SOLUTION BRIEF

FPGAs
Real-Time Analytics on Streaming Data

# Megh Computing Platform Accelerates Real-Time Streaming Analytics with Intel® Arria® 10 FPGAs

## Seamlessly extract more value from streaming data with machine learning and deep learning

"The third wave of computing in the data center has started. Now specialized hardware accelerators based on FPGAs are available on-premise, and in the cloud, to support the needs of real time streaming, machine learning, and deep learning for video analytics and other applications."

**—Prabhat K. Gupta, CEO, Megh Computing**

### Executive summary

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 moving. They are increasingly adopting the Spark Streaming* framework in cloud deployments with machine learning and deep learning libraries to extract information from streaming data. Software-only solutions used today are not able to meet the performance requirements for low latency and high throughput. New integrated hardware and software platforms and tools are needed to deliver the required combination of low latency and high compute capacity in the cloud and at the edge.

Megh Computing provides an innovative platform for accelerating real-time analytics using Spark Streaming and other frameworks like FFmpeg, OpenCV and TensorFlow. The solution runs on an Intel® Arria® 10 FPGA-based hardware and software platform, the Intel® Programmable Acceleration Card (Intel® PAC), and Intel Acceleration Stack for Intel Xeon® CPU with FPGAs, for seamless acceleration of applications that process data streams with machine and deep learning algorithms to extract value. With the Megh and Intel solution, organizations can increase performance at lower latency, lower TCO, and support new use cases and workloads.

### Challenges

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 IP cameras, 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 and deep learning, the demands on compute and infrastructure increase—with high volumes of streaming data requiring new levels of performance, lower latency, and accelerated processing.

Real-time analytics solutions today are based on a software platform using Apache Kafka*, Apache Flink*, or similar frameworks for streaming the data and the Spark Streaming framework as a distributed framework for processing the data. Typical operational stages include (1) extract, transform, load (ETL), which ingests the data and prepares it for processing, and (2) machine learning and/or deep learning libraries to infer information from the data.

These open source or proprietary software-only solutions cannot keep pace with the increasing computation demands of real-time analytics or deliver the low latency required to support the real-time analytics use cases. Moreover, organizations, as they expand the number of nodes to deal with increasing data, find that the performance does not scale linearly.

Lowering latency at high speed often cannot be achieved with the CPU alone. FPGAs can be utilized as part of a heterogeneous CPU and FPGA platform to implement reconfigurable, application-specific accelerators and deliver better performance at lower latency. FPGAs are better suited to these kinds of applications compared to GPUs or other accelerators, as they deliver advantages such as accelerated performance, flexibility, lower power, and programmability. The tradeoff is that FPGAs can be complex to program and manage efficiently and require integration into data center orchestration software.


Current solutions using Spark Streaming* with machine and deep learning often do not meet performance requirements

## Solution

The Megh Computing real-time analytics platform is designed to address the key challenges of real-time analytics for streaming data in diverse formats, including audio, video, and text.

