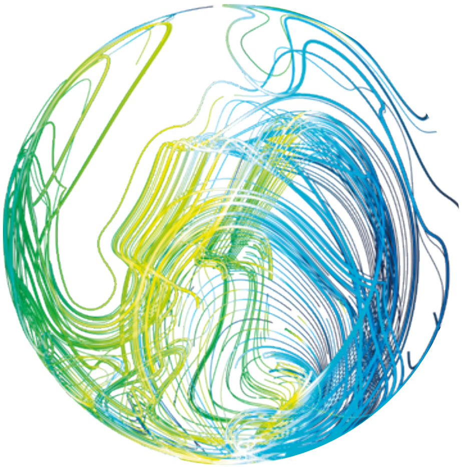
Deloitte.



Hitting the accelerator: the next generation of machine-learning chips

Deloitte Global predicts that by the end of 2018, over 25 percent of all chips used to accelerate machine learning in the data center will be FPGAs (field programmable gate arrays) and ASICs (application-specific integrated circuits). These new kinds of chips should increase dramatically the use of ML, enabling applications to consume less power and at the same time become more responsive, flexible and capable, which is likely to expand the addressable market.

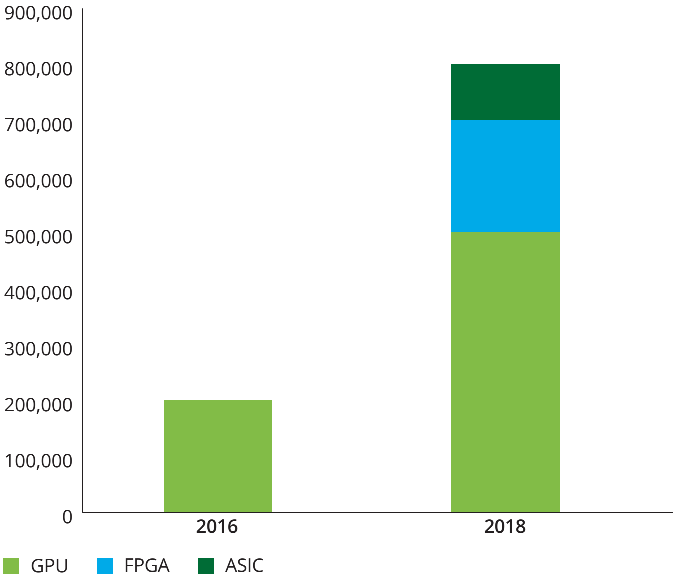
This is a dramatic shift; in 2016, almost all ML involving the artificial neural network (ANN) approach used a combination of standard GPU chips (graphics processing units) and CPU chips (central processing units) in large data centers.

We estimate that about 100,000 to 200,000 GPU chips were sold for ML in 2016.[65] We predict the market for GPUs will be larger in 2018, at over half a million chips. There will also be over 200,000 FPGA and 100,000 ASIC chips sold for ML applications. The dollar value of each kind of chip is different, so Deloitte Global is making a prediction not on the monetary value of each portion of the ML chip market, but merely on the number of chips. One analyst has forecast that the 2022 market for ML accelerator products will be in the admittedly wide range of $4.5 billion to $9.1 billion.[66]

Deloitte Global expects that GPUs and CPUs in 2018 will still be the largest part of the ML chip market, measured by chip units, and will still be growing. But the new kinds of chips may dramatically increase the use of ML, enabling applications to use less power and at the same time become more responsive, flexible and capable, which is likely to expand the addressable market, as can be seen in Figure 13, with chip sales for ML tasks predicted to at least quadruple in only two years.

Growth should be able to continue beyond 2018. The current leader in GPUs for ML in the data center has publicly stated that it anticipates the total available market (TAM) for both training and inference acceleration to be $26 billion by 2020,[67] which would be many millions of chips of various kinds per year, though probably not tens of millions.

