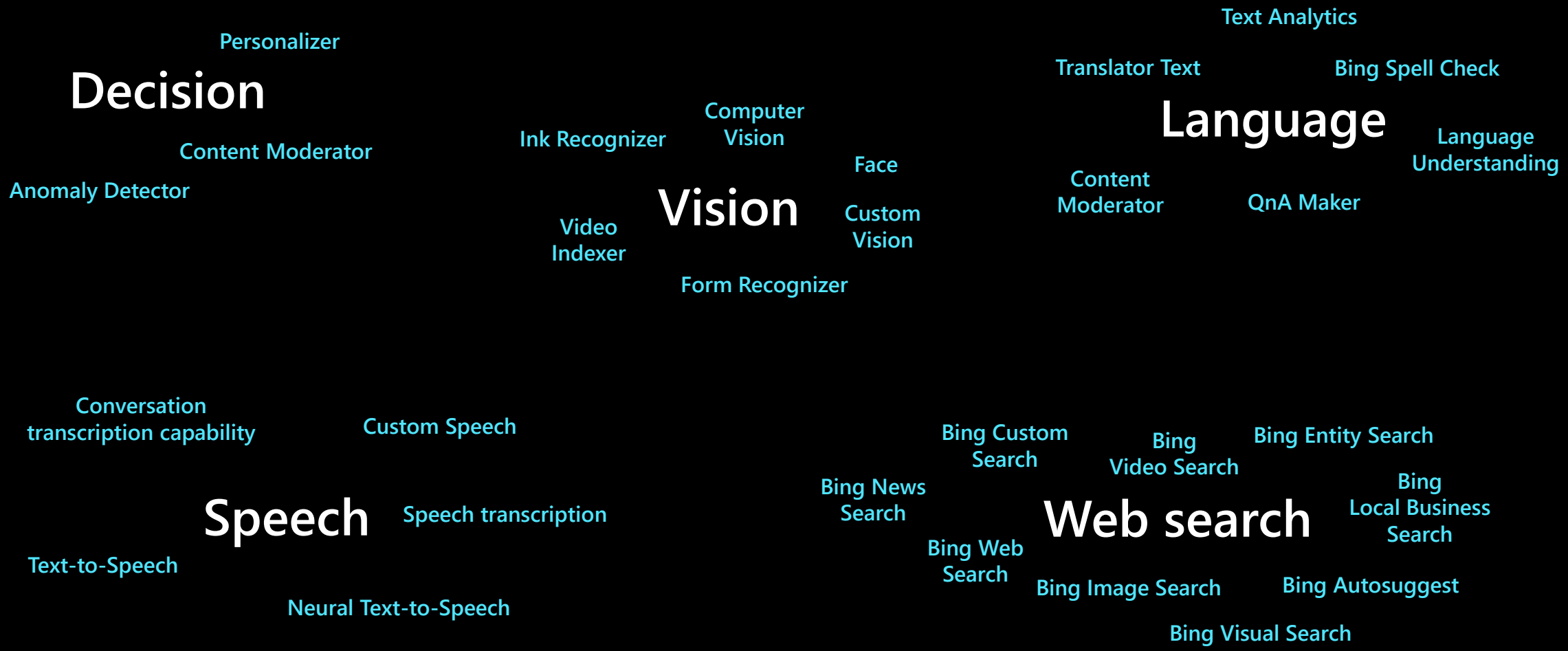
Microsoft

Accelerating Microsoft's AI Ambitions

Eric Chung, PhD
Principal Research Manager
Azure AI & Advanced Architectures

Microsoft: a diverse landscape

AI/ML ubiquitously fuels our technology



Dominant state-of-the-art models evolving rapidly

Classic ML

Deep CNNs

LSTMs/ GRUs

Transformers

2012

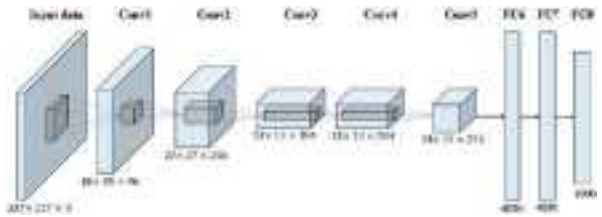
2019+

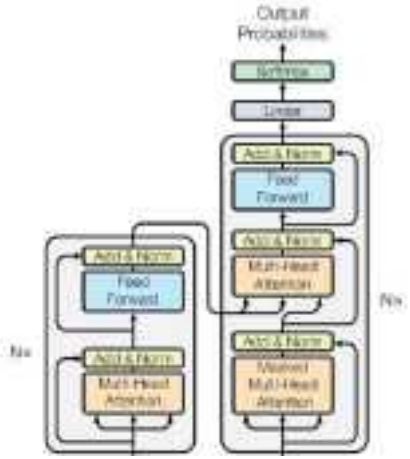
Sparse transformers?

Sparse mixture of experts?

Dynamic, recursive networks?

Graph neural networks?





