WHITE PAPER

GPU vs FPGA Performance Comparison

Image processing, Cloud Computing, Wideband Communications, Big Data, Robotics, High-definition video…, most emerging technologies are increasingly requiring processing power capabilities. The technology selection for each application is a critical decision for system designers. Being GPU power the conservative approach to scale processing capacity, using FPGA for software acceleration is becoming the best option for an increasing number of applications.

This paper evaluates the 2016's state-of-the-art technology for both GPU and FPGA devices, and performs a qualitative and quantitative comparison. The analysis must be considered as a preliminary guideline for technology selection. Some key parameters cannot be directly compared, and different interpretations of the results can be derived introducing other variables.

What is important for your design?

The short answer to this question is that FPGAs are power efficient and GPUs are cost efficient (Figure 1); but taking a design decision based on simple rule-of-thumbs is usually risky.

FPGAs are designed to perform concurrent fixed-point operations with a close-to-hardware programming approach, while GPUs are optimised for parallel processing of floating-point operations using thousands of small cores. Most of the differences between the two technologies, and their applicability to software acceleration, are herein derived from these high-level architectural definitions.

Comparing processing capabilities is not straightforward. GPUs performance is measured in GFLOPS; they are capable to accelerate native CPU algorithms based on floating-point operations, simplifying code adaptation from high-level programming languages. On the other hand, FPGAs processing power is measured in GMACS; they require designing algorithms for fixed-point data types to maximize efficiency, taking massive benefit of bit-wise operations.

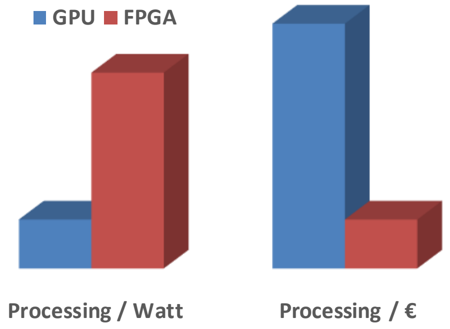

Figure 1. Processing Efficiency

GPUs gain advantage when considering total floating-point processing power, development effort, device cost, and flexibility. However, FPGA is starting to be the logical choice for an increasing number of applications.

FPGA also provides huge processing capabilities with a great power efficiency, reducing thermal management and space requirements. This feature allows the integration of acceleration hardware in small housings, on-board equipment, or extreme temperature environments.

Interfaces are another FPGA's strong point. Being GPUs limited to PCIe, interfacing with devices implementing any other standard or custom interfaces will require additional electronics. FPGA has a huge interface flexibility, recently improved by the integration of programmable logic with CPUs and standard peripherals in SoC devices.

Latency is another parameter to be considered when running processing software in specialised hardware. GPUs improve CPUs performance, but FPGA provides deterministic timing in the order of nanoseconds. This is especially important for encryption, audio coding, network synchronisation or control applications that need to manage small and well-known latencies.

Regarding the price of a software acceleration solution, GPUs are cost efficient both in development and hardware installation. FPGAs require specialised design engineers with knowledge in a number of different technology areas (electronics, HLD, algorithms, communications, etc.). The price comparison for mid-range devices is not a drama considering FPGA power efficiency. The real issue is the engineering effort, which is being mitigated introducing new development environments to add abstraction layers to the lowest-level design. In addition, autocoding techniques are starting to reduce implementation times, although they do not significantly reduce the required know-how.

Finally, RTL-based design enables FPGA to be used as technology path to ASIC development.

Figure 2 and Table 1 summarise this qualitative analysis for a faster understanding of the technology trade-offs.

