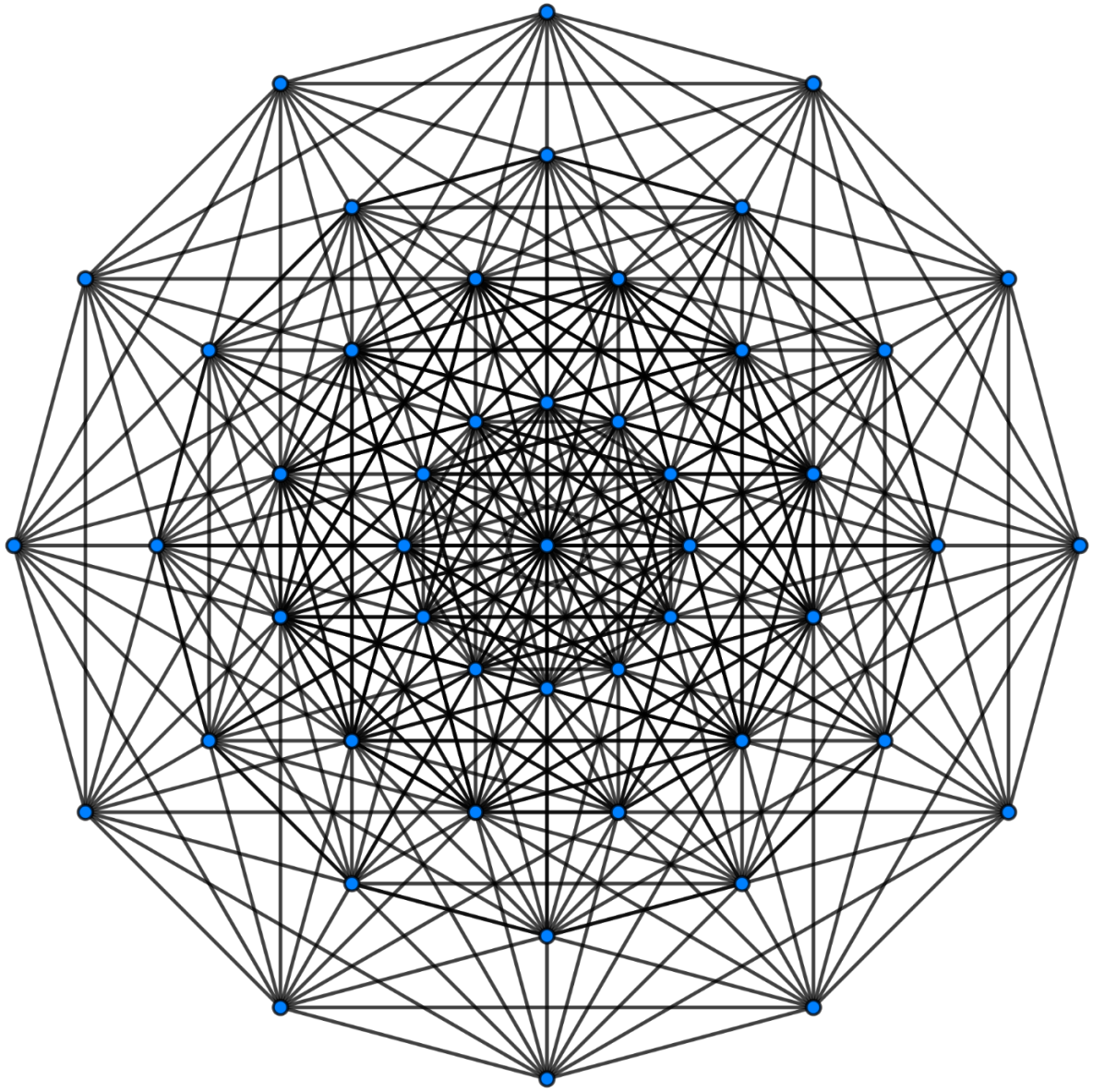
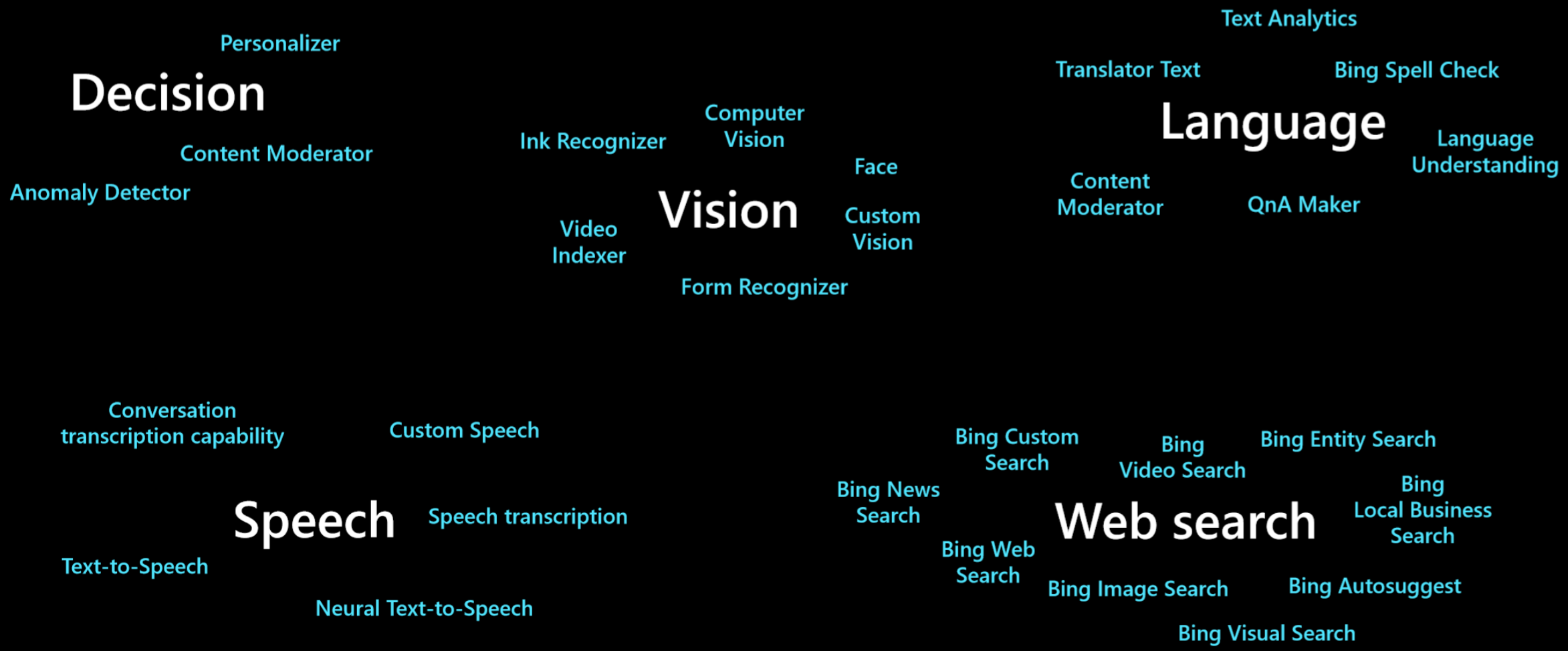
Microsoft Azure

