WHITE PAPER

FPGA

Enabling High-Performance Floating-Point Designs

Unleash high-performance floating-point processing capabilities with Arria® 10 FPGAs and SoCs.

Author

Amulya Vishwanath
DSP Product Marketing Manager
Intel Programmable Solutions Group

Table of Contents

Introduction

A new generation of computationally intensive markets such as 5G, machine learning, data centers, and high-precision radar demand FPGAs and SoCs with enhanced floating-point processing for better numeric precision and lower power consumption. Arria® 10 FPGAs and SoCs are the industry's first FPGAs and SoCs that natively support single-precision floating-point DSP block mode as well as standard- and high-precision fixed-point computations using dedicated hardened circuitry. The single-precision floating-point DSP block mode is IEEE 754 compliant[1] and comprises of an IEEE 754 single-precision floating-point adder and IEEE 754 single-precision floating-point multiplier as shown in Figure 1. The new Arria 10 single-precision floating-point DSP block mode allows you to implement algorithms in floating point with efficiency and power comparable to fixed-point operations. The productivity benefits[2] with this DSP block architecture in Arria 10 FPGAs and SoCs makes them a compelling alternative to graphic processing units (GPUs) for high-performance computing applications.

Performance benchmarks

To demonstrate the single-precision floating-point processing capabilities of Arria 10 devices, this paper explores two digital signal proessing (DSP) applications:

- Poly-Phase Fast Fourier Transform (FFT)
- Single-Precision General Element Matrix Multiplication (SGEMM)

Based on the analysis of sustained DSP performance measured in floating-point operations per second (FLOPS) and power efficiency in FLOPS per watt, Arria 10 devices show a significant boost in performance for these two benchmarks.

Poly-phase FFT

An FFT is a common building block in many DSP applications including wireless and radar. High-precision radar systems require floating-point numeric precision for longer range and to detect low observable targets. Arria 10 floating-point FPGAs and SoCs enable this higher precision processing, which improves system dynamic range, reduces signal loss, and improves the signal-to-noise ratio. This benchmark uses an Intel®-developed poly-phase FFT that can sample data at a rate faster than the clock rate. The poly-phase FFT benchmark is a model-based design implemented in the MathWorks MATLAB*/Simulink* software using the programmable FFT IP core available in DSP Builder for Intel FPGAs.[3]

