Growing AI Diversity and Complexity Demands Flexible Data-Center Accelerators

By Bob Wheeler
Principal Analyst

December 2020

The Linley Group

www.linleygroup.com

Growing AI Diversity and Complexity Demands Flexible Data-Center Accelerators

By Bob Wheeler, Principal Analyst, The Linley Group

AI applications are becoming more diverse, even as models for specific applications rapidly advance. CPUs and GPUs offer the flexibility to handle new models, but they deliver poor throughput or efficiency for real-time inferencing. Purpose-built deep-learning accelerators excel for CNNs but often fare poorly on other model types. SimpleMachines developed a unique "composable computing" architecture that provides both programmability and efficiency. SimpleMachines sponsored this white paper, but the opinions and analysis are those of the author.

Models Grow in Diversity

As artificial intelligence proliferates into new markets, the diversity in applications continues to grow. When processing occurs at aggregation points including data centers and the infrastructure edge, AI systems must be capable of handling a range of model types. This requirement differs from that of endpoint devices, such as a smart camera, that can employ application-specific accelerators for a single function.

Many deep-learning-accelerator (DLA) vendors focus on convolutional neural networks (CNNs), which are commonly used in vision processing. Language processing instead often uses recurrent neural networks (RNNs), such as long short-term memory (LSTM) models, in translation and speech recognition. An alternative to RNNs are attention-based Transformer models including BERT. Multilayer-perceptron (MLP) networks are typically used in recommendation systems.

Despite the availability of proven model types, researchers continue to develop new approaches that improve accuracy or add features. Capsule networks, for example, improve on CNNs by adding spatial relationships. Generative adversarial networks (GANs) use unsupervised learning to train models that generate realistic content. Within a model type, networks are also growing in complexity to improve accuracy. For example, the commonly-cited ResNet-50 CNN has 26 million parameters (weights), whereas the newer YOLOv3 has 62 million. Another measure of complexity is increasing depth, with Inception v4 employing 638 layers compared with 229 for ResNet-50.

Whether for internal workloads or customer workloads, public-cloud providers use multiple model types. Training AI models is a batch-oriented process for which GPUs are popular. Operators can run multiple GPUs in parallel to reduce training time, although some models don't scale well and parallel processing increases software complexity and network demands. When trained models are deployed for real-time inference, batch processing becomes a greater liability, as it increases response time (latency). Both CPUs and GPUs enable inference for many model types, but they become less efficient as batch size decreases.

Purpose-built DLAs are more efficient, but often for only specific types of models, particularly CNNs. In fact, some DLAs for inference handle only specific models, as each model must be optimized for the hardware. Not only does this make the DLA application-specific, but it may also prevent field upgrades to newer models for the same application. Within their limited scope, however, these DLAs can deliver leading performance per watt.

SimpleMachines has developed a new technology that approaches the flexibility of a GPU while delivering the performance of a DLA. It received first silicon of its Mozart inference chip in 4Q20 and is sampling a PCI Express card intended for high-volume servers.

GPUs and DLAs Fall Short

With their massive parallelism, GPUs require batch processing to maximize utilization. Nvidia's A100, based on the Ampere architecture, has 108 shader cores and achieves peak inference throughput with batch sizes greater than 200. For the ResNet-50 v1.0 model, the A100 achieves 38% utilization at a batch size of 211, as Figure 1 shows. At a batch size of eight, which is more appropriate for real-time inferencing, utilization drops precipitously to 14%. Known as *batch=1,* running a single inference at a time delivers minimum latency, but in the absence of parallel operations, utilization falls to a dismal 2.7%. SimpleMachines found that Nvidia's GPUs deliver their lowest utilization on the GNMT RNN, with a T4 card running a batch size of four at only 1.4%.

