AI accelerator

An **AI accelerator** is a class of microprocessor[1] or computer system[2] designed as hardware acceleration for artificial intelligence applications, especially artificial neural networks, machine vision and machine learning. Typical applications include algorithms for robotics, internet of things and other data-intensive or sensor-driven tasks.[3] They are often manycore designs and generally focus on low-precision arithmetic, novel dataflow architectures or in-memory computing capability.[4] A number of vendor-specific terms exist for devices in this category, and it is an emerging technology without a dominant design.

Contents

History of AI acceleration

Computer systems have frequently complemented the CPU with special purpose accelerators for specialized tasks, known as coprocessors. Notable application-specific hardware units include video cards for graphics, sound cards, graphics processing units and digital signal processors. As deep learning and artificial intelligence workloads rose in prominence in the 2010s, specialized hardware units were developed or adapted from existing products to accelerate these tasks.

Early attempts

As early as 1993, digital signal processors were used as neural network accelerators e.g. to accelerate optical character recognition software.[5] In the 1990s, there were also attempts to create parallel high-throughput systems for workstations aimed at various applications, including neural network simulations.[6][7][8] FPGA-based accelerators were also first explored in the 1990s for both inference[9] and training.[10] ANNA was a neural net CMOS accelerator developed by Yann LeCun.[11]

Heterogeneous computing

Heterogeneous computing refers to incorporating a number of specialized processors in a single system, or even a single chip, each optimized for a specific type of task. Architectures such as the cell microprocessor[12] have features significantly overlapping with AI accelerators including: support for packed low precision arithmetic, dataflow architecture, and prioritizing 'throughput' over latency. The Cell microprocessor was subsequently applied to a number of tasks[13][14][15] including AI.[16][17][18]

In the 2000s, CPUs also gained increasingly wide SIMD units, driven by video and gaming workloads; as well as support for packed low precision data types.[19]

Use of GPU

Graphics processing units or GPUs are specialized hardware for the manipulation of images and calculation of local image properties. The mathematical basis of neural networks and image manipulation are similar, embarrassingly parallel tasks involving matrices, leading GPUs to become increasingly used for machine learning tasks.[20][21][22] As of 2016, GPUs are popular for AI work, and they continue to evolve in a direction to facilitate deep learning, both for training[23] and inference in devices such as self-driving cars.[24] GPU developers such as Nvidia NVLink are developing additional connective capability for the kind of dataflow workloads AI benefits from.[25] As GPUs have been increasingly applied to AI acceleration, GPU manufacturers have incorporated neural network specific hardware to further accelerate these tasks.[26][27] Tensor cores are intended to speed up the training of neural networks.[27]

Use of FPGAs

Deep learning frameworks are still evolving, making it hard to design custom hardware. Reconfigurable devices such as field-programmable gate arrays (FPGA) make it easier to evolve hardware, frameworks and software alongside each other.[9][10][28]

Microsoft has used FPGA chips to accelerate inference.[29][30] The application of FPGAs to AI acceleration motivated Intel to acquire Altera with the aim of integrating FPGAs in server CPUs, which would be capable of accelerating AI as well as general purpose tasks.[31]

Emergence of dedicated AI accelerator ASICs

While GPUs and FPGAs perform far better than CPUs for AI related tasks, a factor of up to 10 in efficiency[32][33] may be gained with a more specific design, via an application-specific integrated circuit (ASIC). These accelerators employ strategies such as optimized memory use and the use of lower precision arithmetic to accelerate calculation and increase throughput of computation.[34][35] Some adopted low-precision floating-point formats used AI acceleration are half-precision and the bfloat16 floating-point format.[36][37][38][39][40][41][42]

In-memory computing architectures

In June 2017, IBM researchers announced an architecture in contrast to the von Neumann architecture based on in-memory computing and phase-change memory arrays applied to temporal correlation detection, intending to generalize the approach to heterogeneous computing and massively parallel systems.[43] In October 2018, IBM researchers announced an architecture based on in-memory processing and modeled on the human brain's synaptic network to accelerate deep neural networks.[44] The system is based on phase-change memory arrays.[45]

Nomenclature

As of 2016, the field is still in flux and vendors are pushing their own marketing term for what amounts to an "AI accelerator", in the hope that their designs and APIs will become the dominant design. There is no consensus on the boundary between these devices, nor the exact form they will take; however several examples clearly aim to fill this new space, with a fair amount of overlap in capabilities.

