MOOR
INSIGHTS & STRATEGY

Optimizing Memory With ZeroPoint Technologies

How ZeroPoint IP Drives Energy-Efficient Performance While Lowering TCO

Situation Analysis

Data drives the modern business. While this has always been true, it has perhaps never been more relevant. Data is generated everywhere and at all times, feeding AI models and analytics engines to help organizations do more and do it faster.

The requirement to do more faster is coupled with the requirement to drive down costs. While C-level executives are pressured to reduce time to value, there is an equal and seemingly contradictory requirement to lower costs, all while harnessing the potential in the unprecedented amounts of data being generated, collected, and trained.

One of the main factors limiting data-hungry workloads from performing faster is a server's memory architecture. Current memory architecture has lagged innovation, resulting in significant latency and power consumption. In the broader picture, memory — one of the largest server cost elements — levies a power and performance tax.

This research brief will detail the challenges in data-driven enterprises, including how ineffective memory management costs directly and indirectly burden modern businesses. Further, it will explore some industry initiatives and look at how companies like ZeroPoint have developed technologies that help drive greater efficiencies that can lead to unprecedented performance gains, considerable total cost of ownership (TCO) savings, and untapped revenue opportunities.

The Challenge: Performance, Efficiency, Cost

IT organizations are very thoughtful when evaluating CPU and memory configurations deployed in the server infrastructure that will support applications across an enterprise. Although CPUs are chosen for best overall performance, how performance is measured varies. Industry benchmarks provide insights into per-core or workload-based performance, but requirements vary depending on IT organization and workload.

Higher core counts are sometimes important. Other cases require the absolute best per-core integer performance. Additionally, larger cache sizes are also necessary — sometimes critical — for workload performance. This is especially true in data-driven applications that span high-performance computing (HPC), analytics, and AI.

Also significant is the amount of random-access memory (RAM) supported by a CPU and its bandwidth (i.e., the number of channels). When considering RAM, this memory is configured based on individual DIMM sizes. Data-driven applications and workloads benefit from larger DIMM sizes that can store more data contiguously on a single DIMM so there is less latency when fetching, moving, and processing data.

This means deploying the same amount of RAM across more DIMMs with less capacity forces data to be stored in and fetched from the next level of memory hierarchy. This, in turn, will introduce more latency. While some applications are unbothered by this impact, that is not the case for time-sensitive applications. Consider a GenAI-driven chatbot for customer support. In times of high utilization, a measurable lag can be introduced to customer inquiries, which can result in poor satisfaction or lost sales opportunities.

Because of these factors, servers designed for these applications specify requirements for deploying larger DIMMs to assure the best absolute performance. These DIMMs are extremely expensive but absolutely necessary.

The Inefficiencies of Memory

Depending on the configuration deployed, memory can be the costliest hardware element in a server deployment. In addition to being expensive to acquire, it is also a big part of the TCO equation because it is power hungry and contributes to overall power consumption.

Even with a server configured with CPUs that support the largest cache and RAM, application performance still falls far short of its potential. The reason for this is in the very design of RAM and the lack of innovation.

While CPUs and GPUs have undergone considerable rearchitecting to support the needs of new workloads, from multi-core to chiplets to advanced instruction sets that deliver targeted acceleration, dynamic and static RAM (DRAM and SRAM) have not evolved to the same degree. This performance differential, which results in CPUs sitting idly by waiting to perform (the very definition of latency), will only increase as process nodes from chip manufacturers such as TSMC move from 5nm to 3nm and below.

In addition to the latency introduced by memory inefficiencies, there is also a carbon cost. Fetching data from memory requires more system-level power consumption, directly contributing to the seemingly out-of-control per-server, rack, and data center power consumption. In fact, a recent study from the International Energy Agency (IEA) estimates datacenter consumption to be between 2%-3% of the global power footprint, roughly equal to the power consumption of Australia. This percentage is forecasted to increase four times to 8% by 2030, equivalent to the power consumption of India.

Consider a simple query through a search engine. The IEA estimates that a ChatGPT query consumes nearly 10 times as much power as a Google search (2.9 watt hours vs. 0.3 watt hours). Incorporating such a technology as a mainstay search utility would require an additional 10 terawatt hours of electricity per year. This is mainly due to the sheer amount of data stored in RAM needed to respond to each inquiry.

