WHITE PAPER

In-Memory Computing: Now and Tomorrow

A GridGain Systems In-Memory Computing White Paper

August 2017

Contents

As businesses everywhere struggle to keep up with – and strategically leverage – the massive amounts of data driving their organizations, more and more of them are turning to in-memory computing (IMC). IMC provides a way to store and process data in RAM that is distributed across clusters of computers. This high-performance architecture is rapidly becoming the method of choice for today's big-data and fast-data applications.

The in-memory data revolution is far from over, as demand continues to spur further innovation in in-memory computing technologies. This white paper will examine the factors contributing to increased demand, the innovative directions in which the industry is heading over the next decade, and how the Apache® Ignite™ and the GridGain in-memory computing platforms are leading the way toward the in-memory computing systems of tomorrow.

Why Demand for In-Memory Computing Keeps Growing

Three key reasons are behind the growing movement toward in-memory computing: an explosion of data, heightened performance expectations, and in-memory data technology that keeps getting both better and more affordable.

An explosion of data. With business and personal activities increasingly conducted online using digital wallets, social media, and other apps, the data generated from these activities is growing rapidly. Adding to this glut is the onslaught of data from a growing number of smart devices connected to the Internet of Things (IoT).

Many of these IoT devices are still on the periphery of consumer awareness, but their impact over the next decade will be enormous. Some seem like niche products, such as the smart basketball – containing a battery and sensors sending data to a smartphone app that can analyze player and ball movements to determine why a player might be missing shots. Others are more mainstream, such as the Samsung Family Hub refrigerator, which tracks food expiration dates and lets owners view fridge contents from their smartphones while at the store, alleviating the need to wonder, "Am I out of orange juice?" One such device on its own may send relatively small packets of data. However, as smart devices proliferate, with many of them collecting data nonstop over long time periods (in the manner that Fitbits track people's fitness), the aggregate amount of data will be very large.

Connected vehicles, in particular, are poised to greatly increase the amount of data available about travelers and their circumstances. Connected cars can potentially receive and transmit significant amounts of data about minute-by-minute traffic, as well as road conditions and the car's internal status. It's estimated that each connected car sends between two and four TB of data to the cloud every day. For example, Volvo connected cars send data to the Swedish Transport Authority when the anti-lock brake system is deployed at low temperatures, alerting authorities to frozen roads that need attention. Current Formula One cars have about 300 sensors. A train manufacturer seeking to improve braking efficiency is employing about 400 sensors per train – a number likely to increase to 4,000 over the next five years. With hundreds or thousands of sensors in a vehicle taking readings every millisecond, the output quickly adds up to a massive volume of data.

Ingesting and analyzing all the data from apps and devices can be overwhelming – like trying to drink water from a fire hose. The data industry is racing to keep up with these data volumes, even as it must deal with another daunting challenge: processing the data quickly enough to keep up with today's expectations of real-time response.

Heightened performance expectations. A recent *Computer Business Review Magazine* survey of 200 UK-based CEOs found that ninety percent of the CEOs believed there is a significant gap between the performance level that consumers expect and what IT can deliver. These results echo similar concerns among CEOs in the US.

Consumer expectations of speedy performance are rising and driving business expectations similarly upward. Consumers get very frustrated by applications with poor performance. If a webpage fails to load in three or four seconds, they will abandon a shopping cart and go to a different website. As these users become accustomed to real-time response in consumer apps such as social media and online gaming, they expect the business technologies they use within enterprise and public-sector organizations to deliver the same level of performance. They expect real-time response – once available only for mission-critical applications – and, too often, businesses fall short of these expectations.

To address this performance gap, many organizations are turning to strategies that involve in-memory technology, which is showing up in a wide variety of data platforms. Operational databases, which started off mostly as disk-persisted, have increasingly added in-memory options in the last few years to speed up transaction processing and support analytical insight. Platforms for event and stream processing are using in-memory technology to rapidly ingest, analyze, and filter data on the fly, before they send the data elsewhere for further analysis later on. Many analytic databases and data warehouses rely on in-memory technology in order to quickly run complicated queries on large data sets. In addition, some companies are using in-memory technology to sit on top of Hadoop (which is more batch-oriented than interactive), so that they can access data from their data lakes (made up of Hadoop clusters) more rapidly.

