AMD Powers Alibaba Cloud FaaSwith AI Acceleration Solution for E-Commerce Business

Delivers 75% TCO Savings without Compromising Accuracy

AT A GLANCE:

Alibaba Cloud, a subsidiary of Alibaba Group

Industry: Public cloud service provider

Founded: 2009
Website: www.alibabacloud.com

CHALLENGE:

China is home of the world's largest online retail market and Alibaba is China's largest e-commerce company. The amount of product images that Alibaba Cloud handles from its many 3rd party vendors is staggering. In order to maintain a consistent experience on their e-commerce sites, oversight on images is required and results in a huge AI inference compute workload. To reduce operational expenses, Alibaba Cloud was seeking alternative, cost-effective processor solutions to detect harmful or un-wanted text information embedded in tens of millions of images every day.

SOLUTION:

AMD 16nm Virtex™ UltraScale+™ FPGA powered Alibaba Cloud FaaS and AMD Vitis™ AI development kit (formally called MLSuite).

RESULTS:

Achieved 75% savings in total cost of ownership without compromising accuracy. A single AMD UltraScale+ FPGA delivers hundreds of pictures per-second, representing a 3.5X performance improvement over initial GPU implementation.

CHALLENGE:

Massive AI Workloads to Detect Harmful Images
Alibaba Cloud, the cloud computing and data intelligence arm of Alibaba Group, is the No.1 public cloud service provider in Asia Pacific per market share. Alibaba Cloud provides a comprehensive suite of global cloud computing services to power both international customers' online businesses and Alibaba Group's own e-commerce ecosystem.

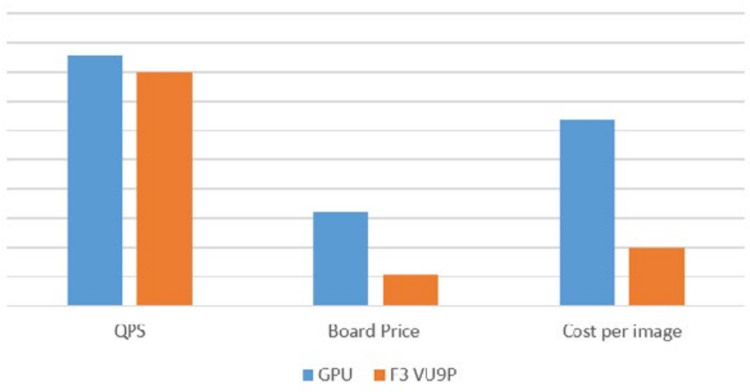
Alibaba Cloud heterogeneous computing FPGA-as-a-Service (FaaS) platform runs a large-scale FPGA instance, the F3 instance, based on AMD 16nm Virtex UltraScale+ VU9P FPGA to support customers inside and outside Alibaba Group.

A large portion of today's internet traffic consists of images. Some images contain harmful unwanted text information such as unpaid advertisements, which have negative impacts on the paid advertisement business. In order to maintain a consistent experience on e-commerce sites, oversight on images is required and creates a large AI inference compute workload.

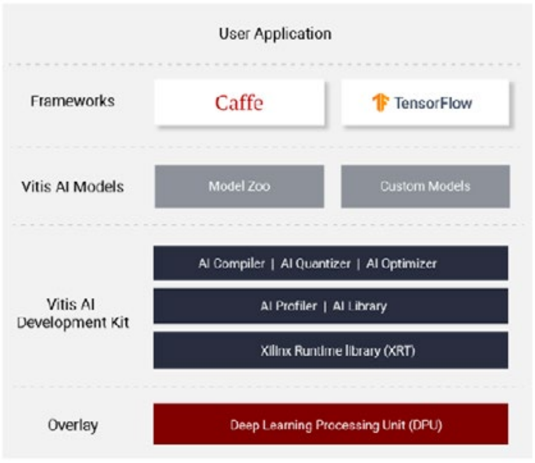
SOLUTION:

Directly Accelerate Yolo-v2 on AMD FPGAs from ML Framework
Alibaba historically used GPUs to run Yolo-v2 Tiny with Float32 data type in order to understand the content in tens of millions of images every day. As the architecture was not well optimized, the GPU could only achieve limited queries per second (QPS) throughput, which resulted in very high costs in power and server footprint. To reduce operation expenses, Alibaba looked for a more cost-effective solution than GPUs for detecting harmful or un-wanted text information.

With the highly adaptable architecture of AMD FPGAs, the Alibaba Cloud FaaS team ran the Yolo-v2 Tiny model at Int16 to achieve superior QPS performance with similar accuracy to GPUs. Inspired by FaaS, with the similar optimization, GPU can achieve similar QPS; however, the AMD solution is much more cost effective per image because the GPU solution has a much higher TCO. In this project, the Alibaba FaaS team also used Vitis AI to expedite their development.



VU9P Vs GPU for YOLO V2 Tiny



AMD Vitis AI (formally MLSuite)

RESULT:

Delivers 75% TCO Savings without Compromising Accuracy
Vitis AI allows developers to optimize and deploy pretrained DNN models to the AMD FPGA without writing any RTL code. The runtime and shell allow them to benefit from AMD hardware acceleration, without needing to be an FPGA expert.

> "Thanks to MLSuite (now part of Vitis AI), the beauty of how Alibaba and AMD developed AI acceleration solutions is that no one at Alibaba had to become an FPGA expert to use the technology."
>
> **- Mr. Zhang**

Mr. Jeff Zhang, Director of Alibaba Cloud FaaS platform, who led the project and successfully implemented AI acceleration into F3 instance, said: "Alibaba Cloud FaaS provides a unified hardware platform and middleware in the cloud.

With the support from AMD Vitis AI, Alibaba FaaS can significantly reduce development and deployment costs of AI accelerators. Accelerator vendors can provide accelerators as a service to users, eliminating the hardware barriers of acceleration technology. Users can use the acceleration services on demand without having to understand underlying hardware details."

Mr. Zhang also pointed out: "At the beginning, many people were not optimistic about the prospect of FPGA in the field of AI. GPU is convenient to use and supports all frameworks. The success of this project proves that FPGA is quite suitable for specific scenarios in this field, and in particular, it has considerable cost-effective advantages for cloud AI inference. For example, shells on the cloud make development much easier; low width and pruning significantly reduces cost and power; IP such as image sharpening, FFT filters bring extra value to some innovative applications

CONCLUSION:

Overall, Alibaba Cloud is pleased with AMD and believes "through the FaaS platform, together with the vast number of ISV and independent IP developers, FPGA has a great opportunity in the AI inference in the cloud." - Mr. Zhang

ADDITIONAL RESOURCES:

Learn More about Vitis AI
Learn more about Virtex UltraScale+ FPGAs

DISCLAIMERS

The information contained herein is for informational purposes only and is subject to change without notice. While every precaution has been taken in the preparation of this document, it may contain technical inaccuracies, omissions and typographical errors, and AMD is under no obligation to update or otherwise correct this information. Advanced Micro Devices, Inc. makes no representations or warranties with respect to the accuracy or completeness of the contents of this document, and assumes no liability of any kind, including the implied warranties of noninfringement, merchantability or fitness for purposes, with respect to the operation or use of AMD hardware, software or other products described herein. No license, including implied or arising by estoppel, to any intellectual property rights is granted by this document. Terms and limitations applicable to the purchase or use of AMD's products are as set forth in a signed agreement between the parties or in AMD's Standard Terms and Conditions of Sale.