

Project Brainwave

Serving DNNs in Real Time at Datacenter Scale with
Project Brainwave

Microsoft – March 2018

ELEC 502

Isabelle André

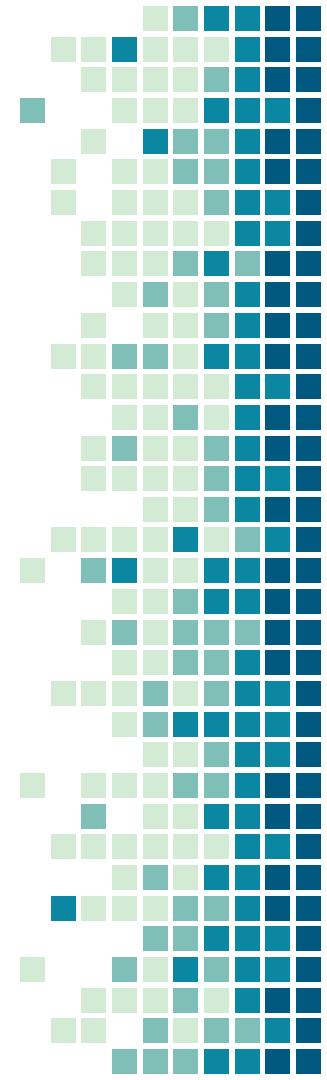


Outline

- Project Brainwave Background
- Brainwave Architecture
 - Layers & Stack
 - Translating high-level models onto Brainwave NPUs
 - Soft NPU Architecture, Microarchitecture
- Brainwave Use Cases within Bing

Background

- Cloud service providers deploy DNN-infused apps ingesting live data streams
 - Web search queries
 - Speech to text
 - Translation
- Real-time latency constraints limit size, complexity, and quality of models that can be deployed on conventional hardware at datacenter scale
- GPGPUs and batch-oriented NPUs

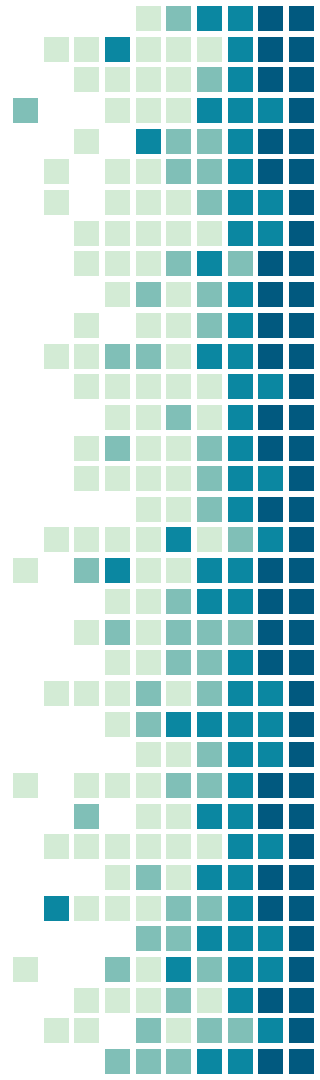


“Project Brainwave

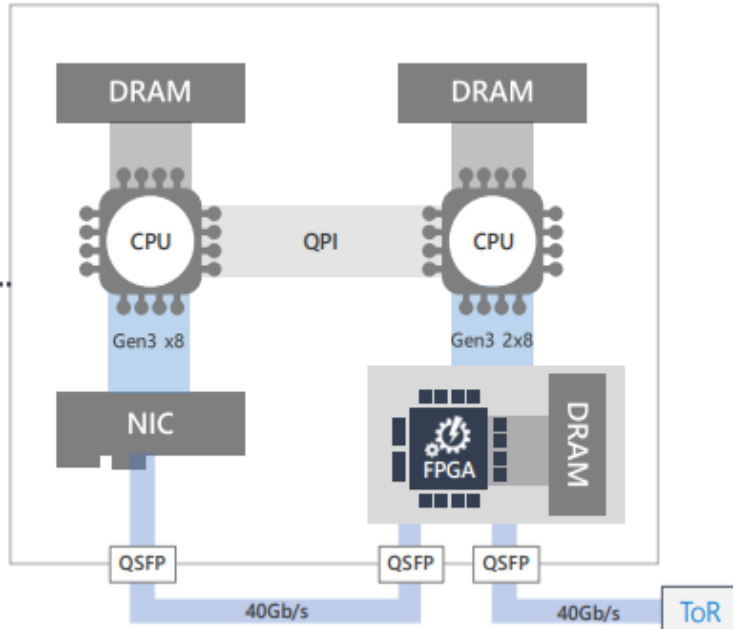
Microsoft's platform for
deployment and accelerated
serving of DNN models in
real-time at low cost

