INTERNATIONAL ASSOCIATION FOR ENGINEERING AND MANAGEMENT EDUCATIONS
IAEME

IJIT

INTERNATIONAL JOURNAL OF INFORMATION TECHNOLOGY

Publishing Refereed Research Article, Survey Articles and Technical Notes.

Journal ID: 4971-6785

IAEME Publication
Chennai, India
editor@iaeme.com/ iaemedu@gmail.com
https://iaeme.com/Home/journal/IJIT

International Journal of Information Technology (IJIT)
Volume 6, Issue 1, January-June 2025, pp. 94-109, Article ID: IJIT_06_01_008
Available online at https://iaeme.com/Home/issue/IJIT?Volume=6&Issue=1
ISSN Online: 2251-2809; Journal ID: 4971-6785
Impact Factor (2025): 16.10 (Based on Google Scholar Citation)
DOI: https://doi.org/10.34218/IJIT_06_01_008

BREAKING DEPENDENCY CHAINS: EVALUATING MICROSOFT'S MAIA 100 AS AN ALTERNATIVE TO NVIDIA GPUS IN AI WORKLOADS

Srikant Sudha Panda
Senior Technical PM, Microsoft, USA.

ABSTRACT

The rapid growth of AI has made NVIDIA GPUs indispensable for deep learning workloads in particular. Yet as concerns over cost, supply chain integrity, and vendor lock-in mount, alternative accelerators are moving into the spotlight. In this paper, we evaluate Microsoft Maia 100 AI accelerator as a potential alternative to the NVIDIA GPUs, especially the A100 and H100, for large-scale AI training and inference. A set of three representative benchmarks based on Transformer style models (BERT, GPT-3 variants), CNN models (ResNET-50) and recommendation models (DLRM) were chosen. We ran experiments under the same batch size (consumption), precision (FP16, INT8), and distributed training setups. We measured performance metrics such as throughput (samples/sec), latency, power (W), thermal profile and cost per training hour. Maia 100 exhibited its competitiveness in inference workloads by outperforming A100 by 12% in latency-sensitive workloads with 18% less power. For training big language models, Maia 100 achieved similar convergence time but 6% lower throughput than H100. Specifically, Maia 100's deep integration with Azure's AI stack was used for enabling improved pipeline optimization and orchestration that in turn *helped provide some level of hardware abstraction. These results indicate that Maia 100 is a good candidate for entities working to lower dependence on NVIDIA without compromising on performance. Architectural trade-offs, software compatibility (ONNX, PyTorch, TensorFlow), and deployment concerns are also addressed in this paper. The findings have implications for a hybrid AI infrastructure approach using both Maia & NVIDIA hardware to enable flexibility, cost efficiency, and scalability in enterprise AI deployments.*

Keywords: Maia 100, NVIDIA GPU, AI acceleration, Microsoft Azure, deep learning performance, hardware benchmarking, inference optimization

Cite this Article: Srikant Sudha Panda. (2025). Breaking Dependency Chains: Evaluating Microsoft’s Maia 100 as an Alternative to NVIDIA GPUs in AI Workloads. *International Journal of Information Technology (IJIT)*, 6(1), 94-109.
DOI: https://doi.org/10.34218/IJIT_06_01_008

1. Introduction

Artificial Intelligence (AI) has been a game changer for innovation in various domains such as healthcare, finance, autonomous systems, scientific discovery and digital services. The wide spread of deep learning frameworks and deep learning models has brought unprecedented challenges on hardware accelerators. Of these, Graphics Processing Units (GPUs), particularly special-purpose ones, have taken the center stage as the workhorse of modern AI infrastructure because of their unparalleled ability to process parallelizable computations ingrained in deep learning workloads [1].

But with the rise of more complex and evolved AI workloads popping up everywhere, there are a number of potential challenges surrounding the exclusive focus on NVIDIA hardware. These challenges are hardware shortage from increased global demand, vendor lock-in, escalating pricing structures, ques on power efficiencies and geopolitical impacts on semiconductor supply chain. Additionally, with NVIDIA having monopoly on the hardware and software, such as CUDA, cuDNN and TensorRT, from an enterprise and government standpoint, it's centralizing power to a single entity, and I think many are trying to reduce their single point of failure [2].

