Architecture Brief

FPGAs
DSP

Agilex™ 5 FPGAs: Enhanced DSP with AI Tensor Block

Authors

Lo Muzio, Pierluigi
DSP Technical Specialist
Altera Corporation

Privitera, Giuseppe
Broadbase Product Portfolio
Marketing Sr. Manager
Altera Corporation

Introduction

The Agilex™ 5 FPGA offers digital signal processing (DSP) blocks with artificial Intelligence (AI) capabilities.

The Enhanced DSP with AI Tensor Block is a flexible and configurable block containing dedicated multipliers and dot products (sum of multiplications) tensor columns with supporting circuitry consisting of adders, subtractors, accumulators, shifters, and registers. The purpose of the Enhanced DSP with AI Tensor Block is to allow efficient implementation of commonly used DSP functions from either a single or a group of DSP blocks. Furthermore, it is also optimized to support Machine Learning (both training and inference) applications by the new added tensor products.

The Agilex 5 FPGA fundamentally on one side, inherits Agilex FPGA DSP block, which already offered some new AI capabilities. On the other side, includes some features inherited from Stratix 10 NX FPGA to address the needs for AI Tensor Operations.

The computing capabilities of Agilex 5 FPGAs are not limited to the inheritance of the advanced features of high-end devices such as Stratix 10 NX and Agilex FPGAs.

The Agilex 5 FPGA Enhanced DSP with AI Tensor Block has been augmented by two important developments, the first one for AI, image, or video processing applications and the second one for DSP-intensive applications, which make usage of complex numbers.

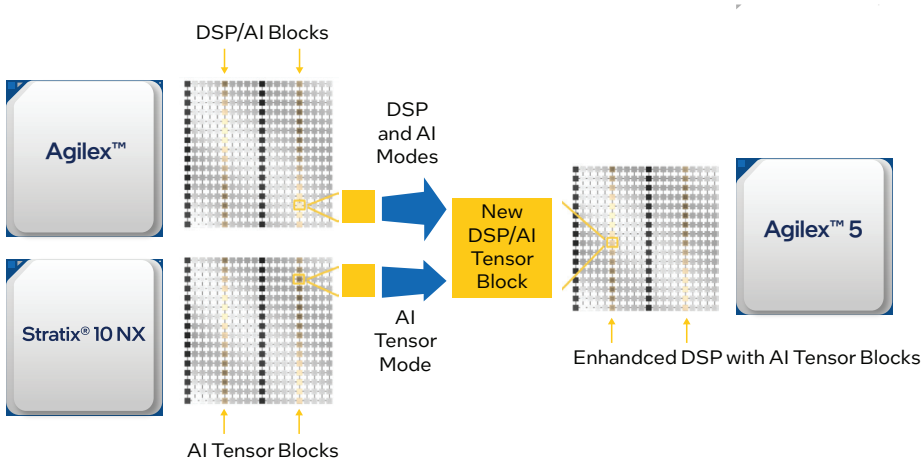


Figure 1. From Agilex FPGA and Stratix 10 NX FPGA to Agilex 5 FPGA

Table of Contents

Fixed-Point and Floating-Point Modes and Enhancements

Agilex 5 FPGAs inherit the Agilex 7 FPGAs DSP variable precision block, which can handle multiple fixed-point and floating-point precisions operations:

- Fixed point: 9x9, 18x19, 27x27 operations
- Floating-point format:
  - Single-precision 32-bit arithmetic FP32 floating-point mode,
  - Half-precision 16-bit arithmetic FP16
  - FP19 floating-point modes
  - BFLOAT16 floating point

The advantage of the variable precision architecture consists of the capability to trade precision with computing performances because the block can be configured for lower precision with the scope to offer more computation units. This capability existed for fixed point in the previous FPGA families and it has been extended to floating point for the first time in Agilex 7 FPGAs. The variable precision architecture block can execute double number of half precision FP16 floating-point multiplications in comparison with single-precision FP32.