Configurable Cloud-Scale DNN Processor for Real-Time AI

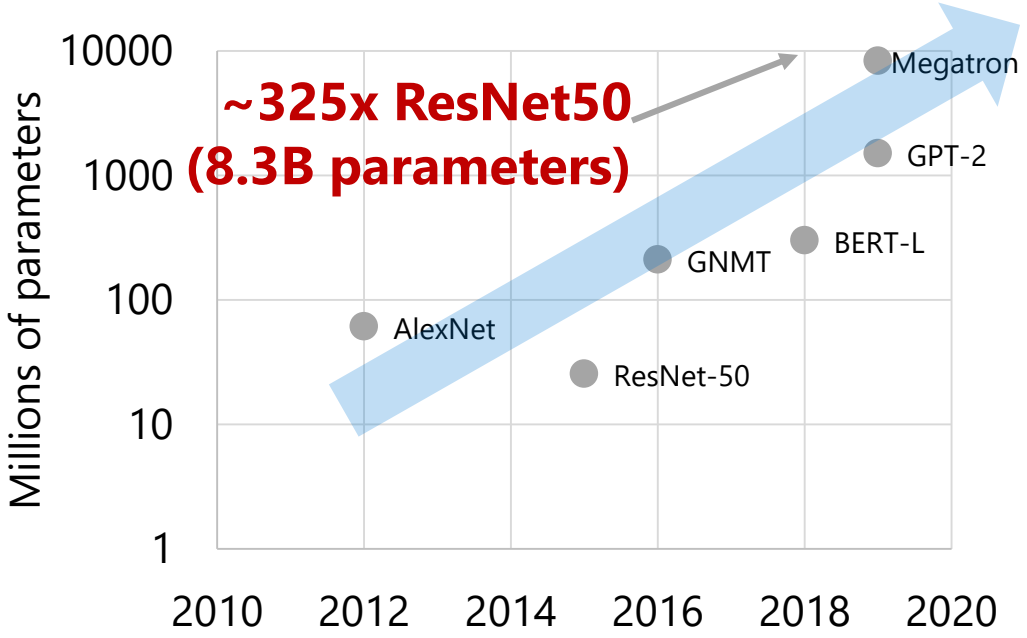


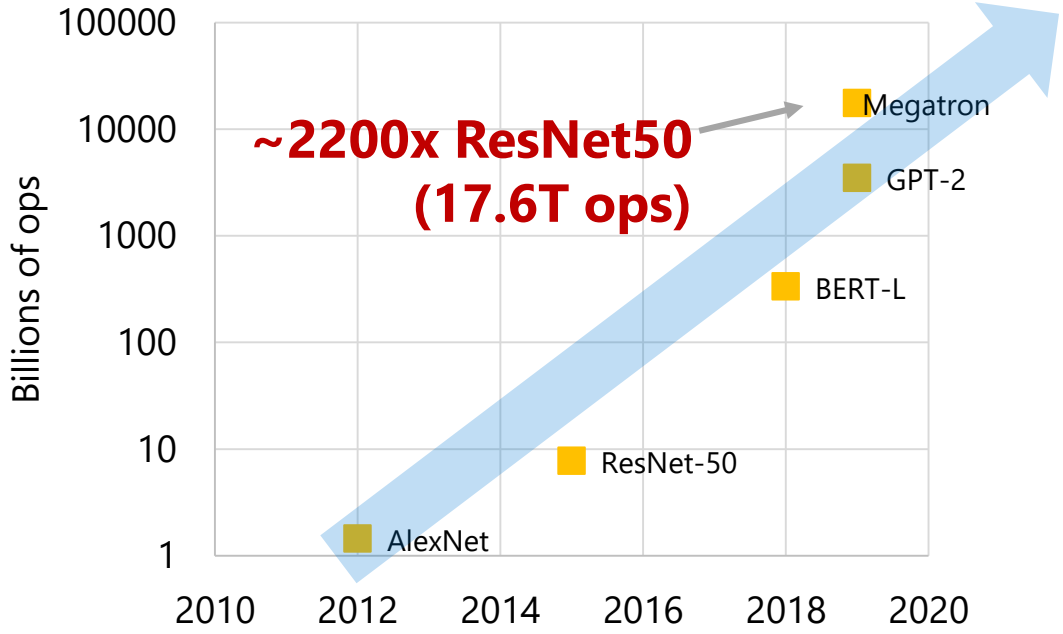
Speaker: Bita Rouhani, Sr. Researcher

AI/ML ubiquitously fuels our technology



Model sizes growing exponentially





Dominant state-of-the-art models also evolving rapidly

Classic ML

Deep CNNs

LSTMs/ GRUs

Transformers

2012

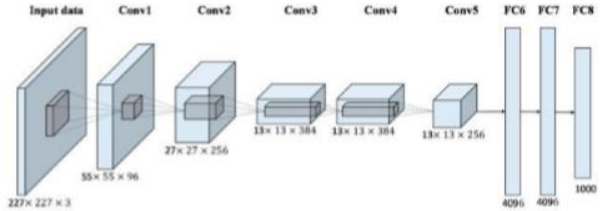
2019+

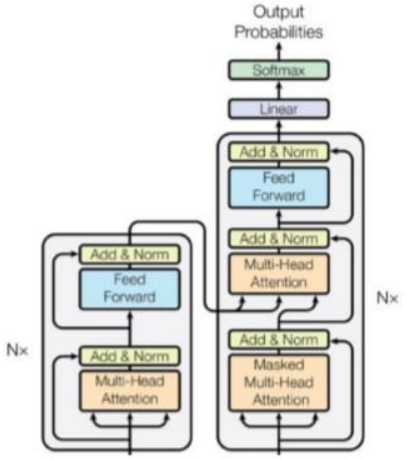
Sparse transformers?

Sparse mixture of experts?

Dynamic, recursive networks?

Graph neural networks?





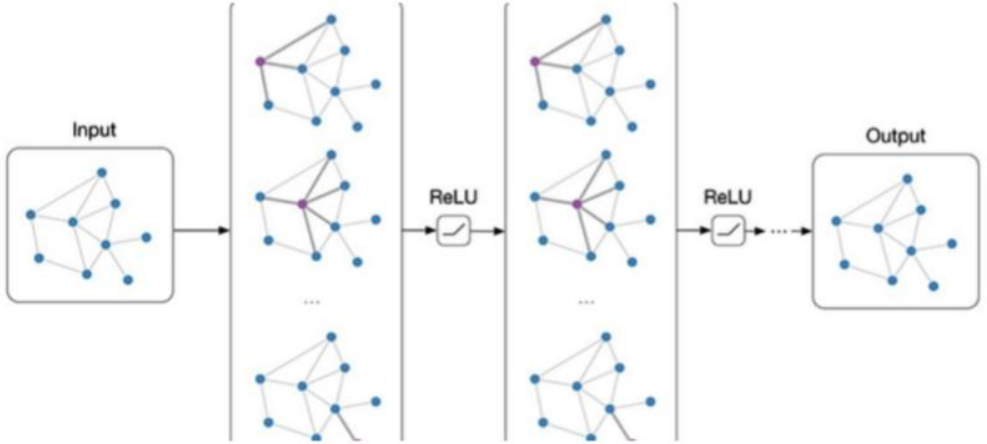


Figure sources:
1. Han et al., Pre-Trained AlexNet Architecture with Pyramid Pooling and Supervision for High Spatial Resolution Remote Sensing Image Scene Classification
2. Vaswani et al., "Attention is all you need"
3. https://tkipf.github.io/graph-convolutional-networks/

Silicon alternatives for AI models

The power of AI on FPGA

Flexibility

- FPGAs ideal for adapting to rapidly evolving AI/DL
- Adaptive numerical precision and custom operators
- CNNs, LSTMs, MLPs, transformers, reinforcement learning, feature extraction, etc.
- Exploit sparsity, etc.

