DELL EMC

Whitepaper

Real Time Streaming Analytics with Megh Computing on Dell EMC PowerEdge Servers

Revision: **1.0**
Issue Date: **10/2/2019**

Abstract

This blog evaluates the performance and efficiency of a video analytics pipeline with Deep Learning inference on Intel Programmable Acceleration Card (PAC) FPGA on Dell EMC PowerEdge R740/R740xd server. The objective is not only to report on real-world inferencing performance but also examine the end-to-end use case of the Intel PAC FPGA solution.

The Internet of Everything (IoE) is transforming industries, applications, and infrastructures. The demand for real-time stream processing is increasing rapidly with the explosion of streaming data from sensors, the web, and other sources. Enterprises want to create business value by processing data as it is generated without moving and this requires new integrated hardware and software platforms and tools to deliver the required combination of low latency and high compute capacity where data resides.

Revisions

Date	Description
02 October 2019	Initial release

Acknowledgements

This paper was produced by the following people:

Name	Role
Bhavesh Patel	Dell EMC
Prabhat K Gupta	Megh Computing
Suchit Subhaschandra	Megh Computing

Overview of Real time Analytics

The streaming analytics market is expected to attain a market size of USD 15.9 billion by 2022, growing at a CAGR of 33.1 percent.[1] The major driver of this growth is the massive surge of structured and unstructured data flowing from sources ranging from IoT sensors and events, social media, and web apps tracking data on usage and behavior to transactional and operational data from a broad spectrum of vertical segments. As businesses transition from reliance on traditional business intelligence (BI) to advanced analytics via machine learning and AI, the demands on compute and infrastructure increase—with high volumes of streaming data requiring new levels of performance, lower latency, and accelerated processing.

Open source or proprietary software-only solutions cannot keep pace with the increasing computation demands of real-time analytics. Organizations are expanding the number of nodes to deal with increasing data, but performance does not scale linearly.

Lowering latency at high volumes often cannot be achieved with the CPU alone. FPGAs can be utilized as part of a heterogeneous CPU and FPGA platform to implement reconfigurable accelerators and deliver better performance at lower latency. FPGAs are better suited for these kinds of applications compared to GPUs or other accelerators. But while FPGAs deliver advantages from accelerated performance to flexibility and programmability, they can be complex to program and manage efficiently. In addition, adoption of FPGAs requires integration into data center orchestration software.

Why FPGA?

FPGAs provide flexibility for AI system architects searching for competitive deep learning accelerators that also support differentiating customization. The ability to tune the underlying hardware architecture, including variable data precision, and software-defined processing allows allow FPGA-based platforms to deploy state-of-the-art deep learning innovations as they emerge. Other customizations include co-processing of custom user functions adjacent to the software-defined deep neural network. Underlying applications are in-line image & data processing, front-end signal processing, network ingest, and IO aggregation.

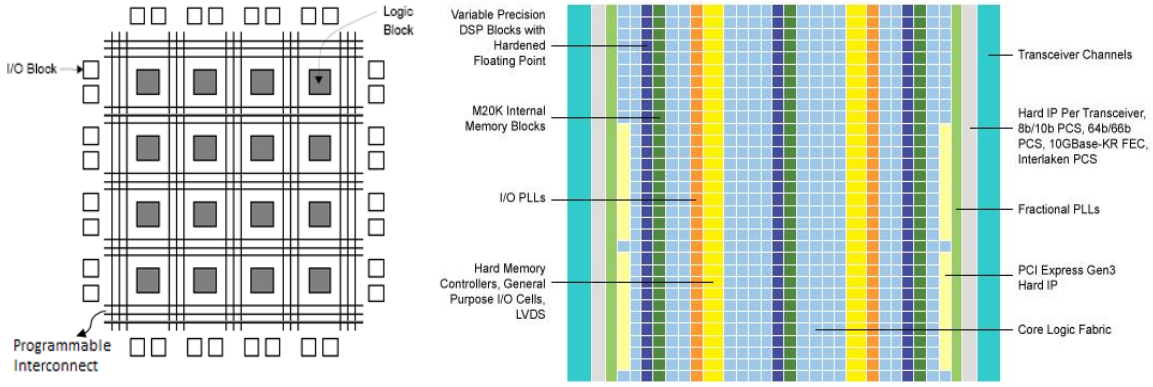


Figure 1. (Left) Arrayed building blocks are connected via interconnect wires; (Right) Fully featured FPGAs include a variety of advanced building blocks.

