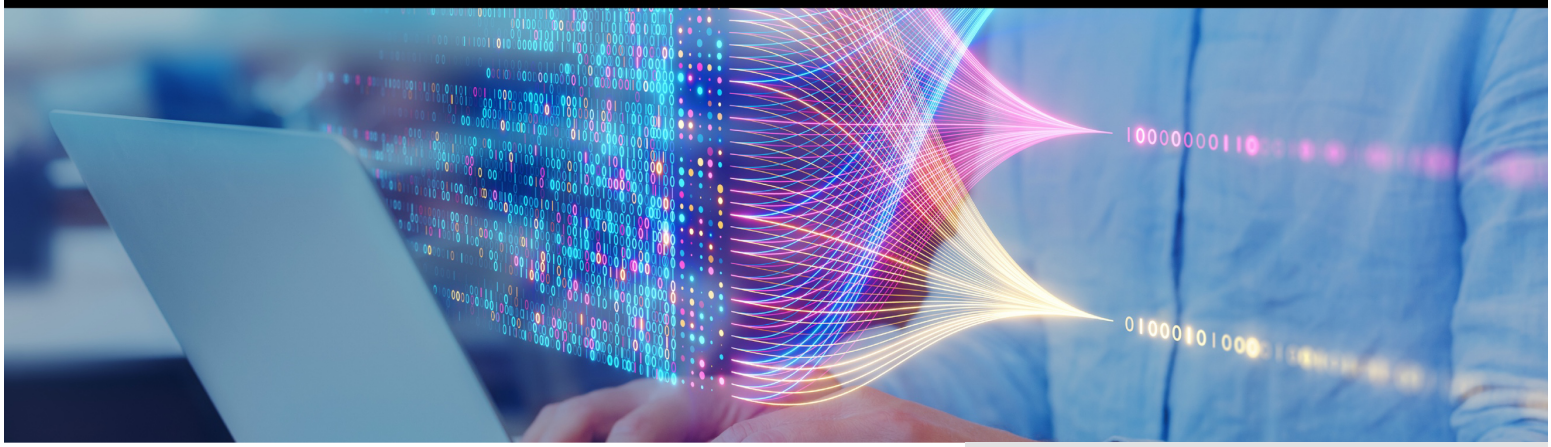

The role of the memory subsystem in achieving AI PC efficiency

Memory is a key enabler along with other hardware accelerators to realize the full potential of AI PCs.

In the evolving artificial intelligence (AI) landscape, the demand for more capable data models across diverse domains has led to the rapid expansion of model sizes. This fast-paced evolution continually increases the size and complexity of AI models, creating unprecedented demands on both compute and the memory subsystem performance to process and integrate vast amounts of data from various inputs — text, audio, video, and more. As AI continues to progress, advanced memory solutions are essential to support this computational growth, not only for large-scale data centers but also for edge devices, including **AI PCs**, which bring AI capabilities directly to personal and professional devices. Optimized memory solutions are instrumental in enabling the next generation of AI-driven innovations across devices and platforms.

This white paper delves into the architectural aspects of AI PCs, focusing on the collaborative efforts of Micron and Lenovo. It provides a comparative analysis of DDR5 and LPCAMM2 memory solutions on a Lenovo AI PC powered by the Intel® Core™ Ultra 9 Processor 185H (formerly code named Meteor Lake).[2] It highlights performance metrics, power savings and the efficiency of these memory solutions in real-world AI workloads. Additionally, it explores the benefits of dual-channel versus single-channel memory and the advantages of higher memory capacities across various AI workloads tested on Compal platforms equipped with the Intel® Core™ Ultra 7 Processor 165U. This comprehensive analysis offers insights into how different memory solutions are influencing and shaping AI PCs. The results serve as a valuable resource for system and product architects, as well as key PC stakeholders such as OEMs and ODMs, guiding them in making informed decisions regarding the selection and integration of the best memory subsystem for their platform to deliver an optimal AI PC experience for their customers.

Key takeaways

20%

Faster inference

When running AI inference on a system with a central processing unit (CPU), integrated GPU (iGPU), and neural processing unit (NPU), a configuration with *dual-channel DDR5 16GB+16GB* is on average 20% faster (inference time) than a *single-channel DDR5 32GB*.[1]

85+%

Lower power consumption

For various benchmarks, LPCAMM2 consumes around 85% less power than DDR5 across CPU, iGPU, and NPU. Subsequently, LPCAMM2 is more power efficient than DDR5, offering up to seven times improvement in power efficiency (performance per watt) for the memory subsystem.

16GB+

More capacity required for AI PC

With 32GB and above, users prevent out-of-memory errors that trigger swapping mechanisms and slow down the system operation speed by up to 50%. LPCAMM2 offers up to 64GB, ensuring future readiness for advanced models.

1. Across various model benchmarks and language models.

2. Data collected from systems in a Micron lab and specified Lenovo systems.

MICRON: Naveen Krishna Yarlagadda, Ranjith Kumar Nagisetty, Sudharshan Vazhkudai, Eric Caward, Evelyn Grevelink, Charlie Wang, Arthur Wang, Jitendra Singh Tomar, Pradeep Kumar Jilagam, James Myers, Viral Gosalia | **LENOVO**: Robin Tan, Judy Zhu, Yongpeng Wang, Jiao Han

Introduction

In the dynamic realm of AI, the advent of AI PC platforms marks a significant leap forward. These platforms are revolutionizing the processing of complex computational tasks on PCs by enabling local data processing. This results in improved performance, enhanced data privacy and security while reducing latency and network congestion, enabling a superior user experience. Just as past technological innovations like personal computers, mobile devices and cloud computing have transformed various aspects of our lives, the recent surge in generative AI is driving unprecedented innovation across every facet of technology, including the emergence of AI PCs.

What is an AI PC?

An **AI PC** is a computing device that integrates AI processing capabilities directly into its hardware and software. Unlike traditional PCs, which rely on application software or cloud services to perform AI tasks, AI PCs have built-in AI processors or accelerators that enable AI by performing matrix multiplication locally.

AI PCs offer several performance advantages over standard laptops, primarily due to their specialized hardware for AI tasks such as processing sensitive user data locally to ensure privacy for tasks like facial recognition or biometric authentication. Here are some key differences:

Heterogeneous compute architecture: AI PCs have adopted heterogeneous computing architecture with a combination of CPU, integrated GPU and NPU, where the **NPU** provides dedicated hardware acceleration for AI tasks and has superior performance per watt as showcased in the following pages .

Local AI processing: AI PCs can run AI models locally with better performance, reducing dependency on cloud-based solutions. This reduces the risk of data interception during transmission and enhances privacy. Standard laptops that do not have an NPU have limited capability for local AI processing.

On-device AI accelerators: AI PCs are optimized for AI applications with faster processing, making them more efficient for complex AI tasks. The NPU in AI PCs is designed to optimally perform matrix multiplications and convolution operations, which are fundamental to AI and machine learning. It leverages parallel arrays of processing to perform multiple operations simultaneously, improving performance with lower power consumption and offloading the CPU to perform general-purpose computing applications and iGPUs to render complex images for other applications such as video editing and gaming, etc. [4]

Power efficiency: AI PCs are more power-efficient for tasks such as video enhancement or image recognition due to their specialized hardware accelerators, which are designed to perform AI tasks efficiently compared to general-purpose CPUs. Because standard laptops lack these accelerators, they are less efficient when running AI workloads.

Software ecosystem: The software ecosystem of AI PCs is optimized for AI applications and frameworks, supporting built-in AI features such as real-time language translation and enhanced video conferencing. Standard laptops have a standard software ecosystem with limited built-in AI features.

Enhanced security: AI PCs equipped with AI-based security systems can analyze data from patterns and behaviors in real time to identify potential threats such as malware and phishing attempts, strengthening overall system security.

For the average user, an AI PC offers enhanced performance, security, privacy and power efficiency. Unlike standard PCs, which transfer data to the cloud for AI tasks and can suffer from network congestion, AI PCs handle these tasks locally. This capability transforms the user experience for AI applications. Professionals, especially in creative and technical fields, can leverage the powerful on-device AI capability of AI PCs to boost productivity, maintain data privacy and streamline complex tasks. For example, writers and content creators can use local generative AI to refine articles and language or rapidly brainstorm ideas.

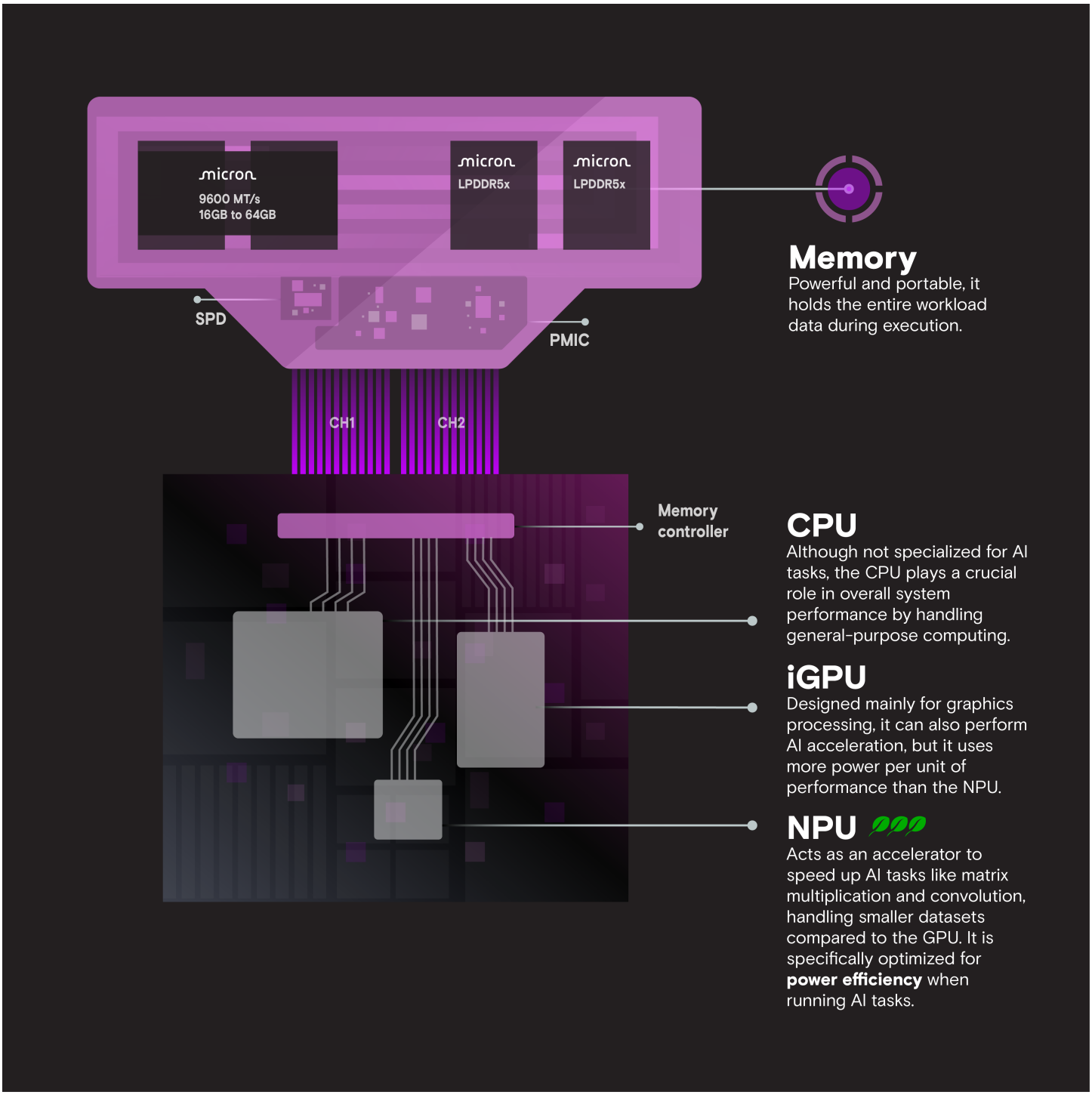

Figure 1: AI PC

AI applications

Below are several examples of AI applications whose performance can significantly benefit when run on an AI PC. These applications profit from the enhanced processing power, reduced latency (data can be processed locally instead of going to the cloud), and improved efficiency provided by AI PCs, making them a valuable tool for both personal and professional use even while on the go.