Brainwave Key Components

- Fabric of FPGAs attached directly to the datacenter network
 - Allocated as shared hardware microservice
 - Callable by any CPU on shared network
 - Scalable architecture
- NPU hosted on each FPGA
 - Adaptable in numerical precision and supported operators

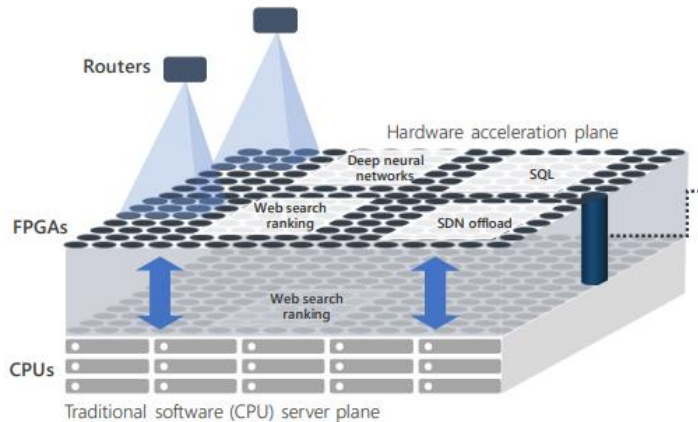


Background: Catapult-Enhanced Servers

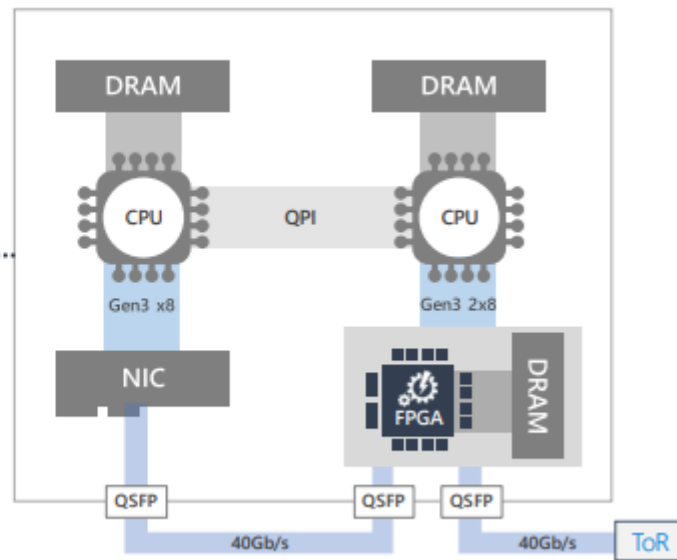
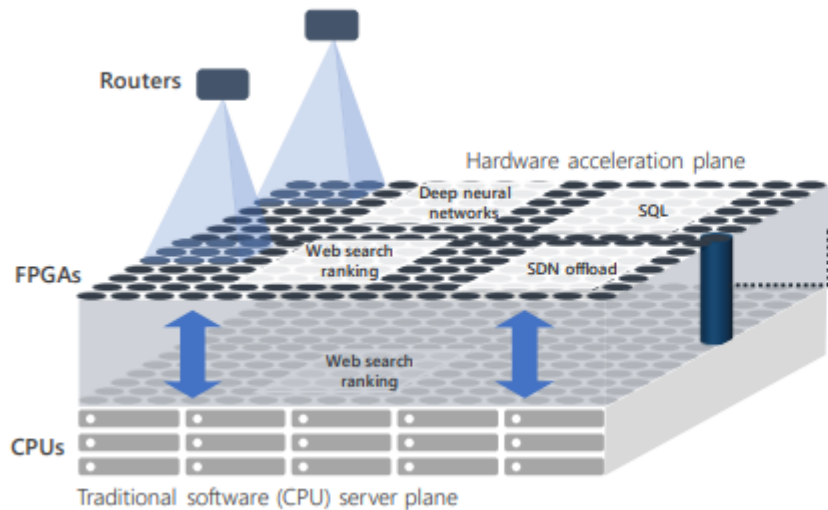