In the past when consumer graphics accelerators emerged, the industry eventually adopted Nvidia's self-assigned term, "the GPU",[46] as the collective noun for "graphics accelerators", which had taken many forms before settling on an overall pipeline implementing a model presented by Direct3D.

Examples

Stand alone products

- Google Tensor processing unit is an accelerator specifically designed by Google for its TensorFlow framework, which is extensively used for convolutional neural networks. It focuses on a high volume of 8-bit precision arithmetic. The initial first generation from 2015 focused on inference, while the second generation announced in May 2017 increased capability for neural network training also. The third-generation TPU was announced on 8 May 2018. On July 2018 the Edge TPU was announced. Edge TPU is Google's purpose-built ASIC chip designed to run its TensorFlow Lite machine learning (ML) models at the edge.[47]
- Adapteva epiphany is a many-core coprocessor featuring a network on a chip scratchpad memory model, suitable for a dataflow programming model, which should be suitable for many machine learning tasks.
- Intel Nervana NNP (Neural Network Processor) (a.k.a. "Lake Crest"), which Intel claims is the first commercially available chip with a purpose built architecture for deep learning. Facebook was a partner in the design process.[48][49]
- Movidius Myriad 2 is a many-core VLIW AI accelerator complemented with video fixed function units.
- Mobileye's EyeQ is a processor specialized for vision processing for self-driving cars[50]
- NM500 is the latest as of 2016 in a series of accelerator chips for radial basis function neural nets from General Vision.[51]

GPU based products

- Nvidia Tesla is Nvidia's line of GPU derived products marketed for GPGPU and AI tasks.
  - Nvidia Volta is a microarchitecture which augments the Graphics processing unit with additional 'tensor units' targeted specifically at accelerating calculations for neural networks[52]
  - Nvidia GeForce 20 series is the first series based on the Turing microarchitecture and features built in "Tensor Cores".[53]
  - Nvidia DGX-1 is a Nvidia workstation/server product which incorporates Nvidia brand GPUs for GPGPU tasks including machine learning.[54]
  - Nvidia Tegra Xavier SoC features their Deep Learning Accelerator (DLA) and Programmable Vision Accelerator (PVA).[55]
- Radeon Instinct is AMD's line of GPU derived products for AI acceleration.[56]
- Qualcomm's Adreno GPUs since the Snapdragon 820 released in March 2015 using their Qualcomm Snapdragon Neural Processing Engine SDK.[57]
- NEC SX-Aurora TSUBASA is NEC's product line for AI applications and machine learning.[58][59]

AI accelerating co-processors

- Qualcomm's Hexagon DSPs since the Snapdragon 820 released in March 2015 using their Qualcomm Snapdragon Neural Processing Engine SDK.[57]
  - Qualcomm's Snapdragon 855 contains their 4th generation on-device AI engine, including a dedicated Tensor Accelerator.
- Cadence's Tensilica IP is a family of neural network processor and neural network-optimized digital signal processor IP core. Such as the Tensilica Vision C5 DSP released in May 2017 and Tensilica Vision Q6 DSP released in April 2018.[60][61] The Tensilica DNA 100 Processor was announced in September 2018.[62]
- Imagination Technologies' PowerVR 2NX NNA (Neural Net Accelerator) is an IP core fromlicensed for integration into chips, first announced September 2017.[63] On December 2018 PowerVR 3NX NNA was announced.[64]
- Apple's Neural Engine is an AI accelerator core within Apple-designed processors. The Apple A11 Bionic SoC[65] released on September 2017 featured a dual core Neural Engine. The Apple A12 Bionic SoC released on September 2018 featured an octa core Neural Engine.
- Cambricon Technologies's Machine Learning Unit (MLU) family of neural processors such as the MLU-100 and MLU-200.[66]
- HiSilicon's Neural Processing Unit is a neural network accelerator within HiSilicon's Kirin SoCs. The Kirin 970[67] with a NPU from Cambricon Technologies was released in October, 2017. The Kirin 980 with a dual core NPU from Cambricon Technologies was released in October, 2018.
- Google's Pixel Visual Core (PVC) is a fully programmable Image, Vision and AI processor for mobile devices. First featured in the Google Pixel 2 released in October, 2017.
- Arm's ML Processor is dedicated IP for neural network model inferencing acceleration. First announced as Project Trillium in January 2018.[68]
- CEVA's NeuPro family of AI processors. The NP500, NP1000, NP2000 and NP4000 were first announced on January 2018. Each containing one programmable vector DSP and one hardwired implementation of 8-bit or 16-bit neural network layers supporting neural nets with performances ranging from 2 TOPS thru 12.5 TOPS.[69]
- Universal Multifunction Accelerator (UMA) by Manjeera Digital Systems in Hyderabad is an accelerator in a proprietary architecture based on Middle Stratum Operations.[70][71][72]