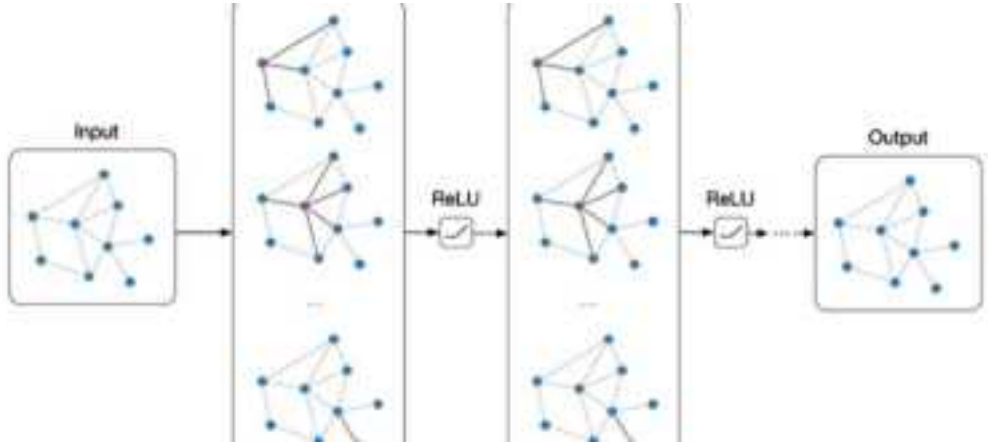
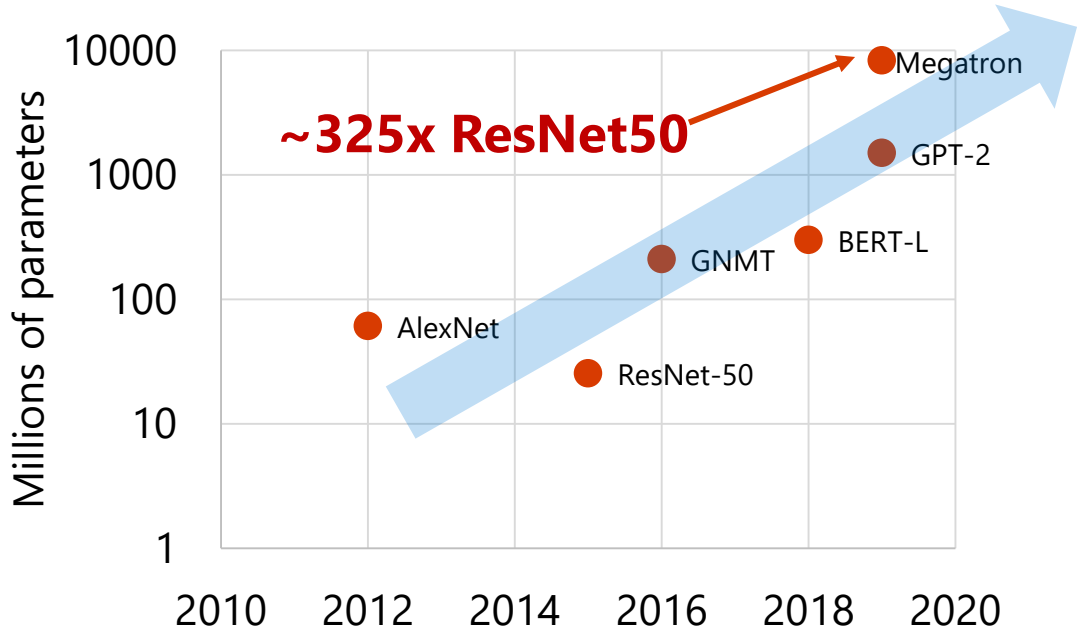
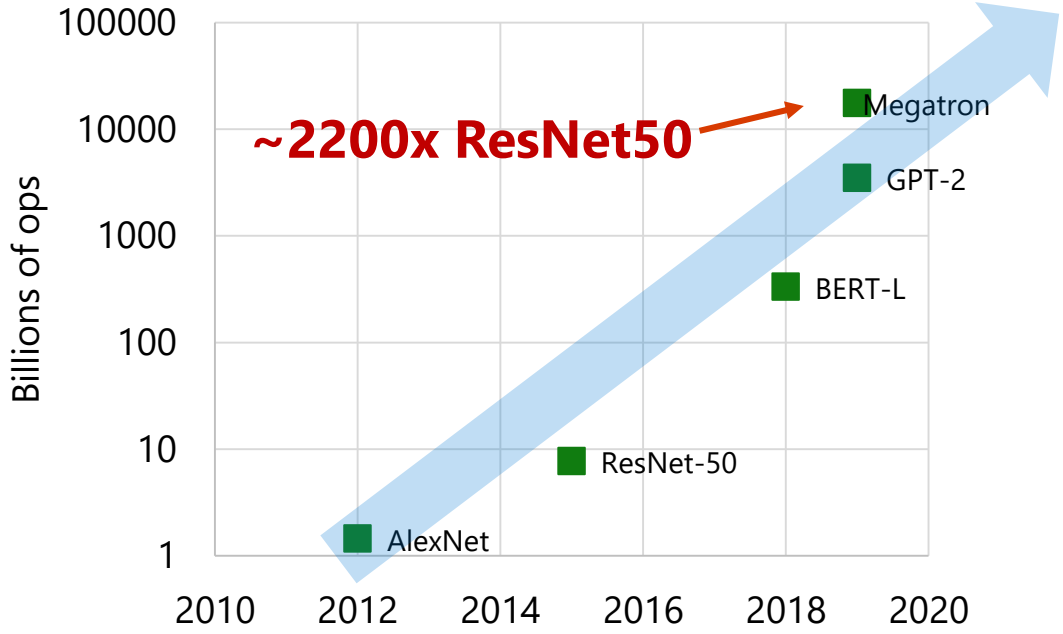


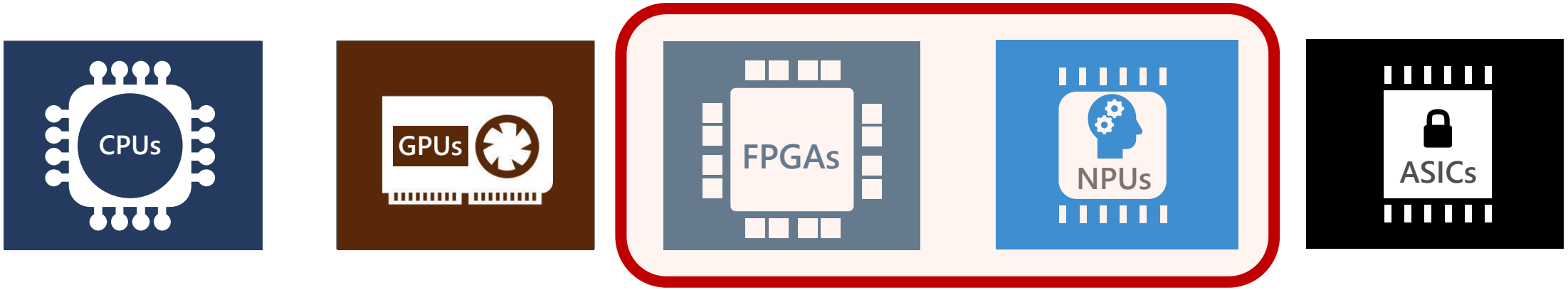
Figure sources:
1. Han et al., Pre-Trained AlexNet Architecture with Pyramid Pooling and Supervision for High Spatial Resolution Remote Sensing Image Scene Classification
2. Vaswani et al., "Attention is all you need"
3. https://tkipf.github.io/graph-convolutional-networks/