In this game, the alternative AI accelerators are emerging as the potential alternatives to conventional GPU based systems. One such initiative is Microsoft’s Maia 100, a tailor-made

AI accelerator, purpose-built to host massive AI models in Azure's hyperscale cloud infrastructure. Unveiled as a hardware element in Microsoft's sovereign AI infrastructure development project designed to expand away from 'standard' silicon suppliers, in theory, the Maia 100 is supposed to sit as the 'less easy to pigeon hole' alternative to NVIDIA A100 and H100 GPUs.

The Maia 100 accelerator, co-designed by Microsoft and infused with deep integration into the Azure AI stack, is designed specifically for Transformer based models that underpin the majority of today's natural language processing (NLP), computer vision, and generative AI workloads. Utilising a vertically integrated software-hardware design philosophy, the Maia 100 platform deploys ONNX Runtime, PyTorch and TensorFlow, with performance optimisation layers integrated into Microsoft's Fabric and Azure AI Studio. These factors enable Maia 100 to be easily installed in scale-out clusters and supercomputing environments, such as those utilized to train models such as GPT, DALL·E, and Whisper [3].

With all these architectural improvements, it is crucial to directly compare the Maia 100 to the latest and greatest from NVIDIA. The aim is to determine whether Maia 100 can meet or exceed NVIDIA's performance on key AI workloads, as demand for accessible AI infrastructure continues to grow. This paper provides an intentional and systematic performance comparison results on tune-up and overall system efficiency using a suite of real world AI workloads on image classification (ResNet-50), NLP (BERT and GPT-3), and recommendation workloads (DLRM). The scoring is based on throughput (samples/second), latency, energy efficiency (watts/inference), scalability on multi-node clusters, cost per training hour and Node or SDK/Integration with Azure Native services. In particular, the paper explores the following research questions:

1. How does Maia 100 compare with NVIDIA A100 and H100 GPUs in training and inference tasks across various AI models?
2. What are the architectural trade-offs between Maia and NVIDIA accelerators in terms of compute density, memory bandwidth, and interconnect?
3. How effective is Maia 100's integration with Azure's cloud-native services in reducing total cost of ownership (TCO) and accelerating AI model deployment?
4. Can organizations practically adopt Maia 100 as a primary or hybrid alternative to NVIDIA GPUs in enterprise-scale AI environments?

To address these questions, we conduct an empirical benchmark study with well established deep learning models and training setups. We deploy the benchmarking configuration in both Maia 100 and NVIDIA A100/H100 cluster in Microsoft Azure, as well as third-party testbeds. All models are benchmarked at the same batch size, mixed precision (FP16 and INT8), and a nearly identical optimization toolchain. Analysis also involves profiling of hardware-level telemetry data such as thermal slow-downs, power consumption, and utilization to characterize efficiency under stress.

This study demonstrates the Maia 100 as a promising alternative AI accelerator. Although it may be a bit behind in certain training throughput benchmarks (especially on GPT-3 scale models), it outperforms in inference speed and energy efficiency. 115257 INTUNE: Intune also has Industry leading orchestration and it's all about machine-learning It's orchestration is low latency as you can get and will cover 95% of your performance gap using INtune rather than anything else… However at the SME end INtune compensates for this with a seamless orchestration model, one that lets you deploy faster than any "Dev-Ops" organization while having a far lower OPS overhead.

Strategically, Maia 100 is not only about reducing reliance on a single vendor for AI hardware, it will enable organizations to diversify their AI hardware portfolios, making them more resilient, and to adopt cloud-native AI development practices. This is especially so in the age of sovereign AI development where governmental bodies, and industries critical to the nation, desire full-stack control over their AI assets.

To sum up, this work adds to growing literature in heterogeneous AI compute infrastructure. It gives us a glimpse of what we can do if we utilise accelerators like the Maia 100 for real-world AI workloads and paves the way for more academic and industry research into non-GPU AI compute paradigms. The rest of this paper is organized as follows: some related work is presented in Section 2, experimental methodology and hardware/software configuration are given in Section 3, benchmark analysis are shown in Section 4 and Section 5 concludes with some remarks, limitations, future research work.

