

Intel® Stratix® 10 NX FPGA

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

UP TO

143 INT8 TOPS

286 INT4 TOPS¹

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:

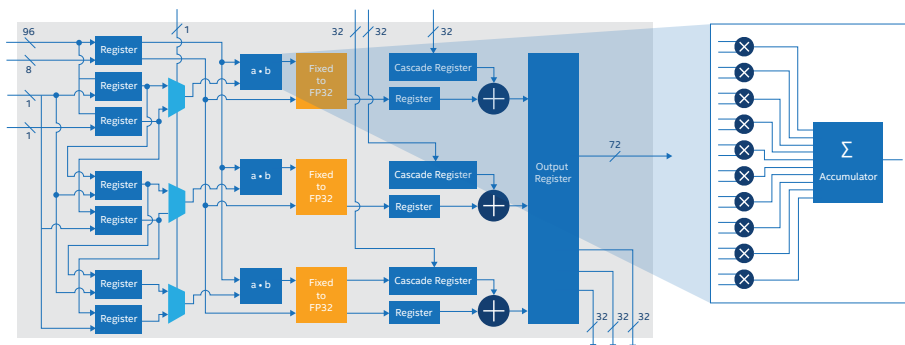
- **High-Performance AI Tensor Blocks**
 - Up to 143 INT8 TOPS or 286 INT4 TOPS at 1 to 2 TOPS/W for AI workloads¹
 - Hardware programmable for AI with customized workloads
- **Abundant Near-Compute Memory**
 - Embedded memory hierarchy for model persistence
 - Up to 16GB of integrated high-bandwidth memory (HBM) with up to 512GB/s of in-package memory bandwidth
- **High-Bandwidth Networking**
 - Up to 57.8 G PAM4 transceivers providing up to 668GB/s of connectivity bandwidth
 - Up to twelve hard 100G Ethernet MAC/PCS/FEC
 - 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. Each block contains three dot-product units, each of which has ten multipliers and ten accumulators, for a total of 30 multipliers and 30 accumulators within an AI Tensor Block. 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.

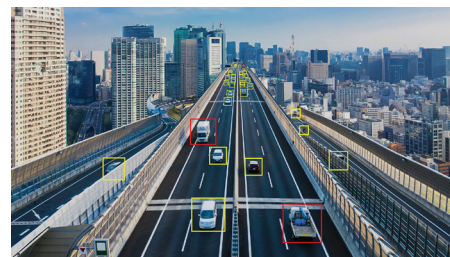
AI Tensor Block High-Level Diagram



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

The AI Tensor Block multipliers have base precisions of INT8 and INT4 and support Block Floating Point 16 (Block FP16) and Block Floating Point 12 (Block 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. The Intel Stratix 10 NX FPGA is estimated to achieve up to 143 INT8/Block FP16 TOPS/TFLOPS or 286 INT4/Block FP12 TOPS/TFLOPS.¹

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



¹ Based on internal Intel estimates.

Tests measure performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.

Intel technologies may require enabled hardware, software or service activation.

No product or component can be absolutely secure.

Results have been estimated or simulated. Your costs and results may vary.

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.