



# Low Latency AI with the Alveo V80 Accelerator

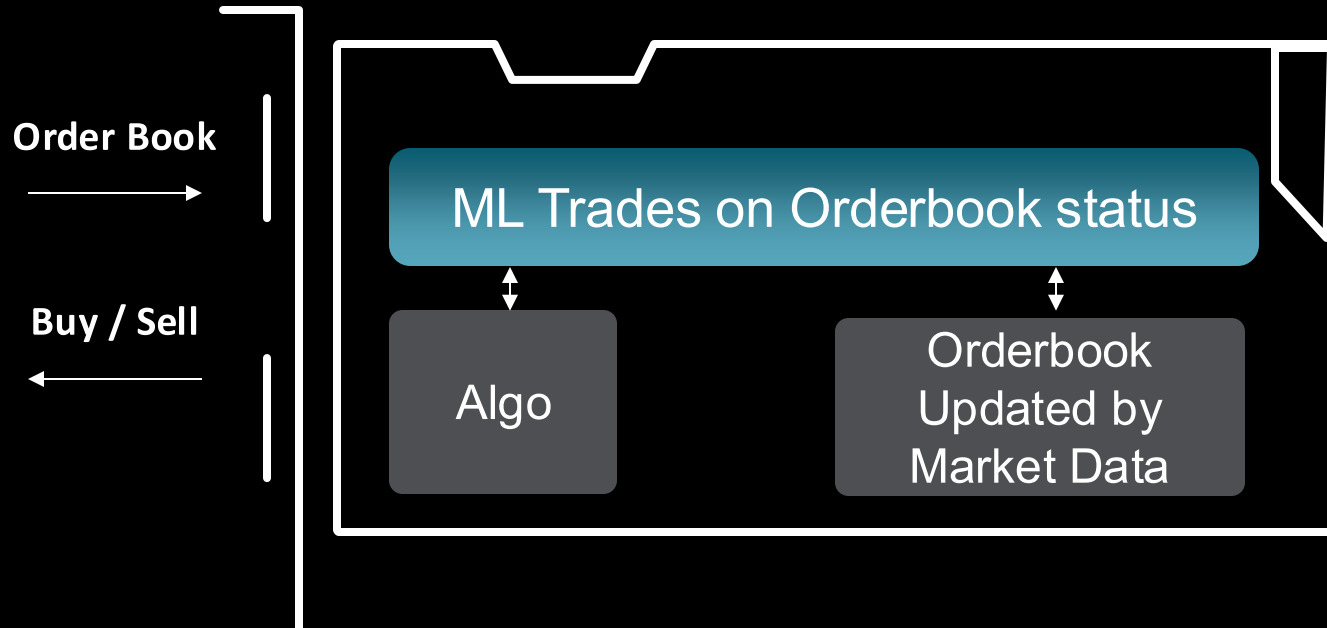
**John Courtney**  
Product Marketing Manager, Fintech



# Electronic Trading is an Ideal Segment for AI/ML

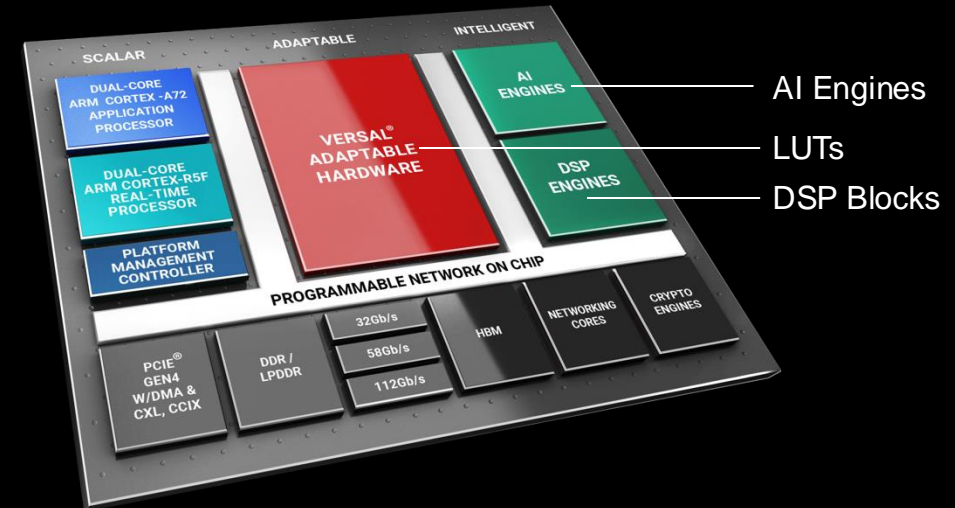
- ML algorithms operate on numerical representations of financial instruments to offer predictive value
- Training data is readily available
- FPGA low latency solutions can integrate in-line ML models within trading pipeline
- Can be implemented in traditional DSP blocks, or AI Engines (vector processor)
- Focus today is on Fabric w/ DSP58s

