Survey of Machine Learning Accelerators

Conference Paper · September 2020

DOI: 10.1109/HPEC43674.2020.9286149

Albert Reuther
Massachusetts Institute of Technology

Survey of Machine Learning Accelerators

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

***Abstract*—New machine learning accelerators are being announced and released each month for a variety of applications from speech recognition, video object detection, assisted driving, and many data center applications. This paper updates the survey of of AI accelerators and processors from last year's IEEE-HPEC paper. This paper collects and summarizes the current accelerators that have been publicly announced with 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 discussed and analyzed. For instance, there are interesting trends in the plot regarding power consumption, numerical precision, and inference versus training. This year, there are many more announced accelerators that are implemented with many more architectures and technologies from vector engines, dataflow engines, neuromorphic designs, flash-based analog memory processing, and photonic-based processing.**

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

I. Introduction

It has become apparent that researching, developing and deploying Artificial Intelligence (AI) and machine learning (ML) solutions has become a promising path to addressing the challenges of evolving events, data deluge, and rapid courses of action faced by many industries, militaries, and other organizations. Advances in the past decade in computations, data sets, and algorithms have driven many advances for machine learning and its application to many different areas.

AI systems bring together a number of components that must work together to effectively provide capabilities for use by decision makers, warfighters, and analysts [1]. Figure 1 captures this system and its components of an end-to-end AI solution. On the left side of Figure 1, structured and unstructured data sources provide different views of entities and/or phenomenology. These raw data products are fed into a data conditioning step in which they are fused, aggregated, structured, accumulated, and converted to information. The information generated by the data conditioning step feeds into a host of supervised and unsupervised algorithms such as neural networks, which extract patterns, predict new events, fill in missing data, or look for similarities across datasets, thereby converting the input information to actionable knowledge. This actionable knowledge is then passed to human beings for decision-making processes in the human-machine teaming phase. The phase of human-machine teaming provides the users with useful and relevant insight turning knowledge into actionable intelligence or insight.

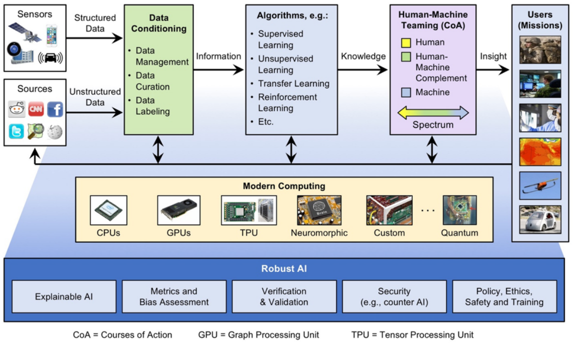

Fig. 1. Canonical AI architecture consists of sensors, data conditioning, algorithms, modern computing, robust AI, human-machine teaming, and users (missions). Each step is critical in developing end-to-end AI applications and systems.

Underpinning this system are modern computing systems, for which Moore's law trends have ended [2], as have a number of related laws and trends including Denard's scaling (power density), clock frequency, core counts, instructions per clock cycle, and instructions per Joule (Koomey's law) [3]. However, advancements and innovations are still progressing in the form of specialized circuits and chips that accelerate often-used operational kernels, methods, or functions. These accelerators are designed with a different balance between performance and functional flexibility. This includes an explosion of innovation in ML processors and accelerators [4], [5]. Understanding the relative benefits of these technologies is of particular importance to applying AI to domains under significant constraints such as size, weight, and power, both in embedded applications and in data centers.

This paper is an update to last year's IEEE-HPEC paper [6]. Quite a number more accelerator chips have been announced and released, and other technologies like neuromorphic architectures, memory-based analog acceleration, and computing with light are gaining attention. There are also some technology categories that were included in last year's paper that will not be included this year, namely most FPGA-based inference instances [7]–[10] and smartphone accelerators. Only FPGA-based offerings that have some dynamic programmability are considered in this paper (e.g., Intel Arria, AImotive, and

This material is based upon work supported by the Assistant Secretary of Defense for Research and Engineering under Air Force Contract No. FA8721-05-C-0002 and/or 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.

Microsoft Brainwave). Smartphone accelerators are not being considered this year because they cannot be used in a different platform without significant re-engineering.

Before getting to the accelerators, we will review a few topics pertinent to understanding the capabilities of the accelerators. We must discuss the types of neural networks for which these ML accelerators are being designed; the distinction between neural network training and inference; the numerical precision with which the neural networks are being used for training and inference, and how neuromorphic accelerators fit into the mix:

- Types of Neural Networks – While AI and machine learning encompass a wide set of statistics-based technologies [1], this paper continues with last year's focus on processors that are geared toward deep neural networks (DNNs) and convolutional neural networks (CNNs). Overall, the most emphasis of computational capability for machine learning is on DNN and CNNs because they are quite computationally intensive [11], and because most of the computations are dense matrix-matrix and matrix-vector multiplies, they are primed to take advantage of computational architectures that exploit data reuse, data locality, and data density.
- Neural Network Training versus Inference – As was explained in last year's survey, neural network training uses libraries of input data to converge model weight parameters by applying the labeled input data (forward projection), measuring the output predictions and then adjusting the model weight parameters to better predict output predictions (back projection). Neural network inference is using a trained model of weight parameters and applying it to input data to receive output predictions. Processors designed for training may also perform well at inference, but the converse is not always true.
- Numerical Precision – The numerical precision with which the model weight parameters and model input data are stored and computed has an impact on the accuracy and efficiency with which networks are trained and used for inference. Generally higher numerical precision representations, particularly floating point representations, are used for training, while lower numerical precision representations, (in particular, integer representations) have been shown to be reasonably effective for inference [12], [13]. However, it is still an open research question whether very limited numerical precisions like int4, int2, and int1 adequately represent model weight parameters and significantly affect model output predictions. Over the past year, more training accelerators have been released that support 16-bit floating point numbers (fp16 and bfloat16), and most inference accelerators now release performance results for 8-bit integer (int8) operands, particularly the inference accelerators geared at edge and embedded processing applications.
- Neuromorphic Computing – The field of neuromophic computing emerged from the neuroscience field, in which researchers and engineers design circuits to model biological and physiological mechanisms in brains. Schuman's survey [14] provides a rich background of all of the significant efforts in the field over the past several decades. Some of the most prominent features of neuromorphic circuits are synthetic neurons and synapses along with spiking synapse signaling, and many agree that the spiking synapse signal is the most prominent feature, especially when implemented in neural network accelerators. In recent accelerators, these spiking signals are usually encoded as digital signals, e.g., IBM TrueNorth chip [15], [16], University of Manchester SpiNNaker [17], and Intel's Loihi [18]. Another recent notable research accelerator is the Tianjic research chip [19], developed by a team at Tsinghua University, which demonstrated the capability of choosing either a spiking neural network (SNN) layer or non-spiking artificial neural network (ANN) layer to instantiate each layer of a DNN model for inference. The team showed that for certain models for audio and video processing a hybrid layer-by-layer approach was most effective, both in accuracy and power consumption. Finally, a startup called Knowm is developing a new neuromorphic computational framework called AHaH Computing (Anti-Hebbian and Hebbian) based on memristor technology [20]. Their goal is to use this technology to dramatically reduce SWaP for machine learning applications.

There are many surveys [12], [21]–[29] and other papers that cover various aspects of AI accelerators; this paper focuses 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, as did last year's paper. Along with this focus, this paper mainly compares neural network accelerators that are useful for government and industrial sensor and data processing applications. Therefore, it will make a distinction between research accelerators and commercially available accelerators, and it will focus comparisons on the latter. Research accelerators are developed by university and industry researchers to demonstrate the art of the possible, often with the intent of influencing commercial attention to technology developments or to attract venture funding to commercialize their developed technology. But before either of these intents are realized, the demonstration of the art of the possible show us opportunities in which commercial products may pursue for improved capabilities and features.

II. Survey of Processors

Many recent advances in AI can be at least partly credited to advances in computing hardware [30], [31], enabling computationally heavy machine-learning algorithms such as neural networks. 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 June 2020) mapping peak performance vs. power consumption. The x-axis indicates peak power, and the y-axis indicate peak giga-operations per second (GOps/s). Note the legend on the right, which indicates various parameters used to differentiate computing techniques and technologies. 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; this is important for showing how much power is consumed, but also how much computation can be packed onto a single chip, a single PCI card, and a full system. Blue corresponds to a single chip; orange corresponds to a card (note that they all are in the 200-300 Watt zone); 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.

