AMD MEXT



ACHIEVING UP TO 3.5X HIGHER PERFORMANCE-PER-DRAM-DOLLAR FOR AGENTIC AI

An analysis of 6th Gen AMD EPYC™ + AMD MEXT™ software for agentic AI

EXECUTIVE SUMMARY

This whitepaper examines how AMD MEXT Predictive Memory™ software can reduce DRAM dependence for agentic AI workloads. AMD internal testing used a custom agentic AI benchmark suite on an AMD reference platform powered by AMD EPYC™ 9996 processors and compared an all-DRAM baseline (~1TB) with configurations retaining 50% and 25% of the baseline DRAM allocation.

With 50% of baseline DRAM retained and 512GB of MEXT Memory™ (AMD MEXT™ software + NVMe flash appearing as memory), mixed-phase geomean performance was nearly the same as the all-DRAM baseline, translating to approximately 1.99x performance-per-DRAM-dollar[1].

With 25% of baseline DRAM retained and a 768GB of MEXT Memory, mixed-phase geomean performance reached 85.3% of baseline, translating to approximately 3.41x performance-per-DRAM-dollar[2].

For separate testing running only a memory-hungry image processing agent, "mtpillow", performance reached 89.7% of baseline for the 50% DRAM / 50% MEXT Memory™ scenario and 88.4% of baseline in the 25% DRAM / 75% MEXT Memory scenario – translating to 1.79x and 3.54x improvements in performance-per-DRAM-dollar.[3,4]

INTRODUCTION

Agentic AI systems can combine multiple specialized agents and large numbers of concurrent workers, creating memory-intensive execution patterns that place pressure on server DRAM capacity. As concurrency and working sets grow, the ability to right-size DRAM while preserving useful throughput can materially affect infrastructure efficiency. This study evaluates how well AMD MEXT™ software can preserve agentic AI benchmark performance by substituting AI-powered, flash-backed memory extension in place of physical DRAM.

HARDWARE CONFIGURATION

- **Platform:** AMD reference platform
- **Processors:** 2× AMD EPYC™ 9996 256-Core Processors
- **CPU resources:** 256 cores and 512 threads per socket; SMT on; NPS4
- **Baseline DRAM allocation:** 1,024.7 GiB/socket with MEXT off
- **MEXT memory-pool storage:** Solidigm SB5PHA7X038T 3.84 TB NVMe

SOFTWARE STACK

- **Application:** AAAI agentic AI benchmark kit, agentic-saturation-replay
- **Operating system:** Ubuntu 24.04.5 LTS
- **Kernel:** Linux 6.17.12-061712-generic
- **Compiler:** GCC 13.3.0
- **MEXT:** AMD MEXT™ software version 1.2

AMD MEXT PREDICTIVE MEMORY SOFTWARE

How It Works

The MEXT software-only solution makes flash storage appear as DRAM-like memory to the OS.

Here's how it works:

- MEXT continuously monitors which memory pages in DRAM are actively being utilized (hot) and which have gone cold
- MEXT offloads the cold memory from DRAM to flash
- MEXT AI engine mitigates the effects of flash latency to keep the system performant
  - AI engine continually predicts which offloaded pages might be requested by the application soon (in other words, which pages are likely to soon go from cold to hot)
  - AI engine transparently moves those pages back into DRAM before the requests are even made
  - As a result, the application stays performant because from its perspective, the relevant memory pages are always already resident in DRAM

THE MARKET: DRAM CHALLENGES

Server memory (DRAM) comes with some major challenges.

First, it is now the most expensive data center component, with cost-per-bit increasing dramatically and showing no signs of slowing.[5]

Second, predicting how much is needed for a particular job is very difficult, and getting it wrong can result in out-of-memory (OOM) errors and job failures; as a result, most teams over-provision to avoid the "memory cliff". Overprovisioning means budget wastage.

Third, for larger memory systems, the granularity of available memory sizes is quite coarse; even if the exact requirement is known, it's impossible to fine-tune the purchased memory to the requirement plus a small buffer (e.g., 10%).

Lastly, even if the precisely correct memory size is somehow implemented, utilization is often quite poor.