ML in FPGA fabric



Multiple Implementation Approaches for ML

Versal™ Adaptive SoC Example



# AMD Alveo™ V80 Specifications

## 800G Networking

- 4x200G
- QSFP56 ports

## 7nm AMD Versal Architecture

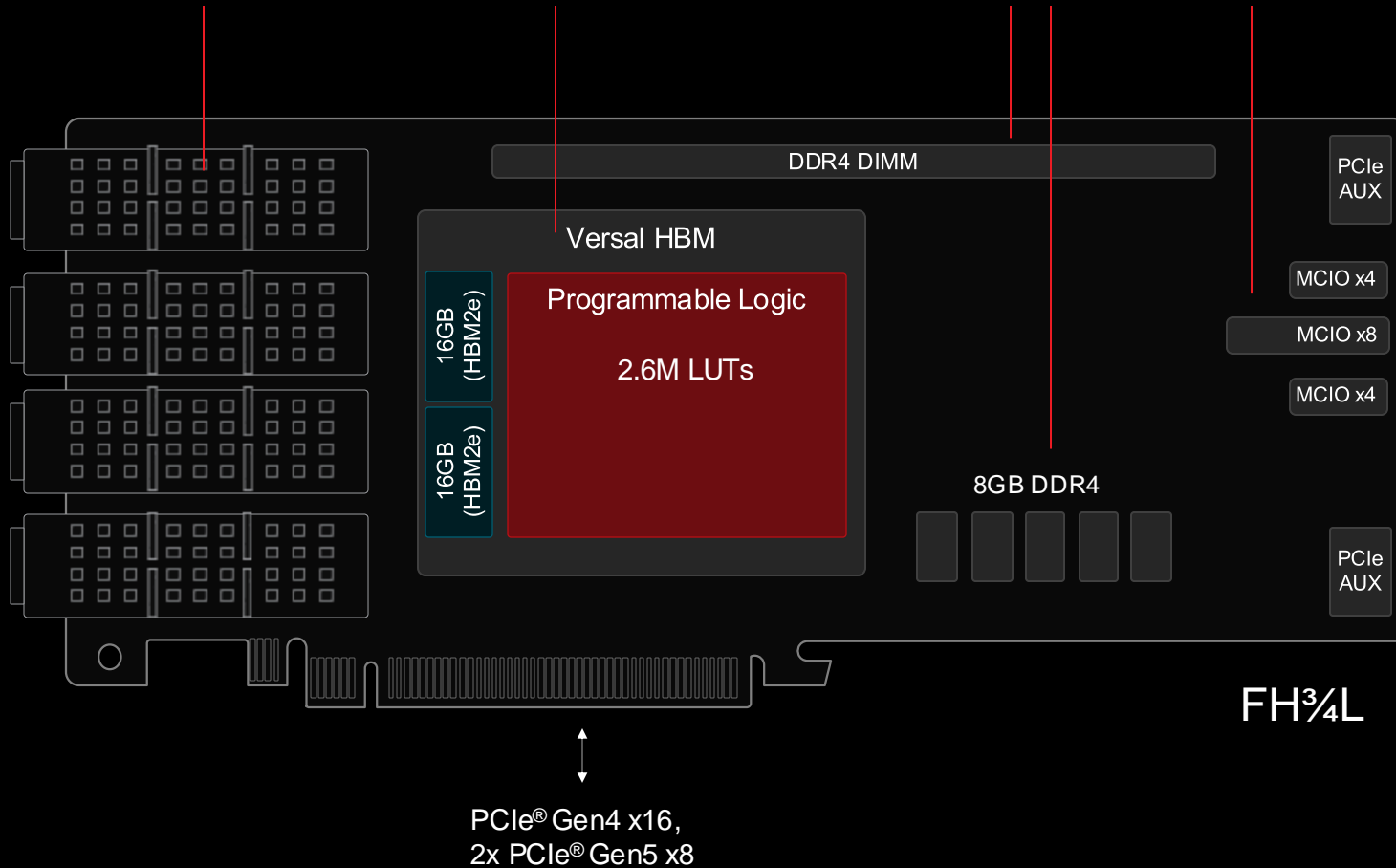
- 2.6M LUTs for flexible compute
- 10.9K DSP slices
- 32GB HBM at 820GB/s

