Taking powerful, efficient inference to the edge

Executive Summary

Artificial intelligence (AI) is now the key driving force behind advances in information technology, big data and the internet of things (IoT). It is a technology that is developing at a rapid pace, particularly when it comes to the field of deep learning. Researchers are continually creating new variants of deep learning that expand the capabilities of machine learning. But building systems that are able to use these deep learning models to analyze real-world data presents a major challenge. Silicon cost and energy consumption are major hurdles to teams keen to put deep learning into edge devices, as well as data centers.

Conventional digital technology is unable to handle the high compute requirements of AI models that need to run in real-time on cost-effective, low-power hardware. This is a problem that needs to be addressed by a change in technology--a change to a hardware platform that employs analog compute-in-memory (CIM). This is the technology pioneered by Mythic, a new generation of AI technology that can propel the next wave of applications that harness deep learning and its many variants.

The challenge of inferencing

Machine learning has already transformed the world of the data center. Rapid progress in deep learning has demonstrated that there are many applications that can take advantage of the technology. However, the energy demands of these applications present challenges to development teams, especially if they want to cost-effectively deploy the technology in systems that require low power consumption.

The key to deep learning’s recent successes was a breakthrough in training technology. Researchers found ways to make the training of multi-layered neural networks a practical reality, although training is likely to continue to rely on high-performance hardware such as graphics processing units (GPUs) and similar accelerators for the foreseeable future.

In principle, when deep neural networks (DNNs) are deployed into the field and used for inferencing – recognizing and classifying data based on a model generated by training – the computational requirements are less severe than those of the training process. One saving lies in the ability to use lower-resolution integer arithmetic instead of floating-point arithmetic. During training, floating-point arithmetic is needed to support the fine-tuning of the network parameters, such as weights. When used in the field for inferencing, many DNNs work well with integer arithmetic based on 8-bit data. The systems still must run billions of computations through the trained network to identify patterns in incoming data, but the compute demand is orders of magnitude lower than that needed for training.

Because they share similar computation hardware requirements, many users in data centers have employed the same basic architectures for inferencing as they have for training. But the silicon and energy costs of these platforms have become burdensome, and these costs limit the expansion of deep learning into new markets. Users want to be able to deploy inferencing into real-time applications, such as machine vision and audio processing. The high frame and sample rates of these systems call for much higher levels of computational throughput, compared to systems that are used to identify and classify objects in static images.

For example, the relatively simple AlexNet image-recognition DNN (often used as a benchmark for hardware) calls for 725 million multiply-accumulate (MAC) operations to handle a single image with a resolution of only 224 x 224 pixels. At 30 frames per second, this calls for a processing throughput of 22 billion MACs per second. And this number excludes the other common operations required in inferencing such as activations, in which the output of each neuron is scaled and compared against a threshold.

Conventional digital processing pipelines can support the throughput required of real-time AI, but the platforms come with high costs in terms of silicon real estate and energy. The energy density of the server blades and compute accelerators is reaching levels where operators have been pushed to employ increasingly complex and costly heat-containment and management strategies. Some have begun to investigate the use of liquid cooling, even to the point of immersing the systems in fluids with high thermal conductivity.

In edge systems, the problems of energy consumption and silicon cost are even more dramatic. For many potential applications, there is no hardware available that supports the AI requirements at the required silicon cost or energy budget. This is why initial deployments of AI for the edge have largely focused on the use of high-bandwidth internet connections to offload much of the deep-learning work to remote cloud servers. In the short term, this strategy proved moderately successful for some target markets, such as in-home smart speakers. These are devices that vendors can expect to be provided with access to a permanent internet connection. Even then, concerns have grown over the privacy of conversations and other user data transferred to the cloud for processing. User acceptance is likely to be much higher for systems that can perform the recognition of speech or images local to the user and are not forced to upload data to the cloud.

For many mobile and real-time industrial systems, pushing inferencing to the cloud is impractical. The systems may not have access to sufficiently high-bandwidth communications links, and even where they are available, the latency incurred by passing data over long distances is a major challenge to responsiveness. Real-time systems cannot afford to have image-recognition functions miss deadlines. They need to have these services provided locally.

