vertical compute

Reshaping the Future of Compute and Memory

Sebastien Couet - Founder & CTO - Vertical Compute

vertical compute

Reshaping the Future of Compute and Memory

Sebastien Couet - Founder & CTO - Vertical Compute

01. Foreword

The field of Artificial Intelligence (AI) is undergoing a revolution, driven by transformer-based models that have enabled the creation of Large Language Models (LLMs) and a new class of applications known as Generative AI (GenAI). At the core of this revolution is AI inference: a process where trained models make predictions from new data.

However, the exponential growth of these data-intensive applications has brought the semiconductor industry to a critical crossroad. Traditional compute architectures have long relied on a hierarchical memory ecosystem: a layered structure where data is moved between different types of memory based on speed, cost, and proximity to the processor. However, as demand for faster, denser, and more energy efficient memory solutions is rising rapidly, this ecosystem is increasingly struggling to keep pace. GenAI workloads nowadays require up to 100x more memory capacity than traditional AI, a demand so enormous that memory itself has become the fundamental bottleneck.

This white paper introduces Vertical Compute's groundbreaking Vertical Integrated Memory (VIM™) technology, a disruptive innovation set to fundamentally alter the landscape of high-performance computing. This white paper will explore the current challenges facing memory and compute workloads nowadays, delve into existing industrial and academic solutions as well as their limitations, and finally, present the VIM™ technology as the new solution that will be able to unlock unprecedented capabilities in the next generation of AI, data processing and beyond.

02.
The Memory Bottleneck in the Data-Compute Era

Computing architectures have long relied on a hierarchy of memory technologies, each optimized for different trade-offs between speed, density, cost and power consumption. At the highest level, integrated Static Random Access Memory (SRAM) provides ultra-fast access times, making it a critical component for high-speed processing tasks. However, SRAM suffers from low density, meaning that only a limited amount of data can be stored on-chip, which poses a challenge for AI workloads that require large datasets to be processed in real-time.

At a lower layer in the memory hierarchy stack, Dynamic Random Access Memory (DRAM) offers significantly higher density, allowing for the storage of much larger datasets. DRAM, however, is an external memory solution that is much slower than SRAM due to the need for constant refreshing and the latency introduced by data movement between the processor and memory modules. This performance gap creates inefficiencies, especially for AI workloads that rely on fast, iterative data processing.

Finally, NAND and NOR flash memory provides the highest density and non-volatility for high-capacity memory storage. However, this comes at the expense of speed and endurance, making it impractical for AI workloads.

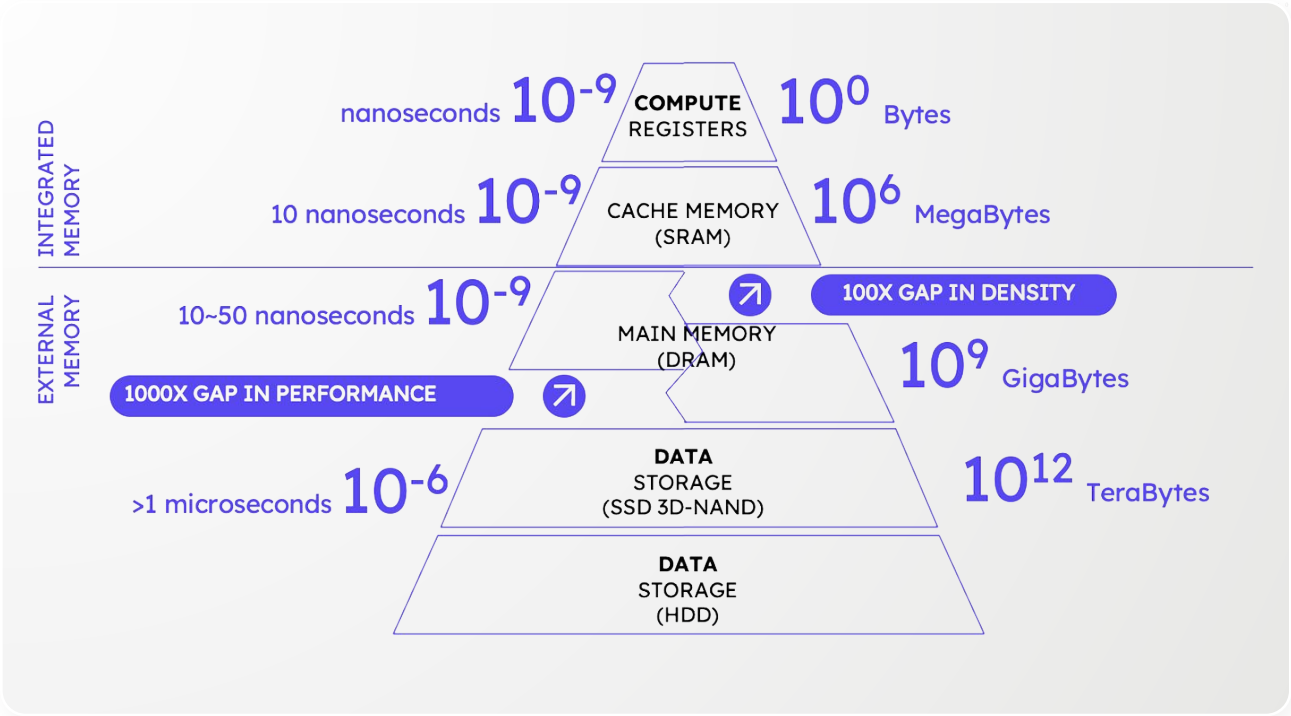

Figure 1 – Illustration of the hierarchy of memory architectures for compute systems.

Throughout the last 50 years, the relentless scaling of compute and transistors has been driven by these three main memory technologies. All three have been well-embedded in virtually every type of application and semiconductor device, from smartphones to data centers. The widespread adoption of these memory technologies is the result of decades of continuous research and development, which was aimed at optimizing their respective trade-offs in speed, density, and endurance. However, the challenge remains that each technology typically excels in only one of these aspects, failing to achieve a balanced combination of high speed, high density, high endurance, low cost and low power.

This period eventually also saw the emergence of several alternative technologies, such as Magnetoresistive RAM (MRAM), Resistive RAM (ReRAM), Phase-Change Memory (PCM), and Ferroelectric RAM (FeRAM). However, none of them achieved the widespread commercial use of SRAM, DRAM, or NAND. While these technologies faced many hurdles, a key limitation for many of them was their insufficient endurance. High-performance workloads simply cannot tolerate early life failures, and therefore, to replace the actual traditional memories with a near-unlimited endurance (i.e. SRAM and DRAM), these emerging technologies often fell short. This endurance gap, along with other challenges in scalability and cost, prevented them from disrupting the established memory technologies.

The first obstacles arose two decades ago when CPU core frequency scaling ended due to the breakdown of Dennard scaling[1]. This scaling principle (in combination with Moore's Law[2]) had allowed processor performances to increase exponentially without a proportional rise in power consumption for many years. However, as transistors continued to shrink, this principle no longer held, and the resulting exponential increase in power and heat made further clock speed improvements physically unfeasible.

With this traditional path for continuous improvement blocked, the industry shifted towards a new strategy: parallelism. This introduced the era of multi-core processors and, ultimately, the widespread adoption of multi-core GPUs for efficient data processing. This shift increased the demand for ever-faster memory buses, as these high-performance, numerous smaller cores were able to process substantially larger data volumes per second compared to the initial general-purpose CPUs.

However, the most significant obstacles have emerged more recently with the rise of AI. Early use cases, such as image recognition, already placed stringent requirements on memory subsystems, and new, even more demanding AI applications are being developed daily, pushing the memory performance to its absolute limits.

