Counterpoint | micron

Memory At The Edge: Powering GenAI's Next Frontier

June 2025

Contents

Executive Summary

The smartphone is undergoing its most profound transformation yet, evolving from a communication hub into an intelligent, context-aware, and increasingly autonomous agent. Driven by the rapid advancement of generative AI (GenAI), particularly agentic AI capabilities, these devices are shifting from responding to commands to proactively understanding intent and acting on the user's behalf. Counterpoint Research defines a GenAI smartphone as one "leveraging large-scale, pre-trained generative AI models to create original content or perform contextually-aware tasks," signifying a fundamental change in mobile computing.

Market adoption is accelerating dramatically. Counterpoint projects GenAI-capable smartphones will grow from approximately 20% of shipments in 2024 to 45% by 2027, surpassing a billion units in the installed base. This surge, led by pioneers like Samsung, Google and now Motorola — with its innovations integrating Perplexity cloud AI to deliver system-level AI experiences — signals how AI is becoming the central pillar of the smartphone experience.

However, realizing the full potential of agentic AI hinges on robust on-device processing. The need for lower latency, enhanced user privacy, cost efficiency, and offline functionality makes edge intelligence an imperative. While techniques like quantization are significantly improving the efficiency and deployability of large language models, they alone cannot overcome the hardware limitations. Running advanced AI workloads still places unprecedented strain on smartphone hardware. While SoCs require significant upgrades (driven by leaders like Qualcomm and MediaTek), **memory systems are emerging as the most critical bottleneck.**

The increasing size and complexity of AI models, the need to run multiple models concurrently, and the demands of real-time contextual processing highlight the key challenges facing memory chipmakers today. Current flagship DRAM capacities (12GB+) are becoming baseline, with requirements potentially reaching **32GB or more**. Bandwidth needs are escalating toward **14.4 Gbps (LPDDR6) and higher** while advanced NAND flash like **UFS 4.x and Zoned UFS (ZUFS)** is essential for fast model loading and caching. And framing all of these needs is power efficiency, which remains paramount.

Delivering on these and future demands requires concerted innovation. Solutions like LPDDR6, processing-in-memory (PIM), Wide I/O, advanced packaging techniques (e.g., heterogeneous packaging), and optimized flash storage are crucial. Industry leaders like Micron, Samsung, SK hynix, and potentially Apple are actively developing these technologies.

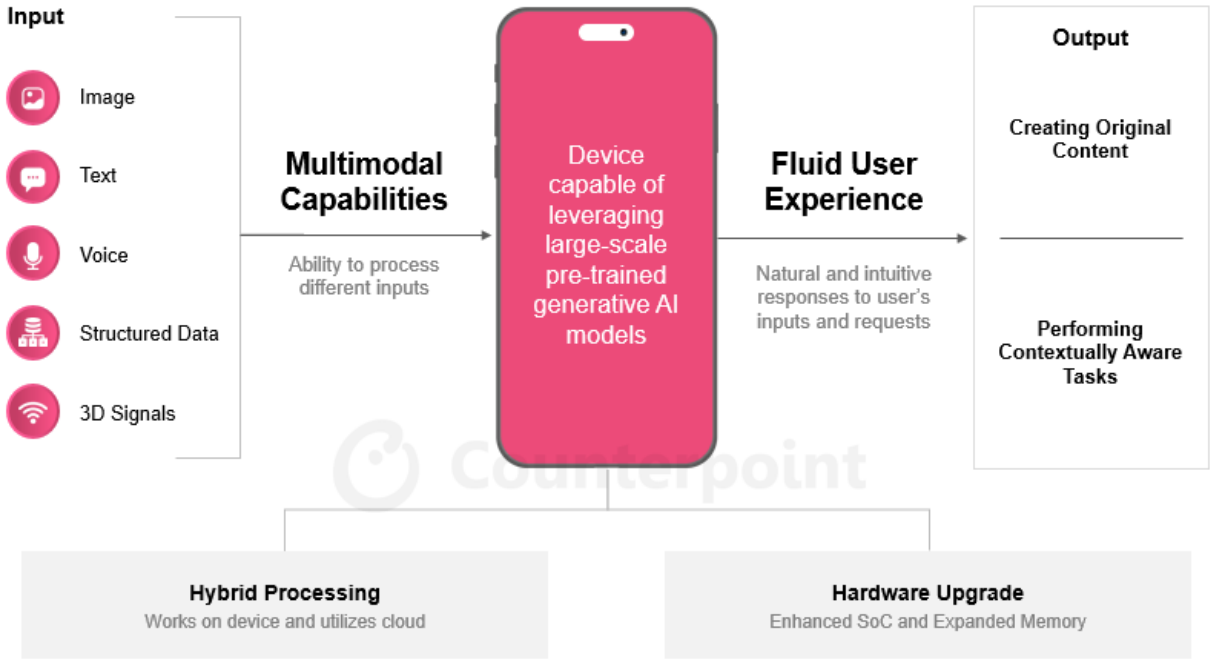
Yet, no single entity can succeed alone. Ecosystem-wide collaboration, robust standardization (e.g., JEDEC), and strategic investments (including government initiatives) are essential to dismantle these bottlenecks and usher in the era of the truly intelligent, agentic smartphone.

