243. "RAMMAGEDDON" IN THE AI ERA: HOW THE AI DATA CENTER BOOM REPRICED MEMORY AND TRIGGERED COMPONENT SCARCITY

Ermachenok S. A.
Belarusian State University of Informatics and Radioelectronics
Minsk, Republic of Belarus

Yeliseyeva V. A. – Senior Lecturer, Master of Arts

This paper examines how rapid growth of artificial intelligence is reshaping the global semiconductor market, particularly the memory sector. The expansion of AI data centers has led to increased demand, resulting in constrained supply and higher prices in both corporate and consumer technology markets.

The artificial intelligence (AI) ecosystem is moving from experimentation to industrial-scale implementation. Modern models require massive datasets, repeated training cycles, and continuous inference for consumer and enterprise services. The physical foundation of this growth is data center: a facility that stores data, moves it at high speeds, and performs large-scale calculations. Data centers existed long before generative AI, but AI has become a new accelerator for the growth of the data center market and a driving force for profound changes in server design.

AI-oriented data centers differ from traditional enterprise facilities in one decisive way: they are based on accelerated computing and unusually high memory demand. This is currently leading to component shortages and higher prices across the electronics economy. Technology media outlets have branded the dynamic random access memory (DRAM) and NAND memory shortages, which have been ongoing since 2024, as "RAMpocalypse" or "RAMmageddon" [1].

An AI-oriented server consists of components such as central processing units (CPUs), graphics processing units (GPUs), specialized AI accelerators, double data rate (DDR) RAM, high-bandwidth memory (HBM), NAND storage, high-speed interconnects, system mainboards, and high-capacity power supply units. Among these elements, memory technologies (particularly DDR, HBM, and NAND) have become a key bottleneck, acting as the primary channel through which AI-driven demand is causing broader market shortages and consumer price inflation [2].

AI workloads are increasingly memory-bound. Large models repeatedly stream parameters and activations between computing units and memory. When bandwidth is limited, accelerators idle and system-level efficiency drops. Consequently, HBM, a multi-layer form of DRAM designed for extreme bandwidth, has become a strategic component of AI clusters.

HBM technology is difficult to scale due to its packaging complexity. It consists of vertically stacked DRAM chips called dies connected by through-silicon vias, with modern standards increasing stack height and creating greater sensitivity to temperature and performance issues.

At the same time, conventional memory is increasingly integrated into AI infrastructure. Host DDR5 is needed for data pipelines, orchestration, and feeding accelerators. However, AI servers require significantly more conventional DDR RAM per system than consumer devices, allowing hyperscalers to absorb a disproportionate amount of the DRAM and NAND memory. Demand is relatively inelastic, as AI companies prioritize rapid deployment over component cost, aggressively securing supply. Capacity is increasingly being reallocated towards high-margin AI-grade memory such as HBM, which consumes more die resources than standard DDR, further tightening supply for personal computers (PCs), smartphones, and other mainstream products.

This reallocation has macro-scale consequences. By 2024, global memory revenue surged over 70 % year-on-year to about US $185 billion, driven by insatiable hyperscale demand that has turned HBM and server DRAM into strategic assets [2]. International Data Corporation (IDC) further notes that, for example, the cost structure of a smartphone is heavily dependent on the memory used. For a mid-range device, memory can represent 15–20 % of the total bill of materials. As memory prices continue to surge, manufacturers will likely have to raise prices significantly, cut specifications or both.

Downstream effects are already reflected in market forecasts. According to IDC, DRAM and NAND supply growth in 2026 is expected to remain below historical averages (approximately 16–17 % year-over-year), with projected contraction scenarios for both the PC and smartphone markets. The PC market is expected to decline by approximately 5–9 %, accompanied by rising average selling prices, while smartphone volumes may decrease by around 3–5 %, particularly in lower-margin segments. At the same time, suppliers are signaling memory price increases of 15–20 % and renegotiation of contracts, indicating a broad industry-wide adjustment to structurally tighter supply conditions [3].

Numerous indicators show that the current crisis is not an isolated incident. Reuters, citing TrendForce, reported in December 2025 that prices in some memory segments more than doubled since February of that year. The same report indicates that DRAM suppliers' average inventory levels have decreased to 2–4 weeks in October 2025 (compared to 13–17 weeks in late 2024), indicating rationing rather than a normal cyclical pattern [4].

The supply outlook remains constrained. Reuters reports that Samsung raised prices of server memory chips by up to 60 % in a single month, and that SK Hynix said its 2026 output was already sold out. Citi research cited by Reuters expects the shortage to persist through late 2027. One of the reasons is the slow supply response: conventional capacity expansions for commodity memory are not expected until 2027–2028. To illustrate the scale of the demand shock, Reuters cited an estimate that a single mega-project could require up to 900,000 dies per month by 2029. This amount is approximately double the current global monthly HBM production [4].

