AI Accelerator Survey and Trends

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—Over the past several years, new machine learning accelerators were being announced and released every 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 AI accelerators and processors from past two 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. This year, we also compile a list of benchmarking performance results and compute the computational efficiency with respect to peak performance.

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

I. INTRODUCTION

Over the past several years, startups and established technology companies have been announcing, releasing, and deploying a wide variety of artificial intelligence (AI) and machine learning (ML) accelerators. The focus of these accelerators has been on accelerating deep neural network (DNN) models, and the application space spans from very low power embedded voice recognition to data center scale training. The announcements of new accelerators has slowed in the past year, but the competition for defining markets and application areas continues. This drive to developing and deploying accelerators has been part of a much larger industrial and technology shift in modern computing.

AI ecosystems bring together a 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. These raw data products are fed into a data conditioning step in which they are fused, aggregated, structured, accumulated, and converted into 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.

This paper updates the survey of AI accelerators and processors from past two years [9], [11]. As in past years, we will review 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]. Taking a page from the system-on-chip (SoC) trends first seen in automotive and smartphones, advancements and innovations are still progressing by developing and integrating accelerators for 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 deep machine learning processors and accelerators [4]–[8]. 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 IEEE-HPEC papers from the past two years [9], [10]. As in past years, 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 and optical 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 accelerators and processors that are geared toward deep neural networks (DNNs) and convolutional neural networks (CNNs) as they are quite computationally intense [11].

- **Neural Network Training versus Inference** – As was explained in the previous two surveys, the 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.

- **Numerical Precision** – 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). But as can be seen in Figure 2, peak performance has been reported for many different numerical formats.

- **Neuromorphic Computing and Photonic Computing** – For this year’s survey, we are going to take a pause on most of the neuromorphic computing and photonic computing accelerators. Several new accelerators have been released, but none of these companies have released peak performance and peak power numbers for them. There have been some relative comparisons of neuromorphic processors to conventional accelerators (e.g., [12]), but there have been no hard numbers. Perhaps next year we will start seeing actual performance numbers that we can incorporate in this survey.

There are many surveys [13]–[22] and other papers that cover various aspects of AI accelerators; 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.

## II. Survey of Processors

Many recent advances in AI can be at least partly credited to advances in computing hardware [23], [24], 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 2021) mapping peak performance vs. power consumption.

The x-axis indicates peak power, and the y-axis indicate peak giga-operations per second (GOps/s), both on a logarithmic scale. Note the legend on the right, which indicates various parameters used to differentiate computing precisions, form factors, and inference/training. 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 200-400 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.

The survey begins with the same scatter plot that we have compiled for the past two years [9], [10]. 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 programmed or “place-and-routed” onto the computational hardware. Processor in memory (PIM) is 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.

