Performance tradeoffs of general-purpose digital hardware and application-specific analog hardware

Performance Tradeoffs of General-Purpose Digital Hardware and Application-Specific Analog Hardware

Carlos Natalino[a], Dan Li[b], Oskars Ozolins[b,c,d], Xiaodan Pang[b,c], and Francesco Da Ros[e]

[a]Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, Sweden.
[b]Department of Applied Physics, KTH Royal Institute of Technology, Isafjordsgatan 22, 164 40 Kista, Sweden.
[c]Industrial Systems Department, RISE Research Institutes of Sweden, 164 40 Kista, Sweden.
[d]Institute of Photonics, Electronics and Telecommunications, Riga Technical University, 1048 Riga, Latvia.
[e]DTU Electro, Technical University of Denmark (DTU), Lyngby, Denmark.

ABSTRACT

The field of artificial intelligence & machine learning (AI/ML) has experienced unprecedented growth over the last decade driven by computationally demanding applications. The computing power has been so far provided by general-purpose digital hardware such as central processing units (CPUs) and graphics processing units (GPUs). As the potential for continuous technological advancements in digital electronics is brought into question, research is focusing on alternative paradigms such as application-specific analog hardware. Both electronics and photonic analog hardware are being actively investigated with promising results showing advantages in terms of processing speed and/or energy efficiency. However, a systematic comparison of these different hardware platforms in terms of high-level computing performance is missing. In this work, we compare these hardware platforms focusing on use cases with different requirements in terms of, e.g., compute capacity, efficiency, and density. The comparison highlights current advantages and key challenges to be addressed in each field.

Keywords: Machine learning, digital electronics, photonic hardware platforms, analog computing.

1. INTRODUCTION

With the introduction of new demanding applications such as natural language processing, image and video recognition, and generative artificial intelligence (AI), the field of artificial intelligence & machine learning (AI/ML) has experienced unprecedented growth over the last decade. This growth happens in two different directions. Firstly, the business prospects for AI/ML companies result in their increasing valuation. Secondly, the complexity of the models also grows as models are capable of more elaborate tasks. For instance, in the pre-deep-learning era, model complexity grew aligned with Moore's law, i.e., double every 2 years. With the introduction of deep learning, this growth rate increased to double every 3-4 months.[1] More recently, the so-called large models have experienced a similar growth rate to deep learning, but with a complexity that is 2 to 3 orders of magnitude higher.[2,3]

The compute capacity requirements have so far been sustained by general-purpose digital hardware such as central processing units (CPUs) and graphics processing units (GPUs). As the potential for continuous technological advancements in digital electronics is brought into question,[4] research is focusing on alternative paradigms such as application-specific analog hardware. Both electronics and photonic hardware platforms are being very actively investigated with promising results showing advantages in terms of processing speed and/or energy efficiency. On the electronic side, impressive demonstrations have been reported using crossbar arrays with

[a]E-mail: carlos.natalino@chalmers.se
[b]E-mail: danl4@kth.se, xiaodan@kth.se
[c]E-mail: oskars.ozolins@ri.se
[e]E-mail: fdro@dtu.dk

memristors, or electronic spiking architectures (TrueNorth, Neurogrid, NorthPole, etc.). On the photonic side, coherent interferometric neural networks, crossbar arrays for in-memory computing, photonic spiking networks, and free-space optical systems have been proposed. However, a systematic comparison of all these different hardware platforms - digital electronics, analog electronics, and analog photonics, in their different flavors - in terms of high-level computing performances, in light of common use cases, is missing. In this work, we compare these hardware platforms focusing on use cases with different requirements in terms of, e.g., compute capacity (and its variation over time), efficiency, and density. The comparison highlights current advantages and key challenges to be addressed in each field.

2. USE CASES AND RELEVANT PERFORMANCE METRICS

AI/ML are currently used in a wide range of use cases, with the increasing potential in several other areas. large language models (LLMs), audio and video processing, surveillance, among others, are currently adopted in real-world environments. It is important, however, to understand that not all these use cases share the same properties, and therefore may require accelerators with different characteristics.