Performance

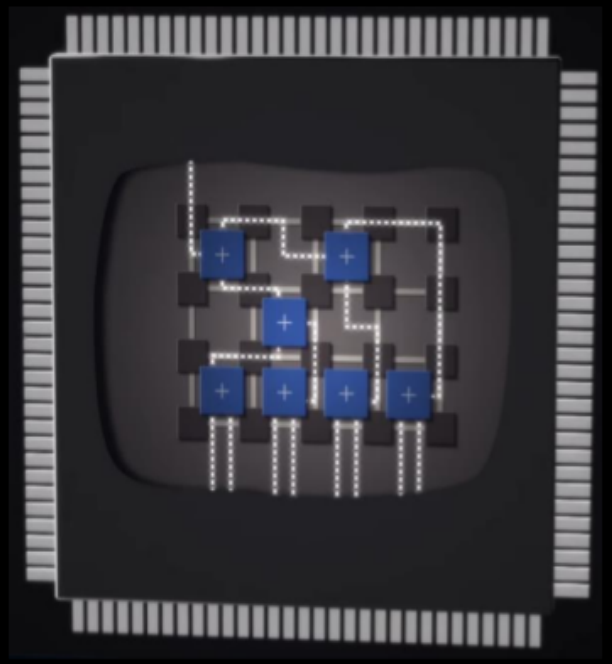
- Excellent inference performance at low batch sizes
- Ultra-low latency serving on modern DNNs
- Scale to many FPGAs in single DNN service

Scale

- Microsoft has the world's largest cloud investment in FPGAs
- Multiple Exa-Ops of aggregate AI capacity
- BrainWave runs on Microsoft's scale infrastructure

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

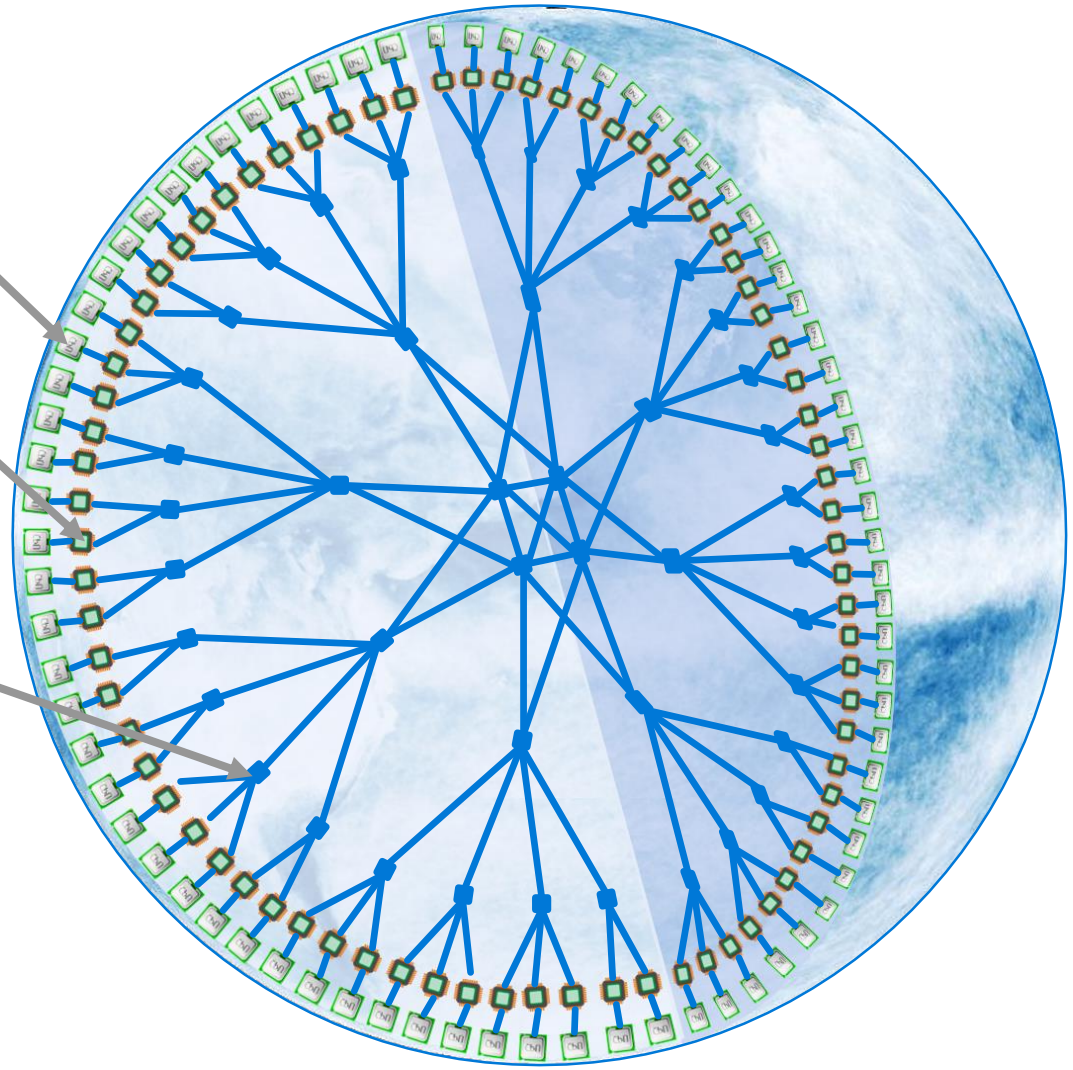
2018: Project Brainwave launched in Azure Machine Learning

Brainwave runs on a configurable cloud at massive scale

CPU compute layer

Reconfigurable compute layer (FPGA)

