FPGA Hardware Acceleration in Electronic Trading, AI, and Data Analytics

Matt Certosimo,
Field Application Engineering Manager, AMD

May 18, 2023

AMD Adaptive Computing Portfolio for Fintech

- Alveo™ X3522PV for low latency (100-1,000ns)* trading and risk analysis with 644MHz F_{MAX}
- Alveo U55C, Alveo U200/250, VCK5000 cards for analytics, accelerated algo trading, and AI

AMD ALVEO

X3522PV

U55C

U200 / U250

VCK5000

	Low Latency Trading	Compute & Analytics	Accelerated Algo Trading	Accelerated Algo Trading
Network Interface	4x 10/25Gb	2x 100G	2x 100G	2x 100G
Logic Resources	1M LUTs[1]	1.3M LUTs	1.2M – 1.7M LUTs	900K LUTs
Form Factor	HHHL	FHHL	Full-Height, ¾ Length	Full-Height, ¾ Length
DDR / HBM Memory	-	16GB HBM	64GB DDR4	32GB DDR4 + 8GB HBM

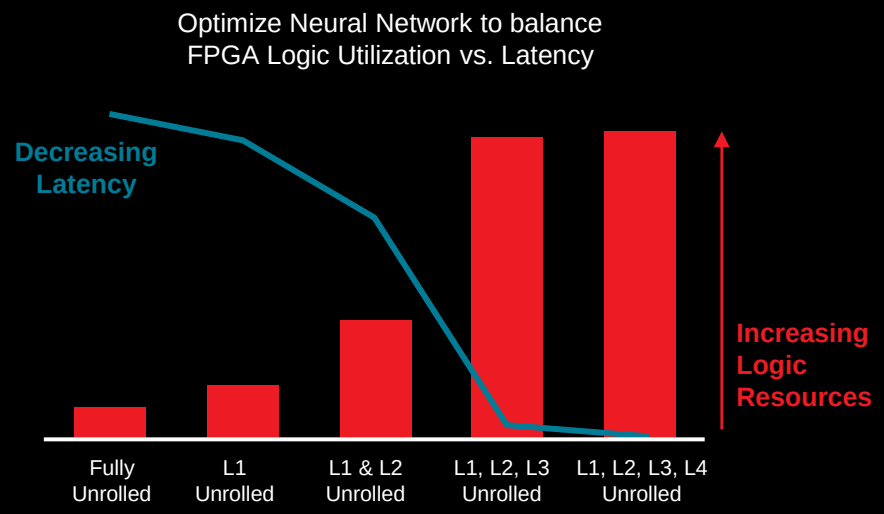
1: -2 screened to -3 speed grade specifications