This challenge has intensified with the emergence of transformer-based algorithms, which have enabled the creation of Large Language Models (LLMs) for generative AI (GenAI) applications. At the core of these models is AI inference: a process where a trained model recognizes patterns and draws conclusions from input data.

[1] R. H. Dennard et al., "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE Journal of Solid-State Circuits*, vol. 9, Oct. 1974.

[2] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, Apr. 1965.

This process is now demanding a level of data bandwidth and memory capacity that traditional solutions cannot provide, leading to a significant divergence in how AI is deployed. This divergence has created two distinct ecosystems, each with its own trade-offs in memory capacity, performance and power:

01.

Edge AI Inference

This approach is used for conventional AI models like Deep Neural Networks (DNNs) and Convolutional Neural Networks (CNNs) for AI-vision inference, where the models are small enough in size to be stored in the limited on-chip memory of a device. To enable high-speed processing and eliminate the delays of off-chip data transfers, the industry has focused on increasing on-chip SRAM capacity for this specific purpose. By keeping data processing locally on the device, edge inference provides the crucial benefits of reduced latency and enhanced privacy for these specific workloads.

02.

Cloud AI (Training and Inference)

The push towards the cloud began with AI-vision training, where the sheer volume of training data and model parameters must be fed to large GPUs or NPUs at immense speeds. This drove the initial adoption of High Bandwidth Memories (HBMs) in data centers. This reliance on the cloud has become even stronger with the rise of GenAI's new class of models. These models allow applications like ChatGPT to create new content (such as text, images, and code) which are becoming increasingly massive in size, with tens to hundreds of billions of parameters that must be stored in memory. This scale, representing a 100x increase in memory capacity, has risen far beyond the capabilities of on-chip memory solutions. Consequently, advanced AI inference for GenAI is effectively "stuck in the clouds," run on expensive, power-hungry, multi-chip systems that are heavily reliant on HBMs.

The Influence of AI Workloads on Memory Architectures

The fundamental building blocks of these AI algorithms rely on computationally simple but massive Multiply-ACcumulate (MAC) operations that work on gigabytes (GBs) of model data. These types of operations are typically at the core of how AI-Processing Units (AI-PUs) perform calculations and come along with the requirement for an unprecedented rate of data delivery to the AI-PUs. This demand has led to a divergence in (new) memory solutions based on the AI workload:

AI-vision training is the process of teaching an AI model to recognize and understand images by feeding it large datasets. The primary challenge here is bandwidth. Training requires repeatedly streaming GBs of data through the system for every learning cycle. This demands extremely high memory throughputs, which is why GPUs from companies like NVIDIA, AMD, and Intel use HBMs paired with massively parallel processors for this type of workload. Without such a solution, the model simply cannot learn effectively.

AI-vision inference happens after training, when the model is actually used to make predictions in the real world. For example, recognizing a face on your phone or detecting objects for a self-driving car. The challenge here is very different, and is focused on achieving low latency and high speed. Instead of massive data streams, inference requires the model to react quickly, often on a small local device. To achieve this, the model's parameters and working data (weights and activations) must fit into the fastest available memory, usually on-chip SRAM. This makes inference much more about responsiveness and efficiency, creating a compact, self-contained processing environment optimized for quick answers.

LLM processing is an entirely different workload. Unlike vision inference, where speed on small devices is key, LLMs demand both enormous capacity and massive bandwidth. A single LLM can have hundreds of billions of parameters which is far beyond the memory capabilities of a single accelerator HBM chip. This means the model must be split across multiple chips in a datacenter environment, with HBM and high-speed interconnects shuttling data between them. This highlights a fundamental challenge: there is no single-chip solution for LLM processing, making it a workload that can only be handled with expensive datacenter infrastructure.

This growing problem is compounded by the fact that the total memory capacity demand over the last 15 years has increased by a factor of one million (as shown in Figure 2), far outpacing the capabilities of existing memory technologies.

The evolution to bigger AI models

The scale of artificial-intelligence neural networks is growing exponentially, a trend best measured by the number of model parameters. These parameters, which function like the connections between neurons in a brain, have expanded from millions to hundreds of billions in just a few years.

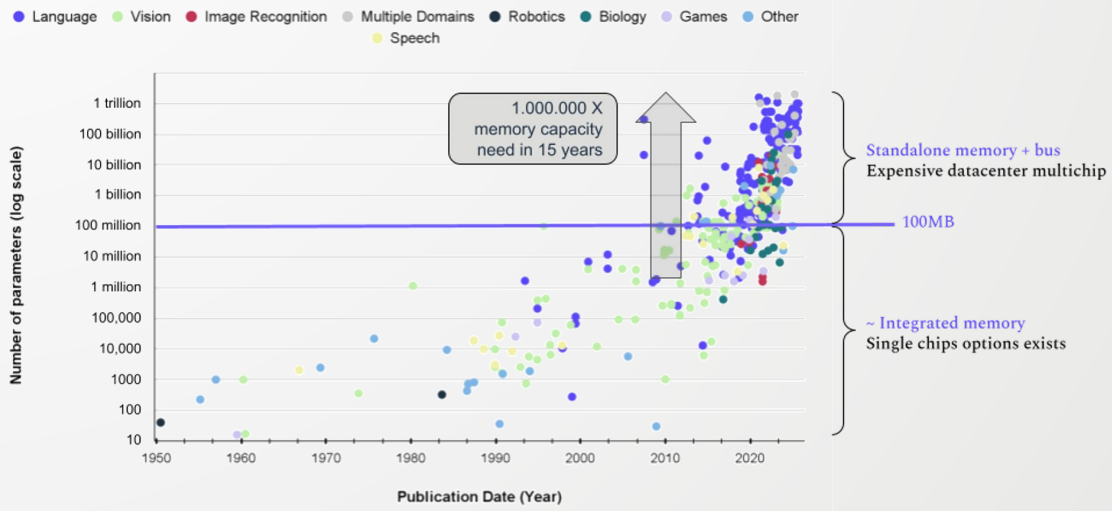

Figure 2 – The exponential growth of AI models, as shown by the number of parameters over time. Data sourced from Our World in Data, based on research from Epoch AI[3].

The LLM Model Size Deluge

The pace of change is so rapid that state-of-the-art model sizes are becoming obsolete within months. For context, while earlier models were in the hundreds of billions of parameters, the recent launch of ChatGPT-5 (mid-2025) is expected to have past the one trillion parameter mark, and even the more recently released DeepSeek V3.1 LLM model (August 2025) requires a massive 685 billion parameters.

[3] Epoch AI (2025) – with major processing by Our World in Data. "Datapoints used to train notable artificial intelligence systems", March 2025, Available at https://archive.ourworldindata.org/20251002-120107/grapher/artificial-intelligence-number-training-datapoints.html

These parameters represent the numerical values, also known as weights, that quantify the influence of one computational node on another within a complex model's network. During training, an LLM model will adjust these weights to minimize errors and improve its ability to predict the next word or generate coherent text. In essence, these billions of weights are the core of the model's "learned knowledge".

This massive data set is roughly equivalent to 685GB of memory, while even smaller versions still demand around 1 billion parameters.

Each inference (i.e. every token generated) requires loading all these parameters from memory to perform computations on. As a result, LLM execution involves storing and transferring GBs of both input and weight data at extremely high speeds. Although the exact impact of memory density on performance is complex, the memory demands far exceed the storage capacity of SRAM (at best about 100MB) and push beyond the data transfer limits of DRAM DDR5 buses, highlighting the urgent need for faster and denser memory solutions.