A particularly important strategy for addressing the big data performance gap is the use of in-memory data grids. Data grids – and more comprehensive in-memory computing platforms that include data grids, such as Apache® Ignite™ and its enterprise version, GridGain – are the second fastest-growing segment of the data industry, after Hadoop. In-memory data grids provide a convenient way to speed up an application's transactions without having to migrate or extensively rework the underlying database. They accomplish this feat by sliding in between existing applications and databases to provide parallel processing of data cached in the RAM of many clustered computers. Clients who have implemented the GridGain in-memory computing platform have found that they can process transactions about 1,000 times faster than they could previously.

Today's data grids and other in-memory options are not only highly useful for providing a performance edge in big-data operations, they are also competitive in another key area: affordability.

Better, cheaper IMC technology. As big data and expectations of real-time data access push businesses toward in-memory data solutions, the decreasing cost and continuing innovation of in-memory technology has made it an increasingly attractive option.

The cost of memory has dropped roughly 30% per year since the 1960s, making it much more affordable. That affordability has flipped the storage paradigm of the past fifty years upside down: instead of a disk-first architecture, with memory used more sparingly to cache small amounts of data for fast access, the data industry is evolving toward a "memory first, disk second" paradigm. This new memory-centric scenario uses memory across many computers in a cluster as the main data system, with slower disk storage used primarily for backup.

As this paradigm shifts, in-memory technology is also evolving and becoming more capable. Over the past few years, in-memory data grids have transitioned from being a simple cache between an application and database to much more capable and comprehensive data platforms in their own right.

The transition of in-memory data grids into mainstream data platforms is just one of several exciting new changes on the horizon for in-memory computing.

What's Next for In-Memory Computing

Over the next decade, in-memory computing will see several transformative developments, including the following:

- Comprehensive IMC platforms becoming systems of record
- First class support for in-memory SQL
- Growing artificial intelligence (AI) use cases for in-memory computing
- Non-volatile memory (NVM) as the preferred storage method
- Hybrid storage models for very large datasets

Let's look at these developments in more detail.

IMC as a system of record. The increased affordability and capabilities of in-memory technology, as mentioned earlier, have made memory-centric architectures the dominant paradigm for today's and tomorrow's data systems. Many businesses (including over 50% of GridGain's clients) use in-memory computing platforms as their systems of record, or the authoritative data source for business-critical records, as they would a database. They are using disk storage or previously existing databases primarily as backup devices.

Contributing to this development is the ability of more and more in-memory data platforms to perform hybrid transactional and analytical processing (HTAP). These platforms, which include Apache Ignite and GridGain, can handle operations and transactions as well as analytics and data warehousing – so there is no need for a business to have separate databases for operational and analytical purposes.

Another factor driving organizations to adopt IMC platforms as their main databases is increasingly robust support for SQL.

First class support for in-memory SQL. About ten years ago, many in the data-processing community thought SQL was dead. With new databases offering a variety of ways to store and query data, many were looking toward NoSQL or talking about developing project-specific query languages.

In recent years, however, the community has recognized the value of SQL and embraced it once again. Not only is SQL the number one paradigm for working with data today, with millions of people familiar with its use, it also offers some important advantages with respect to search performance and querying capabilities.

The ability to index data in SQL provides a significant advantage with respect to searching. When there is no key to the data, without SQL or indexes, searching for a particular piece of data requires a full scan – which is significantly slower.

Also, as use cases of in-memory computing become more and more complex, users need robust querying capabilities, such as complex aggregation functions and joins. No one wants to struggle through a learning curve or complicated setups to create queries in a proprietary language that requires just one line of code in SQL.

For these reasons, first class SQL support (in addition to NoSQL, streaming, and other data-access mechanisms) is a necessity for all serious IMC vendors and projects. Apache Ignite and GridGain have a very strong level of support through the ability to perform distributed processing of ANSI SQL-99 compatible SQL across data stored in memory and on disk. Other popular projects, including Spark and Hadoop, have led to the growth of a whole cottage industry of SQL support. IMC vendors who do not yet offer SQL support are likely to get on board soon, to better handle the new, more complex use cases that are becoming prevalent.

