 128b/130b encoding) compared to the previous generation of PCI Express 2.0 (5 GTps per link using 8b/10b encoding).
- With Intel Integrated I/O Technology, the PCI Express 3.0 controller is integrated into the Intel Xeon processor E5 family. This integration reduces I/O latency and increases overall system performance.

Availability and serviceability

The x3750 M4 provides many features to simplify serviceability and increase system uptime:

- The server offers Chipkill, memory mirroring, and memory rank sparing for redundancy if there is a memory failure.
- The server provides restart recovery for any failed processor. If there is a failure of processor 1, the server connects the south bridge to processor 2 for reboot.
- Tool-less cover removal provides easy access to upgrades and serviceable parts, such as the processor, memory, and adapters.
- The server offers hot-swap drives, supporting RAID redundancy for data protection and greater system uptime.
- The server has up to two redundant hot-swap power supplies and six hot-swap dual-rotor N+N redundant fans to provide availability for business-critical applications.
- The power source independent light path diagnostics panel and individual light path LEDs lead the technician to failed (or failing) components, which simplifies servicing, speeds up problem resolution, and helps improve system availability.
- Predictive Failure Analysis (PFA) detects when system components operate outside of standard thresholds and generates proactive alerts in advance of a possible failure, therefore increasing uptime. These components support PFA:
  - Memory
  - SAS/SATA HDDs

  - Fans
  - VRDs
  - Power supplies
- Solid-state drives (SSDs) offer more reliability than traditional mechanical HDDs for greater uptime.
- The built-in Integrated Management Module Version II (IMM2) continuously monitors system parameters, triggers alerts, and performs recovery actions in case of failures to minimize downtime.
- Built-in diagnostic tests, using Dynamic Systems Analysis (DSA) Preboot, speed up troubleshooting tasks to reduce service time.
- Three-year customer-replaceable unit and onsite limited warranty, 9 x 5 next business day. Optional service upgrades are available.

Manageability and security

Powerful systems management features simplify local and remote management of the x3750 M4:

- The server includes an Integrated Management Module II (IMM2) to monitor server availability and perform remote management. Remote presence support is standard.
- The integrated industry-standard Unified Extensible Firmware Interface (UEFI) enables improved setup, configuration, and updates, and simplifies error handling.
- Integrated Trusted Platform Module (TPM) 1.2 support enables advanced cryptographic functionality, such as digital signatures and remote attestation.
- There is industry-standard Advanced Encryption Standard (AES) NI support for faster, stronger encryption.
- IBM Systems Director is included for proactive systems management. It offers comprehensive systems management tools that increase uptime, reduce costs, and improve productivity through advanced server management capabilities.
- Intel Execute Disable Bit functionality can prevent certain classes of malicious buffer overflow attacks when combined with a supported operating system.
- Intel Trusted Execution Technology provides enhanced security through hardware-based resistance to malicious software attacks, allowing an application to run in its own isolated space, which is protected from all other software running on a system.

Energy efficiency

The x3750 M4 offers the following energy-efficiency features to save energy, reduce operational costs, increase energy availability, and contribute to a green environment:

- Energy-efficient system board components help lower operational costs.
- There are highly efficient 900 W and 1400 W AC power supplies with 80 PLUS Platinum certification at high voltage AC.
- The Intel Xeon processor E5-4600 product family offers better performance over the previous generation while fitting into the same TDP limits.
- Intel Intelligent Power Capability powers individual processor elements on and off as needed to reduce power draw.
- Low-voltage Intel Xeon processors draw less energy to satisfy the demands of power and thermally constrained data centers and telecommunication environments.
- Low-voltage 1.35 V DDR3 memory RDIMMs consume 19% less energy compared to 1.5 V DDR3 RDIMMs.

- SSDs consume as much as 80% less power than traditional spinning 2.5-inch HDDs.
- The server uses hexagonal ventilation holes, which are a part of IBM Calibrated Vectored Cooling™ technology. Hexagonal holes can be grouped more densely than round holes, providing more efficient airflow through the system.
- IBM Systems Director Active Energy Manager™ provides advanced data center power notification and management to help achieve lower heat output and reduced cooling needs.

For more information about the x3750 M4, see *IBM System x3750 M4*, TIPS0881, found at:

http://www.redbooks.ibm.com/abstracts/tips0881.html?Open

Case study: Redline Trading Solutions

With a huge amount of money at stake, financial markets firms are in a constant arms race to minimize end-to-end latency and seize fleeting market opportunities ahead of the competition.

Redline Trading Solutions (Redline) creates ultra-low latency market data and order execution systems that enable firms to excel in today's equities, options, futures, and FX markets. With Redline solutions running on IBM System x3750 M4 servers, trading firms can reliably achieve ultra-low latency and predictable performance.