In summary, the current trend towards data-intensive applications and AI algorithms has exposed the critical limitations of the current memory ecosystem. The fundamental trade-offs in existing memory technologies force system architectures to rely on a complex hierarchy of solutions, leading to an ever growing pressure on interconnect bandwidth and power efficiency. This dependency on inefficient, layered data structures results in significant system-level power consumption - not for computation, but for the constant transfer and buffering of data between different memory types. This significantly increases the energy footprint of AI and the need for costly, advanced infrastructures - including HBMs, high-speed interconnects, and intensive cooling - which results in alarmingly high costs for AI data centers. Consequently, these extreme technical and economic barriers have concentrated the development and deployment of this technology in the hands of a few major market cloud providers and AI companies.

03.
Industry's Attempted Solutions and Their Limitations

These new data-intensive workloads require a significant leap in both memory density and performance, all while keeping power consumption manageable. This demand exposes a growing gap in the traditional memory hierarchy, one that cannot be filled by simply continuing the slow, incremental scaling of existing memory technologies.

To address this, the industry has pursued a number of different approaches over the past decade, each attempting to push the boundaries of what is possible within the existing technological framework. This section will review the current industrial strategies to solve this memory challenge and highlight their inherent limitations.

3.1

The Limits of 2D Scaling

For decades, the continuous improvement in memory performance and cost was driven by shrinking the x and y dimensions of semiconductor components (i.e. 2D scaling driven by lithography). For SRAM, this involved shrinking the six-transistor (6T) cell with each new technology node, directly increasing density and performance. Similarly, DRAM scaling was initially achieved by miniaturizing both the access transistor and the capacitor (1T/1C) of the DRAM memory cell. This cycle held for years providing a predictable and exponential improvement in cost-per-bit and density.

However, this era of 2D scaling is now reaching its physical limits. For SRAM, the 6T cell has become so small that further reductions jeopardize its stability and performance. For DRAM, the capacitor shrank to a point where it could no longer reliably hold a charge, forcing engineers to develop a complex, vertical capacitor structure. This, in turn, has halted further area scaling and caused cost-per-bit improvements to stagnate.

DRAM Bit Cost Evolution

This chart visually demonstrates the plateauing of DRAM cost-scaling, highlighting how the increasing complexity of 2D scaling has halted the historical exponential reduction in cost per bit.

This critical trend underscores the urgent need for new architectural approaches to improve cost per bit.

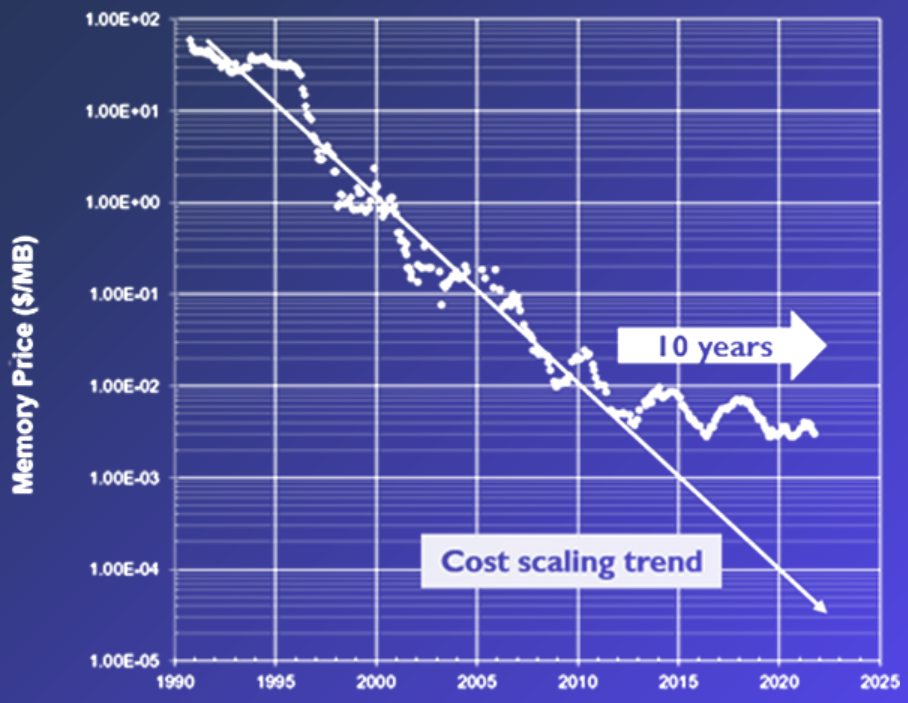

Figure 3 – DRAM bit cost as tracked by John MacCallum based on spot price of components.

3.2

The New AI Hierarchy through 2.5D and 3D Packaging

While this AI evolution was rapidly unfolding, the industry responded by attempting to address these challenges with a mix of 2D, 2.5D, and even 3D memory solutions that are in essence improvements of the traditional memory architectures. This new hierarchy still faces a trade-off, but now for AI workloads.

3.2.1

High Bandwidth Memory (HBM)

HBM is an important memory integration step that has emerged over the last decade as an adaptation of conventional DRAM technology. It represents a significant improvement in memory engineering by implementing a 3D stacking technology where several layers of DRAM dies have been placed on top of each other (i.e. up to 12 layers in some generations). These layers are interconnected by microscopic vertical wires called Through-Silicon Vias (TSV), which allow for ultra-fast signal transmission. Moreover, it allows the HBM memory to be placed directly adjacent to the processor unit (xPU)[4,5].

Such an implementation allows for a much higher number of memory chips to be packed into a smaller space and consequently, reduces the physical distance data must travel between the memory and the processor, resulting in a reduced power consumption. As a result, one of the most important defining features of HBM is its massive memory bandwidth. While early versions delivered over 400GB/s, current standards like HBM3E are already at 1.2TB/s[6], and the forthcoming HBM4 standard is set to push this beyond 2TB/s[7]. This scaling is typically combined with increasing the capacity at the same time. Unlike traditional memory types, HBM achieves this by utilizing a wider memory bus and operating across a larger number of independent channels. This allows for simultaneous data access and significantly expands the amount of data that can be transferred per cycle.

However, DRAM transistors are fundamentally lower performing than the logic transistors used in a CPU or GPU. This makes them unsuitable for directly driving the high-speed communication required by the HBM protocol. To overcome this, a high-performance logic die is added and typically located at the bottom of the stack of memory dies. This logic die, usually manufactured on a more advanced process node (e.g. < N7), acts as the critical interface. It handles the high-speed HBM protocol, manages data flow, and drives the slower DRAM dies above it, essentially acting as a high-speed controller that enables the entire stack to communicate efficiently with the host processor.

[4] S. Ramalingam, "HBM package integration: Technology trends, challenges and applications," *IEEE Hot Chips 28 Symposium (HCS)*, 2016.

[5] C. Y. Lee et al., "3D Integrated Process and Hybrid Bonding of High Bandwidth Memory (HBM)," *Electronics, Magnetics and Photonics*, Springer, 2025.

