Performance solution brief
November 2019

Real-Time Streaming Video Analytics with Hewlett Packard Enterprise, Intel®, and Megh Computing

Key takeaways

HPE ProLiant DL380 Gen10 server with Intel® Arria® 10 GX FPGAs programmed with Megh Computing's real-time streaming Video Analytics Solution:

- 8.3X gain in performance*
- Reduced TCO

Solution components

- HPE ProLiant DL380 Gen10 server
- Intel® Programmable Acceleration Card (Intel® PAC) with Intel® Arria® 10 GX FPGAs
- Megh Computing's real-time streaming Video Analytics Solution

Benchmark configurations

Baseline test:

- Master node: 1 HPE ProLiant DL360 Gen10 server
  - 2 Intel® Xeon® Gold 6218 processors at 2.30 GHz
  - 192 GB memory (12 x 16 GB HPE DDR4 SmartMemory RDIMMs)
  - 2 HPE 480 GB SATA Mixed Use SFF SSDs for OS
  - HPE Ethernet 10 Gb 2-port 562FLR-SFP+
  - HPE Ethernet 1 Gb 4-port 331i
  - Red Hat Enterprise Linux 7.6
- Worker node: 1 HPE ProLiant DL380 Gen10 server
  - 2 Intel Xeon Gold 6254 processors at 3.10 GHz
  - 384 GB memory (12 x 32 GB HPE DDR4 SmartMemory RDIMMs)
  - 2 HPE 480 GB SA Mixed Use SFF SSDs for OS
  - 4 HPE 3.84 TB SATA Mixed Use SFF SC DS SSDs for data
  - HPE Ethernet 10 Gb 2-port 562FLR-SFP+
  - HPE Ethernet 1 Gb 4-port 331i
  - Red Hat Enterprise Linux 7.6
  - Megh Computing's real-time streaming Video Analytics Solution directly installed on the server

Comparison test:

Same configuration as baseline test, but with addition to the worker node of 2 Intel® Programmable Acceleration Cards with Intel® Arria® 10 FPGAs programmed with Megh Computing's real-time streaming Video Analytics Solution.

8.3X analytics platform performance gain with FPGAs

Executive summary

As the demand for video streaming analytics surges, technology innovation is a key requirement to address challenges in many use cases. The newest wave of computing infrastructure is represented by specialized hardware accelerators based on field-programmable gate arrays (FPGAs) to support real-time streaming for machine learning and artificial intelligence for anlaytics. For example, video analytics environments include surveillance, inventory management, and fraud prevention related to mis-scanning of stock keeping units (SKUs). Components consist of video devices, servers, and FPGAs that are pre-programmed with a complete and integratable stack for deep learning.

Deployment of a video surveillance platform in any setting means making decisions about which server and components will maximize the capabilities of the associated software and hardware. Businesses must find the right balance between price and performance. Some require a very low response time in hundreds of miliseconds (ms) and an ideal frame rate. High performance processors will be capable of low latency and desirable frame rate in a real-time streaming analytics scenario, but at the price of a premium processor. Lower-wattage processors are less costly; however, performance then might be inadequate.

A cost-effective, higher-performance solution in a server deployed to run real-time streaming analytics applications can be achieved by pairing lower-cost CPUs such as Intel® Xeon® Gold Scalable processors and Intel FPGAs. Together, the HPE ProLiant DL380 Gen10 server, Intel Xeon Gold 6130 processors, Intel® Arria® 10 GX FPGAs, and Megh Computing's real-time streaming Video Analytics Solution for retail fraud prevention combine to form a solution for outstanding performance and lower total cost of ownership (TCO).

HPE internal lab performance benchmark tests demonstrated the power of this technology solution, showing an 8.3X performance gain utilizing the HPE ProLiant DL380 Gen10 server as the foundation for the Megh Computing real-time streaming analytics retail platform for a retail fraud detection scenario on Intel Arria 10 GX FPGAs.* This paper explains the test results and how the solution can reduce cost over the life of the investment.