Part 1: Future Smartphones — Evolution From Communication Hub to Intelligent Agent

The smartphone has evolved through distinct generations, each adding transformative capabilities. We are now entering a new era defined by the deep integration of generative AI (GenAI), moving beyond simple feature enhancements toward creating truly intelligent, autonomous companions.

Counterpoint Research defines a GenAI smartphone as one "leveraging large-scale, pre-trained generative AI models to create original content or perform contextually-aware tasks." These devices increasingly handle diverse data types — text, images, voice, sensors — to perform sophisticated operations directly at the edge, marking a paradigm shift in personal computing. Achieving this level of functionality on the edge necessitates faster data processing and larger capacity, topics that will be explored in the subsequent sections.

Exhibit 1: GenAI Smartphone Definition



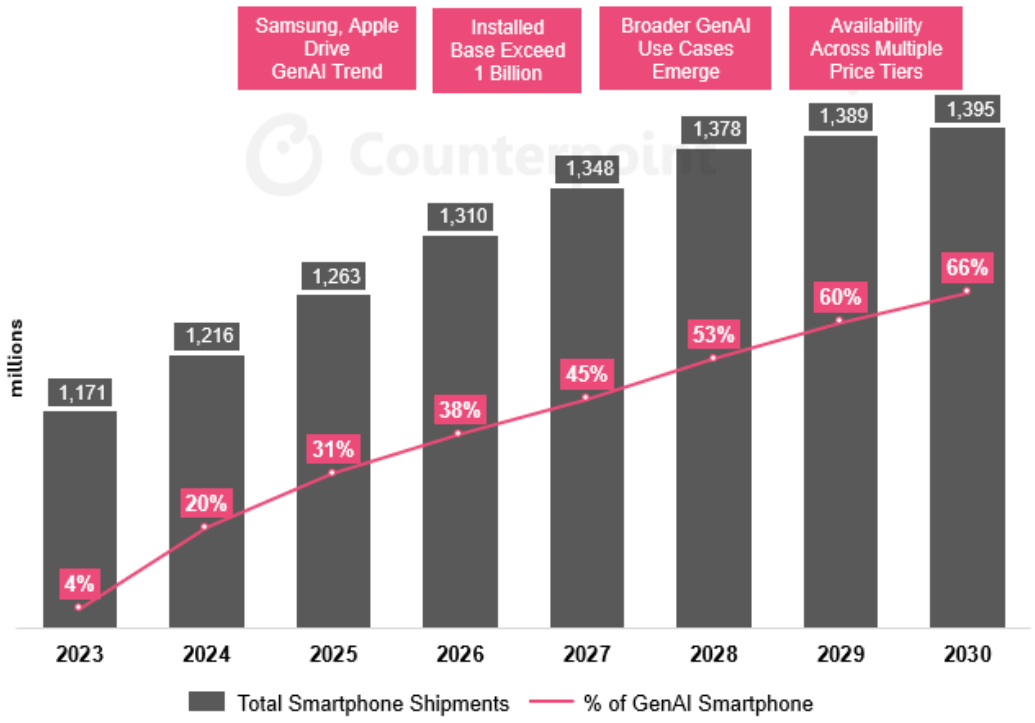
Source: Counterpoint Research AI 360 Service.

Market Outlook: Rapid Proliferation Underway

The adoption of GenAI capabilities in smartphones is accelerating rapidly and the shift from device as a communication hub to an intelligent agent is fully underway.

- **Shipment Growth:** Counterpoint Research estimates GenAI-capable smartphones accounted for over 20% of shipments in 2024 and projects penetration of around 45% in 2027 and nearly two-thirds by 2030. The installed base is expected to exceed one billion units within a few years.
- **OEM Deployment:** Over 15 OEMs have already launched more than 100 GenAI-capable models. Samsung and Google were early movers, Xiaomi is deploying rapidly, and Apple is widely expected to significantly advance the category.
- **SoC Landscape:** Qualcomm and MediaTek currently dominate the GenAI SoC space, providing the core processing power for these capabilities.
- **Segment Penetration:** While initially concentrated in the premium (wholesale $600 and above) segment, GenAI features are quickly cascading down to the mid–tier ($400-$599) through 2027 and are expected to become a baseline expectation across most price bands by 2030.

Exhibit 2: GenAI Smartphone Market Shipment Forecast



Source: Counterpoint Research AI 360 Service.

This surge is not merely about faster processing; it signals a fundamental change in the smartphone's role and how users interact with it.