[6] Micron Technology, Inc., "High-Bandwidth Memory HBM3E." [Online]. Available: https://www.micron.com/products/memory/hbm/hbm3e

[7] JEDEC Solid State Technology Association, "High Bandwidth Memory (HBM4) DRAM Standard." [Online]. Available: https://www.google.com/search?q=https://www.jedec.org/standards-documents/

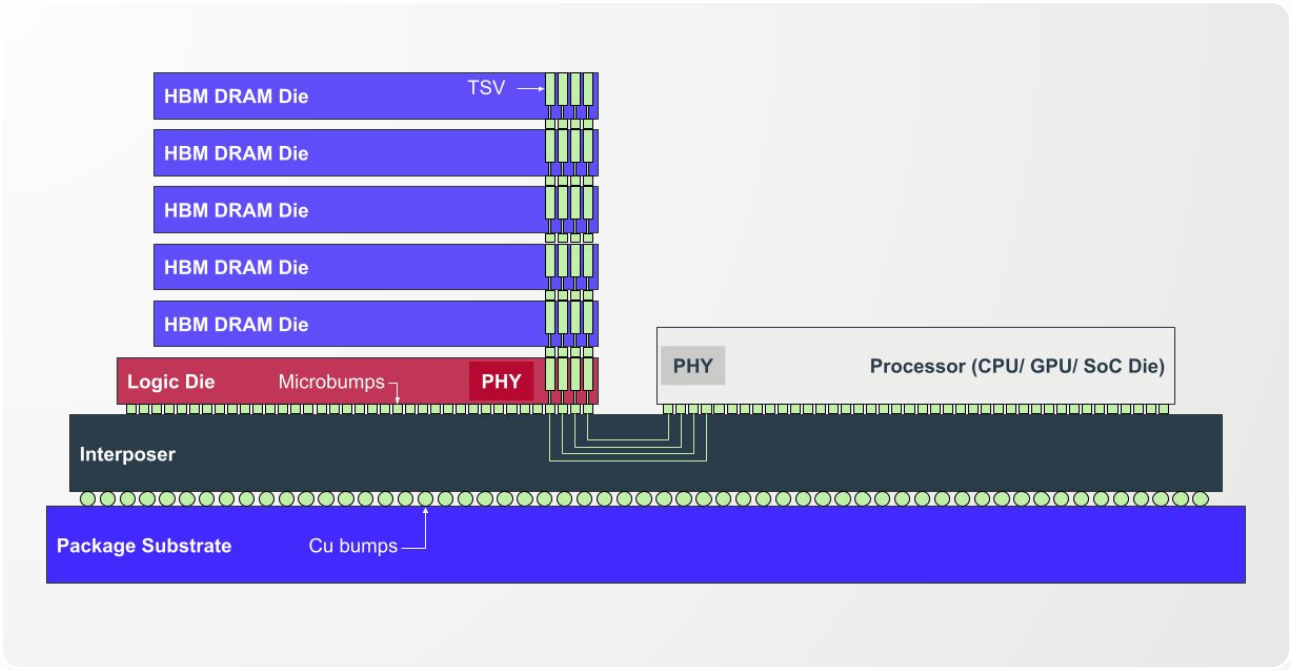

Figure 4 – A cross-section of a HBM module. The diagram illustrates the 3D stacking of DRAM dies on top of a base logic die. This entire memory stack is then connected to a processor (xPU) via an interposer in a 2.5D package configuration.

This elaborate architecture enables a high-performance product that fills the gap between SRAM and DRAM, providing tens of gigabytes of memory near the core with bandwidths up to terabytes per second. This makes HBM the workhorse of AI training and LLM inference in large data centers. However, this is a highly-engineered and expensive solution. The cost is driven by its complex manufacturing process, which involves 3D stacking of many non-cost-scaling DRAM dies, the addition of a high-performance logic die, and numerous extra manufacturing steps that are no longer performed at the wafer level (such as die-to-die bonding and advanced packaging techniques to interconnect the different DRAM dies).

Beyond its manufacturing cost, HBM-based systems also face significant power and thermal hurdles. The TB/s data rates demanded by large AI models still result in a significant power consumption as a result of the continuous memory transfers alone. This power becomes a major source of heat which creates a system-wide thermal challenge. Consequently, these systems require sophisticated and expensive cooling infrastructure, limiting them to specialized data centers.

The high price tag of HBM therefore makes it a premium product, primarily used by large data centers to handle the most demanding AI workloads.

Furthermore, this technology cannot be cost-effectively transferred to consumer products, meaning a lasting reliance on data centers for the more advanced AI inference processing activities. This in turn limits the applicability of real-time AI for critical or comfort-driven interactions due to the inherent latency of network communication. In summary, while HBM is an essential technology that will continue to dominate the data center, the use of HBM-heavy AI compute systems cannot be a long-term solution for client-side inference workloads.

3.2.2

AMD 3D V-Cache

A more recent engineering solution to boost on-chip SRAM capacities was spearheaded by AMD with their 3D V-Cache technology. This is a chiplet-based approach where a separate die, densely packed with SRAM cells, is directly stacked on top of the main CPU (xPU) core using a high-density, copper to copper hybrid bonding technique and through silicon vias (TSVs). Essentially, 3D V-Cache is a larger, secondary L3 cache that sits right on top of the CPU cores. By giving processors access to such a large bank of additional temporary storage, the processor can store much more information than it would normally have immediate access to. While originally developed for gaming applications to improve frame rates, this technology is now also finding its way into data center and high-performance computing (HPC) applications[8].

However, its impact remains limited: with capacities around ~100 MB, it addresses only a fraction of the gap exposed by large AI models. In addition, the technology is expensive and power consuming, limiting its scalability beyond premium products.

Solutions like HBM and 3D V-Cache highlight the industry's immense engineering efforts to go around the memory wall bottleneck. However, the industry is still presented with a core inefficiency: there is currently no single chip solution capable of handling GenAI model inference at a reasonable cost.

[8] AMD, "Elevating Data Center Computing with AMD 3D V-Cache™ Technology." [Online]. Available: https://www.google.com/search?q=https://www.amd.com/en/technologies/3d-v-cache

A fundamental trade-off between capacity, speed, and cost still exists. Table 1, summarizes these divergent memory solutions for various AI workloads.

AI Workload	AI Inference	Memory Type	Challenge
AI-vision (CNNs, DNNs)	Edge	SRAM	Not suitable for GenAI due to low density
GenAI (LLMs)	Cloud	HBM	Too expensive to be deployed at the edge
AI Training	Cloud	HBM, 3D V-Cache	High data volume and throughput requirements

Table 1 – Main memory types used nowadays for different AI workloads and their shortcomings.

3.3

The Push Towards 3D Memory Architectures

While 2D scaling and advanced 2.5D packaging have enabled incredible performance gains, the relentless growth of AI workloads is also pushing these technologies to their practical limits. To continue this momentum, the industry is now looking beyond incremental improvements and toward the next architectural evolution: the development of a true monolithic 3D memory.

A decade ago, the flash memory industry provided a powerful blueprint for such a breakthrough with the introduction of 3D-NAND. For decades, flash memory relied on planar, 2D scaling, similar to SRAM and DRAM, where engineers continuously shrank the memory cell size to pack more bits onto a single die. This process, however, also reached its scaling limit.

This was followed by the revolutionary shift towards a vertical design, more commonly known as 3D-NAND. Instead of shrinking cells horizontally, the architecture moved to stacking them vertically in a "tower" structure. The key enabling feature was the simultaneous patterning of a single vertical via through an oxide-nitride matrix to connect hundreds of horizontal gates at once, which dramatically increased the overall density and significantly reduced the bit cost.

Realizing a similar structure/ feature with SRAM technology seems not possible. The conventional 6T SRAM cell requires a high number of connections which makes 3D routing impractical. By contrast, the DRAM cell has only a few connections and therefore possesses the possibility to be connected in a truly 3D manner.

However, despite various efforts already, the area (and cost) gain would be relatively limited compared to the most advanced 2D-DRAM nodes. The vertical capacitors currently used in the DRAM (2D) cells are quite tall (i.e. ~1 µm) and have to be placed horizontally, limiting the vertical stacking gains due to a poor x-y bitcell area.

Another option being explored by the industry to follow a 3D-NAND-like path for DRAM is to stack several arrays of the convention 2D-DRAM memory technology on top of each other and connect them as a crossbar array as depicted in Figure 5. Such an approach would require a select transistor to be directly integrated in a pillar fashion[9].

