AI and ML Accelerator Survey and Trends

Albert Reuther, Peter Michaleas, Michael Jones, Vijay Gadeppally, Siddharth Samsi, and Jeremy Kepner
MIT Lincoln Laboratory Supercomputing Center
Lexington, MA, USA
{reuther,pmichaleas,michael.jones,vijayg,sid,kepner}@ll.mit.edu

Abstract—This paper updates the survey of AI accelerators and processors from past three years. This paper collects and summarizes the current commercial accelerators that have been publicly announced with peak performance and power consumption numbers. The performance and power values are plotted on a scatter graph, and a number of dimensions and observations from the trends on this plot are again discussed and analyzed. Two new trends plots based on accelerator release dates are included in this year’s paper, along with the additional trends of some neuromorphic, photonic, and memristor-based inference accelerators.

Index Terms—Machine learning, GPU, TPU, dataflow, accelerator, embedded inference, computational performance

I. INTRODUCTION

Just as last year, the pace of new announcements, releases, and deployments of artificial intelligence (AI) and machine learning (ML) accelerators from startups and established technology companies has been modest. This is not unreasonable; for many companies that have released an accelerator report having spent three or four years researching, analyzing, designing, verifying, and validating their accelerator design trade-offs and building the software stack to program the accelerator. For those who have released subsequent versions of their accelerator, they have reported shorter development cycles, though it is still at least two or three years. The focus of these accelerators continues to be on accelerating deep neural network (DNN) models, and the application space spans from very low power embedded voice recognition and image classification to data center scale training, while the competition for defining markets and application areas continues as part of a much larger industrial and technology shift in modern computing to machine learning solutions.

AI ecosystems bring together components from embedded computing (edge computing), traditional high performance computing (HPC), and high performance data analysis (HPDA) that must work together to effectively provide capabilities for use by decision makers, warfighters, and analysts [1]. Figure 1 captures an architectural overview of such end-to-end AI solutions and their components. On the left side of Figure 1, structured and unstructured data sources provide different views of entities and/or phenomenology.

This material is based upon work supported by the Assistant Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Assistant Secretary of Defense for Research and Engineering.
power, both in embedded applications and in data centers.

This paper is an update to IEEE-HPEC papers from the past three years [9]–[11]. As in past years, this paper continues with last year’s focus on accelerators and processors that are geared toward deep neural networks (DNNs) and convolutional neural networks (CNNs) as they are quite computationally intensive [12]. This survey focuses on accelerators and processors for inference for a variety of reasons including that defense and national security AI/ML edge applications rely heavily on inference. And we will consider all of the numerical precision types that an accelerator supports, but for most of them, their best inference performance is in int8 or fp16/bf16 (IEEE 16-bit floating point or Google’s 16-bit brain float).

There are many surveys [13]–[24] and other papers that cover various aspects of AI accelerators. For instance, the first paper in this multi-year survey included the peak performance of FPGAs for certain AI models; however, several of the aforementioned surveys cover FPGAs in depth so they are no longer included in this survey. This multi-year survey effort and this paper focus on gathering a comprehensive list of AI accelerators with their computational capability, power efficiency, and ultimately the computational effectiveness of utilizing accelerators in embedded and data center applications. Along with this focus, this paper mainly compares neural network accelerators that are useful for government and industrial sensor and data processing applications. A few accelerators and processors that were included in previous years’ papers have been left out of this year’s survey. They have been dropped because they have been surpassed by new accelerators from the same company, they are no longer offered, or they are no longer relevant to the topic.

II. SURVEY OF PROCESSORS

Many recent advances in AI can be at least partly credited to advances in computing hardware [6], [7], [25], [26], enabling computationally heavy machine-learning algorithms and in particular DNNs. This survey gathers performance and power information from publicly available materials including research papers, technical trade press, company benchmarks, etc. While there are ways to access information from companies and startups (including those in their silent period), this information is intentionally left out of this survey; such data will be included in this survey when it becomes publicly available. The key metrics of this public data are plotted in Figure 2, which graphs recent processor capabilities (as of July 2022) mapping peak performance vs. power consumption. The dash-dotted box depicts the very dense region that is zoomed in and plotted in Figure 3.