The impact of the ongoing memory shortage is also becoming visible in the gaming console segment, which traditionally relies on high-performance unified memory architectures. As reported by The Verge, manufacturers face increasing pressure as modern consoles depend on large pools of fast memory shared between CPU and GPU, making them particularly sensitive to DRAM price fluctuations [1]. In this context, rising memory costs can directly influence production volumes, retail pricing, and even hardware design decisions, potentially delaying upgrades or limiting availability for end users. This further illustrates that the effects of AI-driven demand are spreading across the entire consumer electronics ecosystem, including entertainment hardware.

These effects are extending beyond computing. Counterpoint's analysis, as reported by The Register, indicates that the amount of memory used in routers and set-top boxes rose almost sevenfold over nine months and now represents more than 20 % of the total in some low-to-mid-range routers (compared to around 3 % a year earlier) [5]. This illustrates a broader point: when AI data centers consume a significant amount of memory, even "inconspicuous" infrastructure hardware becomes more expensive, potentially slowing down broadband rollouts.

In early 2026, major PC manufacturers acknowledged that rising memory costs were already affecting shipments and pricing, passing the burden directly to consumers. The core structural issue is that semiconductor supply responds slowly: Institute of Electrical and Electronics Engineers (IEEE) Spectrum notes that it can take 18 months or more to build and expand a memory factory, and new capacity often arrives after demand has already surged. In the current cycle, additional investments are mainly directed towards HBM and AI-optimized DRAM rather than conventional memory, limiting relief for mainstream markets [6].

At the same time, demand projections remain aggressive. IEEE Spectrum reports forecasts that the HBM market could grow from US$35 billion in 2025 to US$100 billion by 2028, which is described as larger than the entire DRAM market in 2024. Major suppliers' plans include significant production capacity expansions for 2027–2028, Micron targets new HBM capacity in 2027, Samsung's additional production is associated with 2028 timelines, and SK Hynix expansions extend into 2027-2028. Reflecting this, Intel's CEO was quoted as saying: "There's no relief until 2028".

Short-term mitigation is likely to be achieved through incremental capacity expansions by existing DRAM leaders, yield improvements in advanced packaging and a broader diversification of supply chains rather than through building entirely new factories [6]. On the demand side, memory optimization techniques (such as quantization, flash attention, data parallelism) can also reduce memory intensity [7]. However, companies with stronger financial positions and long-term supply agreements remain better insulated from market volatility, while smaller manufacturers face greater volatility, highlighting the unequal distribution of the crisis's impact.

To summarize, AI has transformed memory from a low-cost commodity into a strategic bottleneck. Data from 2024–2026 show a sharp increase in prices, a decrease in inventory, and multi-year forecasts for supply reduction, with consequences for PCs, smartphones, and even broadband infrastructure. The main mechanism is a structural reallocation of wafer and packaging capacity towards HBM and server-grade products, effectively imposing an "AI tax" on consumer electronics. The most credible path to stabilization combines incremental capacity expansion, packaging yield improvements, and memory-efficient AI practices, acknowledging that the transition may redefine the economics of digital devices for several years.

References:

1. The Verge. The RAM shortage is coming for everything you care about – 2026. – URL: https://www.theverge.com/tech/880812/ramageddon-ram-shortage-memory-crisis-price-2026-phones-laptops (date of access: 27.03.2026).

2. Zhang Z., Li W. MEMORY IS THE NEW SILICON GOLD: From Memory Shortages, Price Surges to Technologies in AI Era. – ResearchGate, 2025. – URL: https://www.researchgate.net/publication/399226623_MEMORY_IS_THE_NEW_SILICON_GOLD_From_Memory_Shortages_Price_Surges_to_Technologies_in_AI_Era (date of access: 27.03.2026).

3. IDC. Global memory shortage crisis: market analysis and the potential impact on the smartphone and PC markets in 2026. – 2026. – URL: https://www.idc.com/resource -center/blog/global-memory-shortage-crisis-market-analysis-and-the-potential-impact-on-the-smartphone-and-pc-markets-in-2026/ (date of access: 27.03.2026).

4. Reuters. AI frenzy is driving new global supply chain crisis. – 2025. – URL: https://www.reuters.com/world/china/ai-frenzy-is-driving-new-global-supply-chain-crisis-2025-12-03 (date of access: 27.03.2026).

5. The Register. AI memory router prices surge. – 2026. – URL: https://www.theregister.com/2026/02/13/ai_memory_router_prices/ (date of access: 27.03.2026).

6. IEEE Spectrum. DRAM shortage: why memory is the new bottleneck. – 2025. – URL: https://spectrum.ieee.org/dram-shortage (date of access: 27.03.2026).

7. Kulkarni A. Efficient LLM training: memory–compute optimization. – Medium, 2024. – URL: https://medium.com/@ak16425/efficient-llm-training-memory-compute-optimization-f1aae81c83f4 (date of access: 27.03.2026).