- The Brainwave system is implemented for the Microsoft datacenter architecture with Catapult-enhanced servers.
- Each FPGA operates in-line between server's network interface card (NIC) and top-of-rack (TOR) switch

Background: FPGAs as Shared Microservice



- FPGA proximity to network allows to be sorted into CPU independent resource pools
 - Reclaim underutilized resources for other services by rebalancing subscription between CPUs and FPGAs
 - Support workloads that cannot fit or run effectively on a single FPGA

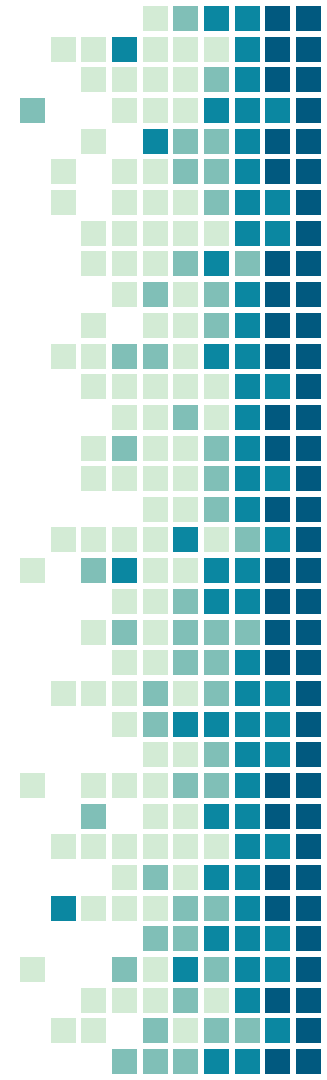


FPGAs allocated as single shared
HW Microservice

Dual CPUs with PCI2-attached
FPGA

Brainwave Architecture

- To achieve ultra-low latency serving of DNN models while preserving accuracy:
 - Model parallelism
 - On-chip pinning at scale
- DNN models are split into sub-graphs to fit into on-chip FPGA memory
- If the FPGA memory is exhausted, more FPGAs can be allocated for the remaining parameters

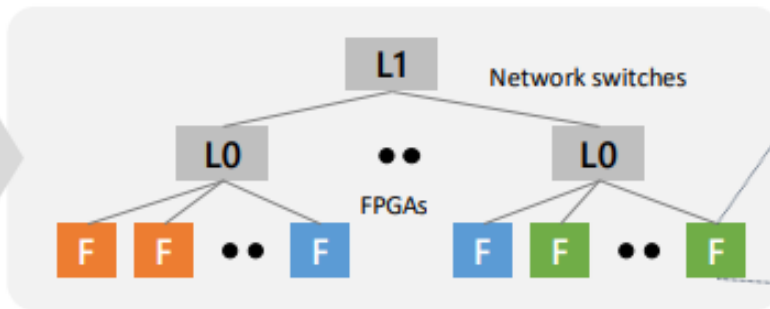


Brainwave Stack

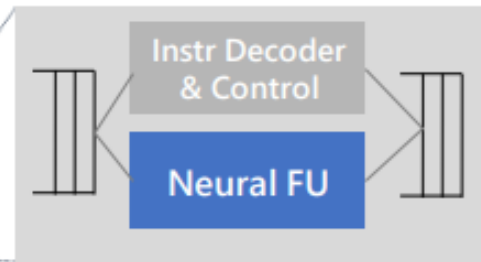
- Flow and runtime for low friction deployment of trained models
- Distributed system architecture mapped onto CPUs and ,microservices
- DNN processing unit synthesized onto FPGA



