IT@Intel White Paper

Intel IT
IT Best Practices
In-Memory BI Platform
November 2013

Configuring an In-Memory BI Platform for Extreme Performance

Performing server-sizing and stress tests helped us find the best price/performance combination of server speed, number of processor cores, cache size, and memory for industry-standard servers based on the Intel® Xeon® processor family.

Ajay Chandramouly
Big Data Domain Owner, Intel IT

Rajeshkumar Ramamurthy
BI Platform Architect, Intel IT

Rama Sheshadri
System Engineer, Intel IT

Prashanth Uk
Capability Engineer, Intel IT

Chandhu Yalla
BI Engineering Manager/Architecture Owner, Intel IT

Executive Overview

To deliver the extreme query responsiveness required for real-time analysis of high-volume data sets, Intel IT conducted tests to find the optimal platform for a cost-effective, high-performance in-memory business intelligence (BI) solution. Performing server-sizing and stress tests helped us find the best price/performance combination of server speed, number of processor cores, cache size, and memory for industry-standard servers based on the Intel® Xeon® processor family.

In-memory BI solutions provide the enhanced access and response capabilities organizations need to deliver the right information to the right decision makers at the right time. Since in-memory BI solutions differ in many ways, we performed tests to determine the best platform for our selected third-party in-memory analytics application.

By determining the optimal configuration for our in-memory platform, we expect to achieve the following advantages:

- Cost-effective performance in an enterprise-class data warehouse built for our in-memory BI solution
- High business value through a solution that enables Intel business groups to achieve real-time visibility into high-volume data sets, faster time to insight, shorter development times, and new self-service BI opportunities
- The ability to easily scale and replicate our solution for future applications of in-memory BI solutions

The exceptional velocity and subsecond latency of an in-memory database is becoming important to Intel IT's overall multiple data warehouse strategy as a way to deliver more powerful analytics capabilities to business groups across Intel. In the near future, we may deploy in-memory BI solutions to big data use cases such as supply-and-demand planning, near real-time identification of new business opportunities, balance-sheet hedging that potentially saves USD millions in foreign currency translation, and real-time supply-chain risk assessment on more than 400,000 parts and USD 1.6 billion in expenditures.

Contents

IT@INTEL

The IT@Intel program connects IT professionals around the world with their peers inside our organization – sharing lessons learned, methods and strategies. Our goal is simple: Share Intel IT best practices that create business value and make IT a competitive advantage. Visit us today at www.intel.com/IT or contact your local Intel representative if you'd like to learn more.

BACKGROUND

Anticipating the increasing demand for faster access to cost-effective business intelligence (BI) from larger data sets, Intel IT is seeking new ways to enable Intel business groups to analyze a broader range of data more quickly, deeply, and efficiently. Our overarching BI goal is to provide the right data to the right people at the right time. This requires a BI strategy that constantly evolves to accommodate a variety of business use cases, providing actionable insights in near real time to high-value business issues.

In recent years, we have advanced our BI data warehouse solutions to include the following:

- An **enterprise data warehouse** (EDW) for enterprise-wide structured data
- An **Apache Hadoop* solution** for raw, unstructured data
- An **extreme data warehouse** (XDW) for structured and semi-structured data
- **Custom, independent data warehouses** for structured, normalized data
- An **in-memory BI platform** for real-time analytics of streaming high-volume data sets

With our in-memory BI platform, the newest addition to this list, we are seeking a way to help Intel business groups quickly analyze high-volume data sets. Intel business use cases that require such rapid results to meet the specific requirements of high business value can justify the expense of an in-memory BI solution. In these cases, the exceptional velocity may yield the highest business value by enabling faster and better informed business decisions.

