Accelerated Face Detection on a cluster of FPGAs using InAccel orchestrator

Scalable deployment of Face Detection on a cluster of Xilinx Alveo cards using InAccel orchestrator

Authors:

Ioannis Stamelos
Elias Koromilas
Chris Kachris
InAccel

If you are responsible for building, testing, or deploying face detection or video analytics:

• **As a business strategist or executive:** You will better understand how to apply the latest technologies for deep learning and face detection to successfully generate increase the performance of your system and reduce the cost significantly.

• **As a technology decision-maker:** You will learn how to incorporate a cost-effective deep learning inference framework into your technology stack and at the same time enjoy
• Higher performance
• Higher performance
• Lower Latency
• Lower cost
• Lower energy consumption
• Instant Scalable deployment
• Multi-tenant deployment

Executive Summary

Face Detection to the process of using a specific function in an image or a video frame to identify a face.

Face Detection is used is many application like security, entertainment, retail and other markets.

In this Solution Brief we show how InAccel orchestrator can be integrated with a widely used Face detection to allow multi-tenant scalable deployment of face detection on a cluster of FPGAs.

We show how InAccel's orchestrator allows **easy deployment, scaling, resource management, and task scheduling** for FPGAs making it easier than ever, the deployment and the utilization of FPGA for Face Detection. The same framework can be applied to any other video analytic application.

Face Detection

Automatic Object Detection using machine learning is one of the most promising technologies in the domain of video classification and detection. Object detection in video is computationally intensive task that requires huge amount of processing power. Hardware accelerators, based on FPGAs, can provide the required processing power to increase the throughput of the application and at the same time to reduce significantly the latency.

InAccel, a world-pioneer in the domain of FPGA-based accelerators, has released today an integrated framework that allows to utilize the power of an FPGA cluster for face detection. Specifically, InAccel has presented a demo in which a cluster of 8 FPGAs are used to provide up to 1700 fps (supporting up to 56 cameras with 30 fps in a single server).

Viola Jones face detection algorithm is a widely-used method for real-time object detection The Viola–Jones object detection framework is the first object detection framework to provide competitive object detection rates in real-time proposed in 2001 by Paul Viola and Michael Jones. Although it can be trained to detect a variety of object classes, it was motivated primarily by the problem of face detection. It uses Haar-like features, which are inner products between the image and Haar templates. A face candidate is a rectangular section of the original image. As images may have faces of different sizes, an image pyramid is constructed by downscaling the image by a constant factor. This multiscale representation of image is then searched for all possible 25×25 faces. The inner product of Haar features requires the sum of different rectangular sections of the downscaled image.

Scalable Deployment of Face Detection on a cluster of FPGAs

Nitish Srivastava et al. from Cornell University have presented an implementation for a Xilinx Zynq device. Based on this implementation InAccel has released an integrated framework targeting the Xilinx Alveo cards that allow to scale out the Face detection application in a cluster of 8 Alveo U200 FPGA cards providing a great performance for video applications.

FPGAs are adaptable hardware platforms that can offer great performance, low-latency and reduced OpEx for applications like machine learning, video processing, quantitative finance, genomics, etc. However, the easy and efficient deployment from users with no prior knowledge on FPGA was challenging.

InAccel provides an FPGA resource manager that allows the instant deployment, scaling and resource management of FPGAs making easier than ever the utilization of FPGAs for applications like machine learning, data processing, data analytics and many more applications. Users can deploy their application from Python, Spark, Jupyter notebooks or even terminals.

In the case of face detection, InAccel FPGA manager was used to scale out the application on a server with 8 FPGA cards. The software developers do not need to change at all the original code and the FPGA manager was used to serialize the request from the video streaming and dispatch the job to the FPGA cluster. Using 8 FPGAs, we managed to achieve up to 1700 fps on a single server. That means that a single server can support up 56 cameras (assuming 30 fps) in a single server and still the CPU processor is free for additional processing (supporting more than 56 videos assuming 30fps).

This application can be further scaled-out to multiple server through the Kubernetes plugin. For example, scaling-out to 8 servers it can support up 13,600 fps on a cluster of 8 servers (64 Alveo U200 FPGA cards). The platform was deployed in a cluster provided by VMAccel.

Figure 1. Scalable deployment of Face Detection on a cluster of Alveo U200 FPGA cards using InAccel orchestrator, Click Here and Here to see the relevant videos
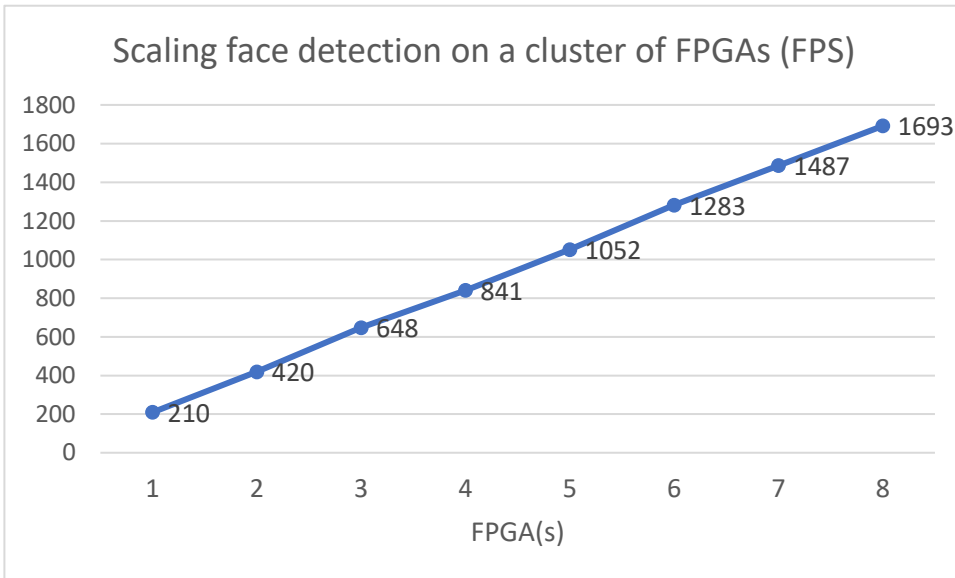
Figure 1. Scaling Face Detection on a cluster of Alveo U200 FPGAs. The linear scaling is based on the efficiency of the InAccel orchestrator.

InAccel helps companies' speedup their applications, with zero code changes using efficiently state-of-the-art accelerators. InAccel provides a unique technology that allows the easy deployment, management, scaling and virtualization of FPGA-based accelerators. InAccel's FPGA orchestrator allows instant deployment and scaling of accelerator for widely-used applications like quantitative finance, big data analytics and machine learning.

Learn more : https://inaccel.com