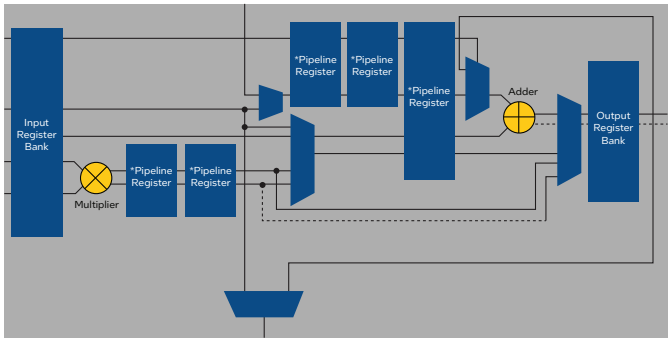


Figure 2. Single-Precision Floating-Point Arithmetic 32-bit

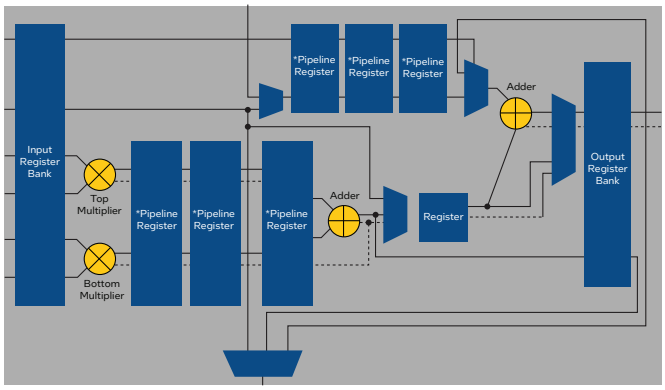


Figure 3. Half-precision point Arithmetic 16-bit

Agilex 5 FPGA replicates this new capability of Agilex 7 FPGA and all the other features, which are listed in the following table.

DSP Block Features	Fixed Point	Floating Point
Input and output register banks	✓	✓
First and second pipeline registers	✓	✓
Pre-adder/subtract, internal coefficients	✓	–
Systolic registers, double accumulation register	✓	–
Exception handling	–	✓
Multipliers, adder, chainin-chainout-adder/accumulator	✓	✓

Table 1. DSP block features

In comparison with Agilex 7 FPGA, Agilex 5 FPGA introduces two new important modes:

- Complex mode
- INT9 vector mode

New Complex Mode

The complex mode doubles the performances for Complex Multiplications. Two DSP block were necessary in the previous families. Agilex 5 FPGAs can execute the 16-bits fixed-point complex multiplication within a single DSP block.

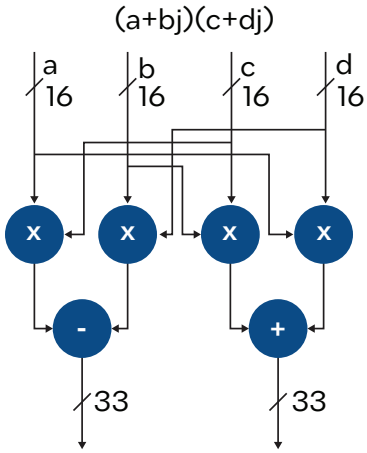


Figure 4. Complex mode

Enhanced INT9 Vector Mode

The other new mode is INT9 vector mode, where one summation of six signed (9x9) multiplications can be executed within the DSP block, instead of four supported by Agilex 7 FPGAs.

A scalar product of two vectors of six elements in a 9-bit fixed point can be executed within the DSP block of an Agilex 5 FPGA. Of course, this mode applies also to INT8 that is quite frequent data format for image processing: a scalar product of two vectors of six INT8 words can be executed within the DSP block of an Agilex 5 FPGA.

*This block diagram shows the functional representation of the DSP block. The pipeline registers are embedded within the various circuits of the DSP block

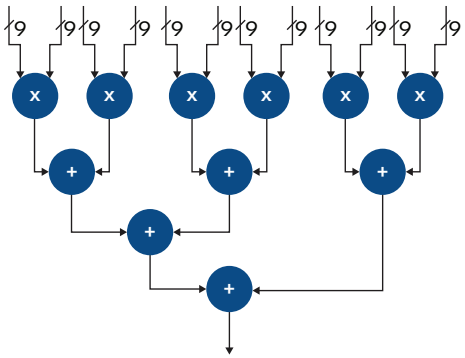


Figure 5. CNT9 Vector Mode

The capabilities of the DSP block for INT9 data format recently evolved from one family to another. Cyclone® V FPGAs offered 3x INT9 multipliers per DSP block, 50% more multipliers than a 18x19 mode. Then the Agilex 7 FPGA improves INT9 mode with four multipliers and Agilex 5 FPGA improves it further to six multipliers.