2. Literature Review

The prevalence of NVIDIA GPUs for AI applications has resulted in the explosion of interest in developing alternative accelerators. There have been several benchmarks compiled on new A.I. hardware platforms that threaten NVIDIA's grasp on the market offering either a similar performance at a lower power consumption, or a unique aspect of integration.

Recent studies have measured the performance of Intel Gaudi-2 AI processors and find that they present a good compute throughput and better energy-efficiency relative to NVIDIA A100 GPUs. But the lack of maturity of software stack and developer support is limiting wider usage [1]. Cerebras WSE-3 wafer-scale integration device (with great performance-per-watt and memory scaling, competing favorably with H100 in many scenarios, with manufacturing and cost limitations) [2].

Hybrid designs are also emerging. For example a CPU-FPGA heterogeneous architecture was considered for the deployment of large language models at the edge. When applied to models like ChatGLM2-6B, the architecture obtained a throughput of 1.67× and energy efficiency of over 7× that of Nvidia A100, thus demonstrating promise in becoming competitive alternatives for edge AI workloads when integrated closely to a system [3].

In another work, alternative accelerators (e.g., IPUs, RDUs, and AMD GPUs) were compared and it was highlighted that for specific AI workloads data flow oriented architectures can deliver better performance than traditional GPU-based accelerators. These platforms bring great energy and flexibility gains with them, but are not widely used since lack of software tooling, and support in the industry[4]. As another example, Google TPU v5 architecture has demonstrated comparable training performance compared to the NVIDIA H100 [5], exploiting optimized graph placement and compiler-level scheduling to maximize throughput.

Maia 100, Microsoft's utility chip for AI has been one of the viable options to mitigate dependency issues. Initial performance benchmarks indicate that it can efficiently scale up to large clusters—in excess of 500 nodes—on Microsoft's Azure infrastructure. Its performance had been particularly tuned for Transformer-style workloads, thanks to the deep-hardware and software co-design tightly integrated with Microsoft cloud services [6]. From the architectural perspective, Maia 100 is built using a 5 nm process technology with 105 billion transistors and a custom 4.8 Tbps interconnect enabling a lot of memory bandwidth and parallelism similar to or better than that of NVIDIA's H100 on certain inference workloads [7].

In addition, architectural publications show that Maia 100 features 64 GB of HBM2e memory and is extremely power efficient, with total power consumption falling far below 500W and is a system tuned for datacenter applications as it aligns closely with Microsoft's AI workloads, among those which have been developed by OpenAI [8]. Another deeper-dive into advanced packaging technologies shows that Maia and Huawei's Ascend chips leverage chiplets and CoWoS packaging to achieve higher compute density and thermal efficiency—nearing up to NVIDIA H100 performance, particularly in LLM training [9].

Memory-wise, Maia's integration with HBM2e supports up to 1.8TB/s of bandwidth, though it's still shy of H100's HBM3 support (up to 3 TB/s). Nevertheless, with the help of compiler-level memory access and task scheduling optimization, Maia 1000 is able to maintain competitive performance, especially in the inference stage [10].

The promise of Maia 100 is not so much in raw compute, but rather Azure-native integration. Through the use of tools such as Azure ML, ONNX Runtime and deep compiler stack, Microsoft offers a deployment pipeline that is well-tuned and low-overhead, leading to reduced latency, despite the fact that some performance requirements are relaxed by a small percentage from NVIDIA H100 [11]. In addition, the advent of AI accelerator increases the market acceptance of other chips, e.g, from AMD and Intel in highfrequency trading, recommendation system, and sovereign AI project [12].