We can make some general observations from Figure 2. First, quite a number of new accelerator chips, cards, and systems have been announced and released in the past year. Each of the five categories have a higher density of entries from last year, and there is a greater diversity of underlying architectures and technologies with which these accelerators have been developed as you will notice in the descriptions below. Also, many more recent accelerators have broken through the 1 TeraOps/W boundary for peak performance. An observation that has not changed from last year is that at least 100W must be employed to perform training; almost all of the points on the scatter plot below 100W are inference-only processors/accelerators. (Cornami is the one exception, but their plot point is based on simulation estimates.) It is generally understood that training requires floating point precision which requires more capable and power-consuming ALUs, datapaths, memories, etc.

When it comes to precision, there is an even wider variety of numerical precisions for which peak performance numbers have been released. There continues to be exploration about how much precision is necessary for neural networks to perform well even when using limited or mixed precision representation of activation functions, weights, and biases [13], [32].

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. In the following listings, the angle-bracketed string is the label of the item on the scatter plot, and the square bracket after the angle bracket is the literature reference from which the performance and power values came. A few of the performance values are reported in frames per second (fps) with a given machine learning model. For translating fps values to performance values, Samuel Albanie's Matlab code and a web site of all of the major machine learning models with their operations per epoch/inference, parameter memory, feature memory, and input size [33] were used.

A. Research Chips

Many research papers have been published that have introduced, evaluated, and compared various architectural elements, organizations, and technologies. The following list contains just a fraction of the research chips, but most of these have been highly cited. And, they have performance and power numbers.

- The NeuFlow chip ⟨NeuFlow⟩ [34] was a project between IBM, New York University and Yale University to explore architectures for efficient edge inference. It was designed with dataflow processing tiles comprised of a matrix multiplier and convolver in a 2-dimensional grid.
- The Stanford-designed energy efficient inference engine (EIE) ⟨EIE⟩ [35] demonstrated the use of sparsity in DNN with compressed neural network models, weight sharing, and on-chip SRAM utilization to design an extremely efficent inference chip.
- Another Stanford chip, the TETRIS ⟨TETRIS⟩ [36], demonstrated highly efficient inference by using 3-dimensional memory and a partitioning scheme that placed weights and data close to their processing elements. This allowed the team to use more chip area for computation than local SRAM memory.
- MIT Eyeriss chip ⟨Eyeriss⟩ [12], [37], [38] is a research chip from Vivienne Sze's group in MIT CSAIL. Their goal was to develop the most energy efficient inference chip possible by experimenting with different circuit computation trade-offs. The reported result was acquired running AlexNet with no mention of batch size.
- The DianNao series of dataflow research chips came from a university research team primarily at the Institute of Computing Technology of the Chinese Academy of Sciences (IST-CAS); this team overlaps with the Cambricon company, which has released several chips as well as the Kirin accelerator that is included in the Huawei smartphone system-on-chip (SoC). They published four different designs aimed at different types of ML processing [39]. The DianNao ⟨DianNao⟩ [39] is a neural network inference accelerator, and the DaDianNao ⟨DaDianNao⟩ [40] is a many-tile version of the DianNao for larger NN model inference. The ShiDianNao ⟨ShiDianNao⟩ [41] is designed specifically for convolutional neural network inference. Finally, the PuDianNao ⟨PuDianNao⟩ [42] is designed for seven representative machine learning techniques: k-means, k-NN, naïve Bayes, support vector machines, linear regression, classification tree, and deep neural networks.

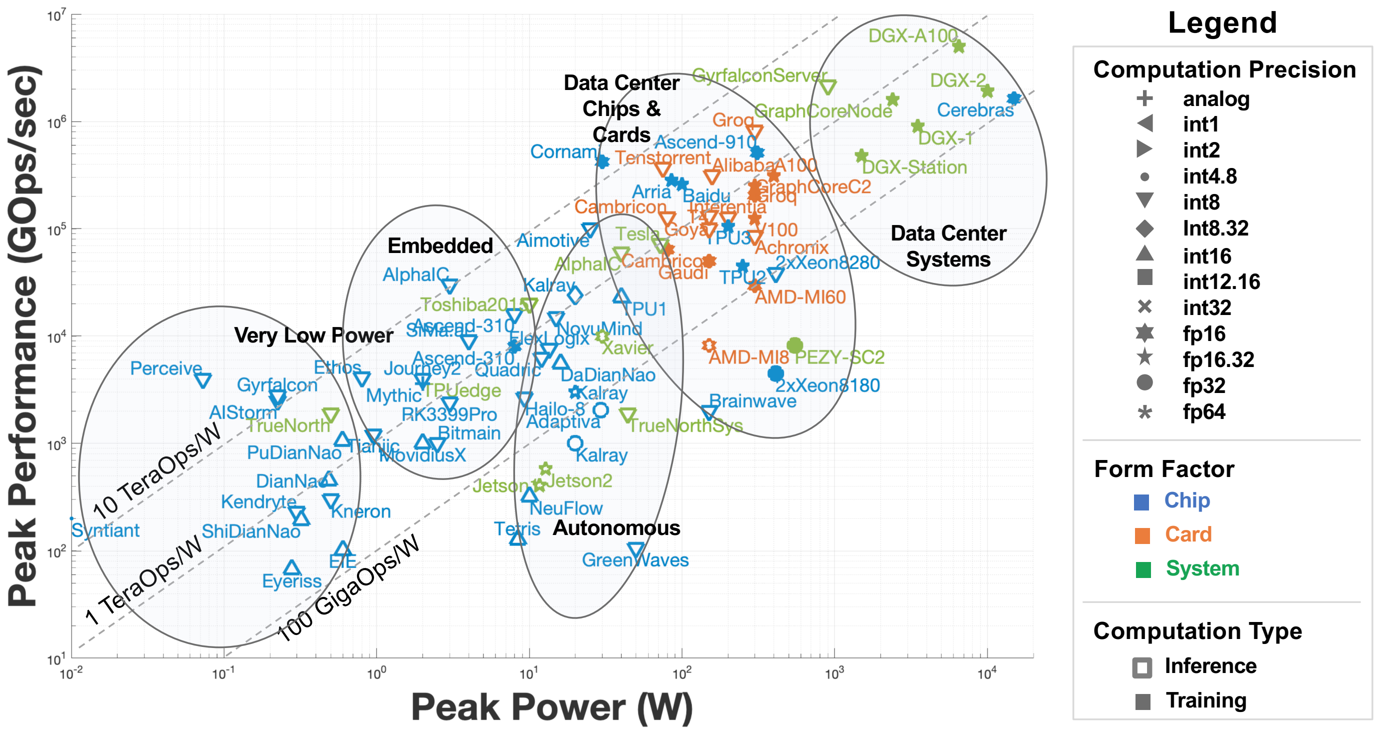

Fig. 2. Peak performance vs. power scatter plot of publicly announced AI accelerators and processors.

- The TrueNorth ⟨TrueNorth⟩ [16], [43], [44] is a digital neuromorphic research chip from the IBM Almaden research lab. It was developed under DARPA funding in the Synapse program to demonstrate the efficacy of digital spiking neural network (neuromorphic) chips. Note that there are points on the graph for both the system, which draws the 44 W power, and the chip, which itself only draws up to 275 mW. The TrueNorth has inspired several companies to develop and release neuromorphic chips.

B. Very Low Power Chips

In the past year, many new chips have been announced or are offering inference-only products in this space.

- The Intel MovidiusX processor ⟨MovidiusX⟩ [45] is an embedded video processor that includes a Neural Engine for video processing and object detection.
- In early 2019, Google released a TPU Edge processor ⟨TPUEdge⟩ [46] for embedded inference application. The TPU Edge uses TensorFlow Lite, which encodes the neural network model with low precision parameters for inference.
- The Rockchip RK3399Pro ⟨Rockchip⟩ [47] is an image and neural co-processor from Chinese company Rockchip. They published peak performance numbers for int8 inference for their ARM CPU and GPU-based co-processor.
- The GreenWaves GAP9 processor ⟨GreenWaves⟩ [48], [49] is designed for battery powered sensors and wearable devices including audio and video processing with respectable performance for such low power consumption. It is capable of computing in 8-, 16-, 24-, and 32-bit integer and 8-, 16-, and 32-bit floating point.
- The Kneron KL520 accelerator ⟨Kneron⟩ [50] is designed for AI smart locks, smart home, people monitoring, and retail applications. The first released systems connect the KL520 with a dual-core ARM M4 processor.
- San Jose startup AIStorm ⟨AIStorm⟩ [51] claims to do some of the math of inference up at the sensor in the analog domain. They originally came to the embedded space scene with biometric sensors and processing, and they call their chip an AI-on-Sensor capability, using technology similar to Mythic and Syntiant, below.
- The Syntiant NDP101 chip ⟨Syntiant⟩ [51], [52] is the neural decision processor inside the Amazon Alexa products for attention phrase recognition. It is a processor-in-memory design, which performs inference weight computations in int4 and activation computation in int8 and consumes less than 200 μW.
- The Mythic Intelligent Processing Unit accelerator ⟨Mythic⟩ [53], [54] combines a RISC-V control processor, router, and flash memory that uses variable resistors (i.e., analog circuitry) to compute matrix multiplies. The accelerators are aiming for embedded, vision, and data center applications.
- The Perceive Ergo processor ⟨Perceive⟩ [55] is meant for battery-powered applications that require high performance. Ergo can run multiple concurrent networks with over 100 million weights in applications such as simultaneous video and audio processing.
- The XMOS Xcore.ai microcontroller ⟨Xcore-ai⟩ [56] is a very low power processor that supports int1, int8, int16,

and int32 computation. It is initially targeted for voice applications such as keyword detection and dictionary functions.