Figure 1 illustrates the variety of building blocks available in an FPGA. The core fabric implements digital logic with Look-up tables (LUTs), Flip-Flops (FFs), Wires, and I/O pads. FPGAs today also include Multiply-accumulate (MAC) blocks for DSP functions, Off-chip memory controllers, High-speed serial transceivers, embedded, distributed memories, Phase-locked loops (PLLs), hardened PCIe interfaces, and range from 1,000 to over 2,000,0000 logic elements.

FPGAs in Mission-Critical Applications

Mission-critical applications (e.g., autonomous vehicle, manufacturing, etc.) require deterministic low-latency. The data flow pattern in such applications may be in streaming form, requiring pipelined-oriented processing. FPGAs are excellent for these kinds of use cases given their support for fine-grained, bit-level operations in comparison to CPU and GPUs. FPGAs also provide customizable I/O, allowing their integration with these sorts of applications.

In autonomous driving or factory automation where response time can be critical, one benefit of FPGAs is that they allow tailored logic for dedicated functions. This means that the FPGA logic becomes custom circuitry but highly reconfigurable, yielding very low compute time and latency. Another key factor may be power – the cost per performance per watt may be of concern when determining long-term viability. Since the logic in FPGA has been tailored for a specific application/workload, the logic is very efficient at executing that application which leads to lower power or increased perf per watt. By comparison, CPUs may need to execute 1000's of instructions to perform the same function that an FPGA maybe able to implement in just a few cycles.

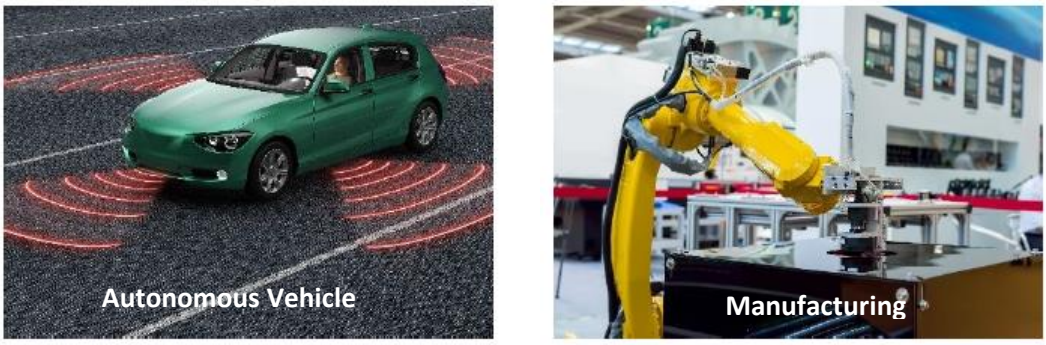


Figure 2. Examples of mission-critical applications require deterministic, fast response.

Intel Programmable Acceleration Card

The Intel Programmable Accelerator Card (PAC) features an Intel Arria 10 FPGA, an industry-leading programmable logic built on 20 nm process technology, integrating a rich feature set of embedded peripherals, embedded high-speed transceivers, hard memory controllers and IP protocol controllers. Variable-precision digital signal processing (DSP) blocks integrated with hardened floating point (IEEE 754-compliant) enable Intel Arria 10 FPGAs to deliver floating point performance of up to 1.5 TFLOPS. Arria 10 FPGAs have a comprehensive set of power-saving features. Combined, these features allow developers to build a versatile set of acceleration solutions.

Figure 3. Intel Programmable Acceleration card.

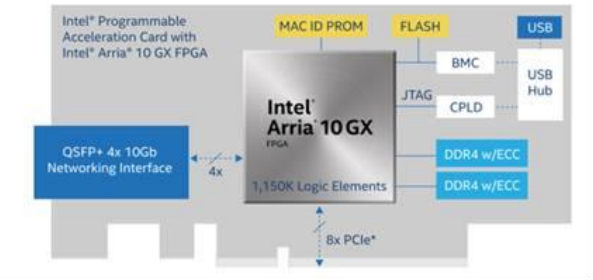


Figure 4. PAC Block Diagram.

Acceleration Stack

The Acceleration Stack for Intel Xeon CPU with FPGAs is a robust collection of software, firmware, and tools designed and distributed by Intel to make it easier to develop and deploy Intel FPGAs for workload optimization in the data center. The Acceleration Stack for Intel Xeon CPU with FPGAs provides multiple benefits to design engineers, such as saving time, enabling code-reuse, and enabling the first common developer interface.

