FPGAs for Advanced Video Processing Solutions (WP022)

Achronix® Data Acceleration

November 17, 2020 **White Paper**

Advanced network infrastructure is deployed not only to address the huge surge in the volume of data transfers, but also to enable data processing in different parts of the network, for instance at the edge, in the core, and in the cloud to name a few. Unsurprisingly, most of the data is either video or images, which is growing exponentially and will continue to do so in the years to come. More computing resources are required to accommodate the massive growth in data (figure following (see page 2)).

Since the types of applications are diverse, there is a wide variety of video or image processing workloads present in data centers. ASIC-based solutions, typically offer higher performance, but cannot be upgraded to support future algorithms. CPU-based solutions are much more flexible; however, clock frequencies has plateaued and drastic improvements in processor performance are not possible anymore. GPUs are another candidate to provide video/image processing solutions, but consume significantly higher power than FPGA-based solutions. FPGAs are an attractive option for video processing and compression since they provide the balanced resources required for implementing innovative video processing algorithms. In addition, FPGAs deliver a flexible solution that shortens time to market and enables continuous upgrades and new feature deployment over the lifetime of the solution.

Table 1: ***Growth in Internet Users and Traffic (Source: Cisco)***

Year	Internet Users	Devices & Connections	Broadband Speeds	Amount of Video Traffic
2017	3.4 billion	18.0 billion	39.0 Mbps	75% of total
2022	4.8 billion	28.5 billion	75.4 Mbps	83% of total

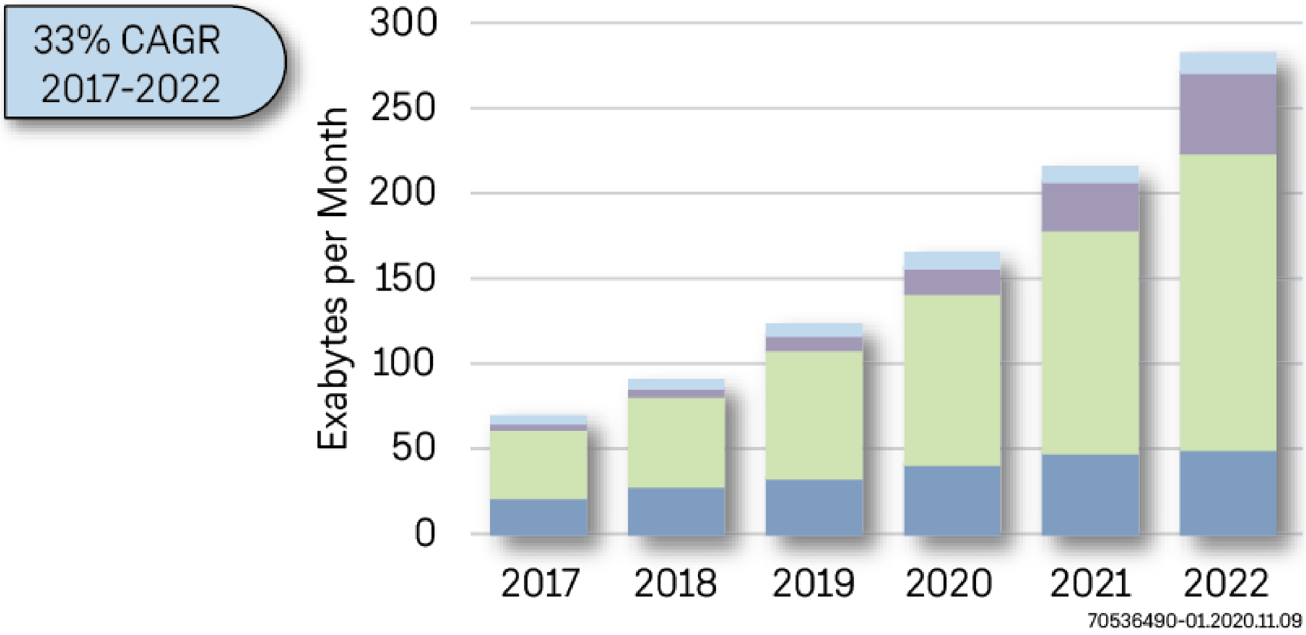


Figure 1: ***Global Internet Video Traffic (Source: Cisco)***

Examples of FPGA-Based Video Solutions

This white paper introduces three typical video applications to show the advantages of FPGA-based solutions for the broadcast industry. These advantages include reduced processing time, lower power consumption and lower costs for service providers and end users.

This white paper describes the benefits of FPGA-based solutions in the following three applications:

- Video streaming
- Video content creation using video editing software
- AI and deep learning – image recognition is a major portion of this application and requires high-performance computing resources

Video Streaming

The demand for video transcoding has dramatically increased to make streaming fast and efficient. Most of the current offerings leverage a software-based approach that can not keep up with the processing requirements for high-bandwidth, broadcast-qualtiy video streaming. Video streaming and/or cloud service providers are challenged by low throughput, high power consumption, long latency and large footprint for their software-based solutions. According to a report from Cisco entitled, “Cisco Visual Networking Index: Forecast and Trends, 2017–2022 White Paper”, video streaming traffic is increasing and will account for 82% percent of whole internet traffic by 2022. The amount of video data traffic will grow steadily year over year in all applications, including video on demand, live streaming and video surveillance.

The rise of video streaming applications such as Netflix and YouTube are driving the demand in video transcoding. The most remarkable differences between traditional broadcasting and video streaming are the amount of content and the number of channels. To support the wide variety of receiving devices, ranging from PCs to smartphones, content must be transcoded to different resolutions and compression formats on the fly. As a result, video streaming consumes enormous amount of computing resources.