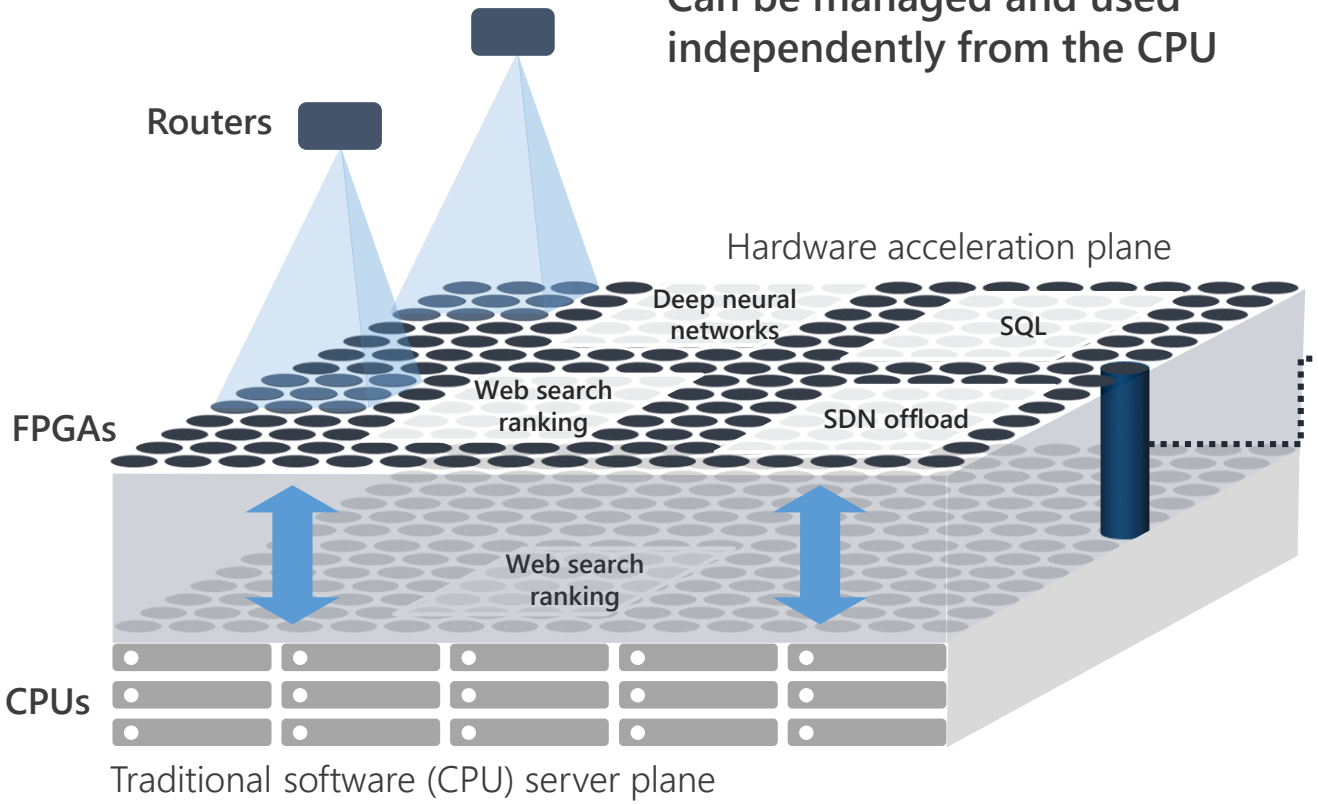
Converged network

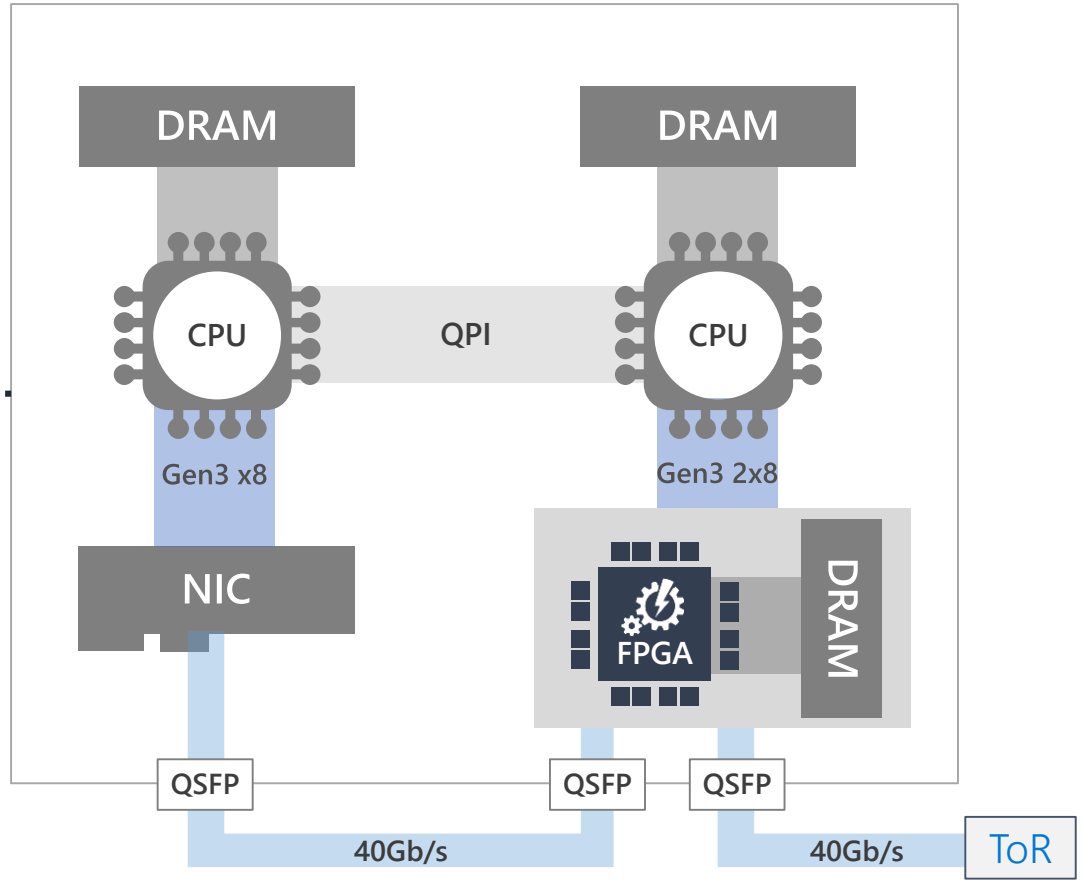


Scalable hardware microservice

Interconnected FPGAs form a separate plane of computation

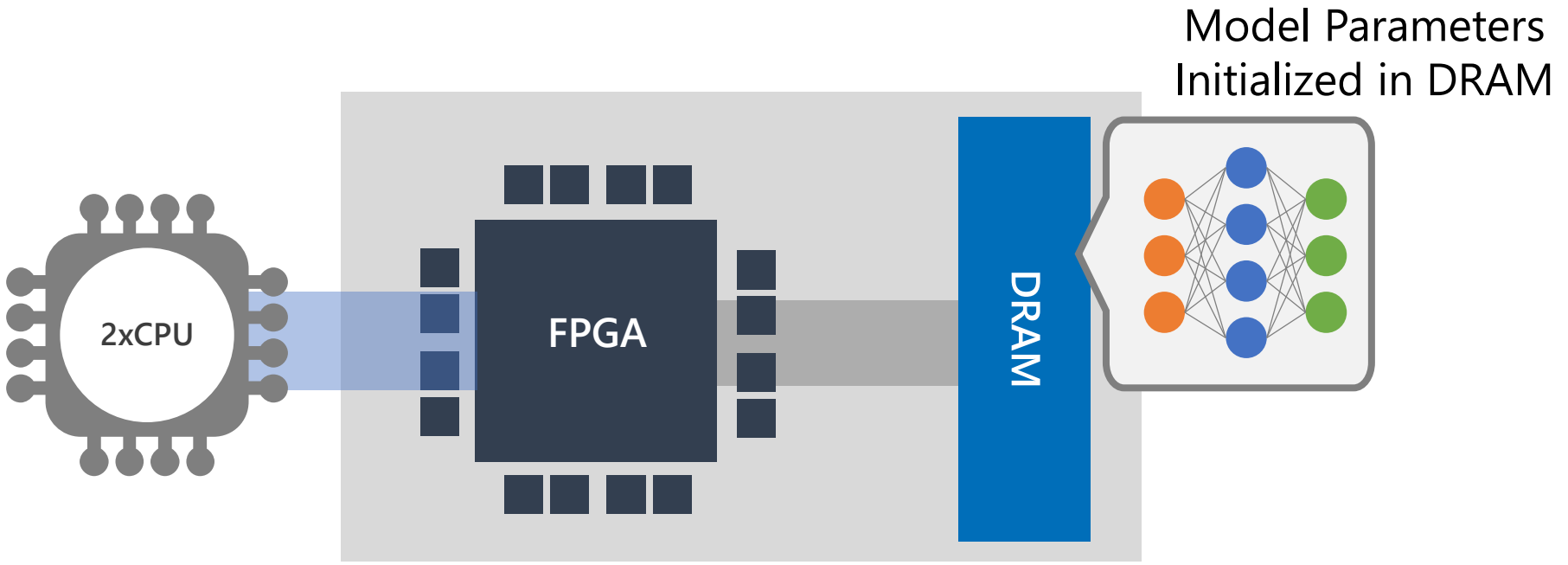
Can be managed and used independently from the CPU