## On-Board DDR

- 8GB for Arm® processor management
- DIMM Expansion slot

## MCIO Expansion

- PCIe® Gen5 connectivity
- Connect to NVMe

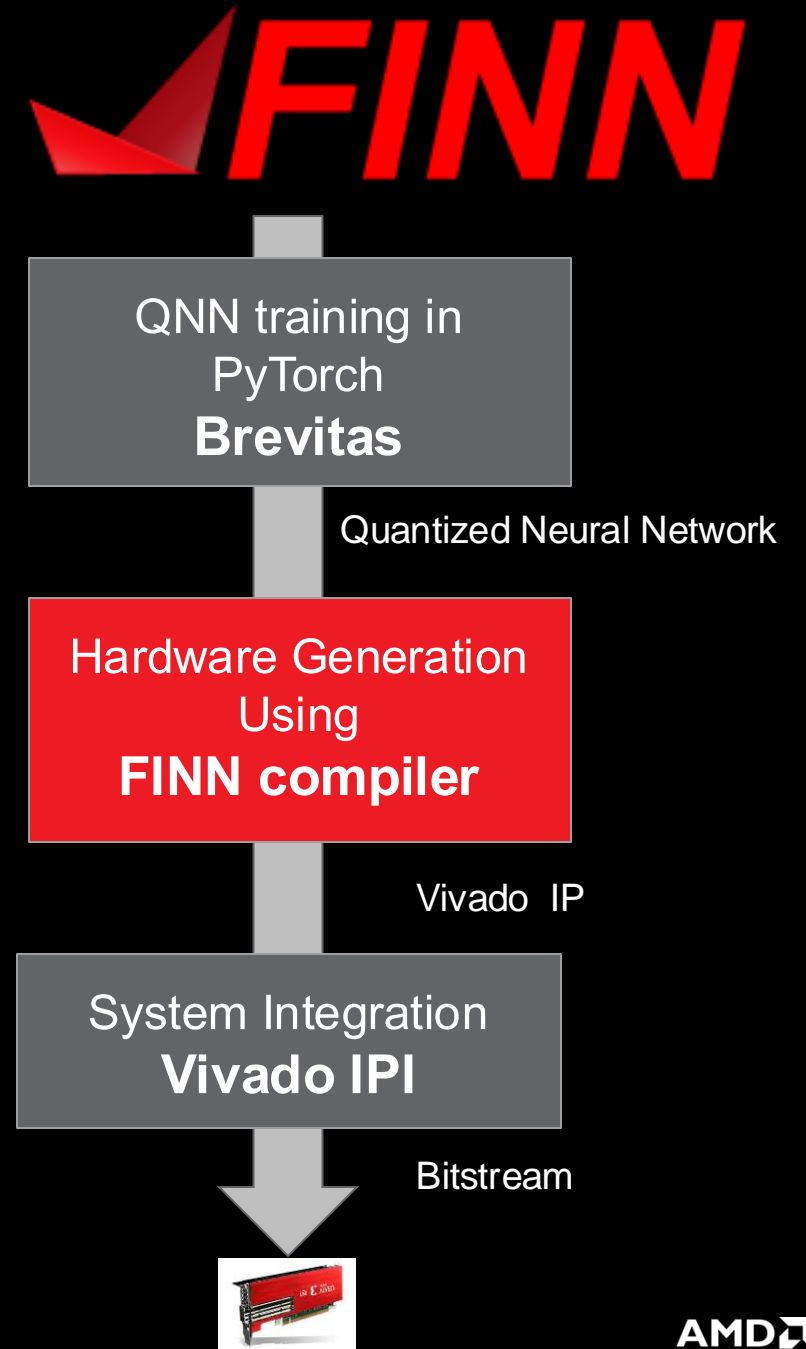


FEATURES	SPECIFICATION
Device Architecture	XCV80 (AMD Versal™ HBM adaptive SoC)
Logic Density	2.6 Million LUTs
HBM Capacity	32GB
DSP58's	10,848
HBM Bandwidth	820GB/s
DDR4 Capacity	32GB
Network Interface	<ul style="list-style-type: none"> <li>• 4x200G (QSFP56)</li> <li>• Per port: 2x100G or 4x 10/25/40/50G</li> </ul>
Expansion	PCIe® Gen5 over MCIO connectors
Form Factor	Full-height, ¾ Length (FH 3/4L), Dual-Slot
PCIe® Interface	PCIe® Gen4 x16, 2x Gen5 x8
Power (Electrical)	300W
Power (Thermal)	190W <sup>1</sup> , passively cooled