Pretrained DNN Model
in CNTK, etc.



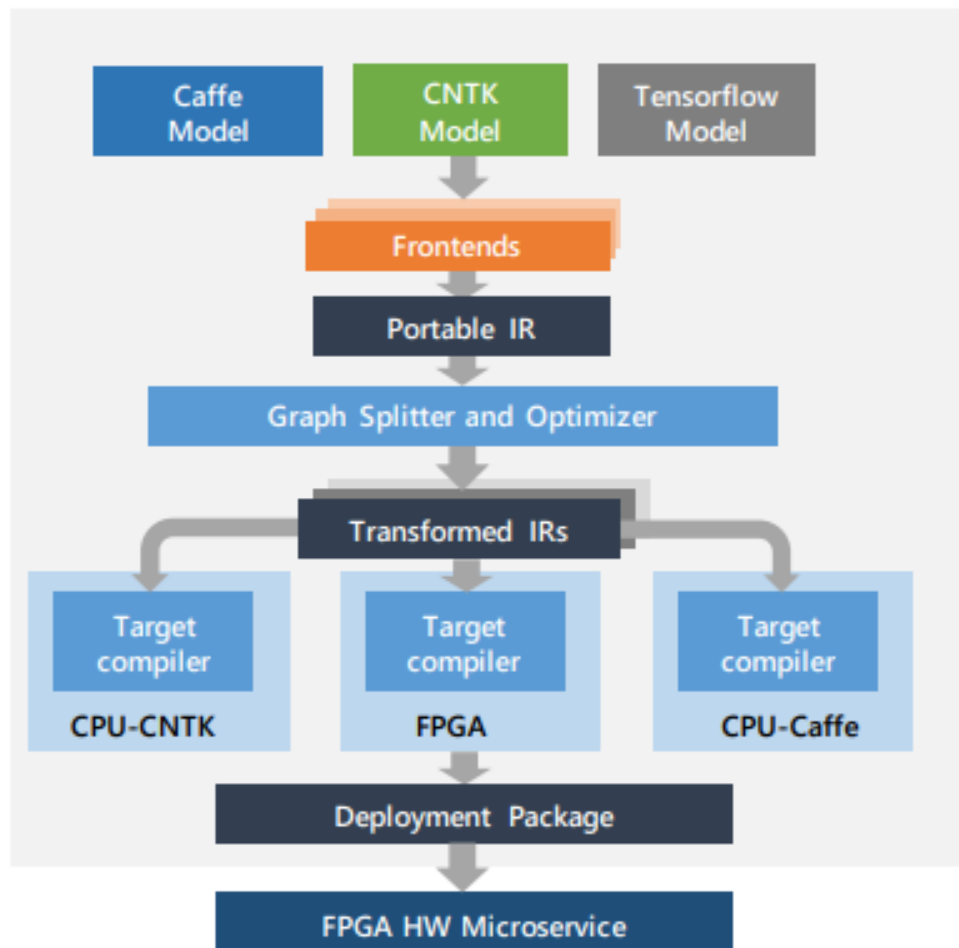
Scalable DNN Hardware
Microservice



BrainWave
Soft NPU

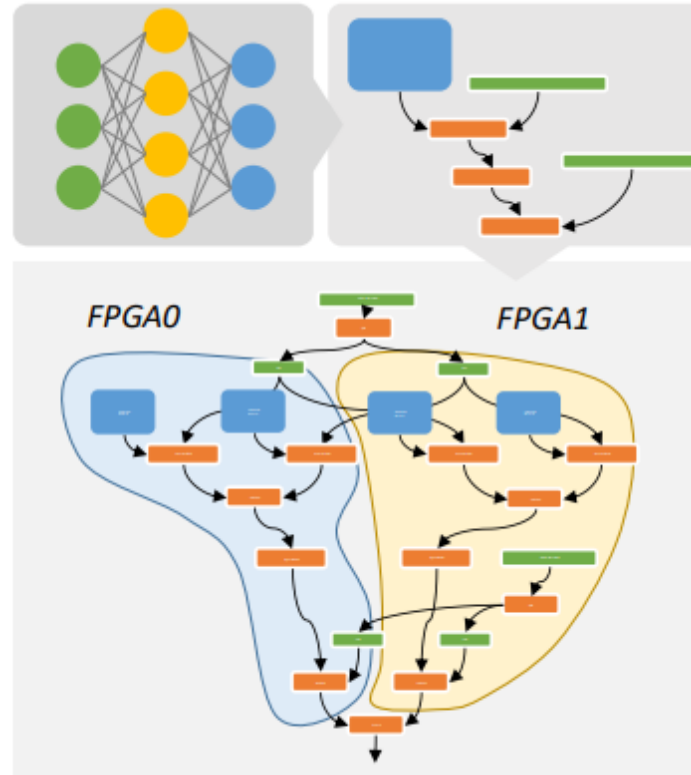
Tool Flow

- Accepts models from different DNN toolchains and frameworks
- Exports into a common graph representation