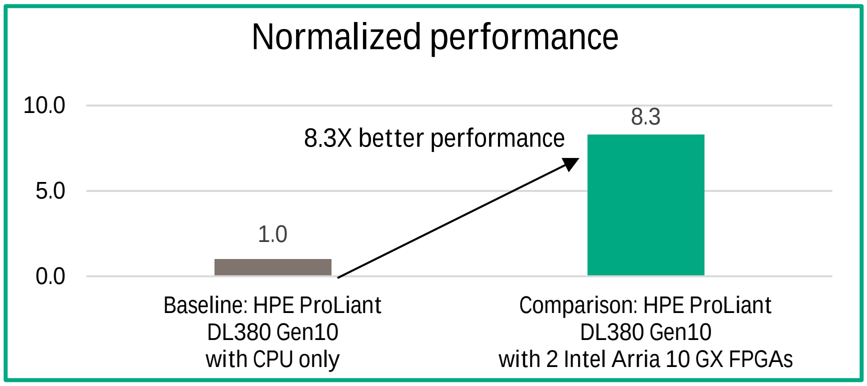

Figure 1. Normalized performance comparison of Megh Computing real-time streaming Video Analytics Solution on an HPE ProLiant DL380 Gen10 server configured with CPU only vs. with Intel Arria 10 GX FPGAs.

* Tests measure performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit intel.com/benchmarks.

Solution overview and test methodology

The solution can be deployed in a cluster consisting of a master node and one or more worker nodes. Each worker node is configured with two Intel Arria 10 GX FPGA cards connected directly by a switch. In the HPE test scenario, the master node was an HPE ProLiant DL380 Gen10, and the worker node consisted of an HPE ProLiant DL360 Gen10 server. The master node and worker node were configured with Intel Xeon Scalable Gold family processors for balanced CPU performance and economics. The benchmark test used a pre-recorded video file stored on a solid-state drive (SSD) in the master node, which fed the video data to the worker node. In the baseline test, there were no FPGAs installed in the worker node, and the inline streaming data was handled by only CPUs with the application installed on the server. In the comparison test, the worker node was configured with two Intel Arria 10GX FPGAs with Megh Computing's Video Analytics Solution pre-programmed on them. The FPGAs directly handled the inline streaming data from the master node, allowing the CPUs to handle other data activities.

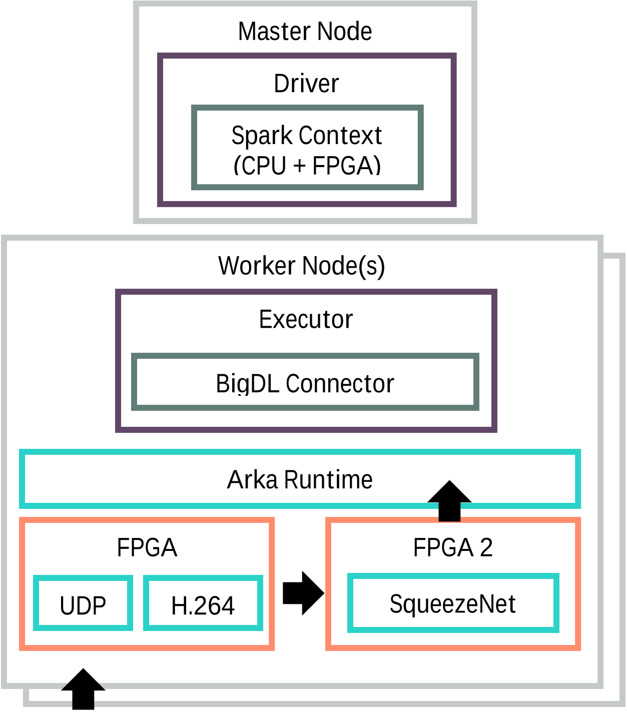

Figure 2. Solution diagram representing the master node server, worker node servers, and two FPGAs programmed with Megh Computing's Video Analytics Solution.

Solution components and benefits

In the benchmark testing conducted by HPE, the HPE ProLiant DL380 Gen10 server was configured with two Intel Arria 10 GX FPGAs pre-programmed with Megh Computing's streaming Video Analytics Solution. Each component contributed powerful benefits to the solution. Figure 2 shows a diagram of the solution components and analytics platform process flow.

HPE ProLiant DL380 Gen10 server

The foundation for the solution was theHPE ProLiant DL380 Gen10 server, a secure 2P, 2U server adaptable for diverse workloads and environments. Designed for superme versatility and resilience, while backed by a comprehensive warranty, the server delivers world-class performance with the right balance of expandability and scalability. The HPE ProLiant DL380 Gen10 server has an adaptable chassis, including HPE modular drive bay configuration options with up to 30 SFF, up to 19 LFF, or up to 20 NVMe drive options along with support for up to three double-wide graphics processing unit (GPU) options. Along with an embedded 4x1GbE, there is a choice of HPE FlexibleLOM or PCIe standup adapters, which offer a choice of networking bandwidth (1GbE to 40GbE) and fabric that adapt and grow to changing business needs.