*Not a STAC benchmark

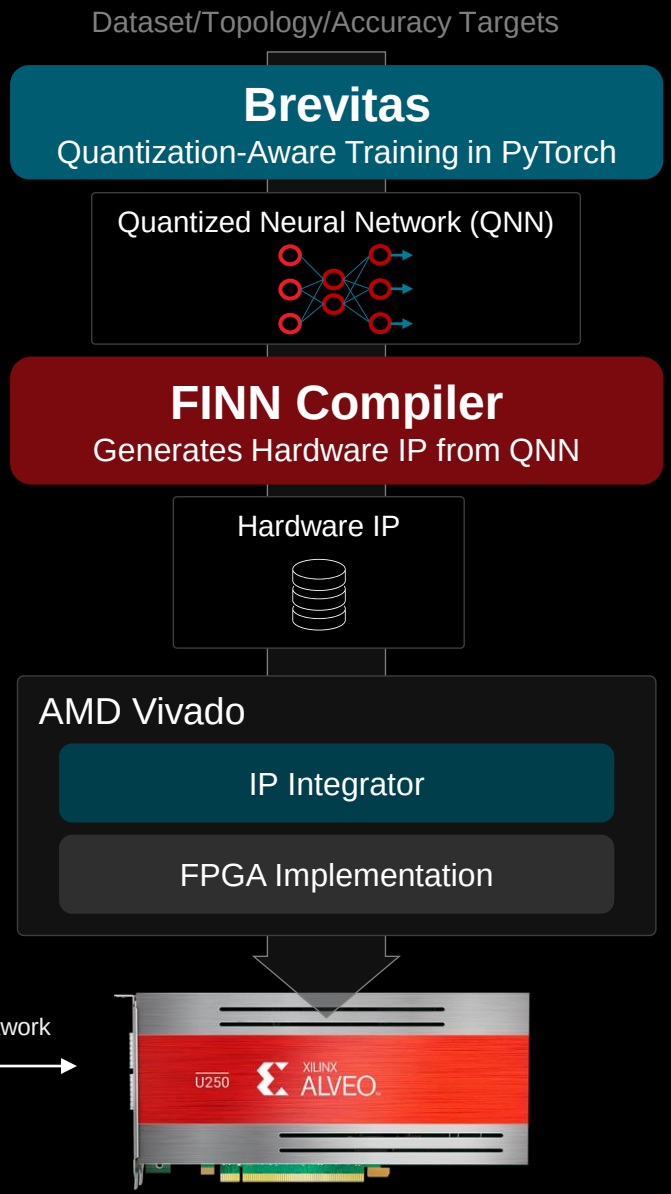
AMD

Generate Low Latency Streaming Neural Networks Using FINN

- FINN AMD Opensource Project[1]
  - Enables ‘streaming’ AI accelerators, integrated directly into datapath
- Brevitas Python Library for Neural Network Training in PyTorch
  - Uses quantization-aware training, customizable for datapath requirements
- FINN compiler generates FPGA IP from Quantized Neural Network
- System Integration with AMD Vivado™ FPGA Flow