- Partition into sub-graphs assigned to different CPUs and FPGAs
- Deployed into a live FPGA hardware microservice



Tool Flow

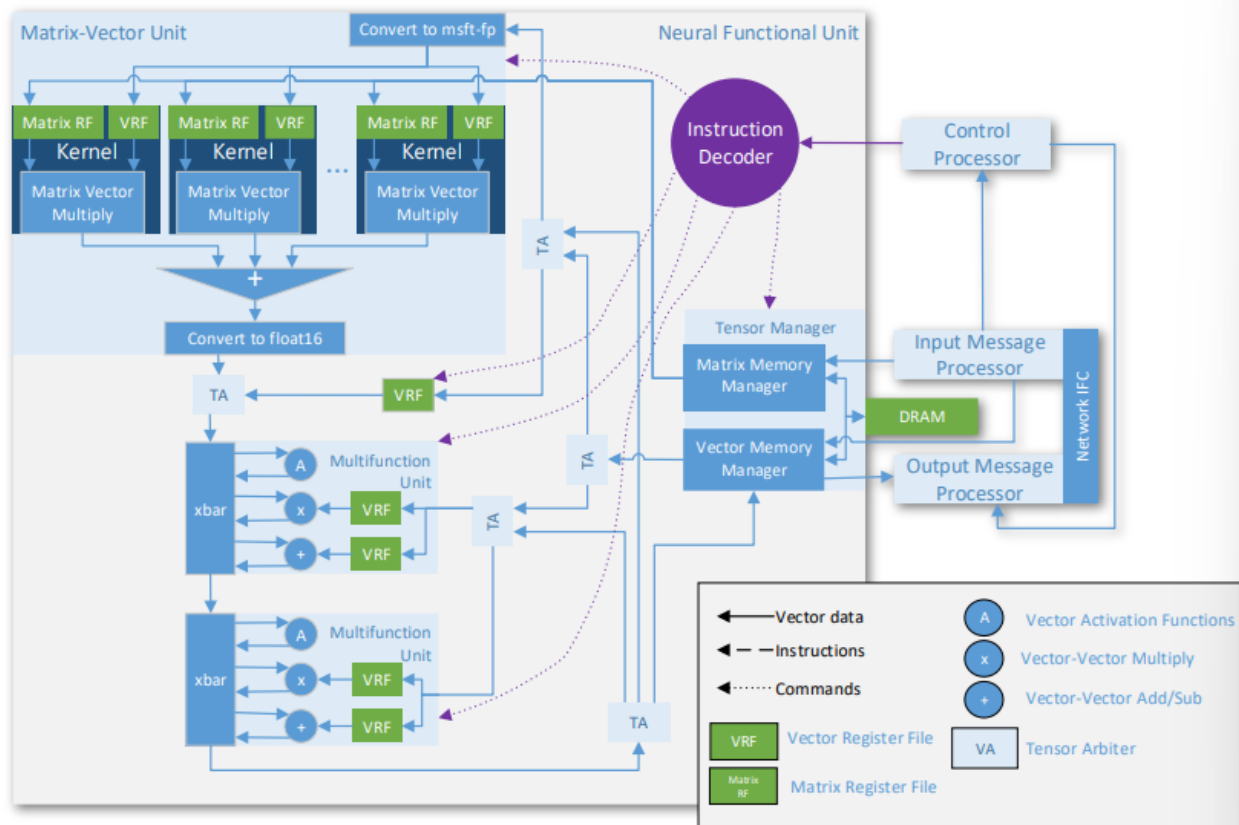
- CNNs mapped to single FPGAs
- RNNs pinned across multiple FPGAs
- Large matrices that cannot fit in a single device are divided into sub-matrices and stored