C. Embedded Chips and Systems

The systems in this category are aimed at embedded applications that require high performance inference relative to power and form factor including embedded camera processors, small UAVs, modest robots, etc.

- ARM has released its Ethos line of machine learning processors that mate one or more of its big.LITTLE ARM cores with MAC Compute Engines (MCEs). The Ethos N77 ⟨Ethos⟩ [57] has four MCEs, each of which are capable of computing 1 TOP/s at 1.0 GHz.
- Bitmain is a Chinese company that has specialized in cryptocurrency mining, but they are broadening their product lineup with the Bitmain BM1880 embedded AI processor ⟨Bitmain⟩ [58]. The processor features two ARM Cortex CPUs, a RISC-V CPU controller, and a tensor processing unit, which is aimed at video surveillance applications.
- The NovuMind NovuTensor chip ⟨NovuMind⟩ [59], [60] is a co-processor for vision inference applications.
- The Lightspeeur 5801 video processing chip from Gyrfalcon ⟨Gyrfalcon⟩ [61] uses a matrix processing engine with processor-in-memory techniques to compute model inference. A server is available (see below) that incorporates 128 Lightspeeur 5801 chips as accelerators.

D. Autonomous Systems

The entries in this category are aimed at inference processing for automotive AI/ML, autonomous vehicles, UAV/RPAs, robots, etc.

- The AImotive aiWare3 ⟨AImotive⟩ [62] is a programmable FPGA-based accelerator aimed at the autonomous driving industry.
- The AlphaIC Real AI Processor-Edge (RAP-E) ⟨AlphaIC⟩ [63] touts agent-based neural network computations, where agents are associated with each computational kernel that is executed.
- The Israeli startup Hailo released some performance details about its Hailo-8 edge inference chip in 2019 ⟨Hailo-8⟩ [64]. They published ResNet-50 inference numbers but no peak performance number or architectural details from which peak numbers could be derived so the peak performance number on the graph comes from the ResNet-50 performance.
- Horizon Robotics has started producing its Journey2 accelerator ⟨Journey2⟩ [65] for autonomous driving vision applications.
- The Huawei HiSilicon Ascend 310 ⟨Ascend310⟩ [66] is an integrated CPU with AI accelerator based on the same Huawei Da Vinci architecture as the data center focused Ascend 910 (see below). While the Ascend 310 reports respectable performance for both int8 and fp16, it is intended for inference applications.
- The Kalray Coolidge ⟨Kalray⟩ [67], [68] parallel processor technology originally was developed for high performance computing for French nuclear simulations. It became apparent to Kalray that the same architecture could be used for AI inference tasks for automotive applications, which has led to a partnership with French chip manufacturer, NXP. The Coolidge chip has 80 64-bit VLIW procesor cores that each have an AI coprocessor.
- The NVIDIA Jetson-TX1 ⟨JetsonTX1⟩ [69] incorporates four ARM cores and 256 CUDA Maxwell cores, and the Jetson-TX2 ⟨JetsonTX2⟩ [69] mates six ARM cores with 256 CUDA Pascal cores. To gain more performance, the NVIDIA Xavier ⟨Xavier⟩ [70] deploys eight ARM cores with 512 CUDA Volta cores and 64 Tensor cores.
- The Quadric q1-64 accelerator chip ⟨Quadric⟩ [71] has a 4,096 fused multiply-add (FMA) array for inference processing. One Quadric card will contain four Quadric q1-64 chips, and they have plans for producing a q1-128 chip with a 16k FMA array for data center applications.
- The SiMa.ai Machine Learning Accelerator (MLA) ⟨Quadric⟩ [71] combines an ARM CPU with a dataflow matrix multiply engine to achieve high computation rates with only 4 Watts.
- The Tesla Full Self-Driving (FSD) Computer chip ⟨Tesla⟩ [72], [73] has two neural processing units (NPUs); each NPU has a 96x96 MAC array with 8-bit multiply and 32-bit add units. The FSD also has a set of 12 ARM Cortex-A72 CPUs and a GPU, but the bulk of the computational capability is delivered by the NPUs.
- The Toshiba 2015 image processor ⟨Toshiba2015⟩ [74] combines two 4-core ARM Cortex-A53 processors, four DSPs, a DNN accelerator, and several application specific accelerators, and it is designed for the autonomous driving market.

E. Data Center Chips and Cards

There are a variety of technologies in this category including several CPUs, a number of GPUs, programmable FPGA solutions, and dataflow accelerators. They are addressed in their own subsections to group similar processing technologies.

1) CPU-based Processors:

- The Intel second-generation Xeon Scalable processors ⟨2xXeon8180⟩ and ⟨2xXeon8280⟩ [75] are conventional high performance Xeon server processors. Intel has been marketing these chips to data analytics companies as very versatile inference engines with reasonable power budgets. The peak performance is computed as the peak of two Intel Xeon Platinum CPUs (8180 for inference and 8280 for training) with the key math units being the dual AVX-512 units per core.
- The Japanese Pezy SC2 massively multicore processor ⟨Pezy⟩ [76] is a 2,048-core chip with 8 threads per core. The processor is designed for scientific HPC installations, but with high parallelism for dense linear algebra, it is also very capable for AI training and inference.
- The Tenstorrent Grayskull accelerator ⟨Tenstorrent⟩ [77] has a 10x12 array of Tensix cores, each of which is

comprised of five RISC cores. These RISC cores can compute in int8, fp16, and bfloat16, and the floating point formats operate at a quarter of the performance of int8.

- Preferred Networks is another Japanese company, and it works with the Tokyo University on chip design projects. The PFN-MN-3 PCIe accelerator card ⟨PFN-MN-3⟩ [77] has four chips, each of which have four die and 32 GB of RAM. Each die has 2048 processing elements and 512 matrix arithmetic blocks, all of which compute at fp16 precision. Indications are that these cards are accelerators for one of the Japanese exascale supercomputers, but they also happen to have the arithmetic units for both training and inference.

2) FPGA-based Accelerators:

- The Intel Arria solution pairs an Intel Xeon CPU with an Altera Arria FPGA ⟨Arria⟩ [78], [79]. The CPU is used to rapidly download FPGA hardware configurations to the Arria, and then farms out the operations to the Arria for processing certain key kernels. Since inference models do not change, this technique is well geared toward this CPU-FPGA processing paradigm. However, it would be more challenging to farm ML model training out to the FPGAs. Since it is an FPGA, the peak performance is equal to the performance the DNN model, the performance peak is reported for using GoogLeNet which ran at 900 fps.
- The Bittware/Achronix VectorPath S7t-VG6 accelerator ⟨Achronix⟩ [80] is a FPGA-based processor on a PCI Express card. The FPGA includes eight banks of GDDR6 memory, a 2000GbE and a 4000GbE network interfaces, and 40,000 int8 multiply-accumulate units.
- Cornami has been developing a reconfigurable AI chip based on FPGA technology. Their FPGA-based prototype posted impressive performance in fp16 precision [81], but their ASIC has not yet taped out.
- The Flex Logix InferX X1 eFPGA/DSP accelerator card ⟨FlexLogix⟩ [82] targets both signal processing and machine learning markets and supports int8, int16, Bfloat16 and fp16 precisions. It has 4,000 multiply-accumulate units and is programmable by TensorFlow Lite and ONNX.
- The Microsoft Brainwave project ⟨Brainwave⟩ [83] is a programmable Intel Stratix 10 280 FPGA that was deployed as part of the Catapult project [84]. It is intended for re-programmable inference processing.

3) GPU-based Accelerators: There are three NVIDIA cards and two AMD/ATI cards on the chart (listed respectively): the Volta architecture V100 ⟨V100⟩ [85], [86], the Turing T4 ⟨T4⟩ [87], the Ampere architecture A100 ⟨A100⟩ [88], the MI8 ⟨MI8⟩ [89], and MI60 ⟨MI60⟩ [90]. The V100, A100, MI8, and MI60 GPUs are pure computation cards intended for both inference and training, while the T4 Turing GPU is geared primarily to inference processing, though it can also be used for training.