Model sizes also growing exponentially





MS deploys many AI accelerators for its high ambition AI workloads



Cloud DNN training and batched inferencing on NVIDIA GPUs (CUDA, PyTorch, TensorFlow)

Cloud and heavy edge inferencing performed on Intel CPUs (ONNX) and MS-NPUs (FPGA)

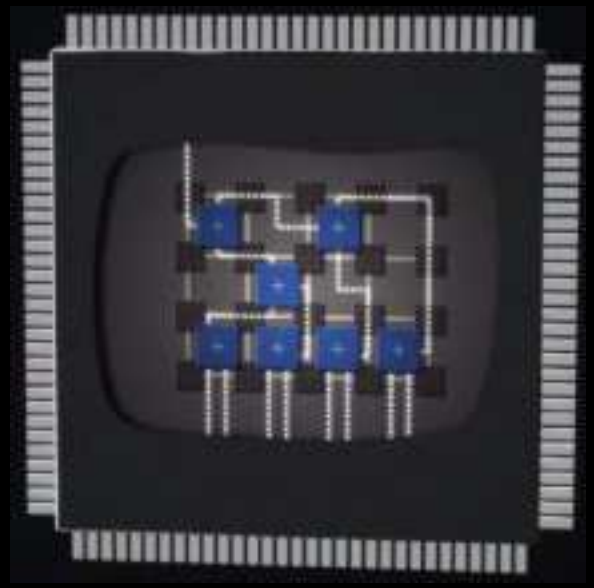
Light edge inferencing on commodity and custom silicon (e.g., Hololens, etc.)

Microsoft

Inside Bing's AI Inference Supercomputer: Project Brainwave

Project Catapult + Brainwave History

Field Programmable Gate Arrays



2011: Project Catapult Launched

2013: Bing pilot runs decision trees 40X faster

2015: Bing ranking throughput increased 2X

2016: Azure Accelerated Networking delivers industry-leading cloud performance

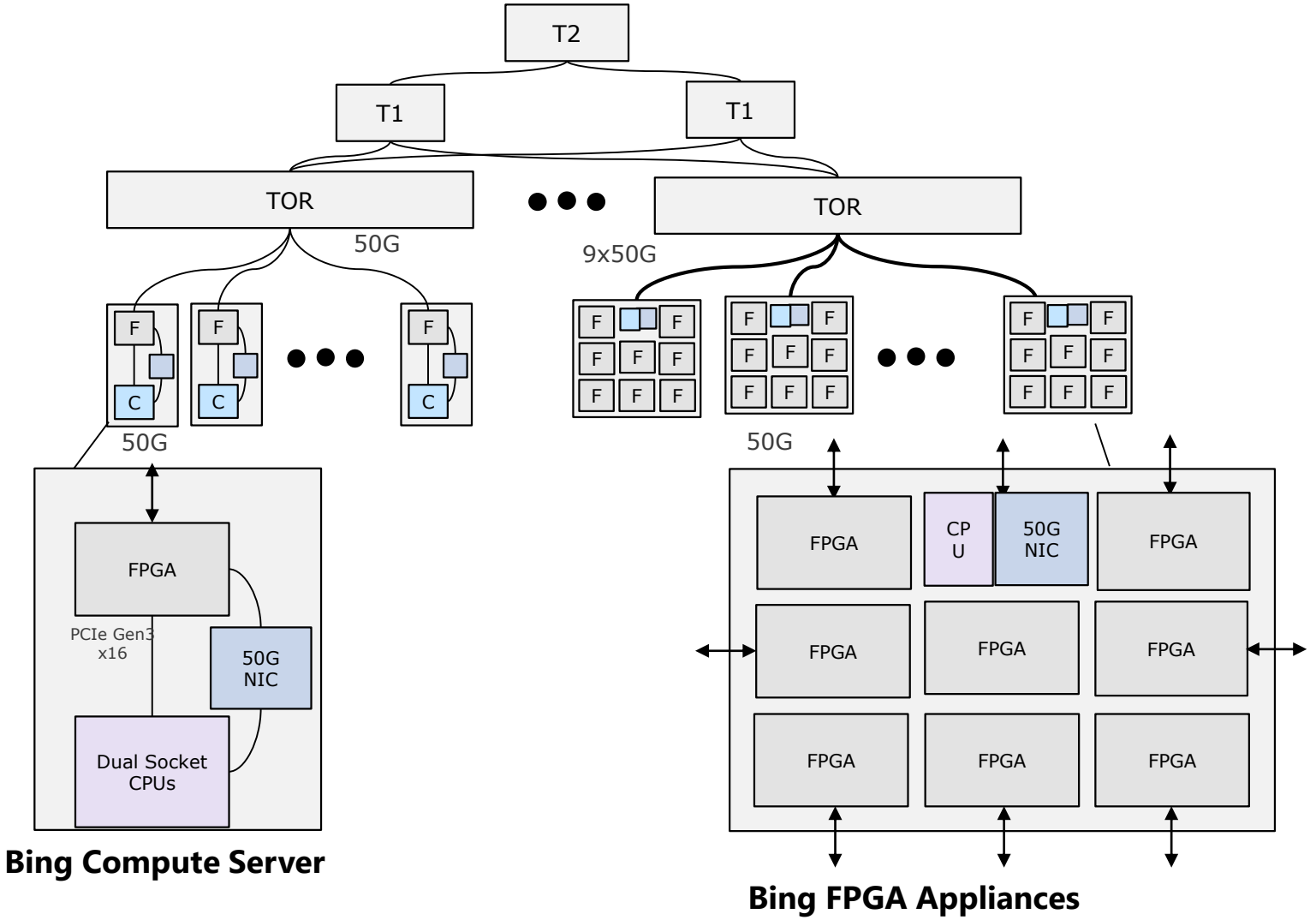
2017: Over 1M servers deployed with FPGAs at hyperscale