Several advantages specific to in-memory databases have increased our interest in in-memory BI analytics:

- Superb performance and extremely low latency compared to standard database management systems, particularly those bogged down by I/O bottlenecks
- Fast response for short queries against large data volumes
- Ability to offload processing from expensive database appliances
- Ability to integrate data from different sources and eliminate or reduce the time spent on performance-tuning tasks such as query analysis, cube building, and aggregate table design
- Easy deployment for self-service analytics, providing intuitive and unconstrained data exploration
- Instant visualization capabilities for complex data sets

Overall, as the velocity and complexity of Intel's business accelerates, in-memory BI platforms will become an important part of our multiple data warehouse strategy. It will enable our business groups to dramatically improve the speed and quality of their decision making to help Intel better compete in global markets.

Potential Return on Investment Benefits from In-Memory BI Solutions

Intel IT's investigation into in-memory business intelligence (BI) solutions began with the recognition of the potential return on investment from the advantages these solutions can deliver to our business groups.

- **Faster time to insight.** With in-memory BI solutions, data is loaded into memory where calculations can be performed by the processor cores much faster than pulling from a hard disk.
- **Real-time visibility.** Traditional BI systems push data from the sources to the data warehouse. In-memory BI solutions provide real-time data replication from enterprise resource planning and similar applications, enabling users to analyze business operations in real time to gain insight.
- **Efficient.** In-memory BI solutions allow business users to handle significantly higher volumes of data faster than traditional analytics tools through the use of columnar compression. Storing information in columns instead of rows enables a database management system structure that provides substantial levels of data compression (up to 10 times), making it more feasible to load large data sets into memory.
- **Agile.** By eliminating the need to aggregate data and build schemas, in-memory BI solutions promote a more agile analytical process that better adapts to changing business requirements compared to traditional BI.
- **Self-service.** The need to build aggregated and pre-calculated data structures limits how a user can explore data. In-memory BI solutions foster self-service by providing greater analytic flexibility without the administrative overhead of index creation and maintenance, thus lessening business user reliance on IT.
- **Improved development time.** Loading detailed data into memory for reporting and analysis reduces the need to build aggregate data structures—a key part of most BI deployments. This makes in-memory BI solutions faster to set up and deploy.

BUSINESS CHALLENGE

While the benefits of in-memory BI solutions are well established, Intel IT found in its early stages of investigation that there is still much to learn about how to design the optimal platform for cost-efficient performance using industry-standard servers.

In-memory computing has been around since the late 1990s and is typically used for transaction and event processing. The high cost of RAM and limited scalability initially hindered widespread development and adoption of in-memory BI solutions. The continual drop of RAM prices and the increasing prevalence of 64-bit OSs currently make in-memory analytics more affordable and scalable. Whereas older 32-bit OSs offered only 4 GB of addressable memory, 64-bit OSs support up to 1 TB of memory, enabling terabyte data sets to be stored in RAM. This capacity increase makes RAM the new "disk" and enables many organizations to load data marts, perhaps even an entire data warehouse, into RAM on a 64-bit OS with 1 TB of addressable memory.

Another factor increasing the viability of in-memory BI solutions is the availability of reasonably priced, high-performance industry-standard blade servers based on the Intel® Xeon® processor family equipped with 1 TB of RAM. Such servers work well for analytical processing and can be easily combined to produce powerful multinode systems. Their multicore processors enable fast, efficient parallel processing of in-memory data.

We have found that the growing popularity of columnar databases as alternatives to conventional row-based relational databases is another factor driving in-memory analytics into the mainstream. Storing information in columns instead of rows enables substantial levels of data compression, making it more feasible to load large data sets into memory.

SOLUTION

Intel IT understands the value of in-memory analytics to Intel's business groups and BI efforts. This factor helped in our decision to analyze server size to determine the optimal platform for cost-effective performance with our selected in-memory analytics application. Our goal was to find the best combination of server speed, core number, memory size, and cache size for this application.

We began our analysis by researching the different categories of in-memory BI tools available for a particular use case, what each tool is best suited for, and its optimal hardware requirements. While the selection of an in-memory analytics application is outside the scope of this paper, our test procedures and results can provide guidance on a platform-sizing strategy for similar applications.

How In-Memory BI Solutions Work

Conventional business intelligence (BI) tools query data on hard drives. Every time a user runs a query with a typical BI data warehouse, the query goes to a database that reads the information from multiple tables stored on server hard drives.