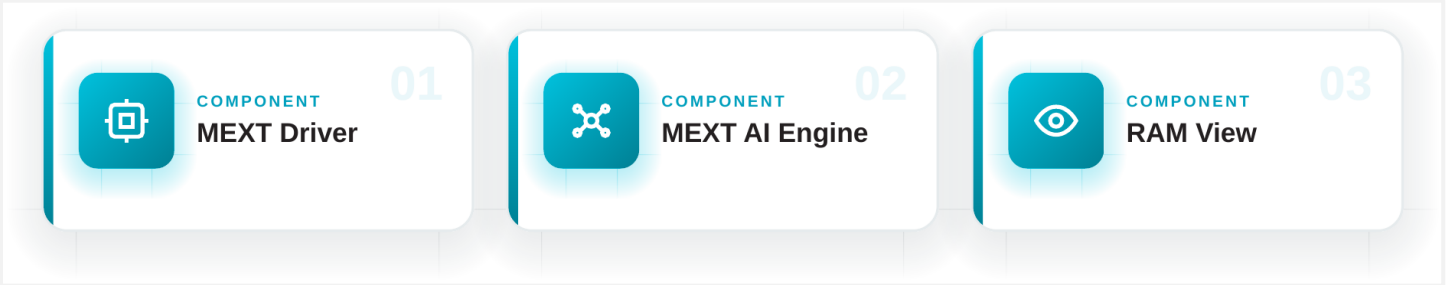
In short, DRAM is extremely expensive, often over-provisioned, hard to fine-tune, and usually poorly utilized.

MEXT Value

MEXT addresses the major DRAM challenges outlined above. First, by leveraging low-cost flash as memory, MEXT enables much more cost-effective computing. Second, MEXT can prevent OOM crashes by enabling applications to go out to flash-as-memory if DRAM ever runs out. Third, MEXT enables more precise right-sizing of memory sizes to true requirements. Lastly, MEXT solves the utilization issue by intelligently tiering across DRAM and flash to keep DRAM mostly all hot.

AMD MEXT SOLUTION COMPONENTS

The solution consists of 3 primary components: the MEXT Driver, the MEXT AI Engine, and the RAM View™ Observability Platform.



Driver

The MEXT Driver is a dynamically loadable kernel module (which does not alter the standard Linux kernel) that sends process and memory page telemetry data to the AI engine.

MEXT AI Engine

The AI engine is a user-space process that feeds predictions of which memory pages should be pushed from flash to DRAM – making predictions / inferences in under a fraction of a second. It runs entirely on the local Linux operating system (on a single CPU core) and does not require a GPU.

It was inspired by modern AI techniques based on neural networks. Instead of using these techniques to predict words or natural language patterns, it predicts sequences of future memory page accesses. It consists of a family of models that work together, combining extremely lightweight heuristic predictors with more powerful neural-network models. For any given workload, it automatically adjusts to use the model or group of models that performs best. Continuous observation of which predicted pages were actually used by the application also enables the engine to acquire real-time feedback regarding model accuracy, supporting ongoing adaptation and self-optimization.

RAM View Observability Platform

RAM View is an observability / visualization tool that illustrates how much memory the application is consuming including real-time insights into page hotness / warmness / coldness. It illustrates the proportion of memory pages that were accessed <15s ago, <1min ago, <2min ago, <3min ago, <5min ago, and >5min ago.

METHODOLOGY

Testing used five agent personas (dev, context, secure, rag and mtpillow), first in solo phases and then in a five-way concurrent mixed phase. Three runs were completed for each scenario. Two metrics are reported. Mixed-phase geomean takes each persona's throughput at the 640-worker step and calculates the geometric mean across the five personas. Solo mtpillow throughput is measured in sessions per second over a 150-second window; the benchmark stepped through worker counts up to 2,560, and the best step of 2,560 workers was repeated three times. Higher throughput is better. The all-DRAM baseline used a 1,024.7 GiB/socket DRAM limit with MEXT disabled. The 50% DRAM-reduction scenario used 511.9 GiB DRAM plus a 512GB MEXT Memory pool. The 75% DRAM-reduction scenario used 255.9 GiB DRAM plus a 768GB MEXT Memory pool.

RESULTS

Test Configuration / Metric	Average Result
All-DRAM baseline - mixed-phase geomean	3.910 tasks/s
50% DRAM reduction + MEXT - mixed-phase geomean	3.921 tasks/s
75% DRAM reduction + MEXT - mixed-phase geomean	3.337 tasks/s
All-DRAM baseline - solo mtpillow	17.78 sessions/s
50% DRAM reduction + MEXT - solo mtpillow	15.94 sessions/s
75% DRAM reduction + MEXT - solo mtpillow	15.72 sessions/s

50% DRAM-Reduction Scenario

Reducing retained DRAM by 50% while adding a 512GB MEXT Memory produced nearly equivalent performance compared to the baseline mixed-phase geomean performance and 89.7% of baseline solo mtpillow throughput. The corresponding performance-per-DRAM-dollar values were approximately 1.99x and 1.79x (taking the performance percentages over 50% of the DRAM costs).[1,3]

75% DRAM-Reduction Scenario

Reducing retained DRAM by 75% while adding a 768GB MEXT Memory produced 85.3% of baseline mixed-phase geomean performance and 88.4% of baseline solo mtpillow throughput. The corresponding performance-per-DRAM-dollar values were approximately 3.41x and 3.54x (taking the performance percentages over 25% of the DRAM costs).[2,4]

CONCLUSION

MEXT can preserve a high proportion of agentic AI workload performance while substantially reducing the amount of physical DRAM needed. Across the different benchmark configurations and scenarios, performance ranged from 85.3% to near-equivalent performance with substantial DRAM cuts, between 50-75%. This means that customers can leverage MEXT to drive significant performance-per-DRAM-dollar gains for agentic AI – enabling more infrastructure efficiency for these critical next-gen workloads.