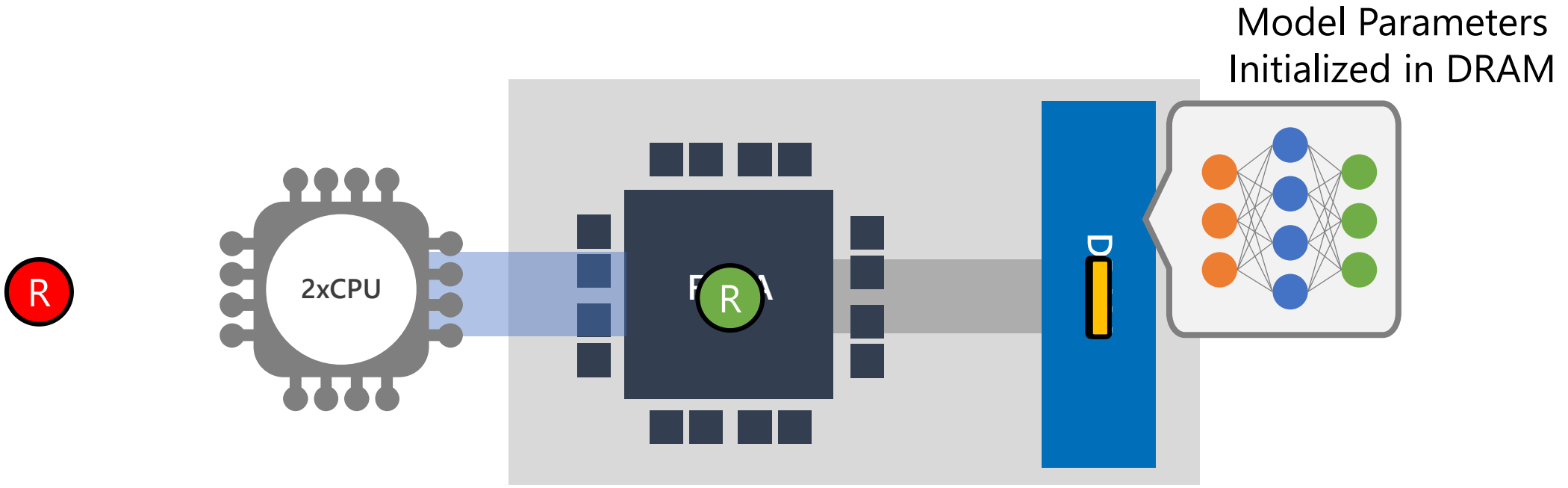
Conventional acceleration approach

Local offload and streaming



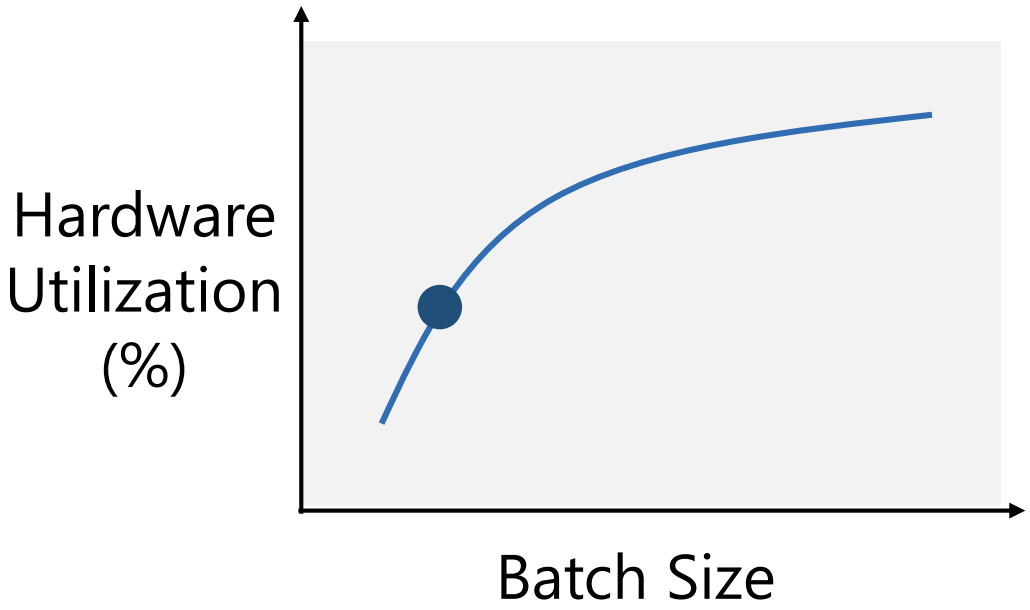
Conventional acceleration approach

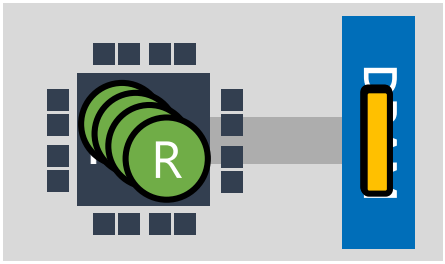
Local offload and streaming



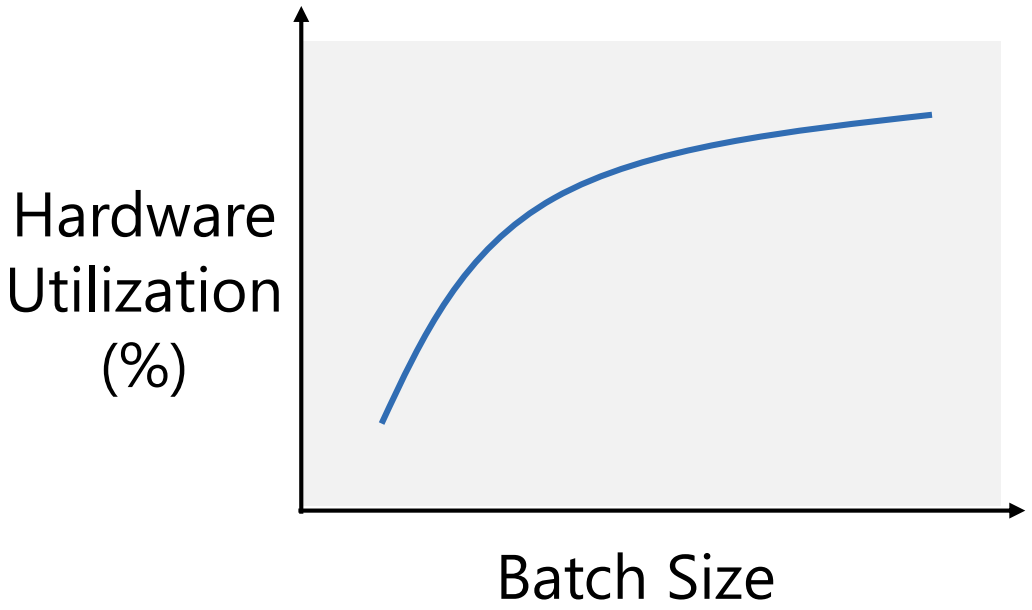
For memory-intensive DNNs with low compute-to-data ratios (e.g., LSTM), HW utilization limited by off-chip DRAM bandwidth