The x-axis indicates peak power, and the y-axis indicate peak giga-operations per second ( GOPs/s ) , both on a logarithmic scale. The computational precision of the processing capability is depicted by the geometric shape used; the computational precision spans from analog and single-bit int1 to four-byte int32 and two-byte fp16 to eight-byte fp64. The precisions that show two types denotes the precision of the multiplication operations on the left and the precision of the accumulate/ addition operations on the right (for example, fp16.32 corresponds to fp16 for multiplication and fp32 for accumulate/ add ). The form factor is depicted by color, which shows the package for which peak power is reported. Blue corresponds to a single chip; orange corresponds to a card; and green corresponds to entire systems (single node desktop and server systems). This survey is limited to single motherboard, single memory-space systems. Finally, the hollow geometric objects are peak performance for inference-only accelerators, while the solid geometric figures are performance for accelerators that are designed to perform both training and inference.

The survey begins with the same scatter plot that we have compiled for the past three years. As we did last year, to save space, we have summarized some of the important metadata of the accelerators, cards, and systems in Table I, including the label used in Figure 2 for each of the points on the graph; many of the points were brought forward from last year’s plot, and some details of those entries are in [9]. There are several additions which we will cover below. In Table I, most of the columns and entries are self explanatory. However, there are two Technology entries that may not be: dataflow and PIM. Dataflow processors are custom-designed processors for neural network inference and training. Since neural network training and inference computations can be entirely deterministically laid out, they are amenable to dataflow processing in which computations, memory accesses, and inter-ALU communications actions are explicitly/statically programmed or “placed-and-routed” onto the computational hardware. Processor in memory (PIM) accelerators integrate processing elements with memory technology. Among such PIM accelerators are those based on an analog computing technology that augments flash memory circuits with in-place analog multiply-add capabilities. Please refer to the references for the Mythic and Gyrfalcon accelerators for more details on this innovative technology.

Finally, a reasonable categorization of accelerators follows their intended application, and the five categories are shown as ellipses on the graph, which roughly correspond to performance and power consumption: Very Low Power for speech processing, very small sensors, etc.; Embedded for cameras, small UAVs and robots, etc.; Autonomous for driver assist services, autonomous driving, and autonomous robots; Data Center Chips and Cards; and Data Center Systems. For most of the accelerators, their descriptions and commentaries have not changed since last year so please refer to last two years’ papers for descriptions and commentaries. There are, however, several new releases that were not covered by past papers that are covered here.

- Acelera, a Dutch embedded system startup, reported the results of an embedded test chip that they have produced [37]. They claim both digital and analog design capabilities, and this test chip was made to test the extent of the digital design capabilities. They expect to add analog (probably flash) design elements in upcoming efforts.
- Maxim Integrated has released a system-on-chip (SoC) for ultra low power applications called the MAX78000 [84]–[86], which includes an ARM CPU core, a RISC-V CPU core and an AI accelerator. The
ARM core is for quick prototyping and code reuse, while the RISC-V core is included to enable optimizing for the lowest power utilization. The AI accelerator has 64 parallel processors and support 1-bit, 2-bit, 4-bit, and 8-bit integer operations. The SoC operates at a maximum of 30mW and is intended for low-latency, battery-powered applications.

- Tachyum came out of startup stealth mode in 2017, and they just recently announced the release of an evaluation board for their Prodigy all-in-one processor [128]. They are promising the functionality of CPUs and GPUs within each core, and it is designed for HPC and machine learning applications. The chip is reported to have “128 high-performance unified cores” running at 5.7 GHz [119].

- NVIDIA announced their next generation GPU called Hopper (H100) in March 2022 [98]. It features even more Symmetric Multiprocessors (SIMD and Tensor cores), 50% higher memory bandwidth, and a 700W power budget for the SXM mezzanine card instance. (PCIe card power budget is 450W).