4) Dataflow Chips and Cards: 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 "placed-and-routed" onto the computational hardware.

- Alibaba Hanguang 800 accelerator ⟨Alibaba⟩ [91] posted the highest inference rate for a chip when it was announced in September 2019. Although ResNet-50 inference benchmarks were released, no peak performance capability was reported nor were any architecture details from which peak performance could be calculated so the peak performance number is their reported ResNet-50 result.
- Amazon Web Services has released a few details of its Inferentia chip ⟨Inferentia⟩ [92], [93]. A peak performance has been published for int8 inference, and it can also compute at fp16 and bfloat16 for compatibility with common DNN models. As an accelerator, it is likely deployed on a card or daughterboard so power consumption is likely in the 150-300 W range.
- Baidu has announced an AI accelerator chip called Kunlun ⟨Baidu⟩ [94], [95]. There are two variants of the Kunlun: the 818-100 for inference and the 818-300 for training. These chips are aimed at low power data center training and inference and are deployed in Baidu's cloud service.
- The Cambricon dataflow chip ⟨Cambricon⟩ [96] was designed by a team at the Institute of Computing Technology of the Chinese Academy of Sciences (IST-CAS) along with the Cambricon company, which came out of the university team. They published both int8 inference and float16 training numbers that are both significant, so both are on the chart.
- Google has released three versions of their Tensor Processing Unit (TPU) [31]. The TPU1 ⟨TPU1⟩ [97] is only for inference, but Google soon made improvements that enabled both training and inference on the TPU2 ⟨TPU2⟩ [97] and TPU3 ⟨TPU3⟩ [97].
- GraphCore.ai has released their C2 card ⟨GraphCoreC2⟩ [98], [99] in early 2019, which is being shipped in their GraphCore server node (see below). This company is a startup headquartered in Bristol, UK with an office in Palo Alto. They have strong venture backing from Dell, Samsung, and others. The performance values for inference benchmarking were achieved with ResNet-50 training for the single C2 card with a batch size for training of 8.
- The Groq Tensor Streaming Processor (TSP) ⟨Groq⟩ [99], [100] is a single processor comprised of over 400,000 multiply-accumulate units along with memory and vector engines. They are organized into sets of Superlanes with each Superlane executing a very long instruction word (VLIW) instruction. Data shuttles between memory units and arithmetic units both within a Superlane and neighboring Superlanes to complete a static program of execution. Each processing unit operates on 8-bit words, and multiple adjacent processing units can be ganged together to execute floating point and multi-byte integer

operations. The Groq TSP currently holds the record for most images per second for ResNet-50 processing.

- The Huawei HiSilicon Ascend 310 ⟨Ascend310⟩ [66] is an integrated CPU with AI accelerator, which is aimed at autonomous systems. It is based on the same Huawei Da Vinci architecture as the data center focused Ascend 910. The Ascend 310 reports respectable performance for both int8 and fp16 and is intended for inference applications. The Ascend 910 ⟨Ascend910⟩ [101] consumes over 300W of power and is the core of the Huawei AI Cluster product line.
- Intel purchased the Israeli startup Habana Labs in 2018, and Habana has released two chips. The Goya chip ⟨Goya⟩ [102], [103] is an inference chip, and its peak performance is an estimate based on the size of the MAC array and 1.5 GHz clock speed typical of 7-nm processors. Habana Labs has also released a training processor called Gaudi ⟨Gaudi⟩ [103], [104]. Intel Habana is planning to release server appliances that include eight Goyas or eight Gaudis.
- The Cerebras CS-1 Wafer Scale Engine (WSE) ⟨Cerebras⟩ [51], [105] is the first wafer-scale processor; it has over 400,000 cores across 84 chips on the same wafer and draws a maximum power of 20kW. Each core has 45 Kbytes of local memory, and the cores are interconnected with a mesh network. While Cerebras has not released clock speeds, numerical precision, or computational performance numbers, there has been discussion that the WSE peak performance exceeds one petaflop. Hence, the graph shows an estimate based on an estimated clock speed of 2GHz for fp16 precision, and that all of the 400,000 cores can execute a fused-multiply-add (FMA) every clock cycle.

F. Data Center Systems

This section lists a number of single-node data center systems.

- There are four NVIDIA server systems on the graph: the DGX-Station, the DGX-1, the DGX-2, and the DGX-A100: The DGX-Station is a tower workstation ⟨DGX-Station⟩ [106] for use as a desktop system that includes four V100 GPUs. The DGX-1 ⟨DGX-1⟩ [106], [107] is a server that includes eight V100 GPUs that occupies three rack units, while the DGX-2 ⟨DGX-2⟩ [107] is a server that includes sixteen V100 GPUs that occupies ten rack units. Finally, the recently announced DGX-A100 ⟨DGX-A100⟩ [108] contains eight A100 GPUs that are networked together with a third generation NV-Link network.
- GraphCore.ai started shipping a Dell/EMC based DSS8440 IPU-Server ⟨GraphCoreNode⟩ [109] in 2019, which contains eight C2 cards (see above). The server power is an estimate based on the components of a typical Intel based, dual-socket server with 8 PCI cards. Training results are on ResNext-50 [110].
- The SolidRun Janux GS31 incorporates 128 Gyrfalcon Lightspeeur 5801 (see above) video processing chips ⟨GyrfalconSaystem⟩ [111], which uses a matrix processing engine with processor-in-memory techniques to compute model inference with a maximum power draw of 900W.

III. Announced Accelerators

In this section, let us pivot towards the future. A number of other accelerator chips have been announced but have not published any performance and power numbers. Below are several companies from whom interesting announcements are expected in the next year or so:

- Qualcomm has announced their Cloud AI 100 accelerator [112], and with their experience in developing communications and smartphone technologies, they have the potential for releasing a chip that delivers high performance with low power draws.
- SambaNova has been hinting at their reconfigurable AI accelerator technology from within stealth mode for a few years, but they have not provided any details from which we can estimate performance or power consumption of their solutions.
- Blaize has emerged from stealth mode and announced its Graph Streaming Processor (GSP) [113], but they have not provided any details beyond a high level component diagram of their chip.
- Enflame has announced it's CloudBlazer T10 data center training accelerator [114], which will support a broad range of datatypes.
- Esperanto is building its Maxion CPU AI processor [115] on the RISC-V open instruction set standard, but they have not released enough information on performance or power consumption to make an assessment.

While there are a few neuromorphic and neuromorphic-inspired chips in the above lists, there have also been a number of announcements from companies that intend to release such accelerators. Here is a list of them.

- Brainchip Akida spiking neural network processor features 1024 neurons per chip, similar to the IBM TrueNorth research chip, that runs on less than one Watt of power [116]. It is expected to use an 11-layer SNN.
- Eta Computing TENSAI Chip was demonstrated using spiking neural networks, but Eta Computing has since released a more conventional inference chip because they determined that spiking neural network chips were not ready for commercial release [117]. It is not clear whether the TENSAI chip will be released commercially.
- aiCTX (pronounced AI cortex) is developing a low-power, low-latency neuromorphic accelerator [118] that executes single sample inference in under a milliwatt and under 10 ms.
- Anaflash chip is an eflash-based spiking neuromorphic chip [119] that encodes 320 neurons with 68 interconnecting synapses per layer. The Univ. of Minnesota research paper that explains this technology is [120], and it provides results on the MNIST written handwriting dataset.

- The Grai Matter Labs chip called NeuronFlow has 1024 neurons per chip that can operate with 8-bit or 16-bit integer precision [121], [122].
- The startup Koniku is modeling circuitry after the brain by designing a co-processor built with biological neurons [123]. The core of the device is a structured micro electrode array system (SMEAS) that they call a Konikore. They have demonstrated that keeping neurons alive is a solvable engineering control problem: living neurons operate in a defined parameter space, and they are developing hardware and algorithms which control the the parameter space of the environment.
- Finally, the Intel Loihi chip [124] is a spiking neural network processor that has been scaled up to 768 chips to simulate 100 million spiking neurons.

Several companies have made technology announcements to use optical processing and silicon photonic networks for AI processing [125]. These companies include Intel [126], Light Matter [127], Lighton [128], Lightelligence [129], [130], Optalysys [131], Fathom Computing [132], and Luminous [133].

As performance and power numbers become available for all of these and other chips, they will be added in future iterations of this survey.

And, as one would expect in such a tumultuous innovation environment, there have also been companies that have cancelled their AI accelerator programs and even a few that have already declared bankruptcy. The most prominent of these was Intel's halt in development of the Nervana NNP-T training chips (codenamed Spring Crest) and curtailing development of the Nervana NNP-I inference (codenamed Spring Hill) in January 2020 [134]. A little further in the past, KnuEdge, the company that announced the KnuPath chip [135] shut its doors in 2018, while Wave Computing, which also owns the MIPS processor technology, folded in early 2020. Finally, TeraDeep, a startup from some Purdue University professors and researchers closed its doors somewhere around 2016 before much of the AI chip race started.