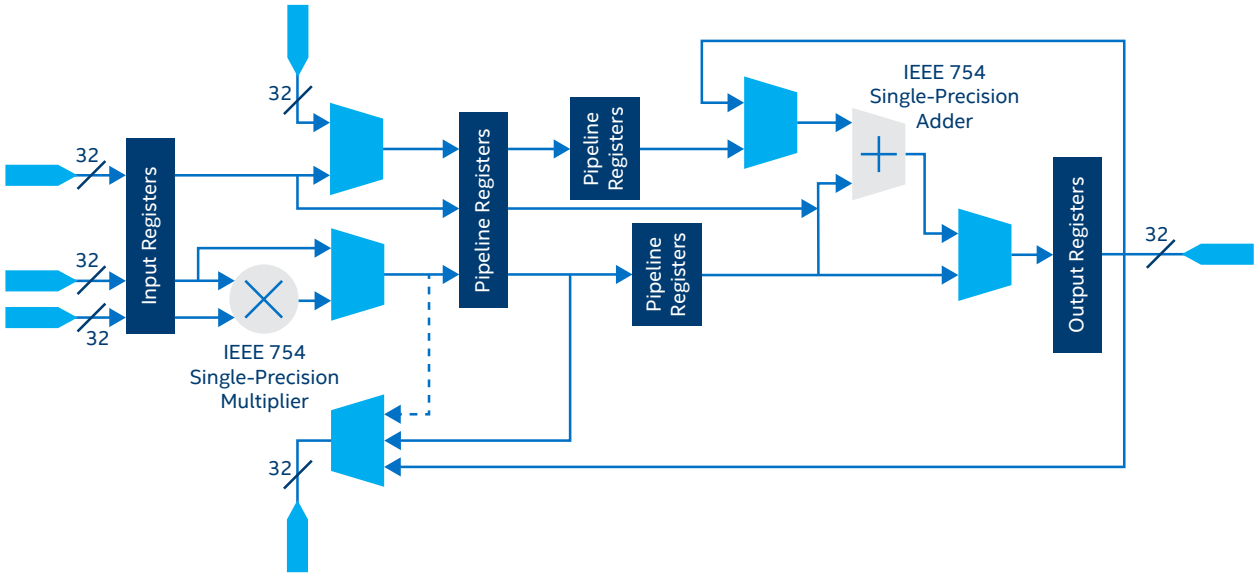

Figure 1. Single-Precision Floating-Point DSP Block Mode in Arria 10 Devices

Design setup

We measured performance metrics for three poly-phase FFT configurations (4K, 16K, and 64K) in hardware using the Arria 10 SoC Development Kit with a 10AX066N2F40E1SG production device in the -1 speed grade (0.95 V). The software package comprised of the MathWorks MATLAB/ Simulink software R2014a, DSP Builder for Intel FPGAs version 16.0, and Intel Quartus® Prime Pro Edition Software version 16.0.2.

Results

The poly-phase FFT benchmark demonstrates that Arria 10 FPGAs can deliver over 1 tera floating-point operations per second (TFLOPS) of sustained floating-point DSP performance for compute intensive applications such as high-precision radar with a power efficiency of around 40 GFLOPS per watt for the 4K FFT configurations shown in Figure 2. Table 1 shows the performance metrics and resource utilization for the three poly-phase FFT configurations.

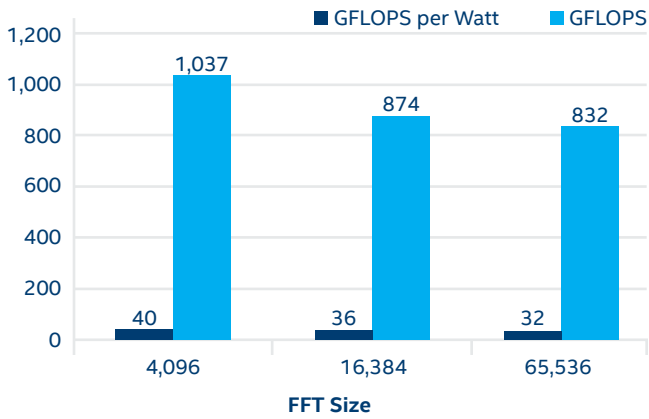

Figure 2. Power Efficiency (GFLOPS per Watt) and Sustained DSP Performance (GFLOPs) for 4K, 16K, and 64K Poly-Phase FFTs

Feature	Design 1	Design 2	Design 3
FFT size	4,096	16,384	65,536
Number of parallel phases	16	16	32
f_{MAX} (MHz)[§]	360	390	325
Number of instances	3	2	1
Throughput (FFTs per second)	4,218,750	761,719	158,691
Sustained DSP performance (GFLOPS)	1,037	874	832
Adaptive logic modules (ALMs) (including System in the Loop [SIL])	113,096 (45%)	89,602 (36%)	113,657 (45%)
DSP blocks (including SIL)	1,687 (100%)	1,384 (82%)	1,616 (96%)
M20K blocks (including SIL)	508 (24%)	617 (29%)	1,175 (55%)
Junction temperature (C)	64	61	63
Power (watts)	26	24	26
Power efficiency (GFLOPS per watt)	40	36	32

[§] Design Space Exploration (DSE) was used with the OPTIMIZATION_MODE set to "Aggressive Performance"

Table 1. Resource Utilization and Results for Three Poly-Phase FFT Configurations

SGEMM

SGEMM is a common operation used in linear algebra, neural networks, and machine learning applications. The SGEMM design developed using the Intel FPGA SDK for OpenCL™ [5] demonstrates a compute architecture with efficient data movement. The configurable routing eliminates results queue storage and data wait times. Figure 3 shows the routing advantage for computation and data movement in an array of two-dimensional (2D) processing elements (PE). The two one-dimensional (1D) feeder arrays each take a bit from load A and load B and the 1D drain array sends the feeder array data to drain C. Isolating the compute core from the feeder arrays provides efficient memory access control. The PE and host are autonomous; the channels move the data efficiently and minimize fanout. The Intel FPGA SDK for OpenCL[4] auto-translates the GEMM algorithm into reconfigurable hardware to perform the dot product operation. Figure 4 shows a four-vector dot product with accumulation.