We can make some general observations from Figure 2. First, a few new accelerator chips, cards, and systems have been announced and released in the past year. The density has clearly increased in the autonomous ellipse and data center...
| Company       | Product                        | Label      | Technology | Form Factor | References |
|---------------|--------------------------------|------------|------------|-------------|------------|
| AMD           | Radeon Instinct MI6            | AMD-MI6    | GPU        | Card        | [32]       |
| AMD           | Radeon Instinct MI60           | AMD-MI60   | GPU        | Card        | [33]       |
| ARM           | Ethos N77                      | Ethos      | dataflow   | Chip        | [34]       |
| Baai          | Haiku 818-300                  | Haiku      | dataflow   | Chip        | [35], [36] |
| Hitmain       | Hitman                        | Hitman     | dataflow   | Chip        | [37]       |
| Blaze         | El Cano                        | Blaze      | dataflow   | Card        | [38]       |
| Cambricon     | MLU100                         | Cambricon  | dataflow   | Card        | [39], [40] |
| Canaan        | Kendryte                       | Kendryte   | CPU        | Chip        | [41]       |
| Cerebras      | CS-1                           | CS-1       | dataflow   | System      | [42]       |
| Cerebras      | CS-2                           | CS-2       | dataflow   | System      | [43]       |
| Cornami       | Cornami                        | Cornami    | dataflow   | Chip        | [44]       |
| Enflame       | Cloudblazer T10                | Enflame    | CPU        | Card        | [45]       |
| Flex Logix    | InferX-X1                      | FlexLogix  | dataflow   | Chip        | [46]       |
| Google        | TPU1                           | TPU1       | dataflow   | Chip        | [47]       |
| Google        | TPU2                           | TPU2       | dataflow   | Chip        | [48], [49] |
| Google        | TPU3                           | TPU3       | dataflow   | Chip        | [50], [51] |
| GraphCore     | C2                             | GraphCore  | dataflow   | Card        | [52], [53] |
| GreenWaves    | GAP9                           | GreenWaves | dataflow   | Chip        | [54], [55] |
| Groq          | Groq Node                      | GroqNode   | dataflow   | System      | [56]       |
| Groq          | Tensor Streaming Processor     | Groq       | dataflow   | Card        | [57], [58] |
| Gyrfalcon     | Gyrfalcon                      | Gyrfalcon  | PIM        | Chip        | [59]       |
| Hana          | Gaudi                          | Gaudi      | dataflow   | Card        | [60], [61] |
| Hana          | Goya-1000                      | Goya       | dataflow   | Card        | [62], [63] |
| Hailo         | Hallao                        | Hallao-80  | dataflow   | Chip        | [64]       |
| Horizon Robotics | Journey2                  | Journey2   | dataflow   | Chip        | [65]       |
| Huawei        | Ascend-310                     | Ascend-310 | dataflow   | Chip        | [66]       |
| Huawei        | Ascend-910                     | Ascend-910 | dataflow   | Chip        | [67]       |
| IBM           | TrueNorth                      | TrueNorth  | neumorphicsystem | [68], [69] |
| IBM           | TrueNorthSys                   | TrueNorthSys | neumorphicsystem | [70]       |
| Intel         | Arria 10 1150                  | Arria      | FPGA       | Chip        | [71], [72] |
| Intel         | MobileEye EyeQ5                | EyeQ5      | dataflow   | Chip        | [73]       |
| Intel         | MentorVX Myriad X              | MentorVX   | manycore   | Chip        | [74]       |
| Intel         | Xeon Platinum 8180             | Xeon8180   | multicore  | System      | [75]       |
| Intel         | Xeon Platinum 8280             | Xeon8280   | multicore  | System      | [76]       |
| Kallay        | Coolidge                       | Kallay     | manycore   | Chip        | [77], [78] |
| Kneron        | KL-520 Neural Processing Unit  | KL-520     | dataflow   | Chip        | [79]       |
| Kneron        | KL-720                        | KL-720     | dataflow   | Chip        | [80]       |
| Microsoft     | Brainwave                     | Brainwave  | dataflow   | Chip        | [81]       |
| Mythic        | M1076                          | Mythic7/6  | PIM        | Chip        | [82], [83] |
| Mythic        | M1018                          | Mythic108  | PIM        | Chip        | [84], [85] |
| NovoMind      | Novo intoler                    | NovoIntoler | dataflow   | Chip        | [86], [87] |
| NVIDIA        | Ampere A100                    | A100       | GPU        | Card        | [88], [89] |
| NVIDIA        | Ampere A40                     | A40        | GPU        | Card        | [90], [91] |
| NVIDIA        | Ampere A30                     | A30        | GPU        | Card        | [92], [93] |
| NVIDIA        | Ampere A10                     | A10        | GPU        | Card        | [94]       |
| NVIDIA        | Pascal Pi100                   | Pi100      | GPU        | Card        | [95], [96] |
| NVIDIA        | T4                             | T4         | GPU        | Card        | [97]       |
| NVIDIA        | Volta V100                     | V100       | GPU        | Card        | [98], [99] |
| NVIDIA        | DOX Station                    | DOX-Station | GPU        | System      | [100]       |
| NVIDIA        | DOX-1                          | DOX-1      | GPU        | System      | [101], [102] |
| NVIDIA        | DOX-2                          | DOX-2      | GPU        | System      | [103]       |
| NVIDIA        | DOX-A100                       | DOX-A100   | GPU        | System      | [104]       |
| NVIDIA        | Jetsol TX1                     | Jetsol1    | GPU        | System      | [105]       |
| NVIDIA        | Jetsol TX2                     | Jetsol2    | GPU        | System      | [106]       |
| NVIDIA        | Jeter Xavier NX                | XavierNX   | GPU        | System      | [107]       |
| NVIDIA        | Jeter XAVI Xvar                | XavierXAV  | GPU        | System      | [108]       |
| Percevis      | Ergo                           | Percevis   | dataflow   | Chip        | [109]       |
| PEZY Computing | PEZY-SC2                   | PEZY-SC2   | manycore   | System      | [110]       |
| Preferred Networks | Preferred-MN-3               | Preferred-MN-3 | manycore | Card | [100], [101] |
| Qualcomm      | Cloud AI 100                   | Qcomm      | dataflow   | Card        | [102], [103] |
| Rockchip      | RK3399Pro                     | RK3399Pro  | dataflow   | Chip        | [104]       |
| SiMa.ai       | SiMa.ai                       | SiMa.ai    | dataflow   | Chip        | [105]       |
| Systiant      | NP1010                        | Symiant    | PIM        | Card        | [106], [107] |
| Tensorgent    | Tensorgent                    | Tensorgent | manycore   | Card        | [108], [109] |
| Tesla         | Tesla Full Self-Driving Computer | Tesla | dataflow | System | [110], [111] |
| Texas Instruments | TDAVM                   | TDAVM      | dataflow   | Chip        | [112], [113] |
| Toshiba       | Toshiba                        | Toshiba    | multicore  | System      | [114]       |
| Unitrother    | TunaAlmi                      | TunaAlmi   | PIM        | Card        | [115]       |
| XAMOS         | xcore.ai                      | xcore.ai   | dataflow   | Chip        | [116]       |
cards and chips ellipse. Further, several cards and chips have
been released that are focused on inference that exceeds a peak
power of 100W, e.g., Intel Habana Goya, NVIDIA Ampere
A10 and A40, Alibaba, and Groq. This is a deviation from
the last few years. This suggests that the power budget for
autonomous vehicles and drones has crept past 100W, and that
these accelerators are aimed at both the autonomous vehicle
and data center inference markets. When it comes to precision,
int8 has become the default numerical precision for embedded,
autonomous and data center inference applications. Along with
int8 for inference, a number of accelerators are also featuring
fp16 and/or bf16 for inference. Finally, the competition for
high-end training nodes shown in the data center systems
eclipse is intensifying. NVIDIA and Cerebras have very high
performing nodes, while Graphcore and Groq also have strong
targets. Google TPUs and SambaNova also are competing
in this space, but they have only been reporting multi-node
benchmark results, rather than single system peak capabilities.