Software	<ul style="list-style-type: none"> <li>• AMD Vivado® Design Suite (RTL)</li> <li>• AMD Alveo Versal Example Design (AVED)</li> </ul>

1: Total thermal power (TDP) is device and server dependent

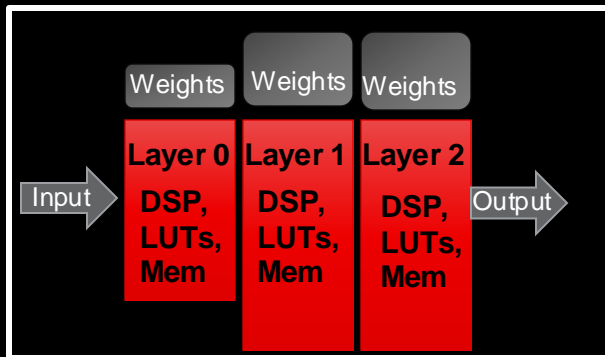
# FINN – Project Mission

- Custom Specialization
  - for creating **high-throughput, ultra-low-latency** DNN inference engines
- End-to-End
  - flow for the **easy** creation of **specialized hardware architectures** for FPGAs
  - Estimates Model latency and resource usage
- Open Source
  - for full **transparency and flexibility** to adapt to end user applications and
  - for easy customer interactions
- Business units providing customer support
  - Lead engineering team: Custom and Strategic Engineering, Dublin