The Acceleration Stack for Intel Xeon CPU with FPGAs provides optimized and simplified hardware interfaces and software application programming interfaces (APIs), saving developer's time so they can focus on the unique value-add of their solution.

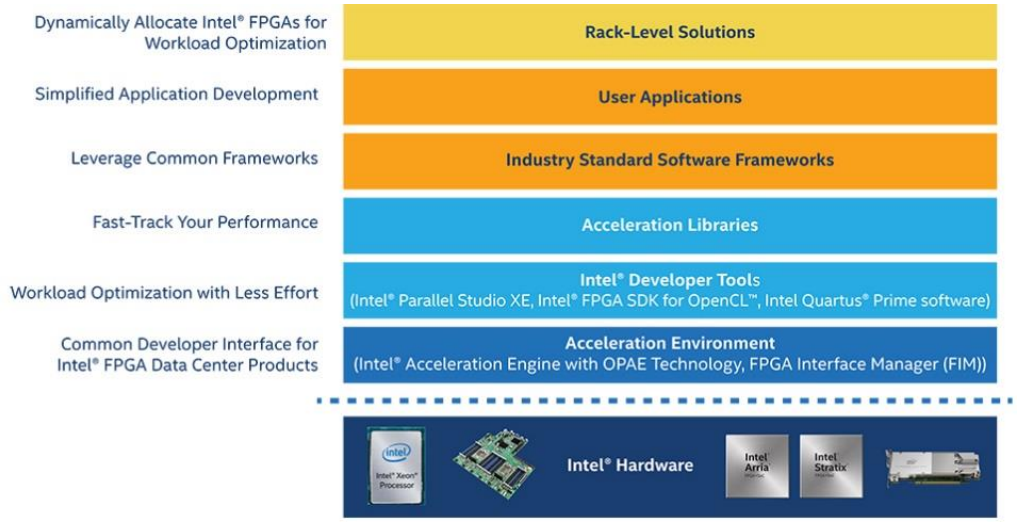


Figure 5. Acceleration Stack for Intel FPGA.

Evaluation Methodology

In this section, we give a brief overview of sample image-classification application, the hardware inferencing infrastructure used, and the deep learning models that were evaluated.

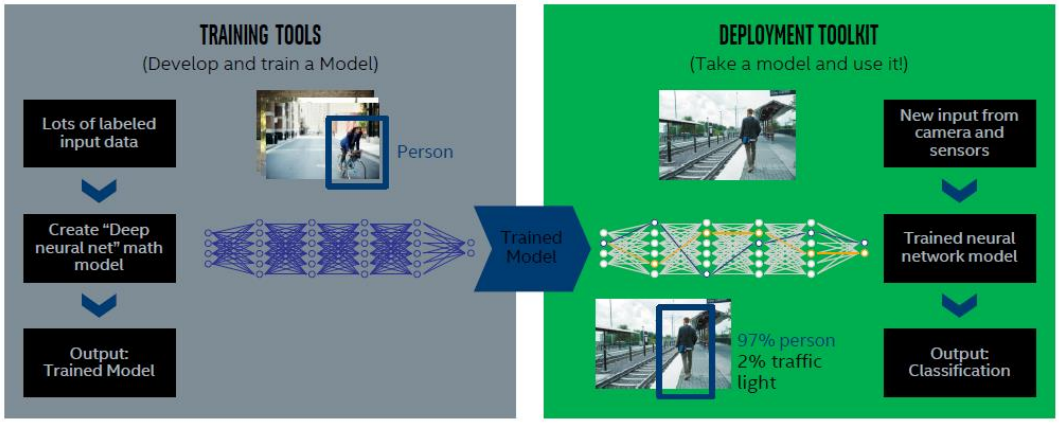


Figure 6. End-to-end deep learning: from trained models to model deployment.