Growing artificial intelligence use cases for IMC. Among the exciting new use cases for in-memory computing are the growing number of artificial intelligence projects involving machine learning (ML) – that is, computers learning from data, rather than through programmed instructions – and deep learning (DL), a type of ML that uses multi-layered neural networks. While ML has been around in some form for over 50 years, only recently has the technology to apply it at high speed to very large data sets become relatively accessible.

Machine learning on small, dense data sets – typically data sets that fit within the storage of a single computer – is a problem that has essentially been solved. There are libraries in the Python, Java, and Scala community that address this problem, as well as projects such as Spark that solve it very efficiently within one computer’s memory. However, applying the statistical algorithms of machine learning and deep learning on large, sparse data sets is a much more complex challenge.

Machine learning and deep learning on large, sparse data sets requires a data management system that can store terabytes of data and perform fast parallel computations. These are tasks for which in-memory computing platforms such as Apache Ignite and GridGain are ideally suited. By storing data in RAM (which is much faster than disk storage) across clusters of computers, and by sending computations directly to the data for parallel processing, these IMC platforms can process the data with the maximum speed and performance allowed by the hardware.

Extensions are currently in development for Apache Ignite and GridGain to provide additional support for in-memory machine learning and deep learning. These tools will be crucial for applying ML and DL algorithms to tens of terabytes of streaming data in the many "fast learning" use cases of the future.

Non-volatile memory (NVM) as the preferred storage method. As in-memory technology continues to evolve, the most important new development is the emergence of NVM, which will have an enormous impact on data storage over the next decade.

The idea of NVM is very simple. DRAM (Dynamic Random Access Memory), the type of RAM currently used in computer systems, loses all of the in-memory data if the power goes out. NVM fixes this problem, retaining its data during a power loss. This development removes the need for similar fault-tolerance on a software level.

Early versions of non-volatile memory were a simple and clunky combination of normal DRAM, flash memory, and a battery to power the copying of compressed data from RAM into flash when the system lost power. Today, we have much more advanced technology in the form of 3D XPoint from Intel and Micron. 3D XPoint features a transistor-free, bit-addressable architecture based on a stackable, cross-gridded data-access array that its makers liken to a three-dimensional checkerboard. This technology can be used both for flash or flash-like devices and for RAM that inserts into the deep sockets in a computer's motherboard.

Industry experts predict increasing adoption of the NVM technology demonstrated in 3D XPoint over the next three to five years. A decade from now, it will probably be the prevalent model of storing data on computer systems. Disk and flash systems might be relegated to the types of backup scenarios for which tape drives are used today. The industry is waiting with bated breath for the development of a supporting infrastructure for NVM technology: firmware, controllers, and system-level software.

As the necessary infrastructure emerges, the adoption of non-volatile memory will bring fundamental changes to in-memory computing. No longer will in-memory computing systems require disk-based storage for fault tolerance. Instead, with NVM retaining data through power outages, IMC systems can be 100% in-memory, entirely composed of memory-based data storage.

To adapt to this major shift in data strategies, forward-looking providers of IMC platforms are making NVM support a high priority. GridGain currently supports NVM within a hybrid storage model that can easily scale up and evolve with the changing storage paradigm.

Hybrid storage models for very large datasets. Once a far-off goal, the idea of storing large databases entirely in memory has become increasingly attainable, but what are the practical limitations? How large can such databases be? Can an in-memory solution support petabytes of data?

Three or four years ago, such a scenario would have seemed out of reach, due to the cost of RAM and the impractical cluster size that would be needed (dozens of computers per cluster). However, technology keeps evolving, and with products such as the Fujitsu M10 server – with 64 terabytes of memory – is raising the bar for memory per single server.

Of course, this is a very expensive server, and 32 of them (adding up to well over a million dollars) would be required to make up a petabyte of combined memory – an amount which, in any case, is more than almost any organization needs today. Still, with a 32-terabyte server now shipping, and with more and more businesses using in-memory computing platforms as their main database or system of record, it makes sense to think ahead to how to support tens of terabytes of data – or even a petabyte, in the future.