A. New Accelerators

For most of the accelerators, their descriptions and com-
mentaries have not changed since last year so please refer to
last year’s paper for descriptions and commentaries. There are,
however, several new releases that were not covered by last
year’s paper that are covered here. 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.

• Blaize has emerged from stealth mode and announced
its Graph Streaming Processor (GSP) [39], but they have
not provided any details beyond a high level component
diagram of their chip.

• Enflame Technology, backed by Tencent, started shipping
its CloudBlazer T10 data center training accelerator PCIe
card [46], [118], which will support a broad range of
datatype including fp32, fp16, bf16, int32, int16, and
int8. The T10 accelerator is focused on data center DNN
training applications.

• Untether announced their TsunAlmi card, which features
four RunAI200 chips, during the Fall of 2020. Their
at-memory design places 250,000 processing elements
within a standard SRAM array. They are targeting the
inference market and expect to ship cards in the first half
of 2021.

• The Texas Instruments TDA4VM chip (TexInst)
is a
feature-rich automotive/autonomous system on a chip
(SoC). It not only includes a 8 TOPS (int8) deep learning
matrix multiply accelerator (MMA) with 4,096 computa-
tional units, but also eight ARM cores, two C7x vector
DSPs, two C66x DSPs, 8 MB of L3 RAM, and several
other audio, video, and security accelerators. [112]–[114]

• Mobileye, a subsidiary of Intel, has released its fifth
generation automotive AI processor, EyeQ5 (EyeQ5). It
includes eight CPU cores and 18 computer vision AI
processors [39]. [119].

• The updated Mythic Intelligent Processing Unit acceler-
ator (Mythic76) [83], [84] combines a RISC-V control
processor, router, and flash memory that uses variable
resistors (i.e., analog circuitry) to compute matrix multi-
plies. The accelerators are aiming for embedded, vision,
and data center applications. It is a smaller, lower-power
76 sq.mm. version of the 108 sq.mm., which is labeled (Mythic108).