Real-time language translation: AI PCs can perform real-time language translation locally, providing faster and more accurate translations without relying on cloud services that require a network connection.

Enhanced video conferencing: AI PCs can improve video conferencing experiences by offering features like real-time background removal, noise reduction and automatic framing.

Content creation: AI PCs are excellent for content creation tasks such as video editing, image enhancement and automated content generation. Their specialized hardware allows them to handle these tasks more efficiently.

Speech-to-text transcription: AI PCs can transcribe speech-to-text in real time, making them useful for meetings, lectures and other scenarios where quick and accurate transcription is needed.

Security applications: AI PCs can run security applications like phishing detection and malware analysis locally, providing enhanced security and privacy.

AI assistants: AI PCs can run AI assistants locally, offering better performance and responsiveness for tasks like scheduling, reminders and information retrieval.

Advanced data analysis: AI PCs can perform complex data analysis tasks more efficiently, making them ideal for applications and professionals in fields like finance, healthcare and research.

Neural processing unit (NPU)

The integration of NPUs into PCs marks a major advancement in computing, enabling more efficient processing of AI tasks and paving a path for new AI-driven applications and features to emerge. Most users are familiar with CPUs and GPUs in their PCs. CPUs are designed for general-purpose computing tasks, while GPUs, including integrated GPUs (iGPUs), are equipped with numerous specialized cores capable of performing multiple operations simultaneously, focusing on performance. This architecture makes GPUs particularly well-suited for image processing, rendering graphics and AI computation.

NPUs, on the other hand, are specifically designed to efficiently handle AI and machine learning operations. They maximize performance per watt (power efficiency), making them ideal for tasks that require continuous processing without draining battery life. With optimal power efficiency, NPUs can deliver robust AI capabilities while enhancing the overall performance of mobile AI-enabled devices. This means AI tasks can be processed efficiently on battery power, allowing users to leverage advanced AI features without needing to be plugged into a power outlet.

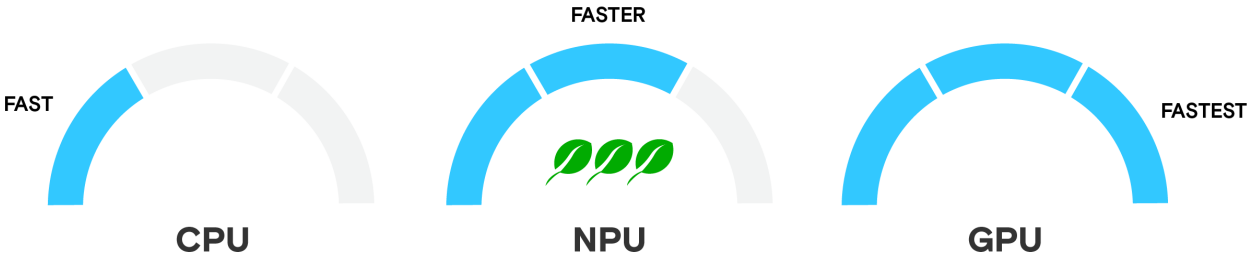

Figure 2: Relative speed of CPU, NPU and GPU for AI tasks

A relationship between performance, power and power efficiency

NPUs are faster than CPUs at computing AI tasks but not as fast as GPUs. However, NPUs use far less power than GPUs, making them power efficient. This is ideal for AI applications that require continuous processing, especially for users working where there is no connectivity. Additionally, while the NPU handles AI-related tasks, the CPU and GPU are freed up to handle their respective tasks, boosting overall system performance. In summary, NPUs complement GPUs by providing a more power-efficient solution for AI tasks, allowing for better overall system performance and longer battery life in AI PCs.

Overview of NPU architecture

The NPU is a specialized processor designed to accelerate neural network computations. It features multiple memory management units (MMUs), direct memory access (DMA) engines and multiply-accumulate (MAC) units, also known as hardware acceleration blocks, to execute multiply-accumulate operations fundamental to neural networks. The number of MAC units in an NPU determines its parallel processing capability, directly impacting its performance, which is measured in tera operations per second (TOPS). Understanding the parameters that contribute to the TOPS metric is crucial for gaining deeper insights into an NPU's performance.

To calculate TOPS, start with operations per second (OPS). Multiply is 1 operation, and accumulate is 1 operation, for a total of **2** operations per MAC unit per clock. Therefore, OPS equals two times the number of MAC units multiplied by their operating frequency. Finally, divide OPS by one trillion to convert OPS to TOPS. [1][5]

$$TOPS = \frac{2 * MAC\ unit\ count * frequency}{1\ trillion}$$

MAC unit count

A MAC operation involves two fundamental operations: 1) multiplication and 2) addition to an accumulator. Each MAC unit can perform one multiplication and one addition per clock cycle, effectively executing two operations per clock cycle (the “2” in the TOPS equation above). A given NPU has a set number of MAC units that can operate at varying levels of precision, depending on the NPU's architecture.

Operating frequency

Frequency refers to the clock speed, or cycles per second, at which an NPU and its MAC units (as well as CPUs and GPUs) operate, directly affecting overall performance. A higher frequency enables more operations per unit of time, resulting in faster processing speeds. However, increasing the frequency also leads to higher power consumption and heat generation, which can negatively impact battery life and user experience. The TOPS value for processors is typically quoted at their peak operating frequency.

Memory and token generation

NPUs rely on external memory to store large datasets and model parameters that exceed the capacity of on-chip memory. Optimized for matrix multiplication, NPUs are particularly effective in generating the initial token for AI and machine learning applications. Depending on power and latency requirements, subsequent processing and generation of tokens can be distributed between multiple CPU cores and the NPU.

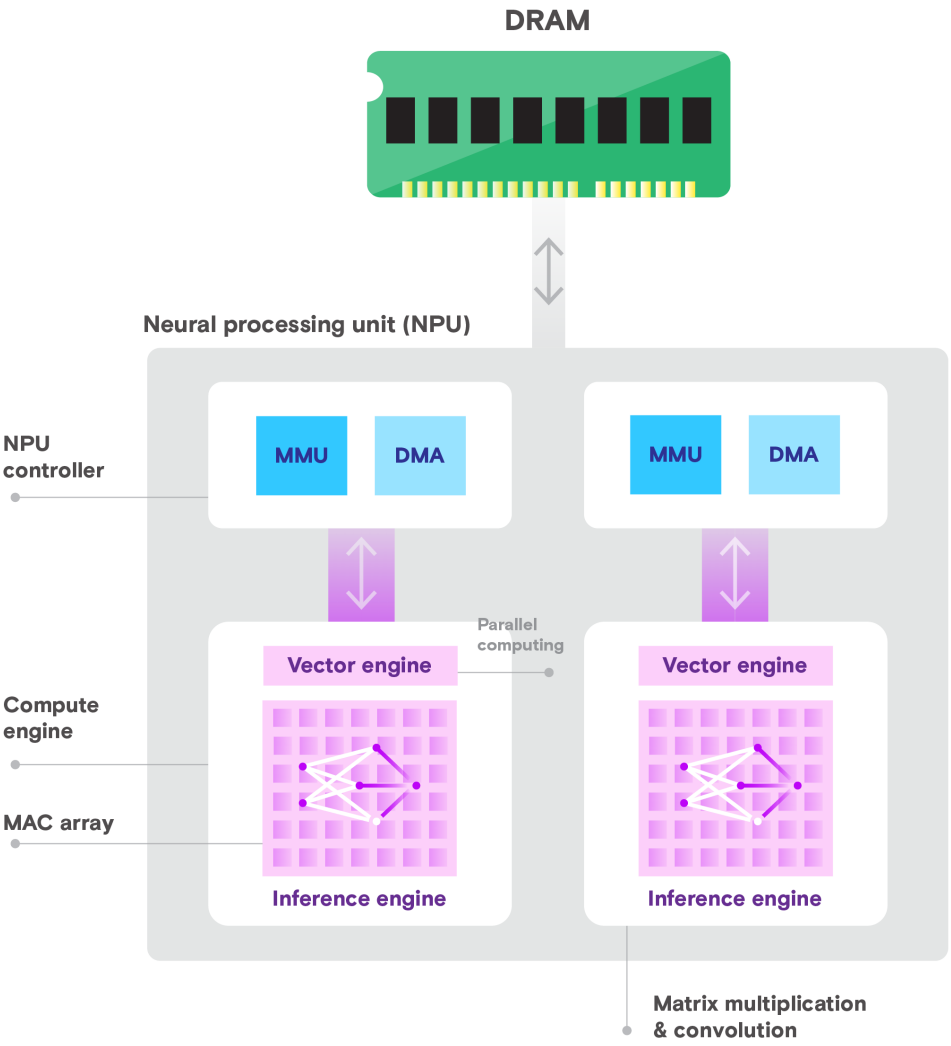

Figure 3: NPU architecture

The Intel® Core™ Ultra 9 Processor 185H NPU brings AI capabilities directly to the chip and is compatible with standardized program interfaces such as Intel® distribution of OpenVINO™ toolkit. It features a multi-engine architecture comprising two neural computing engines: the inference engine and the vector engine. [3][13]

Vector engine: This component is used for parallel computing on the NPU, enhancing its efficiency and performance.

Inference engine: This component executes high-level computation workloads (matmul, convolution), minimizes data movement and focuses on fixed-function operations.