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. Many more accelerators were announced and released, and many of these releases included peak performance and power consumption data. There has been a drive to release 8-bit integer performance for edge/embedded accelerators, while data center accelerators for training often presented 16-bit float performance data. Finally, the diversity of architecture and technologies including neuromorphic, flash-based analog memory processing, dataflow engines, and photonic-based processing is making the competition and performance opportunities very exciting.

Acknowledgement

We are thankful to Masahiro Arakawa, Bill Arcand, Bill Bergeron, David Bestor, Bob Bond, Chansup Byun, Vitaliy Gleyzer, Jeff Gottschalk, Michael Houle, Matthew Hubbell, Anna Klein, David Martinez, Lauren Milechin, Sanjeev Mohindra, Paul Monticciolo, Julie Mullen, Andrew Prout, Stephan Rejto, Antonio Rosa, 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] 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

[5] 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

[6] A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi, and J. Kepner, "Survey and Benchmarking of Machine Learning Accelerators," in *2019 IEEE High Performance Extreme Computing Conference, HPEC 2019*. Institute of Electrical and Electronics Engineers Inc., sep 2019. [Online]. Available: https://doi.org/10.1109/HPEC.2019.8916327

[7] Z. Li, Y. Wang, T. Zhi, and T. Chen, "A survey of neural network accelerators," *Frontiers of Computer Science*, vol. 11, no. 5, pp. 746–761, oct 2017. [Online]. Available: http://link.springer.com/10.1007/s11704-016-6159-1

[8] S. Mittal, "A survey of FPGA-based accelerators for convolutional neural networks," *Neural Computing and Applications*, pp. 1–31, oct 2018. [Online]. Available: http://link.springer.com/10.1007/s00521-018-3761-1

[9] A. G. Blaiech, K. Ben Khalifa, C. Valderrama, M. A. Fernandes, and M. H. Bedoui, "A Survey and Taxonomy of FPGA-based Deep Learning Accelerators," *Journal of Systems Architecture*, vol. 98, pp. 331–345, sep 2019. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1383762118304156

[10] K. Guo, S. Zeng, J. Yu, Y. Wang, and H. Yang, "[DL] A Survey of FPGA-based Neural Network Inference Accelerators," *ACM Transactions on Reconfigurable Technology and Systems*, vol. 12, no. 1, pp. 1–26, apr 2019. [Online]. Available: https://dl.acm.org/doi/10.1145/3289185

[11] A. Canziani, A. Paszke, and E. Culurciello, "An Analysis of Deep Neural Network Models for Practical Applications," *arXiv preprint arXiv:1605.07678*, 2016. [Online]. Available: http://arxiv.org/abs/1605.07678

[12] V. Sze, Y. Chen, T. Yang, and J. S. Emer, "Efficient Processing of Deep Neural Networks: A Tutorial and Survey," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2295–2329, dec 2017. [Online]. Available: https://doi.org/10.1109/JPROC.2017.2761740

[13] S. Narang, G. Diamos, E. Elsen, P. Micikevicius, J. Alben, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, and Others, "Mixed precision training," *Proc. of ICLR,(Vancouver Canada)*, 2018.

[14] 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

[15] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S. K. Esser, R. Appuswamy, B. Taba, A. Amir, M. D. Flickner, W. P. Risk, R. Manohar, and D. S. Modha, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668–673, aug 2014. [Online]. Available: https://science.sciencemag.org/content/345/6197/668

[16] S. K. Esser, P. A. Merolla, J. V. Arthur, A. S. Cassidy, R. Appuswamy, A. Andreopoulos, D. J. Berg, J. L. McKinstry, T. Melano, D. R. Barch, C. di Nolfo, P. Datta, A. Amir, B. Taba, M. D. Flickner, and D. S. Modha, "Convolutional networks for fast, energy-efficient neuromorphic computing." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, no. 41, pp. 11 441–11 446, oct 2016. [Online]. Available: https://doi.org/10.1073/pnas.1604850113

[17] M. Khan, D. Lester, L. Plana, A. Rast, X. Jin, E. Painkras, and S. Furber, "SpiNNaker: Mapping neural networks onto a massively-parallel chip multiprocessor," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. IEEE, jun 2008, pp. 2849–2856. [Online]. Available: http://ieeexplore.ieee.org/document/4634199/

[18] C. Lin, A. Wild, G. N. Chinya, Y. Cao, M. Davies, D. M. Lavery, and H. Wang, "Programming Spiking Neural Networks on Intel's Loihi," *Computer*, vol. 51, no. 3, pp. 52–61, mar 2018. [Online]. Available: https://doi.org/10.1109/MC.2018.157113521

[19] J. Pei, L. Deng, S. Song, M. Zhao, Y. Zhang, S. Wu, G. Wang, Z. Zou, Z. Wu, W. He, F. Chen, N. Deng, S. Wu, Y. Wang, Y. Wu, Z. Yang, C. Ma, G. Li, W. Han, H. Li, H. Wu, R. Zhao, Y. Xie, and L. Shi, "Towards artificial general intelligence with hybrid Tianjic chip architecture," *Nature*, vol. 572, no. 7767, pp. 106–111, aug 2019. [Online]. Available: http://www.nature.com/articles/s41586-019-1424-8

[20] M. A. Nugent and T. W. Molter, "AHaH computing-from metastable switches to attractors to machine learning," *PLoS ONE*, vol. 9, no. 2, p. e85175, feb 2014.

[21] C. S. Lindsey and T. Lindblad, "Survey of Neural Network Hardware," in *SPIE 2492, Applications and Science of Artificial Neural Networks*, S. K. Rogers and D. W. Ruck, Eds., vol. 2492. International Society for Optics and Photonics, apr 1995, pp. 1194–1205. [Online]. Available: http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=1001095

[22] Y. Liao, "Neural Networks in Hardware: A Survey," Department of Computer Science, University of California, Tech. Rep., 2001. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.460.3235

[23] J. Misra and I. Saha, "Artificial neural networks in hardware: A survey of two decades of progress," *Neurocomputing*, vol. 74, no. 1-3, pp. 239–255, dec 2010. [Online]. Available: https://doi.org/10.1016/j.neucom.2010.03.021

[24] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, *Efficient Processing of Deep Neural Networks*. Morgan & Claypool Publishers, 2020. [Online]. Available: https://doi.org/10.2200/S01004ED1V01Y202004CAC050

[25] H. F. Langroudi, T. Pandit, M. Indovina, and D. Kudithipudi, "Digital neuromorphic chips for deep learning inference: a comprehensive study," in *Applications of Machine Learning*, M. E. Zelinski, T. M. Taha, J. Howe, A. A. Awwal, and K. M. Iftekharuddin, Eds. SPIE, sep 2019, p. 9. [Online]. Available: https://doi.org/10.1117/12.2529407

[26] Y. Chen, Y. Xie, L. Song, F. Chen, and T. Tang, "A Survey of Accelerator Architectures for Deep Neural Networks," *Engineering*, vol. 6, no. 3, pp. 264–274, mar 2020. [Online]. Available: https://doi.org/10.1016/j.eng.2020.01.007

[27] E. Wang, J. J. Davis, R. Zhao, H.-C. C. Ng, X. Niu, W. Luk, P. Y. K. Cheung, and G. A. Constantinides, "Deep Neural Network Approximation for Custom Hardware," *ACM Computing Surveys*, vol. 52, no. 2, pp. 1–39, may 2019. [Online]. Available: https://dl.acm.org/doi/10.1145/3309551

[28] S. Khan and A. Mann, "AI Chips: What They Are and Why They Matter," Georgetown Center for Security and Emerging Technology, Tech. Rep., apr 2020. [Online]. Available: https://cset.georgetown.edu/research/ai-chips-what-they-are-and-why-they-matter/

[29] U. Rueckert, "Digital Neural Network Accelerators," in *NANO-CHIPS 2030: On-Chip AI for an Efficient Data-Driven World*, B. Murmann and B. Hoefflinger, Eds. Springer, Cham, 2020, ch. 12, pp. 181–202. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-18338-7_12

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Neural Information Processing Systems*, vol. 25, 2012.