Table 1 enumerates four key properties of use cases that leverage AI/ML models and may impact accelerator requirements. The first property is related to the model complexity. This impacts how much of the accelerator is occupied by the model, e.g., in terms of working memory or compute capacity, and indicates if the model can use a shared accelerator, or requires a dedicated one. For this property, the scale is fairly straightforward, with models ranging from low to high complexity.

Table 1. Properties of use cases with respect to the AI/ML model used.

Model complexity	**Load profile**	**Re-training**	**Deployment**
Low	Constant	Infrequent	Centralized
Medium	Predictable	Occasional	Distributed
High	Hard to predict	Frequent	Federated

The second property is related to the load profile of the use case. This impacts how much usage the accelerator will receive over time, and what would be its underutilization should it serve a single application. This can be measured, for example, by the number of inferences per unit of time, and is a function of the end user of the model. For instance, in industrial automation where the industry works 24/7, the load can be nearly constant, or at least very predictable. Some other applications may have fairly variable loads but with a predictable profile, such as business applications that have a high load during business hours. However, use cases that perform inference depending on end users (e.g., smart city applications) may be at the other end of the spectrum, with very hard-to-predict load.

The third property relates to how frequently the model is re-trained. This is not relevant in cases where the model is dynamically loaded in the accelerator memory, but highly relevant in cases where hardware needs to be reconfigured. Applications such as character recognition have fairly infrequent updates. On the other hand, user-based prediction models, such as traffic prediction, may require frequent re-training to cope with changing user patterns. This aspect directly relates to the training complexity of the accelerator itself, from a hardware perspective.

Finally, the deployment model also impacts accelerator requirements and can be correlated with other use case properties. For instance, a centralized model tends to have a more predictable, if not constant, load profile because all users will request inferences from the same accelerator or group of accelerators. A distributed deployment, on the other hand, may be influenced by hard-to-predict load profiles. Moreover, federated models, i.e., models that are trained in a distributed fashion, may have more frequent re-training procedures than traditional models.

As one can realize, there exists a broad range of properties that each use case may require. However, there is a set of performance metrics that are relevant for the evaluation of AI/ML accelerators used by a wide range of use cases. The most common one is the compute capacity or computing power. Usually, the compute capacity of

general-purpose processing hardware is measured in operations per second (OPS). However, for general-purpose graphics processing units (GP-GPUs) the performance is commonly measured in floating-point operations per second (FLOPS) since AI/ML models usually contain floating-point values. The precision of the floating-point operations (e.g., 16, 32, or 64 bits) is relevant in some cases but has been decreasing in importance as the quantization of models becomes popular. This trend is also increasing the relevance of accelerators based on analog or mixed (hybrid analog and digital) hardware which can generally provide lower bit precision, e.g. for photonic hardware it is generally restricted to up to 7-8 bits[5] considering the current level of technological maturity.

Recently, the cost of electricity has also increased substantially. Combined with the increasing power consumption of AI/ML accelerators, the power efficiency with which the hardware can process and more importantly transfer the information becomes highly relevant. In this case, the power efficiency of AI/ML accelerators is measured in FLOPS per Watt.

With the widespread usage of AI/ML models, these applications are expected to take advantage of edge computing and be deployed close to the end user. This means that the deployment of accelerators will take place not only at large-scale datacenters but also at small edge datacenters. Therefore, the compute density becomes relevant to allow the deployment of high-capacity edge datacenters. In this case, we measure the density in terms of FLOPS per footprint area (e.g., mm^2).

Finally, some secondary metrics are also relevant for certain use cases. For instance, some applications may not use all the compute capacity of the accelerators for the entire time. In such a case, not only the peak power efficiency is important, but also the power efficiency when the accelerator is only partially loaded. Another relevant aspect in this case is the ability for the same accelerator to be used by multiple models. Moreover, some use cases may have a fluctuation in load (e.g., being used only during day or night time). Such cases call for accelerators that can easily be loaded with different models to serve other applications that require compute capacity at the moment.