The result of these memory inefficiencies is higher direct and indirect costs as seen in more expensive server configurations and power expenses.

MEMORY OPTIMIZATION: IS COMPRESSION THE ANSWER?

One way to solve the memory inefficiency challenge is through compression, traditionally done through software. Software-based compression, which has been one of the easier answers for OS and application vendors, means solutions can be deployed with minimal impact on server and memory designs. However, software-based compression can introduce 10,000 nanoseconds (ns) of latency. Further, software-based compression is highly variable in performance due to the data types being compressed and the amount of CPU overhead required.

Hardware-based compression is another method used to address the memory bottleneck challenge. Relative to software-based compression, the hardware implementation of such compression can reduce latency to 2,000–3,000 ns. Additionally, CPU overhead is eliminated because compression is offloaded to dedicated hardware, and applications can achieve consistent performance. However, this 2,000–3,000 ns of added latency becomes too much when compressing data on memory, which has a latency threshold of 100–200 ns.

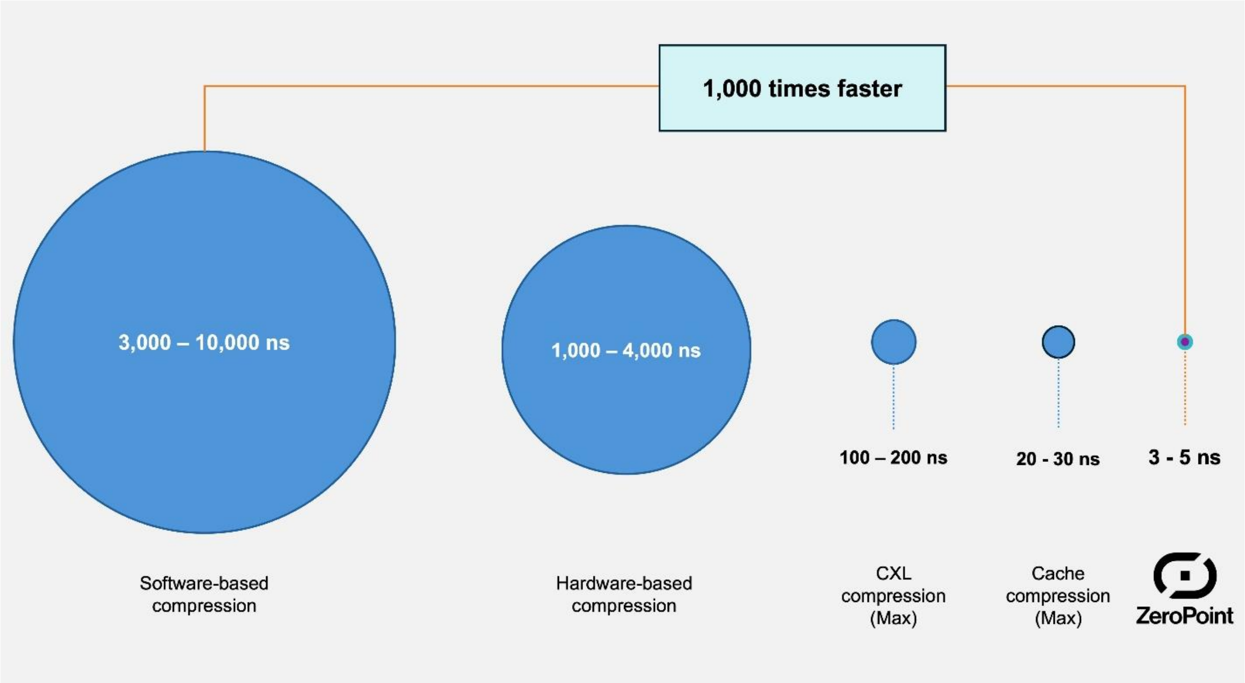
An emerging standard that has support among hardware vendors is the Compute Express Link (CXL). This open-standard interconnect delivers a virtual expansion of memory through a combination of coherence, compression, and increasing bandwidth. The challenges around CXL are tied to complexity and maturity. For software to benefit from CXL, it must be rearchitected. This rearchitecting starts in the operating system and expands into drivers.