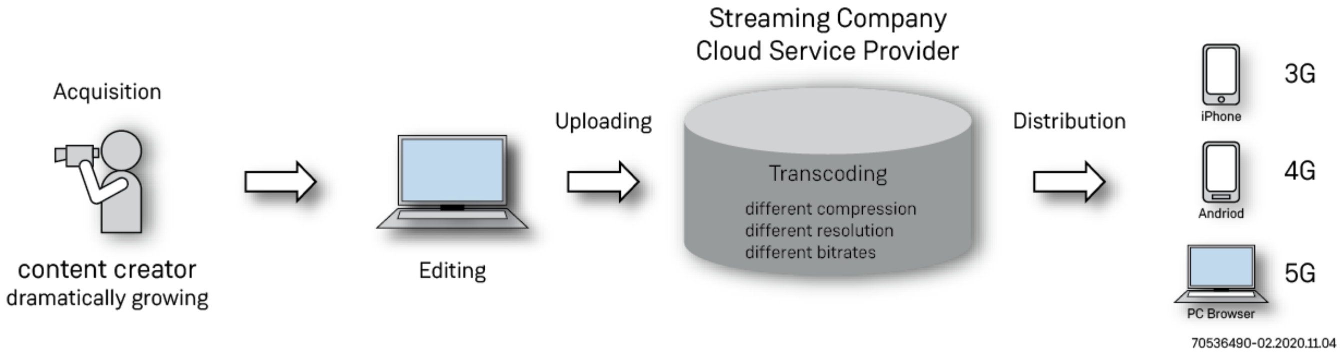


Figure 2: ***Video Transcoding Workflow***

Streaming media and cloud service providers need a solution to ease the strain on compute demand. Achronix Speedster®7t FPGAs hosting IBEX (see page 12) state-of-the-art video processing IP are able to solve this significant issue. This FPGA-based solution can provide high throughput, low power consumption and a small footprint system without sacrificing flexibility. Although powerful, ASIC-based solutions are only able to support the set of features defined at design time and cannot support in-field updates.

Video Content Creation

In the past, HD resolution was dominant in video content creation. Recently, the standard resolution has changed to 4K, pushing to 8K, making video encoding or decoding challenging. The major compression formats used for these higher resolutions are Apple ProRes, Avid DNx and SONY XAVC. Since these formats are proprietary, there are no ASICs or GPUs that supports these formats natively and CPUs deliver poor performance. As a result, FPGAs are the best solution for video content creation at these higher resolutions.

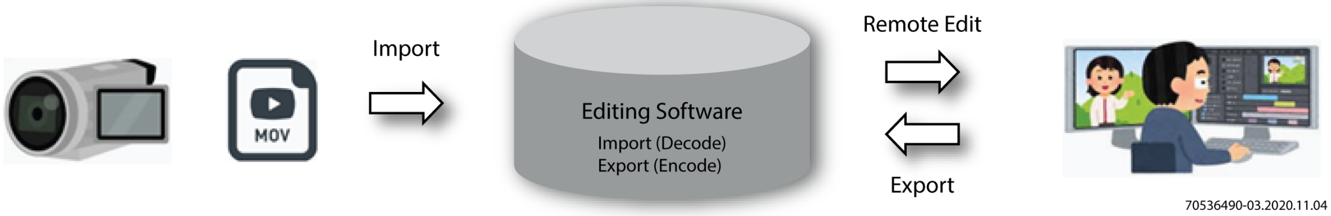


Figure 3: ***Video Editing Workflow***

In a new trend, the concept of remote post-production is becoming common. However, existing PCs do not have sufficient power to process high resolution content (for e.g. 8K) in real time. As a result, editors have started to use cloud infrastructure to gain access to better compute performance. Additionally, COVID-19 has accelerated this trend due to the need for social distancing. A cloud plus FPGA-based solution provides great benefit to editors. Achronix Speedster7t FPGAs' architectural innovations, such as 2D network on chip (NoC), make them uniquely suited to accelerate both the encoding and decoding algorithms.

AI and Deep Learning

Artificial intelligence, machine learning and deep learning are well-known areas that have progressed rapidly in the last few years. Alongside these areas, image recognition is emerging as a major new field that benefits because of AI/ML innovations. For example, advanced driver-assistance systems (ADAS) process captured images using deep learning algorithms. The dash camera installed in the car records the video using H.264 compression. Then the video stream is transcoded to a suitable image format such as JPEG or PNG for deep learning image recognition. Depending on the use case, frame dropping, changing resolution, or other image processing tasks are done at the same time.

There are similar use cases in security cameras in retail or luggage sorting in logistics, where the data flow is the same as the above example — the cameras records video with relatively high-compression formats, such as H. 264 or H.265, transmiting the encoded video stream to the cloud or data center. On the cloud side, the video stream is transcoded from the original format to a suitable format for deep learning, converting the video file to a bank of images.

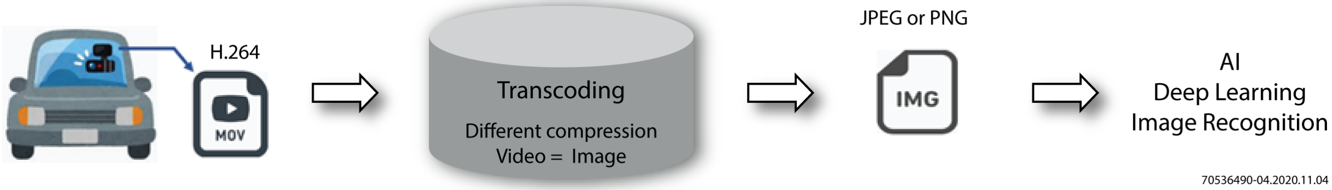


Figure 4: ***Typical Deep Learning Image Data Flow***

Historically, FPGAs has been good at transcoding movies to images. In addition, running image pre-preprocessing using deep learning algorithms hosted in FPGAs not only improves the throughput, but also reduces the volume of data transaction at the system level. Achronix Speedster7t architecture, with its dedicated machine learning processors (MLPs), make it ideal for implementing custom as well as established, deep learning algorithms.

Performance of Representative Video Use Cases on FPGAs

Some of the key comparison metrics between FPGA and CPU implementation of above three representative use cases are shown in the table below.

