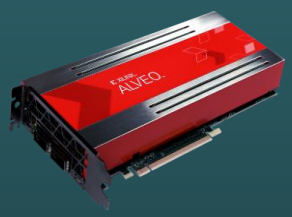
Video & Speech Streaming Analytics

INTRODUCTION

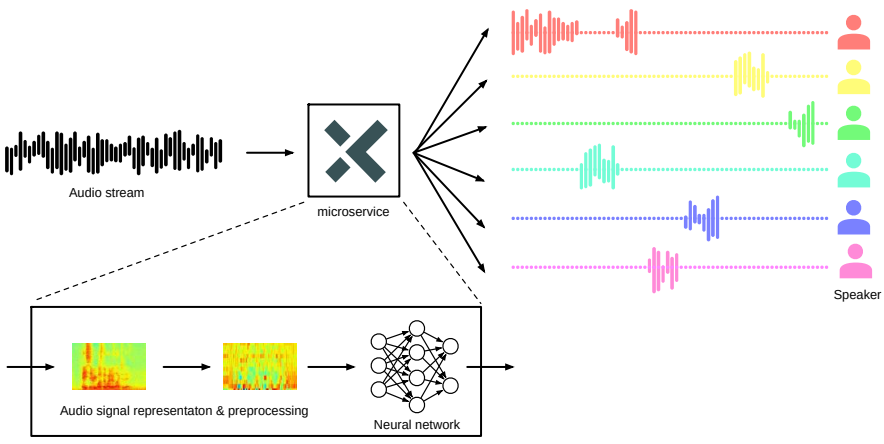
Latency-sensitive applications, such as augmented reality, industrial automation, distributed gaming, real-time analytics and intelligent dialog systems are on the rise, largely promoted by the high-speed connectivity of upcoming 5G networks. These applications are often provided as a Software-as-a-Service (SaaS) in **public clouds**, **on-premises data centers** or **edge clouds** (near-device server resources). Each SaaS application is composed of several building blocks referred to as **microservices**. One commonality of these applications is that they usually require a video or speech streaming analytics microservice. A second commonality is the requirement of a low execution latency and sub-ms latency variance. FPGAs satisfy these latency constraints.



- Video & Speech Streaming Analytics
- Real-time microservice
- Deterministic latency
- Scalable on demand

PRODUCT OVERVIEW

To enable FPGA-based acceleration in the above applications, Xelera Technologies offers the **Xelera Suite Acceleration Software**, one component of which is the **Video & Speech Streaming Analytics plugin**. The plugin is a library of Deep Learning-based inference functions which analyze video frames or audio streams with a low and deterministic latency. These functions serve directly as microservices in the above SaaS applications. Xelera Suite and the Video & Speech Streaming Analytics plugin are available in the public cloud, in on-premises data centers or edge cloud servers. With Xelera Suite's Scale-Out Module, the plugin can be distributed seamlessly on server clusters and scale on demand.



SOLUTION OVERVIEW

Microservice Example: Speaker Diarization

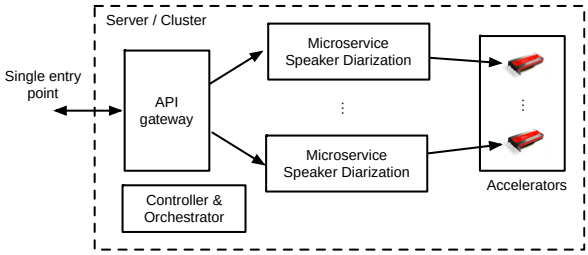
Based on the Video & Speech Streaming Analytics plugin, Xelera provides a real-time audio analytics application, which separates the audio recorded during a discussion of multiple persons into individual audio streams according to the active speaker. At its core, the microservice uses a custom-built neural network to separate and identify individual speakers. The solution scales out to an unprecedented number of concurrent sessions while reliably satisfying the 60 ms latency constraint per session.

ACCELERATE COMPUTING.

XELERA

Video & Speech Streaming Analytics

SCALABLE ON DEMAND



When run in a multi-user SaaS scenario, the number of user sessions that can be served per cloud or edge cloud server is a crucial factor to reduce CAPEX and OPEX. Xelera Suite's Scaling Module invokes dynamically additional accelerators and additional server nodes based on the demanded number of concurrent sessions. The Speaker Diarization microservice has been specifically designed for executing different trained Deep Learning models at the same time.

PERFORMANCE

Edge Cloud Speaker Diarization Use Case Requirements

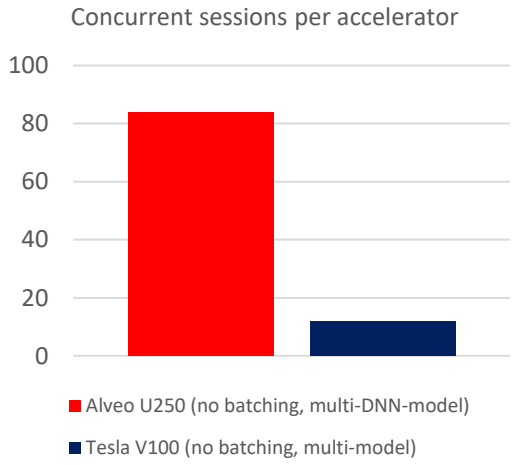
- Multiple user sessions connect asynchronously to the microservices
- Each user session uses a unique set of DNN model weight parameters ('**multi-DNN-model**')
- Each request must be completed within a 60ms latency window
- Maximize the number of concurrent user
- Integration in Keras framework

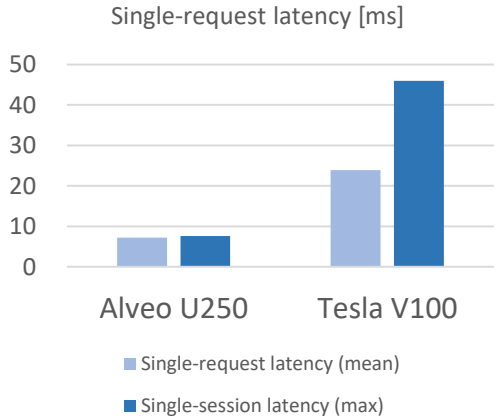
1. Xilinx Alveo U250 accelerator (no batching, multi-DNN-model)

- **84 concurrent sessions** per Accelerator (*)
- Latency per request: **Mean 7.2 ms, Max 7.6 ms** (*)

2. Nvidia Tesla V100 SXM2 (no batching, multi-DNN-model)

- **12 concurrent sessions** per Accelerator (*)
- Latency per request: **Mean 23.9 ms, Max: 45.9 ms** (*)
- Note: The GPU performs better in batching mode (100 concurrent sessions). However, batching does not allow multi-DNN-model setup, which is required in this use case.





COST SAVING

7x better TCO using Alveo U250 compared to Tesla V100 (**)

(*) Benchmark results running on Alveo U250 Dell R740 server vs. NVIDIA Tesla V100 architecture on AWS EC2 p3.2xlarge instance.
(**) Compute to serve up to 1000 concurrent sessions and max 8 Accelerator cards per server

TAKE THE NEXT STEP

Learn more about Xelera Technologies: www.xelera.io
Reach out to the team at info@xelera.io to learn more.

ACCELERATE COMPUTING.