From a maturity perspective, CXL faces the same challenge that all newly implemented standards do: evolution that has a significant downstream impact. As CXL is employed by hardware vendors and deployed (and utilized) by enterprise IT organizations, new requirements and refinements will inevitably arise. This can lead to hardware-level updates that can be extremely costly in terms of design, manufacturing, and validation.

Going even further up in the memory stack to on-chip SRAM, the added latency introduced by a compression algorithm cannot be higher than a few ns. The established approaches to expand memory capacity through data compression are therefore insufficient to address higher and more critical levels of memory, such as directly connected memory or on-chip memory.

Moor Insights & Strategy (MI&S) has reviewed one technology provider that shows significant promise. ZeroPoint Technologies, a memory optimization intellectual property (IP) provider with roots in academia and high-performance computing, has developed a solution that can deliver near latency-free performance (3-5 ns). Its compression, therefore, can outperform existing standards by a factor of 1,000 times when it comes to latency.

FIGURE 1: ZEROPOINT MINIMIZES LATENCY



Current alternatives introduce 1000 times more latency than ZeroPoint's memory optimization
Source: ZeroPoint Technologies

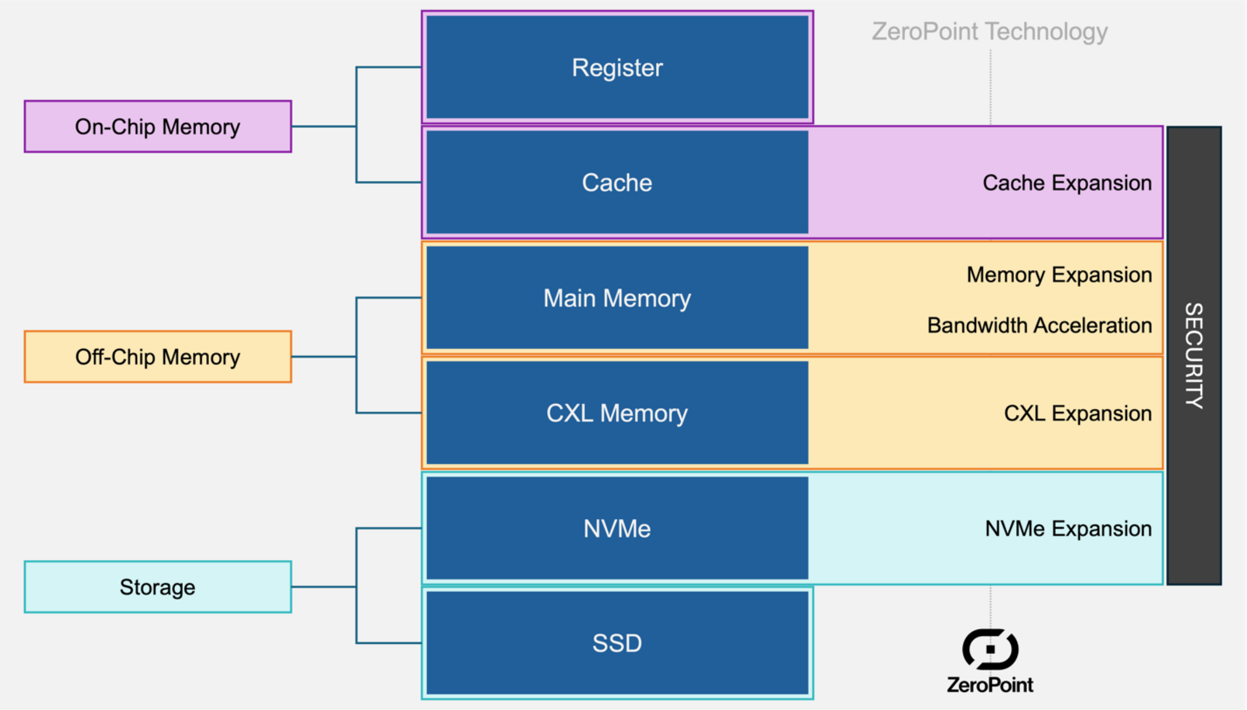
Perhaps the most important aspect of the ZeroPoint value proposition is efficiencies across the memory estate. While ultra-low latency is in a class of its own, other measurements are equally impressive. This includes an improvement in performance per watt of approximately 50%, an 80% increase in CPU productivity, and up to 25% better server TCO. Further, it should deliver considerable savings regarding the total cost of carbon ownership (TCCO).

