AI Requires Many Approaches

By Linley Gwennap
Principal Analyst

July 2018

The Linley Group

www.linleygroup.com

AI Requires Many Approaches

By Linley Gwennap, Principal Analyst, The Linley Group

Artificial intelligence (AI) is being applied to a wide range of problems, so no single processor can support them all. General-purpose processors are the easiest to program, but graphics processors (GPUs) offer greater performance for most workloads. Custom AI architectures optimized for specific workloads can be deployed in FPGAs or ASICs. Different approaches are needed for the data center, for autonomous driving, and for consumer/IoT applications. This paper describes and compares these approaches. Intel sponsored the creation of this white paper, but the opinions and analysis are those of the author. Trademark names are used in an editorial fashion and are the property of their respective owners.

Artificial intelligence (AI) is a transformative technology that is being applied to wide range of applications. Web-based services use it to improve voice interfaces (e.g. Alexa), search results, face recognition, language translation, and spam filtering, among others. Autonomous vehicles use AI for image recognition and path planning. Industries as diverse as healthcare, financial investment, oil exploration, manufacturing, and retail find that this approach can improve essential processes.

Although researchers have worked on AI for decades, it has made great strides just in the past five years due in part to the emergence of deep neural networks (DNNs). Instead of manually building a software program to, say, identify human faces in photographs, researchers simply feed a series of labeled photographs into the network, which then configures itself to recognize the faces. This process is called training. Once the network is trained, it can process new photographs to identify the faces. This deployment mode is known as inference.

Most AI processing occurs in data centers. For example, when you use Alexa or Siri, your request is transmitted to the cloud, where a server interprets your speech and then generates the proper response, which it sends back to your device. But moving data to and from the cloud delays the response, and the entire process fails if your network connection is unavailable. For these and other reasons (e.g., security, privacy), AI processing is beginning to appear in edge and client devices such as smartphones, home gateways, the Internet of Things (IoT), and autonomous cars.

AI processing was initially developed on general-purpose processors, as these devices are ubiquitous and easy to program. In time, researchers discovered that graphics chips (GPUs) could deliver better performance and power efficiency. Recently, the first hardware architectures have emerged that are designed specifically for neural networks. These processors are typically optimized for inference and must also be designed to meet the performance, power, cost, and integration requirements of their target markets. For this reason, different types of AI chips are needed for data centers, autonomous cars, and various consumer and IoT systems.

General-Purpose Processors

General-purpose processors power PCs, servers, and supercomputers, running a wide variety of software programs. As the name implies, deep neural networks are more complex than the simpler networks that preceded them, so researchers turned to the most powerful servers to run these new networks. The most popular processors for these computers are Intel's Xeon products. Data-center operators can use these flexible processors to run standard programs or DNNs, keeping up with changes in demand.

Intel anticipated the popularity of DNNs by developing a technology called AVX-512. This approach, an extension of the company's original AVX design, enables each CPU to process 512 bits of data at a time, eight times as much as a standard CPU. These 512 bits can be allocated in a number of ways, for example, as 16 single-precision floating-point values or 64 eight-bit integers. This flexibility is important because DNN training typically requires 32-bit floating point (FP32), while inference can use smaller integer values (e.g. INT8) to improve throughput and power efficiency.

Intel's newest Xeon Scalable processors, based on the Skylake-SP design, implement up to 28 CPU cores with AVX-512 capability. At a peak speed of 3.2GHz, the top-end 8180 model generates just over 2 trillion FP32 operations or 8 trillion INT8 operations per second at a power rating of 205W (TDP). The chip also includes more than 45MB of cache memory that can hold the key parameters for a large DNN, and it features six channels of DDR4 DRAM to feed the rest of the network. The company also offers a number of other models that address lower price and power levels.

Graphics Processors

To create high-quality images, modern GPUs rely on programmable engines known as shader cores that have powerful floating-point engines. Because these cores focus on computation and omit many features of general-purpose architectures, they are smaller than CPUs, therefore more of them fit onto a chip. The shader cores can be programmed for tasks other than graphics, although their programming model and software-development tools are more difficult to use than those of a standard CPU.

To get around this problem, AI researchers created tools for developing DNNs. These "frameworks" provide a high-level interface for building networks, specifying the number of nodes and connections. Once specified, the network can run on any hardware through an API and a set of drivers. In this fashion, DNN developers don't have to program the underlying hardware. Currently, TensorFlow, Caffe/Caffe2, and Torch/Pytorch are among the most popular DNN frameworks.

Nvidia is the leading vendor of GPU chips for AI applications. It provides drivers for all popular frameworks as well as other tools to facilitate DNN development and training. The company's most recent high-end GPU is the Tesla V100 ("Volta") design, which features 320 shader cores (known as warp cores) that can process a total of 14 trillion FP32 operations per second. The V100 has only 6MB of cache memory; the small on-chip memory can hamper performance on large networks, although the chip uses high-

bandwidth memory to partially offset this shortcoming. The total power for the massive chip and its HBM is 250W TDP.