Table 2: *FPGA vs CPU Performance Comparison*

Compression		Encoder	Decoder	Memo
H.264, H.265	Typical Parameters	↓	↔	Typical parameters mean 8 bits, 420, 2K. Intel QSV, GPU hard core is still powerful especially the encoder side.
	Minor Parameters	↔	↑	Minor parameters mean 10 bits, 422, 3K or 5K. Intel QSV, GPU hard core does not support it.
Intra Frame		↑	↑	Proprietary codec such as Apple ProRes, Avid DNx, SONY XAVC, Panasonic AVC-Intra. Image format such as JPEG, PNG.

Table Note

↑ FPGA deliver better performance.

↔ FPGA and CPU deliver equivalent performance, but an FPGA is the preferred solution to offload the CPU.

↓ FPGA and CPU deliver equivalent performance, but a CPU is the preferred solution.

Video Streaming

In video streaming applications, the prevalent compression format is H.264 or H.265 because the end-point (receiving) devices support these formats natively. The parameters such as bit depth or chroma and resolution are typically 8 bits, 4:2:0 and 1920×1080 or 1280×720. On the decoder side, FPGA-based implementations provide higher throughput than CPU-based systems. In the data plane, an FPGA is more efficient because a CPU is typically under-utilized if it used for any data related tasks besides data processing. However, on the encoder side, the hardened CPU encoder cores are specifically designed for these typical parameters and provide adequate performance.

To get best of both worlds, a combined FPGA and CPU solution with an FPGA taking care of heavy workloads is ideal. Functions that are efficient on FPGA can be ported to run on reconfigurable hardware. For example, the motion estimation algorithm is a suitable workload for an FPGA. On the other hand, the bit-rate control algorithm is more easily handled by CPUs.

Some service providers require the same video quality and streaming format as implemented in their software solutions for x264. A combined FPGA and CPU solution is effective in satisfying these requirements. Using this approach, each function is appropriately assigned with the heavier processing loads moved to the FPGAs. Both video quality and streaming format are similar or better than software-only solutions, but the encoding time is substantially reduced.

The table below lists the x264 profiling results using this approach with the top row showing the motion estimation function (x264_8_me_search_erf) hosted on the FPGA). Motion estimation is as one of the heaviest workloads for CPU and accounts for 21.2278% of total processing time.

Table 3: ***x264 Profiling Result (Obtained via Profiling Software)***

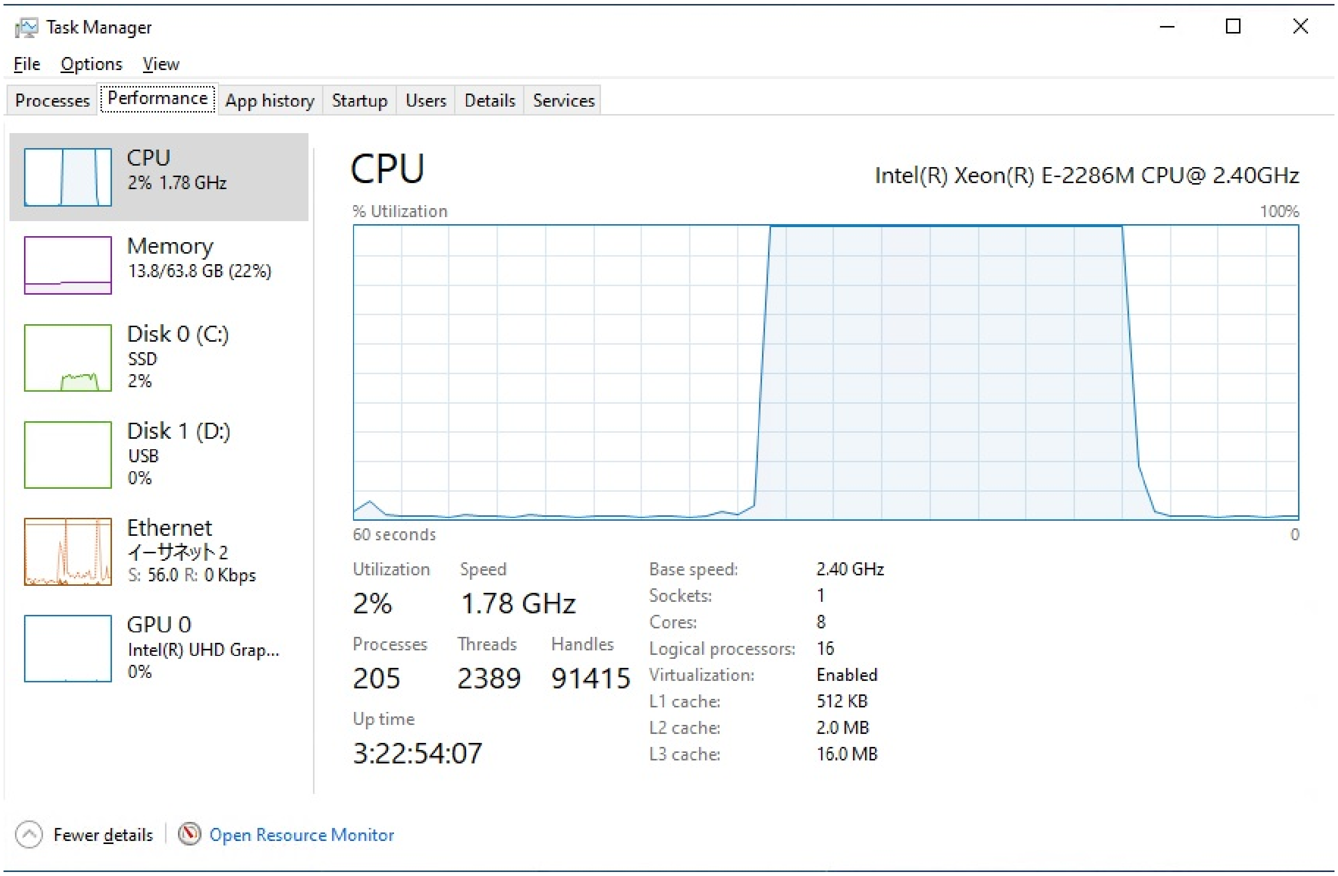
Samples	Percentage	Symbol Name
3679706	21.2278	x264_8_me_search_ref
2078100	11.9883	x264_8_pixel_ads_mvs_ssse3
1453998	8.3880	x264_8_pixel_sad_x3_8x16_sse2
1176121	6.7849	x264_8_picel_sad_x3_16x16_avx2
1156301	6.6706	x264_8_pixel_sad_x3_8x8_sse2
1095731	6.3211	x264_8_pixel_ads2_avx2
868943	5.0128	x264_8_pixel_sad_x3_16x8_avx2
779812	4.4986	x264_8_pixel_ads1_avx2
318990	1.8402	x264_8_pixel_ads_avx2
275943	1.5919	x264_8_quant_4x4_trellis
255712	1.4752	x264_8_trellis_cabc_4x4_psy_ssse3
231397	1.3349	x264_8_pixel_satd_8x8_interval_avx2
187422	1.0812	x264__8_mc_chroma_avx2
168559	0.9724	x264_8_pixel_satd_16x8_interval_avx2
168484	0.9720	x264_8_pixel_sad_8x8_mmx2