What makes ZeroPoint equally interesting is its ability to support virtually any environment, from L3 Cache to fast storage and everything in between. ZeroPoint is perhaps the very definition of data optimization at its foundational level.

ZeroPoint — Differentiated IP and Memory Optimization

For effective and what MI&S would consider perceptibly native performance, memory compression and management must be broad, deep, and transparent. This means compression must span the data pipeline and do so with high efficiency. Additionally, it must be complementary (and invisible) to process nodes, memory technologies, operating systems, and the applications that it benefits. In the case of ZeroPoint, this is precisely what its technology does. It addresses inefficiencies across the memory estate — on-chip (cache expansion), off-chip (memory expansion, bandwidth acceleration), and storage (NVMe expansion).

Figure 2: ZeroPoint Optimizes the Memory Estate



ZeroPoint's technology addresses performance and power efficiency across the memory hierarchy
Source: ZeroPoint Technologies

As seen in the above graphic, ZeroPoint optimizes how data is stored, from the point of fast storage (NVMe) to CXL-connected memory to main memory and cache. It performs this through what is referred to as expansion.

WHAT IS EXPANSION?

Expansion is a process by which the usable capacity in memory is expanded by lossless compression. Essentially, relevant data is compressed and stored in a more organized and accessible fashion, while redundant data is discarded.

While any compression technology will deliver some level of performance improvement and memory–storage expansion, ZeroPoint has shown to be comprehensive in its approach, incorporating technology innovations that span all levels of the cache-storage stack. Its solution is built on several underlying technologies that span the data pipeline:

Cache: ZeroPoint's CacheMX solution delivers cache expansion by compressing on-chip SRAM (cache). The result of implementing CacheMX is a 2-4 times larger cache capacity, which mitigates the scale–real estate challenges introduced as process nodes continue to shrink. Doubling the size of a cache typically means a 30% improvement in computational performance.

Near–Memory (directly connected DRAM): Memory has multiple challenges that span capacity and bandwidth, both of which have a material impact on latency and, therefore, performance. In response to these challenges, ZeroPoint developed two solutions — ZiptilionBW and SuperRAM. ZiptilionBW is a hardware accelerator that, like all ZeroPoint products, is seamless to applications and operating systems, intercepting and compressing data on the fly to deliver faster performance. ZiptilionBW supports both CPUs and GPUs.

SuperRAM is a hardware accelerator integrated directly into the SoC. Specifically, it accelerates two critical compression solutions: zram and zswap. By providing a hardware offload engine for these algorithms, SuperRAM can deliver considerable performance gains while achieving power savings.

Far–Memory (CXL-connected DRAM): The emergence of large-core CPUs further renders near-memory insufficient. As a response, the CXL standard was created and adopted as a common interface to scale memory, accelerators, and storage.

While CXL-connected memory offers efficient memory bandwidth and capacity expansion, hyperscalers Meta and Google have pressed the case for an underlying

CXL hardware accelerator in this proposed Open Compute Project specification to meet the needs of their latency-sensitive workloads.

This is what ZeroPoint Technology delivers: a hardware-accelerated compressed-memory tier in the CXL standardized memory hierarchy. Such technology offers a 2-4 times memory expansion of the nominal memory investment and energy budget. Further, MI&S believes what ZeroPoint has delivered moves beyond hyperscalers in terms of relevance. Any enterprise organization with highly performant business-critical workloads will realize these benefits.

NVMe: ZeroPoint enables NVMe expansion through its FlashMX technology. FlashMX is another hardware accelerator for the zstd compression algorithm that resides on the SoC and intercepts data between the CPU and NVMe storage. Like all other ZeroPoint technologies, FlashMX is transparent to the OS and applications, meaning no low-level architecting is required for applications to take advantage of it. Provided ZeroPoint partners such as Intel and AMD expose this acceleration, it is enabled by default.

This is all wrapped in an off-the-shelf AES-XTS security IP solution, SphinX, which delivers a low-latency, highly secure enclave in which the ZeroPoint technologies operate. SphinX delivers 128- or 256-bit encryption.