Brainwave NPU

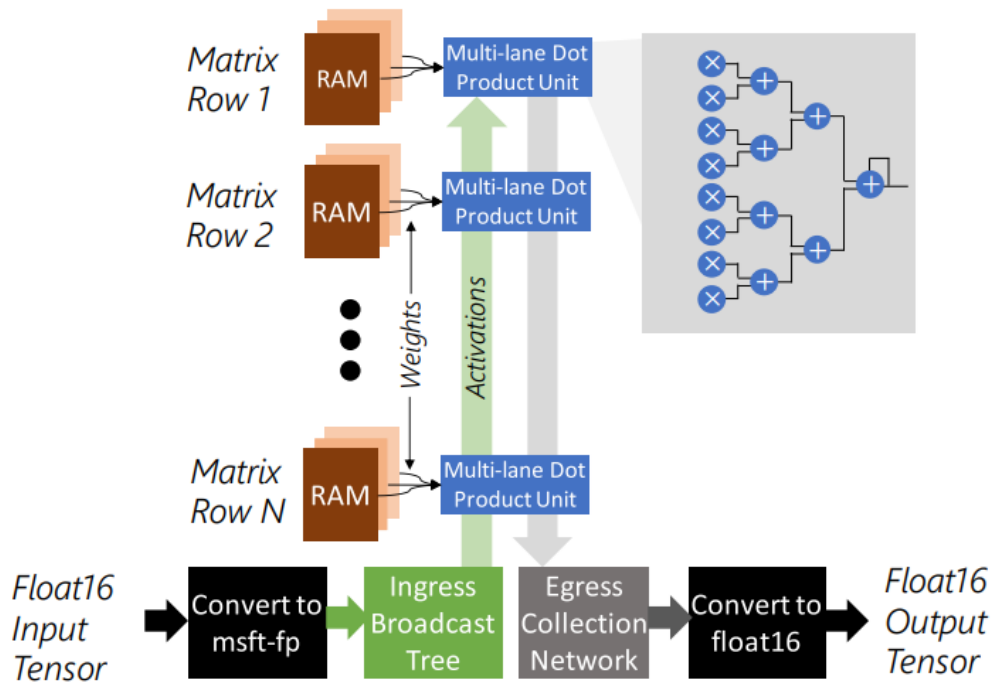
- The Brainwave soft NPU is a parameterized vector processor featuring:
 - Compile-time narrow precision data types
 - Simple single-threaded programming model adaptable to fast-changing DNN algorithms
 - A scalable microarchitecture maximizing hardware efficiency at low batch sizes