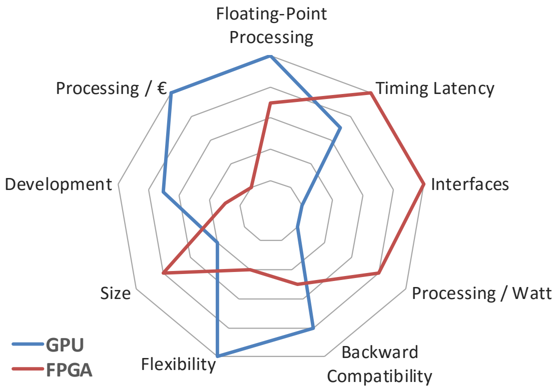

Figure 2. GPU vs FPGA Qualitative Comparison

Feature	Analysis	Winner
Floating-point Processing	The total floating-point operations per second of the best GPUs are higher than the FPGAs' with the maximum DSP capabilities.	GPU
Timing Latency	Algorithms implemented into FPGA provide deterministic timing, with latencies one order of magnitude less than GPUs.	FPGA
Processing / Watt	Measuring GFLOPS per watt, FPGAs are 3-4 times better. Although still far away, latest GPU products are dramatically improving the power burning.	FPGA
Interfaces	GPUs interface via PCIe, while FPGA flexibility allows connection to any other device via -almost- any physical standard or custom interface.	FPGA
Backward Compatibility	Software developed for older GPUs will work in the new devices. FPGA HDL can be moved to newer platforms, but with some reworking.	GPU
Flexibility	FPGA lacks flexibility to modify the hardware implementation of the synthesized code, being a no-problem issue for GPUs developers.	GPU
Size	FPGA's lower power consumption requires less thermal dissipation countermeasures, implementing the solution in smaller dimensions.	FPGA
Development	Many algorithms are designed directly for GPUs, and FPGA developers are difficult and expensive to hire.	GPU
Processing / €	Mid-class devices can be compared within the same order of magnitude, but GPU wins when considering money per GFLOP.	GPU

Table 1. Evaluation of FPGA and GPUs characteristics

GPU performance in numbers

A selection of 28nm graphic cards and FPGA devices are analysed and used for comparison purposes. GPUs performance is derived from commercial graphic cards characteristics.

Table 2 lists a selection of the best cards for the money as representative samples for the 2016's state-of-the-art technology combining older models with newer flagship graphic cards.

GPUs price ranges from less than 100€ to more than 600€, showing huge processing powers over 7,000 GFLOPS for single precision operations, having no rival when evaluating floating-point computational capacity of a single chip.

Since manufacturers are only providing the required power supply, maximum consumption are derived from user analysis in stress conditions where top processing performance is achieved. Burning up to 360W for the high-end models, it demands careful cooling designs (heatsink, fans), heavier power supplies, and users ready-to-pay an increasing electricity bill.

Price efficiency is similar for all the analysed models, ranging from 0.07 to 0.12 €/GFLOPS. High-end graphic cards, however, show improved power efficiency, achieving up to 23 GFLOPS/W. In fact, energy usage is currently the most important constraint to continue increasing graphic cards maximum processing capability, and it is expected to be dramatically improved in the following years.

But as per 2016 state-of-the-art technology, energy burning is the main draw-back of GPUs for software acceleration purposes in a number of applications, and it must be considered together with the relative low price and the huge total processing power. At the end, a high power consumption involves that they cannot be installed in systems with demanding power, space or temperature requirements.

		Nvidia GeForce GT 730	AMD Radeon R7 360	Nvidia GeForce GTX 970	Sapphire Radeon R9 390	Radeon R9 390X	Sapphire Radeon R9 Fury X	Nvidia GeForce GTX 980 Ti
Price (approx.)		80 €	120 €	250 €	400 €	420 €	600 €	700 €
Processing Power	Single	693 GFLOPS	1,612 GFLOPS	3,494 GFLOPS	5,120 GFLOPS	5,913 GFLOPS	7,168 GFLOPS	5,632 GFLOPS
	Double	32 GFLOPS	100 GFLOPS	109 GFLOPS	640 GFLOPS	739 GFLOPS	448 GFLOPS	176 GFLOPS
Technology		28 nm	28 nm	28nm	28nm	28nm	28nm	28nm
GPU		GK208 (Kepler)	Tobago (GCN 1.1)	GM204 (Maxwell)	Grenada (GCN 1.1)	Grenada (GCN 1.1)	Fiji (GCN 1.2)	GM200
Core Clock		902 MHz	1050 MHz	1050 MHz	1000 MHz	1050 MHz	1050 MHz	1000MHz
Power Consumption *Stress Test*		93 W	100 W	242 W	323 W	363 W	358 W	250 W
Price Efficiency		0.10 €/GFLOPS	0.07 €/GFLOPS	0.07 €/GFLOPS	0.08 €/GFLOPS	0.07 €/GFLOPS	0.08 €/GFLOPS	0.12 €/GFLOPS
Power Efficiency		7 GFLOPS/W	16 GFLOPS/W	14 GFLOPS/W	16 GFLOPS/W	16 GFLOPS/W	20 GFLOPS/W	23 GFLOPS/W