The evolution of the number of multipliers per DSP block is summarized in the following table across FPGA families. The FPGA synthesis tool in the Quartus® Prime Software allows the inference of all the DSP operational modes of the new block so that the RTL engineers can easily use the new modes.

The floating-point data formats are easier to be handled in High Level Design flows, as DSP Builder for Altera® FPGAs and oneAPI.

AI Tensor Mode as Stratix 10 NX FPGA

The Agilex 5 FPGA comes with a unique combination of capabilities that provides you with all that is needed to develop a customized hardware with integrated high-performance AI. At the center of these capabilities is a new type of operational mode called the AI Tensor Mode, which is tuned for the common matrix-matrix or vector-matrix multiplications used in AI computations. This mode has the capabilities designed to work efficiently for both small and large matrix sizes. A single Enhanced DSP with AI Tensor Block achieves up to 25X peak, theoretical TOPS improvement in a single DSP block's INT8 operations versus Cyclone V FPGAs.

In 2020 Altera launched the Stratix 10 NX FPGA with a new AI Tensor Block, which offered an enormous amount of AI operations. The Agilex 5 FPGA inherits the Tensor Mode. The scalar product of 10 elements is the fundamental operation of the Tensor Mode.

The Agilex 5 FPGA AI Tensor Block architecture is tuned for common matrix-matrix or vector-matrix multiplications used in a wide range of AI-based workloads with capabilities designed to work efficiently for both small and large matrix sizes.

DSP block Data Format	Cyclone® V FPGA	Arria® 10 FPGA and Stratix® 10 FPGA	Agilex® 7 FPGA	Agilex® 5 FPGA
FP32 IEEE754	–	1	1	1
FP16 IEEE754 FP19 / BFLOAT16	–	–	2	2
INT9	3	–	4 →	6
INT18x19	2	2	2	2
Complex 16x16	1/2	1/2	1/2 →	1
INT27x27	1	1	1	1

Table 2. Number of multipliers or MAC per DSP block comparison between Cyclone V FPGAs, Arria 10 FPGAs, Stratix 10 FPGAs, Agilex 7 FPGAs and Agilex 5 FPGAs.

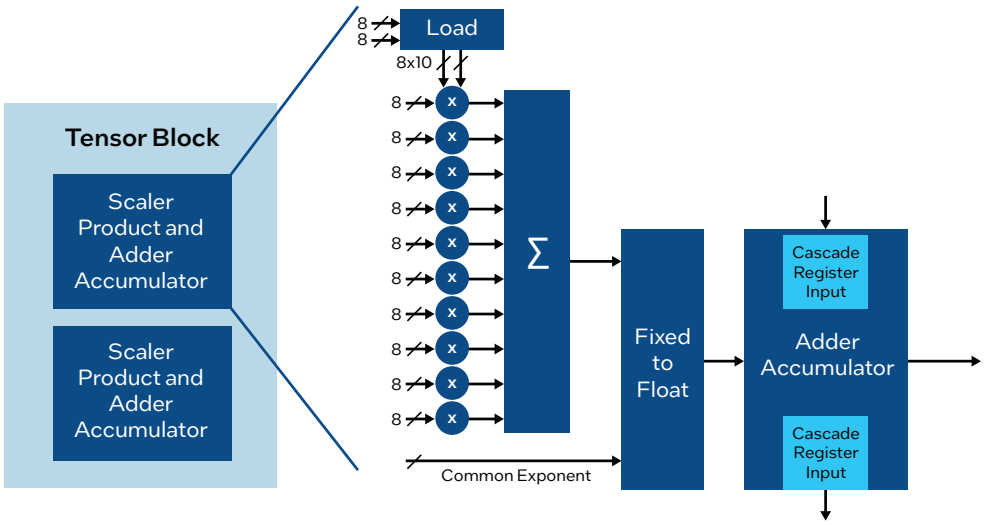


Figure 6. AI Tensor Block Diagram

The width of the input signals is 8 bits in fixed-point mode. An expansion of the input dynamic range is obtained by means of the common exponent that allows block floating point.