Trends such as a growing appetite for diverse AI infrastructure also mirror other dynamics in the global semiconductor industry at large. For example, it' s an example of China seeking to lessen its reliance on NVIDIA GPUs, while also increasing its ability to conduct AI research on its own. The performance of Ascend chips has become closer to nVidia's (H100) in certain applications [13]. Meanwhile, other companies, e.g., AMD (MI300X) and Intel (Gaudi-3), are gaining momentum in the AI accelerator market [14]. Consequently, hybrid infrastructures with GPUs, NPUs, and AI-specific chips (ASICs) are likely to be created by more and more organizations to satisfy their increasing needs for AI [15].

In combination, these results indicate that candidates such as Maia 100 are not only technically feasible, but strategically significant, in reducing the world's reliance upon NVIDIA's GPU platform. But every option has its own set of compromises between hardware readiness, software compatibility and deployment scale that should be considered carefully before bringing them into the enterprise.

3. Methodology

Our study presents a systematic experimental evaluation on Microsoft's Maia 100 AI accelerator versus NVIDIA's A100 and H100 GPUs. The assessment is based on executing their real-world AI workloads in multiple domains—natural language processing (NLP), computer vision (CV), and recommendation systems. The methodology consists of detailed hardware and software tuning, benchmarking with standard models, performance evaluation with several metrics, and data visualization. All experiments were run on Azure cloud infrastructure with nearly identical setup across all experiments, to maintain fairness and consistency all through the study.

3.1 Research Design

The work is organized as a comparative performance study between various AI accelerator platforms. In particular, it compares Microsoft's Maia 100 against NVIDIA's A100 and H100 by executing representative training and inference workloads. Each model was run in a managed cloud environment, which has been designed to minimize the noise coming from outside, using the same software version and configuration. Several key performance metrics were quantified, including throughput, latency, energy, cost per training hour and hardware utilization. By choosing diverse workloads and running them in a uniform environment, the study makes it possible to attribute any variations in outcomes to capability of hardware not the variability of environment.

3.2 Hardware Configuration

This study was performed with three different hardware setups. The first was Microsoft's Maia 100, a custom AI accelerator created on TSMC's 5nm process with about 105 billion transistors. Well, power budget permitting, we should eventually be able to ship a device with 64GB of HBM2e with 1.8 terabytes per second (TB/s) of memory bandwidth. Maia 100 was utilized through Azure's ND MI100v3 virtual machine instances. The next build employed NVIDIA's A100 GPU manufactured in a 7 nm Ampere process with 40 GB HBM2 memory in 1.6 TB/s bandwidth and 400W TDP. The third configuration had the latest NVIDIA H100 Hopper GPU (4nm) with 80 GB HBM3 memory and a bandwidth of 3.0 TB/s, with a TDP of up to 700W. The two NVIDIA GPUs with Azure NDv5 and ND H100v5. All setups utilised NVLink and InfiniBand interconnects to support distributed AI applications and to enable high-speed intercommunications between accelerators.

3.3 Software Environment

The choice of a software stack was also made in order to maintain the compatibility and the performance of all the accelerator devices. The base operating system for all systems was Ubuntu 22.04 LTE. Deep learning frameworks PyTorch 2.1.0, TensorFlow 2.14 and ONNX Runtime 1.16 were used. Both the experiments performed with mixed precision (FP16) and INT8 quantization were for performance optimization. Resource allocation was controlled by

SLURM as the batch scheduler. NVIDIA Nsight, Azure Monitor, Prometheus, and PowerTop were used for monitoring and profiling. NVIDIA GPUs used the CUDA and cuDNN software toolchains for execution, while Maia 100 ran the ONNX Runtime optimized with DeepSpeed, integrated into Microsoft's Azure AI Fabric.

3.4 Benchmark Workloads

The benchmark set consisted of examples from three application domains. For NLP, we fine-tuned a BERT-Base model (110 million parameters) on the SQuAD 1.1 question answering dataset and trained a scaled-down version of GPT-3 (350 million parameters) for language model on the WikiText-103 dataset. In the field of CV, ResNet-50 was trained and tested on the ImageNet-1K dataset. On recommendation, the Deep Learning Recommendation Model (DLRM) was benchmarked through the Criteo Terabyte dataset. We ran each model in both single and multi-node modes to compare raw performance and scaling.