- Qualcomm has announced their Cloud AI 100 accelerator (QComm) [104], 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.
- Several new NVIDIA Ampere data center GPU cards have been released in the past year. The Ampere A40 (A40) and A10 (A10) are follow-on GPUs for data center inference to the Turing T4 card, while the Ampere A30 (A30) is a lower-performance, lower-power, more affordable version of the Ampere A100 training and HPC card [88].
- In June 2021, Google shared details about their fourth generation inference-only TPU4i accelerator (TPU4i) [51]. It features four 16k-element systolic matrix multiply units and was first deployed in early 2020. As with previous TPU variants, TPU4i is available through the Google Compute Cloud and for internal operations.
- Cerebras partnered with the TSMC chip fabrication foundry to scale its tiled wafer scale engine (WSE) from 16-nm feature size to 7-nm. The result is the Cerebras CS-2 (CS-2) [44], which has 850,000 simple arithmetic units. On-board memory scales commensurately along with local memory bandwidth and internal bandwidth. The CS-2 chassis is the same 15-U rack mount system and 12 100-GigE network uplinks, and the system draws up to 23 KW of power.

Finally, we must mention three 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.

- Graphcore has announced its second generation accelerator chip, the CG200, which they are offering in their M2000 IPU Machine computer node. The M2000 incorporates four CG200 accelerators and Graphcore reports that the M2000 is capable of over a petaflop/s of performance [120], [121]. While Graphcore has released some training results for the MLPerf benchmark [122], they have not disclosed any peak power or peak performance values.
- SambaNova has released some impressive benchmark results for their reconfigurable AI accelerator technology, but they still have not provided any details from which we can estimate peak performance or power consumption of their solutions [123].
- The Centaur Technology CNS processor [124], [125] includes eight x86 cores along with an integrated AI accelerator realized as a 4,096 byte-wide SIMD unit. The Centaur AI coprocessor (CT-AIC) will deliver peak performance of 20 TOPS with INT8 precision at 2.5 GHz and can also operate at FP16 and INT16 precisions, though at lower performance. Centaur has not published any power numbers, though [125] predicts peak power to be less than 85 Watts.

Much in the same way we no longer showed most FPGA-based solutions in last year’s survey, we are leaving out the research oriented chips that have not found their way to commercial production this year. Research chips including Eyeriss [16], [126], [127], EIE [128], Tetris [129], TianJie [130], the DianNao family [131], Adapteva [132], [133], and NeuFlow [134] have been important in showing various computational performance, energy efficiency, and computational accuracy gains that could be achieved with specialized accelerator architectures and circuitry.

III. COMPUTATIONAL EFFICIENCY ANALYSIS

In recent years, a number of companies have been reporting actual performance numbers for their chips, cards, and systems. They have been doing so in the context of various benchmarks including MLPerf [135]. Most of the benchmark results have been for inference, where the metrics are images/items per second for throughput and latency to result. There have also been some training benchmark results, where the metric is time to train a particular DNN model [122]. Since fielded defense and national security applications rely heavily on inference, we will focus on inference this year. Further, we will focus on images per second throughput over latency because in our experience, current defense and national security applications often prioritize throughput rate because the images/items from the sensor platform are collected in a consistent stream. Also DNN inference is more straightforward to characterize from a computational and data motion perspective.

Fortunately, the DNN models that are specified in these benchmarks are well defined; that is, computing an inference output for any image (or other input item), the models are consistent and deterministic in the computations and data motion. The most accessible analysis of a wide set of DNN/CNN models is the online document maintained by Dr. Samuel Albanie [136]. His table reports the number of fused-multiply-add (FMA) operations that dominate the inference (forward pass) computation. However, because it only reports the FMA operations, we must keep in mind that it is only an approximation of all of the computations and data motions involved. With FMA computation per single batch inference from Dr. Albanie’s table, we can compute an approximation of the number of operations per second from the reported number of images processed per second. This is captured in Table 1.