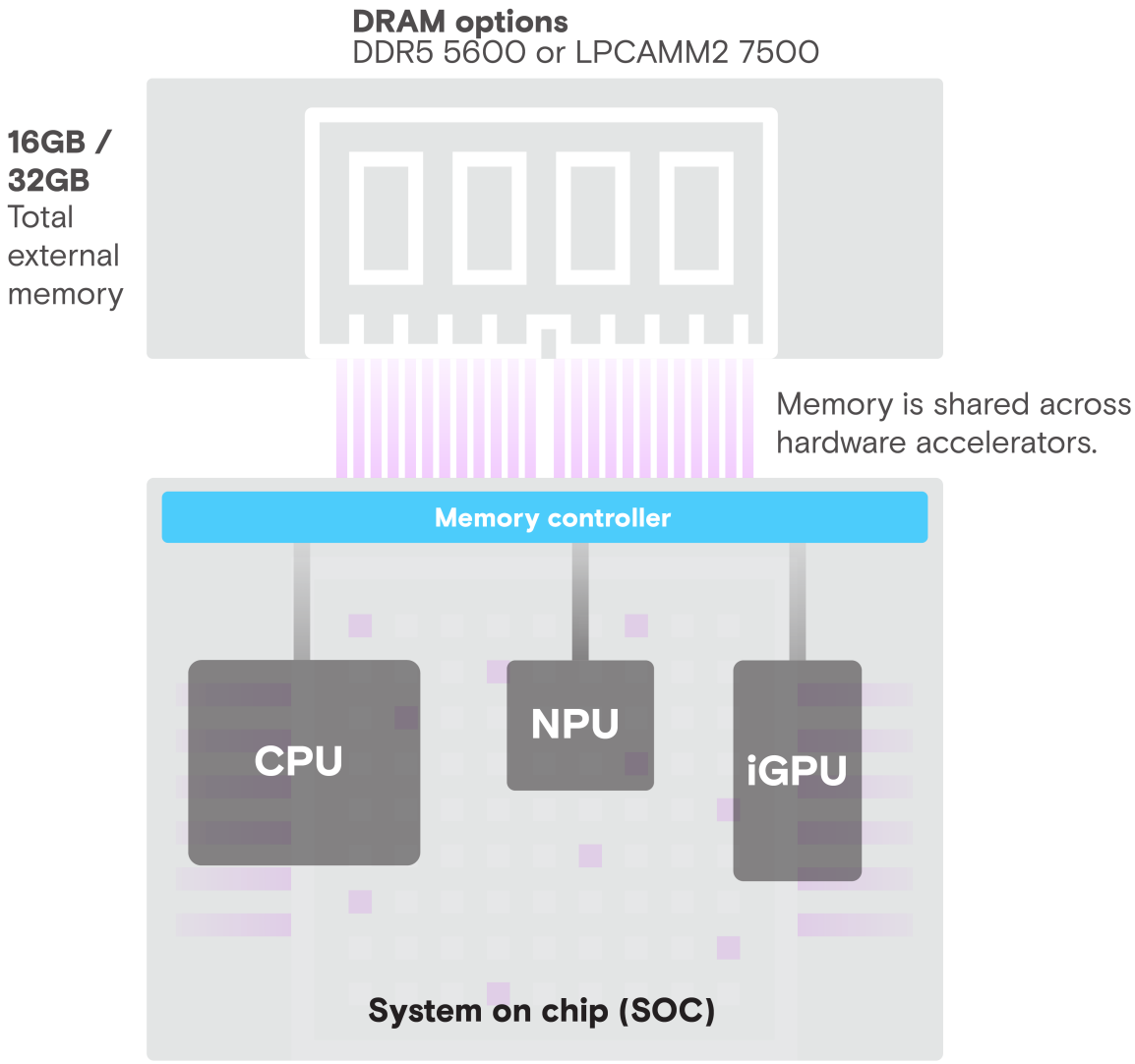

Figure 4: Hardware engines and external memory

The NPU and iGPU share system memory with the CPU as part of the unified memory architecture. In a unified memory architecture, the main memory is shared across all processing units — CPU, GPU and NPU. While very flexible, this shared memory can lead to performance bottlenecks due to the limited memory bandwidth available when these compute elements simultaneously need data from memory. Therefore, designing a memory system that supports higher bandwidth is critical for achieving optimal system performance in heterogeneous compute platforms.

Role of the memory subsystem

The previous section underscores the need for greater compute capabilities to handle AI workloads and the evolution of hardware accelerators for on-device AI inferencing. However, memory performance is as critical in achieving overall system performance as compute performance scales. Most AI and machine learning deployment use cases involve multiple AI models, each specialized for specific tasks, which are loaded into main memory during system boot and remain there. Applications can access these models as needed for AI inferencing. AI models such as large language models (LLMs) and image or video-based models like YOLO and Stable Diffusion, require a significant memory footprint. Swapping these models in and out of memory can adversely impact performance.

Because an AI PC has multiple processing components — where each accelerator, CPU, iGPU, and NPU, runs dedicated tasks in parallel and shares common memory access — this can lead to memory bottlenecks, which limits overall system performance. Additionally, multiple hardware accelerators operating simultaneously increase system power consumption. Therefore, both power consumption and memory performance are key factors to enable AI PCs. To support AI PCs, we offer various modular memory solutions, such as DDR5 SODIMM and LPCAMM2. Selecting the appropriate memory solution is crucial for achieving optimal system performance and battery efficiency. The following section provides a detailed analysis of suitable memory solutions for running AI workloads on AI PCs. The table below provides a high-level comparison between DDR5 and LPCAMM2 memory solutions:

	DDR5	LPCAMM2
Type	Double data rate 5 (DDR5)	Low-power compression attached memory module 2 (LPCAMM2)
Usage	Desktops and high-performance laptops	For thin and light laptops
Speed	Up to 5600 MT/s	Up to 7500 MT/s
Power consumption	Higher as compared with low-power variants	Up to 85% lower power than DDR5 for active cases[3]
Form factor	Small outline dual in-line memory module (SODIMM)	Smaller and thinner than traditional SODIMMs
Upgrade option	Upgradeable in supported systems	Easier to upgrade in thin laptops compared to soldered (BGA) DRAM
Size	Larger	Takes up 64% less space than two SODIMMS of DDR5 memory[4]

Table 1: Comparison of DDR5 and LPCAMM2

3. Based on results from the benchmark power analysis presented in this paper, comparing LPCAMM2 to DDR5.
4. Calculation based on comparison of the total volume of commercially available dual-stacked DDR5 SODIMM (32,808 mm^3) to LPCAMM2 (11,934 mm^3).

Metrics for measuring AI performance

With the rapid evolution of AI and machine learning models in recent years, the computational capability required for systems to seamlessly support those models continues to change. Microsoft recently released guidelines for Copilot+ PCs as a new class of Windows 11 AI PCs that are powered by a neural processing unit (NPU) with a performance of more than 40 trillion operations per second (TOPS). Below are four key metrics that we can use to measure AI performance.

Compute capability

Measured in TOPS (tera operations per second). Quantifies computational power of the compute elements — CPU, NPU or iGPU.

Speed

Measured iterations per second for AI model execution.

Latency

Measured the response time of AI model execution. Lower latencies improve user experience.

Throughput

Measured tokens per second for AI model execution. Often reported as a productivity metric.

Methodology

To understand memory's significance in AI PCs' overall performance, it is crucial to analyze key performance metrics such as bandwidth, latency, power consumption, memory configuration and capacity of memory solutions like DDR5 and LPCAMM2. These metrics impact system performance alongside the compute capabilities of heterogeneous hardware accelerators. Additionally, characterizing memory access patterns with different hardware AI accelerators helps identify bottlenecks in the execution pipeline caused by memory.

Given the multitude of parameters affecting workload performance at the system level, we analyze memory-specific parameters and compute parameters in isolation. This study is divided into two sections:

1. **Impact of SoC compute capabilities on memory**
    - DDR5 vs. LPCAMM2 performance with CPU
    - DDR5 vs. LPCAMM2 performance with NPU
2. **Impact of memory configurations**
    - 1 channel vs. 2 channels
    - 16GB vs. 32GB

The following set of workloads has been identified to comprehensively evaluate the performance of memory solutions and AI accelerators.

Category	Purpose	Workload or benchmark	Hardware accelerator
General-purpose workloads	Performance of general-purpose workloads, like office productivity, video calls, browsing and more. [6]	PCMark® 10 benchmark from UL	CPU
AI workloads	Measures performance of AI/ML workloads across various model architectures running on the CPU.	Geekbench ML, AIMark	CPU
AI workloads	Evaluates the performance of AI and ML models with different hardware accelerators.	Procyon® AI Inference Benchmarks	CPU, GPU, NPU
AI workloads	Analyzes memory-intensive AI use cases involving large data read/write accesses that require a larger memory footprint.	Meta Llama 3 8B, Mistral 7B Instruct	CPU, GPU

Table 2: Benchmarks and workloads

System setup

Power measurement setup

While some software tools can estimate DRAM power consumption, obtaining accurate measurements requires a specific hardware setup. We utilized a data acquisition tool (NI DAQ-6255) with a 16-bit analog input resolution and an 8 µs sampling interval to monitor the DRAM rails. The pinouts for the DRAM rails (VDDQ, VDD1, VDD2) from the interposer board beneath the DRAM module were connected to the DAQ Pre- and post-sanity tests confirmed that these hardware modifications did not cause any variations in performance or power consumption.

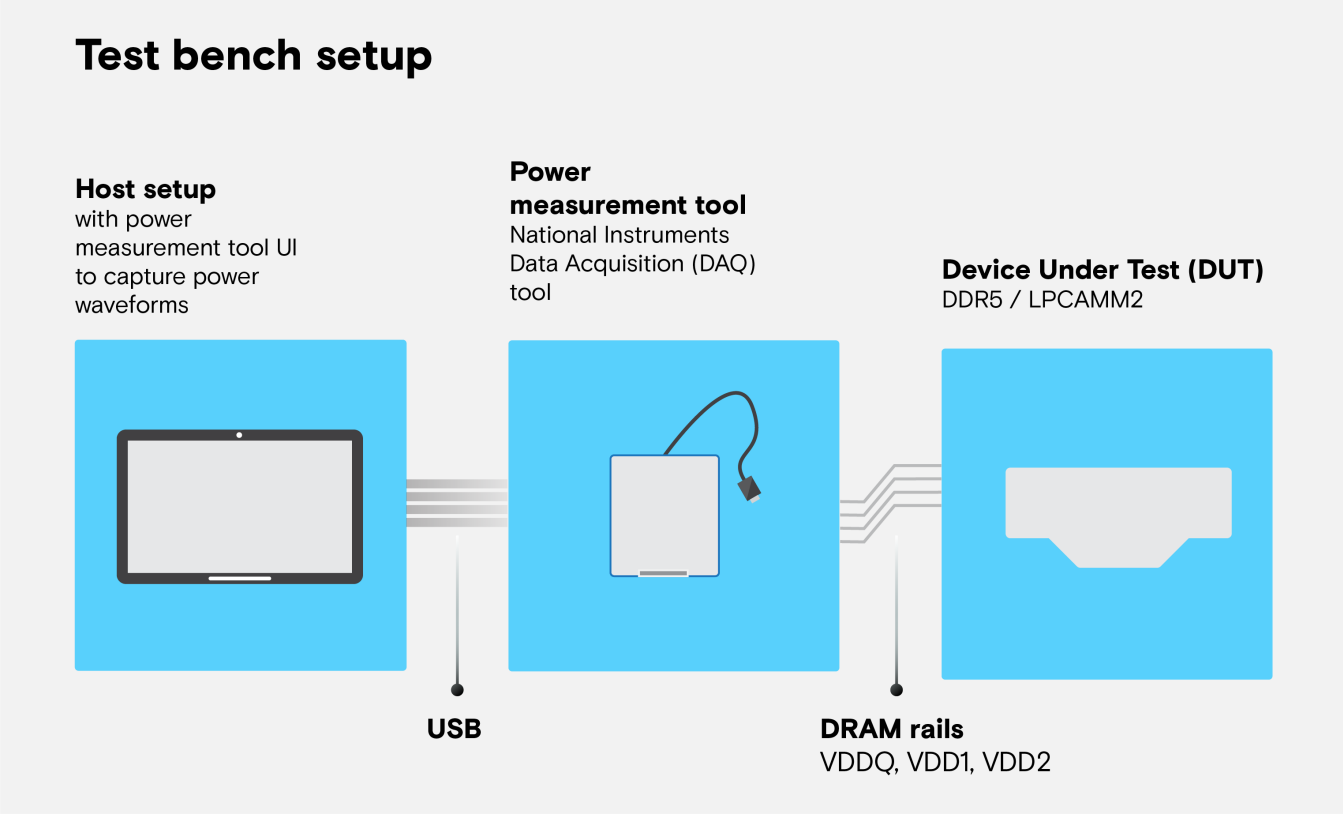

Figure 5: Test bench setup

Note: We compared power consumption in a two-channel (2CH) configuration for DDR5 at 5600 MT/s and LPCAMM2 at 7500 MT/s, both operating at their default system speeds.