Video Content Creation

There is a wide variety of the compression formats supported by video editing software used for content creation, among these are Apple ProRes, Avid DNx, Sony XAVC and Panasonic AVC-Intra. These formats feature a proprietary compression scheme based on an intra-frame structure. Additionally, there are formats that support RAW modes, for example, Apple ProRes RAW, RED RAW, ARRI RAW, Blackmagic RAW, that are supported by camera manufacturers. Due to the ever changing nature of these formats (plus new and emerging ones), an ASIC-based solution is just not practical; therefore, FPGA-based solutions are required.

In the past, the major resolution was HD/2K, and CPUs had sufficient speed to handle these video streams. However, as 4K or 8K resolutions become more common, a CPU plus software only solution can no longer deliver real-time processing. On the other hand, FPGA-based solution can easily handle 4K and 8k resolutions in real-time.

Internal benchmarks determined that an FPGA-based solution is five times faster than the latest CPU plus software solution when compared to even a mid-class FPGA. While a GPU can deliver similar performance as an FPGA, it is at much higher power consumption and much larger solution footprint.



CPU
2% 1.78 GHz

Memory
13.8/63.8 GB (22%)

Disk 0 (C:)
SSD
2%

Disk 1 (D:)
USB
0%

Ethernet
イーサネット 2
S: 56.0 R: 0 Kbps

GPU 0
Intel(R) UHD Grap...
0%

CPU

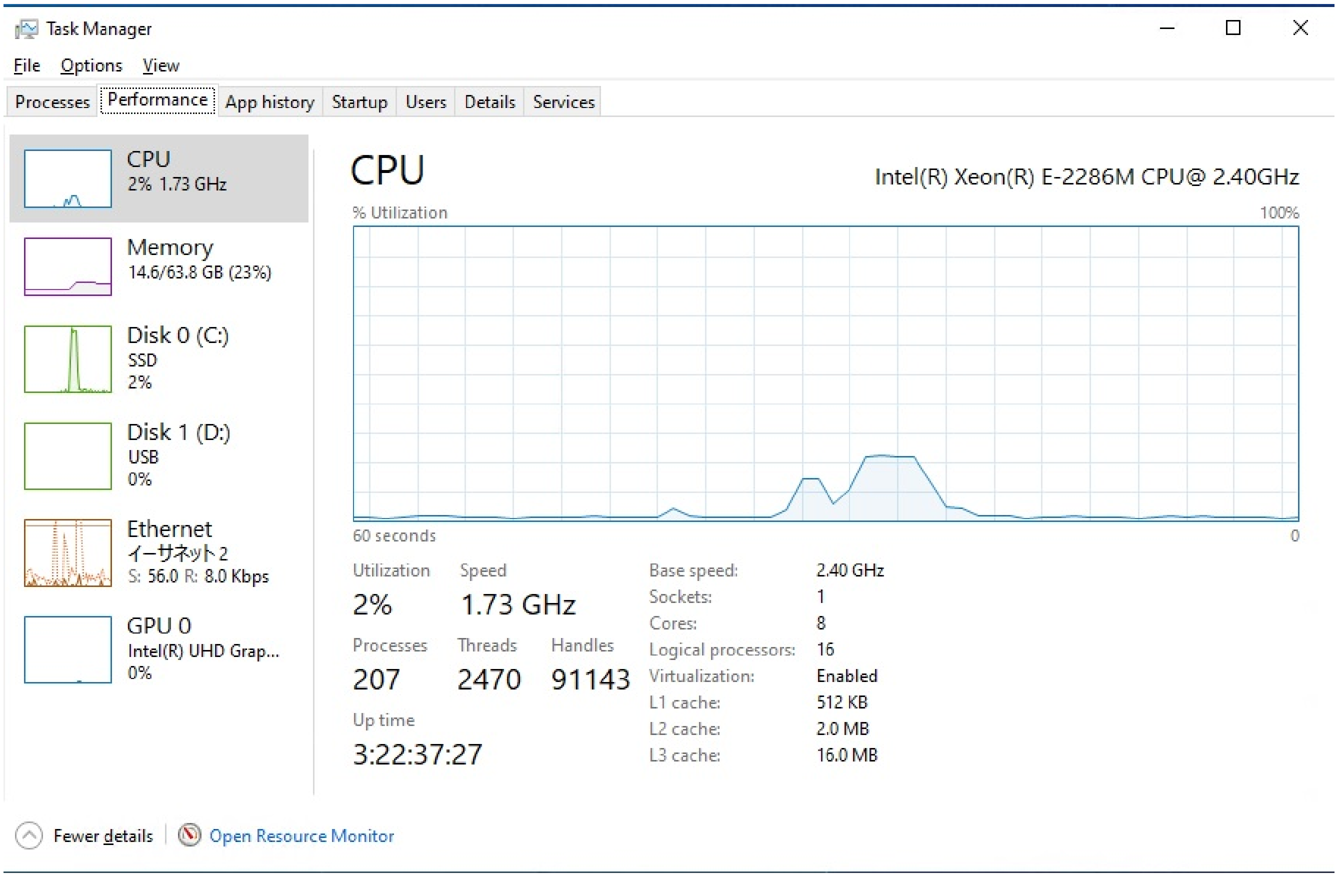
Intel(R) Xeon(R) E-2286M CPU@ 2.40GHz

Utilization 2%
Speed 1.78 GHz
Processes 205
Threads 2389
Handles 91415
Up time 3:22:54:07