Figure 13. Annual minimum sales of ML chips in global data centers (units)



Source: Deloitte Global estimates, 2017, based on publicly available information. See endnotes for full methodology.

Artificial neural networks, machine learning and the associated hardware

Serial processing binary computers, whether made up of tubes or transistors, are capable of many tasks, but there are other computational challenges for which there are better alternatives. Image recognition, for example, is very hard to do using rule-based programming. Inspired by how biological neurons work, scientists in 1943 created a computational model for an artificial neural network.[68]

In subsequent decades, researchers built ANNs in many forms. In the early days, they mostly ran on mainframes and minicomputers, but by the 1980s, they were largely implemented on machines powered by PC-style CPUs.

It is critical to note that ANNs are not exactly like biological neurons; they are merely inspired by certain aspects of how real neurons work. Some chips do work more in the way neurons do, as discussed below, but they should not be confused with ANNs themselves.

In 2009, researchers found that GPUs, the chips that were so good at rendering computer game scenes using highly parallel processing at a reasonable price and great speed, were also very good at machine learning via ANNs. Originally these chips were called not GPUs but “graphic accelerators,” and they had an architecture fundamentally different from that of CPUs, with many independent small processing cores. They excelled at parallel processing tasks, while CPUs were better at serial processing. Not every computing problem was done better in parallel, but rendering video game graphics faster was inherently improved with parallel processing.

For ML tasks, GPUs (with some CPUs in the mix) were found to be between 10 and 100 times as fast as CPU-only ML solutions, depending on the exact task.[69] This acceleration was revolutionary and dramatically expanded the market for ML hardware and solutions. CPUs were still used, but the virtues of GPUs increased the size of the pie, with ML being used much more widely than it had been before 2009.

Machine learning using ANNs can be broadly broken into two primary tasks: training and inference. For example, when trying to develop an image-recognition system to recognize cats, the system is shown hundreds or thousands or millions of images. Some of the images are labeled by humans as “cats,” and others as “not cats.” As the computer is exposed to these labeled images, it generates an algorithm that allows it to detect the presence of a cat in a new image; this is the training portion. Once that algorithm has been created, however, the actual cat-recognition process for a given image is performed through a process called inference. Up until 2016, both training and inference were performed on the same hardware: racks of GPUs and CPUs, usually in large data centers. Although some of the first examples of ML using non-CPU and non-GPU chips were for inference rather than training, it is unclear what the mix will be going forward. As of now, some companies are using FPGAs and ASICs for inference only, and others, for both training and inference.

[65] In calendar-year 2016, the major supplier of GPUs to data centers for machine learning was Nvidia. The company has a January 31 year-end, so its fiscal 2017 figures are close to calendar 2016 but off by one month. In the quarters prior to Q4 fiscal 2016, NVDA reported GPU sales to data centers of about $75 million per quarter, which was largely for the high-performance computing (HPC, aka “supercomputers”) market. In fiscal 2017, they publicly reported total sales to data centers of $830 million, and so for calendar 2016, we estimate that GPU sales to data centers were about $800 million. Of that, about $300 million was HPC and other, and therefore GPU sales for ML were about $500 million. When Nvidia sells to the data center for ML, they don’t just sell a GPU chip but also an entire package, which usually includes memory and cooling. With a very high-end GPU chip, these boards are expensive. The exact price, pricing mix by product (in 2016, Nvidia sold all P4, P40 and K40 boards for ML applications, all at different price points) and discounts for volume purchase are not public information, but Deloitte Global (based largely on the resale market) believes the average price in 2016 was between $2,500 and $5,000 per board, which suggests the total number of GPU chips for ML in the data center (assuming one chip per board) was in the range of 100,000 to 200,000 for calendar 2016.

[66] How large is the deep learning data center market? Seeking Alpha, 26 June 2017: https://seekingalpha.com/article/4084040-large-deep-learning-data-center-market?page=2.

