G00230658

Top 10 Strategic Technology Trends: In-Memory Computing

Published: 3 February 2012

Analyst(s): Massimo Pezzini, Carl Claunch, Joseph Unsworth

In-memory computing (IMC) assumes the primary locus for application data is the computer main memory, not external storage devices. This leads to greater system performance and scalability and holds the potential for dramatic application innovation. Despite multiple challenges, IMC is poised for mainstream adoption over the next five years.

In this research, IT leaders will find a discussion about how evolution of commodity memory technology, system architectures and IMC-enabling application infrastructure will impact users' IT strategies and key recommendations for adoption.

Impacts

- The relentless price declines of memory and the recent advent of solid-state drive (SSD) technology has enabled IMC to become more affordable and impactful for IT organizations.
- IT organizations can cost-justify the use of Flash memory implemented as a discretely addressed form of computer memory to improve the performance of applications and system software in many more cases than with other Flash approaches.
- Enterprises that do not consider adopting in-memory application infrastructure technologies risk being out-innovated by competitors that are early-mainstream users of these capabilities.

Recommendations

- Recognize that Flash memory is fallible and requires data management technologies based on its application workload to drive the most optimal and cost-effective usage.
- When the effective cost of Flash-based SSD is very high, a large storage cache is ineffective in meeting performance targets and it is reasonable to leverage Flash as a new memory layer once the capabilities are mature enough to suit the organization's needs.
- Brainstorm with business leaders about how IMC-enabled application infrastructure can be leveraged to deliver business breakthrough applications, but embrace these technologies on a project-by-project basis, focusing on high-risk/high-reward initiatives or "systems-of-innovation" projects with an ROI of three years or less.

Analysis

Gartner defines IMC as a computing style in which the primary data store for applications is not on electromagnetic rotating disks, but some form of memory (e.g., DRAM or NAND Flash-based). This assumption implies that applications always (at least theoretically) experience negligible data access latency, even if they need to scan large volumes of data, such as in analytics or event-processing scenarios (see "Innovation Insight: Invest in In-Memory Computing for Breakthrough Competitive Advantage").

Multicore 64-bit processors, commodity memory technology, advanced clustering and cloud architectures provide the cost-effective platform for IMC, enabled by powerful, specific, and in many cases sufficiently proven, application infrastructure software. Therefore, IMC is also rapidly becoming affordable and palatable for mainstream users and is no longer the preserve of the most deep-pocketed organizations.

Dramatic business innovation — for example, cloud services — has been enabled by the wholehearted adoption of IMC. Entire industries, such as Web commerce, online gaming, financial trading, social networks or global software as a service (SaaS), could not exist without IMC. The scalability, performance and continuous availability required by those markets would not be possible using traditional computing models and patterns (see Figure 1).

Figure 1. Impacts and Top Recommendations for In-Memory Computing

Impacts	Top Recommendations
The relentless price declines of memory and the recent advent of SSD technology has enabled IMC to become more affordable and impactful for IT organizations.	• Recognize that Flash memory is fallible and requires data management technologies based on its application workload to drive the most optimal and cost-effective usage. • Only select vendors with proven solutions supported with references, which possess capable support and services and are financially stable.
IT organizations can cost-justify use of Flash implemented as a discretely addressed form of computer memory to improve performance of applications and system software more than with other Flash approaches.	• When the effective cost of Flash-based SSD is very high, a large storage cache is ineffective in meeting performance targets and it is reasonable to leverage Flash as a new memory layer once the capabilities are mature enough to suit the organization's needs.
Enterprises that do not consider adopting in-memory application infrastructure technologies risk being out-innovated by competitors that are early-mainstream users of these capabilities.	• Brainstorm with business leaders about how IMC-enabling application infrastructure can deliver business breakthrough applications. • Experiment with these technologies for high-risk/high-reward initiatives or “systems-of-innovation” projects with an ROI of three years or less.

Source: Gartner (February 2012)

Impact: The Relentless Price Declines of Memory and the Recent Advent of SSD Technology Has Enabled IMC to Become More Affordable and Impactful for IT Organizations

The commodity memory industry, DRAM and NAND Flash, has witnessed unprecedented investment over the past 10 years to satiate the performance and portable storage needs of the digital revolution. This resulted in remarkable price declines that, despite the technical and financial challenges, show no signs of abating. In 2011, a gigabyte (GB) of DRAM memory would cost about $10, while NAND Flash would be just over $1 per GB. By 2015, DRAM pricing will fall to about $2 and NAND Flash will fall to about $0.25 per GB, which will have profound implications for data centers when properly implemented and optimized (see "Forecast: Memory, Worldwide, 2005-2015, 4Q11 Update").