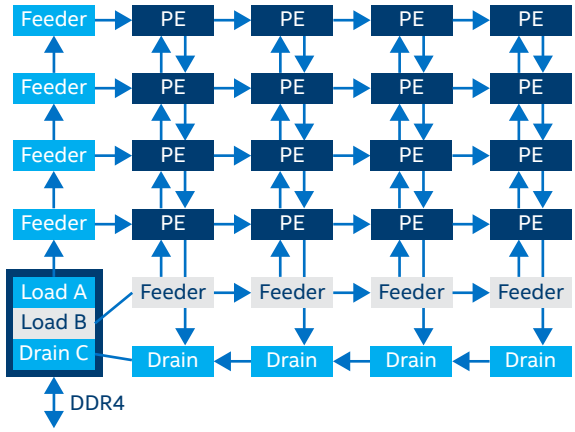

Figure 3. 2D Processing Element Array

Design setup

We obtained performance metrics for a 11R x 16C parameterizable array using an Arria 10 SoC Development Kit with a 10AX115S2F45I1SG production device in the -1 speed grade (0.95 V). The average power is 47 watts at 70 ºC. To compare performance with a previous device family without hardened floating-point capabilites, we implemented a single N-dimension kernel architecture for matrix multiplication on a Stratix® V SoC Development Kit featuring the 5SGSED8K2F40C2 production device in the -2 speed grade (0.95V). The power dissipation was 43.43 watts and the core temperature was 62 ºC. The software package comprised of the Intel FPGA SDK for OpenCL version 16.1, BSP version 14.1, and the Intel Quartus Prime Pro Edition software version 16.1.

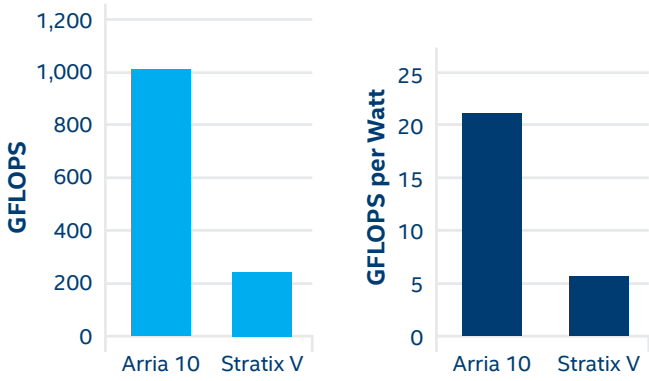

Figure 5. Comparison of Sustained DSP Performance (GFLOPS) and Power Efficiency (GFLOPS per Watt) for GEMM Function

Results

The Arria 10-based GEMM function achieved close to a four-fold increase in DSP performance (GFLOPS) and power efficiency (GFLOPS per watt) compared to the GEMM function compiled on a Stratix V device, as shown in Figure 5. This performance boost is due to efficient data movement and compute structure of the SGEMM function implemented using the Intel FPGA SDK for OpenCL and Arria 10 device. Tables 2 and 3 show the Arria 10 and Stratix V resource utilization and results, respectively.