The tensor modes can use two dot product columns. It requires the weights (coefficients) to be pre-loaded into all columns. Only the activation values are input to the block during processing (the same activation values are used by all columns). A separate configuration of the DSP block can accumulate one or more results of several cascaded blocks in tensor mode. This allows for very efficient/performant implementation of convolution processing in CNN type networks.

The scalar product of the tensor mode allows 8 bits word operations in fixed-point operations, and it also allows block floating-point operations sharing a common exponent at the input. The two vectors of the scalar product are not fed both in parallel to ensure an efficient silicon implementation. Therefore, we can obtain a virtual bandwidth expansion by streaming the data in one dimension and preloading the weights in the other dimension.

The output of the scalar product (32 bits in fixed-point or single-precision floating-point) is cascaded to an additional adder, which allows scalar product on a super block cascading a chain of several dot products of 10 elements, executed in adjacent tensor blocks.

The adder can also be used for accumulation of sequential results of the dot product. It brings flexibility, allowing both fixed- and floating-point operation with 32 bits.

Agilex 5 FPGA tensor block embeds two of such scalar products.

The number of operations of each unit of the tensor block are nine additions + ten multiplications + one final addition/accumulation. So, each tensor block of an Agilex 5 FPGA offers 40 INT8 operations.

Although the Enhanced DSP with AI Tensor Block can always be instantiated in RTL as usual, it makes a lot of sense to use higher level tools especially in the AI context.

The Altera FPGA AI Suite will make heavy use of the Enhanced DSP with AI Tensor Block to implement highly efficient neural network implementations on this new architecture.

Performances

The following table summarizes the performances of the Agilex 5 FPGA E-Series and Agilex 5 FPGA D-Series in terms of number of 18x19 multipliers and TOPS (number of INT8 operations per second).

	A5E005B	A5E007B	A5E008B	A5E013B	A5E028B	A5E043B	A5E052B	A5E065B
18x19 multipliers	130	188	233	376	752	1,104	1,352	1,692
Peak TOPS	1.7	2.46	3.05	4.93	9.85	14.46	17.72	22.17

Table 3. Agilex 5 FPGA E-Series Device Group B

	A5E013A	A5E028A	A5E043A	A5E052A	A5E065A
18x19 multipliers	376	752	1,128	1,352	1,692
Peak TOPS	5.78	11.55	17.33	20.78	25.99

Table 4. Agilex 5 FPGA E-Series Device Group A

	A5D010	A5D025	A5D031	A5D051	A5D064
18x19 multipliers	552	1472	1840	2944	3680
Peak TOPS	8.48	22.61	28.26	45.22	56.22

Table 5. Agilex 5 FPGA D-Series

Device	Maximum TOPS
Cyclone® V FPGA	1
Arria® 10 FPGA	13
Stratix® 10 FPGA (2.8M)	23
Agilex™ 5 FPGA E-Series	26
Agilex™ 5 FPGA D-Series	56

Table 6. Performance comparison between Cyclone V FPGAs, Arria 10 FPGAs, Stratix 10 FPGAs, and Agilex 5 FPGA E-Series

Conclusion

As AI adoption grows, the range of applications and environments in which it runs—from endpoint devices to edge servers, to data centers—will become incredibly diverse. No single architecture, chip, or form factor will be able to meet the requirements of all AI applications. Infrastructure architects must have access to a variety of architectures.

Altera offers four types of silicon enabling the proliferation of AI: FPGAs, GPUs, and ASICs for acceleration, and CPUs for general-purpose computing. Each architecture serves unique needs, allowing infrastructure architects to choose the exact architecture they need to support any AI application.

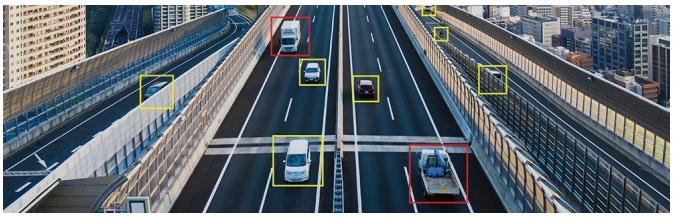
Agilex 5 FPGA E-Series with a breadth of compute types, optimized for power and performance, will always bring the right tools for the job at hand.

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

Medical Imaging / Analytics

- Tissue/Organ recognition
- Drug discovery
- Genome analysis

Figure 7. Enhanced DSP AI Tensor Block Target Applications

Gen-1079-1.3