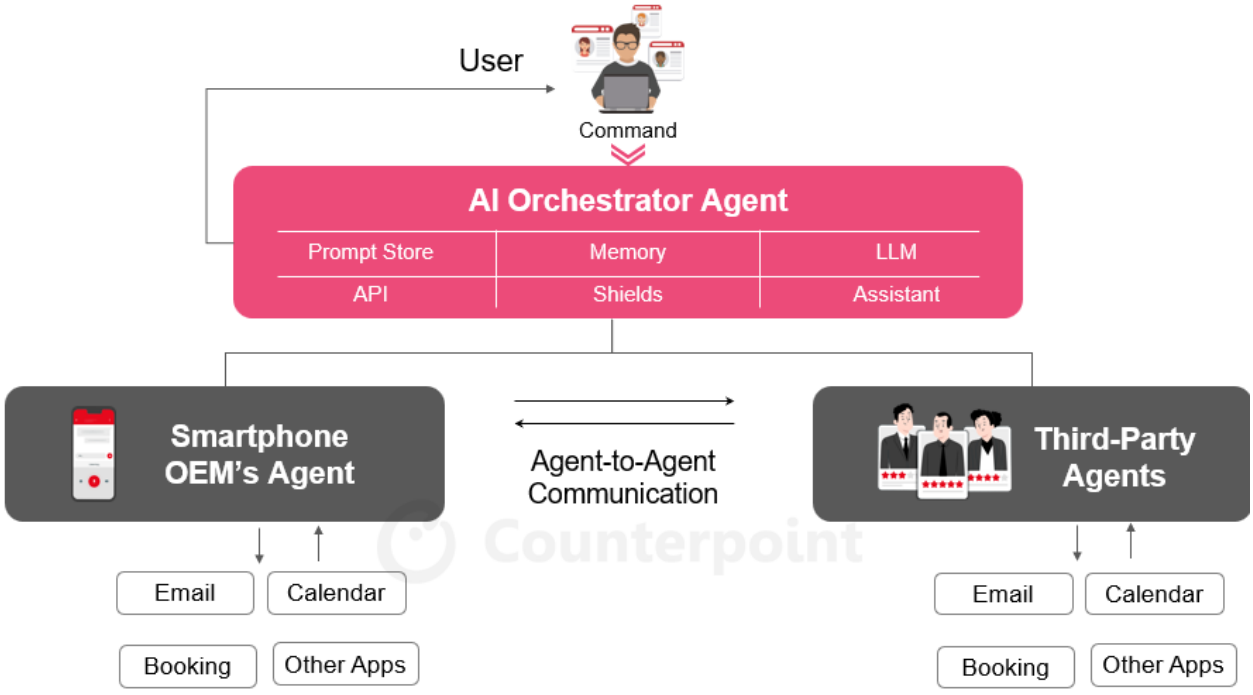
Agentic AI: The Core of the Transformation

The most transformative aspect of GenAI in mobile is the rise of **agentic AI** — systems that don't just respond to explicit commands but proactively understand intent, make decisions, and act autonomously on the user's behalf. These capabilities evolve the smartphone from a tool into a partner.

Instead of users manually navigating between apps, the AI agent acts as an orchestrator, coordinating tasks and information flow seamlessly. Chipset vendors like MediaTek define agentic AI by its autonomy, embedding engines (like AI Stack or Dimensity AI) designed to observe, predict, and execute. At events like MWC, companies like Honor and Qualcomm have demonstrated devices capable of orchestrating daily activities across multiple applications with minimal user input.

A recent commercial example is the new Motorola Razr foldable, which features Perplexity AI deeply integrated not merely as an app but as a proactive, agentic layer within the device experience. Acting as a system-level assistant, it interprets context, answers complex queries, and drives interactions across the interface. It marks one of the first tangible shifts from traditional app-centric design toward an AI-first, large language model (LLM)-centric operating paradigm. In the future, we expect such orchestration "super agents" to proliferate as well as see more on-device implementations like Gemini Nano.

Exhibit 3: Agentic AI Task Orchestration Across Apps



Source: Counterpoint Research AI 360 Service.

The implication is profound: Smartphones are becoming proactive digital companions capable of complex intent recognition and real-time adaptation, fundamentally changing the user experience.

Contextual Awareness: The Engine of Personalization

Agentic AI thrives on deep contextual understanding — moving beyond "you+" (AI assists you) to "you²" (AI is a digital extension of you). This enhancement requires processing a rich stream of data: location, schedule, communication history, sensor readings, user habits, and ambient environment.

Techniques like retrieval-augmented generation (RAG), implemented locally, allow LLMs to access and incorporate this personal context securely, delivering highly relevant and personalized responses without compromising privacy. This deep personalization is key to the value proposition of agentic AI, but it necessitates powerful and efficient on-device processing.

Other techniques such as quantization — which compresses models by reducing numerical precision — helps enable the deployment of advanced AI models on smartphones.

But ultimately, the biggest challenge lies at the hardware level and powering the ever growing universe of AI capabilities within the strict battery, processing and memory limits of a smartphone.