DRAM and Flash memory used in servers has to be customized and optimized to meet performance and quality demands, which adds a significant price premium that must be considered in the evolution of data center costs and technologies. This is especially true for Flash memory, historically designed for less-demanding consumer applications that fueled this industry. Flash memory's tremendous price declines have come at the notable cost of lower quality, particularly in reliability. This quality degradation will accelerate as Flash memory prices continue to plunge in conflict with the stringent enterprise computing requirements. Therefore, Flash memory must be managed and optimized in various types of SSDs (see Note 1, "Emerging Technology Analysis: Enterprise Solid-State Appliances" and "How to Evaluate Solid-State Drives for Enterprise Storage").

Users must first understand the application workload requirements and then select the most cost-effective Flash technology and sophistication of the data management technology accordingly (see "Marketing Essentials: How to Use Application Workloads to Drive Go-to-Market Plans for Enterprise-Grade Solid-State Drives"). As IMC workloads become more understood, monitored and predicted, expect the capacity, reliability and performance demands placed on Flash memory to increase, thus dictating the underlying technology. Ultimately, each IMC workload will be unique, but deployments that are flexible for scalable capacity and high availability will be critical.

An entire industry has developed to address the fallibilities of Flash memory technology by means of an intimate marriage of hardware and software. The deluge of established and new vendors with diverse competencies is driving innovation and competition, but also hype and confusion. This is particularly critical for IMC. Ultimately, the value proposition of intelligently optimizing DRAM and SSD technology in concert is to eliminate the input/output (I/O) storage bottleneck enabling transformational throughput performance with minimal latency, while decreasing power and space requirements. As this approach will be refined in the next two years and tailored specifically to the demands of IMC-enabled application infrastructure technologies, the overall value proposition can be quite compelling.

Recommendations

- Identify IMC application workload performance and high-availability requirements in order to determine the most appropriate and cost-effective memory and SSD technology that will be scalable as IMC needs evolve.
- Integrate storage efficiency and data management technologies to yield the greatest ROI for IMC applications through greater processor utilization, while also defending against the fallibilities of the SSD technology employed.
- Given the recent deluge of system providers, only select vendors with proven solutions supported with references that possess capable support and services and that are financially stable.

Impact: IT Organizations Can Cost-Justify the Use of Flash Memory Implemented as a Discretely Addressed Form of Computer Memory to Improve the Performance of

Applications and System Software in Many More Cases Than With Other Flash Approaches

Flash memory devices can be used in three practical ways: as a substitute for rotating disk (in the form of SSD), as a cache to improve the speed of rotating disk, and as a new type of memory complementing RAM.

Flash as SSD

SSDs are attached to the system using disk interfaces and protocols, and thus appear as file systems on rotating disk drives to the operating system and applications (see Note 2).

When application data won't fit, or the cost to place it solely in RAM can't be justified for performance improvement, it can be moved onto an SSD. The granularity of placement is a single file or a volume. Given its higher cost than rotating disks, only the performance-sensitive data should go into Flash, but if the file or volume is many times bigger, the user will overpay for the performance benefits due to the larger capacity used to store the file, including the small amount of data actually required. If access to the entire file or volume will deliver performance speedups, the effective cost is better than the general case when the percentage of data needed for performance-enhancement is low. If granularity were finer, then the effective cost would be low enough to justify Flash even where a file or volume surrounds the important data with a heap of low-priority elements. When in-memory data is placed solely within the server RAM, it is not constrained to a file or volume and the programmer does not have to use file semantics and APIs, as is usual when the data is on an SSD.

Flash as Disk Cache

Rotating disk drives generally package some form of memory as a cache. Modules with several drives inside may have an additional module-wide cache. Flash technology can be used in this context as a complement to or replacement for DRAM (see Note 3).

Flash as a disk cache transparently benefits disk performance. Users, applications and operating systems see this as a better-performing (although not uniformly) rotating disk drive. When it runs out of space, the cache will dump the oldest data elements by using least-recently-used (LRU) algorithms. This provides the most benefit if the data access pattern fits the LRU algorithm. If the application reads data nearby to where it has very recently read, cache is very effective. But if the access pattern is random or returns to spots after considerable intervening activity, the odds are that the data requested cannot be delivered from the cache. While some adjustments of the caching algorithm are possible, it generally treats all such data as equally important. Thus, performance-critical data may be replaced in the cache with relatively unimportant data unless flagged for priority. Even these flags designate entire files or volumes, thus leading to the same granularity issues as SSDs.