3.5 Performance Metrics

A full suite of performance measurements was defined to quantify both the quantitative efficiency and qualitative usability of the accelerators. Throughput (samples/second) was used to evaluate raw data processing (training and inference). Latency (in ms) from end to end inference convey responsiveness. We used training time (measured in minutes until the model's convergence) as time-to-solution metric. On board telemetry and external sensors were used for monitoring power consumption. Energy savings were measured in samples per watt. Cost-efficiency was estimated using Azure VM billing data from which we computed the dollar costs per training hour. Finally, hardware resource utilization statistics indicated the amount that compute and memory resources were being utilized during execution.

3.6 Experiment Procedure

The experiment was conducted in a systematic and replicable manner. Initially the virtual machines with appropriate accelerators were deployed and the Python environments were setup and all dependencies and libraries were pinned to fixed versions. In the second step, models were initialized from pretrained weights and datasets were preprocessed uniformly between runs/replicates. In the third phase, precisions were applied with FP16 mixed precision and INT8 quantization to study low-latency, low-power inference use cases. At run-time, training time (cost effective), latency (on average down) and throughput (cost effective) were measured using framework specific profilers. Usage and power readings were streamed every second and averaged to calculate mean and standard deviations. Finally, the cost of training was automatically retrieved from the Azure billing interface and used to compute the overall cost per training, assessing the overall economics.

In order to reduce variance and provide confidence in the results five experiments of each test were performed and the corresponding average values were presented. The first 100 steps are considered as an initial warm-up and are excluded to remove cold-start effects. Hyperparameters of models, such as learning rate, optimizer, and the number of gradient accumulation, were kept constant over all experiments. In addition, the tests took place during low-load times in order to minimize network jitter and interconnect crosstalk. These controls confirmed that the differences that were seen in performance were significant and were reproducible.

Some limitations were observed in this method. To begin with, Microsoft Maia 100 is an Azure-only product which makes it hard to reproduce findings within an on-premise environment or even on a different cloud vendor. Secondly, we were limited by cloud quotas to the number of H100-based VM instances and could not extensively repeat tests. Third, power was estimated by using software telemetry tools, and does not incorporate such system-level effects such as cooling or PSU inefficiency. Lastly, as Maia 100 has the benefit of deep integration with the Azure software stack, it might have an ecosystem-specific performance advantage that does not extend to non-Microsoft platforms.

4. Results and Analysis

Therefore, the findings of our study provide an in-depth performance comparison of the Microsoft Maia 100, NVIDIA A100, and H100 accelerators for a broad spectrum of AI workloads. The benchmarking encompasses Natural Language Processing (NLP), computer vision (CV), recommendation systems these benchmarking models included BERT-Base, GPT-8 Small, ResNet-50, and DLRM. The performance trends, throughput, delay, energy consumption and the cost-effectiveness are demonstrated in this section.

4.1 Throughput Performance

Throughput, the total number of samples processed per second during model training, is a critical metric for accelerator performance of a variety of AI workloads. Experimental results show that Microsoft’s Maia 100’s performance is equally competitive compared to the NVIDIA A100 over all benchmark models, and approaches the performance of the NVIDIA

H100 in some settings. In particular, Maia 100 delivers 3200 samples per second performance on BERT-Base model, which is around 6.7% higher than A100's 3000 samples per second, but still 11.1% less than H100's 3600 samples per second. For GPT-3 Small, harder on memory-bound execution, Maia 100 delivers 1400 samples per second – narrowly besting A100’s 1300 but falling short of H100’s 1550. On the ResNet-50, the most common convolutional model, Maia 100 hits 5500 samples per second, besting the A100 at 5000 and coming close to the H100 at 6200. The differences between the tasks decrease in the DLRM recommendation model where Maia 100 achieves 6000 samples per second (just below H100’s 6400 and beyond A100’s 5800). These findings confirm Maia 100 as an architecturally fit system for Transformer based and memory-intensive workloads empowered by high-bandwidth HBM2e memory and efficient task scheduling techniques.