In-memory BI solutions query data in dynamic RAM. This means when a user queries a source database, all its information is loaded into memory for that initial query and all subsequent queries in the session. By looking at data sets held entirely within memory, in-memory tools eliminate repetitive processing and reduce the burden on database servers. By not pulling information from hard drives, in-memory tools can theoretically access data 10,000 to one million times faster and can dramatically reduce I/O cycles and calculation times, greatly improving query response times.

To cost-effectively hold all this data in memory, in-memory BI solutions use techniques such as columnar compression to store the data in highly compressed formats. In-memory BI solutions can achieve as much as a 1:10 data-volume ratio in comparison to traditional on-disk storage.

For disk-based solutions, IT organizations typically must design and build a data layer optimized for query performance. In-memory BI solutions simplify the data analysis process by using a virtual layer to access the data, eliminating the need to build the indexes, aggregated tables, and multidimensional cubes traditionally required by data warehousing and BI applications.

In-memory BI solutions range from spreadsheet-based applications to high-end platforms designed to handle immense amounts of data. In-memory tools ease the data management workload for IT teams and make it easier to develop and run queries. These tools also create opportunities for self-service BI capabilities that open up more analytics activities to end users lacking specialized skills.

From the small number of software-based products available, we chose an in-memory analytics application that facilitated testing on industry-standard servers. This was an important decision as the software's performance demands can be more influential than a particular use case in determining the server-sizing test results.

Test Use Case

The data we used for testing came from an Intel IT BI solution that provides customer insight to our Sales and Marketing group. The customer insight from this BI solution takes advantage of being able to access multiple data containers and integrate several big data technologies to implement a set of business rules for web analytics. Our solution processes raw web data and then integrates this data with traditional transactional internal consumption data available in a different data warehouse.

The ultimate goal is to enable faster decision making on marketing campaigns and approaches for our products and services. To do that, our use case must rapidly analyze data and provide the following insights:

- Identify customer needs, wants, behaviors, and expectations
- Accurately predict which customers are likely to need a specific product or service
- Identify the sales cycle and process (such as viewing a webinar, reading a white paper, or downloading a reference architecture) that lead to a customer purchase

The BI process begins with collecting customer and network usage analytics from visits to Intel.com. This web usage data is then analyzed for marketing and content navigation purposes to better personalize marketing outreach to a customer in the early stages of the sales cycle. In the past, the quantity and variability of the data sources hindered our ability to analyze data fast enough to take meaningful action and make real-time adjustments on product positioning and customer interaction. The goal in using in-memory analytics for this use case was to gain the ability to analyze this data in real time to correctly predict and adjust product positioning or pricing based on the current response to marketing campaigns.

This use case is particularly applicable to in-memory BI analytics because of the growing demand from our business groups for self-service BI and the fast processing of real-time consumption data. Our business users need to be able to view large volumes of data—typically around 7 GB—from different perspectives and drill down through different levels to get the information they need to make well-informed decisions.

Test Objective

The goal of performing server-sizing and stress tests was to find the optimal server configuration that provides a performance level that justifies the increased cost of the server hardware. In our testing, we sought answers to the following questions:

- Will more and faster processor cores deliver better performance for our in-memory BI solution?
- Will larger processor cache increase performance for our in-memory BI solution?
- Will other server or usage aspects, such as server virtualization or multiple concurrent users, affect performance?

Methodology

We identified four key criteria in determining the appropriate platform configuration and sizing for our tabular-based in-memory BI solution:

- **Server speed.** We tested processors with core speeds ranging from 2.4 GHz to 2.93 GHz to determine the effect core speed had on our in-memory BI solution.
- **Number of cores.** Having multiple cores available enables performance options. The number of cores can be used to scale a platform to support more users or to reduce processing time.
- **Cache.** Cache is an important test parameter because as data loads, it is copied to physical memory (RAM), and

then, as the query is executed, data is copied to the processor cache multiple times for decompression (depending on the operations that need to be performed). The larger the processor cache, the better the query performance and the fewer the translation-lookaside-buffer (TLB) lookups.