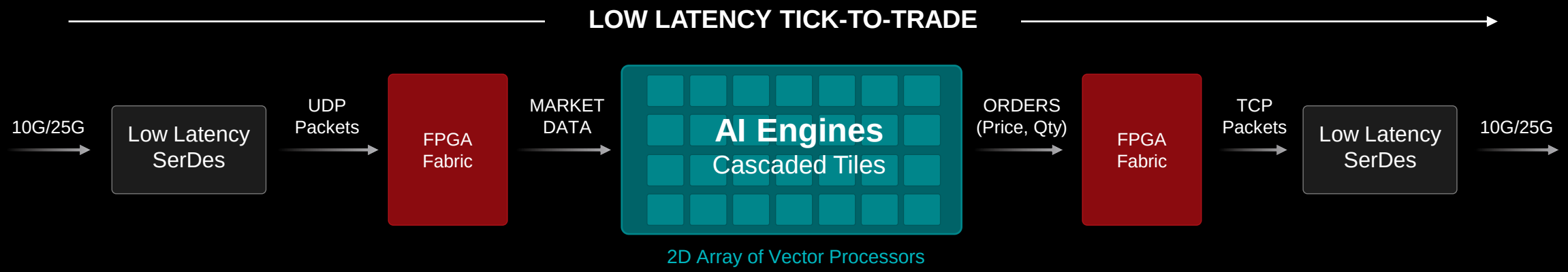


Order book pushed to neural network



1: https://xilinx.github.io/finn/

VCK5000 Card for Low Latency Inference with AI Engines



VCK5000 Development Card

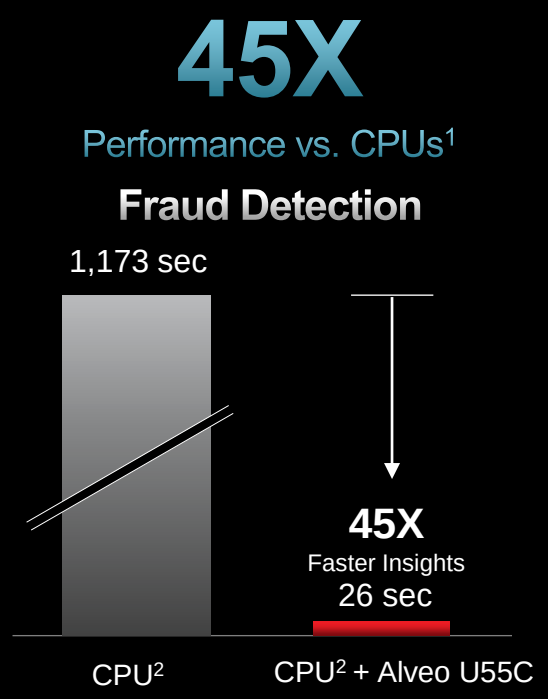
Accelerator Card Features

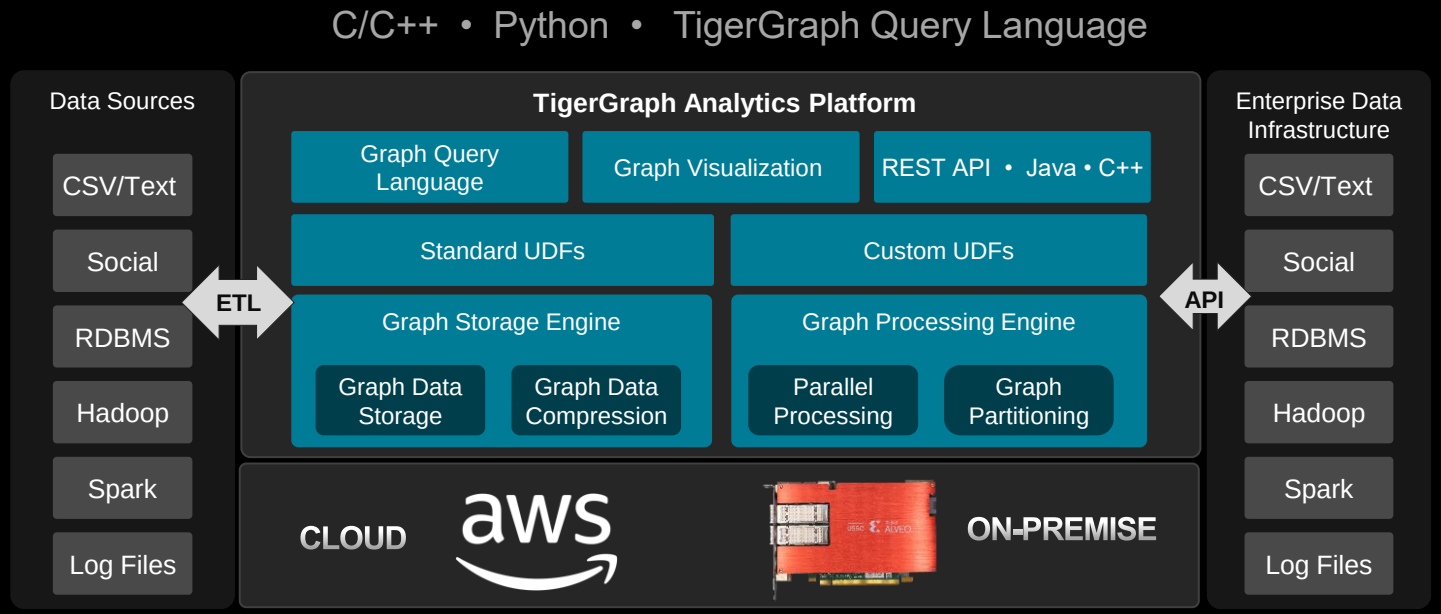
Versal™ Device	Versal AI Core VC1902, -3 screened -2MSE
Logic Resources	900K LUTs
Compute (AI Engines)	145 TOPS (INT8)
Embedded Processors	Dual-Core Arm® Cortex®-A72
Network Interface	2x 100G (QSFP28)
PCIe®	Gen4 x8, Gen3 x16
Power (Max TDP)	225W
Memory	16GB DDR-3200

AMD

Hardware Accelerated, Real-Time Analytics for Faster Insights

- TigerGraph Accelerated Graph Machine Learning (AGML) library on Alveo™ U55C & Amazon EC2 F1 instance
- Accelerates applications such as fraud detection, credit scoring, wealth management, and more
- AGML enables multiple parallel data lookups to accelerator memory which holds the graph database





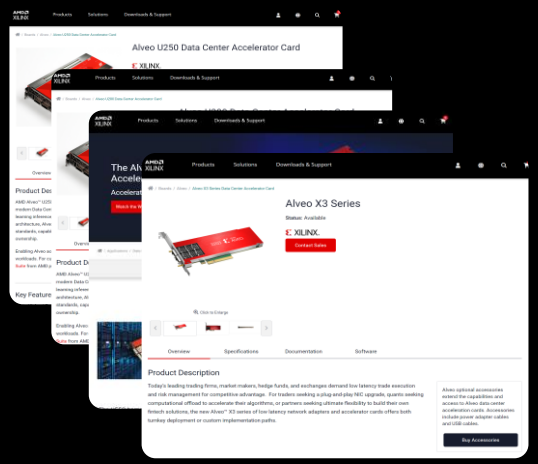
1: Number of Vertices: 125M, Cluster Score: 18% (Louvain Modularity): EPYC 128C / 256T CPU

Not a STAC Benchmark

Get Started Now

Adaptive Accelerator Cards

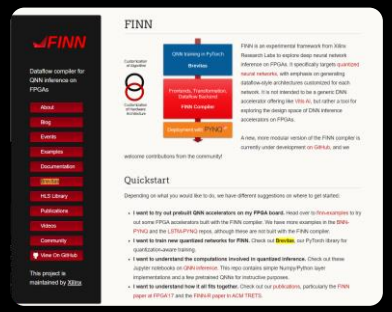
www.xilinx.com/alveo



FINN Project

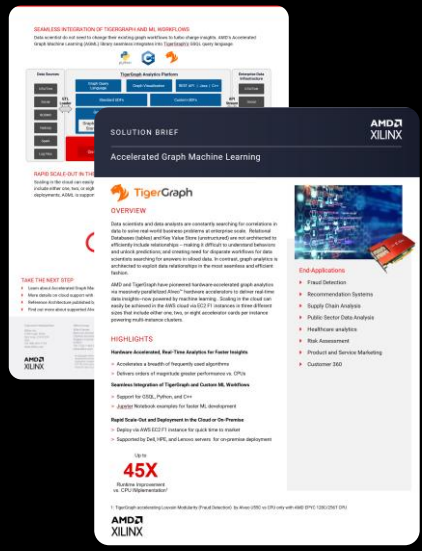
Quantized Neural Networks

https://xilinx.github.io/finn/



Graph Analytics

with TigerGraph




AMD