Application Case Study

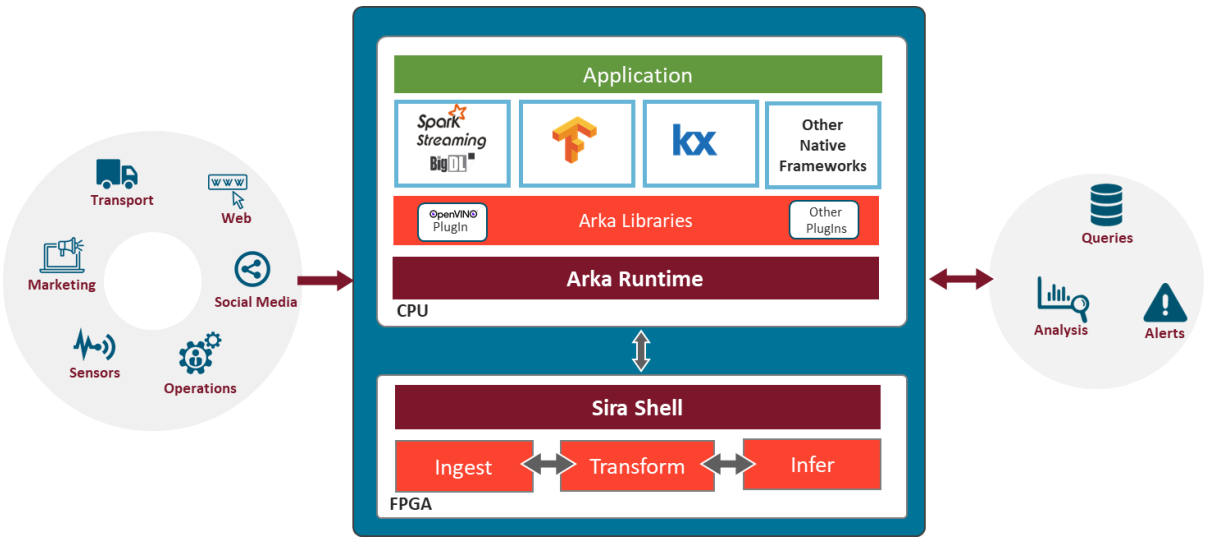
We consider fraud prevention in the retail supply chain as the application case study. Retail inventory loss (or "shrinkage") is a serious problem, totaling about $100 billion annually—almost 1.8% of sales—worldwide. Not surprisingly, retailers are seeking solutions. While some are looking at a broad range of possibilities—facial recognition, motion tracking, reading emotions and gestures, and AR—most want to begin with solutions that are low cost and yield immediate returns. Fortunately, there are effective technological solutions that satisfy these criteria.

Megh has developed the Video Analytics Solution (VAS) to solve this problem. The Solution is targeted for various use cases including fraud prevention in the retail supply chain, inventory tracking in manufacturing, and video surveillance for security.

Solution Description

The Solution maps the entire real time analytics pipeline consisting of the ingestion phase of streaming data input, transformation phase of video decoding and image resizing and the inference phase of object detection and classification into multiple networked FPGAs with integration into user's application.

The Solution runs on Dell EMC PowerEdge Server with Intel Programmable Acceleration Cards. The



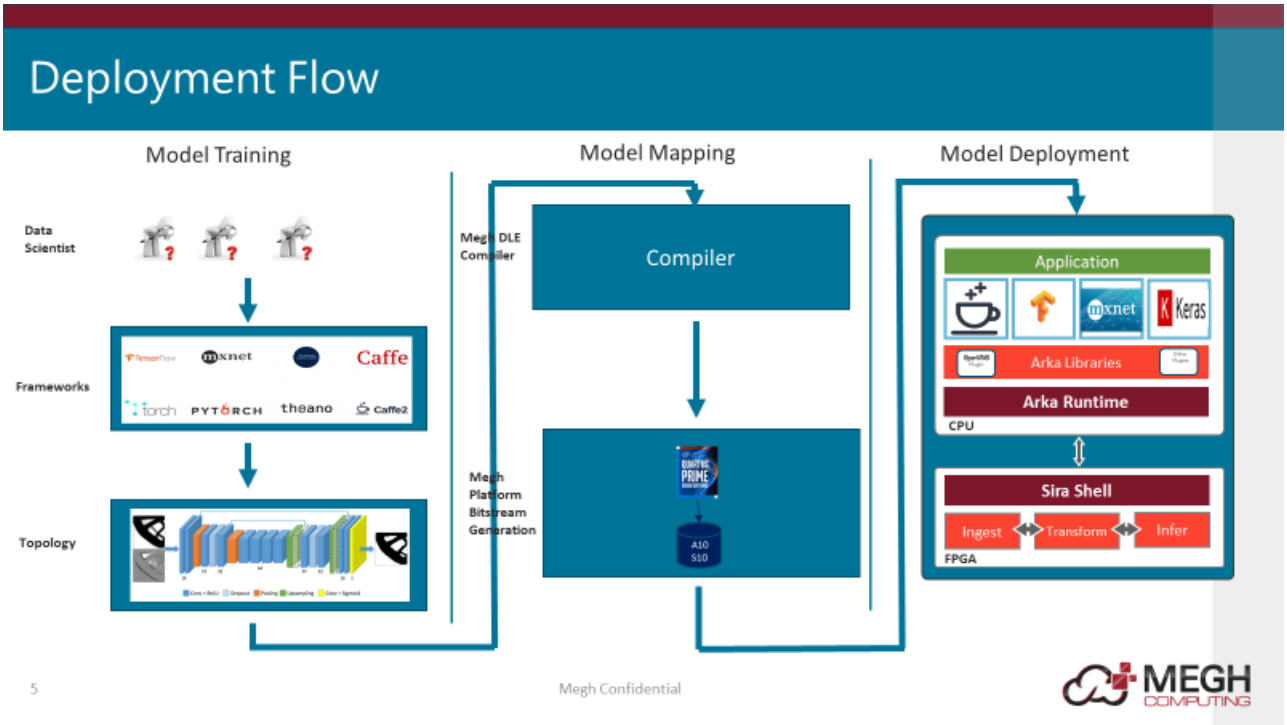
Solution consists of Sira Accelerator Function Units (AFUs) for inline processing of streaming data and offload processing of compute intensive algorithms on multiple networked FPGAs. The AFUs used in this solution include:

- PPE (Packet Processing Engine) for ingesting the data using a direct NIC and extracting the payload

- SPE (Stream Processing Engine) for transforming the data which includes multi-channel H.264 decoder and image re-sizing
- DLE (Deep Leaning Engine) for image classification using CNN topologies like Resent50.

These are managed by the Arka Runtime which exposes Accelerator Functions-as-a-Service via high level APIs for integration with the applications frameworks like Spark, TensorFlow, kdb+, with no changes.

Megh has developed the Deep Learning Engine (DLE) from the ground up for streaming inference. It consists of a library of high performance, mixed precision DL primitives that are drop-in replacements for TensorFlow and PyTorch. Megh provides a DLE compiler that directly parses TensorFlow and PyTorch models, creating an optimal DLE configuration for loading on the FPGA. The deployment flow for the DLE model is shown in the figure below.



Dell EMC PowerEdge Server

We used the Dell EMC PowerEdge R740/R740xd servers to host the PAC boards. The PowerEdge R740/R740xd is a general-purpose platform with highly expandable memory (up to 3TB) and impressive I/O capability to match both read-intensive and write-intensive operations. The Dell EMC PowerEdge R740 is capable of handling demanding workloads and applications such as data warehouses, E-commerce, databases, high-performance computing (HPC), and Deep learning workloads.

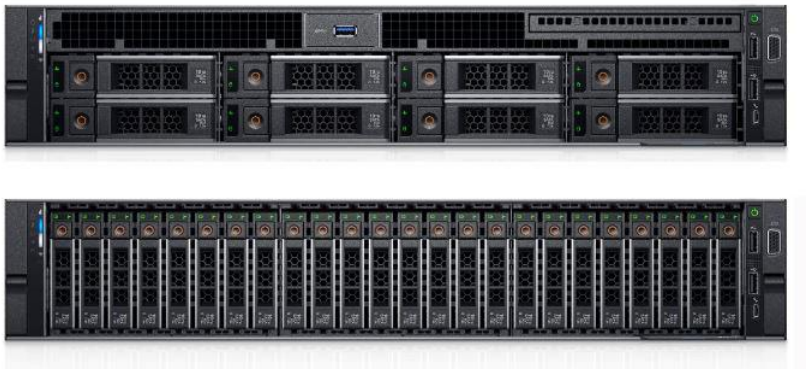


Figure 7. Dell EMC PowerEdge R740/R740xd.

The Dell EMC PowerEdge R740/R740xd is Dell EMC's latest two-socket, 2U rack server designed to run complex workloads using highly scalable memory, I/O capacity, and network options. The R740/R740xd features the Intel Xeon processor scalable family, up to 24 DIMMs, PCI Express (PCIe) 3.0 enabled expansion slots, and a choice of network interface technologies to cover NIC and rNDC. In addition to the R740's capabilities, the R740xd adds unparalleled storage capacity options, making it well-suited for data intensive applications that require greater storage, while not sacrificing I/O performance.