[67] Investor Day Presentation 2017, Nvidia, 10 May 2017: http://files.shareholder.com/downloads/AMDA-1XAJD4/4376258341x0x942329/A2FCD200-F141-4A26-8E5D-215F6F2171E0/NVIDIA_Investor_Day_2017.pdf, slides 24 and 25.

[68] A logical calculus of the ideas immanent in nervous activity, Springer, as accessed on 3 November 2017: https://link.springer.com/article/10.1007%2FBF02478259.

What follows is an overview of the various kinds of chips that are likely to be used for ML in data centers and even outside them.

ML-optimized GPUs: From 2009 to 2016, the GPUs that were sold to data centers and used for ML were essentially the same chips and boards used for computer gaming. As mentioned above, these gaming GPUs, although not designed for ML, were by orders of magnitude better at running ANNs than the CPUs of that era. In 2018, the makers of GPUs are releasing special versions of GPUs that are optimized for ML; for example, Nvidia's Volta architecture is said to be 12 times better at deep-learning training and six times better at inference than the preceding Pascal architecture. We expect these new chips to sell hundreds of thousands of units per year.

ML-optimized CPUs: Meanwhile, we are also seeing CPU companies introduce variants of their standard chips that are specialized for ML. Intel's recent Knights Mill chip offers ML performance[70] four times superior to that of data center CPUs that were not optimized for ML.

ML-optimized FPGAs: FPGA chips are integrated circuits that can be dynamically programmed for applications or functionality. They are currently manufactured by a number of companies in many configurations. The market for these devices represents millions of chips annually and over $4 billion in sales in 2016.[71] A paper published at the beginning of 2017[72] showed that for a subset of deep neural network tasks, FPGAs were able to outperform GPUs by varying degrees in speed and/or power efficiency. Some tasks were only 50 percent faster, while others were 440 percent faster, and some were only slightly faster but 130 percent better in terms of performance per watt (heat often becomes a limiting factor, and so performance per watt can sometimes be critical).

Yet FPGAs are being used well beyond academic circles. One large cloud provider, Microsoft, has said it is using FPGA chips for inference purposes as part of its hosted ML offering, and has publicly disclosed that as of summer 2017, "hundreds of thousands" of the chips were already being used.[73] Amazon Web Services (AWS) and Baidu are also said to be using FPGAs in their data centers for machine learning purposes,[74] although chip volumes are unknown. And of course it matters that Intel, the world's largest maker of CPUs for data centers, purchased the second-largest FPGA company with its 2016 acquisition of Altera. Total 2018 FPGA chip volume for ML would be a minimum of 200,000. The figure is almost certainly going to be higher, but by exactly how much is difficult to predict.

ML-optimized ASICs: ASICs are single-purpose chips and are made by many large manufacturers. Industry revenues are about $15 billion in 2017. CPUs and GPUs are fairly general-purpose chips, manufactured by the millions each year. CPUs and GPUs tend to be fairly expensive on a per-chip basis, and they often use a lot of power. FPGAs tend to be used only when hundreds of chips are needed. They are fast to market, usually better at power efficiency than GPUs and CPUs, and often a good choice if neither the time, budget or volume requirements for an ASIC nor the ability to reprogram the chip dynamically is needed.

In the history of integrated circuit technology, it has been common for certain tasks to be done first on general-purpose processors, then on FPGAs and then on custom ASICs. ASICs often have the best performance, power and therefore efficiency, but designing an ASIC and getting it to the point of manufacturability can cost tens of millions of dollars. Therefore, ASICs are usually used only when a market application has reached a certain critical size at which the advantages of the ASIC solution become compelling. In terms of ML and ANNs, various ASICs seem set to play important roles in 2018 and beyond.