- **RAM size in relation to cores.** Since in-memory BI solutions require databases to reside in RAM, the larger the RAM, the larger the database that can be handled in RAM. A key consideration here is that performing in-memory analytics on a database in RAM requires additional RAM for the processing (cubes/queries). For our use case, the total RAM had to be sized at least 2.5 to 3 times the size of the in-memory database.

Other considerations include physical disks and storage. While storage is not a consideration when processing queries with an in-memory BI solution, it is important in loading the database into memory, server restores, and pagination (storing and retrieving data from secondary storage when used). For high availability and throughput, we used a storage area network (SAN).

Test Types: Platform and Capacity

We performed two types of tests using four server configurations ranging from 8-32 cores (see Table 1). We used **platform tests** to identify the appropriate hardware platform for our tabular-based in-memory analytics solution for single and multiple concurrent users. We used **capacity tests** to determine the appropriate server sizing using different query combinations and large numbers of concurrent sessions. In both cases, we used the actual data and database from our worldwide consumption use case.

PLATFORM TESTING

Our in-memory BI software uses a tabular structure that may hold data or metadata or both for slicing and dicing. This structure includes dimension tables and fact tables (measures). The dimensions include master data relating to employees and geographic location. The measures are numerical data that when aggregated provide insight in reporting or in quantifying items.

For our study, we looked at two query types:

- **Memory-centric queries.** These require reading data mostly from physical memory and minimal processor-intensive computations.
- **Processor-centric queries.** These require extensive processing, including compression and decompression, and thus can greatly benefit from processor cache and additional cores.

Query type provides an important distinction when selecting a processor. With processor-centric queries, all processed data goes through the processor's cache. The smaller the cache or slower the processor, the more time it takes to process the query. The greatest effect on cache size comes with the use of virtual servers where unless the cache is partitioned, there is no control over how the processor's cache is used among the virtual servers sharing the same processor.

For our platform tests, we compared the performance of different platforms executing both types of queries from a single user, from multiple concurrent users (5–10 parallel sessions), as well as with and without data in processor cache.

CAPACITY TESTING

Database performance generally follows the 80/20 rule, where 20 percent of the queries cause 80 percent of performance issues. With this in mind, to test capacity we categorized queries as simple, medium, and complex based on query execution time as follows:

- **Simple** – queries executing in 1-3 seconds
- **Medium** – queries executing in 4-6 seconds
- **Complex** – queries executing in 8-10+ seconds

With this selection of queries, we performed capacity tests across the platforms using different query combinations and concurrencies, simulating different numbers of users making queries at the same time. Using query creation and server stress tools to execute queries in parallel, we performed tests running three loads:

- 1,000 simple queries
- 250 simple, 125 medium, and 125 complex queries
- 200 simple, 50 medium, 50 complex queries

In addition, we used a concurrent load that assumed 1,000 active users and their probable composing time so that at any one time 10 percent, or 100 of them, would be active in concurrent sessions.

Table 1. Test Server Configurations

Server	Core Speed (GHz)	Total Cores	L2/L3 Cache Size per Processor (MB)	Server Memory (GB)
Server 1 – Development Virtual server running on hosts with Intel® Xeon® processor X5670 series	2.93	16 virtual cores	12	32
Server 2 – Production Four-socket server equipped with Intel® Xeon® processor E7330	2.4	16	6	192
Server 3 Four-socket server equipped with Intel® Xeon® processor E5-2650 with Intel® Turbo Boost Technology[1] 2.0	2.7	32	20	96
Server 4 Virtual server running on hosts with Intel® Xeon® processor X5670 series	2.93	8 virtual cores	12	12

Table 2. Optimal Single-Node Configuration. The optimal single-node configuration in our testing was a four-socket server equipped with the Intel® Xeon® processor E5-2650 with Intel® Turbo Boost Technology[1] 2.0.