Base speed:	2.40 GHz
Sockets:	1
Cores:	8
Logical processors:	16
Virtualization:	Enabled
L1 cache:	512 KB
L2 cache:	2.0 MB
L3 cache:	16.0 MB

Figure 5: ***CPU only (Without FPGA Offload)***

The benefit of the FPGA solution is not only acceleration, but also in keeping the CPU less busy. In a CPU only solution, all the CPU cycles are consumed by encoding 4K or 8K content. Whereas using an FPGA to offload the encoding task frees up CPU cycles. As a result, an FPGA accelerator provides the best solution for this application, increasing a video editor's productivity by reducing the processing time required for 4K and 8K video production.



CPU
2% 1.73 GHz

Memory
14.6/63.8 GB (23%)

Disk 0 (C:)
SSD
2%

Disk 1 (D:)
USB
0%

Ethernet
イーサネット 2
S: 56.0 R: 8.0 Kbps

GPU 0
Intel(R) UHD Grap...
0%

CPU
Intel(R) Xeon(R) E-2286M CPU@ 2.40GHz

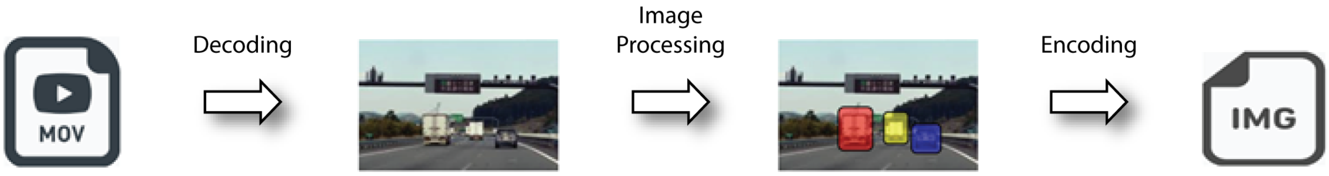
Utilization 2%
Speed 1.73 GHz
Processes 207
Threads 2470
Handles 91143
Up time 3:22:37:27

Base speed:	2.40 GHz
Sockets:	1
Cores:	8
Logical processors:	16
Virtualization:	Enabled
L1 cache:	512 KB
L2 cache:	2.0 MB
L3 cache:	16.0 MB

Figure 6: ***CPU Utilization with FPGA Offload***

AI and Deep Learning

As mentioned before, FPGAs provide equivalent to higher performance when compared to a CPU for H.264/H.265 decoding. An FPGA-based solution provides better performance than a CPU if both the decoder and the Intra-frame encoder (such as JPEG or PNG) are hosted in the same FPGA. Additionally, in deep learning applications, it is common that some image preprocessing is performed before sending the image data to the deep learning processing. A single FPGA can handle of all the processing, including decoding, image processing and encoding (see the figure below) and can deliver high throughput, low latency and fewer data transactions (compared with a CPU). Deep learning technology is expected to be applied to a wide variety of industries or fields now and in the future, and FPGA-based solutions will contribute to its progress.



70536490-07.2020.11.04

Figure 7: ***Typical Data Flow for Video and Image Processing Using Deep Learning***

Speedster7t Architecture — Optimized for Performance

Speedster7t FPGAs were designed specifically to address the highest performance data acceleration applications. The architecture is well suited to address all of the application challenges presented in this white paper. Specifically, Achronix has developed a new innovative 2D network on chip (NoC) which helps to provide a balanced architecture between I/O bandwidth, external memory bandwidth and on-chip performance to ensure the highest overall throughput. In a traditional FPGA architecture, the user needs to design the circuitry to connect the accelerators, resulting in non-optimal placement and routing. Newer FPGA architectures use a network that streams data between processing elements within the logic array and the various on-chip high-speed interfaces and memory ports (figures below).

