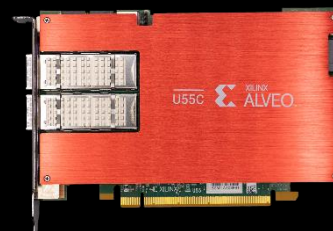# AMD Adaptive Computing Portfolio for Fintech

- 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
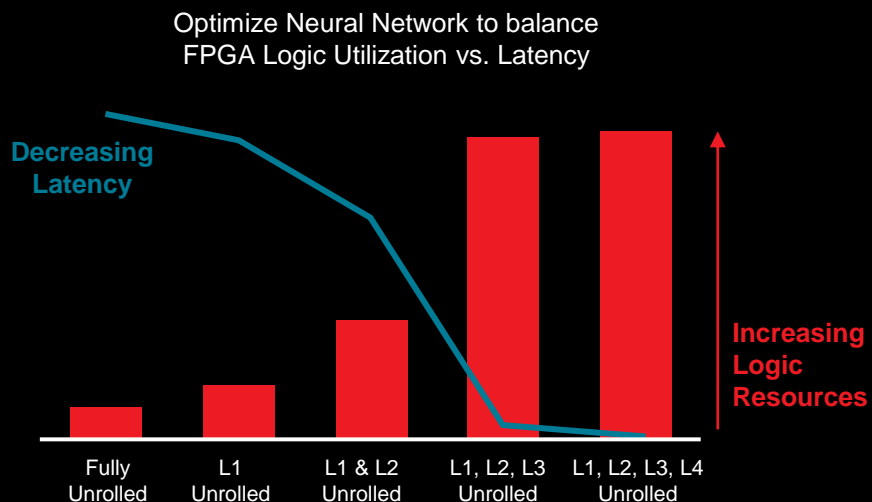
AMD ALVEO

	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

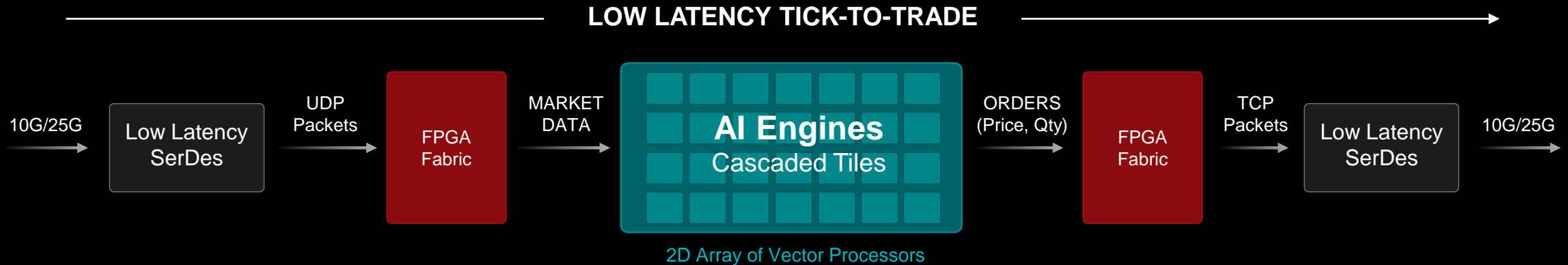1: -2 screened to -3 speed grade specifications

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

Dataset/Topology/Accuracy Targets

Brevitas
Quantization-Aware Training in PyTorch

Quantized Neural Network (QNN)

FINN Compiler
Generates Hardware IP from QNN

Hardware IP

AMD Vivado

IP Integrator

FPGA Implementation

Optimize Neural Network to balance
FPGA Logic Utilization vs. Latency

Decreasing
Latency

Increasing
Logic
Resources

| Fully Unrolled | L1 Unrolled | L1 & L2 Unrolled | L1, L2, L3 Unrolled | L1, L2, L3, L4 Unrolled |

ORDER BOOK
(Market Data)