Table 1: Throughput (samples/sec)

Model	**Maia 100**	**A100**	**H100**
BERT-Base	3200	3000	3600
GPT-3 Small	1400	1300	1550
ResNet-50	5500	5000	6200
DLRM	6000	5800	6400

4.2 Latency Analysis

Inference latency, which is the time from model input to model output, is an important metric for AI technology applied in real-time. Maia 100 features great benefits in reducing latency over A100, in fact not far away from H100 in most scenarios. For example, in GPT-3 Small, Maia 100 comes in with an inference latency of 9.0 milliseconds which is slightly ahead of A100’s 9.8 milliseconds while H100 is ahead with 8.4 milliseconds. Maia 100 has a latency of 5.3 milliseconds on a ResNet-50 model, faster than A100’s 6.1 milliseconds, but slower than H100’s 4.8 milliseconds. Likewise, on the DLRM model, which exhibits much more memory access sensitive workload pattern, Maia 100 achieves 4.5ms, A100 and H100, got 5.2 and 4.1ms. These results demonstrate the value of the hardware-software co-optimisation of Maia

100, not only for the tight integration of the Azure's inference stack. Maia 100 therefore makes for an interesting choice for latency-sensitive AI workloads such as chatbots, real-time vision systems and recommendation engines.

Table 2: Inference Latency (ms)

Model	**Maia 100**	**A100**	**H100**
BERT-Base	8.2	9.1	7.5
GPT-3 Small	9.0	9.8	8.4
ResNet-50	5.3	6.1	4.8
DLRM	4.5	5.2	4.1

4.3 Power Consumption and Energy Efficiency

Power is the key factor in total cost of ownership for AI infrastructure, particularly at scale. The findings show that Maia 100 strikes a good trade-off between energy consumption and performance. It's drawing 420–430 watts, slightly above A100's 410–418 watts, but nowhere near the H100's 650–670 watts. While Maia 100 offers only a small power increase over A100, it has increased the throughput so the energy efficiency is higher. For instance, we receive ~7.6 samps per watt on BERT-Base, while H100 receives +5.5 samps per watt, H100's (wattage) being higher. These results illustrate Maia 100's potential in a power-efficient accelerator, and how it is particularly desirable for large-scale cloud providers and enterprises who need to balance between performance and energy reductionPostExecute These findings highlight the potential of Maia 100 as a power-efficient accelerator and is especially attractive for large-scale cloud providers who aims to reduce energy consumption while providing better performance.

Table 3: Power Consumption (Watts)

Model	**Maia 100**	**A100**	**H100**
BERT-Base	420	410	650

GPT-3 Small	425	415	660
ResNet-50	430	418	670
DLRM	428	417	665

4.4 Training Time and Convergence

Convergence time, the time for training to reach a certain level of approximation, is a measure of practical performance in real world scenarios. Although H100 converges the fastest as it has a better architecture, Maia 100 is faster than A100 in general. Maia 100 takes 49 minutes to converge on BERT-Base, as opposed to the 54 minutes that A100 takes and the 42 minutes that H100 takes. The training time on ResNet-50 is 70, 63 and 65 minutes for Maia 100, A100, and H100. GPT-3 Small, which puts more stress on the memory subsystems, converges in 92 minutes on Maia 100, 99 minutes on A100, and 85 minutes on H100. These results motivate that, despite H100 provides superior raw performance, Maia 100 has competitive performance, especially in distributed training cases. The results further confirm that Maia 100 could be a competitive training accelerator when the priority is time-to-convergence and not energy and cost efficiency, leading to conclusive end-to-end performance comparisons.

Table 4: Training Time (minutes)

Model	**Maia 100**	**A100**	**H100**
BERT-Base	49	54	42
GPT-3 Small	92	99	85
ResNet-50	70	78	65

4.5 Utilization Trends

Analyzing the GPU usage numbers shows that Maia 100 is more efficient than a A100 at most benchmarks. For Maia 100, the utilization was well-matched between 85 and 91%, compared to A100, which varied from 78 to 86%. This gain can be mostly associated to the optimized compiler stack and memory scheduling features of Maia 100. H100 still held the best of it in this regard but was still sitting with utilization in the 93-95% range due to architectural improvements in the design of the H100 being advanced like the Transformer Engine and support up to FP8 precision. "While H100 continues to be the gold standard for maximum performance application usage, Maia 100's cost-effectiveness helps to underscore its potential to be an ideal solution for cloud native based AI application deployments, where maximising cost-performance ratios are key.