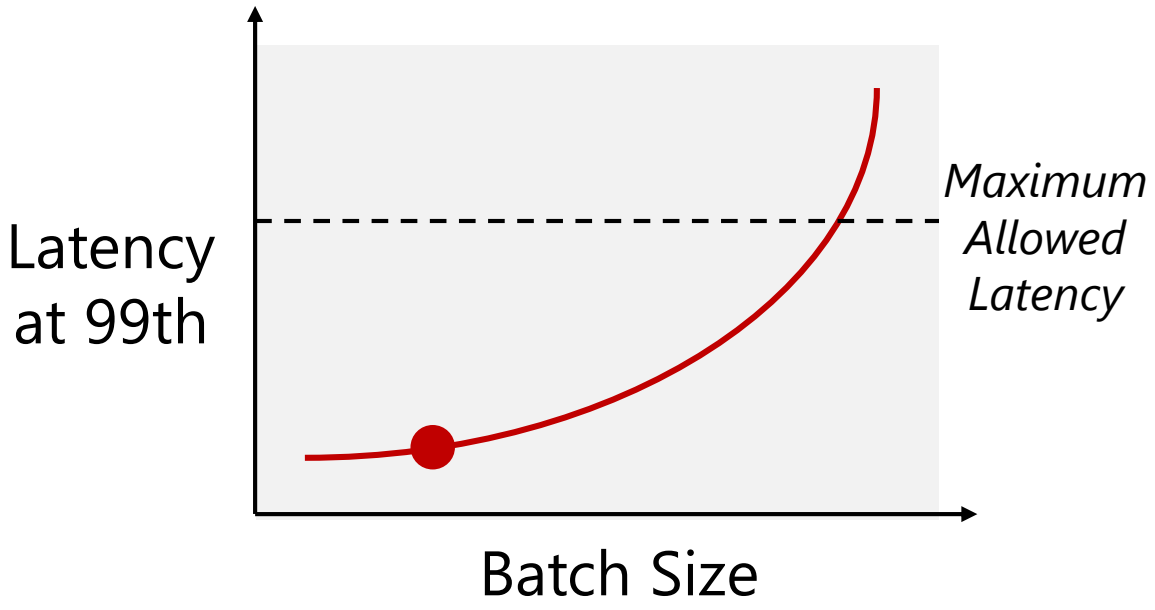
Improving HW utilization with batching





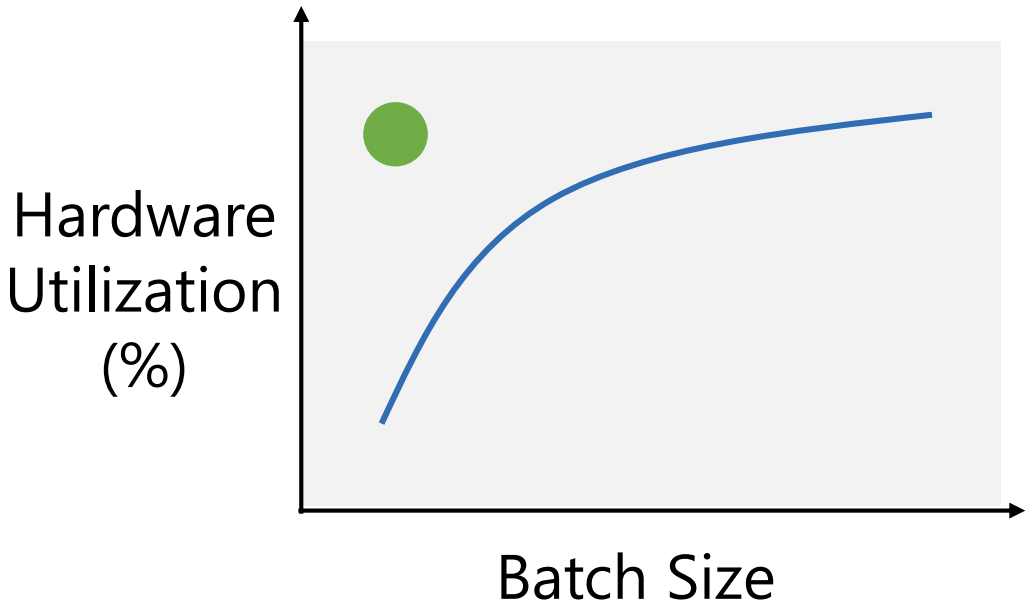
Improving HW utilization with batching

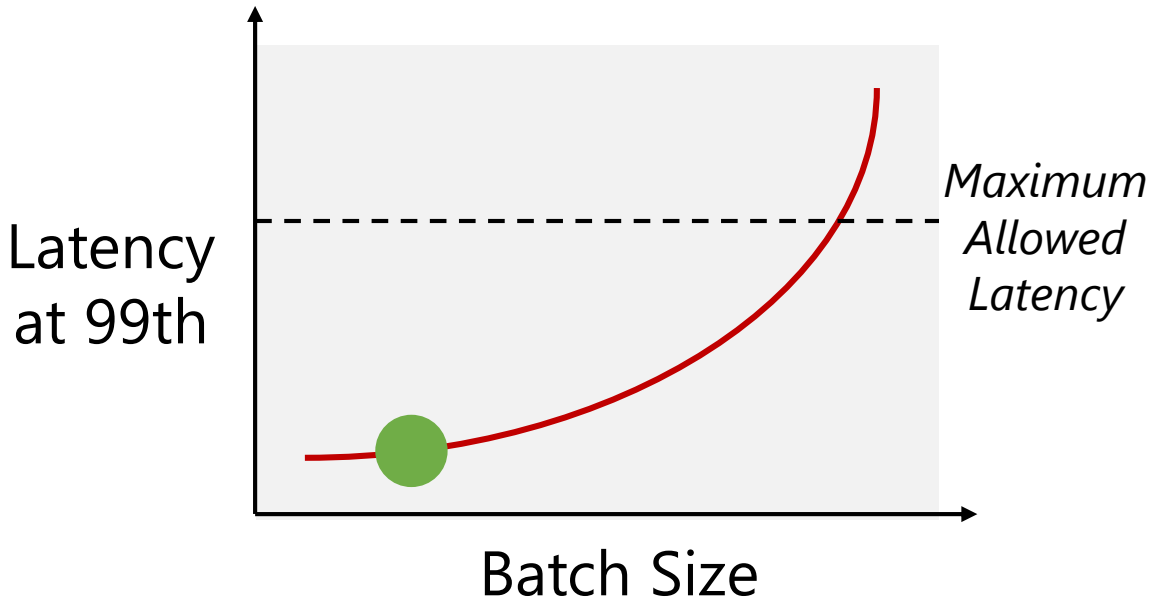