Research and unreleased products

- In December 2017 Tesla Motors confirmed a rumour that it is developing an AI chip for autonomous driving. Jim Keller worked on this project between at least early 2016 and early 2018.[73]
- MIT Eyeriss is an accelerator design aimed explicitly at convolutional neural networks, using a scratchpad memory and network-on-chip architecture.[74]
- Georgia Tech has designed a neuro-inspired processor for performing online reinforcement learning for ultra-low power robotics. It employs mixed-signal design techniques to reduce the operating power.[75]
- Nullhop is an accelerator designed at the Institute of Neuroinformatics of ETH Zürich and University of Zürich based on sparse representation of feature maps. The second generation of the architecture is commercialized by the university spin-off Synthara Technologies.[76][77]
- Kalray is an accelerator for convolutional neural nets.[78]
- SpiNNaker is a many-core design specialized for simulating a large neural network.
- Graphcore IPU is a graph-based AI accelerator.[79]
- DPU, by Wave Computing, a dataflow architecture[80]
- STMicroelectronics at the start of 2017 presented a demonstrator SoC manufactured in a 28 nm process containing a deep CNN accelerator.[81]
- TrueNorth is a manycore design based on spiking neurons rather than traditional arithmetic.[82][83]
- Intel Loihi is an experimental neuromorphic chip.[84]
- BrainChip (https://www.brainchipinc.com) in September 2017 introduced a commercial PCI Express card with a Xilinx Kintex Ultrascale FPGA running neuromorphic neural cores applying pattern recognition on 600 video images per second using 16 watts of power.[85]
- IIT Madras is designing a spiking neuron accelerator for big-data analytics.[86]
- Several memristor-based AI accelerators have been proposed which leverage in-memory computing capability of memristor.[4]
- AlphaICs is designing an agent-based coprocessor called Real AI Processor (RAP) to enable perception and decision making in a chip.[87]

Potential applications

- Autonomous vehicles: Nvidia has targeted their Drive PX-series boards at this space.[88]
- Military robots
- Agricultural robots, for example pesticide-free weed control.[89]
- Voice control, e.g. in mobile phones, a target for Qualcomm Zeroth.[90]
- Machine translation
- Unmanned aerial vehicles, e.g. navigation systems, e.g. the Movidius Myriad 2 has been demonstrated successfully guiding autonomous drones.[91]
- Industrial robots, increasing the range of tasks that can be automated, by adding adaptability to variable situations.
- Health care, to assist with diagnoses
- Search engines, increasing the energy efficiency of data centers and ability to use increasingly advanced queries.
- Natural language processing

See also

- Cognitive computer
- Neuromorphic computing
- Physical neural network
- Hardware acceleration

