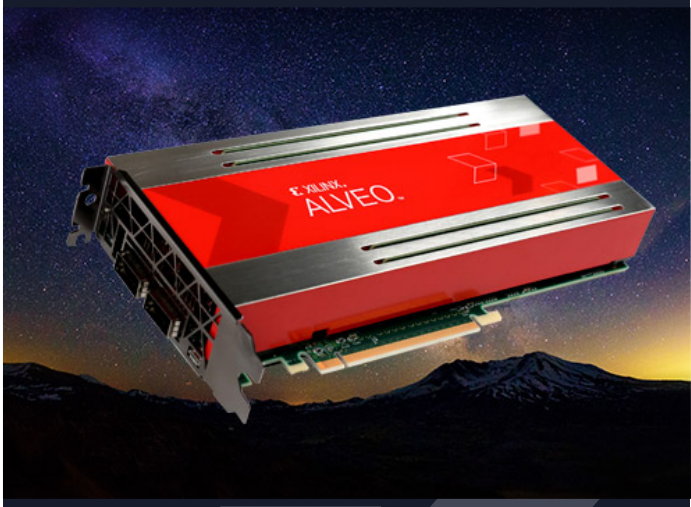
Xilinx ML Suite Accelerates AI/ML on Alveo Data Center Accelerator Cards

SOLUTION BRIEF

XILINX ALVEO™



- 4X Higher Real-Time Inference over High-End GPUs
- 20X Higher Real-Time Inference over High-End CPUs
- Easy-to-Use ML Frameworks, APIs, and Software

INTRODUCTION

Machine Learning is a rapidly evolving technology that is re-inventing traditional applications and simultaneously inspiring new use models. These new applications deployed in the cloud and at on-premises Data Centers must perform in real-time, making it difficult for accelerators to keep pace. Applications including Cloud Surveillance Analytics, Satellite Imaging, Bioinformatics, Financial Modeling, and Network Security require Machine Learning acceleration as well as other application-specific workloads involving OpenCV, Video Transcoding, and Smart NICs.

PRODUCT OVERVIEW

Xilinx's Machine Learning Suite running on Xilinx® Alveo™ Data Center accelerator cards delivers the highest real-time inference available with easy-to-use software tools to quickly deploy any ML application.

SOLUTION OVERVIEW

Xilinx Machine Learning Suite

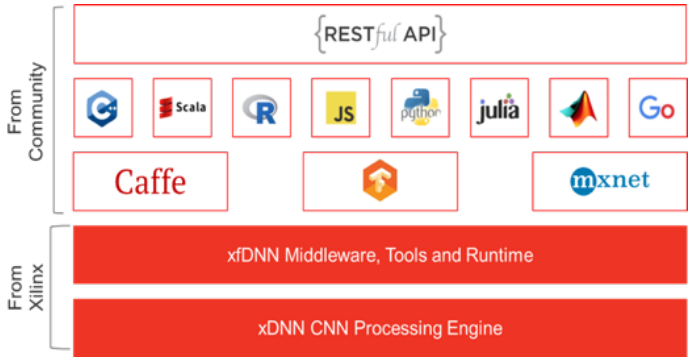
Xilinx ML Suite enables developers to optimize and deploy accelerated ML inference. ML Suite supports the most prevalent machine learning frameworks including Caffe, MxNet, and Tensorflow, as well as Python and RESTful APIs.

ML Suite Components:

> xfDNN Compiler/Optimizer: auto-layer fusing, memory optimization, and framework integration

> xfDNN Quantizer: Improves performance with auto model-precision INT8 calibration

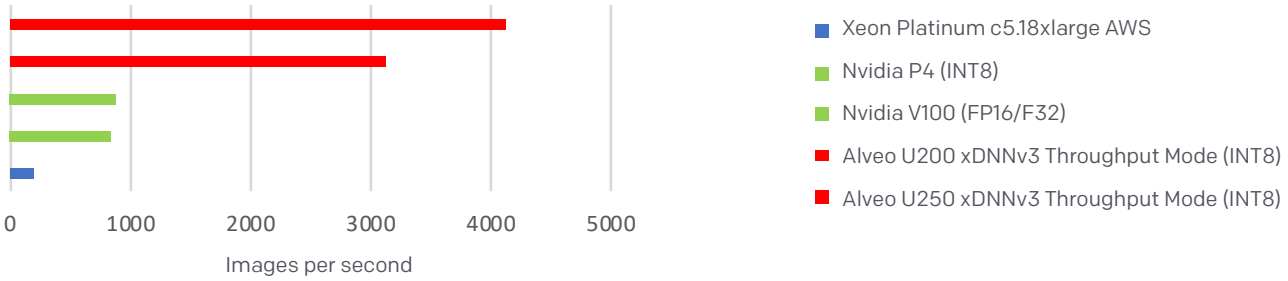
> Deployable on-premises or through cloud services: Amazon, Nimbix, Huawei, Alibaba Cloud, Baidu, and Tencent



HIGHEST REAL-TIME INFERENCE

Xilinx Alveo U200 and U250 accelerator cards are capable of high-performance, energy-efficient real-time DNN inference. An Alveo U250 accelerator card running xDNN processing engines are capable of delivering more than 4,000 images per second of GoogLeNet v1 throughput at low latency without requiring batching. This incredible low latency throughput is unlocked with the new xDNN processing engine, available in the ML Suite.

Real-Time GoogLeNet Throughput



CONCLUSION

Artificial Intelligence and Machine Learning are rapidly evolving disciplines with increasing demand for acceleration. Xilinx Alveo Data Center accelerator cards running Xilinx ML Suite delivers the highest real-time performance available today. ML Suite is adaptable to meet changing needs and delivers easy-to-use software tools and resources to quickly integrate Machine Learning into Data Center applications for rapid deployment.

TAKE THE NEXT STEP

Visit www.xilinx.com/alveo to learn more about the Xilinx ML Suite on Alveo Data Center accelerator cards

For additional Xilinx ML Suite resources, visit www.xilinx.com/ml.

To get started with the Xilinx ML Suite today, download from the Xilinx Github Page, https://github.com/Xilinx/ml-suite.

Corporate Headquarters
Xilinx, Inc.
2100 Logic Drive
San Jose, CA 95124
USA
Tel: 408-559-7778
www.xilinx.com

Xilinx Europe
Xilinx Europe
Bianconi Avenue
Citywest Business Campus
Saggart, County Dublin
Ireland
Tel: +353-1-464-0311
www.xilinx.com

Japan
Xilinx K.K.
Art Village Osaki Central Tower 4F
1-2-2 Osaki, Shinagawa-ku
Tokyo 141-0032 Japan
Tel: +81-3-6744-7777
japan.xilinx.com

Asia Pacific Pte. Ltd.
Xilinx, Asia Pacific
5 Changi Business Park
Singapore 486040
Tel: +65-6407-3000
www.xilinx.com

India
Xilinx India Technology Services Pvt. Ltd.
Block A, B, C, 8th & 13th floors,
Meenakshi Tech Park, Survey No. 39
Gachibowli(V), Seri Lingampally (M),
Hyderabad -500 084
Tel: +91-40-6721-4747
www.xilinx.com