To measure DRAM power consumption, we set up our power measurement on the motherboard at **VIN_BULK**, which is a 5 volt power source for both memory types. Also, by adding a shunt resistor, we calculated the voltage drop proportional to the current flowing through the memory modules. This enabled us to determine the power efficiency by dividing the performance score by the memory power consumption.

Hardware specifications

The table below shows the platform specifications used, which are common across all platforms.

Family	Clock speed	CPU, GPU, NPU	No. of cores, threads	L3 cache	Hyper threading	Memory capacity
Intel® Core™ Ultra 9 Processor 185H	2300 MHz, 5.1 GHz (Turbo Fmax)	6, 18, 11 TOPS	16 cores, 22 threads	Yes	Yes	32GB

Table 3: Hardware specifications Memory platforms [2]

Platform	Model no.	Memory type	Speed grade	Channels	Part number
Lenovo ThinkBook® 16	DVSPA5CP	DDR5	5600 MT/s	4x32bit	MTC8C1084S1SC56BD1
Lenovo ThinkPad® notebook	N8D0IKMI/21KWZC48US	LPCAMM2	7500 MT/s	4x32bit	MTD16C2O325N4FN026CY

Table 4: Memory platforms used for experiments

Synthetic benchmarks

Geekbench ML, AIMark, and PCMark 10

The general-purpose and AI workload benchmarks, as listed in Table 2, are evaluated on the test target using DDR5 and LPCAMM2 memory to compare performance scores and DRAM power consumption. For a clearer interpretation, the performance scores and power consumption results are normalized relative to DDR5 (set to 1.0).

For the benchmark power analysis graph, DRAM power consumption with LPCAMM2 memory is significantly lower compared to the DDR5 SODIMM, at approximately 12–13% that of DDR5 SODIMM, resulting in more than 85% power savings.

Comparing performance scores across the benchmarks, we see that LPCAMM2 has similar performance to DDR5 SODIMM for Geekbench ML and the PCMark 10 benchmark. However, for AIMark, LPCAMM2 scores are 20% higher in performance. Considering both memory parts were validated at the default configuration set by the system and with no custom operating frequencies, the power and performance data indicate that LPCAMM2 is significantly more efficient than DDR5.

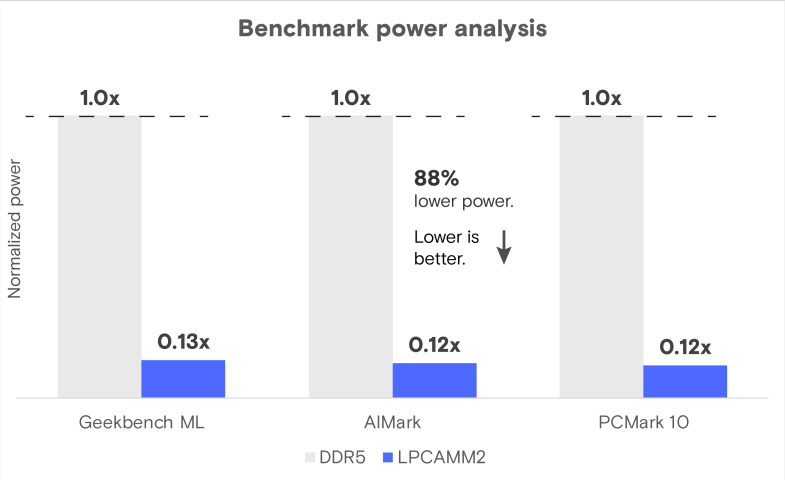

Figure 6: Power analysis for synthetic benchmarks

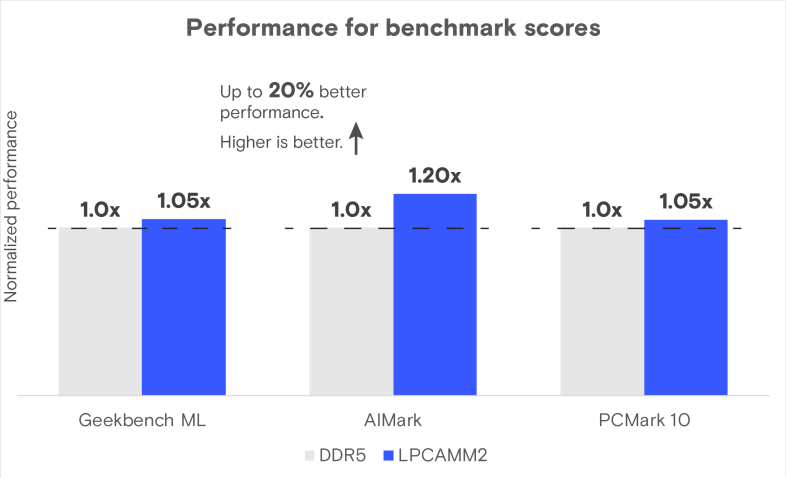

Figure 7: Performance for synthetic benchmark scores

Procyon® AI Computer Vision Benchmark

Procyon® AI Computer Vision Benchmark includes six carefully selected neural network models, each representing a key domain of future AI applications (refer to the table below). These models were chosen to represent foundational tasks that could evolve into a wide array of real-world applications for AI PCs. While the file size of a model correlates to the number of trained parameters, it does not determine the length of the inference time. The architecture of the model is the primary factor influencing inference time.

	MobileNetV3	**Inception-v4**	**ResNet-50**	**DeepLabv3**	**YOLOv3**	**ESRGAN**
Purpose	Image classification	Image classification	Image classification	Image segmentation	Object detection	Super resolution
Parameter	3.9 million	42.6 million	25.6 million	2.1 million	61.9 million	16.7 million
Model Size	14.9MB	162MB	97.8MB	8.06MB	236MB	63.8MB
CPU Inference T	1.0	1.0	1.0	1.0	1.0	1.0
iGPU Inference T	1.2	1.66	1.69	1.64	1.75	1.72
NPU inference T	1.43	1.78	1.71	0.96	1.8	1.77

Table 5: Procyon® AI Computer Vision Benchmark [6]

We use the CPU as a baseline (set to 1.0), as this is our reference for comparison of the other compute elements. We compared the performance gain when running these models with iGPU and NPU. Interestingly, four out of the six models are dramatically faster, about 1.6 times compared to the CPU, with the NPU outperforming the iGPU by a small margin. MobileNet, although showing less gain, still performs up to 1.4 times faster with the NPU. The NPU generally performed well across all models except when running DeepLab, where its performance was almost the same as the CPU. In this category, the iGPU still performed well.

Considering FP16 and INT8 formats

Procyon® AI Inference Benchmarks, utilizing the Intel OpenVINO framework, support execution on various hardware accelerators with different quantization levels. GPUs, optimized for graphics-related tasks, excel at floating-point operations and support integer precision, while NPUs are primarily designed for integer precision.

For completeness, we collected data using both INT8 and FP16 formats. The results from both formats lead to similar conclusions. Therefore, we present only the FP16 results for simplicity. The graphs shown illustrate our evaluation of performance, power consumption and power efficiency, comparing LPCAMM2 against DDR5.

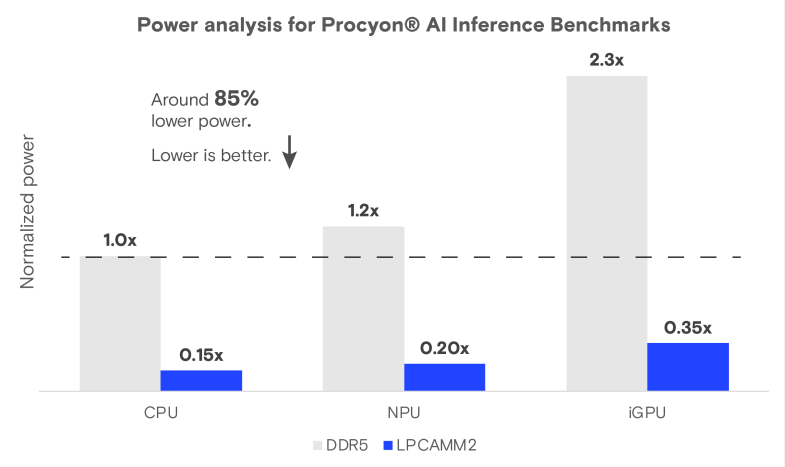

Figure 8: Power analysis for Procyon® AI Inference Benchmarks

The overall system-level performance difference between LPCAMM2 and DDR5 is marginal, However, LPCAMM2 is significantly higher in power efficiency (performance per watt.

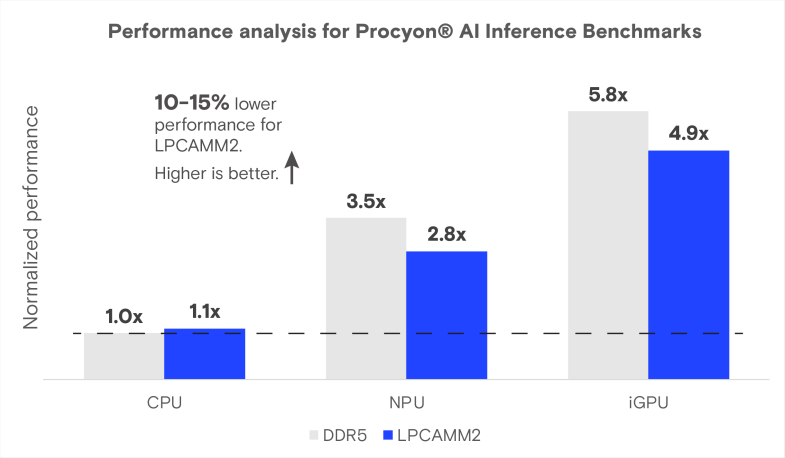

Figure 9: Performance analysis for Procyon® AI Inference Benchmarks

The power efficiency graph normalizes performance scores per watt to compare the power efficiency of LPCAMM2 against DDR5. LPCAMM2 delivers up to seven times better power efficiency than DDR5 for the same power consumption.