References

1. "Intel unveils Movidius Compute Stick USB AI Accelerator" (https://www.v3.co.uk/v3-uk/news/3014293/intel-unveils-movidius-compute-stick-usb-ai-accelerator). 2017-07-21.
2. "Inspurs unveils GX4 AI Accelerator" (https://insidehpc.com/2017/06/inspurs-unveils-gx4-ai-accelerator/). 2017-06-21.
3. "Google Developing AI Processors" (http://www.eetimes.com/document.asp?doc_id=1329715).Google using its own AI accelerators.
4. "A Survey of ReRAM-based Architectures for Processing-in-memory and Neural Networks (https://www.academia.edu/36504841/A_Survey_of_ReRAM-based_Architectures_for_Processing-in-memory_and_Neural_Networks)", S. Mittal, Machine Learning and Knowledge Extraction, 2018
5. "convolutional neural network demo from 1993 featuring DSP32 accelerator" (https://www.youtube.com/watch?v=FwFduRA_L6Q).
6. "design of a connectionist network supercomputer" (http://people.eecs.berkeley.edu/~krste/papers/cns-injs1993.ps).
7. "The end of general purpose computers (not)" (https://www.youtube.com/watch?v=VtJthbiiTBQ).This presentation covers a past attempt at neural net accelerators, notes the similarity to the modern SLI GPGPU processor setup, and argues that general purpose vector accelerators are the way forward (in relation to RISC-V hwacha project. Argues that NN's are just dense and sparse matrices, one of several recurring algorithms)
8. Ramacher, U.; Raab, W.; Hachmann, J.A.U.; Beichter, J.; Bruls, N.; Wesseling, M.; Sicheneder, E.; Glass, J.; Wurz, A.; Manner, R. (1995). *Proceedings of 9th International Parallel Processing Symposium*. pp. 774–781. CiteSeerX 10.1.1.27.6410 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.27.6410). doi:10.1109/IPPS.1995.395862 (https://doi.org/10.1109%2FIPPS.1995.395862). ISBN 978-0-8186-7074-9.
9. "Space Efficient Neural Net Implementation" (https://www.researchgate.net/publication/2318589).
10. "A Generic Building Block for Hopfield Neural Networks with On-Chip Learning" (https://pdfs.semanticscholar.org/63fd/66ff9edb7b5342e4835286d4a2b22e1f2c04.pdf) (PDF). 1996.
11. Application of the ANNA Neural Network Chip to High-Speed Character Recognition (http://yann.lecun.com/exdb/publis/pdf/saeckinger-92.pdf)
12. "Synergistic Processing in Cell's Multicore Architecture" (https://www.semanticscholar.org/paper/Synergistic-Processing-in-Cell-s-Multicore-Archite-Gschwind-Hofstee/9f2a6fc20fb292a5d33eb6bd930e1de9d527ee6b). 2006.
13. De Fabritiis, G. (2007). "Performance of Cell processor for biomolecular simulations". *Computer Physics Communications*. **176** (11–12): 660–664. arXiv:physics/0611201 (https://arxiv.org/abs/physics/0611201). doi:10.1016/j.cpc.2007.02.107 (https://doi.org/10.1016%2Fj.cpc.2007.02.107).
14. "Video Processing and Retrieval on Cell architecture". CiteSeerX 10.1.1.138.5133 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.138.5133).
15. Benthin, Carsten; Wald, Ingo; Scherbaum, Michael; Friedrich, Heiko (2006). *2006 IEEE Symposium on Interactive Ray Tracing*. pp. 15–23. CiteSeerX 10.1.1.67.8982 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.67.8982). doi:10.1109/RT.2006.280210 (https://doi.org/10.1109%2FRT.2006.280210). ISBN 978-1-4244-0693-7.
16. "Development of an artificial neural network on a heterogeneous multicore architecture to predict a successful weight loss in obese individuals" (https://www.teco.edu/~scholz/papers/ScholzDiploma.pdf) (PDF).
17. Kwon, Bomjun; Choi, Taiho; Chung, Heejin; Kim, Geonho (2008). *2008 5th IEEE Consumer Communications and Networking Conference*. pp. 1030–1034. doi:10.1109/ccnc08.2007.235 (https://doi.org/10.1109%2Fccnc08.2007.235). ISBN 978-1-4244-1457-4.
18. Duan, Rubing; Strey, Alfred (2008). *Euro-Par 2008 – Parallel Processing*. Lecture Notes in Computer Science. **5168**. pp. 665–675. doi:10.1007/978-3-540-85451-7_71 (https://doi.org/10.1007%2F978-3-540-85451-7_71). ISBN 978-3-540-85450-0.