2017: Hardware Microservices harness FPGAs for distributed computing

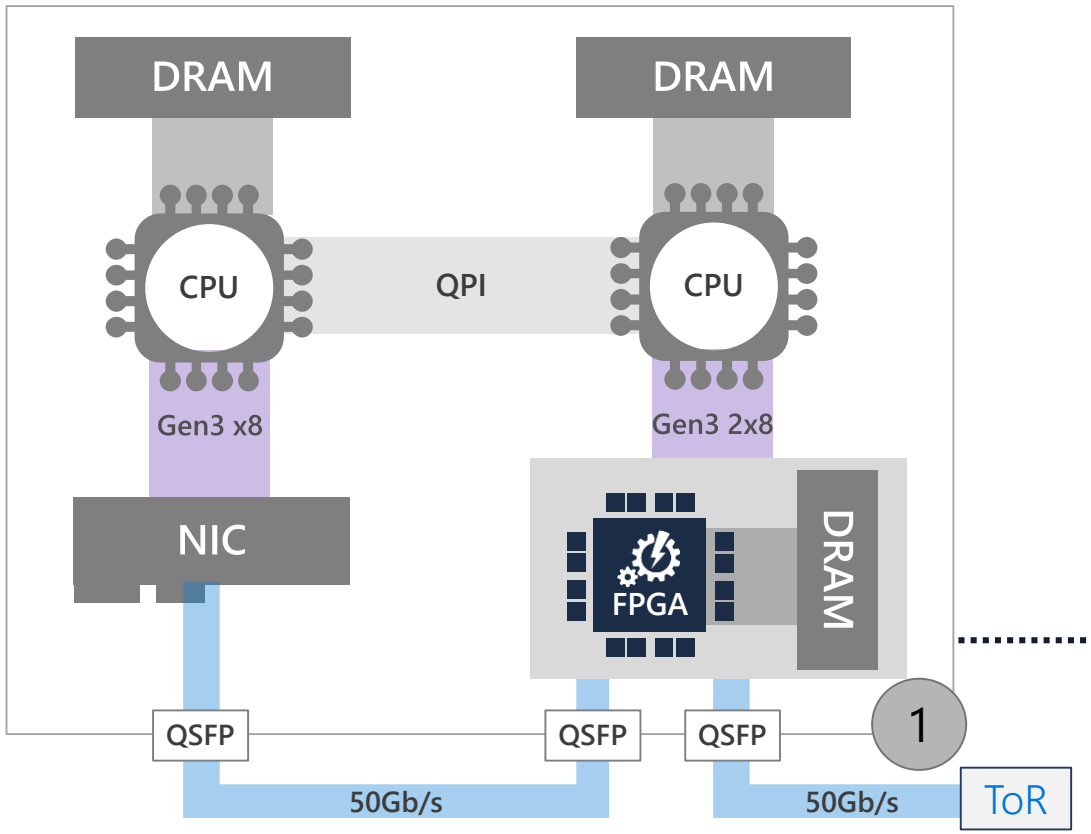
2017: FPGAs enable real-time AI, ultra-low latency inferencing without batching; Bing launches first FPGA-accelerated Deep Neural Network

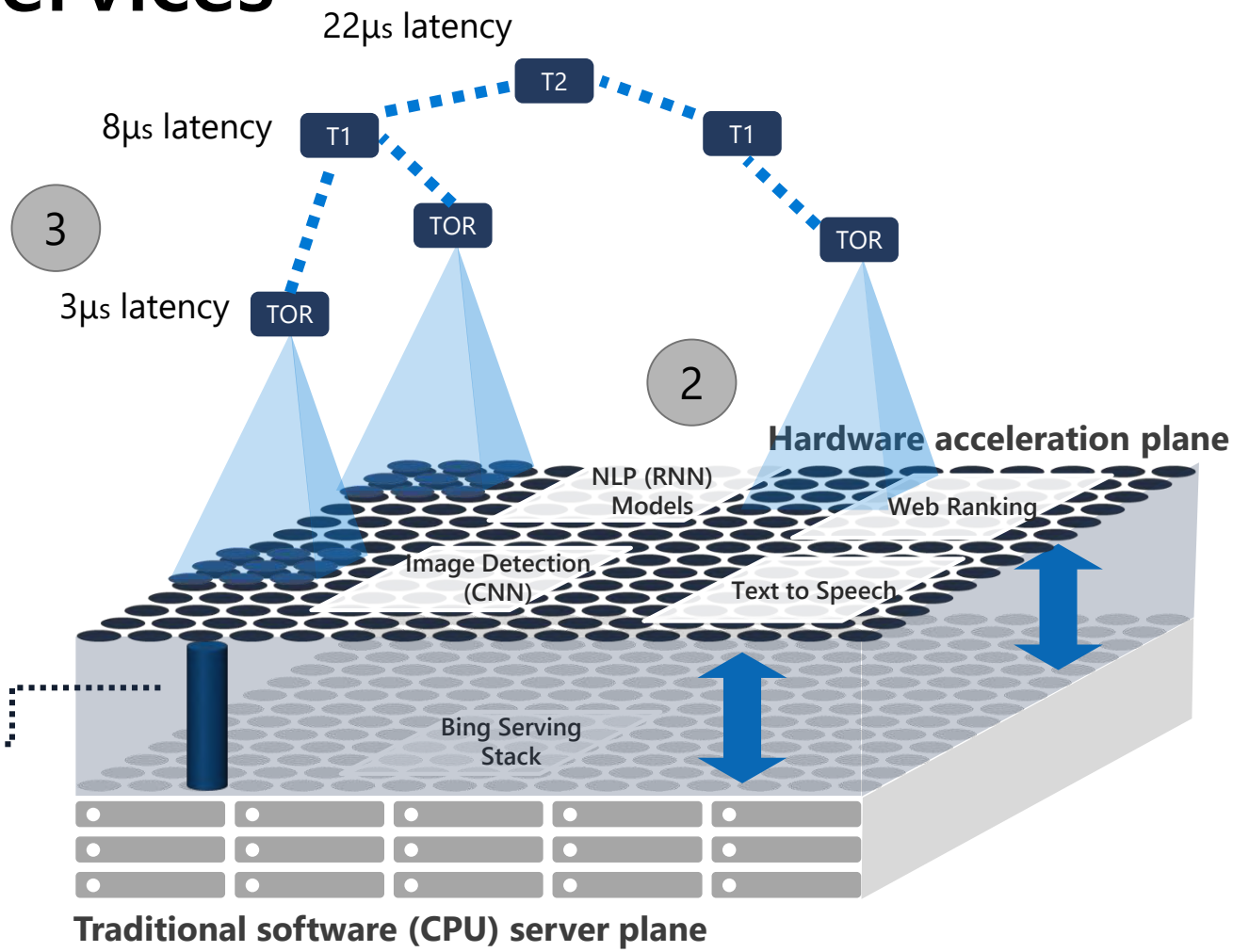
2018: Project Brainwave launched in Azure Machine Learning

