White Paper

Cloud, Data Center
Heterogeneous FPGA Accelerator

Pushing AI Boundaries with Scalable Compute-Focused FPGAs

Authors

Markus Adhiwiyogo
Product Line Manager
Intel Corporation

Rohit D'Souza
AI and Acceleration
Product Marketing Manager
Intel Corporation

Steve Leibson
Senior Marketing Engineering Manager
Intel Corporation

Ronak Shah
Director, Software Product Marketing
and Product Line Management
Intel Corporation

Table of Contents

Introduction 1
Hardware Customization with Integrated AI 1
- Microsoft Bing Search Engine 1
- Increasing AI Model Complexity 1

Introducing the Intel® Stratix® 10 NX FPGA 2
- High Tensor Throughput for Low-Precision AI Inference Workloads 2
- Scalable and Flexible I/O Connectivity Bandwidth 3
- HBM2 Solves Memory Bandwidth Bottlenecks, Delivers Low Latency . . 3

Intel Stratix 10 NX FPGA Applications 4
- Natural Language Processing 4
- Financial Fraud Detection 4
- Smart City and Retail 5

Conclusion 5
For Additional Information 6
References 6

Introduction

FPGAs have enabled customers to quickly innovate and adapt to changing market trends through in-system hardware customization. A key trend today is an increasing and pervasive need for intelligent processing and Artificial Intelligence (AI) alongside the need for customization. Building on the inherent value proposition of FPGAs in fabric and I/O flexibility as well as low and deterministic latency, innovative Intel® FPGAs are addressing hardware customization along with AI.

Hardware Customization with Integrated AI

The concept of hardware customization with integrated AI addresses the growing need to innovate in traditional markets by adding AI capabilities. A great example is Microsoft's Bing search engine application.

Microsoft Bing Search Engine

Microsoft's Bing search engine uses Intel® FPGAs with real-time AI to deliver intelligent and better search results to users [2]. Microsoft deploys machine reading comprehension to understand the meaning behind a user's query to provide answers "based on all the perspectives on the web" [2] in its results. The machine reading comprehension models are run across many web pages and aggregated results are presented to users. This capability, requiring high computation power without slowing the search, is enabled by a deep learning acceleration platform which Microsoft calls 'Project Brainwave' [2].

But the AI landscape is constantly evolving with demand for innovation advancing at a rapid clip. A key, disruptive trend is the exponential increase in AI model size and complexity.

Increasing AI Model Complexity

The increasing complexity of AI models and the explosive growth of AI model size are both rapidly outpacing innovations in compute resources and memory capacity available on a single device. AI model complexity now doubles every 3.5 months or ~10X per year [1] (see Figure 1), which is leading to a rapidly increasing demand in AI computing capability. Memory requirements for AI models are also rising due to an increasing number of parameters or weights in a model. More parameters demand more on-chip storage to maintain model persistence [5]. The increase in required memory also translates to a need for higher I/O bandwidth because more data must be transported. This growth rate is likely to continue.

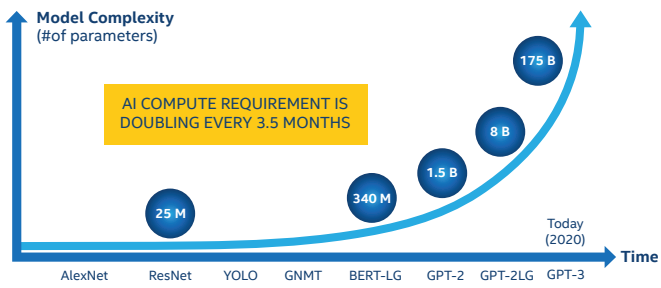

Figure 1. Increasing Complexity of AI Models [1, 6]

Introducing the Intel® Stratix® 10 NX FPGA

The Intel® Stratix® 10 NX FPGA is Intel's first AI-optimized FPGA. Intel has developed the Intel Stratix 10 NX FPGA to enable customers to scale their designs with increasing AI complexity while continuing to deliver real-time results. This new Intel FPGA family addresses the challenges of increasing AI model complexity through the following features:

- A new type of high-density, lower-precision, AI-optimized tensor arithmetic block specifically designed for the needs of AI models
- High-performance transceiver tiles for high-speed networking and low-latency host connectivity
- Integrated, second-generation, high-bandwidth memory (HBM) stacks to meet the extreme memory requirements of the largest AI models

The Intel Stratix 10 NX FPGA is implemented using Intel's chiplet-based architecture, in which the FPGA core die is connected to custom silicon tiles or chiplets in the same package for maximum flexibility and agility. This chiplet-based architecture allows Intel to combine the right mix of functionality using the right process nodes to provide the system functionality that customers need in a single package. For example, transceiver chiplets are used to deliver the exact combination of networking and processor-attach capabilities for system developers in a fraction of the development time and resources to tape out an entirely new chip with integrated transceivers (see Figure 2). This design and manufacturing technology results in faster time to market for system developers. In the case of Intel Stratix 10 NX FPGAs, the FPGA core die's digital signal processing (DSP) block is revised with an AI-optimized tensor arithmetic block.

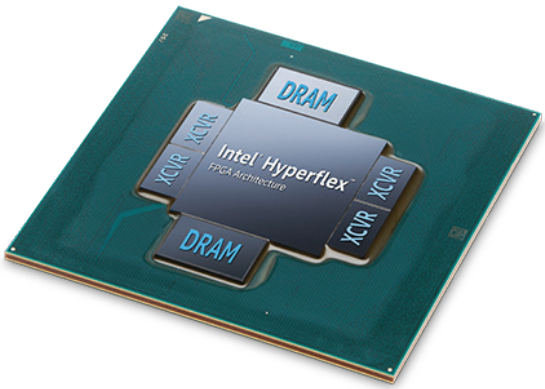

Figure 2. Chiplets in Intel® Stratix® 10 NX FPGA

High Tensor Throughput for Low-Precision AI Inference Workloads

The Intel Stratix 10 NX FPGA fabric includes a new type of AI-optimized tensor arithmetic block called the AI Tensor Block. These AI Tensor Blocks contain dense arrays of lower-precision multipliers typically used for AI model arithmetic. The smaller multipliers in these AI Tensor Blocks can also be aggregated to construct larger-precision multipliers.

The AI Tensor Block's architecture (see Figure 3) contains three dot-product units, each of which has ten multipliers and ten accumulators for a total of 30 multipliers and 30 accumulators within each block. The AI Tensor Block is tuned for the common matrix-matrix or vector-matrix multiplications used for AI computations with capabilities designed to work efficiently for both large and small matrix sizes. The AI Tensor Block multipliers' base precisions are INT8 and INT4 along with shared exponent to support Block Floating Point 16 (Block FP16) and Block Floating Point 12 (Block FP12) numerical formats. Multiple AI Tensor Blocks can be cascaded together to support larger vector calculations. The overall compute performance and efficiency is shown in Table 1.

PRECISION	PERFORMANCE	EFFICIENCY
INT4	286 TOPS	2 TOPS/W
INT8	143 TOPS	1 TOPS/W
Block FP12	286 TFLOPS	2 TFLOPS/W
Block FP 16	143 TFLOPS	1 TFLOPS/W
@600 MHz Max Frequency		

Table 1. Performance and Efficiency of AI Tensor Block [3]

With 3,960 AI Tensor Blocks available, the peak performance numbers for INT8 precision with a maximum target frequency of 600 MHz can be calculated as follows [3]:

*(3,960 AI Tensor Blocks) * (10 inputs * 3 columns) * (2 operations per multiplication)*

*= (3,960 AI Tensor Blocks) * (30 multiplications per AI Tensor Block) * (2 operations per multiplication)*

