(intel)

Intel® Stratix® 10 NX FPGA

AI-Optimized FPGA for High-Bandwidth, Low-Latency AI Acceleration

UP TO

15X MORE INT8
THROUGHPUT [1]

The Intel® Stratix® 10 NX FPGA delivers a unique combination of capabilities needed to implement customized hardware with integrated high-performance artificial intelligence (AI). These capabilities include:
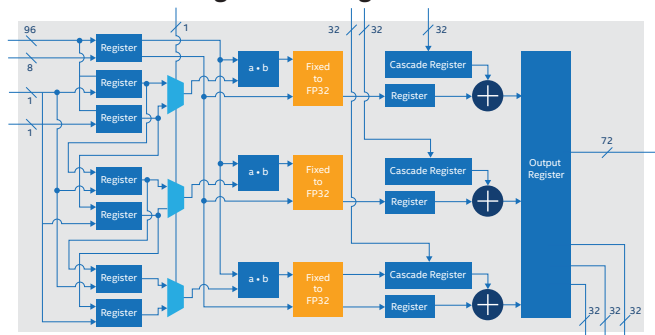
- **High-Performance AI Tensor Blocks**
  - Up to 15X more INT8 throughput than Intel Stratix 10 FPGA digital signal processing (DSP) block for AI workloads[1]
  - Hardware programmable for AI with customized workloads

- **Abundant Near-Compute Memory**
  - Embedded memory hierarchy for model persistence
  - Integrated high- bandwidth memory (HBM)

- **High-Bandwidth Networking**
  - Up to 57.8 G PAM4 transceivers and hard Ethernet blocks for high efficiency
  - Flexible and customizable interconnect to scale across multiple nodes

These three sets of capabilities allow Intel Stratix 10 NX FPGAs to uniquely address the trend towards low latency and larger AI models requiring greater compute density, memory bandwidth, and scalability across multiple nodes as well as reconfigurable custom functions.

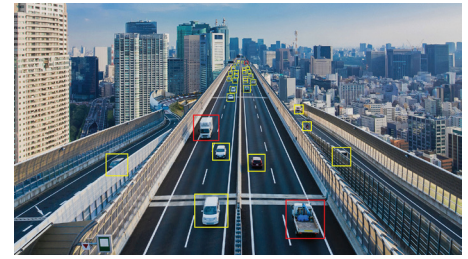### Introducing AI Tensor Block: Enabling Breakthrough Compute Density

The Intel Stratix 10 NX FPGA fabric includes new types of AI-optimized tensor arithmetic blocks called the AI Tensor Blocks. These blocks contain dense arrays of lower-precision multipliers typically used in AI applications. The AI Tensor Block's architecture is tuned for common matrix-matrix or vector-matrix multiplications used in a wide range of AI computations with capabilities designed to work efficiently for both small and large matrix sizes.

AI Tensor Block High-Level Diagram

The AI Tensor Block multipliers have base precisions of INT8 and INT4 and support FP16 and FP12 numerical formats through shared-exponent support hardware. All additions or accumulations can be performed with INT32 or IEEE754 single-precision floating point (FP32) precision and multiple AI Tensor Block can be cascaded together to support larger matrices. A single AI Tensor Block is estimated to achieve up to 15X more INT8 throughput when compared to standard Intel Stratix 10 FPGA DSP block.[1]

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

Extending AI+ for Low Latency and Large Models Across Multi-Node Solution

Natural Language Processing

- Speech recognition
- Speech synthesis

Security

- Deep packet inspection
- Congestion control identification
- Fraud detection

Real-Time Video Analytics

- Content recognition
- Video pre and post processing

Intel® Stratix® 10 NX FPGA – Key Attributes

KEY ATTRIBUTES	
AI Tensor Block	Tuned for AI arithmetic, the AI Tensor Block is estimated to provide up to 15X more INT8 throughput than standard Intel Stratix 10 FPGA DSP block[1] for high-compute density needed for high-throughput AI inference applications.
In package 3D stacked HBM2 high-bandwidth DRAM	Integrated memory stacks allow for large, persistent AI models to be stored on-chip, which results in lower latency with large memory bandwidth to prevent memory-bound performance challenges in large models.
Transceiver data rates	With up to 57.8 G PAM4 transceivers, Intel Stratix 10 NX FPGAs provide the scalability and the flexibility to implement multi-node AI inference solutions, reducing or eliminating bandwidth connectivity as a limiting factor in multi-node designs. The Intel Stratix 10 NX FPGA also incorporate hard intellectual property (IP) such as PCI Express* (PCIe*) Gen3 x16 and 10/25/100G Ethernet media access control (MAC)/physical coding sublayer (PCS)/forward error correction (FEC). These transceivers provide a scalable and flexible connectivity solution to adapt to market requirements.

For More Information

Visit Intel Stratix 10 NX FPGA homepage: www.intel.com/stratix10nx