- Over the past couple of years, NVIDIA has also announced and released several system platforms for automotive, robotic, and other embedded applications that deploy Ampere-generation GPU architecture. Specifically for automotive applications, the DRIVE AGX platform added two new systems: the DRIVE AGX L2 that enables Level 2 autonomous driving within a 45W power envelope and the DRIVE AGX L5 that is intended to enable Level 5 autonomous driving within an 800W power envelope [103]. Similarly, the Jetson AGX Orin and Jetson NX Orin also use an Ampere-generation GPU, and are intended for robotics, factory automation, etc. [100], [101], and they consume a maximum of 60W and 25W peak power.

- Graphcore shared rough peak performance numbers for their second generation accelerator chip, the CG200 [59], [129], [130]. Since it is deployed on a PCIe card, we can assume that the peak power is around 300W. In the past year, Graphcore also announced it’s Bow accelerator, which is the first wafer-on-wafer processor designed in cooperation with TSMC. The accelerator itself is the same CG200 as mentioned above, but it is mated to...
with a second wafer that greatly improves power and clock distribution throughout the CG200 chip [60]. This translates into 40% better performance and 16% better performance-per-Watt.

- Almost a year after Google announced details of their fourth generation inference-only TPUv4 accelerator in June 2021 [54], Google shared details about their fourth generation training accelerator, TPUv4. Very few details were announced, but they did share peak power and performance numbers [55]. As with previous TPU variants, TPU4 is available through the Google Compute Cloud and for internal operations.

Next, we must mention accelerators that do not appear on Figure 2 yet. Each has been released with some benchmark results but either no peak performance numbers or no peak power numbers.

- After last year releasing some impressive benchmark results for their reconfigurable AI accelerator technology [131] and this year publishing two deeper technology reveals [132], [133] and an applications paper with Argonne National Laboratory [134], SambaNova still has not provided any details from which we can estimate peak performance or power consumption of their solutions.

- In May 2022, Intel’s Habana Labs announced the second generation of the Goya inference accelerator and Gaudi training accelerator, named Gaudi2, respectively [135], [136]. Both promised multiple times better performance than their predecessor. Gaudi will be a single-width PCIe card drawing 75W, while the Gaudi2 will continue to be a double-width PCIe card drawing 650W (likely on a PCIe 5.0 slot). Habana released some benchmarking comparisons to Nvidia A100 GPUs for the Gaudi2, but peak performance numbers were not disclosed for either of these accelerators.

- Esperanto has produced a few demo chips for evaluation by Samsung and other partners [137]. The chip is reported to be a 1,000-core RISC-V processor with each core having an AI tensor accelerator. Esperanto has published a few relative performance metrics [138], but they have not disclosed any peak power or peak performance values.

- During the Tesla AI Day event, Tesla gave some details of their custom-built Dojo accelerator and system. They did provide peak performance of 22.6 TF FP32 performance per chip, but they did not report peak power draw per chip. Perhaps these details will come later [139].

Finally, there is one departure to the report this year. Last year, Centaur Technology announced a x86 CPU with an integrated AI accelerator, which was realized as a 4,096 byte-wide SIMD unit. The performance estimates were competitive, but VIA Technologies, the parent company of Centaur, sold off the USA-based engineering team of the processor to Intel, Corp. and seems to have ended the development of the CNS processor [140].

### III. Observations and Trends

There are several observations comments for us to appreciate on Figure 2.

- Int8 continues to be the default numerical precision for embedded, autonomous and data center inference applications. This precision is adequate for most AI/ML applications with a reasonable number of classes. However, some accelerators also use fp16 and/or bfloat16 for inference. For training, has become integer representations.

- Among the very low power chips, what is not captured is the other features beyond the machine learning accelerator on the chip. It is very common in this category and the Embedded category to release system-on-chip (SoC) solutions, which often include low-power CPU cores, audio and video analog-to-digital converters (ADCs),...
encryption engines, network interfaces, etc. These additional features of the SoCs do not change the peak performance metric, but they do have a direct impact on the peak power reported for the chip, so please keep this in mind when comparing them.