Table 2. Graphic Cards Characteristics Compared

FPGA performance in numbers

FPGA technology is evolving fast, with new models implementing 16nm and 20nm, and increasing clock speeds, interfaces bandwidth, on-chip RAM, and fixed- and floating-point processing capacity. For this analysis, we stick to 28nm device not only for a fair technology trade-off, but also for a reasonable price comparison.

Table 3 shows a selection of Xilinx 7-Series devices, including Zynq and representative FPGA integrated circuits. Zynq SoC combines the flexibility of a CPU with the processing power of the FPGA, lowering the entry barrier for software acceleration using programmable logic. It supports embedded operating systems and standard peripherals, being the perfect candidate for mitigating technology draw-backs and opening the FPGA world to a higher number of sectors, applications and end-users.

SoC allows users to implement in the FPGA only the high computation load algorithms and tasks, with a similar approach to how GPUs interface with CPUs for software acceleration.

As mentioned before, estimating the floating-point processing capacity of a device that is designed for fixed-point operations and HDL programming is not direct. The result will depend on the implementation approach and the type of algorithm.

In this paper, processing power for each selected device is estimated considering the peak DSP performance in GMACS, and the single precision floating-point performance and power efficiency claimed by Xilinx.

As expected, the results show a decent total processing power, with an excellent power efficiency always higher than 70 GFLOPS/W. This enables the implementation of current FPGAs devices into small and efficient hardware with more than reasonable thermal dissipation and cooling requirements.

An interesting conclusion is derived from analysing price efficiency, which reaches 0.29 €/GFLOPS for mid-class FPGAs as the Artix-7 200T. Although still far from GPUs mass-production cost, it is a comparable cost budget. This is not true for the costly high-end devices of the family, where prices are a clear draw-back for many software acceleration applications.

Table 3 only considers the device price, which is usually the cost driver for FPGA-based solutions.

	Zynq SoC Z-7020	Zynq SoC Z-7100	Artix-7 200T	Kintex-7 480T	Virtex-7 690T
Price (approx.)	100 €	3,000 €	190 €	2,500 €	11,200 €
Dual ARM® Cortex™-A9 MPCore	Yes	Yes	-	-	-
Programmable Logic Cells	85,000	444,000	215,360	477,000	693,120
Programmable DSP Slices	220	2,020	740	1,920	3,600
Peak DSP Performance	276 GMACs	2,622 GMACs	929 GMACs	2,845 GMACs	5,335 GMACs
Processing Power - single	180 GFLOPS	1,560 GFLOPS	648 GFLOPS	1,800 GFLOPS	3,120 GFLOPS
Technology	28 nm	28 nm	28 nm	28 nm	28nm
PCIe Interface	-	x8 Gen2	x4 Gen2	x8 Gen2	x8 Gen3
Power Consumption	2.5 W	20 W	9 W	25 W	40 W
Price Efficiency - single	0.56 €/GFLOPS	1.92 €/GFLOPS	0.29 €/GFLOPS	1.39 €/GFLOPS	3.59 €/GFLOPS
Power Efficiency - single	72 GFLOPS/W	78 GFLOPS/W	72 GFLOPS/W	72 GFLOPS/W	78 GFLOPS/W

Table 3. Xilinx SoC and FPGA Characteristics Compared

Comparing Key Performance Indicators

Table 4 mixes together representative graphic cards and FPGA devices for a final comparison of the most important parameters quantified in this paper.

The graphical representation of Figure 3 clearly shows the huge processing power of GPUs, and the great power efficiency of the current FPGA technology.

Introducing the price in the equation shows competitive mid-class FPGA, but still far from the GFLOPS per Euro of the mass-market graphic cards.

The cost of the high-end FPGAs limits them to specific niche applications, while the power burning of the high-end GPUs avoids using them for a number of markets and critical systems.