Bing's 500 Petaflops Inference Supercomputer



Scalable Hardware Microservices





1 FPGAs are network connected. Used and managed independently from the CPU.

2 Interconnected FPGAs form a separate plane of computation built on Hardware as a Service (HaaS).

3 Direct FPGA to FPGA communication using Lightweight Transport Layer (LTL) at ultra low latencies.

Rapid Iteration and Deployment

Brainwave v1	Brainwave v2	Brainwave v3	Brainwave v4
Low latency LSTM inference	Narrow Precision Breakthrough	Convolution Optimizations	Generalized ISA, Transformers
2016	2017	2018	2019

Catapult FPGAs + MS-custom algorithms

2016
MS-NPU
(Research)

AlexNet
Batch-1
1.4 Gops/inf
~0.25T effective

2018
MS-NPU
(Production)

ResNet-50
Batch-1
7.7 Gops/inf
5.2T effective
<1.5ms

2019
MS-NPU
(Production)

DeepBench GRU
Batch-1
~47M parameters
70.5 Gops/inf
39T+ effective
<2ms

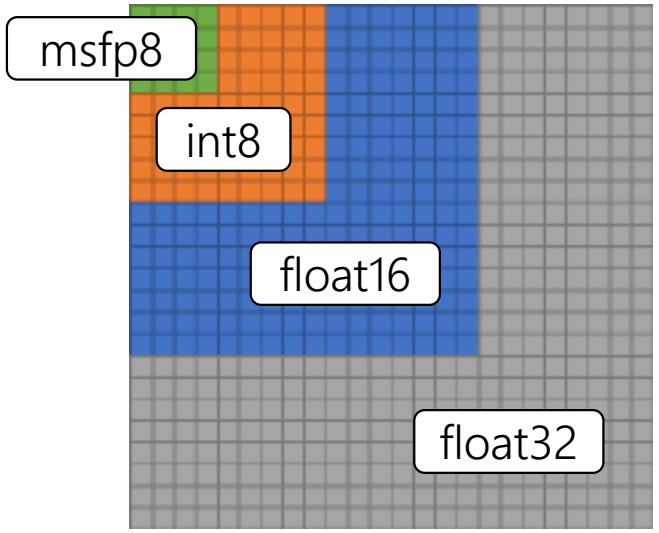
Intel Arria
FPGA (2014)

<70W
TSMC 20nm
Rated for 1 FP32 TFLOPS

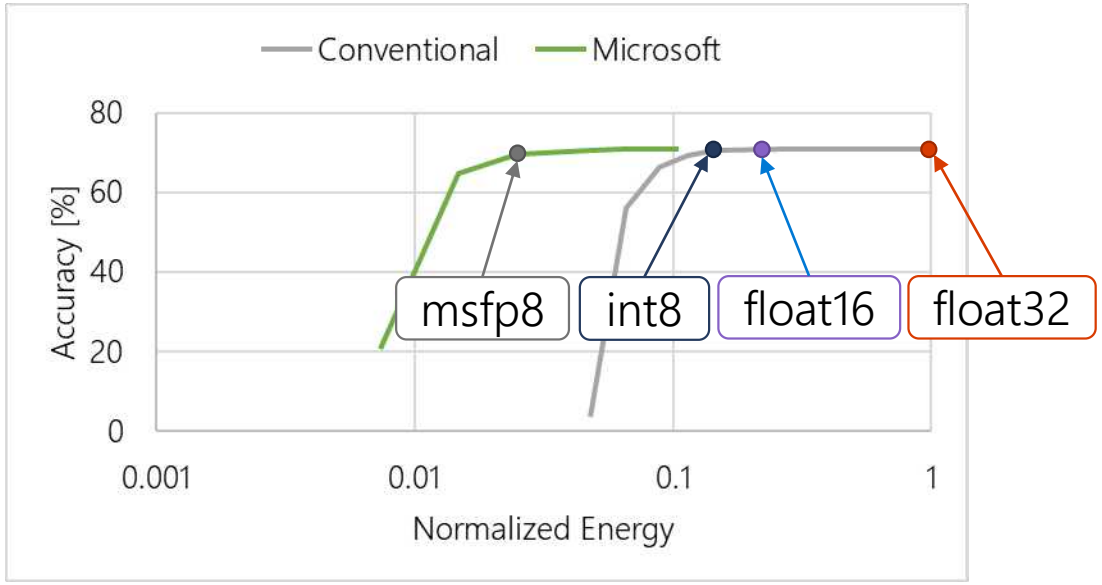
Intel Stratix
10 FPGA
(2017+)