Category and Parameter	Value
Processor	
Cores	32
Speed	2.7 GHz or faster
L2/L3 Cache	20 MB or greater
Memory	
Size	96 GB
Physical Disk	
Size	500–700 GB
Type	Storage area network
RAID	5 or 10
Server	
Type	Physical

RESULTS

We obtained optimal results with the 32-core configuration running 2.7 GHz cores and 20 MB cache. This configuration had the greatest number of cores, nearly the fastest cores, and the largest cache. Applying this platform to the worldwide consumption database used by our Sales and Marketing group, we successfully reduced query execution time by over 50 percent for certain processor-intensive queries, compared to our former BI analytics solution. In addition, the ease of development and faster development time of an in-memory BI solution helped us reduce the time to market for this consumption use case by a factor of 5.

Optimal Configuration

Projections based on our test results indicate that for our in-memory BI solution for the worldwide consumption use case, the optimal server sizing for a single node in a cluster is the configuration described in Table 2.

Advantage of More Cores

In one test, we performed processing first with a server equipped with eight cores. We then added another eight cores for a total of 16 and repeated the same tests.

In our use case, the results for the initial storage load demonstrated that doubling the number of cores from 8 to 16 reduced the processing time for pulling the data from the source and loading it into memory for some query types by over 50 percent, as illustrated by the green line in Figure 1. This task required compression into data cubes. We decreased the overall time to process approximately 100 GB of data and compress it to 10 GB from 3.5 hours to 1.5 hours.

In addition, processing time on queries that require decompression—and are therefore also processor-intensive—was reduced by over 50 percent in some cases when the number of cores doubled from 8 to 16, as illustrated by the blue and black bars in Figure 1. This is a significant runtime improvement for some query types.

An additional benefit of adding cores is that the cores enable scaling the platform to support more users. Our tests demonstrated that increasing cores enables performance to scale for multiple users performing concurrent queries. The largest performance improvement came from faster compression and decompression during processing and greater concurrent query loads.

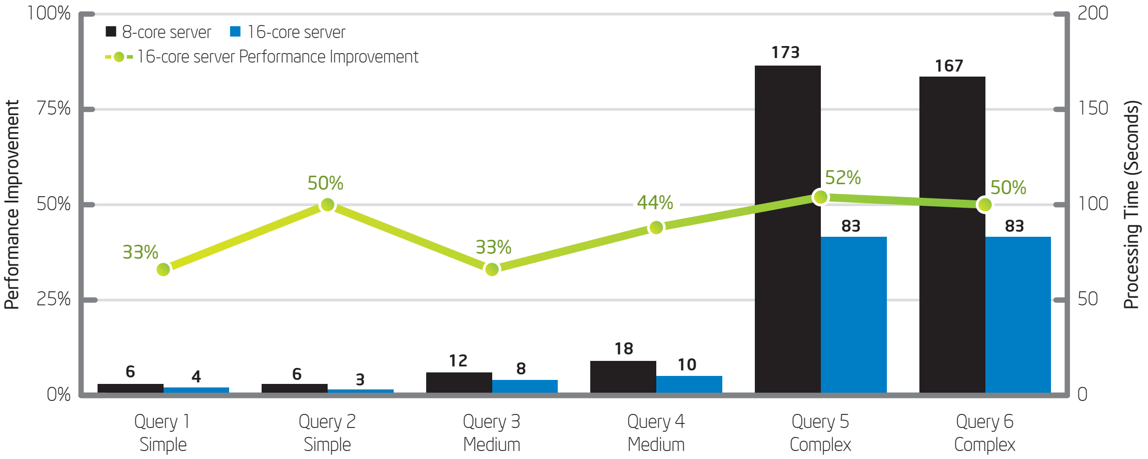

Figure 1. Improvement in processing speed with additional cores. As illustrated by the green line, in queries requiring decompression as part of the processing, doubling the number of cores delivered a greater than 50 percent performance gain in some cases. As illustrated by the blue and black bars, doubling the number of cores reduced processing time by over 50 percent for some processor-intensive queries that require decompression.