One example of an ASIC designed for machine learning is the Tensor Processing Unit (TPU – see below), and others, such as the Nervana chip from Intel, are expected to be available by the beginning of 2018.[75] Fujitsu also plans to launch a chip called Deep Learning Unit (DLU) that will be available in 2018.[76] Unit volumes are difficult to predict; they could be in the tens of thousands or hundreds of thousands.

One large cloud provider, Microsoft, has said it is using FPGA chips for inference purposes as part of its hosted ML offering, and has publicly disclosed that as of summer 2017, "hundreds of thousands" of the chips were already being used.

[69] A tale of two cities: GPU Computing and Machine Learning, Dr. Xiaowen Chu, Department of Computer Science, Hong Kong Baptist University, as accessed on 3 November 2017: https://www.comp.hkbu.edu.hk/~chxw/ppts/hkust_chxw.pptx.

[70] Intel Xeon Phi Knights Mill for Machine Learning, Serve The Home, 21 August 2017: https://www.servethehome.com/intel-knights-mill-for-machine-learning/.

[71] And the winner of the best FPGA of 2016 is, EE Times, 6 March 2017: https://www.eetimes.com/author.asp?doc_id=1331443.

[72] Can FPGAs Beat GPUs in Accelerating Next-Generation Deep Neural Networks? ACM, 22 February 2017: http://dl.acm.org/citation.cfm?id=3021740.

[73] Microsoft unveils Brainwave, a system for running super-fast AI, Venture Beat, 22 August 2017: https://venturebeat.com/2017/08/22/microsoft-unveils-brainwave-a-system-for-running-super-fast-ai/.

[74] Baidu adopts Xilinx to accelerate machine learning applications in the data center, Xilinx, 17 October 2016: https://www.xilinx.com/news/press/2016/baidu-adopts-xilinx-to-accelerate-machine-learning-applications-in-the-data-center.html; As semiconductors' focus on AI grows, Xilinx could be a winner, Barron's, 23 August 2017: http://www.barrons.com/articles/as-semiconductors-focus-on-ai-grows-xilinx-could-be-a-winner-1503506665.

[75] Intel to ship new Nervana Neural Network Processor by end of 2017, TechCrunch, 17 October 2017: https://techcrunch.com/2017/10/17/intel-to-ship-new-nervana-neural-network-processor-by-end-of-2017/.

[76] Fujitsu's 'DLU' AI processor promises 10x the performance of 'The Competition,' Tom's Hardware, 19 July 2017: http://www.tomshardware.com/news/fujitsu-dlu-deep-learning-processor,35037.html.

[77] An open-source software library for Machine Intelligence, Tensor Flow, as accessed on 3 November 2017: https://www.tensorflow.org/.

[78] Tensor processing unit, Wikipedia, as accessed on 3 November 2017: https://en.wikipedia.org/wiki/Tensor_processing_unit.

TPUs: Google has developed a series of ASICs for machine learning, called TPUs. TPUs are optimized to run the open-source ML software TensorFlow (also developed by Google).[77] The first-generation TPU was announced in 2016, and the second-generation chip was introduced in May 2017.[78] As is common in the evolution of chip markets, debate continues about the relative performance of TPUs compared with GPUs. But in tests performed in Google's own data centers on inference tasks, TPUs have shown performance gains over certain GPUs, just as GPUs did compared with CPUs, where the gain was 10 to 50 times. Critically, even when the absolute performance advantage of the TPU over the GPU for a task was not as large, the performance per watt was always considerably superior. For power-constrained applications such as the large server farms where companies do most of their inference, this is likely to be important. The first-generation TPUs appear to have been used only for inference, not for training, although the second-generation devices may be able to do training as well. It is unclear at this time whether the relative performance advantage of TPUs over GPUs for certain inference tasks will be comparable for training tasks. Actual chip volumes have not been disclosed by Google, but estimates suggest around 100,000 units seems likely.[79]