- Not much has changed in the Embedded segment, which probably means that the computational performance and peak power is adequate for the types of applications in this area.
- The density has become very crowded in the Autonomous and Data Center Chips and Cards segments, which required the zoomed in Figure 3. Over the past few years, several established embedded computing microelectronics companies including Texas Instruments have released AI accelerators, while NVIDIA has released and announced several more powerful automotive and robotics application systems as mentioned above. Among the Data Center Cards, the PCIe v5 specification is highly anticipated so as to break through the 300W power limit of PCIe v4.
- Finally, the high-end training systems are not only posting very impressive performance numbers, but those companies have also been announcing highly scalable inter-networking technologies to network thousands of cards together. This is particularly important for dataflow accelerators like Cerebras, GraphCore, Groq, Tesla Dojo, and SambaNova, which are explicitly/statically programmed or “placed-and-routed” onto the computational hardware. It enables these accelerators to accommodate extremely large models like transformers [141].

A. Broader Trends

We also collected release dates, fabrication technology, and peak performance for multiple precisions for a smaller subset of accelerators listed in Table I. We were curious about the trends of peak performance over the past ten years and how numerical precision and fabrication technology influenced it. These data are plotted in Figure 4. Figure 4a plots the release date of a number of accelerators versus their peak performance for one or more precision formats. There are marked gains in peak performance for each of the precision formats, but within each format the maximum gain is 1.5 orders of magnitude over the 10-year period. In Figure 4b, we plot the release date versus the fabrication technology used for the accelerator. The default precision for the peak performance values is int8; however, there are a number of accelerators (e.g., NVIDIA K20, K80 and AMD Mi8) which did not have int8 support. For these accelerators, the peak performance is reported for the lowest precision that the accelerator supported. This plot shows that much performance has been gained over the past ten years by supporting lower precision formats; it is particularly interesting to observe how support for lower precision formats was included in these accelerators as research and industry explore the effectiveness of lower floating point and integer formats in CNN/DNN inference and training.

We have several more observations and trends that are not yet captured in graphs. First, the exploration for the best numerical formats for inference and training continue. For inference, some discussion continues whether int4 will be acceptable for embedded inference, and the Maxim MAX 78000 SoC solution supports 1-bit, 2-bit, 4-bit, and 8-bit integer weights [85]. On the training side, it has been announced that NVIDIA Hopper, Intel Gaudi2 and a future GraphCore accelerator will support the lower precision FP8 numerical format [142]. GraphCore posted an analysis paper on FP8 [143], including trade-off analyses of scaled integer versus floating point representations, different 8-bit floating point representations, and mixed representation DNN model performance.

Another trend that has caught our attention is that mathematical kernels other than DNN/CNN models have been implemented on several dataflow accelerators. These dataflow accelerators generally handle each data item independently (i.e., there are no cache lines), and data movement and computational operations are explicitly/statically programmed or
“placed-and-routed” onto the computational hardware (as mentioned previously). Hence, they are amenable to implementing other mathematical kernels for digital signal processing, physical simulation like computational fluid dynamics and weather simulation, and massive graph processing. Cerebras demonstrated the mapping of fast stencil-code onto their wafer-scale processor [144], while researchers from the University of Bristol demonstrated stencil codes and image processing using a GraphCore IPU [145]. A team from Citadel Enterprise America also reported on a series of HPC microbenchmarks that they ran on GraphCore IPUs [146]. Google Research has been very busy demonstrating their TPs on a variety of parallel HPC applications including flood prediction [147], large scale distributed linear algebra [148], molecular dynamics simulation [149], fast Fourier transforms [150], [151]. MRI reconstruction [152], financial Monte Carlo simulations [153], and Monte Carlo simulation of the Ising model [154]. We see this as a foreshadowing of more interesting research and development in using this high performance accelerators.

IV. SUMMARY

This paper updated the survey of deep neural network accelerators that span from extremely low power through embedded and autonomous applications to data center class accelerators for inference and training. We focused on inference accelerators, and discussed some new additions for the year. The rate of announcements and releases has continued to be consistent and modest.

V. DATA AVAILABILITY

The data spreadsheets and references that have been collected for this study and its papers will be posted at https://github.com/areuther/ai-accelerators after they have cleared the release review process.