= 237,600 operations

Assuming 600 MHz maximum frequency

*(237,600 operations) * (600 MHz)*

= 142.56 TOPS

~143 TOPS

Similarly, the peak performance for INT4 precision is double that of INT8, or approximately 286 TOPS

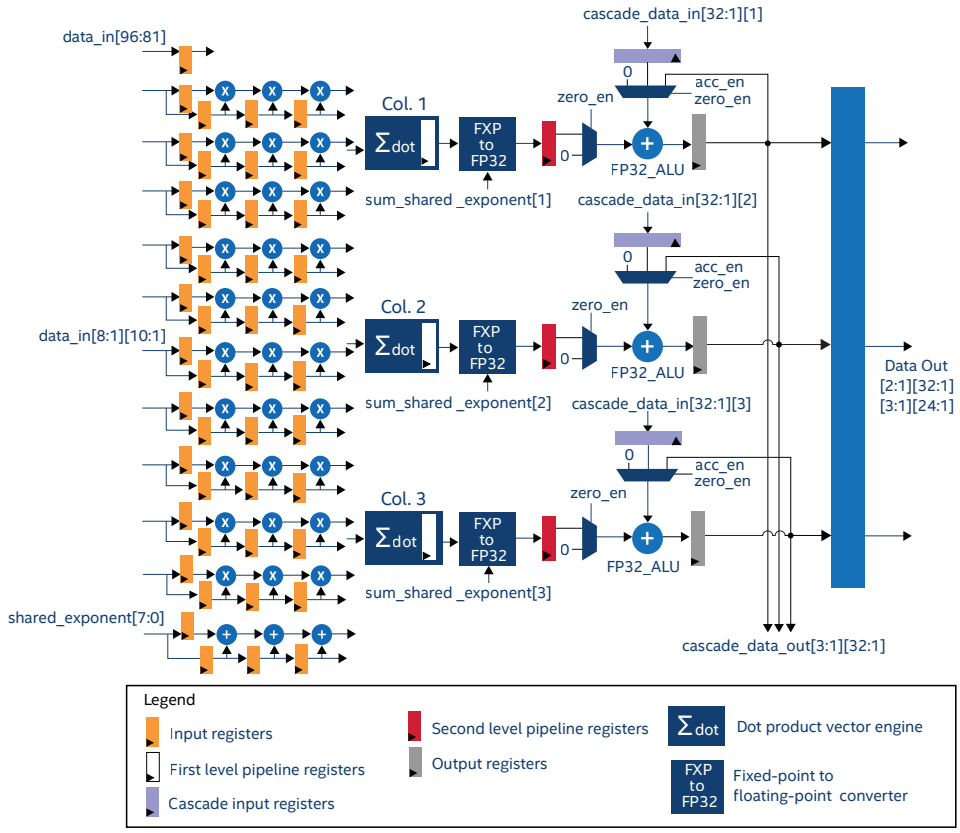

Figure 3. AI Tensor Block architecture

In short, Intel estimates that the AI Tensor Block features and capabilities allow a single-chip implementation of ResNet50 to process images at approximately 7,000 frames per second (FPS) (see Figure 4). This estimate is based on 500 MHz clock rate and assuming 75% utilization of the AI Tensor Blocks [3].

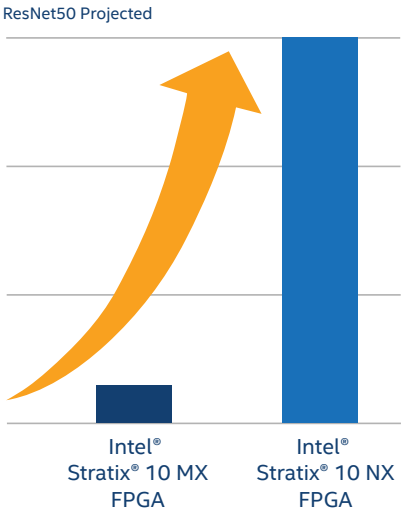

Figure 4. Theoretical performance for ResNet50 topology [3]