19. "Improving the performance of video with AVX" (https://software.intel.com/en-us/articles/improving-the-compute-performance-of-video-processing-software-using-avx-advanced-vector-extensions-instructions). 2012-02-08.
20. "microsoft research/pixel shaders/MNIST" (https://hal.inria.fr/inria-00112631/document).
21. "how the gpu came to be used for general computation" (http://igoro.com/archive/how-gpu-came-to-be-used-for-general-computation/).
22. "imagenet classification with deep convolutional neural networks" (https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf) (PDF).
23. "nvidia driving the development of deep learning" (http://insidehpc.com/2016/05/nvidia-driving-the-development-of-deep-learning/). 2016-05-17.
24. "nvidia introduces supercomputer for self driving cars" (http://gas2.org/2016/01/06/nvidia-introduces-supercomputer-for-self-driving-cars/). 2016-01-06.
25. "how nvlink will enable faster easier multi GPU computing" (https://devblogs.nvidia.com/parallelforall/how-nvlink-will-enable-faster-easier-multi-gpu-computing/). 2014-11-14.
26. "A Survey on Optimized Implementation of Deep Learning Models on the NVIDIA Jetson Platform (https://www.researchgate.net/publication/329802520_A_Survey_on_Optimized_Implementation_of_Deep_Learning_Models_on_the_NVIDIA_Jetson_Platform)", 2019
27. Harris, Mark (May 11, 2017). "CUDA 9 Features Revealed: Volta, Cooperative Groups and More" (https://devblogs.nvidia.com/parallelforall/cuda-9-features-revealed/). Retrieved August 12, 2017.
28. "FPGA Based Deep Learning Accelerators Take on ASICs" (http://www.nextplatform.com/2016/08/23/fpga-based-deep-learning-accelerators-take-asics/). *The Next Platform*. 2016-08-23. Retrieved 2016-09-07.
29. "microsoft extends fpga reach from bing to deep learning" (http://www.nextplatform.com/2015/08/27/microsoft-extends-fpga-reach-from-bing-to-deep-learning/). 2015-08-27.
30. Chung, Eric; Strauss, Karin; Fowers, Jeremy; Kim, Joo-Young; Ruwase, Olatunji; Ovtcharov, Kalin (2015-02-23). "Accelerating Deep Convolutional Neural Networks Using Specialized Hardware" (http://research.microsoft.com/pubs/240715/CNN%20Whitepaper.pdf) (PDF). *Microsoft Research*.
31. "A Survey of FPGA-based Accelerators for Convolutional Neural Networks (https://www.academia.edu/37491583/A_Survey_of_FPGA-based_Accelerators_for_Convolutional_Neural_Networks)", Mittal et al., NCAA, 2018
32. "Google boosts machine learning with its Tensor Processing Unit" (http://techreport.com/news/30155/google-boosts-machine-learning-with-its-tensor-processing-unit). 2016-05-19. Retrieved 2016-09-13.
33. "Chip could bring deep learning to mobile devices" (https://www.sciencedaily.com/releases/2016/02/160203134840.htm). *www.sciencedaily.com*. 2016-02-03. Retrieved 2016-09-13.
34. "Deep Learning with Limited Numerical Precision" (http://jmlr.org/proceedings/papers/v37/gupta15.pdf) (PDF).
35. Rastegari, Mohammad; Ordonez, Vicente; Redmon, Joseph; Farhadi, Ali (2016). "XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks". arXiv:1603.05279 (https://arxiv.org/abs/1603.05279) [cs.CV (https://arxiv.org/archive/cs.CV)].
36. Khari Johnson (2018-05-23). "Intel unveils Nervana Neural Net L-1000 for accelerated AI training" (https://venturebeat.com/2018/05/23/intel-unveils-nervana-neural-net-l-1000-for-accelerated-ai-training/). *VentureBeat*. Retrieved 2018-05-23. "...Intel will be extending bfloat16 support across our AI product lines, including Intel Xeon processors and Intel FPGAs."
37. Michael Feldman (2018-05-23). "Intel Lays Out New Roadmap for AI Portfolio" (https://www.top500.org/news/intel-lays-out-new-roadmap-for-ai-portfolio/). *TOP500 Supercomputer Sites*. Retrieved 2018-05-23. "Intel plans to support this format across all their AI products, including the Xeon and FPGA lines"