Part 2: The On-Device Imperative and the Hardware Challenge

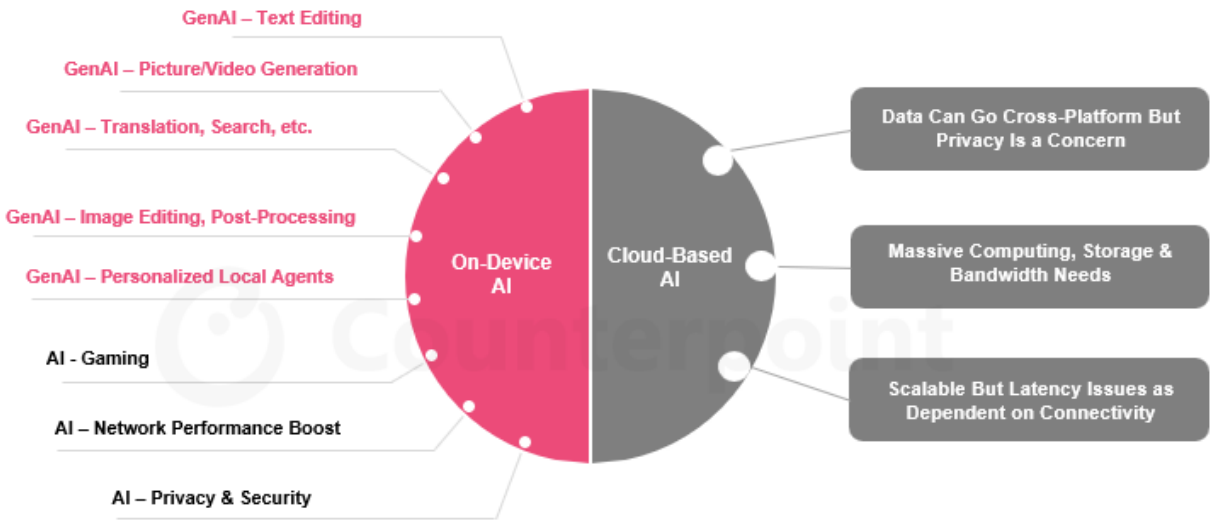
While cloud AI remains relevant, the core functionalities of agentic AI — real-time responsiveness, deep personalization, and proactive assistance — demand significant **on-device** processing. The shift toward edge intelligence is driven by compelling advantages but creates substantial hardware hurdles.

The Case for Edge Intelligence: Why On-Device Matters

Processing AI workloads directly on the smartphone offers critical benefits:

- **Lower Latency:** Eliminates cloud round-trip delays, enabling instantaneous responses crucial for seamless interaction and real-time tasks (e.g., translation, AR overlays).
- **Enhanced Privacy:** Keeps sensitive personal and contextual data on the device, minimizing exposure and building user trust — essential for systems learning intimate user details.
- **Cost Efficiency:** Avoids the significant computational costs associated with running complex models like LLMs in the cloud at scale. (A single cloud GenAI query can be approximately 40 times more expensive than a standard Google search.)
- **Offline Access:** Ensures core AI functionalities remain available even without a reliable internet connection.
- **Bandwidth Efficiency:** Reduces reliance on network bandwidth, saving data costs and improving performance in areas with poor connectivity.
- **Personalization:** Allows models to continuously fine-tune based on private user data stored locally, leading to truly personalized experiences.

Exhibit 4: On-Device vs. Cloud-Based AI Benefits



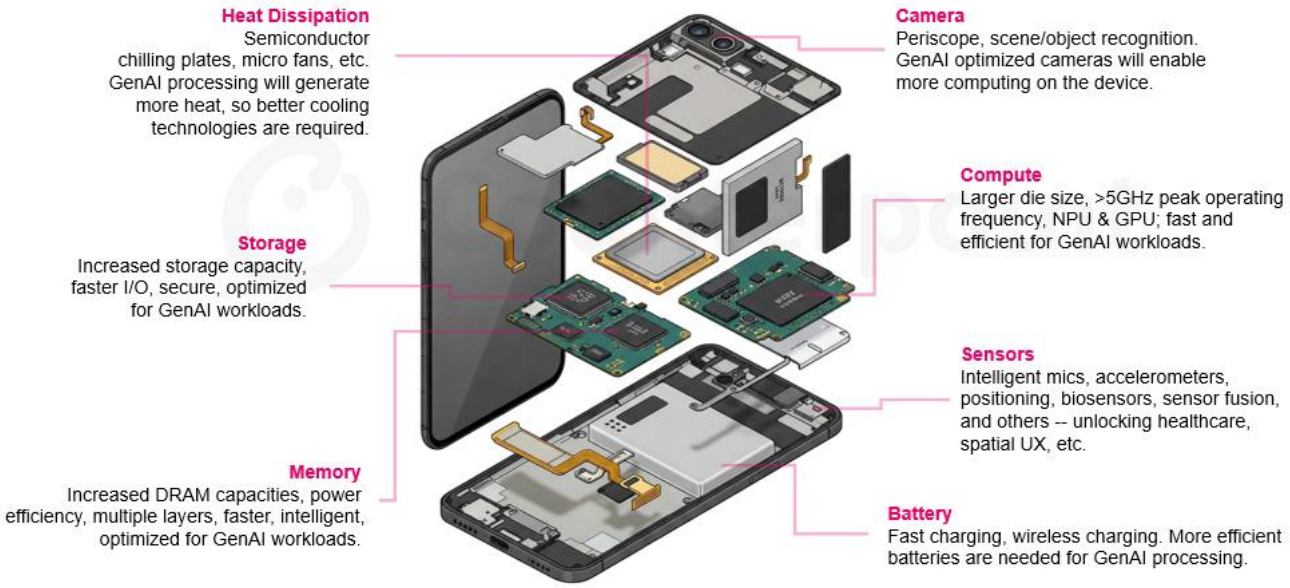
Source: Counterpoint Research AI 360 Service

The consensus is clear: For agentic AI to deliver on its promise, significant computation must happen locally. This requirement, in turn, necessitates a major leap in the capabilities of smartphone hardware components.