Table 1 is sorted by images-per-second (IPS), and the four Google entries at the bottom. In [51], Google reports average computational efficiency results across the eight most utilized models at Google. Almost all of the accelerators are achieving over 20 percent computational efficiency; often it is challenging to achieve 10 percent computational efficiency on dense computational kernels with reasonably high arithmetic intensity [137]. Considering the highly parallel design of these ML accelerators with wide data transfer paths, it is reasonable to expect that further tuning and optimization should bring the utilization of those below 20 percent higher using techniques including strategic data layouts and data transfer latency hiding. Interestingly, there is no correlation between technology type, precision, or application category with computational utilization percentage.
TABLE II
Inference utilization percentage for select accelerators.

| Company/Org | Accelerator | Tech. | DNN Model | IPS | Perf. (TOPS) | Precision | Utilization Percent | References |
|-------------|-------------|-------|-----------|-----|--------------|-----------|---------------------|------------|
| GreenWaves | GAP9        | dataflow | MobileNetV1 | 83.9 | 0.0489 | int8 | 46% | [56] |
| NovuMind   | NovuTensor  | dataflow | ResNet-50 | 697  | 2.8        | int8 | 19% | [86] |
| Mythic     | Mythic108   | PIM     | ResNet-50 | 900  | 5.6        | analog | 10% | [84] |
| NVIDIA     | Jetson Xavier NX | GPU | ResNext-50 | 1,105 | 47.5 | int8 | 22% | MLPerf 1.0-98 |
| NVIDIA     | Jetson T4X Xavier | GPU | ResNext-50 | 2,072 | 8.3 | int8 | 26% | MLPerf 1.0-92 |
| Intel      | 2Xeon8280   | manycore | GoogleNet | 6,000 | 12 | int8:32 | 50% | [77] |
| Qualcomm   | Cloud AI 100 | dataflow | ResNet-50 | 7,807 | 31 | int8 | 8% | MLPerf 1.0-101 |
| NVIDIA     | V100        | GPU     | ResNet-50 | 7,907 | 31.6 | int8 | 56% | [52] |
| Achronix   | Achronix    | dataflow | ResNet-50 | 8,000 | 34.4 | int8 | 40% | [52] |
| RambusCon  | MLI100      | dataflow | ResNet-50 | 10,000 | 40 | int8 | 31% | [52] |
| Habana     | Goya        | dataflow | ResNet-50 | 15,433 | 61.7 | int8 | 62% | [108] |
| Guroq      | TSP         | dataflow | ResNet-50 | 21,700 | 86.8 | int8 | 11% | [139] |
| Tenstorrent| Tenstorrent | manycore | GoogleNet | 22,431 | 89.7 | int8 | 24% | [109] |
| NVIDIA     | A100        | GPU     | ResNet-50 | 38,010 | 152 | int8 | 24% | MLPerf 1.0-29 |
| Google     | TPU1        | dataflow | Google-8 | – | – | int8 | 20% | [51] |
| Google     | TPU2        | dataflow | Google-8 | – | – | int8 | 20% | [51] |
| Google     | TPU3        | dataflow | Google-8 | – | – | int8 | 20% | [51] |
| Google     | TPU4i       | dataflow | Google-8 | – | – | int8 | 20% | [51] |

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 slowed down some, but we are starting to see second generation accelerators that are significantly improving on the capabilities of the first generation. Actual performance benchmark results are being released more, which gives us the opportunity to evaluate computational efficiency for the first time for accelerators for which we have benchmark results and peak performance numbers. Many of these accelerators achieve over 20 percent computational efficiency.

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 are thankful to Masahiro Arakawa, Bill Arcand, Bill Bergeron, David Bestor, Bob Bond, Chansup Byun, Nathan Frey, Vitaly Gleyzer, Jeff Gottschalk, Michael Houle, Matthew Hubbell, Hayden Janathan, 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://iscexplore. ieee.org/document/6757327
[4] C. E. Leiserson, N. C. Thompson, J. S. Emer, 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. Durukakis, 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
[8] Y. LeCun, “Deep Learning Hardware: Past, Present, and Future,” in 2019 IEEE International Solid- State Circuits Conference - (ISSCC), feb. 2019, pp. 12–19.
[9] A. Reuther, P. Micheales, M. Jones, V. Gadepally, S. Samsi, and J. Kepner, “Survey of Machine Learning Accelerators,” in 2020 IEEE High Performance Extreme Computing Conference (HPEC), 2020, pp. 1–12.
[10] ——, “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
[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] S. Ward-Foxton, “Intel Benchmarks Neuromorphic Chip Against AI Accelerators,” dec 2020. [Online]. Available: https://www.eetimes.com/intel-benchmarks-neuromorphic-chip-against-ai-accelerators/
[13] 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
[14] Y. Liao, “Neural Networks in Hardware: A Survey,” Department of Computer Science, University of California, Tech. Rep., 2001.