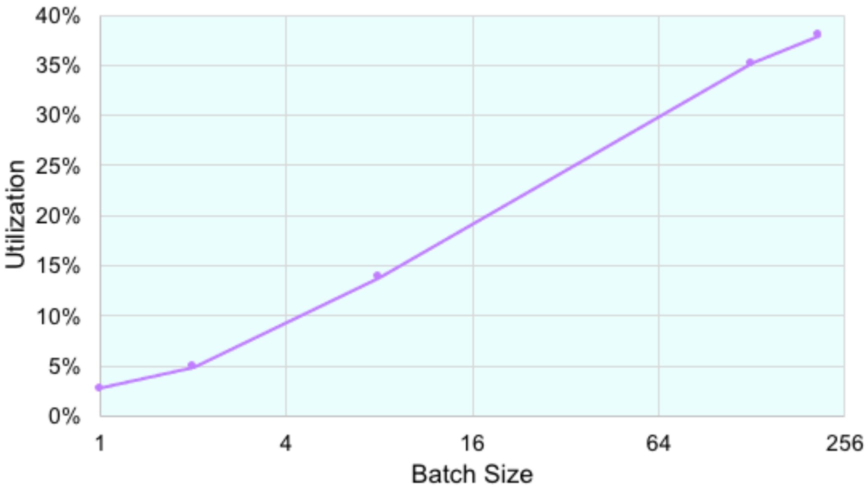

Figure 1. Nvidia A100 ResNet-50 inference utilization versus batch size. GPUs can be quite efficient when batching introduces adequate parallelism, but their efficiency is poor when inferencing with small batch sizes. (Source: The Linley Group)

DLAs implement a wide variety of architectures, so their performance and utilization versus batch size varies widely. The data-center-inference accelerator that achieves the greatest utilization, Qualcomm's AI 100, does so with a batch size of eight (running ResNet-50 v1.0), reaching 50%. Optimized for batch=1, Groq's TSP delivers maximum throughput without batching, but its utilization is only 19%. Still, it delivers about 9x the batch=1 throughput of Nvidia's A100. Clearly, delivering high utilization without batching has proven elusive.

SimpleMachines is applying reconfigurable hardware to AI with the goal of delivering both algorithm adaptability and high utilization. Founded in 2017, the startup is building on a decade of research at the University of Wisconsin-Madison. As a university spin out, it has access to the patents from this research. Professor Karu Sankaralingam is founder, CEO, and CTO of SimpleMachines. Chief architect Greg Wright joined in early 2020 after serving as a senior director of engineering at Qualcomm Research. The company's engineering leadership comes from Qualcomm's server-processor team, which was disbanded in 2018. The startup has raised about $20 million from investors that include Baidu Ventures and IMO Ventures.

Sankaralingam and his fellow researchers bucked the trend of domain-specific architectures (DSAs), despite distinguished computer architects John Hennessy and David Patterson backing that approach. Instead, they observed that four course-grained behaviors can compose most algorithms implemented by various DSAs. Data gather, synchronization, and control comprise three of the common behaviors. The fourth, computation, still presents some specialization tradeoffs. A fifth behavior—parallelism—is a given for the targeted applications.

Conceptually, SimpleMachines' compiler decomposes software source code into the four composable behaviors to generate machine code. Built for these behaviors, the hardware then runs these functions with the efficiency of an application-specific chip. The company estimates its approach reduces overhead to only 15%. Although DLAs might achieve similar efficiency for a given model, SimpleMachines argues that many are obsolete by the time they reach production. Our observation is that most data-center-inference accelerators either fail to deliver compelling performance advantages relative to GPUs or fall short in the breadth of models they accelerate.

Reconfigurable Tiles Distill Four Behaviors

Figure 2 shows a block diagram of SimpleMachines' first chip, the 16nm Mozart. At the top level, it looks similar to many DLAs that use a tiled layout connected in a two-dimensional mesh network. Designed to handle large current models as well as future models, it includes in-package HBM2 memories that offer rapid access to gigabytes of data. The chip maintains a single global address space, and the tiles communicate using a message-based protocol.