Key to the high-volume deployment of inferencing systems at the edge is the ability to reduce the energy demand, as well as the amount of silicon area needed. The metric that matters most is "inferences per watt per dollar" (I/W/$): how many image frames can be processed in a given amount of time and at what energy and capital cost.

Digital solutions hit the wall

At first glance, architectures such as those of GPUs appear to be well suited to handling deep-learning workloads. The bulk of the arithmetic operations performed in DNN layers are based on MAC operations. As GPUs were developed to perform matrix transformations to map 3D models onto 2D surfaces, their pipelines are optimized for MACs. The weakness of the GPU and similar digital architectures lies in the amount of energy needed to perform operations like these for DNNs.

There are two root causes of excess energy consumption that result from the use of today's digital architectures. One lies in the arithmetic process of multiplication. For high throughput, multipliers implemented in digital systems need to employ a large number of logic gates in parallel: their number grows exponentially with data resolution. The use of approximation helps with this for inferencing workloads. A common technique is to reduce the floating-point values for weights used during training to 8-bit values for use during inference. This reduction comes with comparatively little cost to accuracy. In an attempt to cut further the digital processing burden, users have tried coarser levels of approximation for neuron connections that are deemed less important. These optimizations can result in the processing being reduced to using binary or ternary calculations, which also lowers the computational overhead but can lead to a loss of model accuracy in real-world applications.

A bigger problem with existing approaches to DNN processing based on digital logic is that they cannot make full use of the spatial and temporal locality of information that is inherently available in these calculations. Today's digital architectures, such as GPUs, try to make use of locality by caching input data close to the processor. These data elements may be reused many times over when calculating the effect of each neuronal weight on the inputs. A layer may have a thousand input data elements and output a thousand more. But they pass through a process that can involve accessing a million or more individual weights. Each input data element may be reused a thousand times during the process, but an individual weight used just once. As a result, caching the weight data does not make sense in the context of a conventional digital processor, so there is little to stop the million weight data accesses to remote memory from incurring a huge energy burden.

One technique to reduce the number of weight accesses is to prune the neural network. This is a process invoked after training in which paths between neurons that are deemed to have a limited effect on model behavior are removed. The need to employ both approximation and pruning to make DNN workloads run at acceptable performance greatly increases the complexity of development of the neural networks. The quality of results also falls. Errors tend to increase as pruning and approximation are used more aggressively. Development complexity increases because of the need for model tuning to recover from the pruning and the use of binary or ternary approximation to fit within the power and silicon-cost target of the end application without compromising model behavior. The development team is faced with having to re-analyze models extensively, because the pruned versions do not match those trained on a workstation or in the cloud at full floating-point resolution.

An analog compute-in-memory solution

What is required for inferencing at the edge and for low-energy data-center inferencing is an architecture that allows for the seamless transfer of cloud-trained models, but which does not suffer from the high energy cost of memory transfers and high-resolution computation that is common to today’s digital implementations. Mythic’s compute-in-memory solution uses the memory itself to perform computations and so avoids the need to continually move weight data around the system. This slashes the energy per MAC from 10pJ in a typical digital edge inferencing implementation that holds the large weight arrays in DRAM to as low as 0.5pJ. Across the billions of MAC computations required for video-rate inferencing at moderate resolution, the resulting energy savings are dramatic.

In Mythic's architecture, each access to weight memory is essentially energy-free. The main contribution to the processing energy comes from the MAC operation itself, which is implemented simply by passing data through the memory. As a result, this process is much more efficient than that found in logic-hungry digital datapaths. The core of the Mythic implementation is a flash memory technology that is widely used in microcontrollers and other mass-production embedded systems.

The readout circuitry in a conventional flash memory discretizes these programmed charge values into a 1, 2 or 3-bit value. In practice it is possible, in a flash memory made on a mature process such as 40nm, to store reliably a range of charges that correspond to a digital resolution of 8 bits. A further advantage of the flash memory cell is that it is non-volatile. Once programmed it can store an electrical charge for long periods of time without power, a key advantage for systems that need to run off of a battery. But the memory can be erased and reprogrammed at any time to support new or updated models.