38. Lucian Armasu (2018-05-23). "Intel To Launch Spring Crest, Its First Neural Network Processor, In 2019" (https://www.tomshardware.com/news/intel-neural-network-processor-lake-crest,37105.html). *Tom's Hardware*. Retrieved 2018-05-23. "Intel said that the NNP-L1000 would also support bfloat16, a numerical format that's being adopted by all the ML industry players for neural networks. The company will also support bfloat16 in its FPGAs, Xeons, and other ML products. The Nervana NNP-L1000 is scheduled for release in 2019."
39. "Available TensorFlow Ops | Cloud TPU | Google Cloud" (https://cloud.google.com/tpu/docs/tensorflow-ops). *Google Cloud*. Retrieved 2018-05-23. "This page lists the TensorFlow Python APIs and graph operators available on Cloud TPU."
40. Elmar Haußmann (2018-04-26). "Comparing Google's TPUv2 against Nvidia's V100 on ResNet-50" (https://blog.riseml.com/comparing-google-tpuv2-against-nvidia-v100-on-resnet-50-c2bbb6a51e5e). *RiseML Blog*. Retrieved 2018-05-23. "For the Cloud TPU, Google recommended we use the bfloat16 implementation from the official TPU repository with TensorFlow 1.7.0. Both the TPU and GPU implementations make use of mixed-precision computation on the respective architecture and store most tensors with half-precision."
41. Tensorflow Authors (2018-02-28). "ResNet-50 using BFloat16 on TPU" (https://github.com/tensorflow/tpu/tree/master/models/experimental/resnet_bfloat16). *Google*. Retrieved 2018-05-23.
42. Joshua V. Dillon, Ian Langmore, Dustin Tran, Eugene Brevdo, Srinivas Vasudevan, Dave Moore, Brian Patton, Alex Alemi, Matt Hoffman, Rif A. Saurous (2017-11-28). TensorFlow Distributions (Report). arXiv:1711.10604 (https://arxiv.org/abs/1711.10604). Bibcode:2017arXiv171110604D (http://adsabs.harvard.edu/abs/2017arXiv171110604D). Accessed 2018-05-23. "All operations in TensorFlow Distributions are numerically stable across half, single, and double floating-point precisions (as TensorFlow dtypes: tf.bfloat16 (truncated floating point), tf.float16, tf.float32, tf.float64). Class constructors have a validate_args flag for numerical asserts"
43. Abu Sebastian; Tomas Tuma; Nikolaos Papandreou; Manuel Le Gallo; Lukas Kull; Thomas Parnell; Evangelos Eleftheriou (2017). "Temporal correlation detection using computational phase-change memory". *Nature Communications*. **8**. arXiv:1706.00511 (https://arxiv.org/abs/1706.00511). doi:10.1038/s41467-017-01481-9 (https://doi.org/10.1038%2Fs41467-017-01481-9).
44. "A new brain-inspired architecture could improve how computers handle data and advance AI" (https://phys.org/news/2018-10-brain-inspired-architecture-advance-ai.html). *American Institute of Physics*. 2018-10-03. Retrieved 2018-10-05.
45. Carlos Ríos; Nathan Youngblood; Zengguang Cheng; Manuel Le Gallo; Wolfram H.P. Pernice; C David Wright; Abu Sebastian; Harish Bhaskaran (2018). "In-memory computing on a photonic platform". arXiv:1801.06228 (https://arxiv.org/abs/1801.06228) [cs.ET (https://arxiv.org/archive/cs.ET)].
46. "NVIDIA launches the World's First Graphics Processing Unit, the GeForce 256" (http://www.nvidia.com/object/IO_20020111_5424.html).
47. Kundu, Kishalaya (2018-07-26). "Google Announces Edge TPU, Cloud IoT Edge at Cloud Next 2018" (https://beebom.com/google-announces-edge-tpu-cloud-iot-edge-at-cloud-next-2018/). *Beebom*. Retrieved 2019-02-02.
48. Kampman, Jeff (17 October 2017). "Intel unveils purpose-built Neural Network Processor for deep learning" (https://techreport.com/news/32704/intel-unveils-purpose-built-neural-network-processor-for-deep-learning). Tech Report. Retrieved 18 October 2017.
49. "Intel Nervana Neural Network Processors (NNP) Redefine AI Silicon" (https://www.intelnervana.com/intel-nervana-neural-network-processors-nnp-redefine-ai-silicon/?_ga=2.62312428.1380200850.1508486032-2008757629.1504021982). Retrieved 20 October 2017.
50. "The Evolution of EyeQ" (https://www.mobileye.com/our-technology/evolution-eyeq-chip/).
51. "NM500, Neuromorphic chip with 576 neurons" (http://www.general-vision.com/hardware/nm500/).
