Deeper, Faster Learning with FPGA Co-Processors

WRITTEN BY BRANDON LEWIS, TERASIC INC.

Deep learning is a disruptive technology for many industries, but its computational requirements can overwhelm standard CPUs. This problem has led developers to consider alternative architectures, but migrating from familiar CPUs to more esoteric designs is a challenge, to say the least. Rather than revamping existing infrastructure to support deep learning, organizations can turn to hybrid CPU + FPGA compute architectures instead.

Even from a performance standpoint, there are good reasons not to abandon CPUs altogether. For one thing, many of these processors can execute isolated deep learning inferencing workloads efficiently, as individual inferences are sequential operations. But when inferencing operations are presented in batches or high volumes, CPUs struggle to keep up.

GPUs and other massively parallel architectures offer an alternative to serial processing. Massive parallelism is well suited to batch inferencing workloads, as well as the training of deep learning models with sizeable input data sets.

Of course, parallel computers are often inefficient when it comes to sequential processing. For applications that require fast sequential inferencing, such as computer vision for autonomous cars and other time-sensitive applications, GPUs are less than optimal.

To meet both low-volume inferencing and large-batch processing demands, devices that integrate an FPGA with a multicore CPU are an appealing option. Because FPGAs are massively parallel in nature, they can readily execute large deep learning batches. Meanwhile, smaller sequential operations can be handled by the CPU. Alternatively, workloads can be shared between the FPGA and CPU to optimize the efficiency of neural networks.

And, because of the architectural flexibility of this heterogeneous architecture, this optimization can be achieved without having to overhaul existing compute infrastructure.

AI Acceleration With FPGAs: In-Line and Coprocessing

For a better understanding of how FPGAs can accelerate deep learning, let's take a look at how they work with multicore CPUs as in-line and coprocessing compute elements.

As an in-line processor, an FPGA sits in front of a CPU and performs preprocessing tasks like data filtering before passing the output on for further computation. As shown in **Figure 1**, vision systems can use the FPGA for in-line filtering or thresholding before sending pixels to a CPU. Because the CPU processes only pixels from regions of interest determined by the FPGA, overall system throughput is increased.

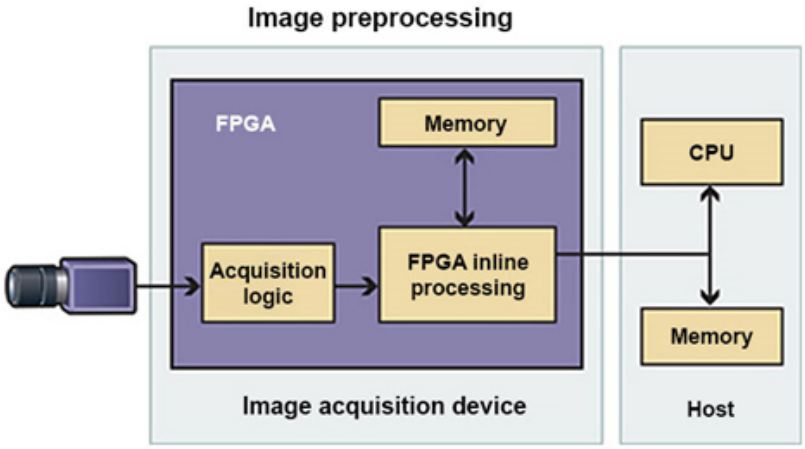

Figure 1. As an in-line processor, FPGAs help increase overall system throughput by filtering data before it reaches a CPU.

As a coprocessor, an FPGA shares the compute workload with a CPU. This can be handled in a number of ways, either by having the FPGA perform parallel processing before sending the output back to the CPU, or by having the FPGA perform all the processing so the CPU can focus on tasks like communications and control.

Continuing with the computer vision example, **Figure 2** shows how a workload can be distributed between an FPGA and CPU with direct memory access (DMA).

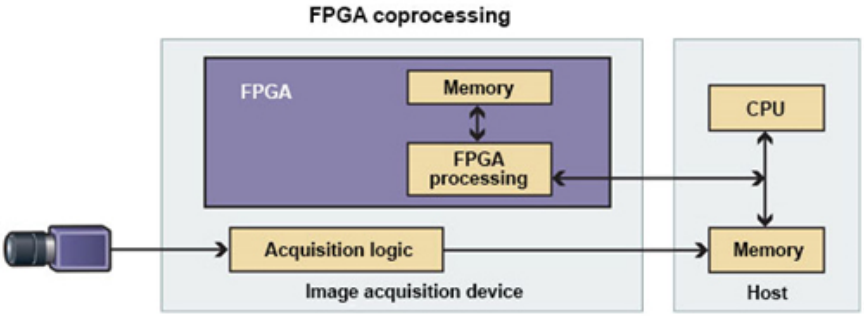

Figure 2. An FPGA coprocessor can share workloads with a CPU via direct memory access (DMA), freeing the CPU for other tasks. (Source: National Instruments)

In summary, pairing FPGAs with multicore CPUs meets the needs of low-volume inferencing and larger-scale batch processing while also increasing system throughput. Still, developers must be able to adopt these solutions with minimal impact on their existing infrastructure.