ACKNOWLEDGEMENT

We express our gratitude to Masahiro Arakawa, Bill Arcand, Bill Bergeron, David Bestor, Bob Bond, Chansup Byun, Nathan Frey, Vitaliy Gleyzer, Jeff Gottschalk, Michael Houle, Matthew Hubbell, Hayden Jananathan, Anna Klein, David Martinez, Joseph McDonald, Lauren Milechin, Sanjeev Mohindra, Paul Monticciolo, Julie Mullen, Andrew Prout, Stephan Rejto, Antonio Rosa, Matthew Weiss, Charles Yee, and Marc Zissman for their support of this work.

REFERENCES

[1] V. Gadepally, J. Goodwin, J. Kepner, A. Reuther, H. Reynolds, S. Samsi, J. Su, and D. Martinez, “AI Enabling Technologies,” MIT Lincoln Laboratory, Lexington, MA, Tech. Rep., may 2019. [Online]. Available: https://arxiv.org/abs/1905.03592
[2] T. N. Theis and H. P. Wong, “The End of Moore’s Law: A New Beginning for Information Technology,” Computing in Science Engineering, vol. 19, no. 2, pp. 41–50, mar 2017. [Online]. Available: https://doi.org/10.1109/MCSE.2017.29
[3] M. Horowitz, “‘Computing’s Energy Problem (and What We Can Do About It),’ in 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC). IEEE, Feb 2014, pp. 10–14. [Online]. Available: http://ieeexplore.ieee.org/document/6757323/
[4] C. E. Leiserson, N. C. Thompson, J. S. Emser, B. C. Kuszmaul, B. W. Lampson, D. Sanchez, and T. B. Schardl, “‘There’s Plenty of Room at the Top: What Will Drive Computer Performance after Moore’s Law?’ Science, vol. 368, no. 6495, jun 2020. [Online]. Available: https://science.sciencemag.org/content/368/6495/eaam9744
[5] N. C. Thompson and S. Spanuth, “The Decline of Computers as a General Purpose Technology,” Communications of the ACM, vol. 64, no. 3, pp. 64–72, mar 2021.
[6] J. L. Hennessy and D. A. Patterson, “A New Golden Age for Computer Architecture,” Communications of the ACM, vol. 62, no. 2, pp. 48–60, Jan 2019. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3310134.3282307
[7] W. J. Dally, Y. Turakhia, and S. Han, “Domain-Specific Hardware Accelerators,” Communications of the ACM, vol. 63, no. 7, pp. 48–57, jun 2020. [Online]. Available: https://dl.acm.org/doi/10.1145/3361682
D. McGrath, “Qualcomm Targets AI Inferencing in the Cloud,”
S. Ward-Foxton, “Qualcomm Cloud AI 100 Promises Impressive...
D. Firu, “Quadric Edge Supercomputer,” Quadric, Tech. Rep., apr 2019.
I. Cutress, “Preferred Networks: A 500 W Custom PCIe-card Using 3000-mm² Silicon,”
D. Firu, “Quadric Edge Supercomputer,” Quadric, Tech. Rep., apr 2019.
D. McGrath, “Qualcomm Targets AI Inferencing in the Cloud,”
Rockchip Released Its First AI Processor RK3399 Pro NPU Performance Up to 2.4TOPS,”
L. Gwennap, “Machine Learning Moves to the Edge,” Microprocessor Report, Tech. Rep., apr 2020.
D. McGrath, “Tech Heavyweights Back AI Chip Startup, oct 21, 2018.
R. Merritt, “Startup Rolls AI Chips for Audio,” feb 2018.
A. Shilov, “Tachyum Teases 128-Core CPU: 5.7 GHz, 950W, 16 DDR5 Channels,” jun 2022.
L. Gwennap, “Tenstorrent Scales AI Performance: Architecture Leads in Data Center Power Efficiency,”
T. P. Morgan, “Intel Pits New Gaudi2 AI Training Engine Against Nvidia GPUs,” may 2022.
D. Schor, “The 2,048-core PEZY-SC2 Sets a Green500 Record,”
R. Prabhakar, S. Jairath, and J. L. Shin, “SambaNova SN10 RDU: A 7nm Dataflow Architecture to Accelerate Software 2.0,” in 2022 IEEE International Solid-State Circuits Conference (ISSCC), vol. 65, 2022, pp. 350–352.
R. Prabhakar and S. Jairath, “SambaNova SN10 RDU: Accelerating Software 2.0 with Dataflow,” in 2021 IEEE Hot Chips 33 Symposium (HCS), aug 2021, pp. 1–37.
M. Emani, V. Vishwanath, C. Adams, M. E. Papka, R. Stevens, L. Florescu, S. Jairath, W. Liu, T. Nama, and J. Suzereth, “Accelerating Scientific Applications With SambaNova Reconfigurable Dataflow Architecture,” Computing in Science & Engineering, vol. 23, no. 2, pp. 114–119, 2021.
O. Peckham, “Intel’s Habana Labs Unveils Gaudi2, Greco AI Processors,” may 2022.
E. Kilgariff, H. Moreton, N. Stam, and B. Bell, “NVIDIA's DGX-2: Sixteen Tesla V100s, 30TB of NVMe-only 400K GB/s and 500 W,”
R. Prabhakar and S. Jairath, “Introducing 2nd Generation IPU Systems for AI at Scale,” jul 2020.
D. Martin, “Samsung, Others Test Esperanto’s 1,000-Core RISC-V Cores,” aug 2021.
S. Ward-Foxton, “TI’s First Automotive SoC with an AI Accelerator Launches,” feb 2021.
E. Kilgariff, H. Moreton, N. Stam, and B. Bell, “NVIDIA’s DGX-2: Sixteen Tesla V100s, 30TB of NVMe-only 400K GB/s and 500 W,”
C. Campa, C. Kawalek, H. Vo, and J. Bessoudo, “Defining AI Innovation with NVIDIA DGX A100,” may 2020.
R. Merritt, “Startup Rolls AI Chips for Audio,” feb 2018. [Online]. Available: https://www.hpcwire.com/2021/08/20/intels-habana-labs-unveils-gaudi2-greco-ai-processors/
Fourier Transform