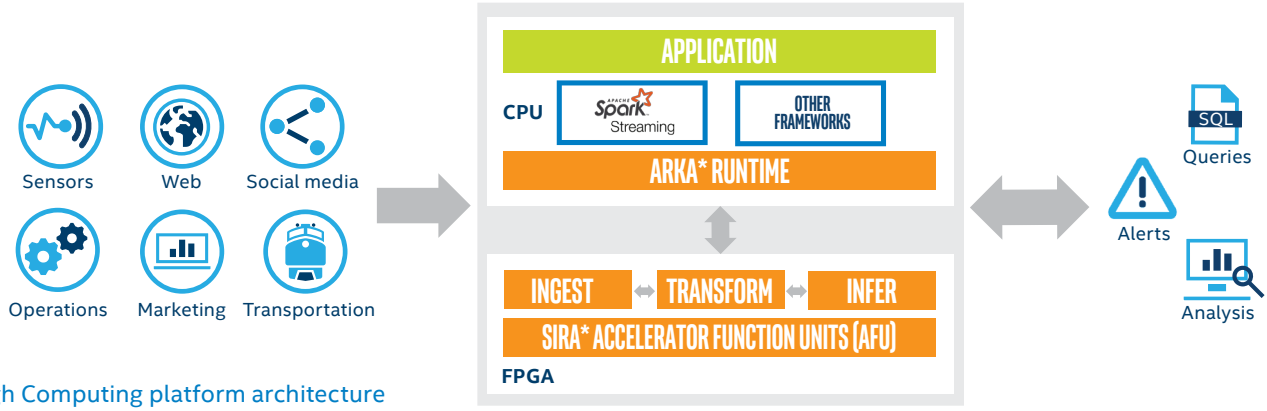
The Megh solution is based on a heterogeneous CPU+FPGA platform using standard high-volume Intel® Xeon® servers coupled with Intel® Arria® 10 FPGA Programmable Acceleration Cards (PACs) and with the Intel Acceleration Stacks. The Megh solution has three main attributes:

- Implementation of the complete real-time analytics pipeline in the FPGA:

  **Ingest:** For streaming data in different formats
  **Transform:** For transforming the data for processing
  **Infer:** For machine learning and deep learning inferencing

- Support for standard and custom data analytics frameworks with machine and deep learning libraries with no code changes. For example, using Spark Streaming and BigDL* libraries for inferencing, or using FFmpeg and OpenCV libraries for video decoding and image processing.
- Customization for different use cases using SDKs for Megh's Arka Runtime software and Sira* FPGA hardware libraries.

With Megh and Intel, you gain the same ease of use as with an ASIC, but with the benefits of FPGA programmability and reconfiguration. Megh's integrated software and hardware solution on the Intel® platform abstracts the complexity of FPGAs by exposing preoptimized libraries via high-level APIs, essentially taking the library calls and remapping them to FPGAs. The Intel Arria 10 FPGAs directly handle the inline streaming data, leaving the CPU available for other data center activities.




Megh Computing platform architecture

| MEGH COMPUTING REAL-TIME ANALYTICS PLATFORM BENEFITS | |
|---|---|
| **Increase performance and efficiency** | • Lower total cost of operations by **more than 2x**[2]<br>• Increase performance efficiencies by **more than 5x to 10x**[2]<br>• More than **10x lower latency**[2] |
| **Speed time to value** | • Support deployment of new real-time analytics use cases via a high-volume hardware platform<br>• Gain the flexibility of open source software frameworks and the ability to customize with Megh SDKs |
| **Simplify deployment** | • Deploy FPGA accelerator-as-a-service using containers or virtual machines (VMs) in public, private, or edge clouds and integrate with the cloud orchestration layer |

## FPGAs for the data economy

Flexible, programmable FPGAs are enabling business and industry to meet the growing need for scale-out accelerators for real-time streaming analytics with machine and deep learning algorithms.

FPGAs provide a reconfigurable sea of hardware gates to:

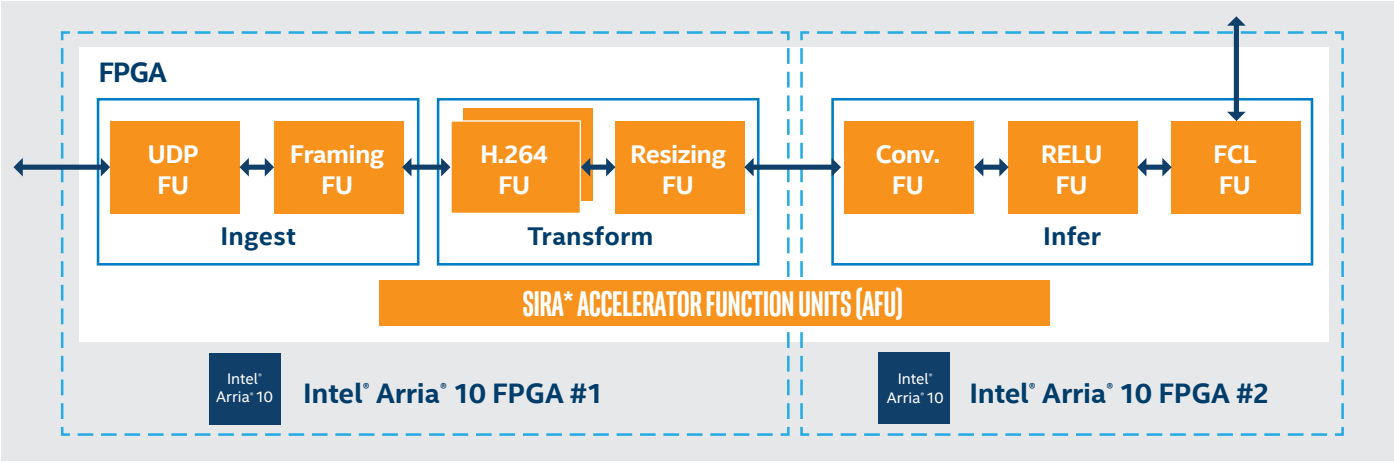
- Design a custom hardware accelerator with direct I/O connectivity and low latency
- Deliver increased performance efficiencies for a single application
- Quickly reconfigure a device as a new accelerator for a different application

The Megh Computing solution takes advantage of these capabilities to implement the complete real-time analytics pipeline in the FPGA. For example, Megh created a video analytics pipeline targeting fraud prevention and other use cases in the retail segment. The complete pipeline consists of various functional units (FUs) to implement the ingest, transform, and infer stages. These functions are mapped across two Intel Arria 10 FPGAs:

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

Data is transferred from the first FPGA to the second FPGA directly. All these functions can be mapped to a single Intel® Stratix® 10 FPGA. Or the solution can use two Intel Stratix 10 FPGAs to offer higher capacity.

In the near future, Megh Computing will build a similar pipeline for text analytics, followed by speech and other pipelines. All these pipelines will leverage the library of Acceleration Function Units (AFUs) provided by Megh and/or integrate FUs developed by third parties.




Megh Computing video analytics pipeline mapping to Intel® Arria® 10 FPGAs

## Sample use case

High volumes of streaming data combined with transactional processing are central to operations and decision-making in areas such as finance, retail, telecommunications, and manufacturing, as well as for the modern enterprise. The Megh and Intel solution is ideal for use cases requiring low latency and fast performance for transmitting, processing, and analyzing streaming data, whether at the edge or cloud.

For example, financial trading analytics must meet a real-time requirement of less than 100 milliseconds. Stock market fluctuations must be visible in real time and portfolios rebalanced based on accurate, up-to-the-minute risk assessments.

Fraud prevention is critical for both finance and retail, requiring transaction validation in milliseconds. Ecommerce and financial companies rely on monitoring of machine-driven algorithms to find variants, helping to detect fraud the moment it happens. In retail, point of sale (POS) terminals must detect inaccurate scans or labeling of SKUs in real time.

IoT is creating myriad new use cases at the edge, including real-time object, image, and facial recognition, augmented and virtual reality, and computer vision applications that require response times in milliseconds to ensure an acceptable quality of experience for end users.

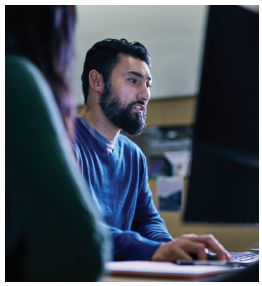
A breadth of applications updating dashboards with real-time data require response times in subseconds, while many traditional operational processes demand nearly instantaneous response times.