When a charge is programmed into a flash memory device, its electric field has an effect on any signal passing through it. In the Mythic architecture, the flash transistor acts as a variable resistor that reduces the signal level passing to the output. That reduction is proportional to the analog value stored in the memory. This simple effect implements the multiplication stage found in DNN calculations. The accumulation process, in which the output from each of those calculations is summed, is handled by aggregating the output of an entire column of memory cells. Thanks to these two properties, the Mythic architecture can process an entire input vector in a single step rather than iterating at high speed as in a digital processor.

Thanks to its ability to streamline MAC processing by orders of magnitude compared to conventional digital processors, Mythic's memory-based accelerator makes it possible to transfer fully trained neural networks directly to a low-energy inferencing platform without any need for pruning or further approximation. But the analog memory-based core is only part of the solution. Functions such as activations and pooling, which form key parts of any DNN are generally best implemented in digital logic. The Mythic architecture accommodates this by including a single-instruction, multiple-data (SIMD) accelerator unit coupled to a RISC-V processor that coordinates operations and local SRAM to hold temporary data. With these components, the Mythic solution has the ability to run a complete DNN model independently.

A key element of the Mythic solution is scalability. The combination of memory-based accelerator, SIMD engine, SRAM and RISC-V processor forms a tile in an array of DNN engines. All the tiles are linked by a high-speed network-on-chip (NoC) routing mesh to allow for the efficient flow of input, output and intermediate data elements. The array of tiles is managed by an on-chip processor and communicates with the system's host processor over a PCIe interface.

The mesh architecture of the Mythic platform provides the ideal substrate for applications such as machine learning. In contrast to the many applications written for conventional architectures, which revolve around the sequential processing of a single stream of code, AI inferencing is a graph-based application. Graph applications are well suited to dataflow architectures where it is straightforward to assign a different compute element to each node of the graph. When one graph node has completed its work, the data output flows to the next graph node for processing. With its combination of different types of compute functions in a mesh, Mythic implements a highly efficient dataflow architecture.

The dataflow architecture also maximizes inference performance by having many of the compute-in-memory elements operating in parallel, pipelining the image processing by handling neural-network layers in parallel in different tiles of the array. By being built from the ground-up as a dataflow architecture, the Mythic solution minimizes the memory and computational overhead required to manage the dependency graphs that define dataflow computing, and it keeps the application operating at maximum performance. The result is an architecture that delivers inferencing performance at breakthrough silicon cost and power levels, while also supporting a wide range of edge and data-center systems.

A new wave of applications

An architecture that offers much greater energy efficiency for applications that rely on deep learning provides developers with the ability to build a much wider range of smart devices, devices that can operate independently of the cloud when required. Applications range from security cameras to robots and drones that need to operate for long periods away from a fixed power source.

Using a highly efficient architecture such as that developed by Mythic, manufacturers will be able to take the next step up from the compute technology found in systems today. Industrial control provides a large number of potential applications for smarter systems. Many processes, from smelting to oil extraction, suffer from long time constants in the systems they attempt to model. These challenge conventional closed-loop control algorithms. Complex patterns in the incoming sensor data can provide early warning of looming problems. Deep learning provides a mechanism to track these patterns. But the cost and power of traditional compute platforms makes deep learning difficult to implement, especially in environments where it is difficult to provide high-bandwidth communications to data centers. The Mythic platform provides the ability build compact, low-power, self-contained control systems that leverage the power of deep learning.

In robotic systems, such as airborne drones, the Mythic platform enables the use of deep learning for image and sensor processing. The result is a dramatic improvement to the situational awareness of these systems. Armed with the ability to classify objects in its camera input, a drone can gain greater autonomy and reduce the need for operators to stay in constant control of the device. This makes it more feasible to use drones for tasks like inspecting infrastructure such as power grids and pipelines over long distances, and for other over-the-horizon sensing applications.

The Mythic platform inside security cameras makes it possible for them to detect anomalies and problems in their field of view without having to pass every single frame of video to a remote server for inspection. This one change can dramatically reduce network bandwidth and power requirements. It also allows the deployment of wireless security cameras powered only by a battery, so that they can be placed in locations that are out of reach of network and power-supply lines.