C. D. Schuman, T. E. Potok, R. M. Patton, J. D. Birdwell, M. E. Dean, K. Yang, Y. F. Chen, G. Roumpos, C. Colby, and J. Anderson, “High Performance Monte Carlo simulation of ising model on TPU clusters,” in International Conference for High Performance Computing, Networking, Storage and Analysis, SC. IEEE Computer Society, nov 2019. [Online]. Available: https://arxiv.org/abs/1903.11714v4

K. Yang, Y. F. Chen, G. Rounplos, C. Colby, and J. Anderson, “High performance Monte Carlo simulation of iseng model on TPU clusters,” in International Conference for High Performance Computing, Networking, Storage and Analysis, SC. IEEE Computer Society, nov 2019. [Online]. Available: https://arxiv.org/abs/1903.11714v4

C. D. Schuman, T. E. Potok, R. M. Patton, J. D. Birdwell, M. E. Dean, G. S. Rose, and J. S. Plank, “A Survey of Neuromorphic Computing and Neural Networks in Hardware,” arXiv preprint arXiv:1705.06963, may 2017. [Online]. Available: http://arxiv.org/abs/1705.06963

C. D. James, J. B. Aimone, N. E. Miner, C. M. Vineyard, F. H. Rothganger, K. D. Carlson, S. A. Mulder, T. J. Draelos, A. Faust, M. J. Marinella, J. H. Naegle, and S. J. Plimpton, “A Historical Survey of Algorithms and Hardware Architectures for Neural-inspired and Neuromorphic Computing Applications,” Biologically Inspired Cognitive Architectures, vol. 19, pp. 49–64, jan 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2212685X16300561

R. F. Service, “Microchips That Mimic the Human Brain Could Make AI Far More Energy Efficient,” may 2022. [Online]. Available: https://www.sciencemag.org/content/article/microchips-mimic-human-brain-could-make-ai-far-more-energy-efficient

G. Orchard, E. P. Frady, D. B. D. Rubin, S. Sanborn, S. B. Shrestha, F. T. Sommer, and M. Davies, “Efficient Neuromorphic Signal Processing with Loihi 2,” in 2021 IEEE Workshop on Signal Processing Systems (SiPS), oct 2021, pp. 254–259.