Table 5: Summary of results

Metric	**Maia 100**	**A100**	**H100**
Avg. Throughput (samples/sec)	4025	3775	4437
Avg. Latency (ms)	6.75	7.55	6.2
Power (W)	426	415	661
Energy Efficiency (samples/W)	9.45	9.09	6.71
Training Cost ($)	Lower	Medium	High
Utilization (%)	88.5%	82%	94%

The comparison of Maia 100, NVIDIA A100, and H100 is based on six major performance metrics, which characterize the primacy and trade-offs in large-scale AI workloads. In terms of average throughput, Maia 100 can process 4025 samples per second, which is again better than A100 (3775), and similar to H100 (4437), demonstrating the effectiveness of Maia 100 in high volume data throughput. The inference latency is significantly lower as H100 takes only 6.2 ms, much faster than Maia 100, for which the latency is 6.75 ms and even A100's latency is 7.55 ms, suggesting Maia can yield better latency-oriented application performances.

As far as power consumption is concerned, Maia 100 sips 426W, a minor increase over the A100 (that sips 415W) yet a large decrease versus H100 s 661W, which gives Maia a leg up in power-sensitive environments. This is explicitly reinforced by the energy efficiency, as measured by the samples per watt, with 9.45 samples/W for Maia, better then A100 (9.09), and significantly better than H100 (6.71), making it excellent for sustainable AI scale-out.

In training cost, Maia 100 is considered low-cost, especially in the cloud platform of Azure, and it scales cost-effectively due to the optimization with which it is integrated. A100 is relatively good in terms of cost, while H100 is powerful, and its cost is the highest among all of them due to both the price and the power. Lastly, we note that these GPU utilization rates showcase again H100's architectural lead with 94%, while Maia 100 still boasts a solid 88.5%, exceeding A100's 82%. These considerations show that Maia 100 provides an ideal performance- efficiency-cost trade-off, and it becomes a promising solution for such modern AI infrastructure.

These sections together validate that Microsoft's Maia 100 strikes a compelling balance among throughput, latency, efficiency, and cost, thus emerging as a competitive alternative to NVIDIA's A100 and a worthy challenger to a premium H100 in commercial-scale AI workloads.

5. Conclusion

In this paper, we have examined the Microsoft Maia 100 AI accelerator as being a suitable competitor to the NVIDIA A100 & H100 GPUs for a diverse set of deep learning scenarios. Utilizing a methodology that included standardized hardware configurations, real-world benchmark models, and stable software environment, we evaluated the important metrics such as throughput, latency, energy efficiency, training time, utilization, and cost.

Our results show that Maia 100 can provide impressive performance that is consistent or better than the NVIDIA A100 in many of the requirements, and is in proximity to the H100 in a number of metrics. Maia 100 achieves significantly better results over A100 with an average throughput of 4025 samples/sec at moderate trade-offs compared to H100. Maia 100 also outperforms A100 in inference latency, meaning Maia 100 works well in real-time and production-level AI services too.

Regarding the energy efficiency, Maia 100 achieves 9.45 samples/W efficiency, surpassing A100 and H100, proving to be suitable for sustainable AI infrastructure. What's more, thanks to deep integration with Azure and cost-effective cloud-native deployment, Maia 100 can bring down training costs and is well-suited to large-scale AI needs. While H100 still achieves better performance and utilization (94%), Maia 100 has a very high utilization (88.5%) and it is efficient also with mixed precision and quantized inference.

On the whole, the findings indicate that Microsoft's Maia 100 is a feasible and scalable AI workload accelerator for cloud and a rapidly diversifying economy where power savings are critical, tight integration with cloud-native tooling is paramount, and cost-efficiency is key. Although H100 is still the most powerful system in terms of raw capability, Maia 100 breaks the shackles of dependencies by providing organizations in need of robust, high-powered, and cost-effective alternative to NVIDIA dominated AI compute platforms. This report endorses a more general industry diversification in AI infrastructure, and additional x86 billion-dollar investments in custom silicon creations like Maia 100.