Escalating Hardware Requirements: A System-Wide Upgrade

Enabling powerful on-device AI requires substantial advancements across the board:

Exhibit 5: Typical Hardware Components for Today's GenAI Smartphone



Source: Counterpoint Research AI 360 Service

- **Processor (SoC):** The heart of AI processing requires increasingly powerful CPUs, GPUs, and especially dedicated **NPUs (neural processing units)** with tens of **TOPS** to drive efficient processing of AI workloads. Leaders like Qualcomm (Snapdragon) and MediaTek (Dimensity) are rapidly enhancing their on device AI capabilities. Innovations in architecture, parallel processing, and advanced process nodes (moving toward 1nm) are crucial.
- **Memory (LPDRAM):** Memory faces immense pressure from larger models, multitasking AI agents and the need to remain low power.
    - Capacity*:* Flagship baseline is moving past 12GB; future and concurrent GenAI experiences will likely require **32GB or more**.
    - Bandwidth*:* This needs to increase dramatically from LPDDR5X speeds ~10.7 Gbps to **LPDDR6 speeds (14.4 Gbps+)** to prevent starving the compute units.

- **Storage (NAND flash):** Storage must be faster and larger to store multiple large AI models, handle rapid loading/caching, and manage the influx of AI-generated data. High-speed interface like **UFS 4.x** is becoming critical.
- **Battery:** Intensive on-device processing demands higher-capacity batteries and smarter power management systems. Fast charging and potentially new battery chemistries are needed.
- **Sensors and Interconnects:** Richer sensor data fuels contextual AI, while efficient internal data pathways are essential. Advanced packaging becomes vital for integrating these components effectively.
- **Thermal Management:** Dissipating the heat generated by sustained AI workloads is critical for maintaining performance. Advanced cooling solutions are necessary.

The simultaneous demand for higher capacity, massively increased bandwidth, faster storage access, and strict power efficiency across memory and storage systems has become a significant bottleneck limiting AI performance at the edge.

Part 3: The Memory-Powered Future of Edge AI

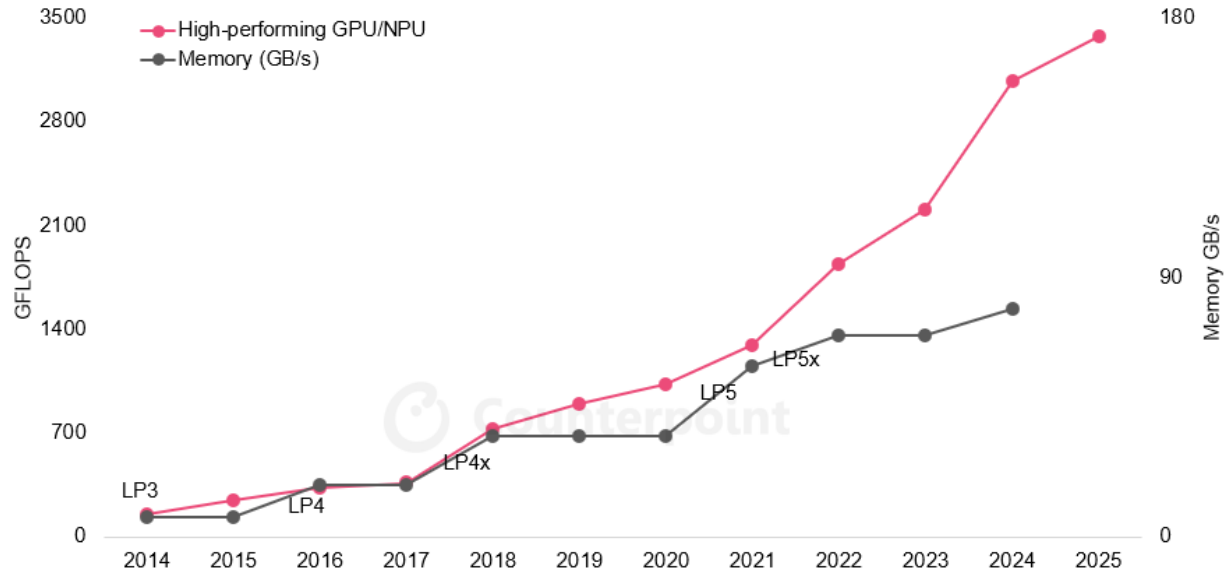
As processing capabilities in SoCs surge forward, requirements on memory subsystems are also rising quickly with the ever increasing need to supply data quickly and efficiently in support of advanced on-device AI. Delivering on these needs is crucial to the user experience as AI model complexity, response needs, and multitasking capabilities increase steeply.

Unlocking Performance: The Role of Memory

OEMs and developers face challenges deploying full, on-device, bandwidth-consuming AI feature sets and models on smartphones with less memory (e.g. 8GB vs. 16GB+). The situation mirrors the HPC/data center space, where innovations like HBM were necessary to feed powerful GPUs. On mobile, the constraints of power, cost, and space make the challenge even more acute.

Notably, “compute” (NPU/GPU) performance growth has outpaced “memory” (DRAM) bandwidth speed growth in mobile, and simply adding more conventional DRAM has not been a viable long-term solution. To meet this challenge, architectural innovation is essential.

Exhibit 6: Accelerating Hardware Innovation



Source: Micron, Qualcomm, Counterpoint Research.