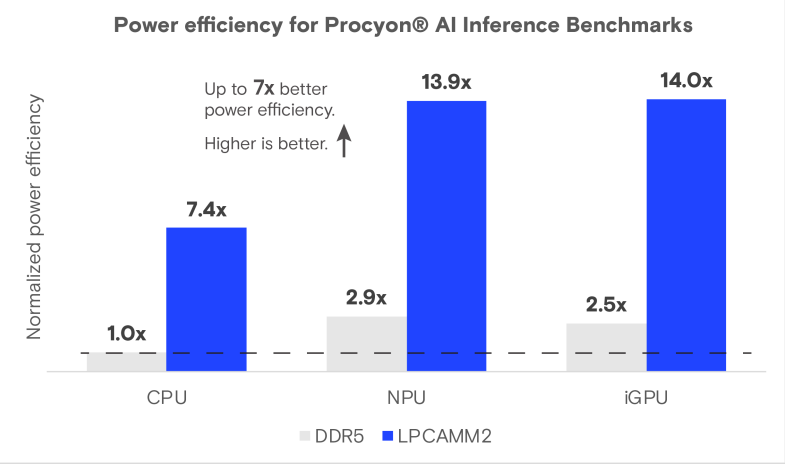

Figure 10: Power efficiency for Procyon® AI Inference Benchmarks

Large language models (LLMs)

In our analysis of large language models (LLMs) for evaluating memory power efficiency, we used LM Studio (0.2.25) to run the models on specific hardware accelerators. [7] This analysis is divided into two subsections: Meta Llama 3 on **CPU** only and Meta Llama 3 on **iGPU.** As the implementation of the Meta Llama 3 8B model for Intel NPU is still a work in progress, it was not considered for the study.

In the first subsection, "Meta Llama 3 and Mistral 7B Instruct on **CPU**," all workloads are executed solely by the CPU. In the second subsection, "Meta Llama 3 on iGPU," the inference tasks are delegated to the GPU.

The system running LPCAMM2 demonstrated four times better memory power efficiency than DDR5 for LLM inferencing workloads. This significant improvement is primarily due to the lower power consumption of LPCAMM2, which uses 57%-61% less active power and up to 80% less standby power compared to DDR5. For more information, refer to the LPCAMM2 product brief.

Meta Llama 3 and Mistral 7B Instruct on CPU

The AI benchmark results discussed earlier indicated that LPCAMM2 outperforms DDR5. However, the models used in the benchmark are not as large and complex as real-world AI models like Meta Llama 3 or Stable Diffusion, which has a memory footprint measuring up to gigabytes. It is crucial to evaluate the efficiency of LPCAMM2 against DDR5 with these more demanding workloads. The graphs to the right show the power and performance metrics measured with DDR5 and LPCAMM2 during CPU-based inference.

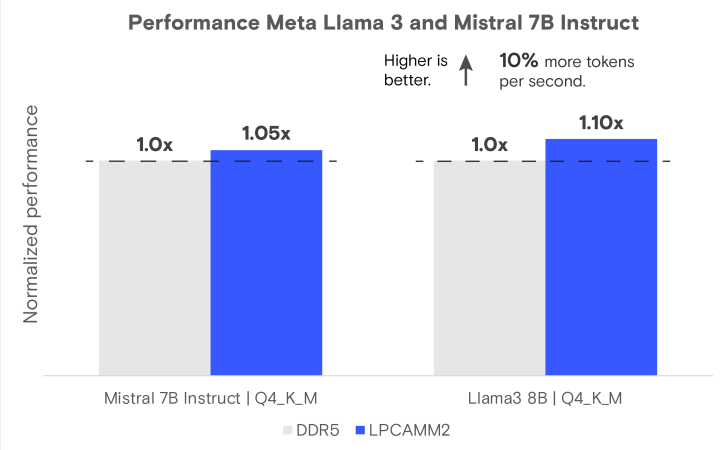

Figure 11: Performance Meta Llama 3 and Mistral Instruct | CPU only

From the performance and power metrics shown in the figures, it is noteworthy that LPCAMM2 delivers performance comparable to DDR5 but with significantly lower power consumption, more than **70%** lower power. The benchmark scores also reveal similar performance for LPCAMM2 and DDR5 on the CPU. The power and performance results are similar across Mistral Instruct and Meta Llama 3 when running inference on the CPU. Thus, for the following section on iGPU only, we investigate only Meta Llama 3.

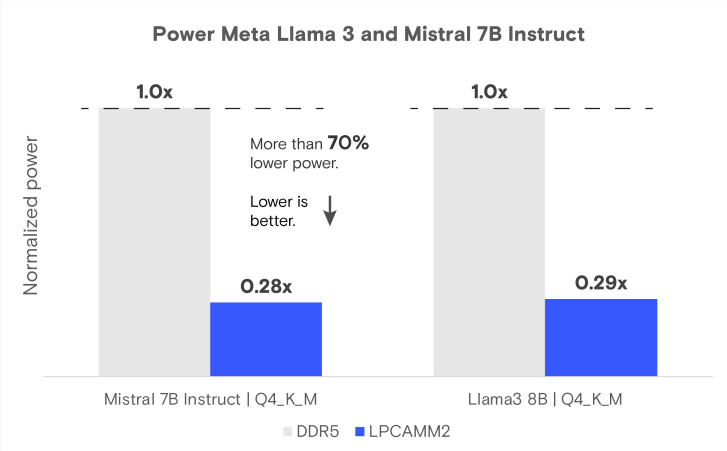

Figure 12: Power Meta Llama 3 and Mistral Instruct | CPU only

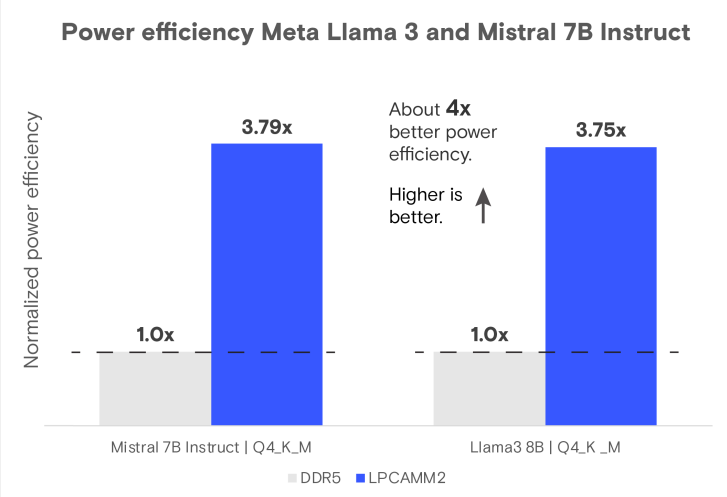

Figure 13: Power efficiency Meta Llama 3 and Mistral Instruct | CPU only

Meta Llama 3 on iGPU

To execute Meta Llama 3 on the built-in Intel® Arc™ GPU in the Intel® Core™ Ultra 9 Processor, we utilized SYCL (a direct programming language) and the Intel® oneAPI Math Kernel Library (oneMKL), a high-performance BLAS library. [8][10]

The integrated GPU (iGPU) leverages host-shared memory, requiring over 5.6GB for the Meta Llama 3 8B model — with 16GB or more of total host memory, where up to half of this memory is allocated to the iGPU).

Note: Detailed guidance for SYCL can be found in llama.cpp. [11]

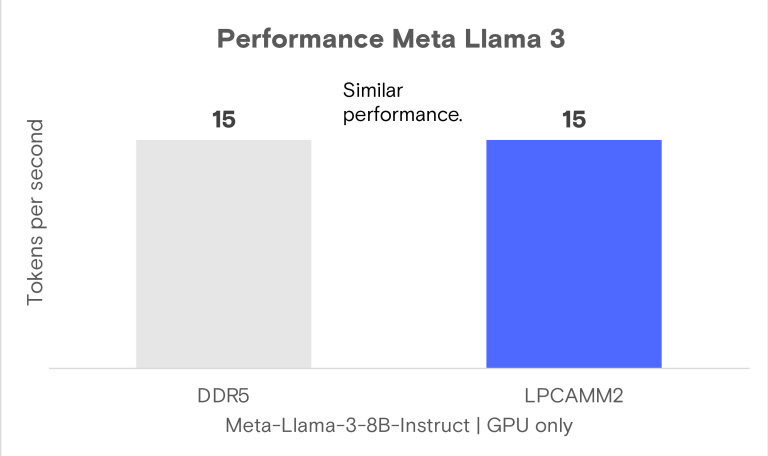

Figure 14: Performance Meta Llama 3 8B | iGPU only

Our study indicates that running inference on the iGPU is nearly twice as fast as running it on the CPU for the Meta Llama 3 8B model. Both DDR5 and LPCAMM2 achieve similar performance (tokens per second) with iGPU inference. However, significant improvements in DRAM power consumption are observed with LPCAMM2. Compared to DDR5, LPCAMM2 consumes **80%** less power.

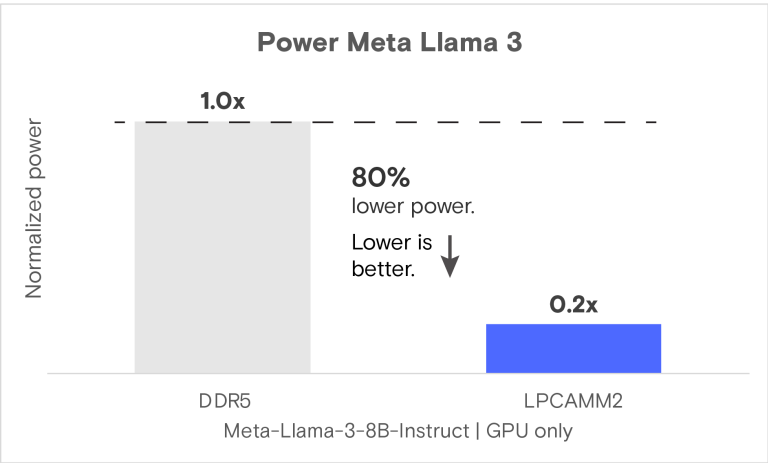

Figure 15: Power Meta Llama 3 8B | iGPU only

Overall, iGPU inference performance for the Meta Llama 3 8B model demonstrates a performance per watt ratio that is about **2.6** times better on LPCAMM2 systems compared to DDR5 systems.

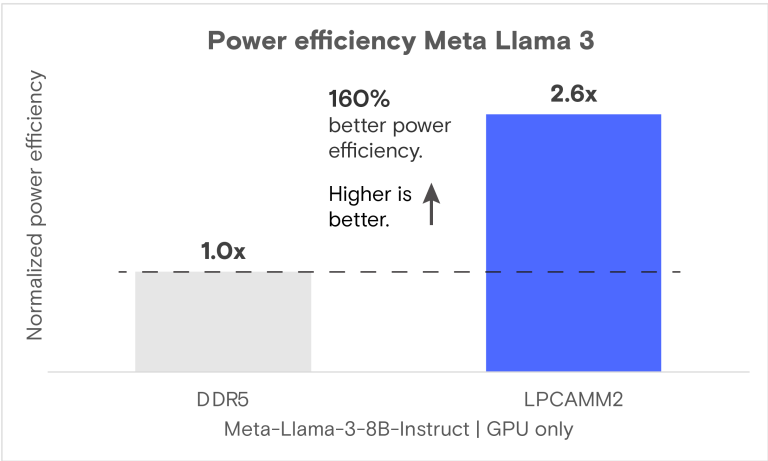

Figure 16: Power efficiency Meta Llama 3 8B | iGPU only

Power and performance of AI accelerators and memory

Based on the DRAM power and performance analysis across various workloads with different accelerators and memory types, we see that GPUs offer the highest performance, albeit with higher power consumption, while NPUs provide high performance at better power efficiency. On the memory front, LPCAMM2 offers considerable power savings and performance comparable to DDR5.