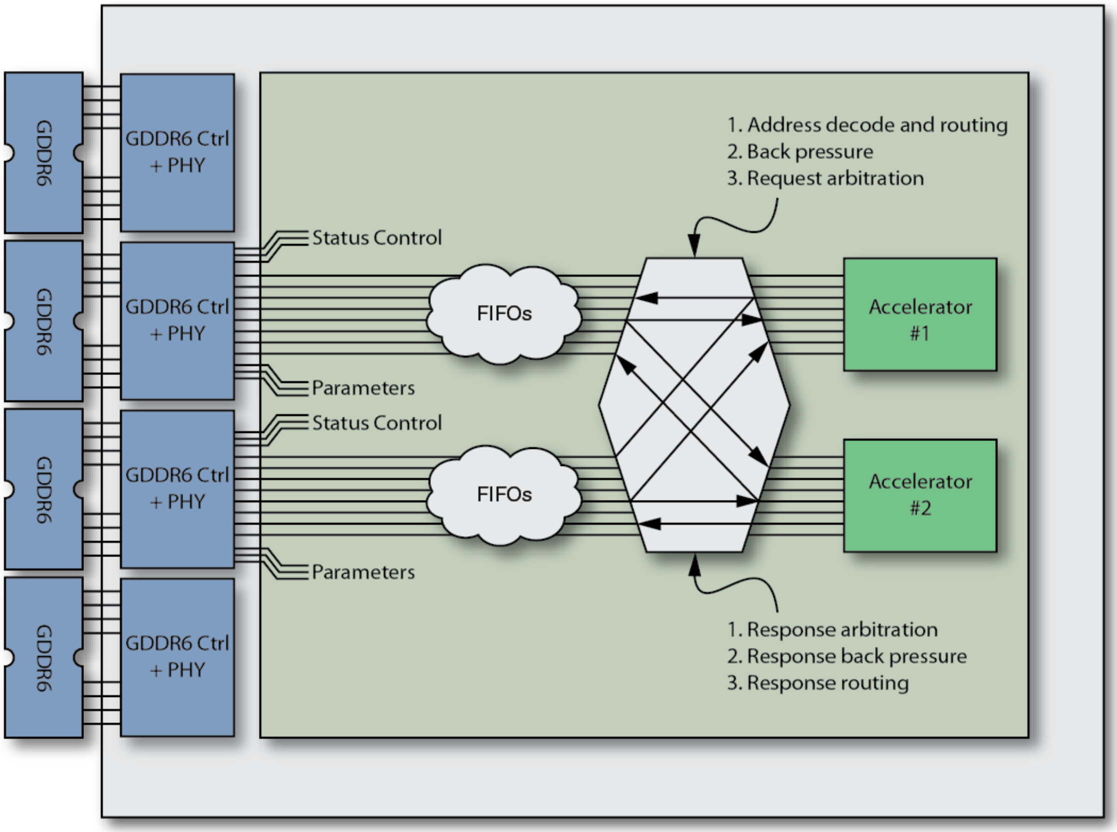


Figure 8: ***Connecting Accelerators in a Traditional FPGA Architecture***

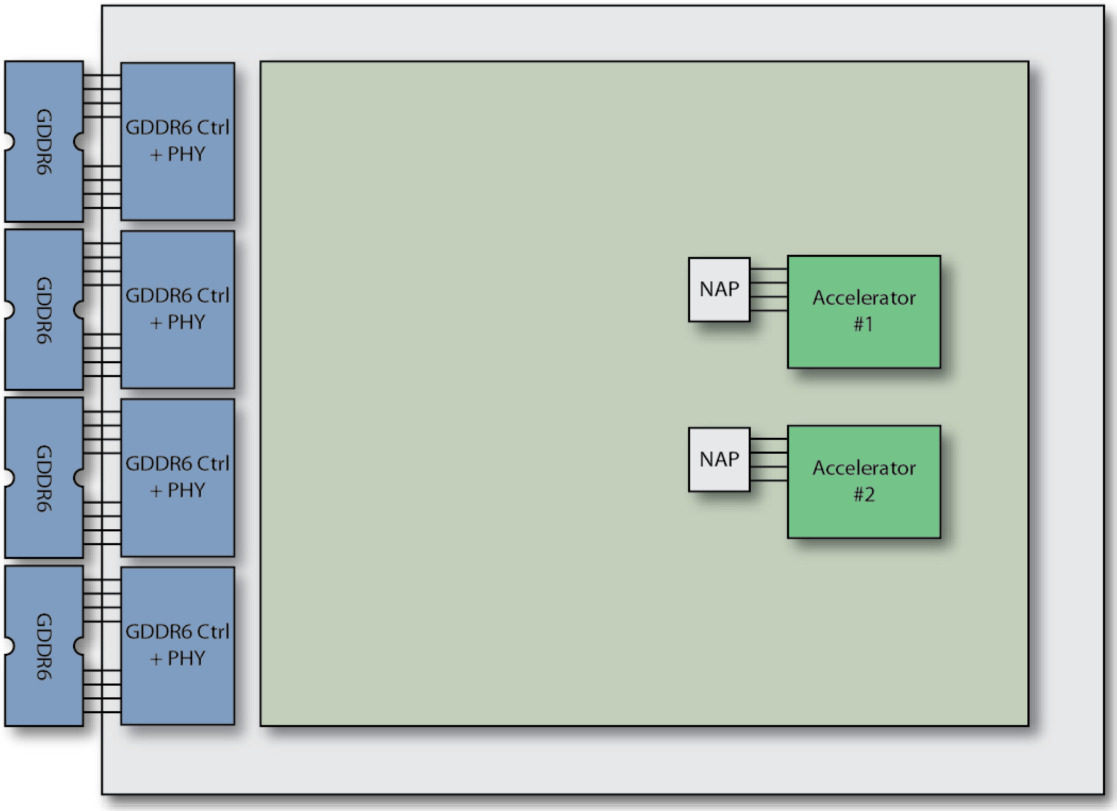


Figure 9: ***Advanced FPGAs Reduce the Amount of Needed Circuitry***

Hardwired architectures greatly improve the latency and energy efficiency of processing, but lack the flexibility to respond to changes in requirements. The first device of Speedster7t FPGA family, AC7t1500, delivers a range of high-speed interfaces that include fracturable Ethernet controllers (supporting rates up to 400G), PCI Gen 5 ports and up to 32 SerDes channels with speeds up to 112 Gbps. Accommodating the needs of encoders that need to buffer bulk data at high speed, the AC7t1500 is the first FPGA to deploy multichannel GDDR6 memory interfaces. These peripherals are interconnected by a smart 2D NoC in addition to the bit-oriented routing structure employed in the programmable-logic fabric. As a result, the Speedster7t FPGA is the first to be able to implement the video processing use cases mentioned above, utilizing a balanced architecture that brings together major improvements in both compute density and data-movement capabilities.

The Speedster7t architecture removes the bottleneck caused by the need to connect high-speed I/O channels directly to programmable logic which operates at much lower clock rates by providing a multi-level NoC hierarchy capable of an aggregate bandwidth of >20 Tbps. Not only does the NoC provide a huge upgrade in speed relative to the FPGA fabric interconnect, but the NoC is also able to move massive quantities of data without consuming any of the FPGA's programmable resources.. The internal NoC not only delivers increased bandwidth, but the smart connection mechanisms in Speedster7t FPGAs also ease the task of getting data into the fabric from the NoC ports.