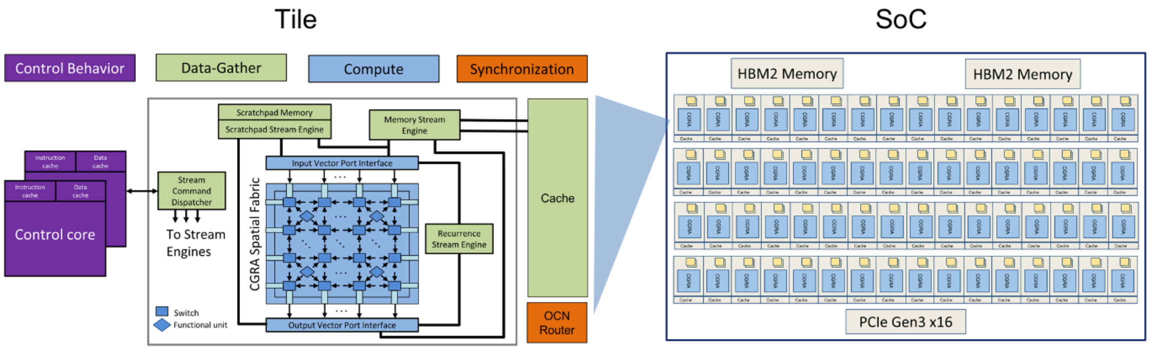

Figure 2. SimpleMachines Mozart block diagram. The tiled top-level architecture is similar to some DLAs, but the computation array in each tile separates the design from other inference chips. (Source: SimpleMachines)

Zooming into the tile level shows that SimpleMachines uses a spatial fabric, separating it from most DLAs. The fabric contains a functional-unit (FU) array, and static routing determines the connections within this array. The compiler determines the array configuration, creating a dataflow architecture at the tile level. This approach is sometimes called a coarse-grained reconfigurable architecture (CGRA), which differs from a fine-grained FPGA. Whereas it can take hours to compile the code for large FPGAs, SimpleMachines compiles its fabric configuration in less than one minute.

Mozart includes a PCIe Gen3 x16 interface, which provides a high-bandwidth host connection. SimpleMachines offers the chip on its Accelerando half-height half-length 75W PCIe card. This form factor is popular for OEMs and cloud vendors because it fits into a standard server and doesn't require an external power connector (that is, it draws all of its power from the PCIe connector). The company also plans to offer a public-cloud service through Equinix, allowing users to rent time on Accelerando-powered systems.

SimpleMachines' software stack is an important enabler of the company's composable-computing vision. It handles running multiple models on a single chip, providing more granular resource allocation for cloud environments. The front-end engine and parser will be compatible with AI frameworks including TensorFlow and ONNX as well as with C++ and Python applications. The behavior compiler decomposes the parsed input into an intermediate representation for each of the four behaviors. A dynamic runtime engine provides the streaming-data host interface and handles board management.

Delivering Performance Breadth

SimpleMachines estimated inference performance for various models at a batch size of four. Some data-center operators require batch=1 processing of certain user requests to minimize response times, and eliminating batching also simplifies software design. On the other hand, the MLPerf Inference benchmarks specify a latency constraint rather than a batch size for relevant multiple-stream and server scenarios. By choosing batch=4 for its estimates, the company struck a balance between these approaches.

Although ResNet-50 is somewhat dated, virtually every vendor reports performance for this model, providing a useful CNN-performance baseline. As Figure 3 shows, SimpleMachines estimates Mozart will deliver 2.4x the throughput of Nvidia's 70W T4 card on this model. Although Nvidia's new A100 exceeds Mozart's performance, it dissipates 400W. Running the BERT model with a sequence length of 384, Mozart is projected to deliver about 60% greater performance per watt than the A100. For the RNN-T model, Mozart is expected to outperform even the A100.

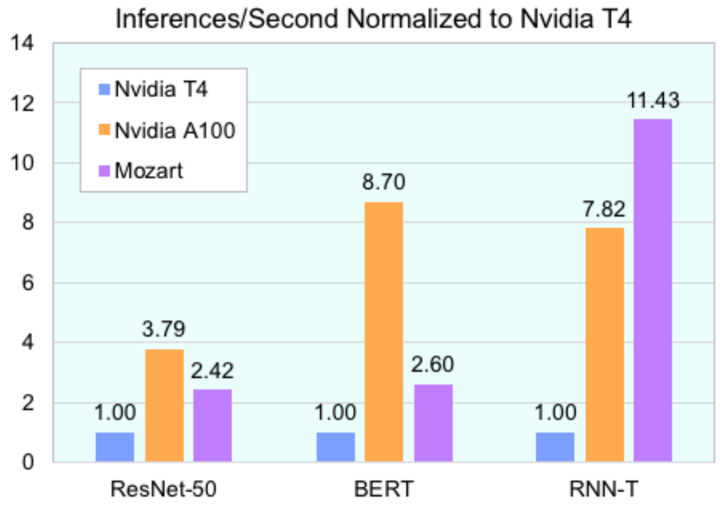

Figure 3. Inference-throughput comparison normalized to Nvidia T4. Mozart performance is estimated for batch=4, whereas all other estimates are batch=4 extrapolations based on public benchmark data. (Data source: SimpleMachines)