Results [*1]

The performance of VAS is summarized in the table below for two configurations:

1. CPU Only: where the complete pipeline is implemented in software. As expected, the performance varies based on the Xeon processor used. The throughput for Resnet50 based inference is 30 fps for a dual socket Xeon Silver server and 150 fps for a dual socket Xeon Platinum processor.
2. CPU+FPGA: where the real time analytics pipeline is implemented in the FPGA. We used two Intel PACs : the first one to ingest the data and do the video decoding and the second one to do the deep learning inferencing.

8 Channels	CPU	CPU + FPGA
CPU	2x Xeon E5 8280 Platinum	Xeon E5 4214 Silver
FPGA		2x Arria 10
Power	2x 400W = 800W	250W (150W + 2x 50W)

Cost*	$30000 * 2 = $60000	$17,000 (2x3500 + 2x $5000)
Throughput (Resnet50 8 bit)	150 * 2 = 300 fps	240 fps required (max ~400 fps)
Latency	> 250ms	< 100 ms

*Cost based on following configurations:

Dell PowerEdge R740 Rack Server with dual Xeon Platinum 8280M processors, 64G Memory, 480GB SSD: $31,441

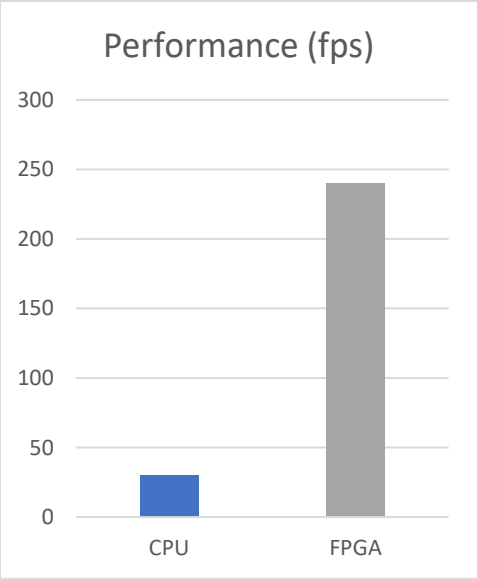
Dell PowerEdge R740 Server with dual Xeon Silver 4214 processors, 16G Memory, 480GB SSD: $3,439

Intel Programmable Acceleration Card with Arria® 10 GX web price: $5,000

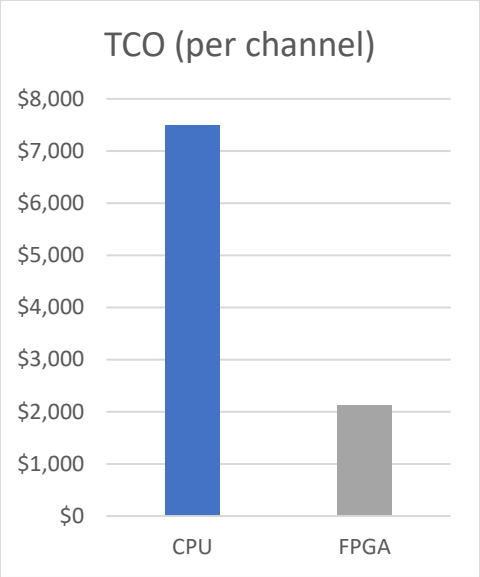
Performance

Using the data above the performance is summarized in the graphs below:

1. Throughput: > 8x improvement from CPU only to CPU+FPGA implementation



2. Cost: >3x improvement in per channel cost for the same performance of 240 fps



Conclusions

Real Time Analytics solution running on Dell EMC Power Edge servers with Intel PACs delivers >8x throughput at <3x lower cost compared to CPU-only solutions to support various Video Analytics use cases for fraud prevention, inventory management and surveillance. The same platform can be applied to support Text Analytics for financial analytics, network analytics and other use cases.

References

[1] H. Zhao, O. Gallo, I. Frosio and J. Kautz, "Loss Functions for Image Restoration with Neural Networks," in IEEE Transactions on Computational Imaging, vol. 3, no. 1, pp. 47-57, March 2017

[2] L. J. Buturovic and L. T. Citkusev, "Back propagation and forward propagation," [Proceedings 1992] IJCNN International Joint Conference on Neural Networks, Baltimore, MD, 1992, pp. 486-491 vol.4.