<150W
Intel 14nm
Rated for 5-10 FP32 TFLOPS

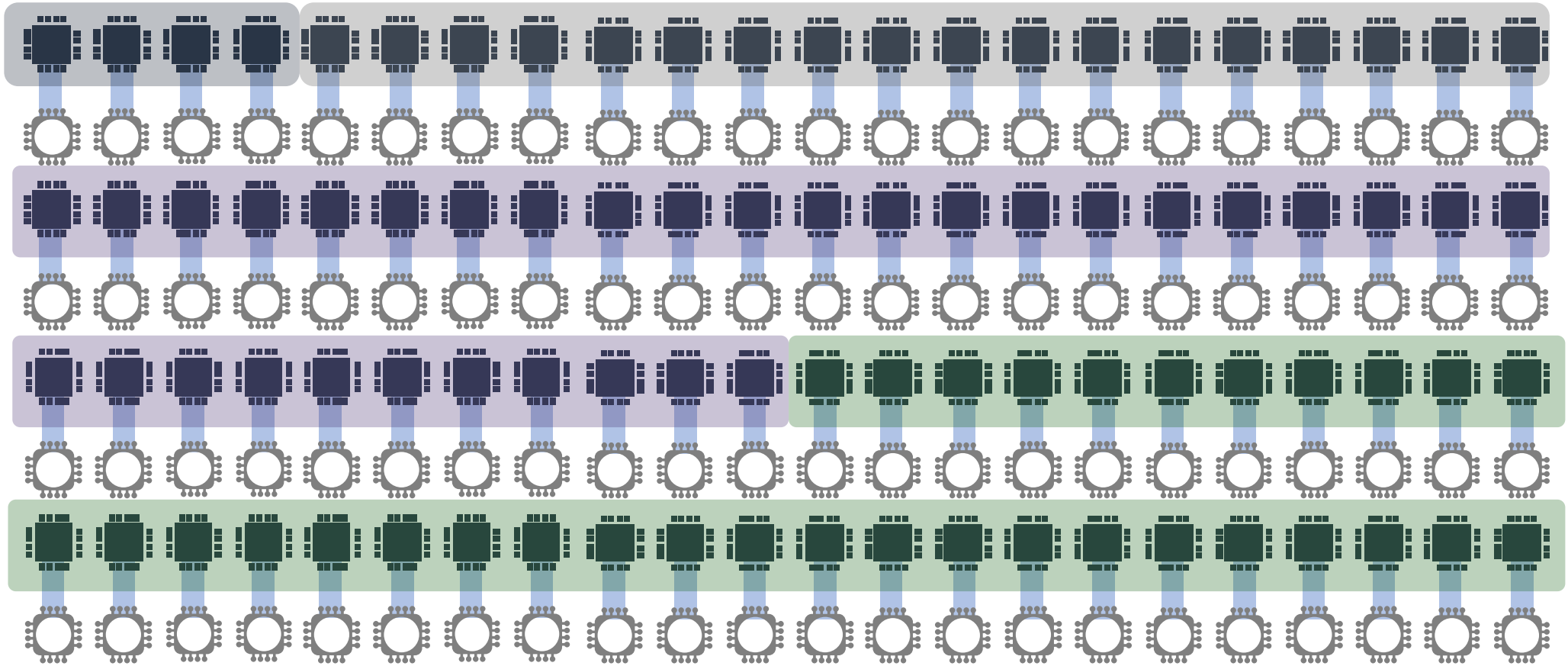
MS-custom datatypes for compute and storage