Brainwave NPU Microarchitecture



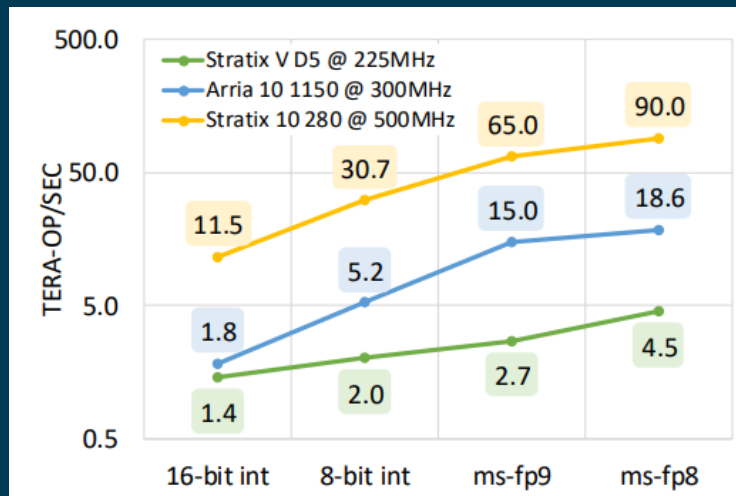
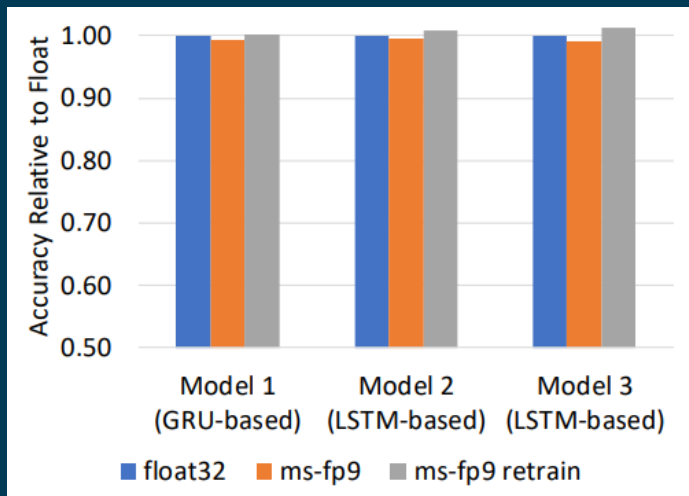
Hardware Organization

- 10s of thousands of MACs in parallel
- Independent SRAM memory port dedicated to every lane of a multi-lane vector dot product unit in the MVU



Narrow Precision

- Neural-optimized data formats based on 8 and 9 bit floating point, ms-fp8 & ms-fp9




Accelerating Bing Intelligent Search

- Bing's intelligent search uses TP1 and DeepScan DNN models to identify relevant set of passages for a query

Bing TP1			
	CPU-only	Brainwave-accelerated	Improvement
Model details	GRU 128x200 (x2) + W2Vec	LSTM 500x200 (x8) + W2Vec	Brainwave-accelerated model is > 10X larger and > 10X lower latency
End-to-end latency per Batch 1 request at 95%	9 ms	0.850 ms	
Bing DeepScan			
	CPU-only	Brainwave-accelerated	Improvement
Model details	1D CNN + W2Vec (RNNs removed)	1D CNN + W2Vec + GRU 500x500 (x4)	Brainwave-accelerated model is > 10X larger and 3X lower latency
End-to-end latency per Batch 1 request at 95%	15 ms	5 ms	

Key Learnings

- Designing a scalable system architecture is as critical as optimizing single chip performance
 - Soft NPUs connected exploit datacenter-scale pinning of models and scale elastically beyond single-chip solutions
- Narrow precision is viable, achieving high performance and efficiency
- Using configurable hardware, a system can be designed without compromising between latency and throughput



Thanks for
Listening!