52. "Nvidia goes beyond the GPU for AI with Volta" (https://www.forbes.com/sites/tiriasresearch/2017/05/10/nvidia-goes-beyond-the-gpu-for-ai-with-volta/).
53. "The NVIDIA Turing GPU Architecture Deep Dive: Prelude to GeForce RTX" (https://www.anandtech.com/show/13282/nvidia-turing-architecture-deep-dive/6). AnandTech.
54. "nvidia dgx-1" (https://images.nvidia.com/content/technologies/deep-learning/pdf/61681-DB2-Launch-Datasheet-Deep-Learning-Letter-WEB.pdf) (PDF).
55. Frumusanu, Andrei. "Investigating NVIDIA's Jetson AGX: A Look at Xavier and Its Carmel Cores" (https://www.anandtech.com/show/13584/nvidia-xavier-agx-hands-on-carmel-and-more). *www.anandtech.com*. Retrieved 2019-02-02.
56. Smith, Ryan (12 December 2016). "AMD Announces Radeon Instinct: GPU Accelerators for Deep Learning, Coming in 2017" (http://www.anandtech.com/show/10905/amd-announces-radeon-instinct-deep-learning-2017). Anandtech. Retrieved 12 December 2016.
57. "On-Device AI with Qualcomm Snapdragon Neural Processing Engine SDK" (https://developer.qualcomm.com/blog/device-ai-qualcomm-snapdragon-neural-processing-engine-sdk). *Qualcomm Developer Network*. Retrieved 2019-02-02.
58. "NEC SX-Aurora TSUBASA" (https://linustechtips.com/main/topic/851286-nec-sx-aurora-tsubasa/).
59. "AI Acceleration-with-NEC's New Vector Computer" (https://www.finanznachrichten.de/nachrichten-2018-10/45108841-ai-acceleration-with-nec-s-new-vector-computer-011.htm).
60. "Cadence Unveils Industry's First Neural Network DSP IP for Automotive, Surveillance, Drone and Mobile Markets" (https://www.cadence.com/content/cadence-www/global/en_US/home/company/newsroom/press-releases/pr/2017/cadence-unveils-industrys-first-neural-network-dsp-ip-for-automo.html).
61. Frumusanu, Andrei. "Cadence Announces Tensilica Vision Q6 DSP" (https://www.anandtech.com/show/12633/cadence-announces-tensilica-vision-q6-dsp). *www.anandtech.com*. Retrieved 2019-02-02.
62. Frumusanu, Andrei. "Cadence Announces The Tensilica DNA 100 IP: Bigger Artificial Intelligence" (https://www.anandtech.com/show/13377/cadence-announces-tensilica-dna-100-a-bigger-nn-ip). *www.anandtech.com*. Retrieved 2019-02-02.
63. "The highest performance neural network inference accelerator" (https://www.imgtec.com/powervr-2nx-neural-network-accelerator/).
64. Oh, Nate. "Imagination Announces PowerVR Series9XTP, Series9XMP, and Series9XEP GPU Cores" (https://www.anandtech.com/show/13671/imagination-announces-powervr-series9x-p-gpus). *www.anandtech.com*. Retrieved 2019-02-02.
65. "The iPhone X's new neural engine exemplifies Apple's approach to AI" (https://www.theverge.com/2017/9/13/16300464/apple-iphone-x-ai-neural-engine). *The Verge*. Retrieved 2017-09-23.
66. Cutress, Ian. "Cambricon, Makers of Huawei's Kirin NPU IP, Build A Big AI Chip and PCIe Card" (https://www.anandtech.com/show/12815/cambricon-makers-of-huaweis-kirin-npu-ip-build-a-big-ai-chip-and-pcie-card). *www.anandtech.com*. Retrieved 2019-02-02.
67. "HUAWEI Reveals the Future of Mobile AI at IFA 2017" (http://consumer.huawei.com/en/press/news/2017/ifa2017-kirin970/).
68. Cutress, Ian. "Hot Chips 2018: Arm's Machine Learning Core Live Blog" (https://www.anandtech.com/show/13253/hot-chips-2018-arm-machine-learning-core-live-blog). *www.anandtech.com*. Retrieved 2019-02-02.
69. "A Family of AI Processors for Deep Learning at the Edge" (https://www.ceva-dsp.com/product/ceva-neupro/).
70. Manjeera Digital System, UMA. "Universal Multifunction Accelerator" (http://manmanjeerads.com/technology.htm). *Manjeera Digital Systems*. Retrieved 28 June 2018.
71. Manjeera Digital Systems, Universal Multifunction Accelerator. "Revolutionise Processing" (http://www.newindianexpress.com/cities/hyderabad/2018/may/11/hyderabad-start-up-looks-to-revolutionise-processor-technology-1813121.html). *Indian Express*. Retrieved 28 June 2018.