3. PERFORMANCE AND TRADEOFFS

In this section, we investigate how relevant performance metrics have been evolving over the past 15 years. We use GP-GPUs as the baseline devices due to their popularity during this period, based on an available dataset updated up to 2023.[6] In addition, we assess how recent works on electronic and photonic analog devices stack against the traditional GP-GPUs. We restrict the analysis to reported results on specific devices. For theoretical scaling analysis considering ideal implementations, we refer the reader to existing literature, e.g.[7] for photonic accelerators. In the end, we comment on some tradeoffs that are currently faced should you adopt these new technologies as application-specific support to GP-GPUs.

3.1 Compute capacity

Fig. 1 shows the evolution of compute capacity. We can observe that GP-GPUs have had a mostly exponential growth over the years. Recent announcements by the industry have predicted performance following the exponential growth illustrated in Fig. 1. However, we can see that electronic and photonic analog devices started with subpar performance, but were able to quickly match or even surpass the capacity of GP-GPUs. Whereas analog demonstrations have yet to match the technological maturity of their digital counterparts, and the reported values rely on proof-of-concept demonstrations, this indicates that these new hardware platforms have a promising future for as high-performance accelerators for AI/ML models. In particular, looking at photonic platforms, the inherently higher degrees of parallelism provided by working at optical frequencies (e.g., in terms of frequency-bandwidth, polarization, spatial mode) has shown very promising preliminary results to scale up computing capacity through parallelism with up to 3 dimensions.[8] This aspect can not only provide scaling factors to increase the compute capacity as shown in Fig. 1 but could potentially address the requirement of providing flexible loads, e.g. by controlling the number of wavelength-multiplexed channels used by the accelerator similarly to what is being applied for optical communications, to provide connectivity on-demand.

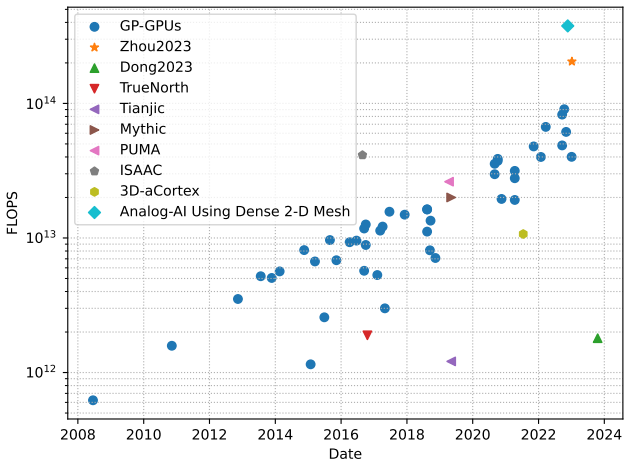

Figure 1. Compute capacity in terms of total floating-point operations per second (FLOPS) for digital (data for general-purpose graphics processing units (GP-GPUs)[6]), analog electronic (data for TrueNorth and Tianjic,[9] Mythic, PUMA, ISAAC, 3D-aCortex and Analog-AI Using Dense 2-D Mesh[3]) and analog photonic (data for Zhou2023,[8] Dong2023[10]) hardware.

3.2 Efficiency

Fig. 2 shows the evolution of compute efficiency over time. We can observe that the growth in efficiency is not as successful as in the case of capacity, with a lower slope of increase over time. This makes it evident that even if the digital electronic devices can scale up their capacity, it becomes increasingly difficult to improve their efficiency.[4] Meanwhile, electronic and photonic analog devices have shown substantial gains in efficiency over GP-GPUs, with some instances being two to three orders of magnitude more efficient. Whereas the estimated efficiencies of analog hardware have not been validated as thoroughly as their digital counterparts, with several results mainly relying on proof-of-concept demonstrations focused on low-scale or partial implementations, or even projected estimates,[11] these results sill indicate that these new devices may become critical for applications where energy efficiency is paramount, e.g., in edge computing applications.