Flash as a New Memory Layer

If Flash can be implemented as a discrete memory, granularity is optimized (see Note 4). These devices will need to have some form of error checking and correction and operating system recovery support, just as with DRAM-based memory, to accommodate the temporary and permanent error rates of Flash chips. Because of the granularity advantage, many more situations will be cost-effective to speed up using Flash as discrete memory, not just those that fit well with SSD or Flash as a disk cache.

This use of Flash is just emerging today, but as yet there is no published API nor hardware access standardization. It will take a few years for the standards to be agreed on and for Flash as a new memory layer to be offered widely by system vendors. Operating systems will have to implement support for the technology and the APIs. Finally, middleware and application providers will need to leverage these APIs to place performance-critical data appropriately. As this approach reaches maturity, it should be added to the arsenal of tools to enhance performance, and should be applied during application design to suit the intended service levels. With a lower cost than RAM memory, Flash as a discrete memory layer may permit a computer to be used with substantially larger memory to support IMC architectures than a RAM-only machine.

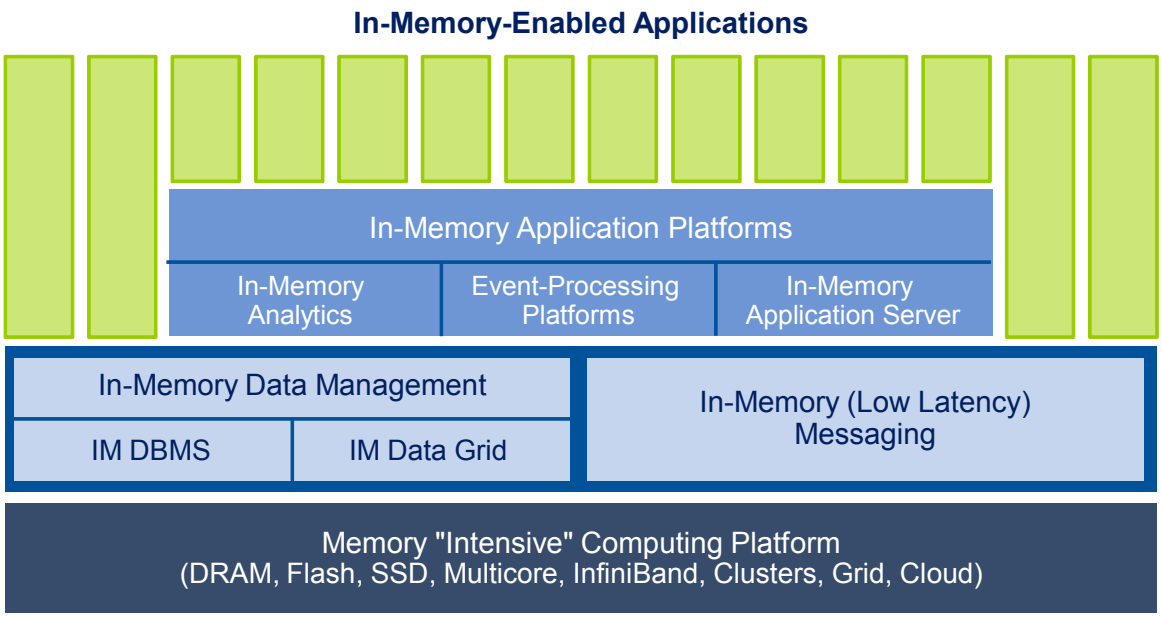
Recommendations:

- When the quantity of data is modest or the majority of the data in the files or volumes will yield a sizeable performance boost when access speed is slashed, consider placing it in Flash-based SSD.
- When the proportion of performance improvement to added cost is too low to justify SSD adoption, evaluate placement of data on storage with a larger cache and verify that the speedup will meet requirements.
- When the effective cost of Flash-based SSD is very high but the need for a performance speedup is critical and alternatives are infeasible or costlier, implement SSD today with the intent to move to Flash as a new memory layer once possible. If it is reasonable to wait, plan to leverage Flash as a new memory layer once the capabilities are mature enough to suit your organizations' needs.

Impact: Enterprises That Do Not Consider Adopting In-Memory Application Infrastructure Technologies Risk Being Out-Innovated by Competitors That Are Early-Mainstream Users of These Capabilities

On the market there are currently several classes of IMC-enabling application infrastructure products that Gartner classifies using the taxonomy in Figure 2.