[31] N. P. Jouppi, C. Young, N. Patil, and D. Patterson, "A Domain-Specific Architecture for Deep Neural Networks," *Communications of the ACM*, vol. 61, no. 9, pp. 50–59, aug 2018. [Online]. Available: http://doi.acm.org/10.1145/3154484

[32] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan, "Deep Learning with Limited Numerical Precision," in *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ser. ICML'15. JMLR.org, 2015, pp. 1737–1746. [Online]. Available: http://dl.acm.org/citation.cfm?id=3045118.3045303

[33] S. Albanie, "Convnet Burden," 2019. [Online]. Available: https://github.com/albanie/convnet-burden

[34] C. Farabet, B. Martini, B. Corda, P. Akselrod, E. Culurciello, and Y. Lecun, "NeuFlow: A runtime reconfigurable dataflow processor for vision," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2011. [Online]. Available: https://doi.org/10.1109/CVPRW.2011.5981829

[35] S. Han, X. Liu, H. Mao, J. Pu, A. Pedram, M. A. Horowitz, and W. J. Dally, "EIE: Efficient Inference Engine on Compressed Deep Neural Network," in *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)*. IEEE, jun 2016, pp. 243–254. [Online]. Available: http://ieeexplore.ieee.org/document/7551397/

[36] M. Gao, J. Pu, X. Yang, M. Horowitz, and C. Kozyrakis, "TETRIS: Scalable and Efficient Neural Network Acceleration with 3D Memory," *ACM SIGARCH Computer Architecture News*, vol. 45, no. 1, pp. 751–764, may 2017. [Online]. Available: https://dl.acm.org/doi/10.1145/3093337.3037702

[37] Y. Chen, J. Emer, and V. Sze, "Eyeriss: A Spatial Architecture for Energy-Efficient Dataflow for Convolutional Neural Networks," *IEEE Micro*, p. 1, 2018. [Online]. Available: https://doi.org/10.1109/MM.2017.265085944

[38] Y. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, jan 2017. [Online]. Available: https://doi.org/10.1109/JSSC.2016.2616357

[39] Y. Chen, T. Chen, Z. Xu, N. Sun, and O. Temam, "DianNao Family: Energy-Efficient Accelerators For Machine Learning," *Communications of the ACM*, vol. 59, no. 11, pp. 105–112, oct 2016. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3013530.2996864

[40] Y. Chen, T. Luo, S. Liu, S. Zhang, L. He, J. Wang, L. Li, T. Chen, Z. Xu, N. Sun, and O. Temam, "DaDianNao: A Machine-Learning Supercomputer," in *2014 47th Annual IEEE/ACM International Symposium on Microarchitecture*. IEEE, dec 2014, pp. 609–622. [Online]. Available: https://doi.org/10.1109/MICRO.2014.58

[41] Z. Du, R. Fasthuber, T. Chen, P. Ienne, L. Li, T. Luo, X. Feng, Y. Chen, and O. Temam, "ShiDianNao: Shifting vision processing closer to the sensor," in *ACM SIGARCH Computer Architecture News*, vol. 43, no. 3. ACM, 2015, pp. 92–104. [Online]. Available: https://doi.org/10.1145/2749469.2750389

[42] D. Liu, T. Chen, S. Liu, J. Zhou, S. Zhou, O. Teman, X. Feng, X. Zhou, and Y. Chen, "Pudiannao: A polyvalent machine learning accelerator," in *ACM SIGARCH Computer Architecture News*, vol. 43, no. 1. ACM, 2015, pp. 369–381. [Online]. Available: https://dl.acm.org/doi/10.1145/2775054.2694358

[43] F. Akopyan, J. Sawada, A. Cassidy, R. Alvarez-Icaza, J. Arthur, P. Merolla, N. Imam, Y. Nakamura, P. Datta, G. Nam, B. Taba, M. Beakes, B. Brezzo, J. B. Kuang, R. Manohar, W. P. Risk, B. Jackson, and D. S. Modha, "TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 34, no. 10, pp. 1537–1557, oct 2015. [Online]. Available: https://doi.org/10.1109/TCAD.2015.2474396

[44] M. Feldman, "IBM Finds Killer App for TrueNorth Neuromorphic Chip," sep 2016. [Online]. Available: https://www.top500.org/news/ibm-finds-killer-app-for-truenorth-neuromorphic-chip/

[45] J. Hruska, "New Movidius Myriad X VPU Packs a Custom Neural Compute Engine," aug 2017. [Online]. Available: https://www.extremetech.com/computing/254772-new-movidius-myriad-x-vpu-packs-custom-neural-compute-engine

[46] "Edge TPU," 2019. [Online]. Available: https://cloud.google.com/edge-tpu/

[47] "Rockchip Released Its First AI Processor RK3399Pro NPU Performance Up to 2.4TOPs," jan 2018. [Online]. Available: https://www.rock-chips.com/a/en/News/Press_Releases/2018/0108/869.html

[48] J. Turley, "GAP9 for ML at the Edge EEJournal," jun 2020. [Online]. Available: https://www.eejournal.com/article/gap9-for-ml-at-the-edge/

[49] "GAP application processors - GreenWaves Technologies," 2020. [Online]. Available: https://greenwaves-technologies.com/gap8_gap9/

[50] S. Ward-Foxton, “Kneron’s Next-Gen Edge AI Chip Gets $40m Boost,” jan 2020. [Online]. Available: https://www.eetasia.com/knerons-next-gen-edge-ai-chip-gets-40m-boost/

[51] R. Merritt, “Startup Accelerates AI at the Sensor,” feb 2019. [Online]. Available: https://www.eetimes.com/document.asp?doc_id=1334301

[52] D. McGrath, “Tech Heavyweights Back AI Chip Startup,” oct 2018. [Online]. Available: https://www.eetimes.com/tech-heavyweights-back-ai-chip-startup/

[53] D. Fick, “Mythic @ Hot Chips 2018 - Mythic - Medium,” aug 2018. [Online]. Available: https://medium.com/mythic-ai/mythic-hot-chips-2018-637dfb9e38b7

[54] N. Hemsoth, “A Mythic Approach to Deep Learning Inference,” aug 2018. [Online]. Available: https://www.nextplatform.com/2018/08/23/a-mythic-approach-to-deep-learning-inference/

[55] J. McGregor, “Perceive Exits Stealth With Super Efficient Machine Learning Chip For Smarter Devices,” apr 2020. [Online]. Available: https://www.forbes.com/sites/tiriasresearch/2020/04/06/perceive-exits-stealth-with-super-efficient-machine-learning-chip-for-smarter-devices/#1b25ab646d9c

[56] S. Ward-Foxton, “XMOS adapts Xcore into AIoT crossover processor’ — EE Times,” feb 2020. [Online]. Available: https://www.eetimes.com/xmos-adapts-xcore-into-aiot-crossover-processor/#

[57] D. Schor, “Arm Ethos is for Ubiquitous AI At the Edge — WikiChip Fuse,” feb 2020. [Online]. Available: https://fuse.wikichip.org/news/3282/arm-ethos-is-for-ubiquitous-ai-at-the-edge/

[58] B. Wheeler, “Bitmain SoC Brings AI to the Edge,” feb 2019. [Online]. Available: https://www.linleygroup.com/newsletters/newsletter_detail.php?num=5975&year=2019&tag=3

[59] J. Yoshida, “NovuMind’s AI Chip Sparks Controversy,” oct 2018. [Online]. Available: https://www.eetimes.com/novuminds-ai-chip-sparks-controversy/

[60] K. Freund, “NovuMind: An Early Entrant in AI Silicon,” Moor Insights & Strategy, Tech. Rep., may 2019. [Online]. Available: https://moorinsightsstrategy.com/wp-content/uploads/2019/05/NovuMind-An-Early-Entrant-in-AI-Silicon-By-Moor-Insights-And-Strategy.pdf

[61] S. Ward-Foxton, “Gyrfalcon Unveils Fourth AI Accelerator Chip — EE Times,” nov 2019. [Online]. Available: https://www.eetimes.com/gyrfalcon-unveils-fourth-ai-accelerator-chip/

[62] “aiWare3 Hardware IP Helps Drive Autonomous Vehicles To Production,” oct 2018. [Online]. Available: https://aimotive.com/news/content/1223

[63] P. Clarke, “Indo-US startup preps agent-based AI processor,” aug 2018. [Online]. Available: https://www.eenewsanalog.com/news/indo-us-startup-preps-agent-based-ai-processor/page/0/1

[64] S. Ward-Foxton, “Details of Hailo AI Edge Accelerator Emerge,” aug 2019. [Online]. Available: https://www.eetimes.com/details-of-hailo-ai-edge-accelerator-emerge/

[65] “Horizon Robotics Journey2 Automotive AI Processor Series,” 2020. [Online]. Available: https://en.horizon.ai/product/journey

[66] Huawei, “Ascend 310 AI Processor,” 2020. [Online]. Available: https://e.huawei.com/us/products/cloud-computing-dc/atlas/ascend-310

[67] B. Dupont de Dinechin, “Kalray’s MPPA® Manycore Processor: At the Heart of Intelligent Systems,” in *17th IEEE International New Circuits and Systems Conference (NEWCAS)*. Munich: IEEE, jun 2019. [Online]. Available: https://www.european-processor-initiative.eu/dissemination-material/1259/