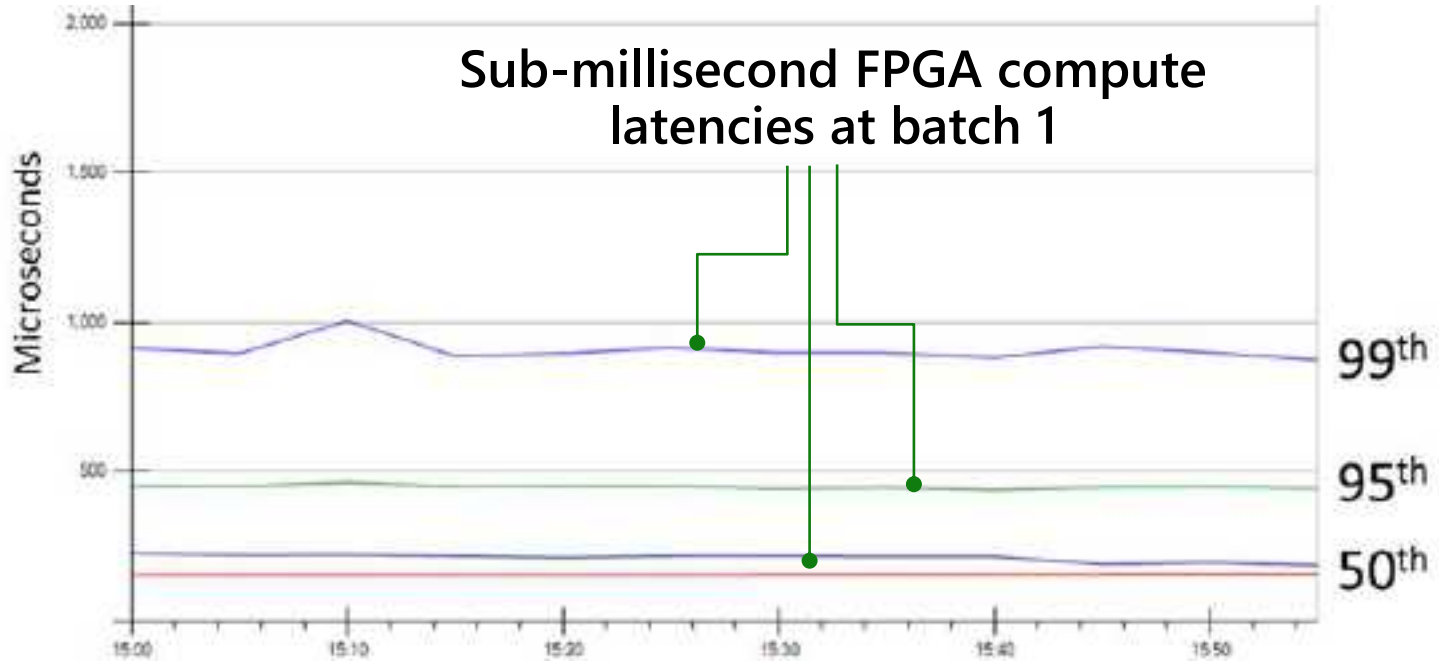
Multiplier Area & Energy



Weight pinning in distributed FPGAs in HW microservices drives high utilization



Brainwave System In Production since 2017

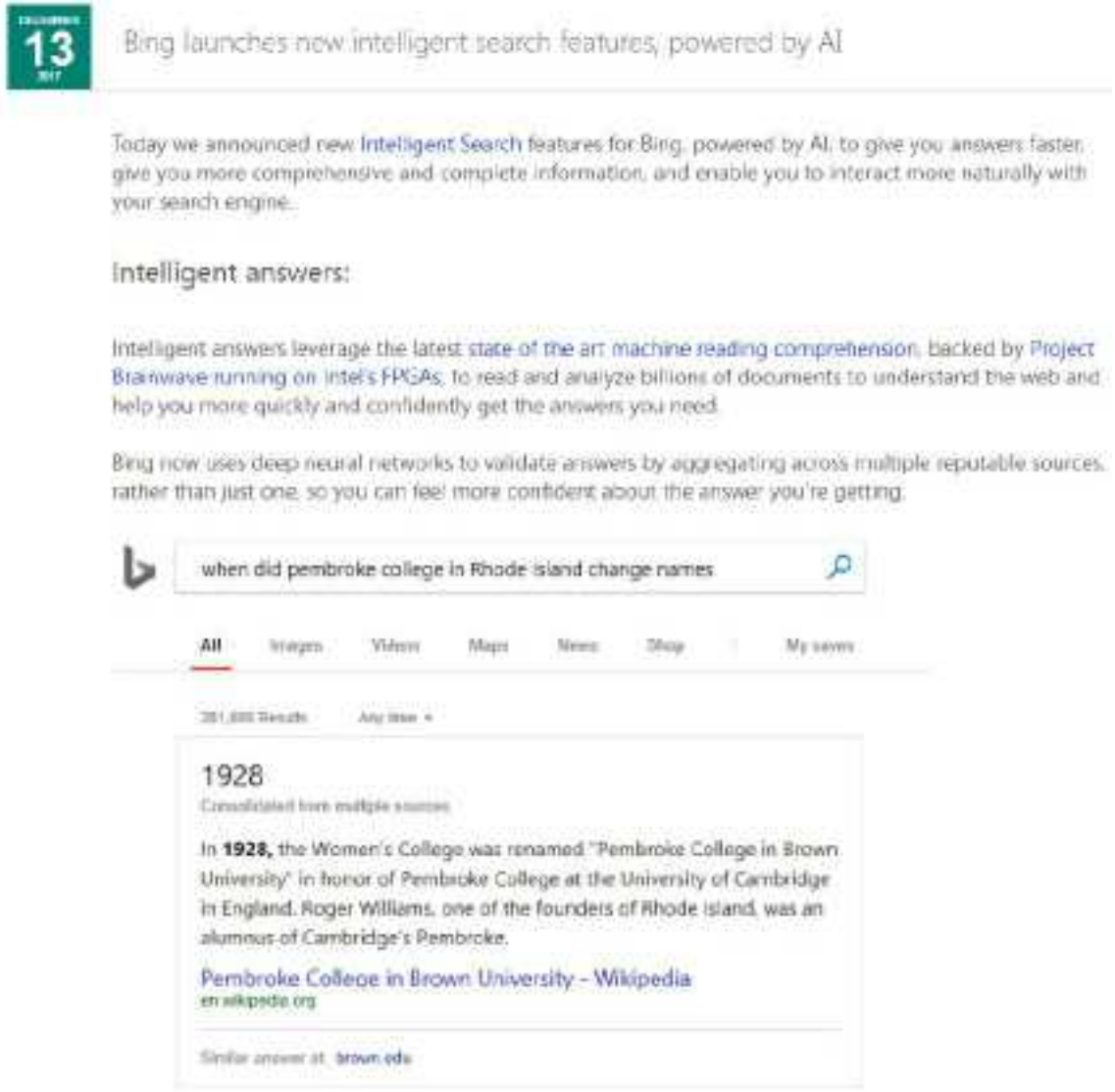


Deployment of LSTM-based NLP model (tens of millions of parameters)

Takes tens of milliseconds to serve on well-tuned CPU implementations

Tail latencies in Brainwave-powered DNN models negligible in E2E software pipelines

Bing Intelligent Search Backed By Project Brainwave



13

Bing launches new intelligent search features, powered by AI

Today we announced new Intelligent Search features for Bing, powered by AI, to give you answers faster, give you more comprehensive and complete information, and enable you to interact more naturally with your search engine.

Intelligent answers:

Intelligent answers leverage the latest state of the art machine reading comprehension, backed by Project Brainwave running on Intel's FPGAs, to read and analyze billions of documents to understand the web and help you more quickly and confidently get the answers you need.