Model		Processing Single	Power Efficiency	Approx. Price	Price Efficiency
GPU	GeForce GT 730	0.69 TFLOPS	7 GFLOPS/W	80 €	0.10 €/GFLOPS
	Radeon R9 390X	5.91 TFLOPS	16 GFLOPS/W	420 €	0.07 €/GFLOPS
	Radeon R9 Fury X	7.17 TFLOPS	20 GFLOPS/W	600 €	0.08 €/GFLOPS
FPGA	Artix-7 200T	0.65 TFLOPS	72 GFLOPS/W	190 €	0.29 €/GFLOPS
	Kintex-7 480T	1.80 TFLOPS	72 GFLOPS/W	2,500 €	1.39 €/GFLOPS
	Virtex-7 690T	3.12 TFLOPS	78 GFLOPS/W	11,200 €	3.59 €/GFLOPS

Table 4. GPU vs FPGA - Key Performance Indicators

Conclusion

This paper compares key performance indicators of 28nm GPUs and FPGAs devices, focussed on their applicability to software acceleration purposes. The trade-off analysis is a preliminary guideline for technology selection, a design decision that is a fundamental performance and cost driver.

In addition to processing capacity, power efficiency and cost, other parameters such as latency, development effort, flexibility or interfaces are discussed.

The analysis shows cost efficient GPUs with huge floating-point processing capacity, and power efficient FPGAs with flexible interfaces and deterministic latency. The availability of SoC integrating CPUs with FPGA and standard peripherals is a step forward mitigating FPGA draw-backs, i.e., high development efforts, and limited flexibility and backward compatibility.

In the near future, we can envisage GPUs with improved power efficiency, and FPGAs with increased computational capacity and lower cost. After that, technology selection will be focussed only on architectural design considerations and specific application requirements, which definitely will be good news for the engineer's community.

But for the time being, the selection of the device will remain linked to the end-user application, available budget, and development capacity.

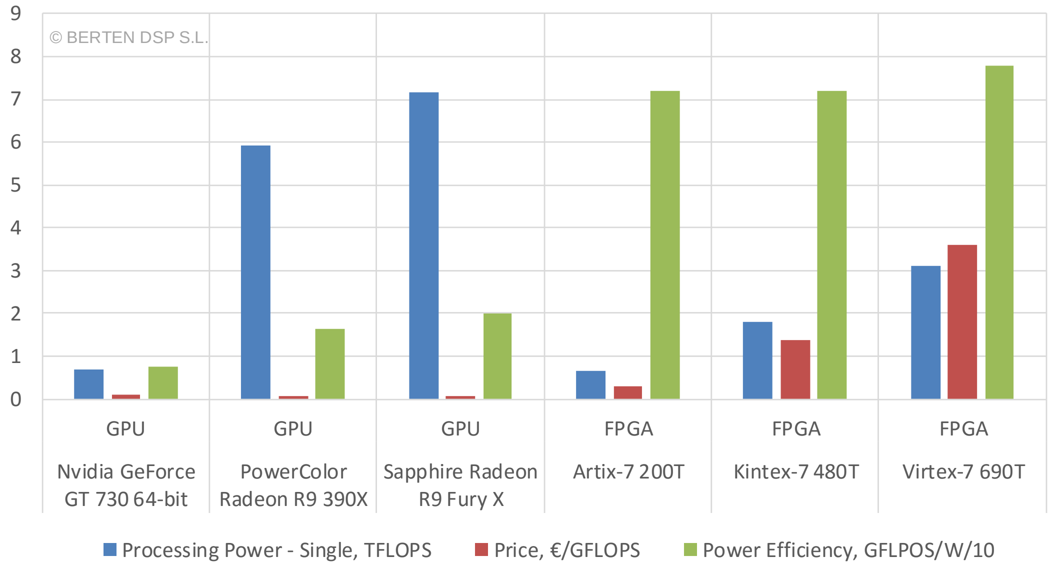

Figure 3. GPU vs FPGA Performance Comparison

All the information provided in this paper is believed accurate and reliable. BERTEN does not accept any responsibility or liability for the accuracy, content, completeness, or reliability of the information contained in this document.
The use of the content of this paper is permitted if properly cited or if the company name (BERTEN) is clearly referred.
Contact us at berten@bertendsp.com for any comment, suggestion, or question.