Batching improves HW utilization but also increases latency

Improving HW utilization with batching

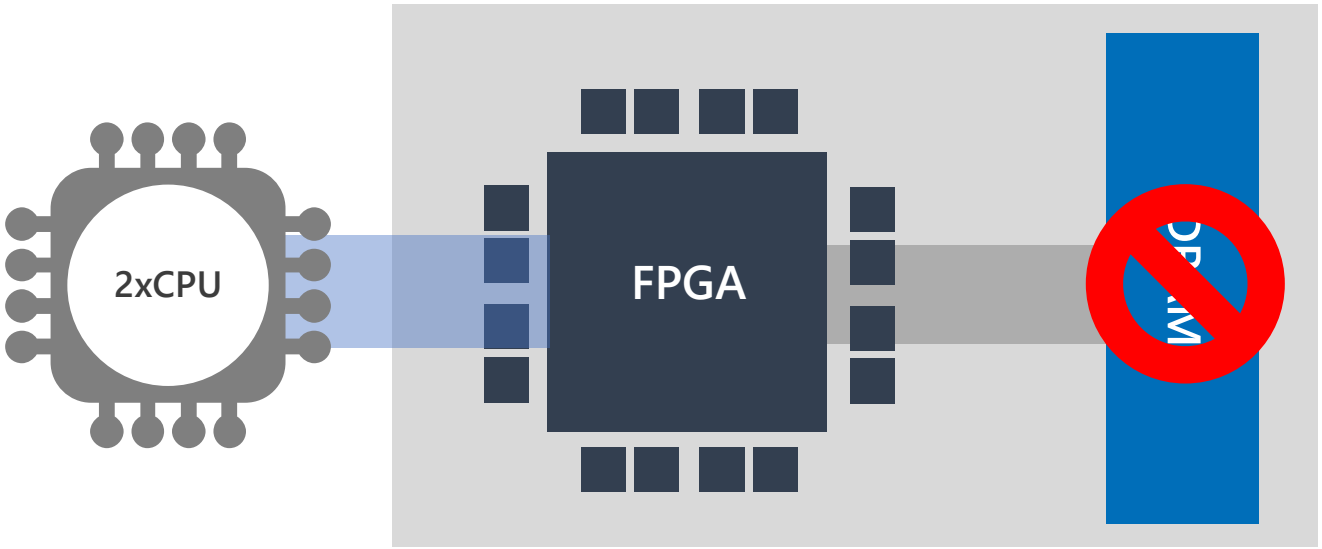




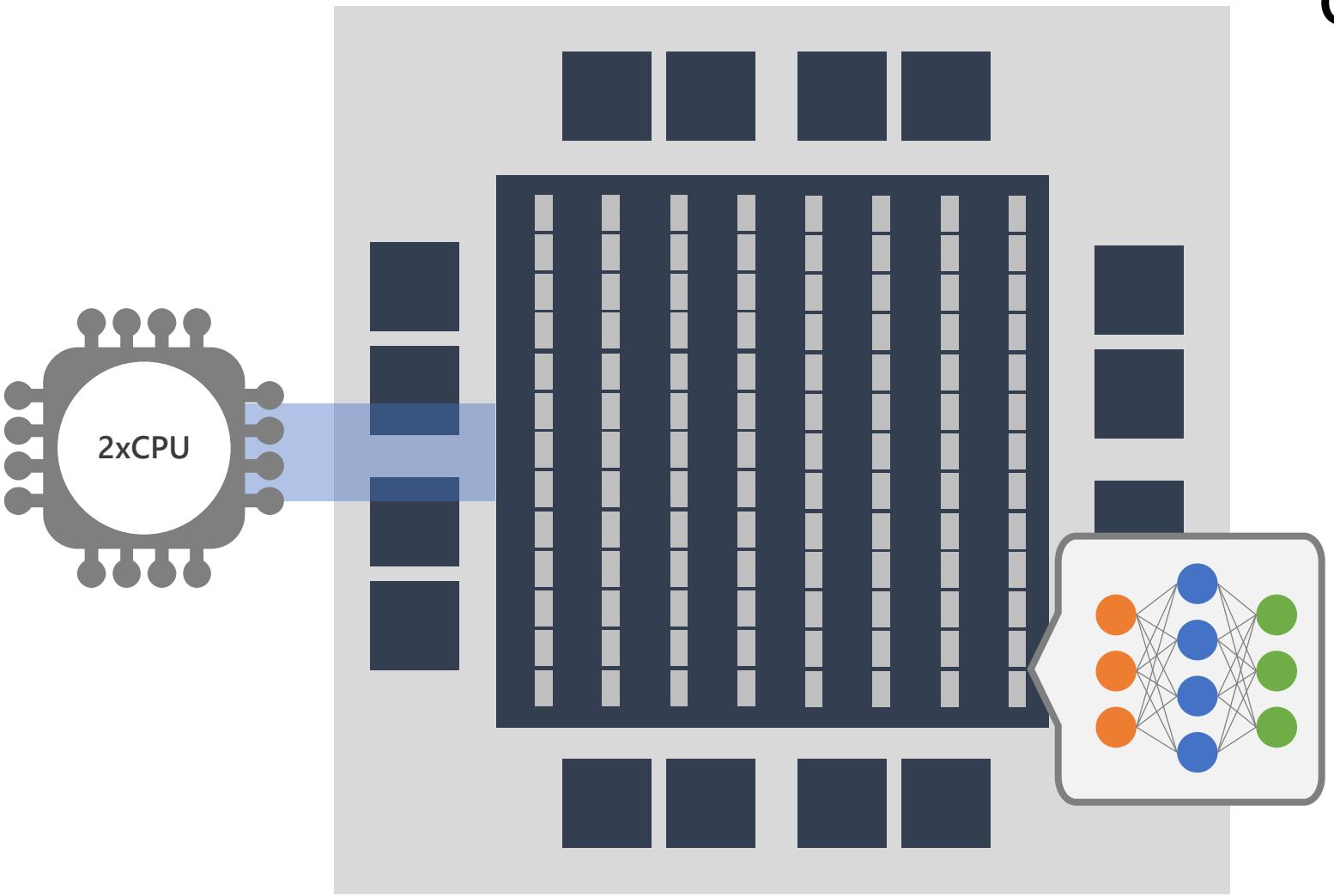
Batching improves HW utilization but increases latency