The benefit of using the x3750 M4 in such environments is extreme low latency and deterministic performance, which enable firms to spot and act on profitable trading opportunities faster than competitors.

Redline speaks highly of the x3750 M4:

> "'Intel's rapid progression in core-count per socket is great news for firms who seek ultra-low latency in the smallest possible physical package,' says Lee Fisher, VP, Technical Marketing, Redline Trading Solutions. 'With servers like the IBM System x3750 M4, they can pack more processing into rented co-located rack space, reducing costs and latency. IBM has enormous expertise in tuning systems for low latency, and offers support in financial centers worldwide.'
>
> 'To help financial markets clients benefit from best practices in deploying leading-edge technologies, tuning and configuring their systems, and benchmarking industry-standard financial test suites such as STAC, IBM has established the Wall Street Center of Excellence, a global initiative that focuses on low latency and high performance technologies.'
>
> Fisher adds, 'We test our software extensively on IBM System x, configured according to guidance from the IBM Wall Street Center of Excellence. Our IBM clients can also configure their servers in the same recommended way to achieve the expected performance. In our experience, IBM System x servers deliver predictable high throughput and low latency based on their high-performance design, build-quality, and integrated network adapters.'"[1]

For more information, see the System x Case Study on Redline's use of x3750 M4 systems at the following website:

http://ibm.com/software/success/cssdb.nsf/CS/STRD-99UF2Z?Open

[1] Source: http://ibm.com/software/success/cssdb.nsf/CS/STRD-99UF2Z?Open

Benchmark technical requirements

This section describes the technical requirements for performing the benchmark.

Database

The STAC-M3 benchmark does not specify a particular database software to test. Instead, it is the decision of the team that is running the benchmark to select the correct database software for the test. The database is one of the items that contributes to the overall performance of the system. Therefore, the database is a critical component. It must support high performance I/O to read in the large data set that this test defines.

In general, each database *language* requires its own STAC-M3 implementation (called a STAC Pack). Generally, the database vendor creates the STAC Pack for their database. Kx Systems kdb+ is one of the most popular databases for this test, and they provide the necessary scripts to run the STAC-M3 benchmark.

Kdb+ is both an in-memory and disk-based database, so it offers the highest performance possible, using all the available RAM and high-speed disk storage. The database maps from disk to memory as required, and the in-memory portion is used for real-time data, such as collecting feeds from exchanges.

Kx delivers a combination of kdb+, which is an ultra high-performance database with a unified format for real-time and historical data, and q, which is an exceptionally efficient proprietary language, which enable trading operations to manage risk and implement sophisticated trading strategies in real time. Kx simultaneously supports thousands of real-time custom queries and analyses on historical / in-memory data. kdb+ is a scalable, column-oriented database for processing massive data volumes.

Operating system

IBM has found that performing well on STAC-M3 with kdb+ requires a platform that supports a large SMP system, large memory size, and high performance I/O. Intel -based systems are commonly used for this type of environment, so Linux and Windows are both options for the base operating system. In addition, the database that is selected must support the operating system.

Although both Linux and Windows are suitable, Linux is often the choice because of its popularity with high-performance computing (HPC) applications. In this test, Red Hat Enterprise Linux is a common choice. We used RHEL 6.4 in the benchmark and this version has many options to tune the system to maximize performance. In this benchmark, we took advantage of these tuning options.

Server

As Table 1 on page 6 shows, the STAC-M3 benchmarks are primarily I/O-intensive, but do include some compute-intensive elements. Therefore, this test performs well on systems that have both good I/O subsystems, large memory capacity, and large symmetric multi-processing (SMP) capabilities.

Systems with four processors, such as the IBM System x3750 M4, offer an ideal mix for this type of workload. The x3750 M4 can support up to 32 processor cores (64 with hyperthreading) and up to 1.5 TB of system memory. The system also supports the latest PCIe 3.0 interfaces for high I/O controller bandwidth.

The IBM System x3750 M4 provides advanced features and capabilities in a dense 2U design. These include support for up to four sockets and 48 DIMMs, mix and match internal HDD or SSD storage, dual power supplies, and integrated 1 Gigabit Ethernet (GbE) and 10 GbE networking with options for fiber or copper. The unique 2+2 socket design enables pay-as-you-grow processing and memory expansion to help lower cost and manage growth. The 5+3 PCIe socket design allows you to pay for I/O capabilities as needed.

The x3750 M4 capabilities and performance allow clients to reduce total cost of ownership (TCO) by up to 52% over 4 years by consolidating multiple 2-socket servers into fewer 4-socket x3750 M4 servers.

Storage

Because of the I/O-intensity of these tests, the ability to transfer large amounts of data quickly is important. Performance from the storage system is reflected directly in the results of this benchmark. In previous runs of this test, there were various storage systems that were used, from traditional HDDs to SSDs to DRAM subsystems.