72. AI Chip, UMA (10 May 2018). "AI Chip from Hyderabad" (https://telanganatoday.com/first-indian-chip-for-ai-applications-developed-in-hyderabad) (News Paper). Telangana Today. Retrieved 28 June 2018.
73. Lambert, Fred (December 8, 2017). "Elon Musk confirms that Tesla is working on its own new AI chip led by Jim Keller" (https://electrek.co/2017/12/08/elon-musk-tesla-new-ai-chip-jim-keller/).
74. Chen, Yu-Hsin; Krishna, Tushar; Emer, Joel; Sze, Vivienne (2016). "Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks" (http://www.mit.edu/~sze/eyeriss.html). *IEEE International Solid-State Circuits Conference, ISSCC 2016, Digest of Technical Papers*. pp. 262–263.
75. "Mixed-signal Processing Powers Bio-mimetic CMOS Chip to Enable Neural Learning in Autonomous Micro-Robots | IEN" (http://www.ien.gatech.edu/news/mixed-signal-processing-powers-bio-mimetic-cmos-chip-enable-neural-learning-autonomous-micro).
76. Aimar, Alessandro; et al. (2017). "NullHop: A Flexible Convolutional Neural Network Accelerator Based on Sparse Representations of Feature Maps". arXiv:1706.01406 (https://arxiv.org/abs/1706.01406) [cs.CV (https://arxiv.org/archive/cs.CV)].
77. "Synthara Technologies" (https://www.synthara.ch/).
78. "kalray MPPA" (http://www.hotchips.org/wp-content/uploads/hc_archives/hc27/HC27.24-Monday-Epub/HC27.24.20-Multimedia-Epub/HC27.24.240-DeDinechin-KalrayMMPA-Kalray-v4.pdf) (PDF).
79. "Graphcore Technology" (https://www.graphcore.ai/technology).
80. "Wave Computing's DPU architecture" (https://www.nextplatform.com/2017/08/23/first-depth-view-wave-computings-dpu-architecture-systems/). 2017-08-23.
81. "A 2.9 TOPS/W Deep Convolutional Neural Network SoC in FD-SOI 28nm for Intelligent Embedded Systems" (https://reconfigdeeplearning.files.wordpress.com/2017/02/isscc2017-14-1visuals.pdf) (PDF).
82. "yann lecun on IBM truenorth" (https://www.facebook.com/yann.lecun/posts/10152184295832143).argues that spiking neurons have never produced leading quality results, and that 8-16 bit precision is optimal, pushes the competing 'neuflow' design
83. "IBM cracks open new era of neuromorphic computing" (http://www.extremetech.com/extreme/187612-ibm-cracks-open-a-new-era-of-computing-with-brain-like-chip-4096-cores-1-million-neurons-5-4-billion-transistors). "TrueNorth is incredibly efficient: The chip consumes just 72 milliwatts at max load, which equates to around 400 billion synaptic operations per second per watt — or about 176,000 times more efficient than a modern CPU running the same brain-like workload, or 769 times more efficient than other state-of-the-art neuromorphic approaches"
84. "Intel's New Self-Learning Chip Promises to Accelerate Artificial Intelligence" (https://newsroom.intel.com/editorials/intels-new-self-learning-chip-promises-accelerate-artificial-intelligence/).
85. "BrainChip Accelerator" (http://www.brainchipinc.com/products/civil-surveillance-solutions/brainchip-accelerator).
86. "India preps RISC-V Processors - Shakti targets servers, IoT, analytics" (http://www.eetimes.com/document.asp?doc_id=1328790&page_number=2). "The Shakti project now includes plans for at least six microprocessor designs as well as associated fabrics and an accelerator chip"
87. "AlphaICs" (http://www.alphaics.ai/).
88. "drive px" (http://www.nvidia.com/object/drive-px.html).
89. "design of a machine vision system for weed control" (http://abe.ufl.edu/wlee/publications/icame96.pdf) (PDF).
90. "qualcomm research brings server class machine learning to every data devices" (https://www.qualcomm.com/news/onq/2015/10/01/qualcomm-research-brings-server-class-machine-learning-everyday-devices-making). October 2015.
91. "movidius powers worlds most intelligent drone" (https://www.siliconrepublic.com/machines/movidius-dji-drone). 2016-03-16.

External links

- http://www.nextplatform.com/2016/04/05/nvidia-puts-accelerator-metal-pascal/
- http://eyeriss.mit.edu
- http://www.alphaics.ai/