G. Orchard, W. Wild, G. Orchard, Y. Sandamirskaya, G. A. F. Guerra, P. Joshi, P. Plank, and S. R. Risboud, “Advancing Neuromorphic Computing With Loihi: A Survey of Results and Outlook,” Proceedings of the IEEE, vol. 109, no. 5, pp. 911–934, may 2021.

M. Barnell, C. Raymond, M. Wilson, D. Isereu, and C. Cicotta, “Target Classification in Synthetic Aperture Radar and Optical Imagery Using Loihi Neuromorphic Hardware,” in 2020 IEEE High Performance Extreme Computing Conference (HPEC), 2020, pp. 1–6.

A. Viale, A. Marchisio, M. Martina, G. Masera, and M. Shafique, “CarSNN: An Efficient Spiking Neural Network for Event-Based Autonomous Cars on the Loihi Neuromorphic Research Processor,” in 2021 International Joint Conference on Neural Networks (IJCNN), jul 2021, pp. 1–10.

S. Ward-Foxton, “Innatera Unveils Neuromorphic AI Chip to Accelerate Spiking Networks,” jul 2021. [Online]. Available: https://www.eetimes.com/innatera-unveils-neuromorphic-ai-chip-to-accelerate-spiking-networks/

M. Levy, “Innatera’s Spiking Neural Processor,” apr 2021. [Online]. Available: https://www.linleygroup.com/newsletters/newsletter_detail.php?num=6302&year=2021&tag=3

V. Ostrovskii, P. Fedoseev, Y. Bobrova, and D. Butusov, “Structural and Parametric Identification of Known Memristors,” Nanomaterials, vol. 12, no. 1, jan 2022. [Online]. Available: /pmc/articles/PMC8746671/pmc/articles/PMC8746671/

S. Ward-Foxton, “Optical Compute Promises Game-Changing AI Performance,” aug 2020. [Online]. Available: https://www.eetimes.com/optical-compute-promises-game-changing-ai-performance/?utm_source=eetimes&utm_medium=networksearch

—, “Optical Chip Solves Hardest Math Problems Faster than GPUs,” dec 2021. [Online]. Available: https://www.eetimes.com/optical-computing-chip-runs-hardest-math-problems-100x-faster-than-gpus/

J. Launay, I. Poli, K. Müller, I. Carron, L. Daudet, F. Krzakala, and S. Gigan, “Light-in-the-Loop: Using a Photonics Co-Processor for Scalable Training of Neural Networks,” arXiv preprint, jun 2020. [Online]. Available: https://arxiv.org/abs/2006.01475v2

E. Cottle, F. Michel, J. Wilson, N. New, and I. Kundi, “Optical Convolutional Neural Networks – Combining Silicon Photonics and Fourier Optics for Computer Vision,” arXiv preprint, dec 2020. [Online]. Available: https://arxiv.org/abs/2103.09944v1

D. Schneider, “A Neural-Net Based on Light Could Best Digital Computers,” jun 2019. [Online]. Available: https://spectrum.ieee.org/optical-compute-promises-game-changing-ai-performance/?utm_source=eetimes&utm_medium=networksearch

R. F. Service, “Microchips That Mimic the Human Brain Could Make AI Far More Energy Efficient,” may 2022. [Online]. Available: https://www.sciencemag.org/content/article/microchips-mimic-human-brain-could-make-ai-far-more-energy-efficient

G. Orchard, E. P. Frady, D. B. D. Rubin, S. Sanborn, S. B. Shrestha, F. T. Sommer, and M. Davies, “Efficient Neuromorphic Signal Processing with Loihi 2,” in 2021 IEEE Workshop on Signal Processing Systems (SiPS), oct 2021, pp. 254–259.

G. Orchard, W. Wild, G. Orchard, Y. Sandamirskaya, G. A. F. Guerra, P. Joshi, P. Plank, and S. R. Risboud, “Advancing Neuromorphic Computing With Loihi: A Survey of Results and Outlook,” Proceedings of the IEEE, vol. 109, no. 5, pp. 911–934, may 2021.