There are considerations of performance, capacity, and total cost. HDDs provide the best cost for capacity, but many spindles are required to achieve the performance that is required. SSDs provide faster access per drive, and the price of SSDs are rapidly decreasing. But, traditional SSD disks use SAS controller interfaces, so multi-controllers are necessary to maintain high bandwidth. Specialized solid-state controllers are also a consideration because these subsystems can take advantage of a higher bandwidth interface, both internally with PCI and externally with InfiniBand. IBM FlashSystem™, formerly Texas Memory Systems®, is a good example of this type of system. There are also dedicated DRAM-based solutions, but cost can be a factor when you try to accommodate large amounts of historical data.

For more information about IBM FlashSystem, go to the following website:

http://ibm.com/systems/storage/flash/

Another consideration for storage is a tiered subsystem, which uses different classes of storage to optimize higher-cost, high-performance disks or memory with more cost-efficient lower-cost “deep” drives for more data storage. This approach requires intelligent management of the tiers based on data access, which can be managed through the file system or at the storage subsystem.

In the most recent test, the server was equipped with Intel SSDs. The Intel Solid-State Drive DC S3700 Series delivers superior performance with fast random read performance and random write performance. With excellent input/output operations per second (IOPS) distribution and low maximum latencies, the DC S3700 Series ensures quick and consistent command response times. All this performance is delivered with low active power consumption.

The DC S3700 Series has the following features:

- *Full End to End Data Protection* protects your data from the time it enters the drive to the time it leaves. The DC S3700 uses an advance error correction scheme that ensures data integrity by protecting against possible data corruption in the NAND, SRAM, and DRAM memory. The DC S3700 also protects the data in transit through several techniques, such as parity checks, Cyclic Redundancy Checks (CRC), and LBA tag validation. After an error is detected, an immediate attempt is made to correct it, and any uncorrectable error is reported to the host. To further improve data assurance, the Intel SSD DC S3700 provides an array of surplus flash memory that caches data to minimize potential data loss.
- *Enhanced Power-Loss Data Protection* reduces potential data loss by detecting and protecting data from an unexpected system power loss. The drive saves all cached data while it is being written before shutting down, thereby minimizing potential data loss.

Interfaces (network, SAN, and connectivity)

Another important component of any benchmark is the connectivity for network and storage. Although STAC-M3 is neutral about the number of servers that are used, most published STAC-M3 benchmarks have used a single server, meaning client/server and server-to-server networking is not a major consideration. However, if STAC-M3 were used to test a multi-server system, then a high-performance, low latency interconnect is important.

Because storage is a key component, the interface to the storage subsystem is critical. In this case, the usage of local SSD dictates internal SAS connectivity, but external drives might be necessary to increase the amount of storage for capacity and performance. SAS can still be used because of its low cost and good performance (6 Gbps) For higher performance environments, InfiniBand, Fibre Channel, and 10 Gb Ethernet also might be considered.

File system

The database that is used for the STAC-M3 benchmark is on the disk subsystem and typically uses flat files on a file system. Therefore, the file system is also an important component of this test.

The database is large and can contain many files. The file system should support high capacity and many files. Also, the file system must support streaming data and fast metadata access. If the operating system is Linux, there are many file systems to choose from. EXT4, XFS, and BTRFS are native open source file systems that are supported by most distributions. If a cluster is required for multi-system access, then there are a number of clustered file systems to consider. Gluster and Luster are two open source clustered file systems.

IBM GPFS™ is a proprietary clustered file system that is designed for high performance and multi-system access. In “Performance tuning” on page 15, we list the tuning parameters for common file systems. Because the data is read sequentially, large block sizes (greater than 1 MB) should be used for the creation of the file system. For more information, go to the following website:

http://ibm.com/systems/software/gpfs/

To achieve the capacity and bandwidth that is required for this test, many disk spindles are required. Good performance requires that the data uses all of the disks during the I/O operations. There are many to achieve this task:

- Create many small file systems, with a single RAID group per file system. The database distributes the data across all of the file systems that are created. This is the simplest approach, but can be the most difficult to manage as the database changes or grows. The simple design requires the least amount of file system tuning and a basic file system, such as EXT3 or EXT4, can be used, but there can be many file systems to manage.
- Use a volume manager to create a single large volume by striping many RAID groups. This creates one large file system for the database. The advantage is that there is only one file system to manage and the volume manager can be used to grow or modify the file system.

  This design requires more advanced planning for stripe sizes and file system parameters to optimize the I/O. Some of the tuning and planning can be moved to the disk subsystem if the storage array can handle the disk striping across many disks. Storage controllers, such as IBM Storwize® V7000, IBM XIV®, and IBM SAN Volume Controller, all can use a many disks to create a single volume. In this case, the controller can optimize performance across all of the disks that are provided. In either case, the file system must efficiently use the very large volume that is created. XFS is a good choice in this example.