Intel® Programmable Acceleration Card with Intel® Arria 10 GX FPGAs (Intel PAC with Intel Arria 10 GX FPGAs)

Intel® Programmable Acceleration Cards (Intel® PACs) are PCIe-based acceleration adapter cards that offer both inline and lookaside acceleration. Intel Arria 10 devices have a 20 nm Arm®-based SoC with up to 1.5 GHz bandwidth and hard floating-point digital signal processing (DSP) blocks with speeds up to 1.5 tera floating-point operations per second (TFLOPS). This FPGA family implements publicly-available OpenCore designs. Intel Arria 10 GX FPGAs are enabled with up to 96 full-duplex transceivers with data rates up to 17.4 Gbps chip-to-chip, 12.5 Gbps backplane, and up to 1,150K equivalent logic elements (Les).

The Intel Arria 10 GX FPGA is ideal for a broad array of applications such as communications, data center, military environment, broadcast, automotive, and other end markets.

Megh Computing real-time streaming analytics platform

The Megh Computing real-time streaming analytics platform is designed to analyze streaming data in audio, video, and text formats. The Megh analytics platform abstracts the complexity of FPGAs via plugin libraries that work with open source and custom frameworks. The platform subtype described in this paper, Megh's Video Analytics Solution, creates a streaming video analytics pipeline targeting fraud prevention and other retail segment uses cases.

The software component is composed of the Spark Streaming framework with BigDL libraries combined with Megh libraries for the Arka Runtime framework which manage Sira Accelerator Function Units (AFUs). The pipeline consists of functional units (FUs) to implement the three analytics stages:

- Ingestion of streaming data input,
- Transformation of video decoding, and image resizing, and
- Inference phase of object detection and classification.

These functions are mapped across the two FPGAs as follows:

- The ingest and transform stages are implemented on the first FPGA.
- The inference stage is implemented on the second FPGA.

The Arka Runtime exposes the accelerator-as-a-service functions, manages the Intel FPGAs and supports software fallback for the AFUs.

The Sira AFUs deliver the actual acceleration and are implemented as libraries that get downloaded to the Intel FPGA. Applications run unmodified using standard application programming interfaces (APIs).

TCO impact of benchmark configuration

Together, the various solution components in this benchmark testing scenario can help lower the cost of ownership over the life of the solution compared to using a server or multiple servers with high-end processors without accelerators, or deploying generic GPUs that then will need to be programmed locally, adding time and cost.

The high-performing Intel Arria 10 GX FPGA becomes a purpose-built component when coded with Megh Computing's real-time streaming Video Analytics Solution. The card can simply be plugged into a server for deployment in a real-time streaming analytics environment, thus avoiding labor-intensive local programming. The card can be reprogrammed to serve a different use case, for excellent investment protection.

Megh Computing's Video Analytics Solution itself is designed to enable low latency, high throughput, and scalability. These capabilities increase the productivity of hardware and software solutions, contributing to greater return on investment (ROI).

Additional economic benefit is gained from the performance-boosting and highly adaptable design of the HPE ProLiant DL380 Gen10 Server with Intel Xeon Scalable Gold 6xxx processors, which offer a balance of performance and cost, as well as a lower wattage for long-term energy cost savings. HPE options, along with HPE management solutions and HPE Pointnext technology services that combine to maximize customer return on investment. HPE Pointnext technology services provide global support with important capabilities offered as a standard, and proactive optional support plans. HPE has a broad range of technology partners and expertise at the edge, with artificial intelligence (AI), and in the cloud to help customers successfully implement transformational technologies in their environment.

Bottom line

Customers can achieve dramatically increased performance and reduced TCO by running the Megh Computing real-time streaming video analytics platform on the HPE ProLiant DL380 Gen10 server with Intel Arria 10 GX FPGAs compared to a configuration without FPGAs.

For more information

HPE ProLiant DL380 Gen10 Server
Intel Arria 10 GX FPGA
Megh Computing real-time streaming analytics platform

Our solution partners

a50000486enw, November 2019