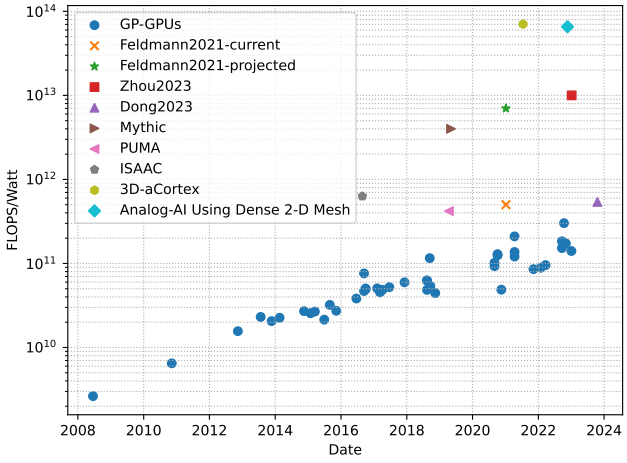

Figure 2. Compute efficiency in terms of floating-point operations per second (FLOPS) per Watt for digital (data for general-purpose graphics processing units (GP-GPUs)[6]), analog electronic (data for Mythic, PUMA, ISAAC, 3D-aCortex and Analog-AI Using Dense 2-D Mesh[3]), and analog photonic (data for Feldmann2021-current and Feldmann2021-projected,[11] Zhou2023,[8] Dong2023[10]) hardware.

3.3 Density

Compute density is another area where electronic and photonic analog devices excel, at least in proof-of-concept demonstrations. Fig. 3 shows that GP-GPUs have not been able to scale exponentially their density, especially due to challenges in heat dissipation from very small dies. In comparison, for electronic and photonic analog devices higher compute densities have been reported. In particular, a potentially strong advantage of analog photonic is the lower heat dissipation of optical interconnects which reduces the challenges for thermal management affecting digital (and analog) electronics, as well as the added benefit of waveguide crossings in the designs, a feature that can enable more compact designs.[5]

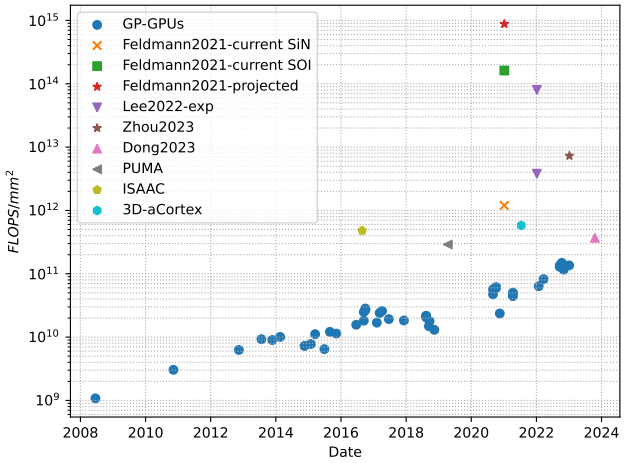

Figure 3. Compute density in terms of floating-point operations per second (FLOPS) per mm^2 for digital (data for general-purpose graphics processing units (GP-GPUs)[6]), analog electronic (data for PUMA, ISAAC, and 3D-aCortex[3]), and analog photonic (data for Feldmann2021-current SiN, Feldmann2021-current SOI and Feldmann2021-projected,[11] Lee2022-exp,[12] Zhou2023,[8] Dong2023[10]) hardware.

3.4 Tradeoffs

As commented in Sec. 2, the load profile varies greatly depending on the use case. For instance, scalability is a critical property for use cases with variable load profiles which can take advantage of the variable performance of GP-GPUs. This is achieved through the deactivation, when not necessary, or blocks of compute units. The result is a compute efficiency that scales almost linearly to the load. At the same time, the performance of the inferences that are still being executed is not impacted, i.e., the inference time is not impacted by the deactivated compute units.

Currently, electronic and photonic analog devices utilize a different approach more similar to the ones adopted in CPUs. In this approach, each compute unit is able to execute a different instruction over an independent piece of data. This process is also known as multiple instructions over multiple data (MIMD). The issue with this approach is that the only way to reduce power consumption is by reducing the clock of the device, which will negatively impact the inference time of the requests being executed. However, photonic devices have the potential to implement something similar to single instruction over multiple data (SIMD) through the use of multiplexing in the physical layer through, for instance, wavelength-division multiplexing (WDM), frequency-division multiplexing (FDM), or space-division multiplexing (SDM). In this case, multiple channels (i.e., multiple data) would traverse the same set of photonic devices concurrently, and instead of changing the clock, some channels could be deactivated.