  For more information about the IBM SAN Volume Controller, see the following website:

  http://www-03.ibm.com/systems/storage/software/virtualization/svc/
- Use a cluster file system. Clustered file systems, such as IBM GPFS, can intelligently manage many RAID groups or JBOD disks. The management and performance tuning is then handled by the file system. This requires more planning and tuning in the file system design, but is a flexible architecture because storage arrays and disks can be used. This design can be significant cost savings because this is a software design.

Performance tuning

This section describes the tuning actions that should be taken before STAC-M3 benchmarking to ensure optimal performance.

System tuning

Because the STAC-M3 is primarily an I/O-intensive test, it is a preferred practice to design and tune the system for streaming I/O and test the system performance with artificial load generators to measure the systems performance before starting the benchmark. Also, the database must be optimized for both large SMP and I/O, so optimization for multi-threading is required.

Hardware architecture

The hardware design must be optimized for streaming data and multi-threading. Tuning starts with the setup of the disk subsystem. Multiple paths to the storage array are important and I/O controllers should be distributed over multiple PCIe slots. Many Intel based PCIe slots may be grouped by I/O controllers (south bridge design) and in this case, I/O interface cards should be evenly distributed over the controllers. In newer Intel based systems, the *south bridge* design is replaced, and the I/O is directly connected to the processors. In this case, there might be different processors that control the various PCIe slots. Distributing the I/O over the processors can be a performance benefit.

Hardware tuning

Part of the setup of the system is setting the hardware tuning parameters in the BIOS or uEFI firmware. There can be many hardware tuning options that can have a significant impact on performance. These parameters are accessed through system POST by pressing the F1 key or by using the command-line tools of the operating system. IBM System x servers support a tool that is called the Advanced Systems Utility (ASU), which is part of IBM Toolscenter:

`http://www-947.ibm.com/support/entry/portal/docdisplay?lndocid=tool-center`

This tool allows for scripting changes to hardware options, which is convenient for large system deployments.

Some of the more important system settings for this type of benchmark are listed below. A complete list of ASU/uEFI settings can be found at the following website:

`http://ibm.com/support/entry/portal/docdisplay?lndocid=MIGR-5083207`

ASU settings

The following settings can result in better performance:

- `uEFI.TurboModeEnable=Enable`
- `uEFI.PerformanceStates=Enable`
- `uEFI.PackageCState=ACPI C3`
- `uEFI.ProcessorC1eEnable=Disable`
- `uEFI.DDRspeed=Max Performance`
- `uEFI.QPISpeed=Max Performance`
- `uEFI.EnergyManager=Disable`
- `uEFI.OperatingMode=Performance Mode`

Hyper-Threading Technology

Hyper-Threading (HT) Technology is the Intel proprietary simultaneous multithreading (SMT) implementation that is used to improve parallelization of computations (doing multiple tasks concurrently.

For each processor core that is physically present, the operating system addresses two virtual or logical cores, and shares the workload between them when possible. The main function of HT is to decrease the number of dependent instructions in the pipeline. It takes advantage of superscalar architecture (multiple instructions operating on separate data in parallel). They appear to the OS as two processors, so the OS can schedule two processes at once. In addition, two or more processes can use the same resources.

In earlier implementations of HT, it was a preferred practice to disable this feature because some threads could become *processor starved*. In the current Intel processors, HT is more optimized and large SMP benchmarks can benefit from HT being enabled. It is worth testing with it on and off.

Operating system tuning

Correct setup and tuning of the operating system is critical to good performance. Both Linux and Windows have many tuning options. Microsoft, Red Hat, and SUSE all publish system tuning guides. Because this test is frequently run on Red Hat Enterprise Linux, we list some of the tuning options for RHEL 6. Detailed tuning for Red Hat Enterprise Linux can be found on the Red Hat website:

https://access.redhat.com/site/documentation/en-US/Red_Hat_Enterprise_Linux/6/html/Performance_Tuning_Guide/index.html

For the STAC-M3 test, the following sections provide a list of tuning preferred practices.

tuned-adm

This command line utility (**# tuned-adm list**) allows user to switch between user-defined tuning profiles. Several predefined profiles are included for some of the more common cases. There are three profiles that are suitable for this benchmark, depending on the type of storage that is used:

- Throughput-performance

  A server profile for typical throughput performance tuning. It disables **tuned** and **ktune** power-saving mechanisms, enables **sysctl** settings that improve the throughput performance of your disk and network I/O, and switches to the deadline scheduler.

- Latency-performance

  A server profile for typical latency performance tuning. It disables **tuned** and **ktune** power-saving mechanisms and enables **sysctl** settings that improve the latency performance of your network I/O.