Order book pushed to neural network

XILINX
U250 ALVEO

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

3

VCK5000 Card for Low Latency Inference with AI Engines

LOW LATENCY TICK-TO-TRADE

10G/25G → **Low Latency SerDes** → UDP Packets → **FPGA Fabric** → MARKET DATA → **AI Engines** Cascaded Tiles → ORDERS (Price, Qty) → **FPGA Fabric** → TCP Packets → **Low Latency SerDes** → 10G/25G

2D Array of Vector Processors

VCK5000 Development Card

AMD

Accelerator Card Features

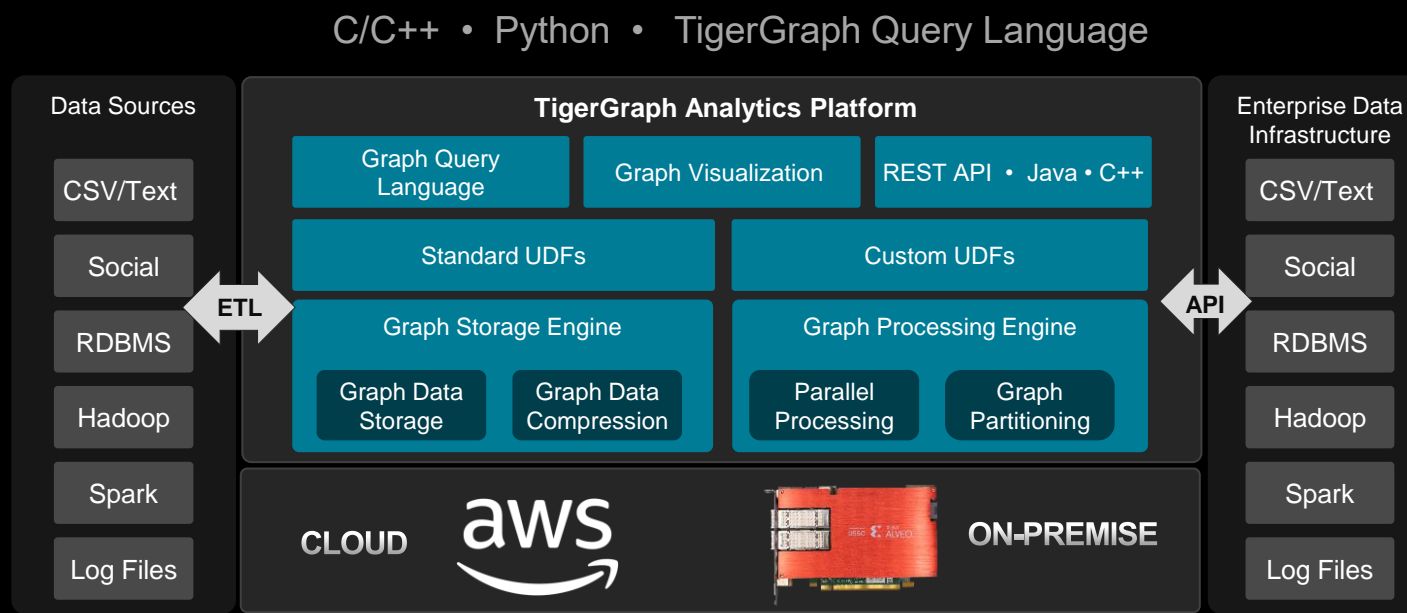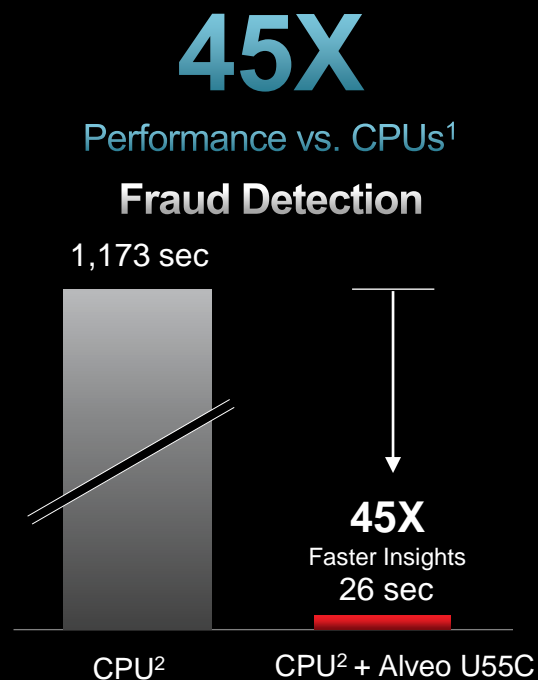|	
Versal™ Device	Versal AI Core VC1902, -3 screened -2MSE
Logic Resources	900K LUTs
Compute (AI Engines)	145 TOPS (INT8)
Embedded Processors	Dual-Core Arm® Cortex®-A72
Network Interface	2x 100G (QSFP28)
PCIe®	Gen4 x8, Gen3 x16
Power (Max TDP)	225W
Memory	16GB DDR-3200