While the systems and applications market is rich with solutions that promise to drive the lowest latency, MI&S sees ZeroPoint as unique in its approach to driving application performance, energy efficiency, and TCO savings at the foundational level. It ensures that benefits are realized regardless of application architectures and other variables that can impact other solutions.

Differentiated IP Stems From a Differentiated Team

The technology market is filled with bright ideas and interesting concepts from companies that the venture capital community has backed before producing a product. Indeed, the AI era rivals the early days of the Internet in terms of startup funding for ideas that may never see productization.

ZeroPoint is different in a couple of ways. First and foremost, it is a company that has taken its concept to product and can demonstrate significant performance gains and project cost savings. Additionally, it is a company filled with technology professionals who have been addressing challenges in the systems and memory market for years — a team with distinction and roots in academia and industry. It is this theoretical, practical, and relevant expertise that has guided the design of a solution to a problem that hinders nearly every organization.

MI&S sees ZeroPoint as a point technology worthy of consideration for chip makers and systems designers looking to create differentiation in latency-sensitive markets such as HPC and AI. This is particularly the case for GenAI-driven applications supporting chatbots and other natural language processing functions.

SUMMARY

The modern business is data driven. This data feeds AI models and analytics engines to help companies do more faster, to allow governments to deliver critical services to constituents faster and more accurately, and to enable consumers to acquire and consume services more efficiently.

However, the memory architectures used to temporarily store and move data were designed in a different era to support applications that had a different way of measuring performance. These memory architectures have not kept pace with CPU architectures.

While compression technologies have existed for some time, their performance has been substandard. Software compression has delivered some latency relief but without consistency of performance. Legacy hardware compression, although an order of magnitude more effective, has still fallen short and is challenging to deploy.

While CXL and other advancements help solve some of the challenges faced in driving data in and out of CPUs faster, latency can still be greatly improved. And a lack of maturity in terms of real-world operation can lead to challenges as the CXL standard is implemented more broadly.

ZeroPoint delivers hardware-based compression that is transparent (requiring no modifications to the OS) and performant (near latency-free). Further, its approach to addressing bottlenecks both broadly and deeply assures that applications reliant on this deluge of data are constantly fed, meaning CPUs, GPUs, and xPUs run at maximum performance and energy efficiency. The result is a reduction in latency by up to 1,000 times, along with an increase in memory and cache by a factor of 2-4 times.

Application performance is not the only value ZeroPoint’s portfolio delivers. Because memory is managed more efficiently, it is estimated that the company’s compression technology being deployed on processors can result in approximately 25% TCO reduction due in part to CapEx and power savings.

Because of these factors, MI&S sees ZeroPoint as a potentially key partner for any silicon provider looking to deliver best-in-class performance of applications such as HPC or GenAI. We see inference as a prime workload that can benefit from ZeroPoint's technology portfolio. If your memory vendor, CPU of choice, or cloud provider isn't employing ZeroPoint technology, ask why.

For more information on ZeroPoint and its IP, visit zeropoint-tech.com.

CONTRIBUTOR

Matt Kimball, Vice President and Principal Analyst, Servers

PUBLISHER

Patrick Moorhead, CEO, Founder and Chief Analyst at Moor Insights & Strategy

INQUIRIES

Contact us if you would like to discuss this report, and Moor Insights & Strategy will respond promptly.

CITATIONS

Accredited press and analysts can cite this paper, but it must be cited in context, displaying the author's name, title, and "Moor Insights & Strategy." Non-press and non-analysts must receive prior written permission from Moor Insights & Strategy for any citations.

LICENSING

Moor Insights & Strategy owns this document, including any supporting materials. It may not be reproduced, distributed, or shared in any form without its prior written permission.

DISCLOSURES

ZeroPoint commissioned this paper. Moor Insights & Strategy provides research, analysis, advice, and consulting to many of the high-tech companies mentioned in this paper. No employees at the firm hold equity positions with any of the companies cited in this document.

DISCLAIMER

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Moor Insights & Strategy disclaims all warranties regarding such information's accuracy, completeness, or adequacy and shall have no liability for errors, omissions, or inadequacies. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The views expressed herein are subject to change without notice.

Moor Insights & Strategy provides forecasts and forward-looking statements as directional indicators, not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on the future, they are subject to risks and uncertainties that could materially cause actual results to differ. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of this document's publication date. Please remember that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.