Illustrated below is the power and performance landscape of memory types with each of the available hardware accelerators (CPU, NPU and iGPU) based on the results obtained from the Procyon® AI Computer Vision Benchmark. The Y-axis represents performance, and the X-axis represents DRAM power consumption. The size of the bubble indicates overall power efficiency, with larger circles denoting better power efficiency. The optimal quadrant is the top left, representing the highest performance with the lowest power consumption. The data indicates that AI acceleration performed on either the NPU or the iGPU, in conjunction with LPCAMM2, achieves the highest memory power efficiency which is optimal for an AI PC.

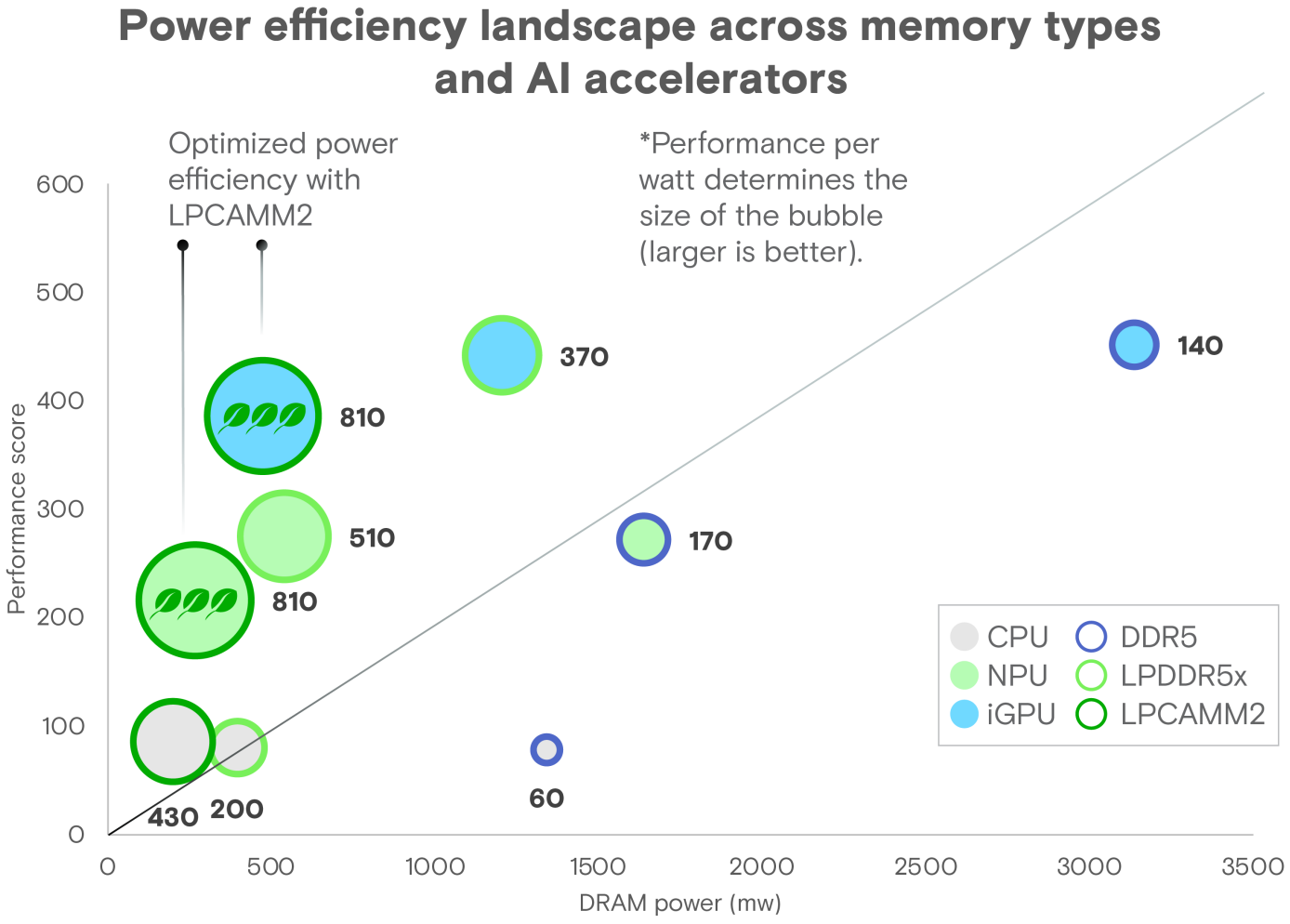

Figure 17: DDR5 versus LPDDR5/LPCAMM2 across AI accelerators for Procyon® AI Computer Vision Benchmark

Analysis of AI inference at the system level

Based on the power and performance analysis discussed earlier, LPCAMM2 demonstrated significantly higher power efficiency than DDR5 across AI workloads and hardware accelerators (refer to Figures 10 and 13). However, to determine which hardware accelerator excels in terms of power, performance and scalability, we need a deeper microarchitectural analysis of the system pipeline. For this, we use the Intel® VTune™ Profiler to capture the microarchitectural pipeline usage of the CPU. Though it only provides the pipeline usage for the CPU, we use this tool to characterize the various hardware accelerators by analyzing the microarchitectural usage of the CPU when it is accompanied in the system with either an NPU or iGPU. In this way, we can profile the overall efficiency of a *CPU-only* system or a *CPU + NPU* system or a *CPU + iGPU* system.

This investigation helps us to better understand where the bottlenecks occur during program execution and how they shift between memory and compute as the hardware accelerators become more capable. We captured the microarchitecture pipeline usage of the **Procyon® AI Computer Vision Benchmark** on each hardware accelerator (CPU, iGPU and NPU) with LPCAMM2. Our findings indicate that as the compute capabilities of the hardware accelerators increase, the workloads become more memory-bound, which requires a more capable memory subsystem.

Microarchitecture analysis using Intel® VTune™ Profiler

CPU utilization with hardware accelerators: An analysis of CPU utilization for the Procyon® AI Computer Vision Benchmark shows a significant reduction when using GPU and NPU accelerators. Specifically, CPU utilization decreases by 85% with the GPU and by 90% with the NPU.

This lower CPU utilization in CPU + NPU and CPU + iGPU configurations occurs because many compute tasks are offloaded to these accelerators, freeing up CPU cycles for other system tasks. This capability is crucial for an AI PC to support concurrent workloads, such as running traditional workloads at full speed alongside AI workloads.

Additionally, the CPU's average operating frequency increases when using the GPU accelerator, while it decreases by 10% (from 3.0 GHz to 2.7 GHz) when using the NPU. This indicates that a system with an NPU is likely to have a more favorable power profile. In summary, based on the analysis of CPU utilization and operating frequency, the NPU stands out as the optimal choice.

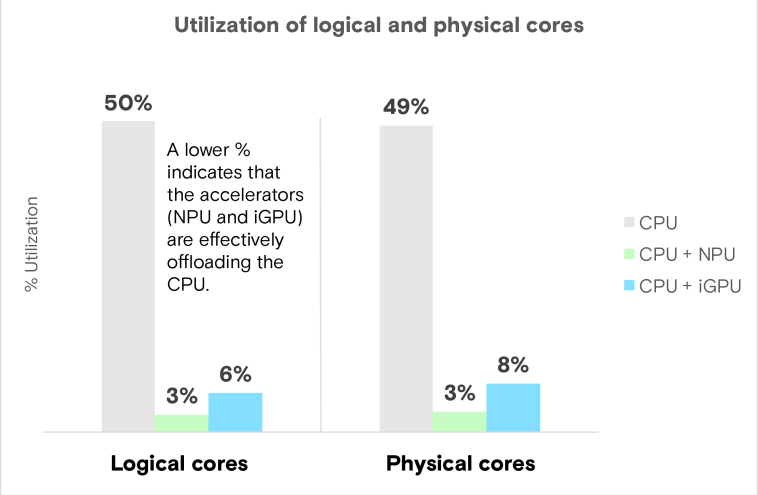

Figure 18: Utilization of logical versus physical cores on the CPU when the system has CPU-only, CPU + NPU and CPU + iGPU

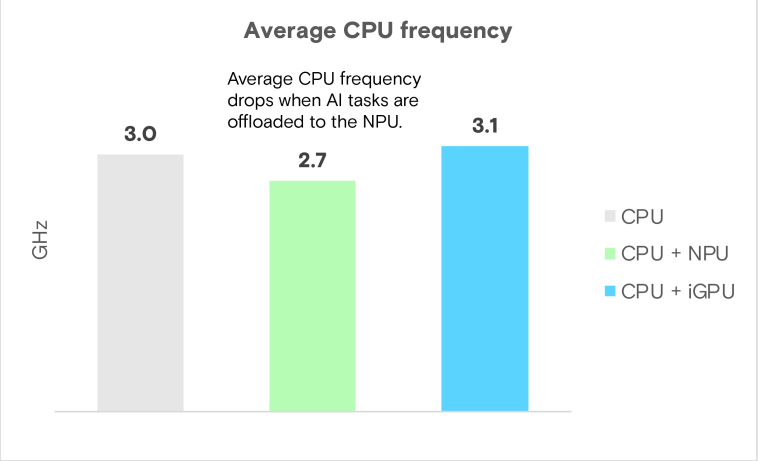

Figure 19: Average CPU frequency on the CPU when the system has CPU-only, CPU + NPU and CPU + iGPU

Validating NPU efficacy by CPU core usage: To further validate the effectiveness of the NPU, we analyzed the usage of the CPU's performance cores (P-cores), efficiency cores (E-cores), and low power efficiency cores (LPE-cores) while tasks were offloaded to the accelerators.

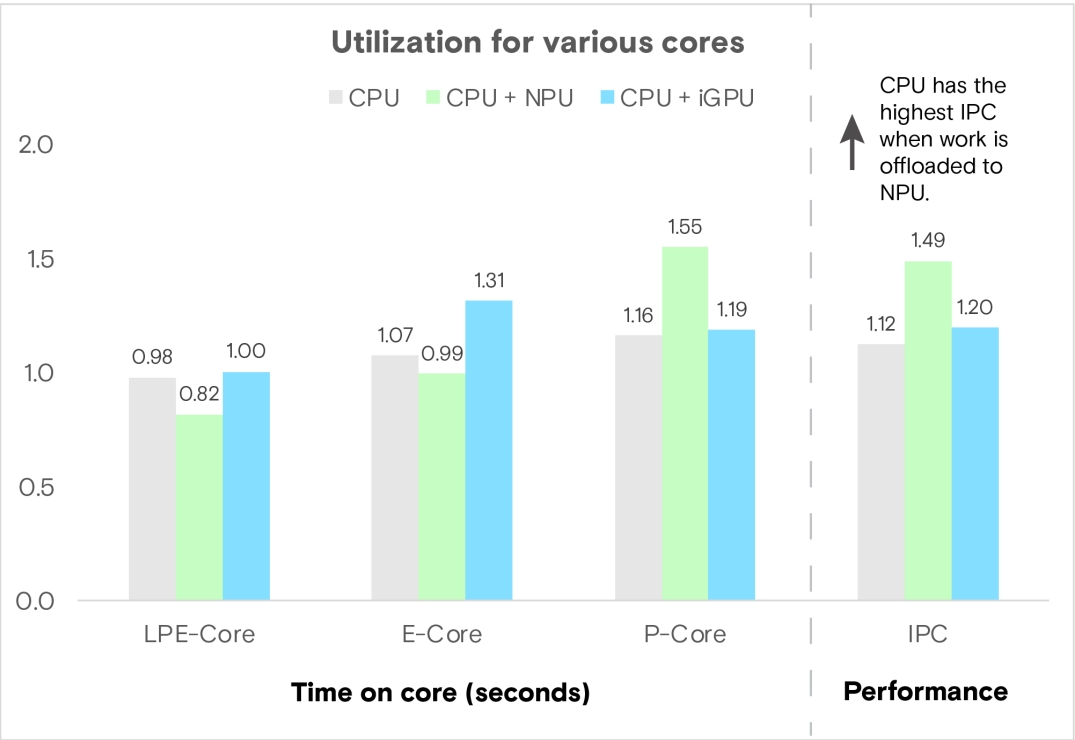

Figure 20: P-, E- and LPE-core utilization on the CPU when the system has CPU-only, CPU + NPU and CPU + iGPU

Analysis of core utilization and performance bottlenecks: From the chart above, we observe that the GPU primarily utilized the E-cores, while the NPU mostly used the P-cores during execution. Despite the frequent use of the P-cores by NPU acceleration, the average CPU frequency is lower compared to the CPU frequency with GPU acceleration. Additionally, LPE-core utilization is significantly lower in the NPU scenario compared to both the CPU-only and GPU scenarios.