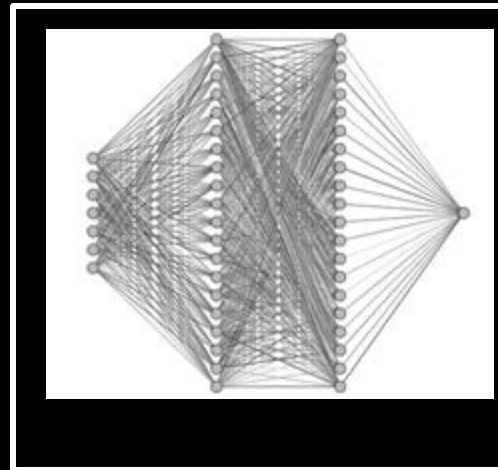
# FPGA Dataflow Architecture Value

## FPGA Neural Net (FINN)

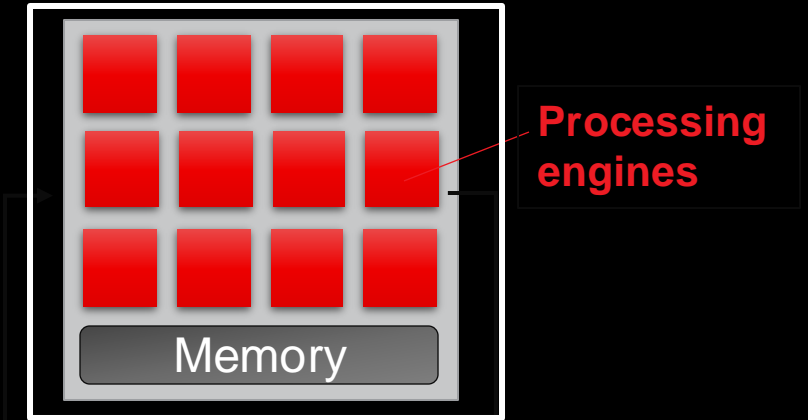


- FPGA fabric solution
- Optimized for low latency inference
  - Latency = ( #layers \* clk cycles per layer)
  - FPGA embedded memory ideal for weights
- Very high throughput
- Dataflow architecture
  - Each layer feeds the next layer (no layer to layer buffering)
- Fully unfolded = FPGA DSPs dedicated to individual neuron
- Partially folded = FPGA DSPs implement more than 1 neuron
- Ideal for mixed precision

## Generic Neural Net



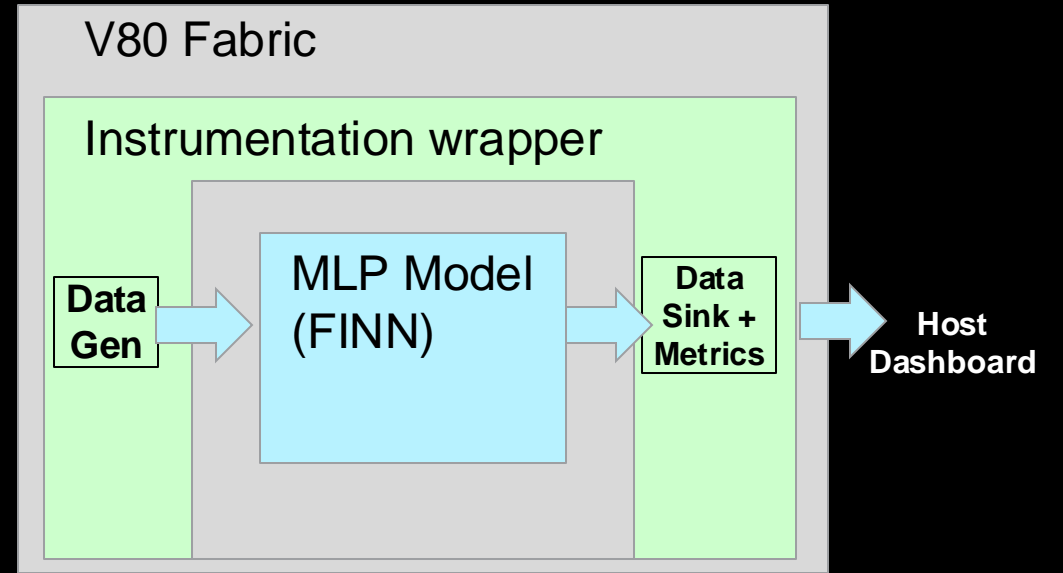
## Matrix of Processing Engines (AI Engine)



- ASIC, GPU generic solution
- Iterative “layer-by-layer” compute loaded onto fabric and results stored in memory
- Specialized processing engines
  - Operators
  - ALU types: tensor, matrix or vector based
- Designed to run any DNN
- Well suited for deep models (large)

# FINN Example on V80

- Trained model (external framework) compiled with FINN for target to FPGA.
- Model: MLP which receives orderbook status messages (input) and generates buy or sell as output
- Instrumentation wrapper simulates the data source and sink (for convenience)
  - Reports actual Latency and throughput from hardware
- Model performance and evaluations supported:
  - Latency
  - Time to first inference
  - Throughput (inferences per second)
  - What is biggest model supported on device
  - What is the biggest model that has a latency of < 500ns