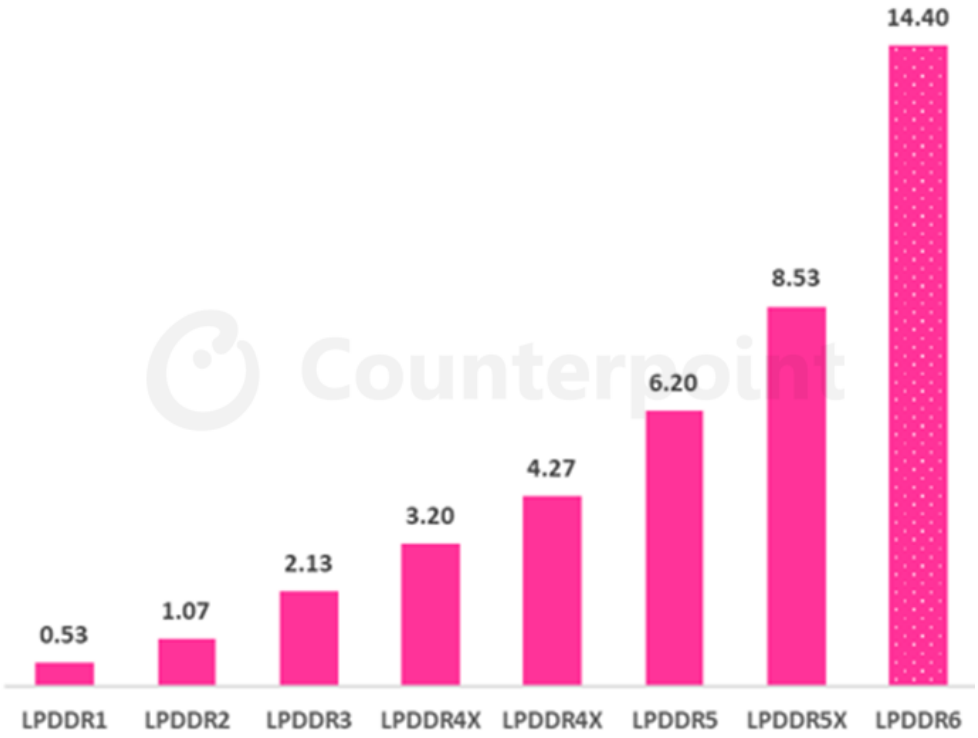
Powering Tomorrow's GenAI Smartphones: Next-Generation Memory Solutions

The industry is actively developing several key technologies to ensure memory solutions dovetail with the escalating needs of the GenAI smartphone:

1. Advanced LPDDR (LPDDR5X, LPDDR6)

- **LPDDR5X:** Current standard, offering speeds up to ~10.7 Gbps
- **LPDDR6:** The next major DRAM standard under development by JEDEC, targeting significantly higher bandwidth (**up to 14.4 Gbps or more**) and improved power efficiency, crucial for feeding demanding AI accelerators.

Exhibit 7: LPDDR RAM Speed Evolution (Gbps)

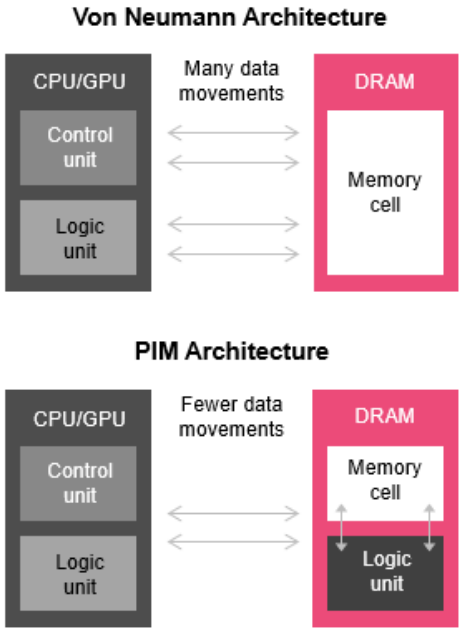


Source: Counterpoint Research Memory Service

2. Processing-In-Memory (PIM) Architecture

PIM fundamentally challenges current Von Neumann architectures by embedding compute capabilities directly within or near memory arrays. This placement drastically reduces data movement for specific operations common in AI (like vector math, matrix operations), resulting in significant latency and power-saving benefits. While standardization and ecosystem support are still evolving, PIM holds considerable promise for accelerating specific AI workloads.

Exhibit 8: Von Neumann Architecture vs. PIM Architecture

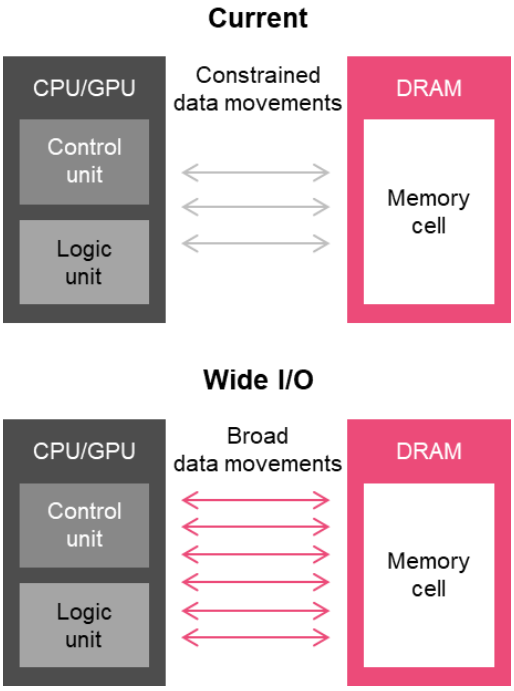


Source: Micron

3. Wide I/O Interface and Advanced Packaging

- **Wide I/O:** Wide I/O increases the physical number of data connections between memory and the SoC, boosting raw bandwidth. It's often implemented using advanced packaging.
- **Advanced Packaging:** Techniques like 3D stacking (using through-silicon vias, or TSVs) or heterogeneous packaging (integrating chipsets) are crucial. They enable Wide I/O and allow for better thermal management and potentially higher pin counts. We could see some OEMs and supply chain partners exploring strategies such as offloading DRAM into separate packages, which could improve performance for AI workloads considerably.