- Enterprise-storage

  A server profile to improve throughput performance for enterprise-sized server configurations. This switches to the deadline scheduler and disables certain I/O barriers, dramatically improving throughput.

Default installation

With RHEL 6, the default base installation is now a minimal installation with limited packages. Select this installation and set up the RHEL 6 YUM repository. Then, you must install only the packages that are needed for the applications that the server runs. This eliminates processing impact.

I/O schedulers

There are several I/O schedulers that are included with RHEL 6. They can be applied globally through /etc/grub.conf on the boot option line by using **elevator=<*scheduler*>**. This applies to all devices, or specific devices can be assigned individually by using **/sys/block/<device>/queue/scheduler**. To set the deadline scheduler to /dev/dm-1, run the following command:

```
# echo deadline > /sys/block/dm-1/queue/scheduler
```

There are three main I/O schedulers: completely fair queuing (CFQ), NOOP (first in, first out), and deadline. Deadline is good for high-throughput applications. NOOP is good for random I/O work loads. Deadline is appropriate for this benchmark. The I/O optimized profiles of **tuned-adm** set the scheduler to deadline.

HBA queue depth

If the storage subsystem is access through a SAN fabric and FC HBAs are used, there is an HBA queue depth. Increasing the queue depth can increase the performance of multipathed storage. You can increase the storage in Linux by using an HBA module option. As of RHEL 6, this is set in `/etc/modprobe.d/`.

Multipath configuration

The `/etc/multipath.conf` file controls which devices are managed by the multipath daemon (**`multipathd`**). Multipath must be enabled by running the following command:

```
#chkconfig multipathd on
```

Red Hat provides a default `multipath.conf`, which has base disks that are *blacklisted* by default. If the system in not booting from SAN, the root disk device should be blacklisted, otherwise any changes to multipath require a reboot.

IRQ load balancing

IRQ for a device in Linux is mapped to a specific interrupt, which can be listed by running **`cat /proc/interrupts`**.

IRQ load balancing enables mapping specific interrupts to specific processors. This is enabled by using **`/proc/irq/<interrupt number>/smp_affinity`**, where **`smp_affinity`** is a bit mask for the number of processors in the system. **`cat ffffff > /proc/irq/25/smp_affinity`** enables interrupt 25 on all processors. A more efficient usage of the system is to assign a specific interrupt to specific processors. Network, HBA, and RAID controllers can be masked to different processors.

Service management

Disables and stops all unnecessary services. There are a number of tasks and services that are enabled by default that are not needed for many server environments. For example, the `cups` print service is enabled, as is `isdnd` and others.

Task processor affinity

This **`#taskset -p mask pid`** command enables the assignment of processes to specific processors. The command uses either a bid mask to specify the processors or uses a processor list. The Kx kdb+ database allows for thread and process control. Each set of database threads can be assigned to a specific set of processors. Table 2 provides a description of these assignments.

Table 2 Database thread and process control

DB process	Processors
1	`taskset -c 0-9`
2	`taskset -c 10-19`
3	`taskset -c 20-29`
4	`taskset -c 30-39`
5	`taskset -c 40-49`
6	`taskset -c 50-59`
7	`taskset -c 60-69`
8	`taskset -c 70-79`

numactl

As of the current generation of Intel processors, the system architecture is based on non-uniform memory access (NUMA). In this design, each processor socket has its own bank of memory. Then, the processors are interconnected by a high speed link (QPI). Therefore, when memory is accessed from the local processor's memory, access latency is lower than accessing memory cross the processor interconnect. Linux now includes the **numactl** utility. **numactl** allows for placement of processes on specific processors and memory. This can impact performance for applications that are latency sensitive. **numactl** was not used on the benchmark that is described in this paper, but could be an option for future tests.

Control groups (cgroups)

With RHEL 6, there is a new kernel feature that is known as *control groups* (cgroups). This mechanism allows the allocation of system resources, processors, memory, network, and I/O for a group of tasks. Cgroups are controlled by the **cgconfig** service. Cgroups are defined for a set of processes and are hierarchical.

Blockdev

The **blockdev** command allows for changes to the I/O controls (**ioctl**) to a blockdev. If an application is read intensive, **blockdev** can adjust the read ahead depth in sectors. The default is 32 and it can be increased to 2048.

Security Enhanced Linux

Security Enhanced Linux (SELINUX) is on by default on RHEL 6. It can be disabled by editing /etc/sysconfig/selinux. Unless the server has specific security requirements by the user, disabling SELINUX enhances performance. Also, some applications do not run with SELINUX enabled, so many software vendor suggest disabling SELINUX.

Huge pages

As a preferred practice, disable Transparent Huge Pages.

Swap file

As a preferred practice, allocate a swap file at least as big as the physical memory.

Performance monitoring

Monitoring system performance is a key part of system performance tuning. There are many utilities in Linux to monitor performance under load. Table 3 lists the tools for I/O usage and how to use them.