[68] P. Clarke, “NXP, Kalray demo Coolidge parallel processor in ’BlueBox’,” jan 2020. [Online]. Available: https://www.eenewsanalog.com/news/nxp-kalray-demo-coolidge-parallel-processor-bluebox

[69] D. Franklin, “NVIDIA Jetson TX2 Delivers Twice the Intelligence to the Edge,” mar 2017. [Online]. Available: https://developer.nvidia.com/blog/jetson-tx2-delivers-twice-intelligence-edge/

[70] J. Hruska, “Nvidia’s Jetson Xavier Stuffs Volta Performance Into Tiny Form Factor,” jun 2018. [Online]. Available: https://www.extremetech.com/computing/270681-nvidias-jetson-xavier-stuffs-volta-performance-into-tiny-form-factor

[71] D. Firu, “Quadric Edge Supercomputer,” Quadric, Tech. Rep., apr 2019. [Online]. Available: https://quadric.io/supercomputing.pdf

[72] “FSD Chip - Tesla,” 2020. [Online]. Available: https://en.wikichip.org/wiki/tesla_(car_company)/fsd_chip

[73] E. Talpes, D. D. Sarma, G. Venkataramanan, P. Bannon, B. McGee, B. Floering, A. Jalote, C. Hsiong, S. Arora, A. Gorti, and G. S. Sachdev, “Compute Solution for Tesla’s Full Self-Driving Computer,” *IEEE Micro*, vol. 40, no. 2, pp. 25–35, mar 2020. [Online]. Available: https://doi.org/10.1109/MM.2020.2975764

[74] R. Merritt, “Samsung, Toshiba Detail AI Chips,” feb 2019. [Online]. Available: https://www.eetimes.com/samsung-toshiba-detail-ai-chips/

[75] J. De Gelas, “Intel’s Xeon Cascade Lake vs. NVIDIA Turing: An Analysis in AI,” jul 2019. [Online]. Available: https://www.anandtech.com/show/14466/intel-xeon-cascade-lake-vs-nvidia-turing

[76] D. Schor, “The 2,048-core PEZY-SC2 sets a Green500 record — WikiChip Fuse,” nov 2017. [Online]. Available: https://fuse.wikichip.org/news/191/the-2048-core-pezy-sc2-sets-a-green500-record/

[77] L. Gwennap, “Tenstorrent Scales AI Performance: Architecture Leads in Data-Center Power Efficiency,” Microprocessor Report, Tech. Rep., apr 2020. [Online]. Available: https://www.tenstorrent.com/wp-content/uploads/2020/04/Tenstorrent-Scales-AI-Performance.pdf

[78] N. Hemsoth, “Intel FPGA Architecture Focuses on Deep Learning Inference,” jul 2018. [Online]. Available: https://www.nextplatform.com/2018/07/31/intel-fpga-architecture-focuses-on-deep-learning-inference/

[79] M. S. Abdelfattah, D. Han, A. Bitar, R. DiCecco, S. O’Connell, N. Shanker, J. Chu, I. Prins, J. Fender, A. C. Ling, and G. R. Chiu, “DLA: Compiler and FPGA Overlay for Neural Network Inference Acceleration,” in *2018 28th International Conference on Field Programmable Logic and Applications (FPL)*, aug 2018, pp. 411–4117. [Online]. Available: https://doi.org/10.1109/FPL.2018.00077

[80] G. Roos, “FPGA acceleration card delivers on bandwidth, speed, and flexibility,” nov 2019. [Online]. Available: https://www.eetimes.com/fpga-acceleration-card-delivers-on-bandwidth-speed-and-flexibility/

[81] “Cornami Achieves Unprecedented Performance at Lowest Power Dissipation for Deep Neural Networks,” oct 2019. [Online]. Available: https://cornami.com/1416-2/

[82] V. Mehta, “Performance Estimation and Benchmarks for Real-World Edge Inference Applications,” in *Linley Spring Processor Conference*. Linley Group, 2020.

[83] T. P. Morgan, “Drilling Into Microsoft’s BrainWave Soft Deep Learning Chip,” aug 2017. [Online]. Available: https://www.nextplatform.com/2017/08/24/drilling-microsofts-brainwave-soft-deep-leaning-chip/

[84] D. Chiou, “The microsoft catapult project,” in *2017 IEEE International Symposium on Workload Characterization (IISWC)*. Institute of Electrical and Electronics Engineers (IEEE), dec 2017, pp. 124–124.

[85] “NVIDIA Tesla V100 Tensor Core GPU,” 2019. [Online]. Available: https://www.nvidia.com/en-us/data-center/tesla-v100/

[86] R. Smith, “16GB NVIDIA Tesla V100 Gets Reprieve; Remains in Production,” may 2018. [Online]. Available: https://www.anandtech.com/show/12809/16gb-nvidia-tesla-v100-gets-reprieve-remains-in-production

[87] E. Kilgariff, H. Moreton, N. Stam, and B. Bell, “NVIDIA Turing Architecture In-Depth,” sep 2018. [Online]. Available: https://developer.nvidia.com/blog/nvidia-turing-architecture-in-depth/

[88] R. Krashinsky, O. Giroux, S. Jones, N. Stam, and S. Ramaswamy, “NVIDIA Ampere Architecture In-Depth,” may 2020. [Online]. Available: https://devblogs.nvidia.com/nvidia-ampere-architecture-in-depth/

[89] ExxactCorp, “Taking a Deeper Look at AMD Radeon Instinct GPUs for Deep Learning,” dec 2017. [Online]. Available: https://blog.exxactcorp.com/taking-deeper-look-amd-radeon-instinct-gpus-deep-learning/

[90] R. Smith, “AMD Announces Radeon Instinct MI60 & MI50 Accelerators Powered By 7nm Vega,” nov 2018. [Online]. Available: https://www.anandtech.com/show/13562/amd-announces-radeon-instinct-mi60-mi50-accelerators-powered-by-7nm-vega

[91] T. Peng, “Alibaba’s New AI Chip Can Process Nearly 80K Images Per Second,” 2019. [Online]. Available: https://medium.com/syncedreview/alibabas-new-ai-chip-can-process-nearly-80k-images-per-second-63412dec22a3

[92] J. Hamilton, “AWS Inferentia Machine Learning Processor,” nov 2018. [Online]. Available: https://perspectives.mvdirona.com/2018/11/aws-inferentia-machine-learning-processor/

[93] C. Evangelist, “Deep dive into Amazon Inferentia: A custom-built chip to enhance ML and AI,” jan 2020. [Online]. Avail-

able: https://www.cloudmanagementinsider.com/amazon-inferentia-for-machine-learning-and-artificial-intelligence/

[94] R. Merritt, “Baidu Accelerator Rises in AI,” jul 2018. [Online]. Available: https://www.eetimes.com/baidu-accelerator-rises-in-ai/

[95] C. Duckett, “Baidu Creates Kunlun Silicon for AI,” jul 2018. [Online]. Available: https://www.zdnet.com/article/baidu-creates-kunlun-silicon-for-ai/

[96] I. Cutress, “Cambricon, Maker of Hauwei’s Kirin NPU IP, Build a Big AI Chip and PCIe Card,” may 2018. [Online]. Available: https://www.anandtech.com/show/12815/cambricon-makers-of-huaweis-kirin-npu-ip-build-a-big-ai-chip-and-pcie-card

[97] P. Teich, “Tearing Apart Google’s TPU 3.0 AI Coprocessor,” may 2018.