### Retail

- Fraud prevention due to mis-scans of SKUs
- Image classification of fresh produce for quality control
- Better inventory management by avoiding product misplacement

### Finance

- Fraud prevention in credit card transactions
- Real-time monitoring of trading transactions

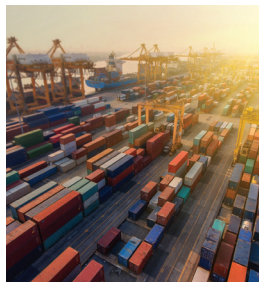


### Telecommunications

- Analyze and adjust network parameters based on network probes
- Real-time monitoring of network data for security checks

### Industrial manufacturing

- Real-time monitoring of production line for operational efficiency
- Predictive asset management



### Enterprise

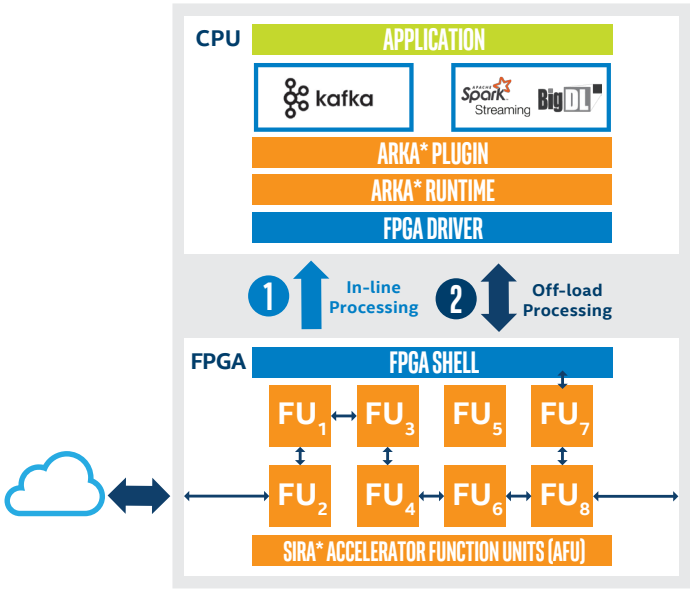
- Security event management
- Analytics of real-time and historical data for improved decision-making

## Solution architecture

In essence, the Megh Computing solution abstracts the complexity of FPGAs via plugin libraries that work with open source and custom frameworks.

Megh combines software and hardware to provide the essential combination of performance and flexibility needed to enable real-time analytics processing of streaming data. The hardware platform is comprised of Intel® Programmable Acceleration Cards (Intel® PACs) with Intel Arria 10 or Intel Stratix 10 FPGAs running on Intel® CPU-based servers. For software, the Spark Streaming framework with BigDL libraries is combined with Megh's libraries for Arka Runtime and Sira Accelerator Function Units (AFUs).

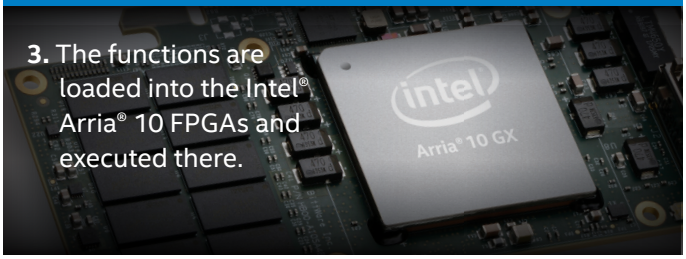
- Arka Runtime exposes the accelerator-as-a-service functions. It manages the Intel® FPGAs and supports software fallback for the AFUs. It also helps build custom data-flow accelerator pipelines and plugins for the frameworks.
- The Sira AFUs deliver the actual acceleration and are implemented as libraries that get downloaded to the Intel FPGA. Applications run unmodified using standard or custom APIs.




The Megh platform acts as an intermediary layer between the application and the Intel® Arria® 10 FPGA, simplifying and accelerating real-time analytics processing on streaming data

**1.** Highly optimized Sira* Accelerator Function Units are exposed to the application by the Arka* Runtime framework via high-level APIs.