Buying many terabytes of DRAM is not a viable option for most businesses. Therefore, the developers of Apache Ignite and GridGain have come up with a versatile solution to the problem of extremely large data sets: a hybrid storage model with a universal interface to all storage media (RAM, flash, or disk, plus NVM in the future). Users do not need to know details of where their data is stored. They can use the same full, unified API (including both SQL and key-value access) to address the data. Meanwhile, in the background (if using the optional Persistent Store feature) Apache Ignite or GridGain will move data to the fastest layer available. Once RAM is saturated, data will be pushed into flash memory, and all the data will live on disk.

This hybrid storage model gives businesses the flexibility to easily adjust storage strategy and capacity – as well as data-processing performance – without changing data-access mechanisms. To increase the speed of systems, users simply add RAM to their clusters. To add capacity at a lower cost, they can add flash memory or disk space.

While this hybrid storage model may sound simple, developing it took years due to the architectural differences between RAM, flash, and disk. Most of the algorithms within Apache Ignite and GridGain have been optimized for RAM, which is byte-addressable, whereas flash and disk are block-addressable. To read one byte of data from disk or flash, the entire block of data must be read. This requires extra time and leads to a large difference in latency compared to RAM. Accessing a byte of data in RAM takes nanoseconds, compared to milliseconds with flash and disk. Reconciling the latency differences among these storage media was a significant challenge – which explains why very few in-memory computing solutions have successfully addressed this problem, apart from Apache Ignite and GridGain.

Businesses looking toward petabytes of data in the future can prepare for that future today by choosing an in-memory solution with a hybrid storage model. The hybrid storage feature in Apache Ignite and GridGain is available now, and developers of other in-memory solutions are working on similar features because they know that hybrid storage is the logical next wave for in-memory computing.

GridGain: The IMC Platform for Today and Tomorrow

The GridGain in-memory computing platform, the enterprise version of Apache Ignite, uniquely addresses both today's and tomorrow's big-data and fast-data needs. Built on a memory-centric architecture, the solution solves today's needs with industry-leading in-memory SQL support, the ability to be used as the system of record, and a track record of more than half of GridGain customers using it to accelerate their main, high-performance database. Looking toward tomorrow, GridGain addresses the transformative developments expected in in-memory computing over the next decade, with NVM support and extensions for machine learning and deep learning currently in development as part of Apache Ignite. Plus, the optional hybrid storage architecture available in Apache Ignite and GridGain provides a flexible framework for easily adjusting a system's mix of RAM, flash, and disk to support petabytes of data in the future.

As a complete in-memory computing platform, GridGain users can consolidate on a single high-performance, highly scalable solution for transactions and analytics, providing a lower TCO. Advanced SQL functionality and API-based support for common programming languages enable rapid deployment. These features, along with the rapidly decreasing cost of memory, boost ROI for in-memory computing initiatives, enabling companies to build less expensive systems that perform thousands of times better.

The GridGain in-memory computing platform offers a comprehensive feature set, including the following:

- **A unified high-performance architecture.** The GridGain in-memory computing platform consists of multiple features connected by a clustered, in-memory file system. The In-memory Data Grid, Compute Grid, Distributed SQL, Streaming Grid and Service Grid are interconnected. Computations occur as close as possible to the data. Additional features such as high throughput, low latency, load balancing, caching, in-memory indexing, streaming, Hadoop acceleration and other performance improvements are crucial to success in real-time modeling, processing, and analytics.
- **Scalability**. The GridGain platform is massively scalable, allowing companies to add cluster nodes and memory in real-time with automatic data rebalancing. Because it is hardware-agnostic, users can choose their own preferred hardware for scaling up.
- **Full SQL support**. GridGain is ANSI SQL-99 compliant and the Distributed SQL capabilities support DML and DDL, so users can leverage their existing SQL code using the GridGain JDBC and ODBC APIs. Users with existing code bases that are not based on SQL can leverage their existing code through supported APIs for Java, .NET, C++, and more.
- **A Distributed ACID and ANSI-99 SQL-Compliant Disk Store.** The optional GridGain Persistent Store is a distributed ACID and ANSI-99 SQL-compliant disk store available in Apache Ignite. It may be deployed on spinning disks, solid state drives (SSDs), Flash, 3D XPoint or other similar storage technologies. If used, the optional Persistent Store keeps the superset of data and all the SQL indexes on disk, which allows GridGain to be fully operational from disk. The combination of this new feature and the platform's advanced SQL capabilities allows GridGain to serve as a distributed transactional SQL database, spanning both memory and disk, while continuing to support all of the traditional in-memory only use cases. Persistent Store allows organizations to maximize their return on investment by defining their own optimal tradeoff between infrastructure costs and application performance by adjusting the amount of data that is kept in-memory.
- **High availability**. The GridGain in-memory computing platform provides essential high-availability features such as data-center replication, automatic failover, fault tolerance, and quick recovery on an enterprise-level scale.

- **Transaction processing**. The GridGain platform supports ACID-compliant transactions in a number of user-configurable modes.
- **Security features**. GridGain supports authentication, authorization, multiple encryption levels, tracing, and auditing.
- **Open Source framework.** GridGain is based on Apache® Ignite™, a popular Apache Software Foundation open source project. GridGain Systems was the original creator of the code contributed to the Apache Software Foundation that became Apache Ignite and fully supports the technology behind Apache Ignite. The GridGain Enterprise Edition extends the features in Apache Ignite to provide enterprise-level capabilities and services, such as additional security, data center replication, auditing mechanisms, a GUI for management and monitoring, network segmentation, and a recoverable local store. The GridGain Ultimate Edition includes all the features of the GridGain Enterprise Edition plus a Cluster Snapshots feature for automated backups when using the GridGain Persistent Store feature in production environments.
- **Production Support**. GridGain Systems Support, available to GridGain Professional, Enterprise, and Ultimate users, includes rolling updates, faster availability of all releases and patches, and 24/7 enterprise-level support.

The Future of In-Memory Computing

As businesses cope with an explosion of data and users who increasingly expect real-time performance, many have turned toward newly affordable, increasingly full-featured in-memory data solutions. As a result, in-memory computing platforms are becoming the systems of record for a growing number of organizations, allowing them to keep most or all of their data in memory for the fastest possible performance – and providing them with a flexible platform which is ready to support the coming waves of in-memory innovation.

The in-memory platforms of the future will embrace these trends and go further. Not only will they offer the key capabilities that database users expect, such as SQL support, they will also provide a bridge to emerging use cases, such as machine learning and deep learning, and transformative new storage technologies, such as non-volatile memory. In other words, they will follow in the footsteps of Apache Ignite and GridGain. With full ANSI SQL-99 support, extensions for ML and DL, NVM support, and a hybrid storage model that enables flexible expansion to supports petabytes of data, Apache Ignite and GridGain are leading the way toward a vibrant in-memory future.

Contact GridGain Systems

To learn more about how GridGain In-Memory Data Fabric can help your business, please email our sales team at sales@gridgain.com, call us at +1 (650) 241-2281 (US) or +44 (0) 7775 835 770 (Europe), or complete our contact form to have us contact you.

About GridGain Systems

GridGain is revolutionizing real-time data access and processing by offering an in-memory computing platform built on Apache® Ignite™. GridGain solutions are used by global enterprises in financial, software, ecommerce, retail, online business services, healthcare, telecom and other major sectors, with a client list that includes Barclays, ING, Sberbank, Misys, IHS Markit, Workday, Silver Spring Networks and Huawei. GridGain solutions can connect data stores (RDBMS, NoSQL and Apache® Hadoop®) with web-scale applications or can function as a standalone transactional SQL database to enable massive data throughput and ultra-low latencies across a scalable, distributed cluster of commodity servers. GridGain is the most comprehensive, memory-centric, in-memory computing platform for high volume ACID transactions, real-time analytics, and hybrid transactional/analytical processing. For more information, visit gridgain.com.