Scalable and Flexible I/O Connectivity Bandwidth

Intel Stratix 10 NX FPGAs contain up to 96 serializer/deserializer (SERDES) transceivers. The transceivers operate at a maximum data rate of 58.8 Gbps PAM4 or 28.9 Gbps non return to zero (NRZ). Each variant also includes PCI Express* (PCIe*) Gen3 x16 and 100G Ethernet media access control (MAC) or physical coding sublayer (PCS) hard intellectual property (IP) cores. The transceivers allow designers to connect Intel Stratix 10 NX FPGAs to a variety of other devices using many standard or custom protocols. Taking all these features into account, Intel Stratix 10 NX FPGAs provide up to 668 GB/s of connectivity bandwidth and allow designers to implement multi-node AI inference solutions with higher connectivity bandwidth requirements. This results in a scalable connectivity solution with the flexibility to adapt to changing bandwidth requirements.

HBM2 Solves Memory Bandwidth Bottlenecks, Delivers Low Latency

Intel Stratix 10 NX devices integrate two HBM2 memory stacks with a high-performance FPGA fabric in a single package, which allows these FPGAs to effectively address the AI memory-bandwidth bottleneck challenge. Each HBM2 stack provides up to 256 GB/s of bandwidth, providing as much as 512 GB/s of aggregate bandwidth in a single package - significantly more bandwidth than delivered by four externally connected DDR4 memories.

Devices in the family include 8 GB or 16 GB of memory in two HBM2 memory stacks, so both HBM2 capacity options have the same aggregate bandwidth. The integrated HBM2 memory stacks allow large AI models to persist in the large on-chip HBM2 memories, lowering access latency compared to off-chip memory. By delivering large memory bandwidth, the HBM2 memory stacks mitigate the memory bottleneck for memory-bound workloads, ultimately leading to lower overall latencies compared to devices without HBM.

This benefit is illustrated in Figure 5 where the increase in model size reduces the rate of latency increase in the Intel Stratix 10 NX FPGA with HBM2 DRAM compared to an FPGA that uses external DDR4 SDRAM. Figure 5 shows the latencies associated with the FPGA without HBM (blue line) increase dramatically after consuming all of the FPGA's internal SRAM resources, forcing the use of external DDR4 SDRAM. The latencies for the Intel Stratix 10 NX FPGA with HBM2 (orange line) increase at a much lower rate after consuming all of the FPGA's on-chip SRAM resources [4]. In addition, integration of the HBM2 memory stacks lowers system power consumption and reduces the board form factor compared to exclusive use of external DDR4 SDRAM, resulting in lower total cost of ownership (TCO).

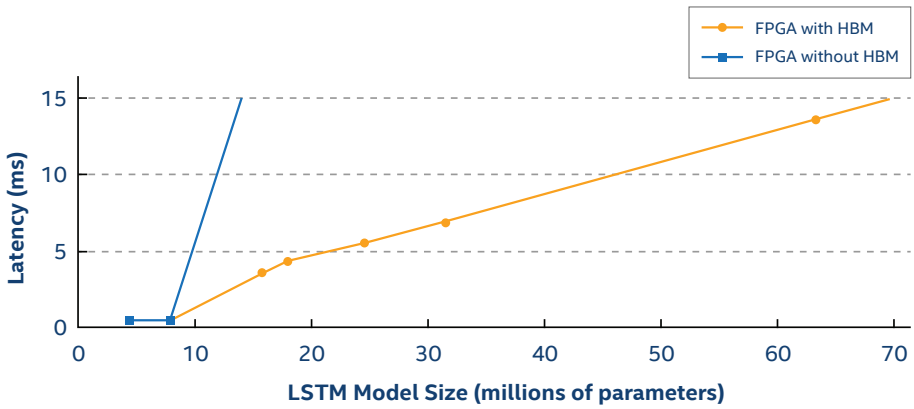

Figure 5. Latency comparison between FPGAs with and without HBM [4]

Intel Stratix 10 NX FPGA Applications

In general, FPGAs excel in AI applications requiring real time, low batch, and low latency. The following sample applications demonstrate how hardware customization with integrated AI using the Intel Stratix 10 NX FPGA can play a critical role in delivering to real-time requirements.