Secondly, GP-GPUs, due to their digital nature, can easily (i.e. using well-established programming routines) and quickly (i.e., in a matter of few seconds) change context and start processing a different model. This is particularly interesting in use cases where the AI/ML model is distributed, and may serve a limited number of sparse requests. In this case, a single device can be used to serve multiple models, loading them in memory and executing the inference almost instantaneously, depending on the need.

Meanwhile, current analog devices do not offer clear programming procedures. Several promising techniques have been proposed, either relying on offline (in silico) or online (in situ) methods.[13] However, the former category could allow for implementing the AI/ML model onto several hardware devices but it relies heavily on the availability of a physically accurate model of the hardware and generally suffers from a mismatch between a digital model and the analog hardware,[14] even when hardware-aware modeling is considered;[15,16] while the latter category offers potentially higher accuracy but it suffers from poorer scalability in terms of model size and normally requires lengthy sequential calibration procedures, additional hardware monitoring, and re-optimization for every training step.[17] A substantial research effort is dedicated to this challenge which is currently limiting the performance and usability of analog hardware platforms compared to their digital counterparts.

Finally, the precision and correctness of data representations play an important role in several AI/ML models for industries that require precise and reliable data processing. Although requirements for precision have been relaxed through e.g. quantization and addressed in models such as LLMs, as mentioned in Sec. 2, high-precision operations are still relevant in many industries. GP-GPUs offer flexible bit precision for floating-point operations, often ranging from 8 to 64 bits.[18] Moreover, they offer at least error detection, with industry-grade GPUs offering also error correction. In general digital hardware is more prone to allow for signal regeneration than its analog counterparts.

Currently, it is challenging to accurately control the precision of the operations performed in analog photonic hardware, and even analog electronic faces challenges unless mixed processing or redundant coding is considered.[19] Moreover, there are no error detection and correction capabilities inherently built-in into these devices. Therefore, attention needs to be paid to mitigate these issues for these devices to be applicable (if at all) in high-precision use cases.

4. FINAL REMARKS

This work provided a comparison among digital electronics, analog electronics, and analog photonic AI/ML accelerators. To do so, we enumerated key properties of AI/ML use cases commonly found in the industry. Then, we compared the reported performance of these types of devices over the past 15 years in terms of compute capacity, efficiency, and density. Finally, we analyzed the tradeoffs among these different solutions when it comes to three critical aspects of AI/ML models and their use in the real world. The comparison shows that electronic and photonic analog hardware is a promising solution to accelerate the execution of AI/ML models. However, they still require substantial advancements in key areas such as scalability, programmability, and precision.

ACKNOWLEDGMENTS

This work was supported by Vetenskapsrådet, The Swedish Research Council, grant no. 2022-04798, the Villum Foundations (grant no. VIL29334, OPTIC-AI).

REFERENCES

[1] Mehonic, A. and Kenyon, A. J., "Brain-inspired computing needs a master plan," *Nature* **604**, 255–260 (4 2022).

[2] Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., and Villalobos, P., "Compute trends across three eras of machine learning," in [*International Joint Conference on Neural Networks (IJCNN)*], 1–8 (2022).

[3] Aguirre, F., Sebastian, A., Gallo, M. L., Song, W., Wang, T., Yang, J. J., Lu, W., Chang, M.-F., Ielmini, D., Yang, Y., Mehonic, A., Kenyon, A., Villena, M. A., Roldán, J. B., Wu, Y., Hsu, H.-H., Raghavan, N., Suñé, J., Miranda, E., Eltawil, A., Setti, G., Smagulova, K., Salama, K. N., Krestinskaya, O., Yan, X., Ang, K.-W., Jain, S., Li, S., Alharbi, O., Pazos, S., and Lanza, M., "Hardware implementation of memristor-based artificial neural networks," *Nature Communications* **15**, 1974 (3 2024).

[4] Hennessy, J. and Patterson, D., "A new golden age for computer architecture: domain-specific hardware/software co-design, enhanced," *ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)* (2018).

[5] McMahon, P. L., "The physics of optical computing," *Nature Reviews Physics* **5**, 717–734 (10 2023).