This is typically referred to as moving from a $6F^2$ (current) to a $4F^2$ (pillar/crossbar) cell architecture, which provides roughly a 50% or 1.5x gain. Moreover, this approach would not provide a breakthrough cost scaling like 3D-NAND, as the 2D-array modules would have to be repeated in the manufacturing flow, adding significant cost and complexity.

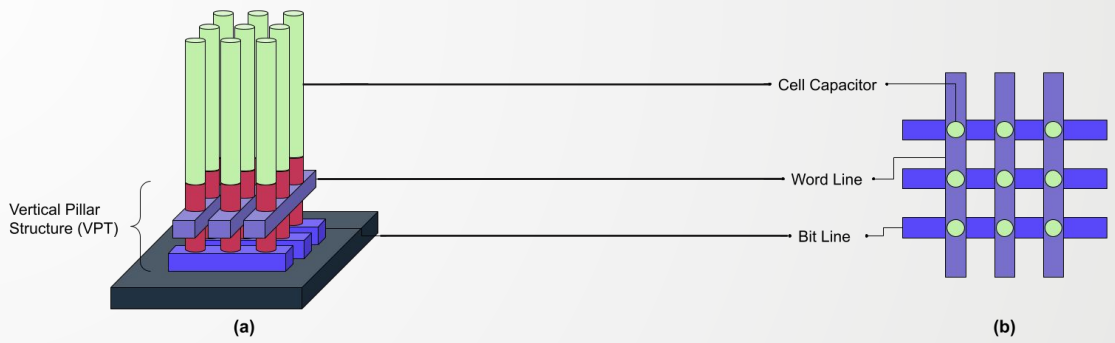

Figure 5: – (a) Schematic diagram of VPT $4F^2$ cell array, and (b) a top view of the $4F^2$ cell layout[9].

While $4F^2$ and 3D-DRAM solutions are currently being pursued by the industry, they are expected to enable only a continuation of the DRAM density scaling roadmap with potentially reasonable cost savings (still to be confirmed once produced). However, none of these options enable the truly '3D-NAND'-like wafer cost structure that will be required by AI compute and other large workload applications in the next decade.

[9] H. Chung et al., "Novel 4F2 DRAM cell with Vertical Pillar Transistor (VPT).", *Proceedings of the European Solid-State Device Research Conference (ESSDERC),* 2011.

3.4

Alternative Emerging Memory Technologies

3D integration solutions using conventional memory technologies have not yet resolved the memory bottleneck. Therefore, the industry and academia have been researching and developing a variety of alternative memory technologies for some time, with varying levels of maturity. However, no single existing technology has been capable of combining the necessary attributes so far to truly address the needs of modern (and future) AI applications.

To enable the next breakthrough for AI application workloads, a new memory solution must combine the following critical aspects: high speed (≤10 ns latency), high capacity (GBs to TBs), extremely high endurance, and the lowest possible cost and active power.

In a first place, most alternative memory technologies that were originally developed for embedded-flash replacement in microcontroller applications - such as PCM, ReRAM, and Spin-Transfer Torque MRAM (STT-MRAM) - do not meet the endurance criteria required for AI. This is because embedded flash memory is primarily used for firmware storage, which involves infrequent read/ write cycles. In contrast, AI workloads often require billions of read/ write cycles per second, similar to the demands placed on an L3 cache or HBM memory. These emerging technologies, while offering non-volatility, are not built for the extremely high-cycling applications of high-performance AI, which require endurance levels closer to DRAM or SRAM.

In a second place, another class of memory technologies is being explored that could potentially meet the high-endurance and speed requirements for AI applications. However, even in their early stages of development, they face significant architectural, scaling, and manufacturing challenges that may limit their ability to deliver the cost and density breakthroughs necessary to solve the memory bottleneck.

SOT-MRAM is a strong contender for tackling the memory bottleneck, offering write times as low as 1 ns, making it a promising alternative to SRAM technology. Its key building block is a magnetic tunnel junction (MTJ), which consists of a thin dielectric layer (typically MgO) sandwiched between two ferromagnetic layers (often CoFeB-based). One of these ferromagnetic layers (i.e. the fixed layer) has a fixed magnetization, while the other's magnetization is free (i.e. the free layer) to switch, representing a stored bit (“0” or “1”).

The memory cell is read by applying a current through the MTJ and measuring the junction's Tunnel Magnetoresistance (TMR).

This resistance is either high or low depending on whether the magnetization of the free layer is parallel or antiparallel to the fixed layer.

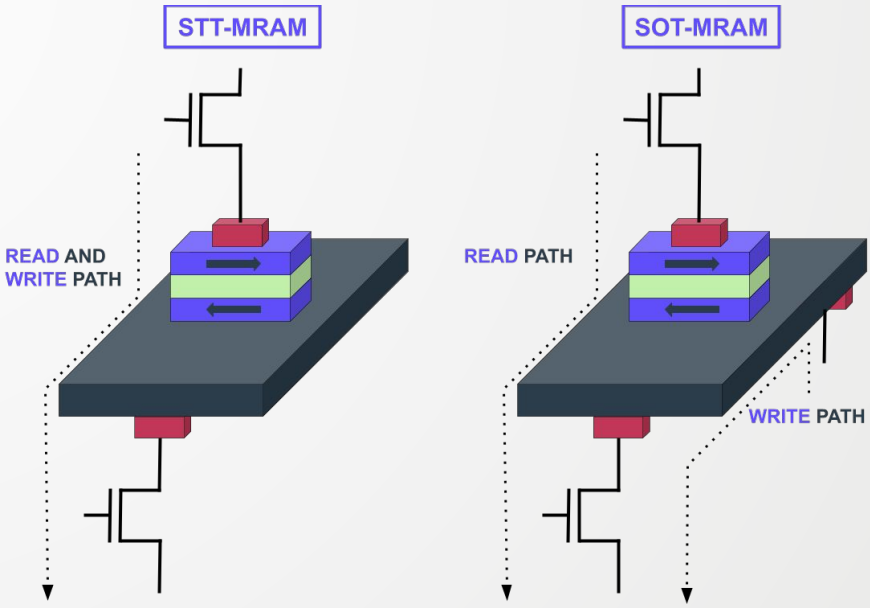

Figure 6 – STT-MRAM and SOT-MRAM bit cell architecture.

What truly sets SOT-MRAM apart is its write operation, which is fundamentally different from its predecessor, STT-MRAM. In an SOT-MRAM device, the bit is written by injecting a lateral, in-plane current into an adjacent SOT layer (typically a heavy metal like tungsten) positioned underneath the MTJ. This current generates a spin-polarized current that switches the magnetization of the free layer. This key innovation decouples the read and write paths, ensuring reliable operation and avoiding a high current tunneling perpendicularly through the delicate MTJ barrier, which limited the reliability and endurance of STT-MRAM. Furthermore, this decoupled architecture makes the probability of a "read disturb" - where a read operation accidentally flips the bit - much lower.

This separation allows SOT-MRAM to offer an unlimited endurance, making it suitable for high-cycling applications like SRAM. The manufacturing capabilities for such devices are also available with leading Tier-1 foundries, making the technology potentially mass-producible. However, the limited area-scaling gains (a maximum of 1.5-2x versus SRAM) make it difficult to predict whether the industry will adopt it at a pace fast enough to disrupt the market.