Natural Language Processing

Millions of users interact with cloud-based devices like Siri or Alexa and expect a real-time, interactive, and intelligent response (see Figure 6). To respond to millions of simultaneous customer queries in real time, the system must process each request in small batches while maintaining low latency. The AI Tensor Blocks along with the customizable memory hierarchy and pipelined parallel architecture enables the Intel Stratix 10 NX FPGA to do just that. By building the entire system in a pipelined fashion across multiple nodes, system designers can create a real-time interactive solution.

Figure 6. Speech-to-Text and Text-to-Speech - Natural Language Processing Application

Myrtle.ai, an Intel partner, has demonstrated a real-time text-to-speech synthesis application on the Intel Stratix 10 NX FPGA. The application deploys WaveNet, an AI model that produces state-of-the-art audio quality. For more details of this demonstration, check out the following resources:

- Demo video: www.intel.com/text-speech-fpga-demo
- White paper: www.intel.com/text-speech-fpga-paper

Financial Fraud Detection

Financial fraud detection is an application where every millisecond counts. The goal is to detect and deter a fraudulent credit card transaction before it is processed. This goal can be achieved through the Intel Stratix 10 NX FPGA's ability to create low- and deterministic-latency custom hardware as well as the ability to support model persistence via high-bandwidth networking and memory accesses.

In a real-world example of financial fraud detection system (see Figure 7), the total roundtrip latency requirement from the time a credit card is swiped at a point-of-sale (POS) terminal to authenticating the transaction must be less than 400 milliseconds. However, the AI algorithm has only 10 milliseconds to run the AI inference algorithm and determine whether a transaction is fraudulent or not. This application is run at batch size of 1, which is where FPGAs shine. The Intel Stratix 10 NX FPGA, with its high-performance AI Tensor Blocks and high-bandwidth HBM2 memory can easily meet the stringent low latency requirements in financial fraud detection applications.

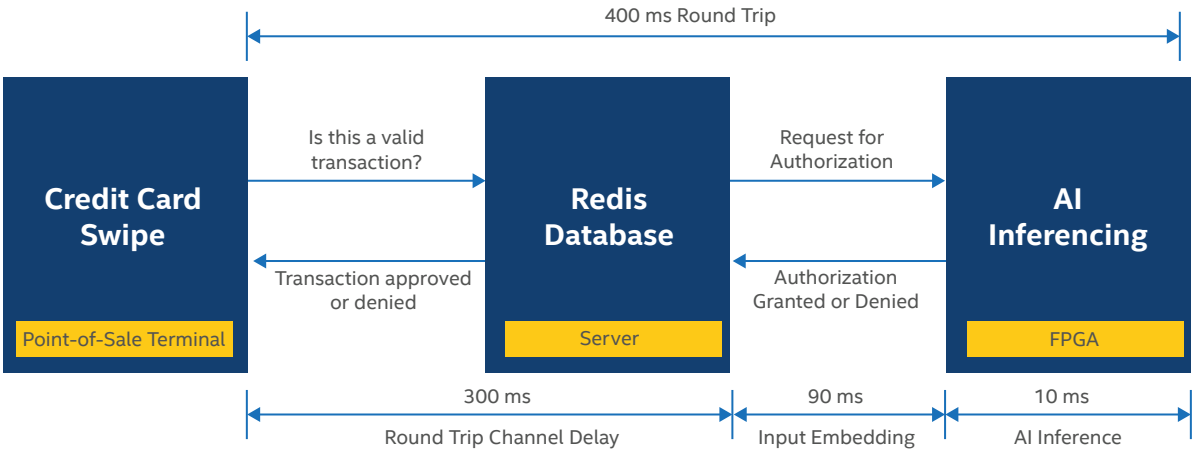

Figure 7. A high-level view of a financial fraud detection system and associated latencies [3]