This architecture makes it possible to take designs even further, for example, the inclusion of the matrix-oriented arithmetic units that enable the machine learning use case described above. Using techniques such as deep learning or simpler statistical techniques, the equipment can analyze traffic patterns to observe and enhance the flow of packets through the network and react quickly to changing conditions. At a high level, the following three Speedster7t innovations enable better FPGA designs for the mentioned use cases:

High-Speed Memory Interfaces

The choice of memory interfaces made by the Speedster7t architects reflects the massive bandwidth that the Ethernet and NoC connections provide. One possible approach would have been to design a family of products that employ the upcoming HBM2 interface. Although such an interface could deliver the level of performance needed, HBM2 is an expensive option and would force customers to wait for the necessary components and integration technology to become available.

The Speedster7t family instead employs the GDDR6 standard, which delivers the highest performance available for off-chip memories today. Speedster7t FPGAs are the first devices in the market to support this interface. Each on-chip GDDR6 memory controller can sustain 512 Gbps of bandwidth. With up to eight GDDR6 controllers in a single AC7t1500 device, a Speedster7t FPGA delivers an aggregate memory bandwidth of 4 Tbps.

PCIe Gen 5 Support

Alongside the Ethernet and memory controllers, the PCIe Gen 5 support on Speedster7t FPGAs allow tight integration with a host processor to support high-performance accelerator applications. The PCI Gen 5 controller makes it possible to read and write data stored within the memory hierarchy of the FPGA, including the many block RAMs located within the fabric as well as external GDDR6 and DDR4 SRAMs attached to the FPGA's memory controllers. Data movement controllers, such as DMA engines, instantiated in the FPGA array can similarly access memory shared with the host processor over the PCIe Gen 5 bus. This high-bandwidth connection is achieved without consuming any FPGA fabric resources and nearly zero design time. The user only needs to enable the PCIe and GDDR6 interfaces in order to send transactions via the NoC.

The direct connection between the PCIe subsystem and any of the GDDR6 or DDR4 memory interfaces is shown in the figure below.

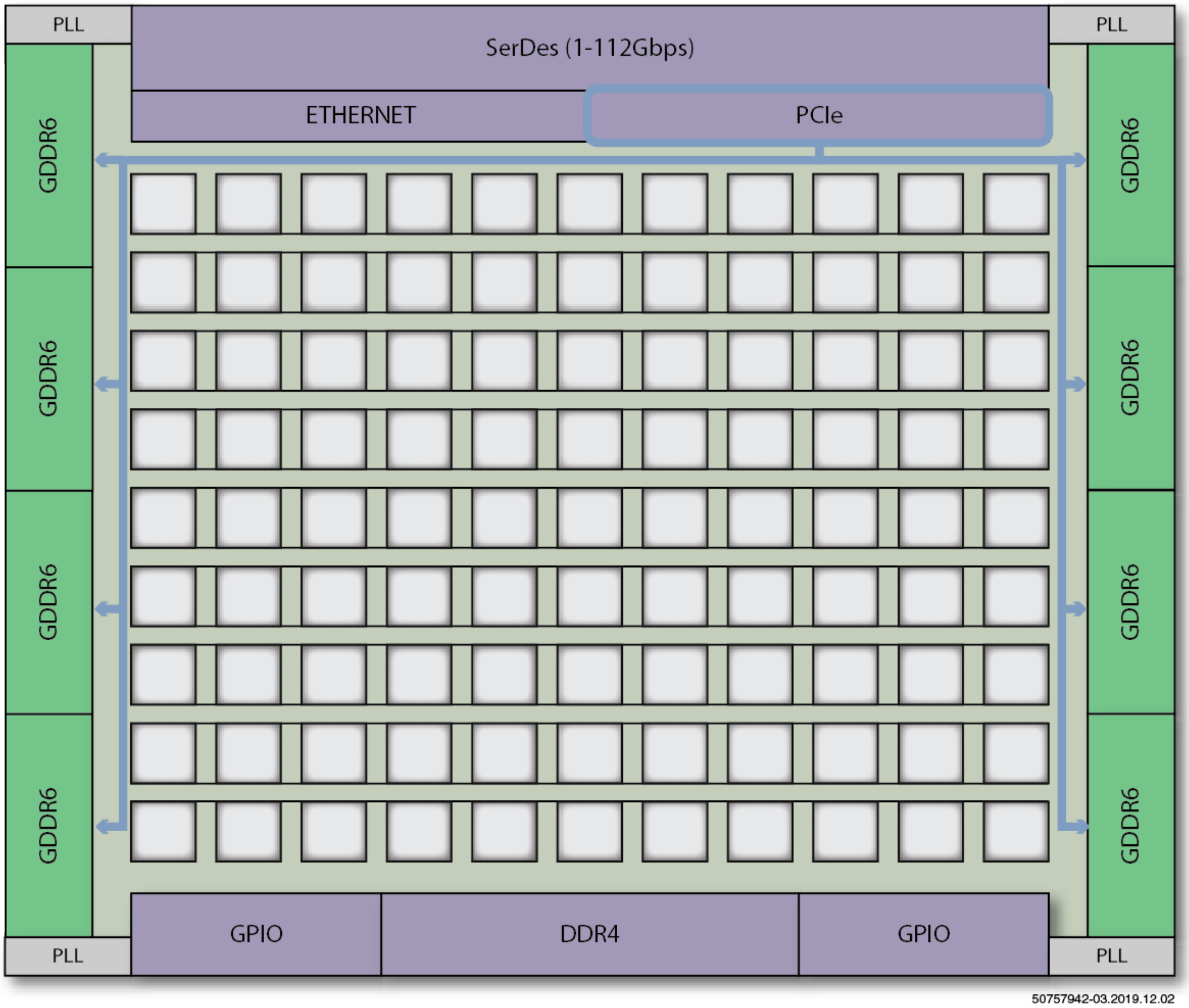


Figure 10: ***Data Transfer Between PCIe and GDDR6 Without Fabric Intervention***

Machine Learning Processor

For computationally intensive tasks, the Speedster7t machine learning processors (MLPs) deployed across a Speedster7t FPGA are flexible and factorable arithmetic units. MLPs are high-density multiplier arrays with floating-point and integer MAC blocks supporting multiple number formats. MLPs have integrated memory blocks that can perform operand and memory cascade functions without using any FPGA resources. MLPs are suited for a range of matrix-math operations, ranging from beamforming calculations for 5G radio controllers to the acceleration of deep-learning applications, such as traffic pattern and packet content analysis required for video processing systems.