2T0C DRAM bit cells, typically also referred to as gain cell or capacitor-less DRAM, is another new memory technology attempting to tackle the memory bottleneck. This cell architecture functions like a DRAM but has essentially redesigned its core functionality. It is composed of two thin-film transistors (2T, one for read, one for write) and no capacitor (0T), storing the charge on the gate of one of the transistors[10]. Such a design makes it highly attractive for high density applications thanks to the elimination of the large capacitor that was constraining further DRAM technology scaling.

However, in pure silicon, this design is limited by very short storage times due to gate leakage, a fundamental issue that has prevented its successful implementation. To overcome this, current research focuses on replacing silicon with a lower-leakage material. Oxide semiconductors, such as Indium-Gallium-Zinc-Oxide (IGZO), are a promising candidate due to their wide bandgap, which results in an extremely low off-current and dramatically improves the memory retention time, refresh rate, and power consumption[11].

The low-temperature process compatibility of IGZO with back-end-of-line (BEOL) fabrication is a key enabler for new DRAM architectures. It allows for the stacking of memory arrays directly on top of the logic circuits, which reduces the overall chip footprint. It also paves the way for monolithic 3D stacked configurations, where transistors can be monolithically integrated into vertical plugs in a manner similar to 3D NAND. While this opens a new path for aggressive DRAM scaling, long-term reliability and endurance remain key challenges, with the material itself being highly sensitive to defects. While significant progress has been made in understanding and modeling these issues, the path to an industry-viable, high-volume product is still being paved.

Ferroelectric Memories (FeRAMs) are another class of memory technology being explored. They rely on ferroelectric materials that possess a permanent electric dipole (i.e. a separation of positive and negative charges), which can be switched with an applied electric field. This allows them to store information in a non-volatile manner using voltage rather than current. This voltage-driven operation offers the potential for a significant lower power compared to other memory technologies.

This new memory technology is mainly attracting attention in two different types of implementations, each with its own advantages and limitations:

[10] A. Belmonte et al., "Capacitor-less, long-retention (>400s) DRAM cell paving the way towards low-power and high-density monolithic 3D DRAM.", *IEEE International Electron Devices Meeting (IEDM)*, 2020.

[11] A. Belmonte et al., "Tailoring IGZO-TFT architecture for capacitorless DRAM, demonstrating >10^3s retention, >10^11 cycles endurance and Lg scalability down to 14nm.", *IEEE International Electron Devices Meeting (IEDM)*, 2021.

FeRAM (as a 1T1F cell):

This architecture consists of a standard transistor (1T) connected to a ferroelectric capacitor (1F). A bit is written by applying a voltage across the capacitor, which creates an electric field that aligns the dipoles in a specific direction (e.g. up for a “1” and down for a “0”). To read the stored bit, a small voltage is applied. The resulting current pulse, which is measured as the dipoles switch, indicates the stored state. This process is typically “destructive”, meaning that the data state is changed during the read operation and therefore, must be rewritten afterwards.

It is important to note, however, that the technology faces challenges in 2D scaling. More precisely, as the capacitor pillar shrinks, its surface area (and thus its stored charge) decreases, which can impact the overall reliability of the bit cell. In addition, the material also exhibits limitations in terms of endurance and fatigue, issues that must be addressed before it can be used as a replacement for high-cycling SRAM applications.

FeFET (Ferroelectric Field-Effect Transistor):

A more advanced architecture finds its inspiration in the 3D-NAND memory architecture. A FeFET implementation uses a transistor where the gate’s insulating layer is replaced with a ferroelectric material. A bit is written by applying a voltage to the gate, which aligns the ferroelectric dipoles and in turn modulates the transistor threshold voltage. This change in threshold voltage allows it to permanently store a bit (e.g. a low threshold for a "1" and a high threshold for a "0"). To read the bit, a small voltage is applied to the gate (similar to FeRAM). The transistor will either turn on or stay off depending on its stored threshold voltage state. However, it must be noted that this read operation is non-destructive in the end, which is a major advantage compared to its FeRAM counterpart. An advanced FeFET can be built using a similar monolithic 3D stacking technique as with 3D-NAND. As such, the ferroelectric material which stores the data can be integrated into a vertical transistor channel or pillar. This allows for many layers of these vertical channels to be stacked one on top of the other, enabling a density gain that is limited by the number of layers, not the horizontal footprint of a single cell. Hence, FeFET has the potential to actually enable a truly high-density memory solution for higher speeds.

Despite breakthroughs in hafnium dioxide (HfO2)-based ferroelectrics enabling low-voltage operation (<1V), the technology's manufacturability, endurance and reliability challenges remain significant.

While a high-density FeFET could fill the cost gap between DRAM and NAND, its limitations would likely prevent it from being a high-performance, high-density memory alternative to SRAM or DRAM.

	3D SRAM silicon based	3D DRAM OSC channel (1T1C)	3D DRAM OSC channel (2T0C)	3D Ferroelectric capacitor (1TnC)	3D Ferroelectric FET & OSC channel	VIM™
Bit density	1× (ref.)	1.5×	2×	2×	8×	20×
Latency [ns]	~50	~50	~5	~20	~1000	~5
Endurance [cycles]	$>10^{15}$	10^{12}	10^{12}	10^{9}	10^{5}	$>10^{15}$
CMOS BEOL compatible	No	No	Yes	Yes	No	Yes
Main challenge	Reliability of stacking Integration challenges	OSC reliability	OSC reliability Limited retention time	Endurance Uniformity	OSC reliability, Endurance Uniformity	New Concept
Key players	Memory IDMs	Memory IDMs Neo-Semiconductor	imec	Intel	Sunrise	Vertical Compute

Table 2 – Overview of active research solutions for 3D memory technologies.

04.
Vertical Integrated Memory (VIM™): A New Paradigm

Having explored the critical limitations of the existing memory landscape, it is clear that incremental improvements and complex packaging solutions are not enough to meet the demands of modern AI. A fundamental architectural shift is required to break the historical trade-offs between speed, density, and cost.

The core promise of **Vertical Compute's groundbreaking VIM™ technology** is to deliver this breakthrough. Our mission is to fundamentally change the architecture of compute systems by proposing a new memory solution that combines the speed and endurance of SRAM with the high density and cost-efficiency of 3D-NAND.