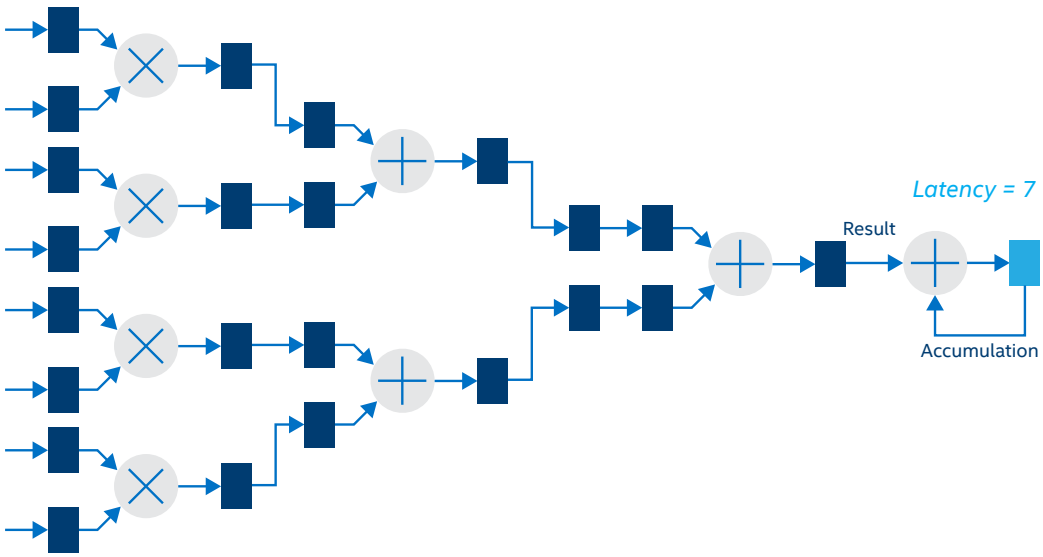

Figure 4. Four-Vector Dot Product with Accumulation

[5] OpenCL and the OpenCL logo are trademarks of Apple Inc. used by permission by Khronos

Kernel f_{MAX} (MHz)	ALMs	Registers	DSP Blocks	M20K (RAM Blocks)	Sustained DSP Performance (GFLOPS)	Power Efficiency (GFLOPS per Watt)
365.36	234,420 (55%)	710,542	1,408 / 1,518 (93%)	2,176 / 2,713 (80%)	1,010	21.5

Table 2. Arria 10 SGEMM + BSP Resource Utilization and Results

Kernel f_{MAX} (MHz)	ALMs	Registers	DSP Blocks	M20K (RAM Blocks)	Sustained DSP Performance (GFLOPS)	Power Efficiency (GFLOPS per Watt)
173.13	219,953 (84%)	445,089	768 / 1,963 (39%)	1,334 / 2,567 (52%)	260.85	6.006

Table 3. Stratix V GEMM + BSP Resource Utilization and Results

Conclusion

Arria 10 FPGAs and SoCs demonstrate higher performance and performance per watt than previous generations for real-world floating-point applications.

- The Arria 10-based poly-phase FFT design received a significant performance boost for a given power budget, achieving 832 to 1,037 GFLOPS sustained DSP performance for 4K, 16K, and 64K poly-phase FFTs, respectively, with power efficiency from 32 to 40 GFLOPS per watt.
- The Arria 10-based SGEMM design achieved 1,010 GFLOPS sustained DSP performance for a 11R x 16C parameterizable array, which is a 4 four-fold increase compared to Stratix V devices.

Arria 10 devices can achieve single-precision sustained DSP performance up to 1.5 TFLOPS to support next-generation computationally intensive applications.

Note:
The designer needs access to the original Intel design files, software, and hardware platform to achieve these benchmark results.

References

[1] http://www.bogdan-pasca.org/resources/publications/2015_langhammer_pasca_fp_dsp_block_architecture_for_fpgas.pdf
[2] https://www.altera.com/en_US/pdfs/literature/po/bg-floating-point-fpga.pdf
[3] https://www.altera.com/products/design-software/model---simulation/dsp-builder.tablet.html
[4] https://www.altera.com/products/design-software/embedded-software-developers/opencl/overview.html

Where to Get More Information

For more information about Intel and Arria 10 FPGAs, visit **https://www.altera.com/products/fpga/arria-series/arria-10/overview.html**

For more information on floating-point processing with Arria 10 devices, visit **https://www.altera.com/products/fpga/features/dsp/arria10-dsp-block.html**