Table 3 Performance monitoring tools

Utility	Description
sysstat	This is the RPM package that contains the **sar**, **iostat**, **mpstat**, and **pidstat** utilities. This package is not part of the base RHEL 6 installation. **nfsstat** is part of nfs-utils, **dstat** is its own RPM, and **procp** is the RPM that contains **vmstat**, **top**, and **free**, and is installed with the base RHEL 6 installation.
System Activity Report (SAR)	This is the RPM package the contains **sar**, **iostat**, **mpstat**, and **pidstat**. This package is not part of the base RHEL 6 installation. **nfsstat** is part of nfs-utils, **dstat** is its own RPM, and **procp** is the RPM that contains **vmstat**, **top**, and **free**, and is installed with the base RHEL 6 installation.
vmstat (Virtual Memory Statistics)	This utility interactively displays virtual memory statistics and I/O data. For disk I/O, use the **–d** flag.

Utility	Description
iostat	Displays processor, disk, and NFS data interactively.
top	Real-time display of the processes that are running. It is useful for tracking process usage, but **nmon** can perform the same function and display other data.
nfsstat	This is the NFS statistics viewer. It is not as useful as **iostat** for real-time NFS activity.
netstat	Network configuration details. It is more useful as a debugging tool than as a monitoring tool. Much of the data is cumulative since the last reset or reboot.
ifstat	Real-time display of network traffic. It is not part of RHEL, but **dstat** has the same data.
dstat	This is part of RHEL 6 (dstat.rpm) and combines the information from **iostat**, **vmstat**, and **ifstat** in to one tool. Also, the display of **dstat** is concise and easy to read.
nmon (IBM)	This is a comprehensive real-time and batch performance display tool from IBM AIX®. It has been ported to Linux. Although it is not part of RHEL 6, it can easily be downloaded and run on RHEL.
iotop	This utility is included in RHEL 6 as an optional package (iotop.rpm), **iotop** displays, in real time, the top processes in order of I/O.
vnstat	This utility provides both summary and real-time output of network traffic. It is not part of RHEL 6 distribution, but can be found online and installed.
atop	An interactive monitor to view the load.
htop	An interactive process viewer.

Performance validation

After the system is configured and the tuning options are set, it is a preferred practice to validate the performance of the configuration before starting the STAC-M3 benchmark.

Artificial load generation tools are useful for creating an I/O load on the system, and then the system performance can be monitored. For a Linux system, IORATE (http://iorate.org/) is a command-line load generator. IORATE is a good tool to test the system with because it can generate I/O to multiple devices, both RAW and formatted, and the number of threads per device is a configuration option.

For initial tests with IORATE, use 1 MB blocksizes, and create one thread per processor core. The system generally performs well with multiple I/O devices or files. Therefore, for a 32 core system with HT enabled, use eight devices or files with eight threads each. During the load test, the I/O to each device can be monitored from Linux by running **iostat –d –x –m 1**.

nmon is also useful because it can also display, both in graphical and text-based form, the system I/O performance. Ultimately, the aggregate system throughput should be in the multiple GB per second range. This is not a guarantee of good results, but poor I/O throughput generally yields poor results on the STAC test.

Results

As noted earlier, the official record of these results is the STAC Report, which is available to the public at http://www.stacresearch.com/node/13956. For detailed benchmark versions and other information, see the STAC Report. The results are summarized here for convenience.

The baseline STAC-M3 Benchmark tests include:

- "Light-Compute benchmarks"
- "Post-Trade Analytics benchmarks" on page 22
- "Research Analytics benchmarks" on page 23
- "NBBO benchmark" on page 24
- "Multi-day/multi-user VWAB benchmark" on page 25

Kdb+ version: This run of STAC-M3 was implemented with kdb+ 2.8. Since this test, Version 3.1 of kdb+ was released. This new version changes how the database reads data from the I/O subsystem, which can result in significant improvement in the overall results of the STAC-M3 test. Therefore, results with kdb+ 2.x should not be compared to Version 3.x for evaluation of the hardware performance.

Light-Compute benchmarks

The Light-Compute benchmark tests include a high bid, single client thread test and a write test for one day of data.

High bid (one client thread requesting)

Table 4 shows the results for the high bid for a certain 1% of symbols over varying time frames.

Table 4 High bid benchmarking results

Latency (milliseconds)	**MEAN**	**MAX**
Yearly - Last-result latency	12,591	12,871
Yearly - Last-result latency[a]	3,334	3,391
Quarterly - Last-result latency	3,124	3,378
Monthly - Last-result latency	1,198	1,226
Weekly - Last-result latency	342	379
MB per second	**MEAN**	**MAX**
Yearly - Bytes-read per second[b]	1,545	1,573
Yearly - Bytes-read per second[ab]	0	0
Quarterly - Bytes-read per second[b]	1,490	1,536
Monthly - Bytes-read per second[b]	1,292	1,313
Weekly - Bytes-read per second[b]	1,137	1,180

[a] Shows the year-high bid run a second time without clearing the cache.