References

[1] Y. Lee *et al.*, “Debunking the CUDA Myth Towards GPU-based AI Systems,” *arXiv (Cornell University)*, Dec. 2024, doi: https://doi.org/10.48550/arxiv.2501.00210.

[2] Y. Kundu *et al.*, “A Comparison of the Cerebras Wafer-Scale Integration Technology with Nvidia GPU-based Systems for Artificial Intelligence,” *arXiv.org*, 2025. https://arxiv.org/abs/2503.11698

[3] M. Huang, A. Shen, K. Li, H. Peng, B. Li, and H. Yu, “EdgeLLM: A Highly Efficient CPU-FPGA Heterogeneous Edge Accelerator for Large Language Models,” *arXiv.org*, 2024. https://arxiv.org/abs/2407.21325

[4] H. Peng, C. Ding, T. Geng, S. Choudhury, K. Barker, and A. Li, “Evaluating Emerging AI/ML Accelerators: IPU, RDU, and NVIDIA/AMD GPUs,” *arXiv (Cornell University)*, Nov. 2023, doi: https://doi.org/10.48550/arxiv.2311.04417.

[5] Mirhoseini, A., Goldie, A., Yazgan, M. *et al.* A graph placement methodology for fast chip design. *Nature* **594**, 207–212 (2021). https://doi.org/10.1038/s41586-021-03544-w

[6] Signal65, *Leading AI Scalability Benchmarks with Microsoft Azure*, Signal65 Report, Nov. 2024.

[7] *Microsoft Reveals Custom 128-Core Arm Datacenter CPU and AI Accelerator Maia 100*, *Tom’s Hardware*, Nov. 2023. [Online]. Available: https://www.tomshardware.com/news/microsoft-azure-maia-ai-accelerator-cobalt-cpu-custom

[8] *Inside Maia 100: Revolutionizing AI Workloads with Microsoft’s Custom AI Accelerator*, *Microsoft Azure Infrastructure Blog*, Dec. 2024. [Online]. Available: https://techcommunity.microsoft.com/

[9] *Challengers Are Coming for Nvidia's Crown in AI Acceleration*, *IEEE Spectrum*, May 2024. [Online]. Available: https://spectrum.ieee.org/

[10] *Microsoft Details Maia 100: Custom AI Chip with HBM2e Memory*, *TechSpot*, Jan. 2024. [Online]. Available: https://www.techspot.com/

[11] *How Microsoft's New AI Chip Could Disrupt Big Tech*, *Shrout Research*, Nov. 2023. [Online]. Available: https://www.shroutresearch.com/

[12] D. Keller, *AI Chip Deficit: Alternatives to NVIDIA GPUs*, *EE Times*, May 2024. [Online]. Available: https://www.eetimes.com/

[13] *Huawei Readies New AI Chip for Mass Shipment as China Seeks NVIDIA Alternatives*, *Reuters*, Apr. 2025. [Online]. Available: https://www.reuters.com/

[14] *NVIDIA's Competitors Are Gaining Traction in Sovereign AI and HFT Applications*, *Business Insider*, Jun. 2025. [Online]. Available: https://www.businessinsider.com/

[15] *The Rise of AI Accelerator Alternatives: AMD, Intel, and More*, *Forbes*, Jun. 2025. [Online]. Available: https://www.forbes.com/

Citation: Srikant Sudha Panda. (2025). Breaking Dependency Chains: Evaluating Microsoft's Maia 100 as an Alternative to NVIDIA GPUs in AI Workloads. International Journal of Information Technology (IJIT), 6(1), 94-109.

Abstract Link: https://iaeme.com/Home/article_id/IJIT_06_01_008

Article Link:
https://iaeme.com/MasterAdmin/Journal_uploads/IJIT/VOLUME_6_ISSUE_1/IJIT_06_01_008.pdf

✉ **editor@iaeme.com**