The instructions per cycle (IPC) on the CPU, due to NPU acceleration, are 25% higher than with GPU acceleration and 30% higher than in the CPU-only scenario. This analysis further confirms the ability of the NPU to efficiently utilize the CPU pipeline.

Shifting performance bottlenecks: As accelerator capabilities increase, performance bottlenecks occur on both compute and memory. To investigate this, we analyzed the microarchitectural pipeline slot utilization with these accelerators to determine if the memory subsystem can effectively support the increasing compute capability of the accelerators.

A typical CPU execution pipeline is characterized by the number of slots a) retired, b) discarded due to bad speculation, or c) bound by front-end or back-end operations. Back-end boundedness can be either bound by memory or core (compute). Memory limitations can further be categorized into those bound by DRAM or cache, and within DRAM, it can be bound by bandwidth or latency.

This detailed analysis helps us to understand how the memory subsystem copes with the demands of advanced accelerators and where potential improvements can be made.

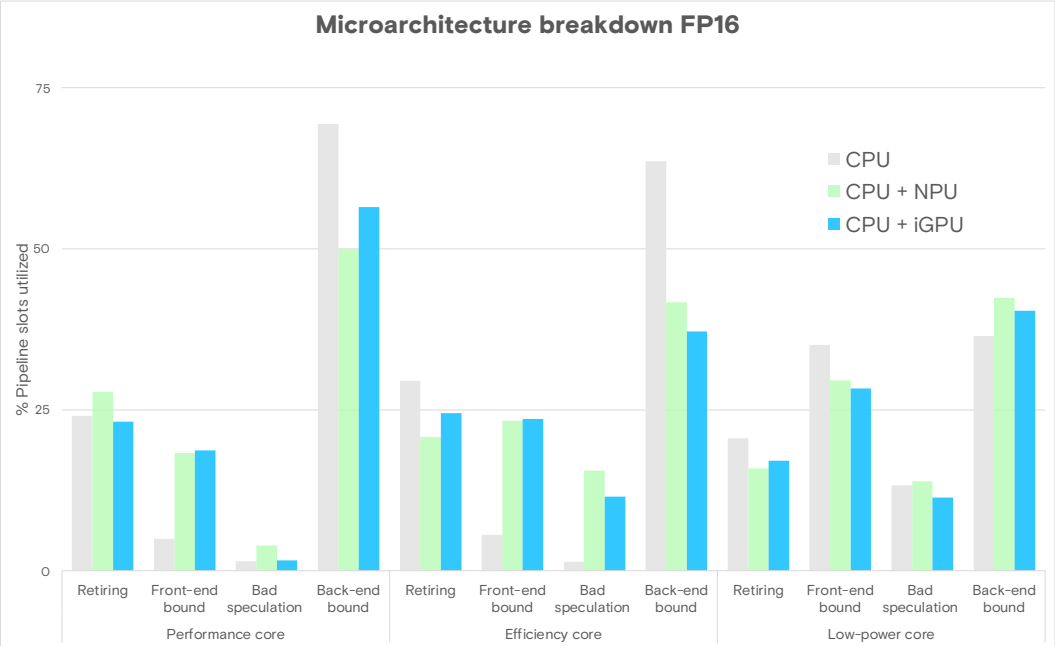

Figure 21: Microarchitecture pipeline analysis for FP16 precision and using OpenVINO

Analysis of core utilization and performance bottlenecks

Figure 21 shows that as execution shifts to high-performance cores, the number of pipeline slots blocked due to *front-end* operations and bad speculation decreases, while the number of retiring slots increases. However, the number of slots blocked due to back-end operations significantly increases. The number of retiring operations scales with CPU core performance in the case of NPU acceleration. Since processing back-end operations is a clear bottleneck with performance cores, we need to further analyze cache and DRAM latencies.

Figure 22 illustrates the distribution of clock ticks consumed due to reading data from the memory bus (bandwidth bound) and the clock ticks spent waiting for the data (latency bound). The number of clock ticks for latency is higher than for bandwidth in the case of iGPU and NPU, indicating that the CPU spends more time waiting for responses than accessing the data bus.

Overall, this suggests that as the compute capabilities of the AI accelerator increase, the workloads become more memory-bound (by both bandwidth and latency). Thus, memory plays a key role in achieving optimal performance for AI accelerators.

Given that the NPU yields better power efficiency (performance per watt) and LPCAMM2 offers up to 80% DRAM power savings compared to DDR5, the combination of LPCAMM2 and NPU is well-suited for AI PCs.

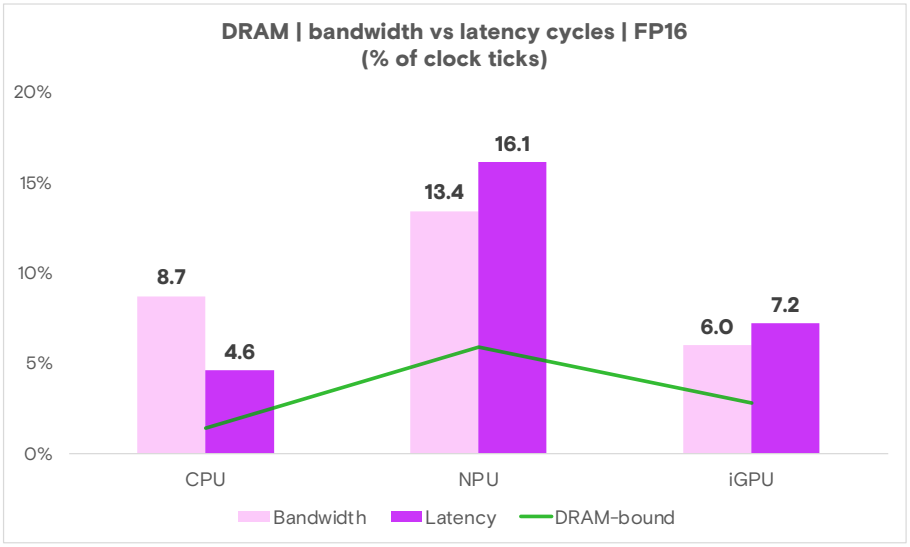

Figure 22: Clock ticks for DRAM bandwidth vs latency, showing boundedness

Memory utilization of AI models

In this section, we examine the system memory utilization across various AI applications and workloads. For systems running Windows OS 24H2, the baseline memory usage is approximately 6GB during idle states. When executing tasks in the Procyon® AI Computer Vision Benchmark, we used the CPU as a baseline (set to 1.0) for comparison among various models. We compared the memory utilization when each model is loaded onto the iGPU and NPU. Notably, the NPU consumes the most memory when running the task, with a GEOMEAN of 1.37x compared to the CPU.

	MobileNetV3	Inception-v4	ResNet-50	DeepLabv3	YOLOv3	ESRGAN
Purpose	Image classification	Image classification	Image classification	Image segmentation	Object detection	Super resolution
Parameter	3.9 million	42.6 million	25.6 million	2.1 million	61.9 million	16.7 million
Model size	14.9MB	162MB	97.8MB	8.06MB	236MB	63.8MB
CPU memory util	1	1	1	1	1	1
iGPU memory util	1.25	1.20	1.14	1.14	1.13	1.35
NPU memory util	1.33	1.33	1.29	1.36	1.44	1.47

Table 6: Procyon® AI Computer Vision Benchmark

Next, we simulate the execution of language models on an edge PC device using LM Studio to load the Meta Llama 3 7B model. We observe a substantial increase in memory utilization from an idle state of 6GB to 15GB when both models are loaded, resulting in a total memory usage increase of approximately 9GB.

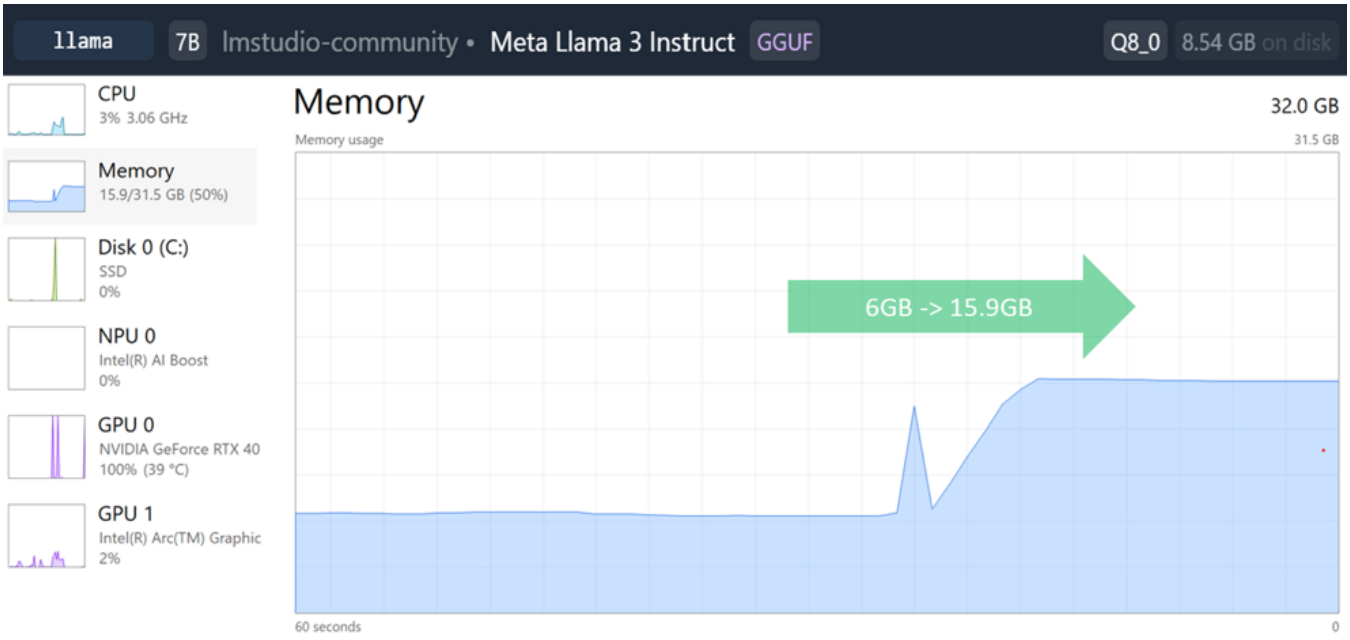

Figure 23: Meta Llama 3 Instruct memory utilization

Stable Diffusion

Finally, we evaluate the Stable Diffusion model with Intel® Core™ Ultra 7 Processor 165U across three distinct laptop power profiles selected in the OS:

- **Best power efficiency**: Activates battery saver mode, prioritizing power conservation over performance.
- **Balanced**: Offers an optimal trade-off between performance and power consumption, making it ideal for day-to-day workloads on battery power.
- **Best performance**: Configures the system for maximum performance, with higher power consumption expected. This mode is best suited for use when connected to the grid.