[b] Bytes read per second from persistent media, according to `iostat`. Cache hits do not count as bytes read.

Write test

Table 5 shows the results for the Basic Data Generation Algorithm for one day of data.

Table 5 Write test benchmarking results

Latency (milliseconds)	MEAN	MAX
Write-completion latency	10,687	11,645

Post-Trade Analytics benchmarks

The Post-Trade Analytics Benchmark tests include a Volume Weighted Average Price (VWAB) benchmark on a single client thread for one day of data, a theoretical profit and loss (P&L) benchmark, and a market snapshot with 10 requesting client threads.

VWAB on one day of data (one client thread requesting)

Table 6 shows the results for the return of approximately a 4-hour volume-weighted bid over a single day for certain 1% of symbols.

Table 6 Write test benchmarking results

Latency (milliseconds)	MEAN	MAX
First-result latency	503	509
Last-result latency	503	509

Theoretical P&L (10 client threads requesting)

Table 7 shows the results for each of 10 client threads querying a unique set of 100 trades, finding the amount of time until 2x, 4x, and 20x the size of each trade was traded in the market, and returning the Volume Weighted Average Price (VWAP) and total volume over those times intervals.

Table 7 Theoretical P&L benchmarking results

Latency (milliseconds)	MEAN	MED	MIN	MAX	STDV
First-result latency	1,418	1,417	1,357	1,474	27
Last-result latency	1,418	1,417	1,357	1,474	27

Market snapshot (10 client threads requesting)

Table 8 shows the results for each of 10 client threads querying a unique date, time, and set of symbols (1% of the total symbols), and returning the price and size information for the latest quote and trade for each symbol.

Table 8 Market snapshot benchmarking results

Latency (milliseconds)	MEAN	MED	MIN	MAX	STDV
First-result latency	1,003	992	924	1,078	45
Last-result latency	1,003	992	924	1,078	45

Research Analytics benchmarks

The Research Analytics benchmark test includes a volume curves benchmark, an aggregated stats benchmark, and a benchmark that measures statistics from requesting client threads that are measured at random intervals.

First result = last result: For kdb,+ the distinction between "time first result arrives" and "time last result arrives" is not relevant because everything arrives at once. This is why the first and last results in the tables in these sections are the same. In STAC terminology, the LAT1 value always equals LAT2.

Volume curves (10 client threads requesting)

Table 9 shows the results for each of 10 client threads querying a unique set of 20 dates and set of symbols (10% of the total symbols) and returning the average proportion of volume that is traded in each minute interval for each symbol across the date set.

Table 9 Volume curve benchmarking results

Latency (milliseconds)	**MEAN**	**MED**	**MIN**	**MAX**	**STDV**
First-result latency	20,723	20,839	17,927	22,435	797
Last-result latency	20,723	20,839	17,927	22,435	797

Aggregated Stats (10 client threads requesting)

Table 10 shows the results for each of 10 client threads querying a unique exchange, date, and start time, and returning basic statistics that are calculated for the entirety of the 100-minute time range following the start time. Time ranges always cross a date boundary.

Table 10 Aggregated Stats benchmarking results

Latency (milliseconds)	**MEAN**	**MED**	**MIN**	**MAX**	**STDV**
First-result latency	60,429	58,836	30,277	100,701	15,677
Last-result latency	60,429	58,836	30,277	100,701	15,677

Statistics over unpredictable intervals (variable client threads requesting)

Table 11 shows the results for each of some client threads querying a unique exchange, date, and start time, and returning basic statistics that are calculated for each minute interval in a 100-minute time range following the start time. Start times are offset from minute boundaries by a random amount. Time ranges always cross a date boundary. Tests must be run with 1, 10, 50, and 100 client threads. Tests with other numbers of client threads
are optional.

Table 11 Statistics over unpredictable intervals benchmarking results

Latency (milliseconds)	MEAN	MED	MIN	MAX	STDV
1 client thread-First-result latency	11,925	11,824	11,716	12,435	259
10 client threads-First-result latency	30,947	31,795	26,147	35,673	2,990
50 client threads-First-result latency	87,048	87,468	21,408	151,486	35,985
100 client threads-First-result latency	147,626	148,109	21,300	277,365	70,838
1 client thread-Last-result latency	11,925	11,824	11,716	12,435	259
10 client threads-Last-result latency	30,947	31,795	26,147	35,673	2,990
50 client threads-Last-result latency	87,048	87,468	21,408	151,486	35,985
100 client threads-Last-result latency	147,626	148,109	21,300	277,365	70,838

NBBO benchmark

The National Best Bid and Offer (NBBO) benchmark test calculates NBBO across all exchanges for all symbols on one day. Table 12 shows the results.