DISCLAIMER

©2026 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD Arrow logo, AMD EPYC, and combinations thereof are trademarks of Advanced Micro Devices, Inc. Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies. Supported features may vary by operating system. Please confirm with the system manufacturer for specific features. No technology or product can be completely secure.

ENDNOTES

1. 9xx6-082: Testing conducted by AMD October 20th, 2026. Testing used the agentic AI benchmark kit (agentic-saturation-replay 0.3.0-pre, commit 29e26c8) on an AMD reference platform with an AMD EPYC 9996 processor. The mixed-phase geomean all-DRAM baseline averaged 3.910, compared with 3.921 for the 50% DRAM / 50% MEXT memory scenario. Performance was 99.7% of baseline. Higher throughput is better. Results are averages of three runs per scenario. The mixed-phase geomean is the geometric mean of the five agent personas' throughput values at the 640-worker step. This translates to approximately 1.99x performance-per-DRAM-dollar using the tested DRAM fractions (0.997 / 0.50). Performance-per-DRAM-dollar claims are not meant to equate to performance-per-system-dollar claims or overall TCO claims. They represent performance normalized to the all-DRAM baseline divided by the retained DRAM fraction for the test. Results may vary based on workload, memory capacity and utilization, software versions, operating-system settings, storage configuration, system configuration and other factors.
2. 9xx6-084: Testing conducted by AMD; October 20th, 2026. Testing used the agentic AI benchmark kit (agentic-saturation-replay 0.3.0-pre, commit 29e26c8) on an AMD "Nigeria" reference platform with an AMD EPYC 9996 processor. The mixed-phase geomean all-DRAM baseline averaged 3.910, compared with 3.337 for the 25% DRAM / 75% MEXT memory scenario. Performance was 85.3% of baseline. Higher throughput is better. Results are averages of three runs per scenario. The mixed-phase geomean is the geometric mean of the five agent personas' throughput values at the 640-worker step. This translates to approximately 3.41x performance-per-DRAM-dollar using the tested DRAM fractions. Performance-per-DRAM-dollar claims are not meant to equate to performance-per-system-dollar claims or overall TCO claims. They represent performance normalized to the all-DRAM baseline divided by the retained DRAM fraction for the test. Results may vary based on workload, memory capacity and utilization, software versions, operating-system settings, storage configuration, system configuration and other factors.
3. 9xx6-083: Testing conducted by AMD; October 20th, 2026. Testing used the agentic AI benchmark kit (agentic-saturation-replay 0.3.0-pre, commit 29e26c8) on an AMD "Nigeria" reference platform with an AMD EPYC 9996 processor. The solo mtpillow throughput all-DRAM baseline averaged 17.78 sessions/s, compared with 15.94 sessions/s for the 50% DRAM / 50% MEXT memory scenario. Performance was 89.7% of baseline. Higher throughput is better. Results are averages of three runs per scenario. Solo mtpillow throughput was measured in sessions/s over a 150-second window; the best worker step, 2,560 workers in every run, was repeated three times. This translates to approximately 1.79x performance-per-DRAM-dollar using the tested DRAM fractions. Performance-per-DRAM-dollar claims are not meant to equate to performance-per-system-dollar claims or overall TCO claims. They represent performance normalized to the all-DRAM baseline divided by the retained DRAM fraction for the test. Results may vary based on workload, memory capacity and utilization, software versions, operating-system settings, storage configuration, system configuration and other factors.
4. 9xx6-085: Testing conducted by AMD; October 20th, 2026. Testing used the agentic AI benchmark kit (agentic-saturation-replay 0.3.0-pre, commit 29e26c8) on an AMD "Nigeria" reference platform with an AMD EPYC 9996 processor. The solo mtpillow throughput all-DRAM baseline averaged 17.78 sessions/s, compared with 15.72 sessions/s for the 25% DRAM / 75% MEXT memory scenario. Performance was 88.4% of baseline. Higher throughput is better. Results are averages of three runs per scenario. Solo mtpillow throughput was measured in sessions/s over a 150-second window; the best worker step, 2,560 workers in every run, was repeated three times. This translates to approximately 3.54x performance-per-DRAM-dollar using the tested DRAM fractions. Performance-per-DRAM-dollar claims are not meant to equate to performance-per-system-dollar claims or overall TCO claims. They represent performance normalized to the all-DRAM baseline divided by the retained DRAM fraction for the test. Results may vary based on workload, memory capacity and utilization, software versions, operating-system settings, storage configuration, system configuration and other factors.
5. CounterPoint Research: https://counterpointresearch.com/en/reports/memory-price-tracker-feb-2026