[6] Hobbhahn, M., Heim, L., and Aydos, G., "Trends in machine learning hardware," (2023). Accessed: 2024-03-20.

[7] Youngblood, N., "Coherent photonic crossbar arrays for large-scale matrix-matrix multiplication," *IEEE Journal of Selected Topics in Quantum Electronics* **29**(2: Optical Computing), 1–11 (2022).

[8] Zhou, W., Dong, B., Farmakidis, N., Li, X., Youngblood, N., Huang, K., He, Y., David Wright, C., Pernice, W. H. P., and Bhaskaran, H., "In-memory photonic dot-product engine with electrically programmable weight banks," *Nature Communications* **14**, 2887 (may 2023).

[9] Reuther, A., Michaleas, P., Jones, M., Gadepally, V., Samsi, S., and Kepner, J., "AI and ML accelerator survey and trends," in [*IEEE High Performance Extreme Computing Conference (HPEC)*], 1–10 (2022).

[10] Dong, B., Aggarwal, S., Zhou, W., Ali, U. E., Farmakidis, N., Lee, J. S., He, Y., Li, X., Kwong, D.-L., Wright, C. D., Pernice, W. H. P., and Bhaskaran, H., "Higher-dimensional processing using a photonic tensor core with continuous-time data," *Nature Photonics* **17**, 1080–1088 (12 2023).

[11] Feldmann, J., Youngblood, N., Karpov, M., Gehring, H., Li, X., Stappers, M., Le Gallo, M., Fu, X., Lukashchuk, A., Raja, A. S., Liu, J., Wright, C. D., Sebastian, A., Kippenberg, T. J., Pernice, W. H. P., and Bhaskaran, H., "Publisher correction: Parallel convolutional processing using an integrated photonic tensor core," *Nature* **591**, E13–E13 (mar 2021).

[12] Lee, J. S., Farmakidis, N., Wright, C. D., and Bhaskaran, H., "Polarization-selective reconfigurability in hybridized-active-dielectric nanowires," *Science Advances* **8**(24) (2022).

[13] Buckley, S. M., Tait, A. N., McCaughan, A. N., and Shastri, B. J., "Photonic online learning: a perspective," *Nanophotonics* **12**, 833–845 (Jan. 2023).

[14] Modha, D. S., Akopyan, F., Andreopoulos, A., Appuswamy, R., Arthur, J. V., Cassidy, A. S., Datta, P., DeBole, M. V., Esser, S. K., Otero, C. O., et al., "Neural inference at the frontier of energy, space, and time," *Science* **382**(6668), 329–335 (2023).

[15] Moralis-Pegios, M., Mourgias-Alexandris, G., Tsakyridis, A., Giamougiannis, G., Totovic, A., Dabos, G., Passalis, N., Kirtas, M., Rutirawut, T., Gardes, F. Y., Tefas, A., and Pleros, N., "Neuromorphic silicon photonics and hardware-aware deep learning for high-speed inference," *Journal of Lightwave Technology* **40**, 3243–3254 (May 2022).

[16] Cem, A., Yan, S., Ding, Y., Zibar, D., and Da Ros, F., "Data-driven modeling of mach-zehnder interferometer-based optical matrix multipliers," *Journal of Lightwave Technology* **41**, 5425–5436 (Aug. 2023).

[17] Pai, S., Sun, Z., Hughes, T. W., Park, T., Bartlett, B., Williamson, I. A. D., Minkov, M., Milanizadeh, M., Abebe, N., Morichetti, F., Melloni, A., Fan, S., Solgaard, O., and Miller, D. A. B., "Experimentally realized in situ backpropagation for deep learning in photonic neural networks," *Science* **380**, 398–404 (Apr. 2023).

[18] Reuther, A., Michaleas, P., Jones, M., Gadepally, V., Samsi, S., and Kepner, J., "Lincoln ai computing survey (laics) update," in [*IEEE High Performance Extreme Computing Conference (HPEC)*], 1–7 (2023).

[19] Garg, S., Lou, J., Jain, A., Guo, Z., Shastri, B. J., and Nahmias, M., "Dynamic precision analog computing for neural networks," *IEEE Journal of Selected Topics in Quantum Electronics* **29**(2: Optical Computing), 1–12 (2022).