Bing now uses deep neural networks to validate answers by aggregating across multiple reputable sources, rather than just one, so you can feel more confident about the answer you're getting.

when did pembroke college in Rhode Island change names

All Images Videos Maps News Shop My saves

1928

In **1928,** the Women's College was renamed "Pembroke College in Brown University" in honor of Pembroke College at the University of Cambridge in England. Roger Williams, one of the founders of Rhode Island, was an alumnus of Cambridge's Pembroke.

Pembroke College in Brown University - Wikipedia

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	

https://www.microsoft.com/en-us/research/uploads/prod/2018/03/mi0218_Chung-2018Mar25.pdf

https://blogs.bing.com/search/2017-12/search-2017-12-december-ai-update

Microsoft

Hardware for Future AI

Success Drivers for Future AI Hardware

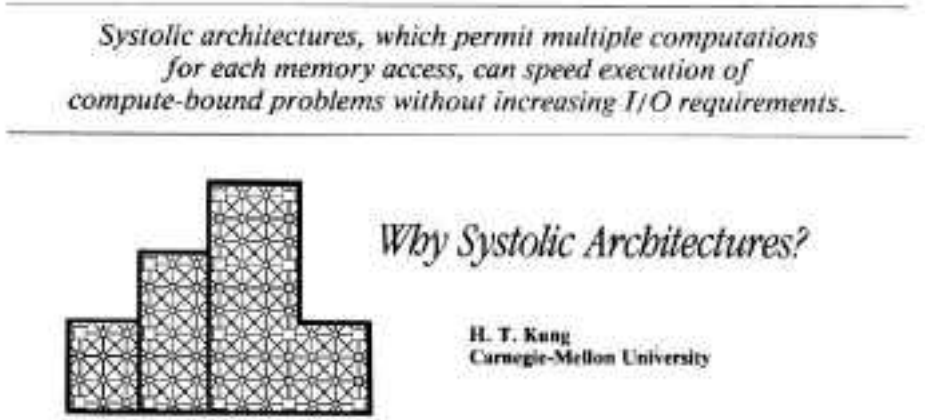
- Must solve real customer problems – solutions including non-AI pieces, not just AI components
- Must be differentiated E2E including system overheads
- Want durable and "horizontally-capable" architectures with long shelf lives (3-5 years)
- Compatible and friendly to deploy in diverse environments (SKUs, datacenters, etc)
- Must be easy to develop software/models for and integrate seamlessly with AI tools ecosystem
- $ Improved cost of ownership at system-scale vs general-purpose commodity hardware

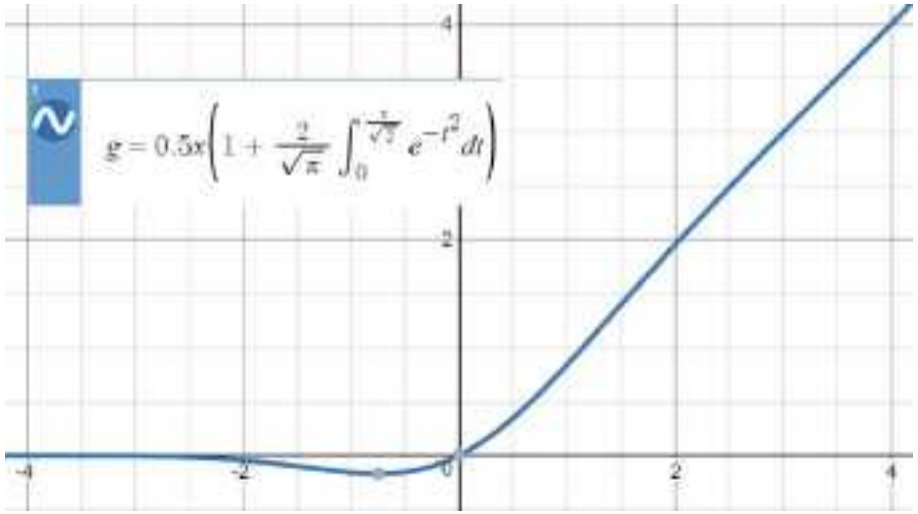
AI/ML != matrix multiply

Technology for accelerating dense matrix arithmetic now well understood

Fast-evolving non-matrix computations can become bottleneck

- Batch normalization
- Loss normalization
- Softmax
- RELU, Leaky RELU, GELU
- Beam search
- K-nearest neighbor
- Top-k sorting
- …

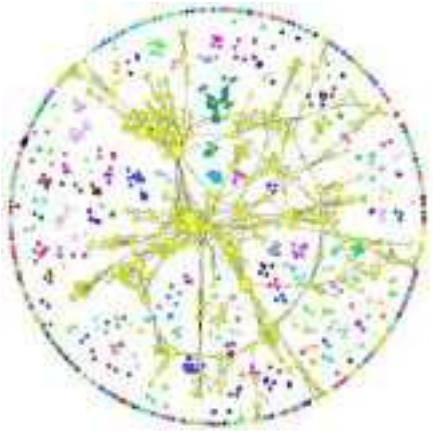




1. H.T. Kung, “Why Systolic Arrays?”, 1982
2. https://datascience.stackexchange.com/questions/49522/what-is-gelu-activation

Emerging Trends

AI innovation demands expressiveness, flexibility, and efficiency



Language-Level Expressiveness

Dynamic networks, control flow, and recursion

Efficiency by Exploiting Sparsity

Increasing large and sparse networks

Closing thoughts and predictions

Current industry tech has “solved” AI 1.0 – i.e., dense float matrix arithmetic

But AI is still fast-evolving, no signs of slowing down

HW+SW+Algorithm innovation needed to fuel differentiation for AI 2.0 beyond (e.g., novel data types, sparsity, SGD/backprop replacement, etc.)

Today’s AI machines still orders-of-magnitude less efficient than biology ➔ plenty runway left …

The upcoming wave(s) of algorithmic disruption coupled with hardware innovation will decide our AI destiny

Microsoft

Q/A & Discussion

erchung@microsoft.com