When the deep-learning functions enabled by Mythic are not required, the devices can be powered down to save energy and maintain a high level of battery autonomy. The non-volatile nature of the flash memory ensures the millions of weights required for the neural-network layers are preserved and immediately available when required.

The scalability of the Mythic platform makes it easy to tune the implementation to the requirements of the model. Individual tiles or groups of tiles inside the device can be dedicated to different models. For example, in a system intended for use in drones or robots, one group of tiles can be assigned to camera image enhancement to reduce the effect of lighting changes on object recognition. Other tiles can be dedicated to scene segmentation and object tracking, each feeding their results back to the host processor.

In the data-center environment, support for interchip communication between Mythic devices allows the creation of large-scale inferencing engines that are able to handle the complex workloads now encountered in this environment. Compared to existing platforms, Mythic offers significant advantages in terms of cost and energy efficiency. Because the core MAC operations consume a fraction of the power of that required by GPU, FPGA or digital ASIC platforms, the Mythic solution can support extremely high processing densities without the need for advanced cooling or power-supply infrastructure. The result is a platform that scales from the simplest edge sensor node to highly sophisticated cloud based DNN inferencing.

Delivering practical inferencing

Mythic's platform enables the rapid development of deep-learning applications, as it supports models trained using readily available frameworks, such as ONNX and TensorFlow, without requiring that the developer perform complex runtime optimizations. Mythic has adopted a software strategy that will ensure the process of transferring models to the end device is as seamless as possible through the use of tools that convert the trained models produced by a variety of machine-learning frameworks into executables that can be downloaded to a single Mythic product or an array of them. Integration into end applications will be supported by libraries that provide application programming interfaces (APIs) to the software running inside the device, making it easy to employ the Mythic silicon as a coprocessor to conventional microcontroller-based designs.

The software developed for the Mythic platform optimizes the neural network for execution through a simple two-stage process. The first step, using the Mythic Optimization Suite, transforms the trained network into a form that is compatible with analog CIM, including quantization from floating-point representations to 8-bit integer. This suite checks that the quantization will not degrade performance below acceptable levels. The tools also include retraining flows for applications that have strict accuracy requirements or more aggressive performance and power targets or a combination of all three. Quantization-aware and analog-aware retraining builds resiliency into layers that are more sensitive to the lower bit-depths of quantization and to analog noise.

The Mythic Graph Compiler follows and performs automated mapping onto the target array, packing, and support-code generation. The output is a packaged binary that contains everything that the host driver needs to program and control the Mythic device to perform inferencing in a real-time environment. Longer term, Mythic envisions an SDK with a suite of powerful tools to help developers evaluate tradeoffs and identify the best solution within the constraints of power and cost.

As well as integrating easily with today's popular tools, it is important to be able to take advantage of the rapid pace of development in the field of machine learning. As new layer types and network topologies are invented, software and hardware support should be straightforward. The Mythic platform ensures this through the modular design of the software SDK, leveraging a large amount of generic matrix compute capabilities rather than architecture-specific accelerators.

Flexibility is just as important to the hardware architecture. The scalable, tiled nature of the Mythic platform greatly eases development and integration into an end system. During prototyping, it is likely that there will be multiple iterations of model tuning and training to ensure that the core deep-learning engine can deal with real-world problems and is unaffected by problems that may be caused by a skewed training set. For example, if the prototyping stage determines that additional models are needed to handle changes in lighting or environmental conditions, a move to a larger array can easily be made. Similarly, the developers may find opportunities for cost reduction that allow the use of a smaller model and a more appropriate device.

Mythic intends to provide a range of implementations that scale in size and number. The platform will be made available not just in IC form but as accelerator cards that may use multiple devices. The tiled architecture eases the production of derivatives based on market demand and so supports the evolution of deep learning in edge and low-power data-center systems as new applications emerge.

By tackling the core aspects of deep learning with an architecture optimized for low-power operation, Mythic is poised to drive a new wave of smart applications that unleash the power of AI.