**2.** Applications can run unmodified and call the APIs.




**3.** The functions are loaded into the Intel® Arria® 10 FPGAs and executed there.

### Megh's solution offers the following advantages:

- Complete real-time analytics pipeline implemented in the FPGA
  - Fixed low latency to implement the real-time use cases
  - High throughput resulting in lower TCO
- Software abstraction layer for the FPGA to ease deployment using different frameworks
  - Manage and configure FPGAs
  - Expose FPGA accelerators-as-a-service
- Customization tools for rapid time to market
  - Arka and Sira SDKs to develop custom accelerators and pipelines
  - Support services available from Megh Computing

### Intel Arria 10 and Intel Stratix 10 FPGAs

As the demands for big data and AI increase, the reprogrammable technology of the FPGA meets the processing requirements and changing workloads of data center applications. With reconfigurable logic, memory, and digital signal processing blocks, FPGAs can be programmed to execute any type of function with high throughput and real-time performance, making them ideal for many critical enterprise and cloud applications.

The acceleration stack for the Intel® Xeon® processor with Intel FPGAs works with industry-leading operating systems and virtualization and orchestration software, providing a common interface for software developers to get fast time to revenue, simplified management, and access to a growing ecosystem of acceleration workloads.

Like the Intel PAC with Intel Arria 10 FPGA, the newest Intel PAC with Intel Stratix 10 FPGA supports IP to accelerate a wide range of application workloads. The Intel® PAC with Stratix® 10 SX FPGA is a larger form factor card built for inline processing and memory-intensive workloads, like streaming analytics and video transcoding. The smaller form factor Intel PAC with Arria 10 FPGA is ideal for backtesting, database acceleration, and image processing workloads.

## Moving forward

The Megh Computing real-time analytics platform targets use cases that benefit from Intel FPGA acceleration for lower latency and higher throughput. Designed to work with standardized hardware and software, the solution simplifies scaling from a single node to large cluster, as well as deployment in public, private, and edge clouds.

With Megh and Intel, organizations can realize and accelerate the benefits of machine and deep learning while leveraging their current infrastructure.

## Learn more

Discover the Megh Computing real-time analytics platform at megh.com or call us at +1 (888) 428.2396.

Find out more about Intel Programmable Acceleration Cards and Acceleration Stack.

## Solution components

Megh Computing real-time analytics platform and libraries

**Arka* Runtime**

- Expose and manage accelerators-as-a-service
- Support for continuous streams
- Support for multiple FPGAs

**Sira* Accelerator Function Units (AFUs)**

- Packet processing engine (PPE): UDP packet processing and TCP bypass
- Stream Processing Engine (SPE): Data transformation
- Deep learning engine (DLE): Inline inferencing for different topologies

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

Qualified Intel® processor-based servers

## About Megh Computing

Megh Computing's mission is to enable the third wave of computing in the data center with scale-out FPGA-based accelerators. We were founded in 2017 by a team from Intel that pioneered the use of FPGAs in the data center. Over the past ten years, this team has taken the promise of heterogeneous computing with CPU and FPGA platforms from concept to production. Our current focus is leveraging our expertise in full stack development to deliver a platform that accelerates real-time analytics using FPGA accelerators in the public, private, and edge cloud.

megh.com

MEGH COMPUTING

1. marketresearchandstatistics.com/ad/global-streaming-analytics-market.
2. The CPU-only configuration is based on an Intel® Xeon® processor 8000 system with memory and one executor; 46 cores. Power is estimated at 400 W with performance of 182 frames/sec for inferencing using ResNet-50 model with the BigDL library. The FPGA system consists of the Intel® Xeon® processor 3000 system with two Intel® Arria® 10 FPGA cards. Power is estimated at $150/W for the CPU and 3x 50W for the FPGA with performance of 939 frames/sec with ResNet-50 model running on the FPGA.

Performance results are based on testing as of December 2018, and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit intel.com/benchmarks.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com.



0219/MG/CMD/PDF 338557-001US