Qualcomm's 7nm AI 100 is a recent example of a 75W DLA designed for data-center inference. At a batch size of eight, it delivers 4.4x Mozart's estimated ResNet-50 throughput. Qualcomm claims its chip can handle other types of models but hasn't disclosed any benchmarks, leaving an open question regarding how well it handles Transformer and RNN types. Another example comes from startup Tenstorrent, which offers its Grayskull DLA for 75W cards. For ResNet-50, that chip running batch=1 delivers nearly twice Mozart's throughput. The company has also demonstrated BERT with a sequence length of 128 at 2,830 sentences per second, which exceeds Mozart's estimated throughput by 33% but requires Grayskull to run 10 copies of the model in parallel. Tenstorrent hasn't published RNN benchmarks.

SimpleMachines' performance estimates suggest Mozart will deliver good CNN performance for small batch sizes, whereas it will excel in handling other model types. In addition to the models shown, the company ran a customer's proprietary pattern-matching workload in simulation and estimates Mozart delivers 30x the throughput of T4. Its flexibility makes it comparable to Google's TPUv4, which in training (not inference) benchmarks delivers similar throughput to Nvidia's A100 across CNN, RNN, and Transformer models. Like Google's TPUs, Mozart uses HBM, whereas some DLAs rely on only on-chip SRAM or slower DRAM. These designs can perform well on small models but suffer from lower throughput once a model exceeds on-chip-SRAM capacity.

Mozart is only a starting point for SimpleMachines' architecture. Because the startup combines distributed control with a tiled physical design, it can easily scale its architecture up and down in power and performance. It can also enhance the functional units to handle different data types. The startup's next chip, Bach, will move to 7nm technology and address higher power levels, increasing both throughput and efficiency.

Conclusion

Backed by billions of dollars of investment, the semiconductor industry has developed a large number of application-specific DLA architectures. Yet for public-cloud deployment, merchant competitors have been thus far unable to unseat the more general-purpose Nvidia GPUs. Arguably, Google's captive TPU is the only chip that has meaningfully displaced GPUs, and even Google Cloud offers accelerator-optimized compute instances using Nvidia's A100. The reason is obvious: only GPUs offer enough breadth to accelerate the wide variety of AI workloads that end customers require. Cloud providers may deploy CNN-specific DLAs for certain internal workloads, but doing so would reduce the flexibility of their compute resources.

At the other end of the performance spectrum, some DLAs are integrated into systems on a chip (SoCs). These SoCs can be embedded into a specific type of system, such as smart cameras. The DLAs can thus be highly optimized for the CNNs used in vision processing, and these CNNs are rarely replaced in the field. Client-edge applications often prioritize cost and power over accuracy or performance, so purpose-built DLAs are a good fit in embedded SoCs.

SimpleMachines developed a unique architecture that blends the benefits of GPUs with the efficiency of DLAs. Although Mozart primarily targets AI inference, the startup's composable-computing approach can address AI training as well as other workloads that benefit from a dataflow implementation. With Mozart now available, customers can begin to test their own models and make independent performance assessments. SimpleMachines' initial estimates indicate Mozart should shine for a variety of models, including some that perform relatively poorly on GPUs. Although the chip should also outperform GPUs for real-time CNN inferencing, that's merely table stakes to play in the accelerator market. SimpleMachines' promise is a breakthrough design that can finally displace GPUs for a wide variety of AI workloads.

Bob Wheeler is a principal analyst at The Linley Group and a senior editor for Microprocessor Report. *The Linley Group offers the most-comprehensive analysis of microprocessors and SoC design. We analyze not only the business strategy but also the internal technology. Our in-depth articles cover topics including embedded processors, mobile processors, server processors, AI accelerators, IoT processors, processor-IP cores, and Ethernet chips. For more information, see our website at www.linleygroup.com.*