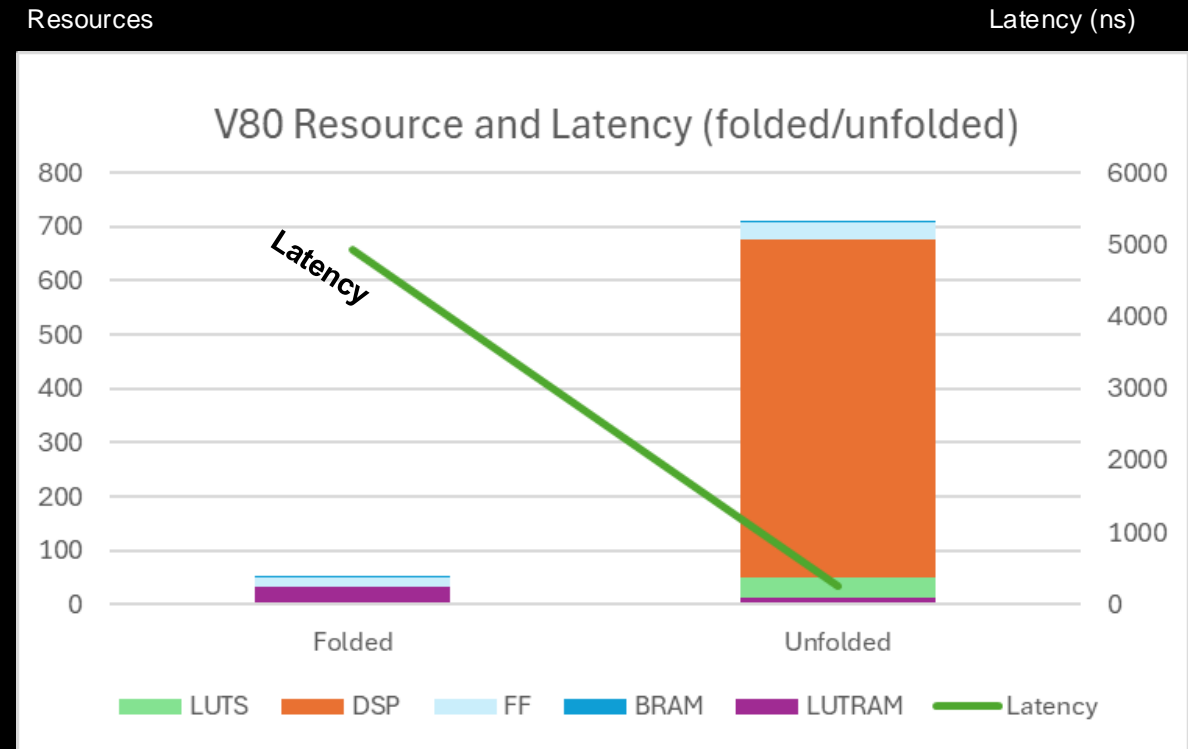
## Example Output

```
FINN Instrumentation Summary
-----
Samples      :          10
Status       :           00
Latency avg  :    220 cycles |      733 ns
Latency min  :     51 cycles |      170 ns
Interval avg :     12 cycles |       40 ns
Checksum    :      7FED85
-----
```

Resource	Utilization	Available	Utilization %
LUT	13261	2574208	0.52
LUTRAM	651	1287104	0.05
FF	43519	5148416	0.85
BRAM	67.50	3741	1.80
DSP	368	10848	3.39
MMCM	1	13	7.69

# Results on V80 – Latency and unfolding

- MLP with 3 Fully Connected layers (7 : 20 : 20 : 20 :1)
- Fully folded
  - Lowest resources consumed
  - Highest latency
- Fully unfolded
  - Higher resources consumed
  - Lowest latency
- Development and testing is on-going
  - Latency achieved
    - Unfolded: 49 clk cycles = 245ns @200M and 156ns @320M
    - Folded: 987 clk cycles = 5us @ 200M and 3.2us @320M
  - Lowest latency achieved with
    - Maximum unfolding - lowest clock cycles per layer
    - Highest clock frequency



\* FF and LUTS are in 1k units

**AMD** 