Data-Center Accelerators

Although GPUs provide an efficient solution for DNNs, particularly when floating-point math is required, they were originally designed for graphics and not for AI. Now that the demand for AI processing is large (we estimate the revenue from AI processors exceeded $2 billion in 2017, nearly double that of the previous year), companies are developing custom chips that are optimized for DNN acceleration. This approach should deliver the best performance per watt, although the outcome varies depending on the implementation.

Because the development of these AI accelerators is still in its infancy, companies are applying many different architectures to this problem. Some common features are starting to emerge, however. For example, products intended for inference can eliminate floating-point support and focus on smaller integer values. An integer MAC unit uses about half the die area of a comparable floating-point unit; furthermore, an 8-bit MAC uses one-fourth the area of a 32-bit MAC. Thus, shifting from FP32 to INT8 can reduce the computation area by as much as 8x. The smaller area saves power as well. Studies have shown the difference in DNN accuracy between FP32 and INT8 is small for inference. Researchers are evaluating other numeric formats to optimize the tradeoff between accuracy and die area/power.

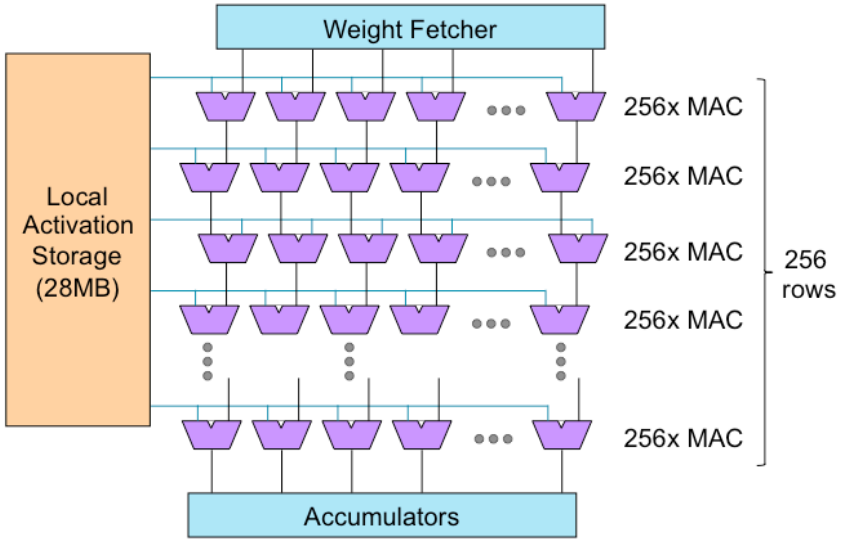

Figure 1. Systolic MAC array. In this structure, weights flow down the array while activation data flows across, enabling a large number of multiply-accumulate (MAC) calculations per cycle. This example is from Google's TPU.

Another common feature is the systolic MAC array. As Figure 1 shows, a systolic array can compute up to 256 MACs at once, then the data flows immediately to the next row for the next computation. In contrast, a CPU or GPU core decodes an instruction, loads 512 bits from a register, computes up to 16 MACs, then stores the results into a register. A systolic array eliminates this extra instruction decoding and register access, simpli-

fying the chip design. Finally, AI processors often include special hardware to compute common DNN functions such as activations, normalization, and pooling; these functions often require several instructions on a standard CPU or GPU.

In 2016, Intel acquired Nervana Systems, a startup developing a new architecture optimized for DNNs. Intel expects the Nervana design to support both training and inference in the data center. This design uses on-chip HBM for massive memory bandwidth. The company hasn't announced further details of the first Nervana products, but we expect they will outperform CPUs and GPUs in performance per watt when they enter production next year.

FPGAs

While custom architectures show great promise for AI acceleration, researchers continue to develop new DNN algorithms, new activation functions, and new data formats. Implementing a custom architecture in an ASIC or other fixed design can take years, and once the design is released, the hardware can't be changed. Thus, if the designers guess wrong or fail to anticipate new algorithms, their product may fail in the market.

One solution is the use of FPGAs, flexible chips that can implement new designs on the fly. Instead of carving a new architecture into an ASIC and waiting months for the silicon to come back from the fab, a company can burn the same architecture into an FPGA and be up and running in minutes. Furthermore, if the company wants to fix a bug or make a minor improvement to the architecture, it can update the same FPGA chip, again in a matter of minutes, instead of having to buy a new chip.

FPGAs are well suited to neural networks because, in addition to their configurable logic, they include numerous MAC units known as DSP blocks. For example, Intel's Stratix 10 FPGAs have as many as 5,760 DSP blocks that can generate 23 trillion integer operations per second. (Stratix 10 uses 18-bit integers that are more accurate than 8-bit integers.) The FPGA can also be configured to produce 9 trillion FP32 operations per second. Intel has measured the FPGA's power, including memory, at 125W when processing these DNN calculations.