Lower-power ML accelerator chips: Over time, Deloitte Global believes that other chips, optimized for machine learning at even lower power, will see increased deployment in non-data-center markets, specifically for sensor networks, Internet of Things devices and gateways, and medical technologies. Deloitte Global predicts there will be over half a billion mobile chips running ML inferences on smartphones, tablets and other devices in 2018.[80] One example outside the smartphone world would be the Movidius chip from Intel, which is specifically used for ML acceleration for vision processing.[81]

When looking at Internet of Things applications that are mobile or not connected to the power grid, power requirements need to be measured in milliwatts at most. By contrast, GPUs for machine learning frequently consume over 250 watts per chip, and even TPUs require around 75 watts. Inside a data center, on a rack of cards cooled with fans connected to massive power lines and in a building with an air-conditioning plant capable of cooling kilowatts of heat, the energy consumed and the heat produced are difficult challenges.

For applications such as sensor networks, power draw would likely need to be below 10 milliwatts. Equally, any ML chip that needs to work inside the human body cannot use much power or produce much heat; its power consumption may need to be measured in microwatts or less. While there are commercial chips in smartphones and other mobile devices that are at the high end of the range, there is nothing that works at the low end. That is unlikely to change in 2018, but over the next year or two, there may be significant progress in low-power ML chips; in early 2017, one university laboratory produced an ML chip that consumes only 288 microwatts.[82]

Other ML accelerators: There are a number of companies looking to develop their own ASICs (or new computing architectures) that will be optimized for artificial intelligence and machine learning. At the time of writing, these companies have received hundreds of millions of dollars in funding, and have written papers claiming their solutions will be better than the current GPU/CPU solutions, especially for low-precision arithmetic. None seems to be selling these solutions in commercial volumes yet, so the impact in 2018 is unlikely to be large. But in 2019 and beyond, these devices may capture some part of the market.

Neuromorphic chips: There is an additional class of chips that do not fit into the conventional classifications above. IBM's True North chip is one of a class called neuromorphic chips, which are potentially capable of accelerating ML tasks and being very energy-efficient.[83] At this time, there do not appear to be any commercial-scale uses of these chips in data centers, although the US military has stated that it is exploring the technology for ML applications.[84] It is difficult to predict neuromorphic chip volumes for 2018, but it seems likely to be below 100,000 units and possibly even below 10,000.

[79] Will ASIC chips become the next big thing in AI? Forbes, 4 August 2017: https://www.forbes.com/sites/moorinsights/2017/08/04/will-asic-chips-become-the-next-big-thing-in-ai/#19ebf7ed11d9.

[80] The Deloitte Global 2017 prediction was for 300 million smartphones to ship with onboard ML chips. These tend to be found only in the higher-end phones ($400 and up), and we assume that another 300 million to 400 million smartphones will have onboard ML chips in 2018. But not all 2017 phones will still be in use by the end of 2018 (breakage, upgrades, etc.), so we assume that there will be at least 500 million, likely more than 600 million, phones with these chips by the end of the year. https://www2.deloitte.com/content/dam/Deloitte/xe/Documents/technology-media-telecommunications/predicitons2017/ME-Predictions-2017-Brains-at-the-edge.pdf.

[81] Intel unveils neural compute engine in Movidius Myriad X VPU to unleash AI at the edge, Intel, 28 August 2017: https://newsroom.intel.com/news/intel-unveils-neural-compute-engine-movidius-myriad-x-vpu-unleash-ai-edge/.

[82] Speck-size computers: now with deep learning, IEEE, 28 March 2017: http://spectrum.ieee.org/semiconductors/processors/specksize-computers-now-with-deep-learning.

[83] True North, Wikipedia, as accessed on 3 November 2017: https://en.wikipedia.org/wiki/TrueNorth; IBM Finds Killer App for TrueNorth Neuromorphic Chip, Top 500, 24 September, 2016: https://www.top500.org/news/ibm-finds-killer-app-for-truenorth-neuromorphic-chip/.