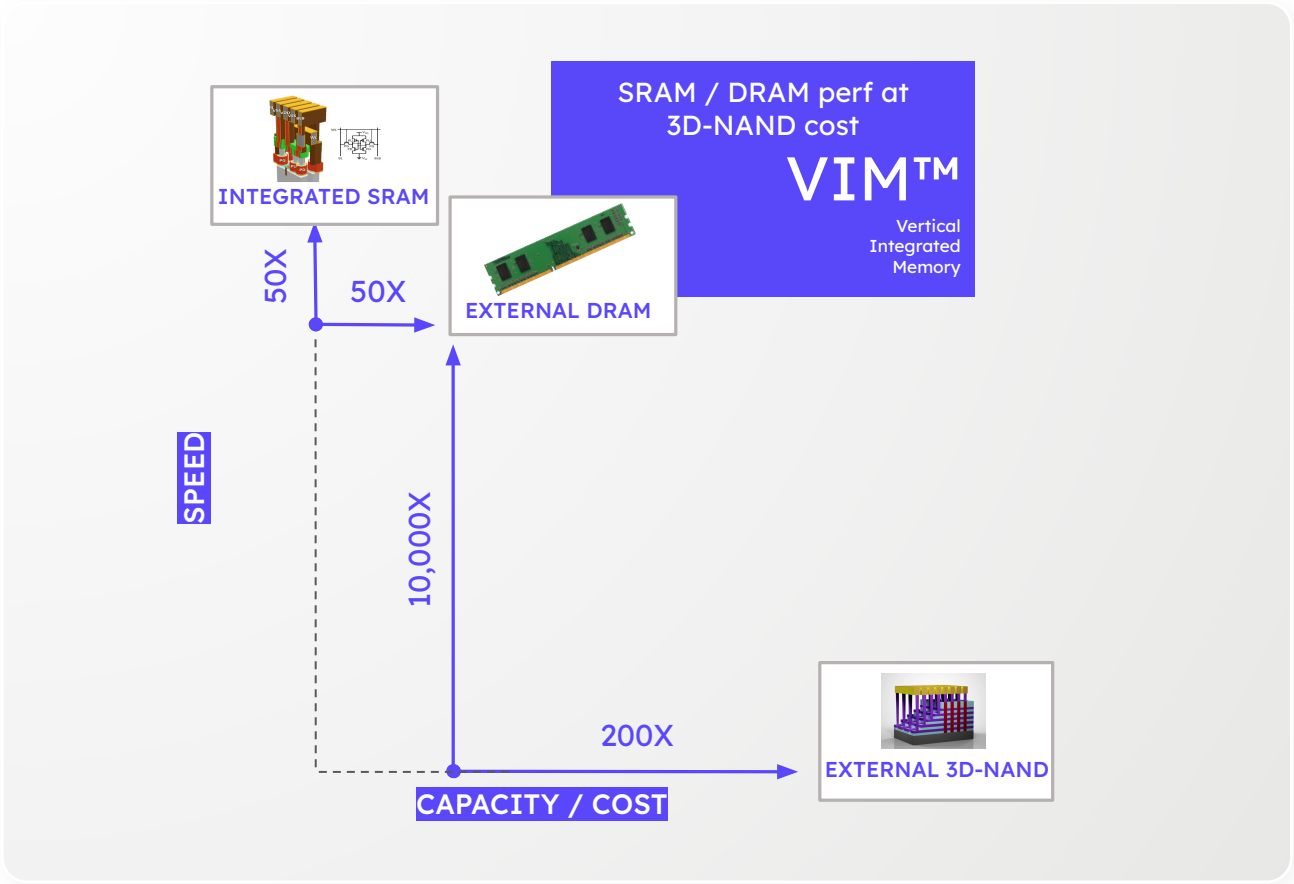

Figure 7 – VIM™ to break the gap between speed and density/cost.

The approach of Vertical Compute involves a vertical memory architecture that allows memory cells to be stacked directly above processing units in the CMOS back-end. Such a design allows to reduce the physical distance between computation and memory, thereby cutting down on data transfer times and energy consumption at the same time.

To achieve this, VIM™ merges the best principles of three innovative worlds. The core is a vertical domain wall memory device that stores bits magnetically. This can be done very fast, with speed similar to SRAM, and with high retention times, i.e. the memory cell is non-volatile. The VIM™ structure can be fabricated with very similar technologies that are currently mass produced. Without the need for a staircase connection to each individual domain, the overall array architecture and connectivity can be made drastically simpler, cheaper and denser. The second portion consists of a domain wall injector line to write and store bits in the vertical structure. Finally, conventional MTJ technology is used for providing the read out state. A key feature of the VIM™ technology is to use the MTJ only for read-out at (very) low currents, enabling an effective unlimited read and write endurance. The manufacturing capabilities for such devices are available with Tier-1 leading foundries, making this structure capable of transitioning in a fast manner to high volume manufacturing.

The unique combination of these techniques makes a radical revolution possible in memory devices. By leveraging commercial semiconductor processes, the cost of our VIM™ device will be very low (similar to the cost structure of 3D-NAND memory). As such, the resulting VIM™ technology is capable of redefining the memory hierarchy and unlocking new possibilities that were previously impossible. This includes enabling truly local on-chip, always-on, private AI assistants, reducing the enormous cost and power consumption of data centers, and paving the way for a more powerful and accessible next generation of computing. The expected outcome of Vertical Compute's proprietary technology has been compared to the state-of-the-art in Table 3 below.

	SRAM	DRAM	HBM-DRAM	3D-NAND	VIM™
Density [Gb/mm2]	0.01-0.02	0.5	5-10	20-30	1-30
Latency [ns]	~1	~50-100	~10	~100,000	<10
Bandwidth [GB/s]	500+	~100	1000+	~1	2000+
Density/cost	Very low	Medium	Low	High	High
Endurance [cycles]	10^{15}	10^{15}	10^{15}	10^{3}-10^{5}	10^{15}
Energy efficiency [pJ/bit]	~0.1-1	~1-5	~1-10	~10-100	0.2

Table 3 – Overview of competing memory technologies with respect to high-performance compute.

05.
VIM™: The Revolutionary Design and Working Principle

The VIM™ architecture, as briefly introduced in the previous section, is depicted in Figure 8.

Bits are stored in the heart of the memory cell, a vertical magnetic wire. Vertical Compute has invented an efficient way to write, order and move magnetic domains (representing 0's and 1's) within the vertical wire with high vertical density (vertical spacing <50 nm). Storing many bits vertically requires manufacturing multiple layers on top of each other, and having the ability to insert, shift and read magnetic field packages. The concept of vertical magnetic wires isn't new. For example, the idea has been proposed by IBM in 2008 by S. Parkin. However, Vertical Compute's breakthrough architecture is fundamentally different from previous solutions and, for the first time, enables mass-volume manufacturing. The concept has been invented by Vertical Compute's founder and CTO - Sebastien Couet - and has been protected through a broad IP portfolio.

The read and write operations are performed through two other integrated components: the SOT and MTJ structures. First, the SOT layer, borrowed from the most promising MRAM architectures, provides a high-speed write mechanism and is positioned on top of the vertical magnetic wire. Second, the MTJ structure is positioned on top of the SOT layer (i.e. atop the VIM™ device) and is responsible for the read operation specifically.

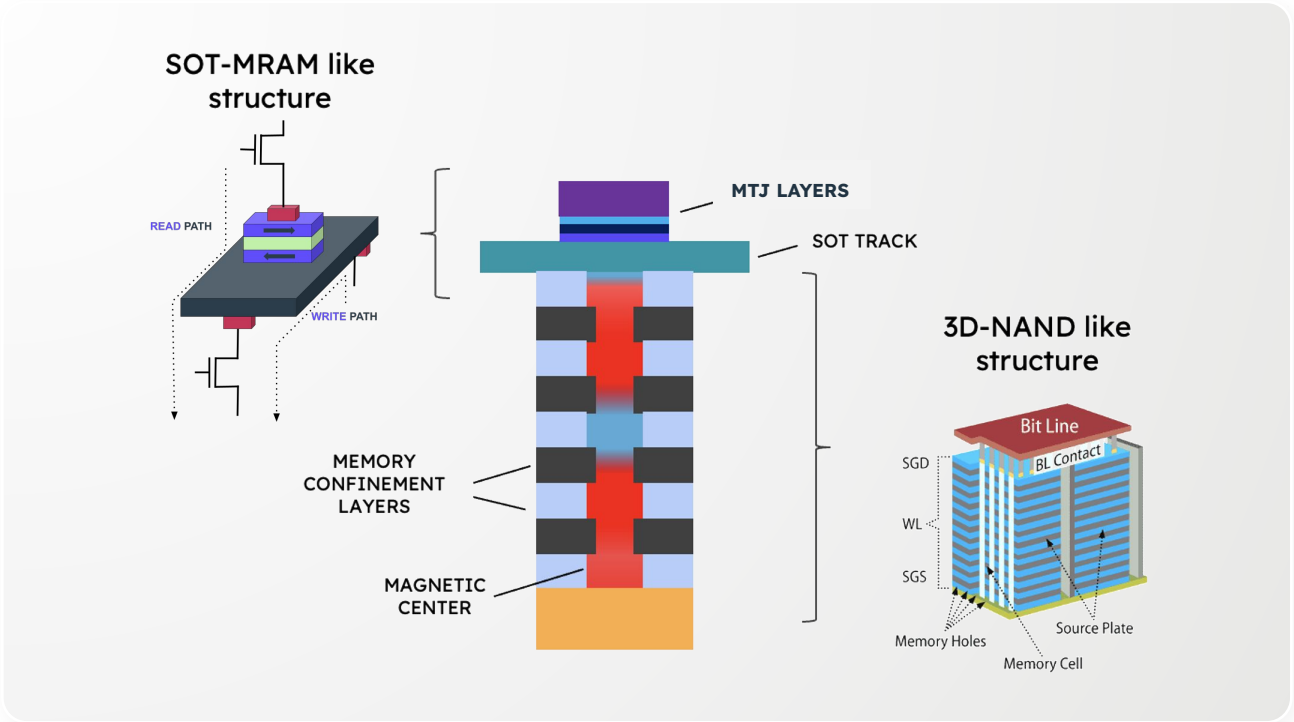

Figure 8 – Vertical Compute's VIM™ architecture.