Microsoft created an architecture called Brainwave that it instantiates in Intel Stratix FPGAs, accelerating inference in its data centers. The company makes changes on a weekly basis, rolling them out to thousands of FPGAs at once. Using this iterative approach, it developed and tested a custom 9-bit floating-point format (FP9) before settling on an 8-bit format (FP8) that doubles performance over standard INT8. To meet the needs of its data centers, Microsoft also optimized Brainwave for low latency, maintaining high efficiency even with small numbers of requests. These customizations demonstrate the advantages of using FPGAs for DNNs.

Autonomous Driving

Every major carmaker is developing autonomous-driving technology that will eventually allow passengers to safely ride in a car with no human driver. This technology can also apply to trucks, airplanes, and drones. Some vehicles available

today have semiautonomous technology, but they require driver supervision. By 2020, we expect mass production of so-called Level 4 vehicles that don't require a human driver within a limited geographical area (e.g., a city) and good weather.

This capability requires sophisticated AI algorithms and processors. Autonomous vehicles must analyze in real time data from cameras and other sensors, identifying landmarks (e.g., lane markers, signage, buildings) as well as potential hazards (e.g., vehicles, pedestrians, road debris). DNNs excel at such image recognition. Once the environment around the vehicle is understood, other sophisticated algorithms must determine the optimal path toward the destination while ensuring safety. A combination of DNNs and traditional software will likely handle path planning.

Despite their large bodies and powerful engines, automobiles have more power and size constraints than a data center. Customers don't want the driving system to fill the entire trunk, and automakers limit the power drain to about 40W to avoid reducing engine performance and mileage. New processors must meet these constraints while delivering the high performance required for Level 4 and Level 5 autonomous driving.

In 2017, Intel acquired Mobileye, the leading supplier of processors for Level 2 and 3 driving (ADAS). Mobileye specializes in vision processing, that is, the ability to analyze camera images and identify objects. The company's current EyeQ4 processor can generate 2 trillion integer operations per second while using only 3W, a tiny fraction of the power of a high-end data-center processor. Intel recommends combining two of these processors with a low-power Xeon chip that handles path planning; this combination can fit into a car's 40W power budget.

Consumer/IoT Products

AI processing applies to many consumer products. Voice assistants such as Alexa typically appear today in smart speakers, but we are starting to see voice-enabled televisions, ovens, smartwatches, lighting, and even toilets. While users may accept a cloud-processing delay for general queries, a voice command to adjust the lighting or change the channel should be processed locally for a fast response. A neural network that can recognize a small number of basic voice commands requires relatively little memory and processing power.

Most IP security cameras send video to the cloud for processing, but this approach consumes lots of bandwidth, particularly for high-resolution cameras. Some "smart cameras" instead integrate a processor that analyzes the images and sends them to the cloud only when something unusual appears. Drones can use similar processors to analyze their surroundings and avoid obstacles or track the user across a ski slope or skateboard park.

Processors for consumer applications must deliver the desired AI performance at the lowest possible cost. Many applications also require small chips that sip as little power as possible. Even if a microcontroller has enough performance to handle basic voice recognition, an AI chip can perform the same function using a fraction of the power. This power savings is critical in smartwatches and other battery-powered devices.

To meet these consumer needs, some companies have developed low-power AI processors. For example, Intel offers Myriad 2, a vision-processing chip developed by its Movidius subsidiary. This chip is rated at 1 trillion operations per second, yet it consumes about 1W (typical). It features 12 custom SHAVE cores that can process 128 bits per cycle as well as a new Neural Compute Engine that contributes most of the DNN acceleration. To simplify software development, Myriad 2 is also available in a USB stick that plugs directly into a PC.

Conclusion

Since the advent of deep neural networks, AI processing is sweeping through the tech industry and spilling into many other fields, improving both business processes and consumer lifestyles by automating simple tasks. To fully instantiate these improvements, AI processing cannot be trapped in the cloud; instead, it is moving closer to the end user. Efficiently implementing these AI tasks requires new types of processors for cars, drones, security cameras, consumer appliances, wearables, medical devices, and other IoT systems. Data centers will continue to play an important role, particularly for training DNNs, so cloud-service providers also seek better processors for their AI needs.

No single processor can satisfy this broad range of applications, which have different performance, power, and cost requirements. To better address the AI market, Intel has developed a diverse set of products. Its Xeon Scalable processors with AVX-512 can quickly switch from standard data-center workloads to AI processing. Its new Nervana processor is fully optimized for accelerating DNN calculations. Customers designing their own AI accelerator can implement it in Stratix FPGAs that can be easily reprogrammed for changing algorithms. Intel's Mobileye platform supports self-driving cars, and its Myriad 2 processor meets the low-power requirements of consumer and battery-powered devices. No other vendor supplies such a broad range of AI processors.

Linley Gwennap is principal analyst at The Linley Group and editor-in-chief of Microprocessor Report. *The Linley Group offers the most comprehensive analysis of microprocessor and SoC design. We analyze not only the business strategy but also the internal technology. Our in-depth reports also cover topics including deep-learning processors, Ethernet chips, and processor IP cores. For more information, see our web site at www.linleygroup.com.*