[84] US Military Sees Future in Neuromorphic Computing, The Next Platform, 26 June, 2017: https://www.nextplatform.com/2017/06/26/u-s-military-sees-future-neuromorphic-computing/.

[85] Blogging the Periodic Table: Aluminum, Slate, 30 July 2010: http://www.slate.com/articles/health_and_science/elements/features/2010/blogging_the_periodic_table/aluminum_it_used_to_be_more_precious_than_gold.html.

[86] It's Elemental: Aluminum, Jefferson Lab, Accessed on 6 November 2017: https://education.jlab.org/itselemental/ele013.html.

The bottom line

When it comes to machine learning, big changes to the machine (in this case, the chips) are likely to cause big changes in the industry. After moving from CPU-only to CPU-plus-GPU solutions, the industry exploded in usefulness and ubiquity; using chips that are 10 to 50 times better will do that. If the various FPGA and ASIC solutions offer similar order-of-magnitude improvements in processing speed, efficiency, price or any combination thereof, a similar explosion in utility and adoption seems probable.

That said, there are certain tasks that ML is good at and others where it has its limitations. These new chips are likely to allow companies to perform a given level of ML using less power at less cost. But on their own, they are not likely to give *better* or more accurate results.

If the only accomplishment of these new chips is to make machine learning 10, 100 or 1,000 times less expensive, that could be more revolutionary than it seems. Famously, when aluminum was first purified and produced, it was so expensive that it was used instead of gold on the Washington Monument, and a French emperor had cutlery made out of the new and almost priceless material while less-important guests had to make do with solid-gold utensils.[85] In the 1880s, new processes for refining aluminum from bauxite ore were invented, and the price dropped by orders of magnitude.[86] Nothing about the metal itself had changed; it was the same, but much cheaper. As a result, it became not an object of ostentatious display but an extremely useful and much-used material in many industries. A change in the price of machine learning seems likely to produce similarly disruptive effects.

However, it isn't just the chips that are getting better. Deloitte Global has identified what we believe are important vectors of progress that promise to unlock more intensive use of ML in the enterprise. Some of these advances make ML easier, cheaper or faster (or a combination of all three). This will have the effect of expanding the market for ML, just as Economics 101 would predict. Other advances enable applications in new areas, which will also expand the market.

The key improvements are found in the companion prediction *Machine learning: things are getting intense* and include (in addition to the chip improvements we discuss above) automating data science, reducing the need for training data, explaining the results of ML better and deploying local ML. Taken together, these improvements will double the intensity with which enterprises are using ML by the end of 2018, and they promise over the long term to make it a fully mainstream technology, one that will enable new applications across industries where companies have limited talent, infrastructure or data to train the models.

Deloitte.

Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited, a UK private company limited by guarantee (“DTTL”), its network of member firms, and their related entities. DTTL and each of its member firms are legally separate and independent entities. DTTL (also referred to as “Deloitte Global”) does not provide services to clients. Please see www.deloitte.com/about for a more detailed description of DTTL and its member firms.

Deloitte provides audit, consulting, financial advisory, risk management, tax and related services to public and private clients spanning multiple industries. Deloitte serves four out of five Fortune Global 500® companies through a globally connected network of member firms in more than 150 countries bringing world-class capabilities, insights, and high-quality service to address clients’ most complex business challenges. To learn more about how Deloitte’s approximately 225,000 professionals make an impact that matters, please connect with us on Facebook, LinkedIn, or Twitter.

This communication contains general information only, and none of Deloitte Touche Tohmatsu Limited, its member firms, or their related entities (collectively, the “Deloitte Network”) is, by means of this communication, rendering professional advice or services. Before making any decision or taking any action that might affect your finances or your business, you should consult a qualified professional adviser. No entity in the Deloitte Network shall be responsible for any loss whatsoever sustained by any person who relies on this communication.

Designed and produced by The Creative Studio at Deloitte, London. J14285