Advantage of Faster Cores

Our particular in-memory BI solution is a single-threaded formula engine. This means that to scale for more concurrent users, we needed more cores with faster core speeds. Faster cores help us increase the speed of data processing—encoding, compression, and decompression—in a linear fashion.

In our testing, we compared a node with 16 virtual cores running at 2.93 GHz (Server 1, blue bars) to a node with 16 cores running at 2.4 GHz (Server 2) for both single-user (black bars) and multiple concurrent user (gray bars) queries. We recorded more than double the performance for single-user and multiple concurrent user queries with the 2.93-GHz cores (see Figure 2).

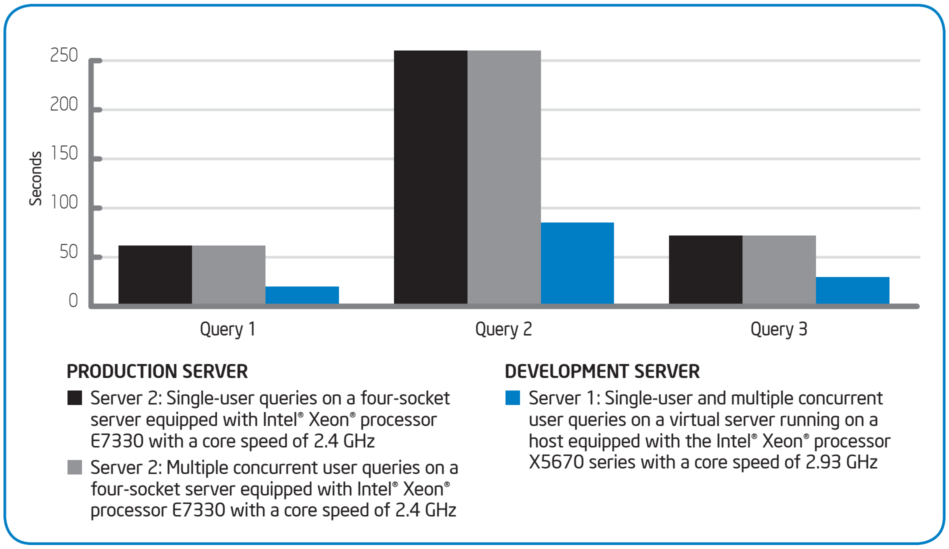

Figure 2. Single-user and multiple concurrent user queries on production servers versus a virtual development server. Query performance more than doubled across all three queries when using server platforms with cores running at 2.93 GHz.

Advantage of Larger Cache

Processors look first for data in processor cache before reading data from RAM. The larger the cache and the more data already loaded and available in it, the faster the query performance, particularly in regard to Server 3 (see Figure 3).

We began to see the advantage of a larger cache erode in situations where a node is a virtual server sharing a host with 19 other virtual servers with no controls placed on cache usage. In these cases, a 16-core virtual server sharing 12 MB cache with 19 other virtual servers (Server 1) has little advantage over an 8-core virtual server sharing 12 MB cache with 19 other virtual servers (Server 4) in our multiple concurrent user tests (see Figure 3).