Exhibit 9: Current Architecture vs. Wide I/O Interface



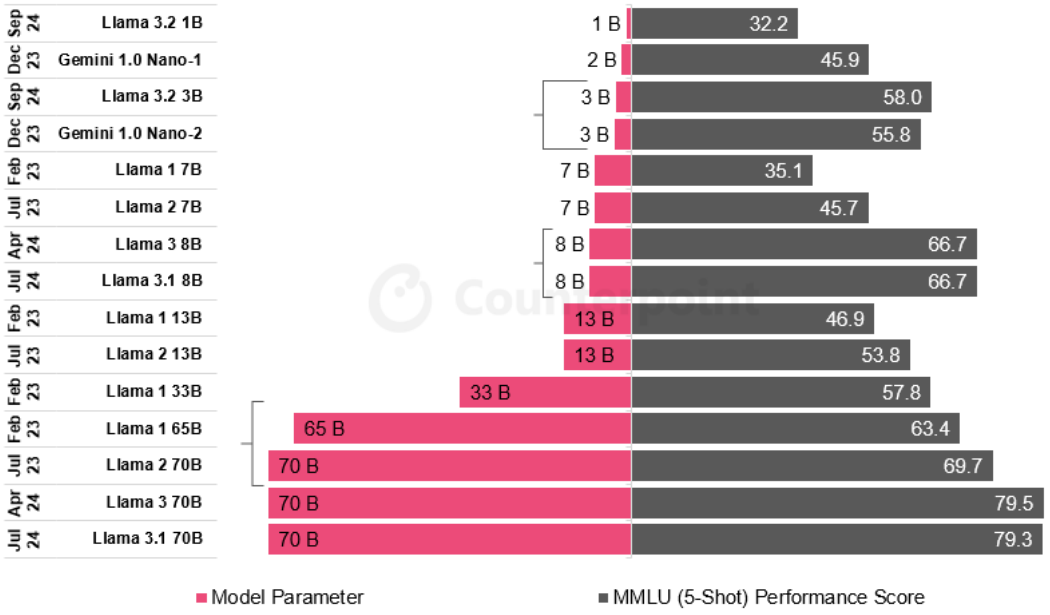
Source: Micron.

Quantization and Small Language Models

Quantization is a pivotal technique in the advancement of AI model efficiency, particularly for on-device applications. By reducing the precision of a model's parameters from 32 bits to 16 or even 8 bits, quantization significantly decreases the memory footprint and computational requirements without compromising accuracy.

This process is essential for enabling sophisticated AI functionalities on resource-constrained devices like smartphones. Additionally, the development of small language models (SLMs) has revolutionized the landscape of AI. These models, despite having fewer parameters, can perform on par with their larger counterparts. For instance, SLMs with just 8 billion parameters can now closely match the performance of older models, which had 70 billion parameters.

Exhibit 10: Large Cloud GPT vs. Small Parameter LlaMA Performance Comparison



Source: Company announcements.

This efficiency is achieved through techniques — such as data filtering during training, which uses cleaner datasets to improve model quality — and the mixture of experts approach. In the latter technique, the model is divided into specialized chunks that work together, activating only the necessary parts at any given time. Microsoft's BitNet b1.58 takes quantization to the extreme and shows how effectively it can reduce average bit requirements per parameter — in Bitnet's case to an astonishing 1.58 bits.

These innovations make it feasible to run advanced AI models on smaller, low-power devices, thereby enhancing the capabilities of edge intelligence and contributing to the proliferation of GenAI smartphones.

Conclusion — Industry Action and a Call for Collaboration

The evolution of the smartphone into an agentic, AI-native platform is rapidly moving from concept to reality. Real-world examples like the Perplexity AI integration in the Motorola Razr, alongside advancements from Samsung, Google, and Honor clearly signal that the future mobile experience will be orchestrated by AI, anticipating user needs rather than simply reacting to taps and swipes.