The Stable Diffusion process involves four key steps: text device, u-net device, u-net-neg device, and variational autoencoder (VAE) device. Each step plays an important role in ensuring the diffusion process is stable and accurate, where the most intensive steps are u-net device and u-net-neg device. There are three options, depending on the power profile selected by the user: best performance, best power efficiency or balanced power. With the advent of AI PCs, AI workloads such as Stable Diffusion can intelligently select the appropriate compute resource (CPU, iGPU or NPU) to align with user preferences (for example, optimized performance). This means workloads are intelligently placed rather than randomly selected, delivering a superior user experience. [9]

Power modes

	Best power efficiency	Balanced	Best performance
Text device	CPU	CPU	CPU
U-net device	NPU	GPU	GPU
U-net-neg device	NPU	NPU	GPU
VAE device	GPU	GPU	GPU

Table 7: Stable Diffusion power modes and compute element

DDR5 16GB | one channel (1CH): Running the Stable Diffusion AI model under different memory configurations reveals significant performance differences. With a 16GB memory configuration, the system's memory usage can exceed capacity, leading to memory swapping where less frequently used data is transferred to the solid state drive (SSD). This increases SSD traffic and results in slower performance due to the slower access speed of SSDs compared to RAM, causing longer loading and image generation times.

DDR5 32GB | dual channel (2CH): In contrast, a 32GB memory configuration allows the system to load the model entirely into DRAM, eliminating the need for memory swapping along with increasing memory bandwidth. This results in smoother and faster processing with minimal delays. Overall performance improved by 50%, resulting in quicker response times and more efficient handling of Stable Diffusion tasks. This illustrates the need for sufficient memory capacity to extract optimal performance for AI workloads on an AI PC.

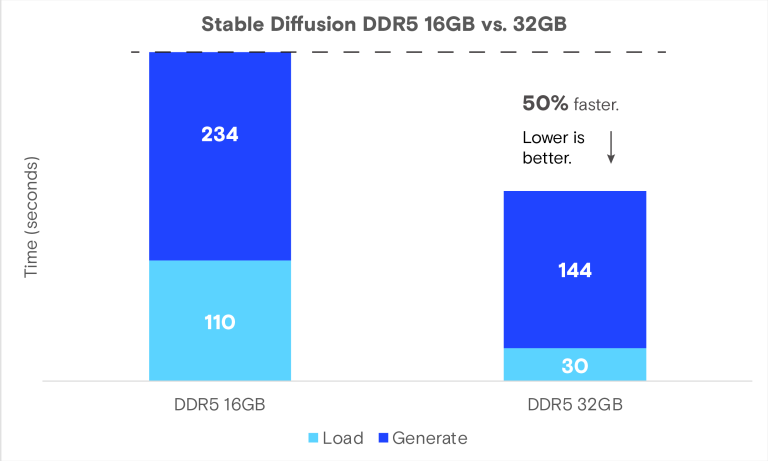

Figure 24: Stable Diffusion with DDR5 16GB (1CH) vs 32GB (2CH)

Single channel versus dual channel

For our experiment comparing single versus dual channel, our test system was equipped with Intel® Core™ Ultra 7 Processor 165U (code named Meteor Lake) and DDR5 SODIMM memory running at 5600 MT/s. Using the Procyon® AI Computer Vision Benchmark, we explored the impact of memory channel configuration (single versus dual) on inference performance. We ran tests across all hardware engines, including CPU, iGPU and NPU, on Intel OpenVINO.

CPU name	DRAM type	Data rate (MT/s)	Memory channel	CPU	NPU	iGPU
Intel® Core™ Ultra 7 Processor 165U	DDR5 SODIMM	5600	Single	109.0	448.0	314.0
			Dual	153.0	537.0	384.0
% increase for dual channel				40%	20%	22%

Table 8: Single vs. dual channel for DDR5 for Procyon® AI Computer Vision Benchmark

Channel count: As the overall DRAM bandwidth increases with the increase in the memory channels, the result is an increase in overall performance. We observed a **20-40%** increase in performance with 2CH DDR5 as compared to 1CH configuration. Increasing the memory bandwidth with two channels greatly enhances the system's ability to process AI workloads more efficiently and this improvement is vital for applications requiring high data throughput and quick access times. However, adding more memory channels would add to the overall cost and the complexity of the motherboard.

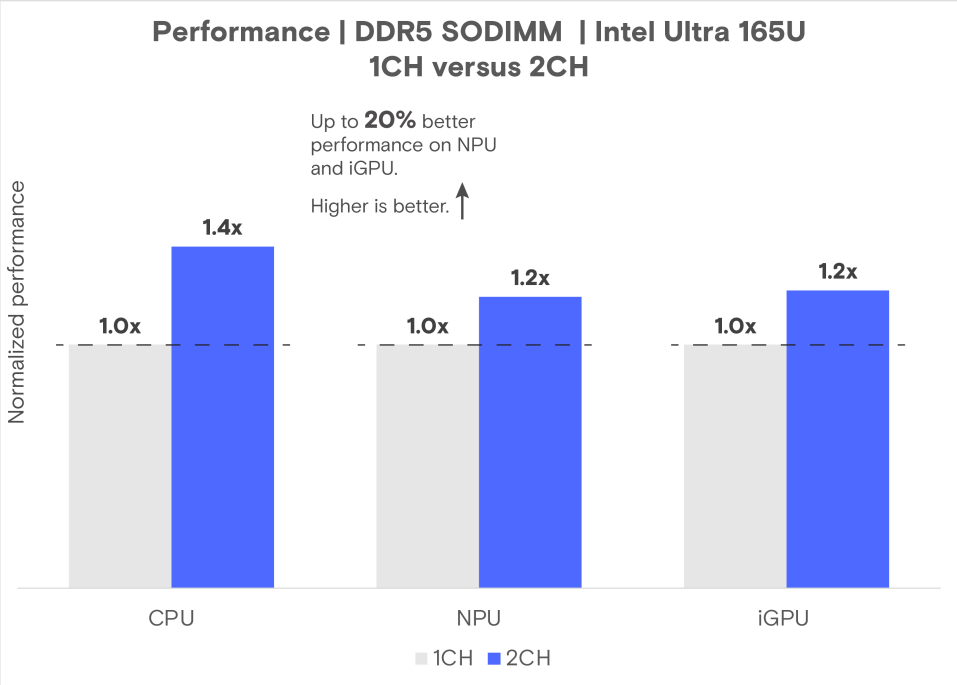

Figure 25: Performance of DDR5 SODIMM single-channel (1CH) versus dual-channel (2CH)

Conclusion

As AI models continue to evolve, growing in complexity and size, the importance of sophisticated memory solutions that can provide high performance and capacity while providing power efficiency will only increase. Specifically, advanced memory technologies like LPCAMM2, characterized by superior power efficiency and performance comparable to DDR5, represents a key advancement in meeting these demands and driving the evolution of AI PC architecture. Particularly important for AI PCs is their need to run multiple complex models simultaneously without excessive power consumption. This is essential for maintaining battery life and ensuring that users can run AI tasks seamlessly alongside more general applications. The complementary nature of pairing LPCAMM2 with an NPU is noteworthy, as it allows AI tasks to be processed efficiently on battery power. This enables users to leverage advanced AI features without needing to be plugged into a power outlet, making AI PCs outfitted with LPCAMM2 not only powerful but portable.

Moreover, as AI workloads become more sophisticated and model sizes increase, the demand for more memory capacity and bandwidth in AI PCs is becoming increasingly critical. Memory capacities greater than 16GB are essential to handle the intensive data processing and model training required by modern AI applications. From our analysis, we observed that many AI workloads are memory-bound and benefit from higher memory bandwidth and increased capacity. For example, as seen in Stable Diffusion tasks, which rely on large datasets and complex computations, increasing bandwidth (1-channel to 2-channel) and using higher memory capacity (going from 16GB to 32GB) significantly improves computation time. With more memory, these tasks can be executed more efficiently, reducing latency and enhancing overall system performance. It is important to note that system architecture plays a critical role in determining the optimal memory configuration for AI PCs. The choice between DDR5 and LPCAMM2 memory types depends on the specific requirements of the system architecture and the intended use cases. Key metrics to consider include power efficiency, memory latency and bandwidth and AI accelerator utilization.

For more information, refer to the LPCAMM2 product brief. Check out the product offerings for LPCAMM2 at: https://www.micron.com/products/memory/dram-components/lpddr/lpcamm2

References

[1] Burns, P. (2024, April 24). A guide to AI TOPS and NPU performance metrics. Qualcomm. https://www.qualcomm.com/news/onq/2024/04/a-guide-to-ai-tops-and-npu-performance-metrics

[2] Intel® Core™ Ultra 9 Processor 185H. Intel. https://www.intel.com/content/www/us/en/products/sku/236849/intel-core-ultra-9-processor-185h-24m-cache-up-to-5-10-ghz/specifications.html

[3] Lam, C. (2024, April 22). Intel Meteor Lake's NPU. Intel. https://chipsandcheese.com/p/intel-meteor-lakes-npu

[4] Toback, M. (2024, December 24). NPUs vs GPUs in Mini PCs: 5 Powerful Real-World Use Cases. Mini PC Technology. https://minipctech.com/npus-vs-gpus/

[5] The next step for power-efficient memory performance in client laptops. Micron.
https://www.micron.com/content/dam/micron/global/public/documents/products/product-flyer/lpddr5x-camm2-technical-brief.pdf

[6] AI computer vision benchmark. UL solutions. https://benchmarks.ul.com/procyon/ai-inference-benchmark-for-windows

[7] LM Studio. https://lmstudio.ai/docs

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., Polosukhin, I. (2023, August 2). Attention is all you need.
https://arxiv.org/pdf/1706.03762

[9] Alammar, J. (2022, October 4). The Illustrated Stable Diffusion. https://jalammar.github.io/illustrated-stable-diffusion/

[10] Zacarias, F., Palli, K., Vazhkudai, S., Grevelink, E. A memory perspective: The effects of fine-tuning LLMs with high-bandwidth memory. Micron.
https://www.micron.com/content/dam/micron/global/public/documents/products/product-flyer/llm-training-engineering-report.pdf

[11] GitHub. https://github.com/ggerganov/llama.cpp

[12] GitHub. https://github.com/ggerganov/llama.cpp/blob/master/docs/backend/SYCL.md

[13] Quick Overview of Intel's Neural Processing Unit (NPU). Intel. https://intel.github.io/intel-npu-acceleration-library/npu.html

micron.com/ddr5