ALVEO for Machine Learning Inference

AVNET SILICA

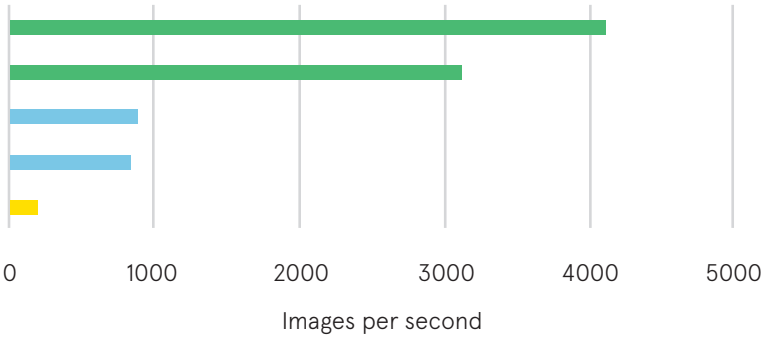
ADAPTABLE ACCELERATOR CARDS FOR MACHINE LEARNING INFERENCE IN THE CLOUD OR ON-PREMISES

Xilinx Alveo U50, U200, U250 and U280 accelerator cards are capable of delivering high-performance, energy-efficient real-time DNN inference. In addition to the DDR Memory, Alveo U50 and U280 offer 8 GB of integrated HBM2 memory delivering very high memory bandwidth (460 GB/s with U280 and 316 GB/s with U50). An Alveo U250 accelerator card running xDNN (Xilinx Deep Neural Network) processing engines can deliver more than 4,000 images per second of GoogLeNet v1 throughput at low latency. This incredible low latency throughput is unlocked with the new xDNN processing engine, available in the ML Suite.
The Xilinx ML Suite (Machine Learning Suite) enables developers to optimize and deploy accelerated ML inference. ML Suite supports popular machine learning frameworks including Caffe, MxNet and Tensorflow, and offers Python and RESTful APIs.

KEY BENEFITS

- Deliver highest real-time inference
- Machine learning inference for video processing with any workload using the same accelerator card
- Reconfigurable hardware allows to adapt to evolving algorithms and provides a future proof architecture
- Deploy solutions in the cloud or on-premises interchangeably, scalable to application requirements
- Applications are available for common workloads
- Developer Tools are available for C/C++ and OpenCL/OpenCV

Real-Time GoogLeNet throughput



PARTS

Alveo U50	Alveo U200		Alveo U250		Alveo U280	
• 16,2 Peak INT8 TOPs • 8 GB HBM2 Memory • 316 GB/s HBM2 Memory Bandwidth • 24 TB/s Internal SRAM Bandwidth • 872K LUTs	• 18,6 Peak INT8 TOPs • 64 GB DDR Memory • 77 GB/s DDR Memory Bandwidth • 31 TB/s Internal SRAM Bandwidth • 892K LUTs		• 33,3 Peak INT8 TOPs • 64 GB DDR Memory • 77 GB/s DDR Memory Bandwidth • 38 TB/s Internal SRAM Bandwidth • 1.341K LUTs		• 24,5 Peak INT8 TOPs • 32 GB DDR Memory • 38 GB/s DDR Memory Bandwidth • 8 GB HBM2 Memory • 460 GB/s HBM2 Memory Bandwidth • 30 TB/s Internal SRAM Bandwidth • 1.079K LUTs	
A-U50-P00G-PQ-G U50 Acceleration card with passive cooling	A-U200-P64G-PQ-G	U200 Acceleration card with passive cooling	A-U250-P64G-PQ-G	U250 Acceleration card with passive cooling	A-U280-P32G-PQ-G	U280 Acceleration card with passive cooling
	A-U200-A64G-PQ-G	U200 Acceleration card with active cooling	A-U250-A64G-PQ-G	U250 Acceleration card with active cooling	A-U280-A32G-PQ-G	U280 Acceleration card with active cooling

Visit **https://xilinx.com/alveo** and **https://github.com/Xilinx/ml-suite**

Contact your local Avnet Silica office or send your request to **artificial-intelligence@avnet.eu**

XILINX

FSC Mixed Sources
Product group from well-managed forests and other controlled sources
www.fsc.org Cert no. IC-COC-100065
© 1996 Forest Stewardship Council