Figure 2. IMC-Enabling Application Infrastructure Taxonomy



Source: Gartner (February 2012)

These technologies enable transformational applications that were virtually impossible to implement in the past, but they can also deliver incremental performance, latency and scalability benefits to established applications.

IMC-enabling application infrastructure is still relatively expensive and difficult to justify because of factors such as:

- Lack of commonly agreed-on, mature standards, which make products hard to replace (thus increasing vendor lock-in risks) and skills difficult to transfer
- Complexity in design, implementation, testing, management and monitoring
- New high-availability and disaster recovery challenges
- New security vulnerabilities
- Lack of best practices
- Hard-to-find skills

Many of these challenges (for example, skills and best practices) are likely to be addressed over the next five years, whereas others (for example, standards) may take longer to overcome. However, in many instances, products are mature enough and able to address issues such as:

- Reducing applications' running costs via database/legacy application offloading
- Improving transactional applications' latency, performance and scalability

- Boosting analytical applications' response times
- Maximizing the benefits of applications' co-location by minimizing interapplication messaging latency
- Dramatically shortening batch process execution times
- Enabling real-time, self-service analytics by reducing the effort required to build and populate data warehouses and data marts
- Implementing situation awareness and management-by-exception strategies

Examples of the use of IMC-enabling application infrastructure abound in multiple vertical sectors (such as capital markets, media and entertainment, travel, e-commerce, military and defense, SaaS/cloud services, social media and others) and geographies. Nevertheless, most IT organizations are unaware of the dramatic potential for breakthrough innovation carried by these technologies and only the most leading-edge IT organizations have dared to consider the previously unthinkable applications that these technologies enable (for examples of such applications, see Note 5).

However, through dramatic and relentless reductions in the cost of memory-intensive hardware, the growing popularity of IMC-enabling application infrastructure products, and the increasing utilization of these technologies as enablers for packaged applications, SaaS offerings and other application infrastructure products are paving the way for widespread adoption by early-mainstream organizations.

Recommendations:

- Brainstorm with business leaders about how IMC-enabled application infrastructure can be leveraged to deliver business breakthrough applications (for example, real-time fraud detection, inventory analytics, supply chain tracking, algorithmic trading, revenue leakage management, smart metering and smart grid).
- To gain technical knowledge about IMC-enabling application infrastructure, task a small team of architects and engineers to investigate the short- and midterm business value opportunities of this set of technologies.
- Embrace IMC-enabling application infrastructure technologies on a project-by-project basis. Balance the anticipated business benefits with the costs and risks typically associated with the adoption of emerging, leading-edge technologies. To mitigate risks, initially experiment with IMC-enabling application infrastructure in systems-of-innovation-type projects with an ROI of three years or less.

Recommended Reading

Some documents may not be available as part of your current Gartner subscription.

"Innovation Insight: Invest in In-Memory Computing for Breakthrough Competitive Advantage"

"Predicts 2012: Cloud and In-Memory Drive Innovation in Application Platforms"

"Predicts 2012: Cloud Computing and Event Processing Will Be the Key Advances in Application Architecture"

"Emerging Technology Analysis: Enterprise Solid-State Appliances"

"Forecast: Memory, Worldwide, 2005-2015, 4Q11 Update"

"Marketing Essentials: How to Use Application Workloads to Drive Go-to-Market Plans for Enterprise-Grade Solid-State Drives"

"How to Evaluate Solid-State Drives for Enterprise Storage"

"How Flash Memory in Servers Delivers Higher Value as a Uniquely Addressable Memory Type"

"Taxonomy, Definitions and Vendor Landscape for Application Platform Products"

"What CIOs Need to Know About In-Memory Database Management Systems"

"Business Reasons to Implement Event Processing"

"Need for Speed Powers In-Memory Business Intelligence"

"Hype Cycle for Application Infrastructure, 2011"

"Hype Cycle for Business Intelligence, 2011"

"Hype Cycle for Business Process Management, 2011"

"Hype Cycle for Data Management, 2011"

Note 1 Flash Memory and SSD Defined

Flash memory devices are nonvolatile, which means that they can store data persistently without the need for power (as opposed to DRAM, which requires constant power). Flash memory performance, while not as fast as DRAM, is considerably cheaper, yet compared with HDD is much more expensive but has significant performance, latency, power and size advantages. SSD for data centers is defined as a nonvolatile (Flash-based) or volatile (RAM-based) semiconductor memory-based device with integrated management capabilities to achieve high performance and reliability; these can vary based on the technology employed in the SSD. Although SSDs often mimic HDDs today, SSDs are defined by how the host system accesses it, whether as a storage, memory or as some other paradigm. Also, SSDs are not limited to a particular form factor or interface, as these continually evolve.