Operating Vertical Compute's VIM™ Architecture

The operational principle of a VIM™ device can be compared to a chain of magnetic toy cars, as shown in Figure 9. Each car represents a magnetic domain (i.e. a storage bit), with red and blue colors indicating opposite magnetic polarities. When a new car is pushed from the top of the wire downward, the entire chain shifts downward, with each car locking into place through the pinning layers.

Bit select operation - A new bit to be written into the wire can be prepared by sending a current pulse through the SOT layer, which sets the magnetic orientation of the topmost domain.

Push & store - Next, a vertical current pulse is sent through the magnetic wire to push the entire stack of magnetic domains (~magnetic cars) down by one position. This also makes room for the next bit to be written.

Read Operation - Finally, the state of the bit at the top of the stack can be read through the MTJ structure, using a sensing mechanism identical to that of conventional MRAM. To read a bit further down the stack, the device simply performs a series of push operations until the desired bit reaches the top.

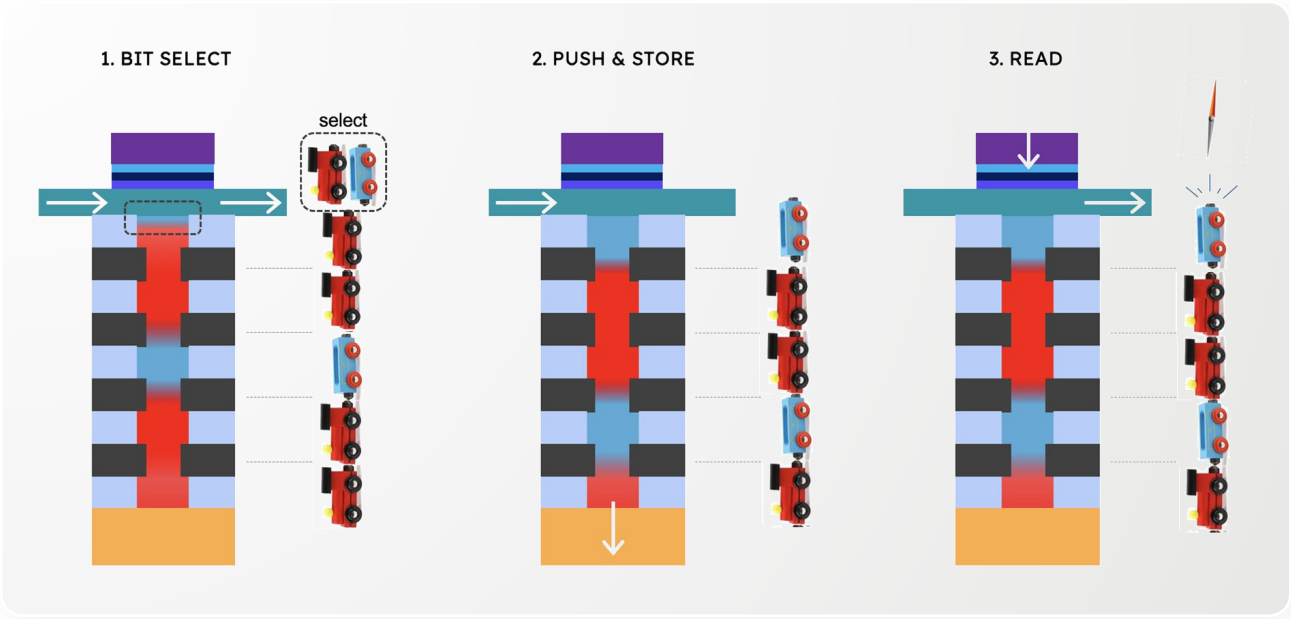

Figure 9 – Working principle of the VIM™. White arrows indicate a current flow to operate the device.

06.
VIM™: Reshaping the Memory Hierarchy

Leveraging this groundbreaking technology, Vertical Compute aims to deliver high-density, low-cost, and high-speed memory solutions tailored for advanced computing systems. These solutions will address critical challenges across multiple domains: reducing the enormous cost and power consumption in data centers, enabling LLM execution at the edge on devices like laptops and IoT systems, and integrating VIM™ directly into AI-PUs for optimal performance.

Like other memory technologies, continuous innovation and a structured roadmap will drive progressive advancements. By enabling a vertical scaling path similar to 3D-NAND, VIM™ is set to offer a sustainable, long-term density roadmap that will be capable of driving the evolution of next-generation compute systems.

In its initial two generations, Vertical Compute is set to challenge the performance of both SRAM and DRAM. VIM™ will offer up to 100× higher memory density at a cost more than 10× lower compared to SRAM, all while maintaining competitive performance across all key specifications.

In the next phase, Vertical Compute aims to expand into a broader market by advancing the technology further. By adopting more advanced process nodes and a more mature VIM™ technology, we expect another 10× increase in density, making it a strong competitor to DRAM. With an estimated 40× lower equivalent cost than DRAM when the roadmap fully materialises, this breakthrough will open the door for adoption in high-performance computing (HPC) systems and the wider DRAM market.

07.
Conclusion

The current memory landscape, built on decades of incremental improvements to SRAM, DRAM, and NAND, has reached a critical bottleneck in the face of modern AI. The demands of LLMs and other data-intensive applications have exposed a fundamental flaw in the traditional memory hierarchy, forcing the industry to rely on a complex, expensive, and power-hungry mix of technologies. From the scaling limits of 2D memory to the highly-engineered solutions like HBM, the market is urgently seeking a disruptive architectural shift that can finally break the historical trade-offs between performance, density, and cost.

Vertical Compute's VIM™ technology offers a clear path forward by uniquely combining the best aspects of three distinct memory paradigms into a single, cohesive solution. By leveraging a high-density, vertical architecture with the high-speed and unlimited endurance of SOT-MRAM, VIM™ is set to deliver a game-changing solution that achieves the speed of SRAM and the density of 3D-NAND at an unprecedented low cost. This innovation is not just another incremental step, it represents a fundamental rethinking of memory design that will redefine the possibilities of next-generation compute systems.

The successful implementation of VIM™ technology promises to unlock new capabilities across the entire computing ecosystem. By enabling high-capacity, high-performance, and non-volatile memory directly on-chip, VIM™ will empower local, always-on AI assistants, dramatically reduce the energy footprint of data centers, and lower the barriers to entry for advanced AI. As we embark on this ambitious roadmap, Vertical Compute is set to lead the charge, turning a theoretical breakthrough into a commercial reality and paving the way for a future where powerful and accessible computing is no longer constrained by the limitations of memory.