[98] D. Lacey, “Preliminary IPU Benchmarks,” oct 2017. [Online]. Available: https://www.graphcore.ai/posts/preliminary-ipu-benchmarks-providing-previously-unseen-performance-for-a-range-of-machine-learning-applications

[99] L. Gwennap, “Groq Rocks Neural Networks,” Microprocessor Report, Tech. Rep., jan 2020. [Online]. Available: http://groq.com/wp-content/uploads/2020/04/Groq-Rocks-NNs-Linley-Group-MPR-2020Jan06.pdf

[100] D. Abts, J. Ross, J. Sparling, M. Wong-VanHaren, M. Baker, T. Hawkins, A. Bell, J. Thompson, T. Kahsai, G. Kimmell, J. Hwang, R. Leslie-Hurd, M. Bye, E. R. Creswick, M. Boyd, M. Venigalla, E. Laforge, J. Purdy, P. Kamath, D. Maheshwari, M. Beidler, G. Rosseel, O. Ahmad, G. Gagarin, R. Czekalski, A. Rane, S. Parmar, J. Werner, J. Sproch, A. Macias, and B. Kurtz, “Think Fast: A Tensor Streaming Processor (TSP) for Accelerating Deep Learning Workloads,” in *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*, may 2020, pp. 145–158. [Online]. Available: https://doi.org/10.1109/ISCA45697.2020.00023

[101] Huawei, “Ascend 910 AI Processor,” 2020. [Online]. Available: https://e.huawei.com/us/products/cloud-computing-dc/atlas/ascend-910

[102] L. Gwennap, “Habana Wins Cigar for AI Inference,” feb 2019. [Online]. Available: https://www.linleygroup.com/mpr/article.php?id=12103

[103] E. Medina and E. Dagan, “Habana Labs Purpose-Built AI Inference and Training Processor Architectures: Scaling AI Training Systems Using Standard Ethernet With Gaudi Processor,” *IEEE Micro*, vol. 40, no. 2, pp. 17–24, mar 2020. [Online]. Available: https://doi.org/10.1109/MM.2020.2975185

[104] L. Gwennap, “Habana Offers Gaudi for AI Training,” Microprocessor Report, Tech. Rep., jun 2019. [Online]. Available: https://habana.ai/wp-content/uploads/2019/06/Habana-Offers-Gaudi-for-AI-Training.pdf

[105] A. Hock, “Introducing the Cerebras CS-1, the Industry’s Fastest Artificial Intelligence Computer - Cerebras,” nov 2019. [Online]. Available: https://www.cerebras.net/introducing-the-cerebras-cs-1-the-industrys-fastest-artificial-intelligence-computer/

[106] P. Alcorn, “Nvidia Infuses DGX-1 with Volta, Eight V100s in a Single Chassis,” may 2017. [Online]. Available: https://www.tomshardware.com/news/nvidia-volta-v100-dgx-1-hgx-1,34380.html

[107] I. Cutress, “NVIDIA’s DGX-2: Sixteen Tesla V100s, 30TB of NVMe, Only $400K,” mar 2018. [Online]. Available: https://www.anandtech.com/show/12587/nvidias-dgx2-sixteen-v100-gpus-30-tb-of-nvme-only-400k

[108] C. Campa, C. Kawalek, H. Vo, and J. Bessoudo, “Defining AI Innovation with NVIDIA DGX A100,” may 2020. [Online]. Available: https://devblogs.nvidia.com/defining-ai-innovation-with-dgx-a100/

[109] “Dell DSS8440 Graphcore IPU Server,” Graphcore, Tech. Rep., feb 2020. [Online]. Available: https://www.graphcore.ai/hubfs/Lead gen assets/DSS8440 IPU Server White Paper_2020.pdf

[110] D. Lacey, “Updated Graphcore IPU benchmarks,” jun 2020. [Online]. Available: https://www.graphcore.ai/posts/new-graphcore-ipu-benchmarks

[111] “SolidRun, Gyrfalcon Develop Arm-based Edge Optimized AI Inference Server,” feb 2020. [Online]. Available: https://www.hpcwire.com/off-the-wire/solidrun-gyrfalcon-develop-edge-optimized-ai-inference-server/

[112] D. McGrath, “Qualcomm Targets AI Inferencing in the Cloud — EE Times,” apr 2019. [Online]. Available: https://www.eetimes.com/qualcomm-targets-ai-inferencing-in-the-cloud/#

[113] J. Yoshida, “Blaize Fires up GSP for AI Processing,” nov 2019. [Online]. Available: https://www.eetimes.com/blaize-fires-up-gsp-for-ai-processing/

[114] “Enflame Technology Announces CloudBlazer with DTU Chip on GLOBALFOUNDRIES 12LP FinFET Platform for Data Center Training,” dec 2019. [Online]. Available: https://www.hpcwire.com/off-the-wire/enflame-technology-announces-cloudblazer-with-dtu-chip-on-globalfoundries-12lp-finfet-platform-for-data-center-training/

[115] L. Gwennap, “Esperanto Maxes Out RISC-V High-End Maxion CPU Raises RISC-V Performance Bar,” Microprocessor Report, Tech. Rep., 2018.

[116] “BrainChip Showcases Vision and Learning Capabilities of its Akida Neural Processing IP and Device at tinyML Summit 2020,” feb 2020. [Online]. Available: https://www.design-reuse.com/news/47498/brainchip-akida-neural-processing-ip-tinyml-summit-2020.html

[117] S. K. Moore, “Low-Power AI Startup Eta Compute Delivers First Commercial Chips - IEEE Spectrum,” 2020. [Online]. Available: https://spectrum.ieee.org/tech-talk/semiconductors/processors/lowpower-ai-startup-eta-compute-delivers-first-commercial-chips

[118] “Baidu Backs Neuromorphic IC Developer,” nov 2018. [Online]. Available: https://www.eetimes.com/baidu-backs-neuromorphic-ic-developer/

[119] P. Clarke, “AI chip startup offers new edge computing solution,” dec 2018. [Online]. Available: https://www.smart2zero.com/news/ai-chip-startup-offers-new-edge-computing-solution#

[120] M. Kim, J. Kim, C. Park, L. Everson, H. Kim, S. Song, S. Lee, and C. H. Kim, “A 68 Parallel Row Access Neuromorphic Core with 22K Multi-Level Synapses Based on Logic-Compatible Embedded Flash Memory Technology,” in *Technical Digest - International Electron Devices Meeting, IEDM*, vol. 2018-Decem. Institute of Electrical and Electronics Engineers Inc., jan 2019, pp. 15.4.1–15.4.4. [Online]. Available: https://doi.org/10.1109/IEDM.2018.8614599

[121] N. Dahad, “Startup Launches Its First Low Latency Edge AI Chip,” oct 2019. [Online]. Available: https://www.eetimes.eu/startup-launches-its-first-low-latency-edge-ai-chip/

[122] P. Clarke, “GrAI Matter research gives rise to AI processor for the edge,” jan 2020. [Online]. Available: https://www.eenewsanalog.com/news/grai-matter-paris-research-gives-rise-ai-processor-edge

[123] O. E. Agabi, “Cell Culture, Transport and Investigation,” pp. 1—-51, 2016. [Online]. Available: https://patents.google.com/patent/US20170015964A1/en

[124] N. Hemsoth, “First Wave of Spiking Neural Network Hardware Hits,” sep 2018. [Online]. Available: https://www.nextplatform.com/2018/09/11/first-wave-of-spiking-neural-network-hardware-hits/

[125] J. Dunietz, “Light-Powered Computers Brighten AI’s Future,” *Scientific American*, jun 2017. [Online]. Available: https://www.scientificamerican.com/article/light-powered-computers-brighten-ai-rsquo-s-future/

[126] M. Feldman, “The Silicon Photonics Key to Building Better Neural Networks,” 2019. [Online]. Available: https://www.nextplatform.com/2019/05/21/the-silicon-photonics-key-to-building-better-neural-networks/

[127] ——, “Photonic Computing Company Takes Aim at Artificial Intelligence,” 2018. [Online]. Available: https://www.top500.org/news/photonic-computing-company-takes-aim-at-artificial-intelligence/

[128] “Photonic Computing for Massively Parallel AI,” LightOn, Paris, Tech. Rep., may 2020. [Online]. Available: https://lighton.ai/wp-content/uploads/2020/05/LightOn-White-Paper-v1.0-S.pdf

[129] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljacic, “Deep learning with coherent nanophotonic circuits,” *Nature Photonics*, vol. 11, no. 7, pp. 441–446, jun 2017. [Online]. Available: https://www.nature.com/articles/nphoton.2017.93

[130] P. Clarke, “Startup reveals prototype optical AI processor, says report,” apr 2019. [Online]. Available: https://www.eenewseurope.com/news/startup-reveals-prototype-optical-ai-processor

[131] M. Feldman, “Optalysys Claims AI Breakthrough Using Optical Processing Technology,” mar 2018. [Online]. Available: https://www.top500.org/news/optalysys-claims-ai-breakthrough-using-optical-processing-technology/

[132] ——, “Optical Computing Startup Demos Training of Neural Networks,” feb 2018. [Online]. Available: https://www.top500.org/news/optical-computing-startup-demos-training-of-neural-networks/

[133] M. Giles, “Bill Gates just backed a chip startup that uses light to turbocharge AI,” jun 2019. [On-

line]. Available: https://www.technologyreview.com/2019/06/13/867/ai-chips-uses-optical-semiconductor-machine-learning/

[134] K. Freund, “Intel Lays Out Strategy For AI: It’s Habana,” jan 2020. [Online]. Available: https://www.forbes.com/sites/moorinsights/2020/01/31/intel-lays-out-strategy-for-ai-its-habana/#64ca5b294dd3

[135] P. Clarke, “Military startup aims large with neural processor chip,” jun 2016. [Online]. Available: https://www.eenewseurope.com/news/military-startup-aims-large-neural-processor-chip