Smart City and Retail

Certain video analytics applications have stringent real-time response requirements. These applications typically require integration of video ingestion and processing with customized algorithms. FPGAs excel in these unique video analytics applications because of their hardware customization ability, which allows implementation of custom processing and custom I/O protocols. An example of video analytics in a smart retail application appears in Figure 8. In this application, real-time video images of items being purchased are recognized and matched to a scanned barcode. This application is designed to reduce financial losses when customers scan one item but pass a different, more expensive item into their shopping bag. The entire pipelined operation takes less than 50 milliseconds, due to the direct video input and pipelining of the video-ingest, transformation, and the AI inference stages – all enabled by the FPGA.

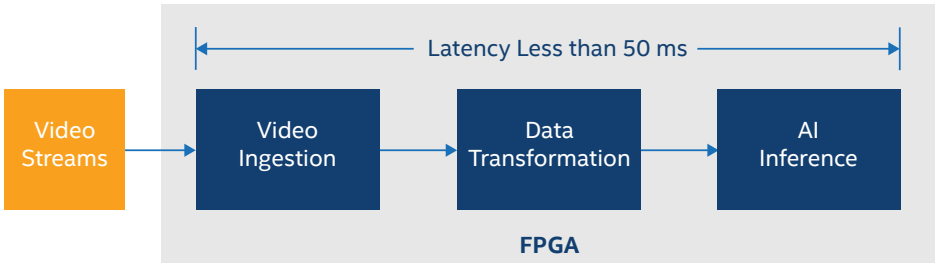

Figure 8. Pipelined implementation of an AI-powered application for checkout in Smart Retail [7][3]

Conclusion

The Intel Stratix 10 NX FPGA delivers a unique combination of features needed to implement high-performance AI systems. It holistically addresses increasing AI model complexity in all four of the following categories:

- **High-Compute Density**: Hardened AI Tensor Blocks deliver as much as 286 INT4/Block FP12 or 143 INT8/Block FP16 TOPS/TFLOPS [3] leaving the available soft logic for application support.
- **Fast Interconnect**: Up to 668 GB/s of high-bandwidth networking is available on each Intel Stratix 10 NX FPGA enabling multi-node applications across multiple Intel Stratix 10 NX FPGAs
- **Abundant Near-Compute Memory**: In addition to on-chip SRAM, 8 or 16 GB of HBM2 memory is available for storing large AI models and datasets.
- **Hardware Customization with Integrated AI**: Supports custom applications and AI on the same device.

Due to these unique capabilities, Intel Stratix 10 NX FPGAs are being adopted to address the trend towards larger AI models requiring greater compute density, memory bandwidth, and scalability across multiple nodes.

For Additional Information

Visit the Intel Stratix 10 NX FPGAs web page at www.intel.com/stratix10nx

References

[1] https://s21.q4cdn.com/600692695/files/doc_presentations/2019/11/intel-ai-summit-keynote-slides.pdf

[2] www.intel.com/content/www/us/en/programmable/b/bing-intelligence-search-with-intel-fpgas.html

[3] Based on internal Intel estimates

[4] E. Nurvitadhi et al., "Scalable Low-Latency Persistent Neural Machine Translation on CPU Server with Multiple FPGAs," 2019 International Conference on Field-Programmable Technology (ICFPT), Tianjin, China, 2019, pp. 307-310.

[5] E. Nurvitadhi et al., "Why Compete When You Can Work Together: FPGA-ASIC Integration for Persistent RNNs," 2019 IEEE 27th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM), San Diego, CA, USA, 2019, pp. 199-207.

[6] https://medium.com/@Synced/openai-unveils-175-billion-parameter-gpt-3-language-model-3d3f453124cd

[7] https://megh.com/video-analytics-solution/

WP-01301-1.0