



FPGAS ACCELERATING AI FOR FINANCIAL SERVICES

Intel® Network & Custom Logic Group (NCLG)

Intel[®] AI Hardware – Device, Edge, and Multi-Cloud

OPTIMIZED SOFTWARE
STACK

CPU

FPGA

GPU

ACCELERATORS



WORKLOAD BREADTH

AI SPECIALIZATION

Multi-purpose foundation for artificial intelligence (AI)

Real-time deep learning inference and more

Highly-parallel media, graphics and compute

Multi-modal deep learning inference

Edge media and vision inference

Deep learning training

All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.

¹Unified software stack development in progress

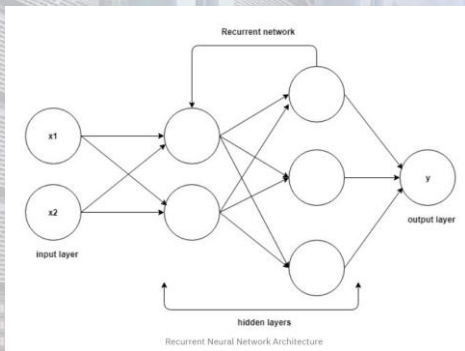
www.intel.ai/technology

Scale Your Innovation



RNNs for Financial Applications

Recurrent neural networks (RNNs) are neural networks with memory



Fraud detection

Anti-money laundering

Speech recognition



Requirements

- Low latency for real-time response
- High memory-to-compute ratio for increased performance

Intel® FPGAs Well Suited to Address RNN Workloads

Intel® FPGAs enabling technologies

- Pipelining
- Many large independent local memories
- Independent DSP



What this means for RNN applications:

Delivering batch 1 performance **54 TOPs at 34.9 W**

High memory bandwidth **Up to 58TB/s**

Delivering unstructured sparsity **96% sparsity**

If you would like to hear more, stop by our stand today

RNN Demo at Booth

	Myrtle results
Platform	Intel® FPGA PAC D5005 ¹
Sparsity (%)	96
Batch Size	1
Effective Throughput (TOPS)	54.0
Power (W)	34.9
Performance per Watt (Effective GOPS/W)	1547
Latency per 1s input audio (mS)	0.343

Not STAC Benchmarks

1. Intel® Programmable Acceleration Card (Intel PAC) measurements taken in conjunction with Intel i7-7700K at 4.20 GHz, RAM 4 * 16 GB at 2,800 MHz, 1 TB M.2 PCIe* SSD, PRIME Z270-P motherboard, 650 W PSU, Ubuntu
2. Peak throughput of 53.37 TOPS measured over shorter input duration of 200 ms, When measuring latency over a 1s input period, peak throughput drops to 23 TOPS

Speech Transcription

This demo showcases Speech to Text conversion using Myrtle's Recurrent Neural Network accelerator running on an Intel® Stratix® 10 FPGA. [Details](#)

Inference stream	On demand inference
49.93 TOPS	3843.5X 35 W 4.65% WER
the ideas also remain but they have become types in nature forms of men animals birds fishes	the ideas also remained but they have become types in nature forms of men animals birds fishes
so for the hundredth time she was thinking today as she walked alone up the lane back of the barn and then slowly down through the bottoms	so for the hundredth time she was thinking to day as she walked alone up the lane back of the barn and then slowly down through the bottoms
the analysis of knowledge will occupy us until the end of the thirteenth lecture and is the most difficult part of our whole enterprise	the analysis of knowledge will occupy us until the end of the thirteenth lecture and is the most difficult part of our whole enterprise
thought the fir tree and believed it all because the man who told the story was so good looking well well	thought the fir tree and believed it all because the man who told the story was so good looking well well

Legal Notices, Copyrights, and Disclaimers

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No product or component can be absolutely secure. Check with your system manufacturer or retailer or learn more at [intel.com](https://www.intel.com).

Intel, the Intel logo, are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries.

*Other names and brands may be claimed as the property of others.

© Intel Corporation