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

45X
Performance vs. CPUs[1]
Fraud Detection

1,173 sec

45X
Faster Insights
26 sec

CPU[2]   CPU[2] + Alveo U55C

C/C++ • Python • TigerGraph Query Language

Data Sources	TigerGraph Analytics Platform	Enterprise Data Infrastructure
CSV/Text	Graph Query Language · Graph Visualization · REST API • Java • C++	CSV/Text
Social	Standard UDFs · Custom UDFs	Social
RDBMS	Graph Storage Engine · Graph Processing Engine	RDBMS
Hadoop	Graph Data Storage · Graph Data Compression · Parallel Processing · Graph Partitioning	Hadoop
Spark		Spark
Log Files		Log Files

ETL   API

CLOUD   aws   ON-PREMISE

1: Number of Vertices: 125M, Cluster Score: 18% (Louvain Modularity): EPYC 128C / 256T CPU
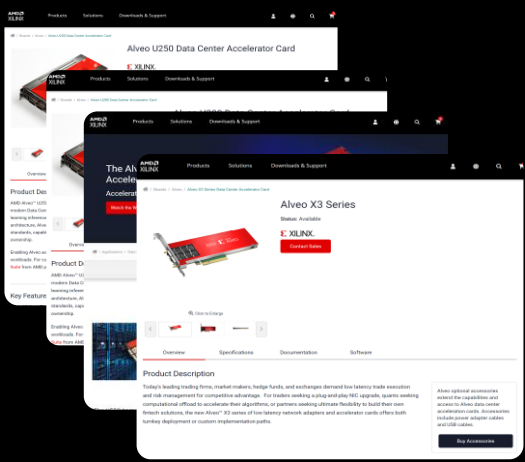**Not a STAC benchmark**

5

AMD
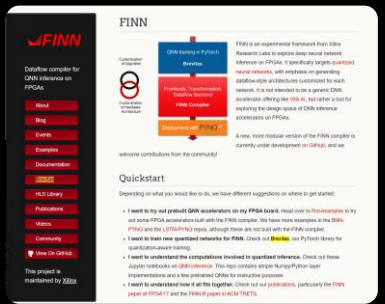
Get Started Now

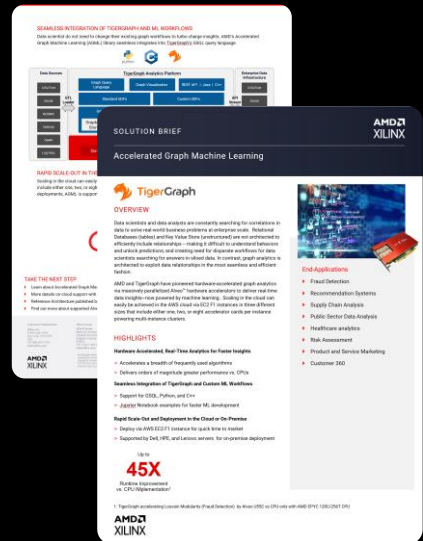## Adaptive Accelerator Cards
www.xilinx.com/alveo

FINN Project
Quantized Neural Networks
https://xilinx.github.io/finn/

Graph Analytics
with TigerGraph

AMD