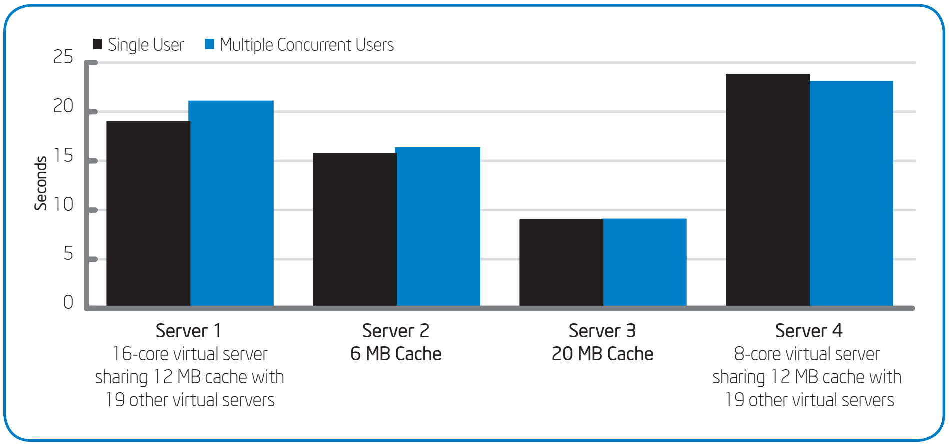

Figure 3. Advantage of larger cache. Our testing showed that the advantage of a larger cache decreases when a virtual server shares a host with 19 other virtual servers. In multiple user tests, the performance increase of Server 1 over Server 4 is negligible.

CONCLUSION

The results of our server-sizing and stress tests demonstrate the importance of properly configuring servers for use in an in-memory BI platform to achieve optimal results. Based on our tests, we found that increasing core number per node, core speed, and cache size all affect an in-memory BI platform's performance in handling heterogeneous queries consisting of memory-intensive and processor-intensive queries.

Our results indicate an extreme performance platform configuration would include the following:

- Four-socket motherboard configured with 32 high-speed cores
- High quantity of L2/L3 cache per processor
- Dedicated cache when sharing a host to avoid cache competition
- Server memory 2.5 to 3 times the size of the compressed database

Core number was particularly important on processor-intensive queries that required decompression, providing up to 52 percent faster performance when we doubled the cores. High core number also helped maintain performance with multiple concurrent users. Faster cores more than doubled the performance in both single and multiple concurrent user tests, demonstrating their importance in helping provide near real-time results.

Using industry-standard servers will enable us to easily scale and replicate our solution for additional applications of in-memory BI solutions. Based on our results, we plan to continue to analyze the business, financial, and BI benefits of in-memory BI solutions on a case-by-case basis as part of our multicontainer strategy for better BI analytics. As memory prices continue to drop and more in-memory BI software products become available, we see increasing reliance on in-memory BI solutions helping enable Intel business groups to achieve real-time visibility into high-volume data sets, faster time to insight, shorter development times, and new self-service BI opportunities.

FOR MORE INFORMATION

Visit www.intel.com/IT to find white papers on related topics:

- "Integrating Apache Hadoop* into Intel's Big Data Environment"
- "Using a Multiple Data Warehouse Strategy to Improve BI Analytics"

ACRONYMS

EDW	enterprise data warehouse
ROI	return on investment
SAN	storage area network
XDW	extreme data warehouse

For more information on Intel IT best practices, visit www.intel.com/IT.

[1] Requires a system with Intel® Turbo Boost Technology. Intel Turbo Boost Technology and Intel Turbo Boost Technology 2.0 are only available on select Intel® processors. Consult your PC manufacturer. Performance varies depending on hardware, software, and system configuration. For more information, visit www.intel.com/go/turbo

Intel processor numbers are not a measure of performance. Processor numbers differentiate features within each processor family, not across different processor families. To learn about Intel Processor Numbers, go to www.intel.com/content/www/us/en/processors/processor-numbers.html

Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing. For more information on performance tests and on the performance of Intel products, reference www.intel.com/performance/resources/benchmark_limitations.htm or call (U.S.) 1-800-628-8686 or 1-916-356-3104.

THE INFORMATION PROVIDED IN THIS PAPER IS INTENDED TO BE GENERAL IN NATURE AND IS NOT SPECIFIC GUIDANCE. RECOMMENDATIONS (INCLUDING POTENTIAL COST SAVINGS) ARE BASED UPON INTEL'S EXPERIENCE AND ARE ESTIMATES ONLY. INTEL DOES NOT GUARANTEE OR WARRANT OTHERS WILL OBTAIN SIMILAR RESULTS.

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

Intel, Intel logo, and Intel Xeon are trademarks of Intel Corporation in the U.S. and other countries.

*Other names and brands may be claimed as the property of others.

Printed in USA ♻ Please Recycle 1113/ACHA/KC/PDF 329337-001US