New FPGAs Deliver Performance Boost, Integration Flexibility

Intel® Stratix® 10 FPGAs offer a path to accelerated deep learning performance and simple integration with deployed systems. These FPGAs integrate as many as 5.5 million logic elements alongside a quad-core 64-bit Arm Cortex-A53 CPU. They also provide programmable I/O pins that allow the FPGAs to interface easily with standard networking and compute technologies.

On the performance front, Intel Stratix 10 devices were designed using the new Intel® HyperFlex™ FPGA Architecture. This architecture introduces hyper-register technology, which places bypassable registers into every routing segment of the device core and at all functional block inputs (**Figure 3**).

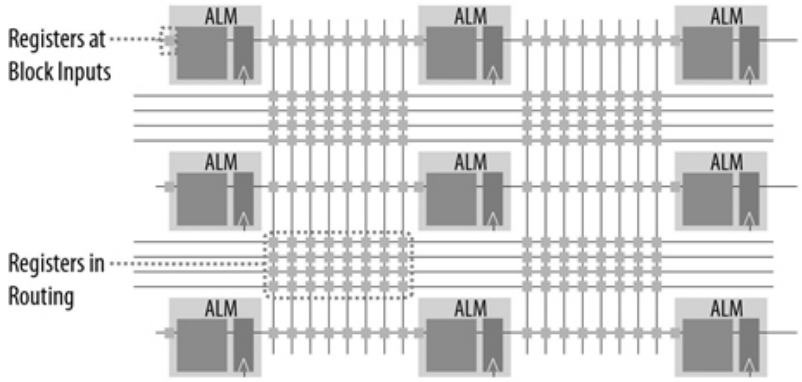

Figure 3. Hyper-Registers place registers at every routing segment and all functional block inputs to enable doubling of the clock frequency. (Source: Intel® Corporation)

The bypassable registers optimize the flow of data across the FPGA fabric, which helps the chips achieve maximum performance. As a result, Intel Stratix 10 devices deliver twice the clock speed of previous-generation FPGAs at 70 percent lower power consumption. This remarkable achievement makes the FPGAs a good fit for performance-hungry but power-constrained applications.

In terms of platform integration, Intel Stratix 10 FPGA devices support both serial and parallel flash interfaces. These memory types—which are common in networking platforms—have great utility for deep learning, as they allow developers to choose a configuration that best suits their workload. The DE10-Pro Stratix 10 GX/SX PCIe Board from Terasic Inc., for instance, supports multiple types of memory for various applications (**Figure 4**):

- QDR-IV memory module for high-bandwidth, low-latency applications
- QDR-II+ memory module for low-latency memory read/write
- DDR4 for applications that require the largest memory capacity possible

Figure 4. The DE10-Pro Stratix 10 GX/SX PCIe Board from Terasic, Inc. supports multiple memory types for different deep learning use cases.

The DE10-Pro includes x16 PCIe Gen 3 lanes for chip-to-chip data transfer speeds of up to 128 Gbps, while four QSFP28 connectors all support 100 Gigabit Ethernet. These interfaces enable tremendous data offload capabilities, as well as quick read-and-write memory access. In server or data center environments, this means workloads can be shared between banks of compute and memory resources to scale deep learning performance as needed.

Finally, from a software perspective, the DE10-Pro Stratix 10 GX/SX PCIe Board supports the Intel® Open Visual Inference & Neural Network Optimization (Intel® OpenVINO™) toolkit. OpenVINO is a development suite for heterogeneous execution architectures that is based on a common API that abstracts the complexity of FPGA programming.

OpenVINO includes a library of functions, kernels, and optimized calls for OpenCV and OpenVX, and has demonstrated performance enhancements of up to 19x for computer vision and deep learning workloads (**Figure 5**).

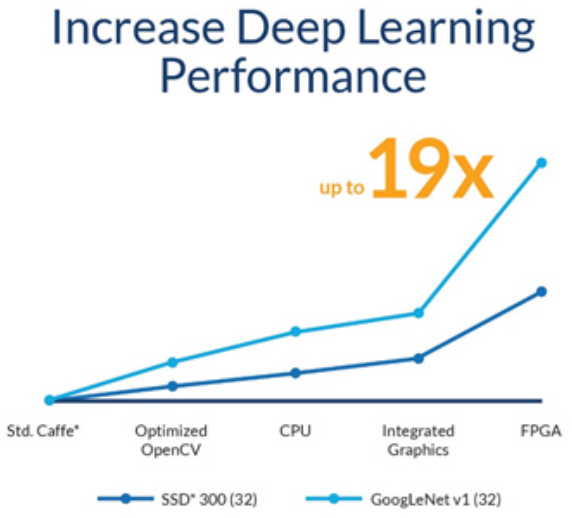

Figure 5. The Open Visual Inference & Neural Network Optimization (OpenVINO™) toolkit has demonstrated significant performance optimizations. (Source: Intel® Corporation)

Accelerate Faster

Deep learning workloads are prompting innovation across the technology sector in general, and in the processing market in particular. Industry is currently investigating new ways to compute deep learning workloads using processors designed specifically for neural network execution.

FPGAs with integrated multicore CPUs provide the flexibility and performance to execute deep learning workloads where, when, and how the highest throughput can be achieved. They also offer a migration path to future demands, be they in artificial intelligence, next-generation networks, or any segment that can be addressed by high-performance computing (HPC).