Note 2 Flash as SSD

The software does nothing specific to place data in an SSD — instead, it is accessed as a file, using the same programming interface and methods as when the file is placed on rotating disk. Similarly, the operating system views the SSD as a disk drive connected to an interface, such as Serial

Advanced Technology Attachment (SATA) or Small Computer System Interface (SCSI), with a driver managing the SSD activity. While nothing explicit is necessary to use an SSD, the realities are that SSDs have uniform random access times and benefit when updates are minimized. Thus, if the operating system is SSD-aware, it might disable unnecessary and detrimental activities such as compaction, as well as bunch together changes to a "block" to lower the number of writes made to the device.

Note 3 Flash as Disk Cache

Disk caches have been implemented with DRAM chips using a small amount of RAM to improve the effective speed of the disk. Data from any section of the disk that was recently read may still be in the RAM, allowing it to be read by the computer much faster than the native disk speed. Since the total size of cache is small and newer requests tend to replace older information, a request to read data in a spot that had been read recently may not work because it cannot be found in the cache and will instead transfer it at disk speed. Any write to the disk drive would be made to the cache but also directly to the rotating disk, since DRAM is not persistent. To guarantee persistence for any updates to the drive, it will have to wait until the rotating disk write is complete before signaling completion to the computer. Module-wide caches might have batteries installed, in which case they can maintain power to the RAM and keep the data persistent even if external power fails. This allows those modules to signal completion after the data is only written to the cache, with the module moving those updates to rotating disks at some later point. Therefore the computer sees the write as much faster. Flash memory, with its inherent persistence, allows for a cache to guarantee persistence more readily, but only if the disk drive has enough capability within it to ensure that the updates will eventually get written to rotating disk. For caches within SSD drives this typically is not implemented, but could be. For a module-wide cache, this is a more common capability. On top of this benefit, Flash memory is less expensive than DRAM, thus lowering costs or allowing for larger cache sizes at a given price point. For reads and for writes in the absence of the appropriate persistence mechanisms, Flash has no speed advantage over the same capacity of a DRAM-based cache.

Note 4 Flash as a New Memory Layer

When Flash is used as a new memory layer, only the specific data elements that are important for performance need be in Flash. Thus, the expensive Flash memory is used only for data that will really benefit from its use, making the effective cost low. Even when the data elements we want would have been part of some larger collection like a file, the application or system software can put only the important data into memory. The software would not access this data through the file programming interfaces. The operating system would not see this as a disk drive nor access it through a driver. Ideally, a new application programming interface is used to get and put data from this new memory layer. Because Flash memory is persistent, the software can put information there that will be preserved even when a power failure brings down the computer, avoiding the need to write it out to a file on some disk drive to gain the persistence. Thus, the user pays for only the capacity of Flash that delivers high performance benefits and the design of the software takes advantage of the unique character of Flash memory — persistence, speed and cost intermediate between RAM and disk.

Note 5 Examples of Breakthrough IMC-Enabled Applications

- Online gaming/entertainment
- SaaS and PaaS
- Risk management
- Real-time fraud detection
- Algorithmic trading
- Global Web commerce
- Ad hoc analysis/pivoting
- "What if" analysis
- Security intelligence
- Profitability analysis
- Inventory forecasting
- Supply chain tracking
- Communication service delivery
- Telecom/media revenue leakage management
- Sales incentive promotions management
- Airline, railway and fleet operation management
- Smart metering and smart grid operations

Regional Headquarters

Corporate Headquarters
56 Top Gallant Road
Stamford, CT 06902-7700
USA
+1 203 964 0096

Japan Headquarters
Gartner Japan Ltd.
Aobadai Hills, 6F
7-7, Aobadai, 4-chome
Meguro-ku, Tokyo 153-0042
JAPAN
+81 3 3481 3670

European Headquarters
Tamesis
The Glanty
Egham
Surrey, TW20 9AW
UNITED KINGDOM
+44 1784 431611

Latin America Headquarters
Gartner do Brazil
Av. das Nações Unidas, 12551
9° andar—World Trade Center
04578-903—São Paulo SP
BRAZIL
+55 11 3443 1509

Asia/Pacific Headquarters
Gartner Australasia Pty. Ltd.
Level 9, 141 Walker Street
North Sydney
New South Wales 2060
AUSTRALIA
+61 2 9459 4600