Momentum Builds: Industry Actions and Investments Across the Ecosystem

This transformation is underpinned by tangible actions across the ecosystem:

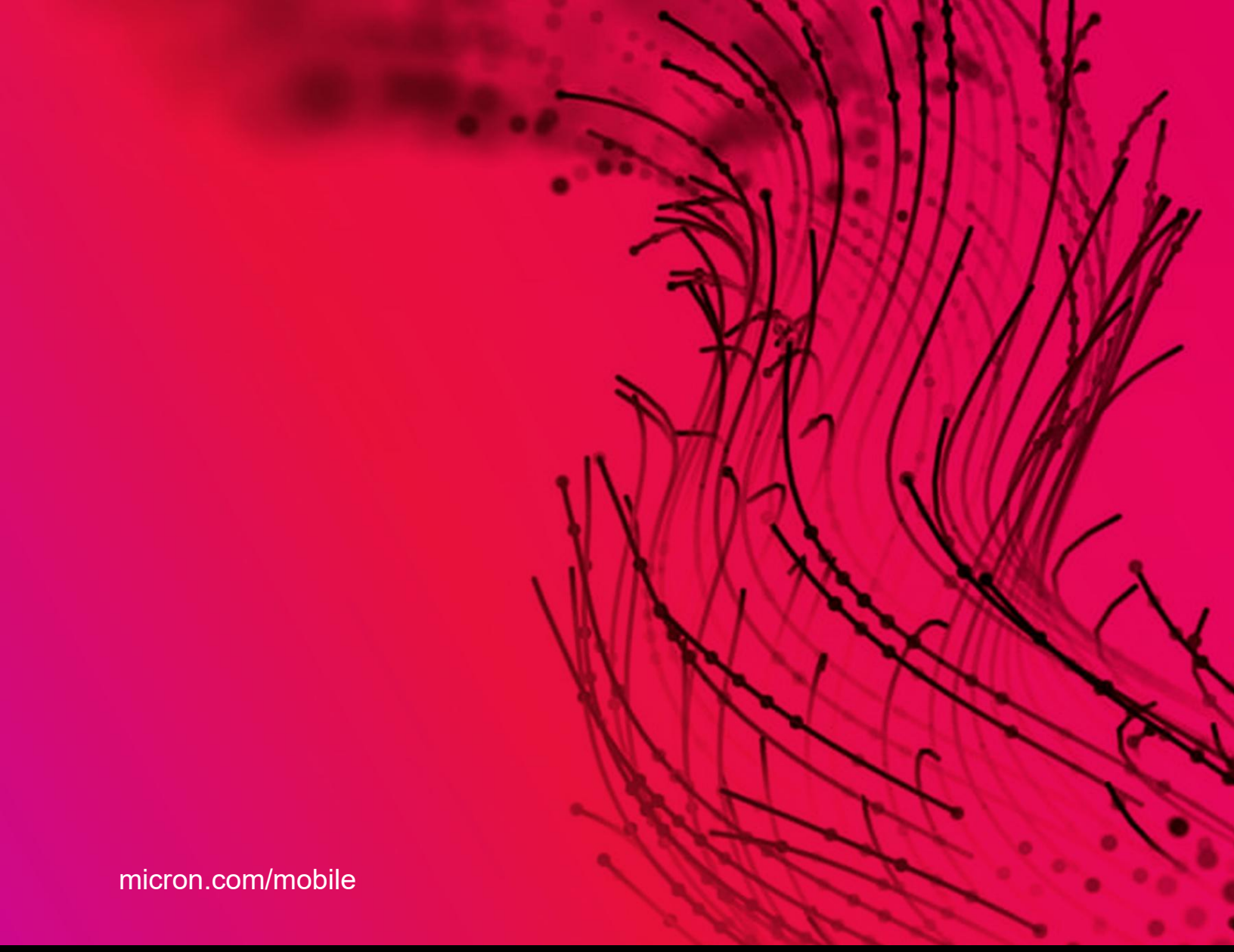
- **Memory Vendor Innovation:** Companies like Micron, Samsung, and SK hynix are actively developing and sampling next-generation technologies like LPDDR6, PIM variants, and ZUFS, highlighted at industry events like MWC.
- **OEM Architectural Shifts:** Major device manufacturers appear to be fundamentally rethinking device architectures, particularly around memory subsystems, to accommodate future AI demands.
- **Government Support:** Recognizing the strategic importance of AI hardware, governments in the U.S., E.U., India, and elsewhere are investing significantly in domestic semiconductor manufacturing and R&D.

Message to the Industry: Align, Accelerate, Collaborate

Intelligence is only as powerful as the system supporting it. While compute (NPUs/GPUs) garners headlines, **memory performance is the core enabler for the agentic AI revolution on mobile.** Meeting AI's growing memory demands is not just a task for memory vendors. It also requires unified industry efforts, including actions like these :

- **Deep Collaboration:** Unprecedented cooperation is needed between SoC designers, memory/storage vendors, OEMs, OS developers, and AI researchers to co-optimize hardware and software for edge AI workloads.
- **Accelerated Standardization:** Robust, timely standards from bodies like JEDEC are vital for LPDDR6, future UFS iterations, and potentially PIM/packaging interfaces, ensuring interoperability and fostering innovation.
- **Shared Vision and Investment:** A collective focus on anticipating future requirements and investing in next-generation memory, storage, and packaging technologies is essential to keep pace with AI's exponential demands.

This is more than a race for faster processors; it's a call to action for the industry to unlock the transformative potential of agentic AI, elevating smartphones into truly intelligent partners that enrich our lives. The future of mobile is not just smart — it is autonomous. And it starts now.



COUNTERPOINT RESEARCH

Seoul | Taipei | San Jose | San Diego | Mumbai | London
Los Angeles | Delhi | Beijing | Shenzhen | Shanghai

Contact Us:

www.counterpointresearch.com
info@counterpointresearch.com

@CounterPointTR

@Counterpoint Technology Market Research

Counterpoint