Table 12 NBBO benchmarking results

Latency (milliseconds)	MEAN	MAX
Write-completion latency	146,109	154,686

Multi-day/multi-user VWAB benchmark

The multi-day/multi-user VWAB benchmark tests each of 100 client threads querying unique symbol sets, and returning a 4-hour volume-weighted bid for 12 random days per thread for 1% of symbols per thread. Table 13 shows the results.

Table 13 Multi-day/multi-user VWAB benchmarking results

Latency (milliseconds)	**MEAN**	**MED**	**MIN**	**MAX**	**STDV**
First-result latency	42,001	41,985	6,384	109,193	21,487
Last-result latency	42,001	41,985	6,384	109,193	21,487

Summary

The IBM System x3750 M4 with 8-core Intel Xeon E5-4650 processors and eight 800 GB Intel S3700 800 GB SSDs represents some of the most current hardware technology that is available. This solution is optimized as a cost-competitive solution, balancing high performance and low latency with price.

In the financial markets, the goal is to continuously adopt the latest technology to gain the performance advantage, which also means optimizing the technology as efficiently as possible to gain the most effective solution. Choosing the correct software is key to a good solution. In-memory databases are evolving and updates continue to yield better results. System software is also being enhanced. Newer versions of Linux and Windows provide new optimizations to take advantage of newer hardware. Storage technologies are also changing, and the trend of software-defined storage combines higher performance with lower costs.

Finally, system tuning and optimization is critical for achieving the best performance. In this paper, we described some of the preferred practices of performance tuning as it relates to high performance I/O.

In summary, as newer technologies are released, they will be tested for these types of workloads and this benchmark will be updated to reflect the current offerings from all of the vendors that are involved.

For more information

This benchmark was sponsored by the IBM Wall Street Center of Excellence. For more information about the Center, or to arrange a visit, visit the following website:

http://ibm.com/systems/services/briefingcenter/wscoe/

Additional resources

- *IBM Case Study: Redline Trading Solutions delivers reliable ultra-low latency for financial markets firms*

  http://ibm.com/software/success/cssdb.nsf/CS/STRD-99UF2Z
- Intel Financial Services Industry Community

  http://software.intel.com/en-us/financial-services
- *Introducing uEFI-Compliant Firmware*

  http://ibm.com/support/entry/portal/docdisplay?lndocid=MIGR-5083207
- *Introducing UEFI-compliant firmware addendum for Intel Xeon E5 family of servers*

  http://ibm.com/support/entry/portal/docdisplay?lndocid=MIGR-5089627
- *A Performance Guide For HPC Applications On the IBM System x iDataPlex dx360 M4 System*

  https://www.ibm.com/developerworks/community/wikis/home?lang=en#!/wiki/Welcome%20to%20High%20Performance%20Computing%20%28HPC%29%20Central/page/Performance%20Tips%20and%20Whitepapers

Authors

This paper was produced by a team of specialists from around the world working at the International Technical Support Organization, Raleigh Center.

Joshua Blumert
Dave Weber
Stephen Smith

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an ITSO residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Stay connected to IBM Redbooks

- Find us on Facebook:

  http://www.facebook.com/IBMRedbooks

- Follow us on Twitter:

  http://twitter.com/ibmredbooks

- Look for us on LinkedIn:

  http://www.linkedin.com/groups?home=&gid=2130806

- Explore new IBM Redbooks® publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

  https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm

- Stay current on recent Redbooks publications with RSS Feeds:

  http://www.redbooks.ibm.com/rss.html

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.

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 references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

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 measurements 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.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs.

This document REDP-5029-00 was created or updated on September 25, 2013.

Send us your comments in one of the following ways:

- Use the online **Contact us** review Redbooks form found at:
  ibm.com/redbooks
- Send your comments in an email to:
  redbooks@us.ibm.com
- Mail your comments to:
  IBM Corporation, International Technical Support Organization
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400 U.S.A.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at http://www.ibm.com/legal/copytrade.shtml

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

AIX®
Calibrated Vectored Cooling™
FlashSystem™
GPFS™
IBM®
IBM FlashSystem™
IBM Systems Director Active Energy Manager™
Redbooks®
Redpaper™
Redbooks (logo) ®
Storwize®
System x®
XIV®

The following terms are trademarks of other companies:

"STAC", STAC logos, and all STAC names are trademarks or registered trademarks of the Securities Technology Analysis Center, LLC.

Texas Memory Systems, and the Texas Memory Systems logo are trademarks or registered trademarks of Texas Memory Systems, an IBM Company.

Intel, Intel Xeon, Intel logo, Intel Inside logo, and Intel Centrino logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.