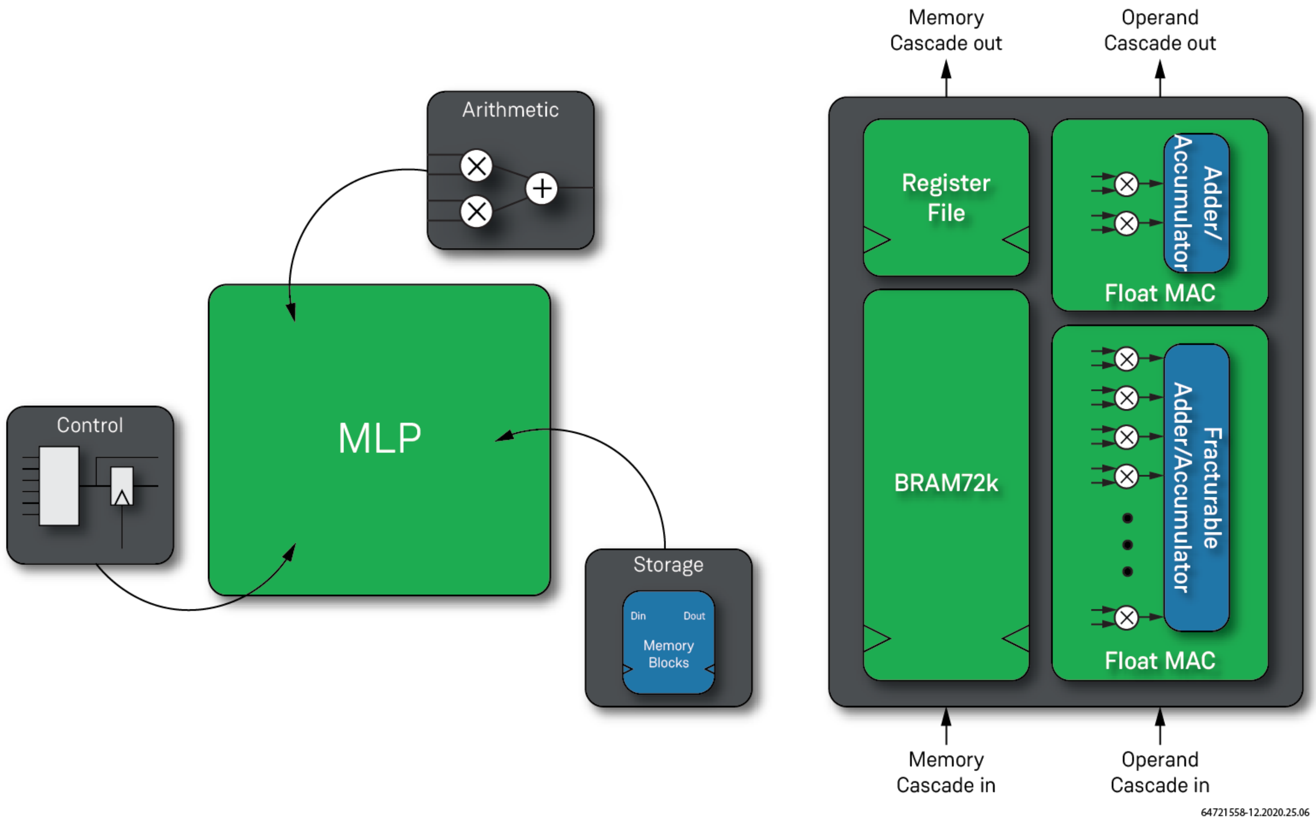


Figure 11: ***Machine Learning Processor Block Diagram***

Conclusion

While the performance of an ASIC is typically high, it supports only the feature set conceived of at design time and is not field upgradable. A CPU is the most flexible and easiest to design; however, clock frequencies have plateaued, and the era of dramatic improvements in performance are over. The workloads are increasing year by year — CPUs will not be able to keep pace. FPGAs represent a good balance between performance and flexibility. Video encoding, decoding, and image processing algorithms lend themselves to FPGA implementation due to the need of a lot of parallel processing. In conclusion. FPGA-based solutions reduce time to market, are highly customizable and can be effectively used to implement evolving algorithms.

IBEX Technology Introduction

IBEX Technology, based in Japan, provides Video Codec IP, IC design services and consulting such as ASIC and large-scale FPGA design, PCBs, etc. IBEX has developed a wide variety of video codecs such as MPEG-2, H.264/AVC, Apple ProRes, Avid DNxHD, SONY XAVC, Panasonic AVC-Intra and H.265/HEVC. IBEX Technology has track record mainly in the broadcast equipment industry, which requires high-quality, stability and reliability.

IBEX was founded in 1985 in Atsugi, Kanagawa, Japan and is recognized as a leader in the fields of LSI and video systems design. In the past, their major customers included Japanese television manufacturers and broadcasters, providing video codec IP for ASICs. In the last 10 years, the company's focus has shifted to providing FPGA-based solutions, due to performance and cost competitiveness of FPGAs. IBEX supplies IP for FPGAs to the customers worldwide.

Achronix®

Data Acceleration

Achronix Semiconductor Corporation

2903 Bunker Hill Lane
Santa Clara, CA 95054
USA

Website: www.achronix.com
E-mail : info@achronix.com

Notice of Disclaimer

The information given in this document is believed to be accurate and reliable. However, Achronix Semiconductor Corporation does not give any representations or warranties as to the completeness or accuracy of such information and shall have no liability for the use of the information contained herein. Achronix Semiconductor Corporation reserves the right to make changes to this document and the information contained herein at any time and without notice. All Achronix trademarks, registered trademarks, disclaimers and patents are listed at http://www.achronix.com/legal.