Ideally want high HW utilization at low batch sizes

Alternative: “persistent” neural nets



Alternative: "persistent" neural nets



Observations

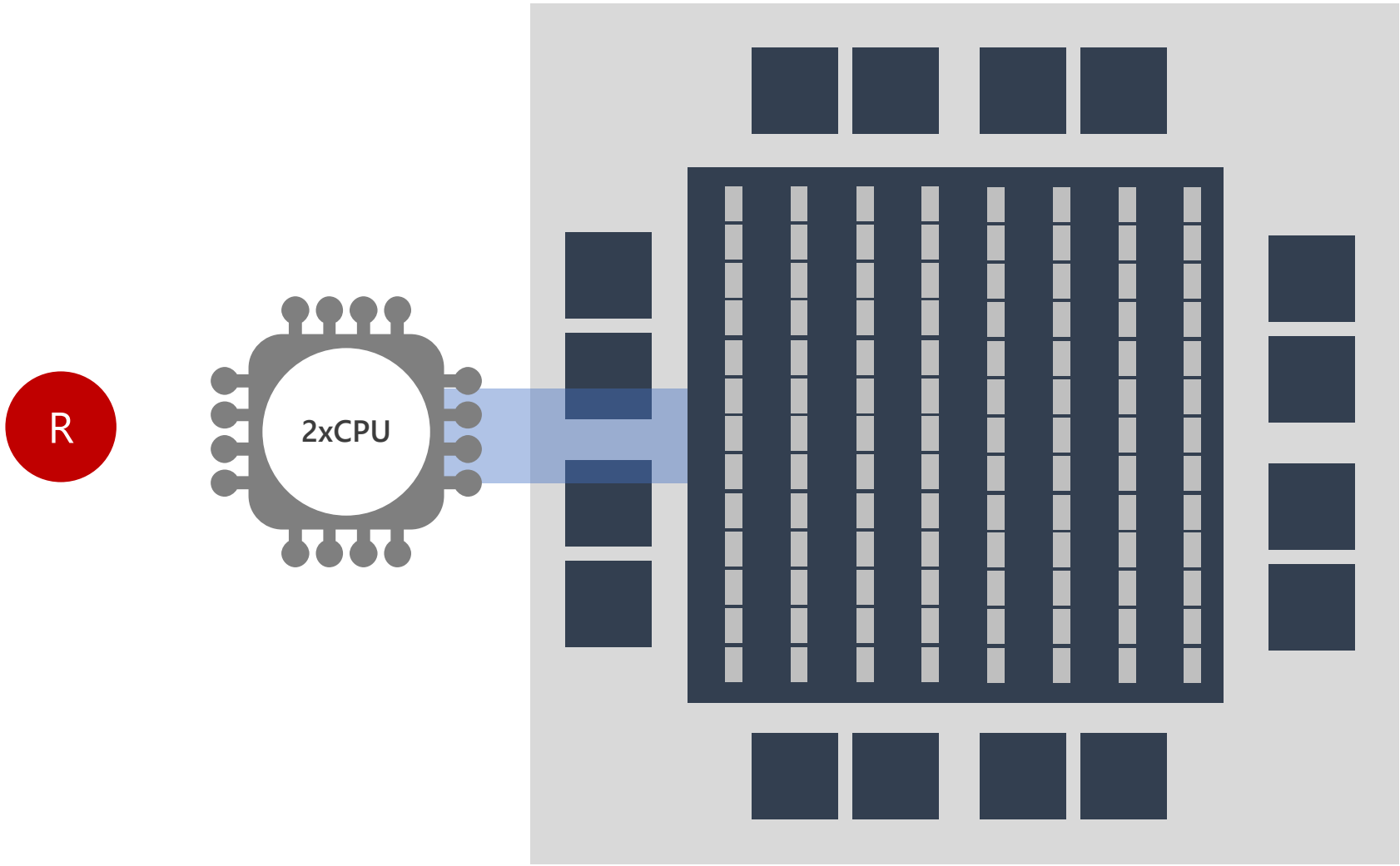
State-of-art FPGAs have O(10K) distributed Block RAMs O(10MB)

➔ Tens of TB/sec of memory BW

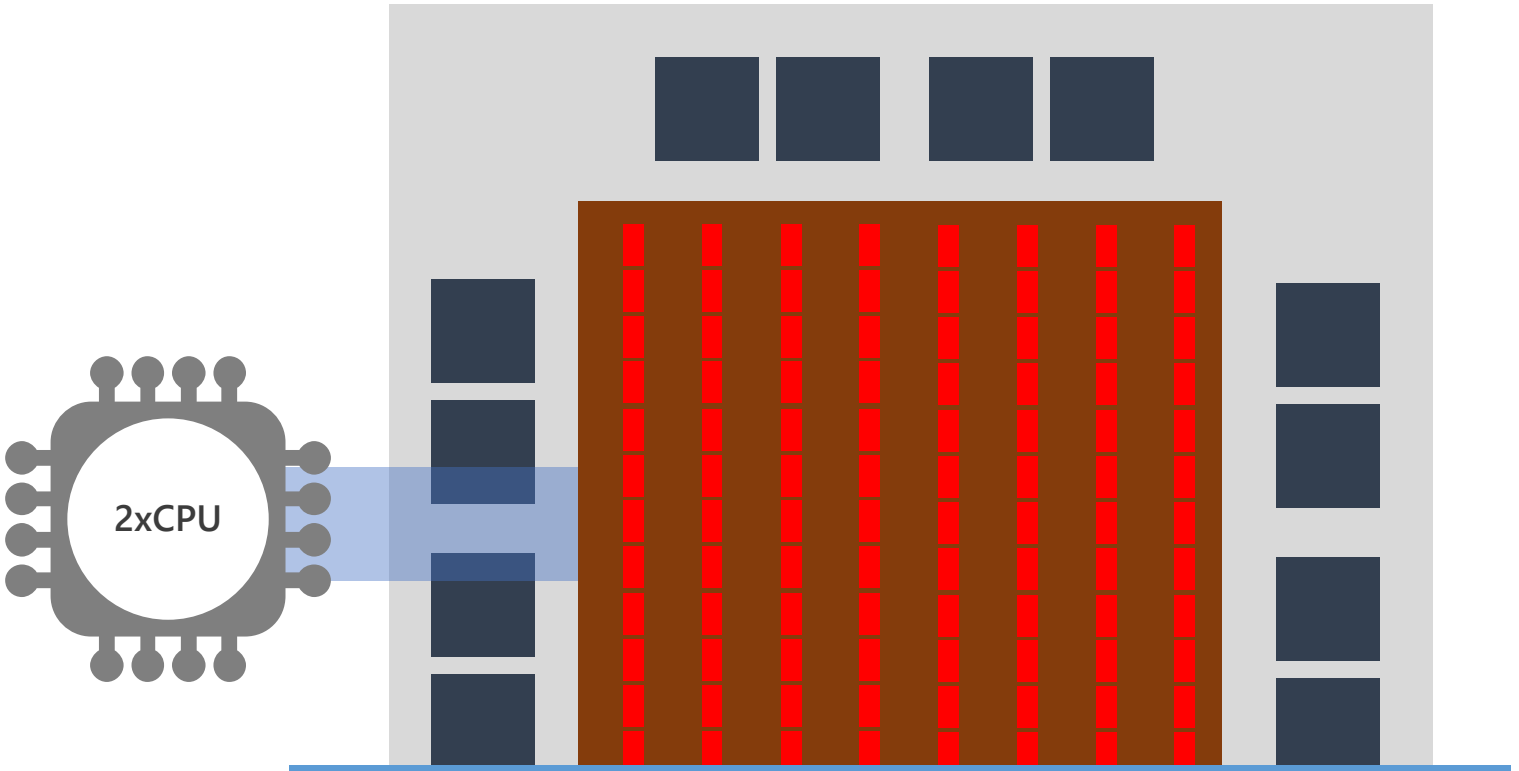
Large-scale cloud services and DNN models run persistently

Solution: persist all model parameters in FPGA on-chip memory during service lifetime

Alternative: “persistent” neural nets

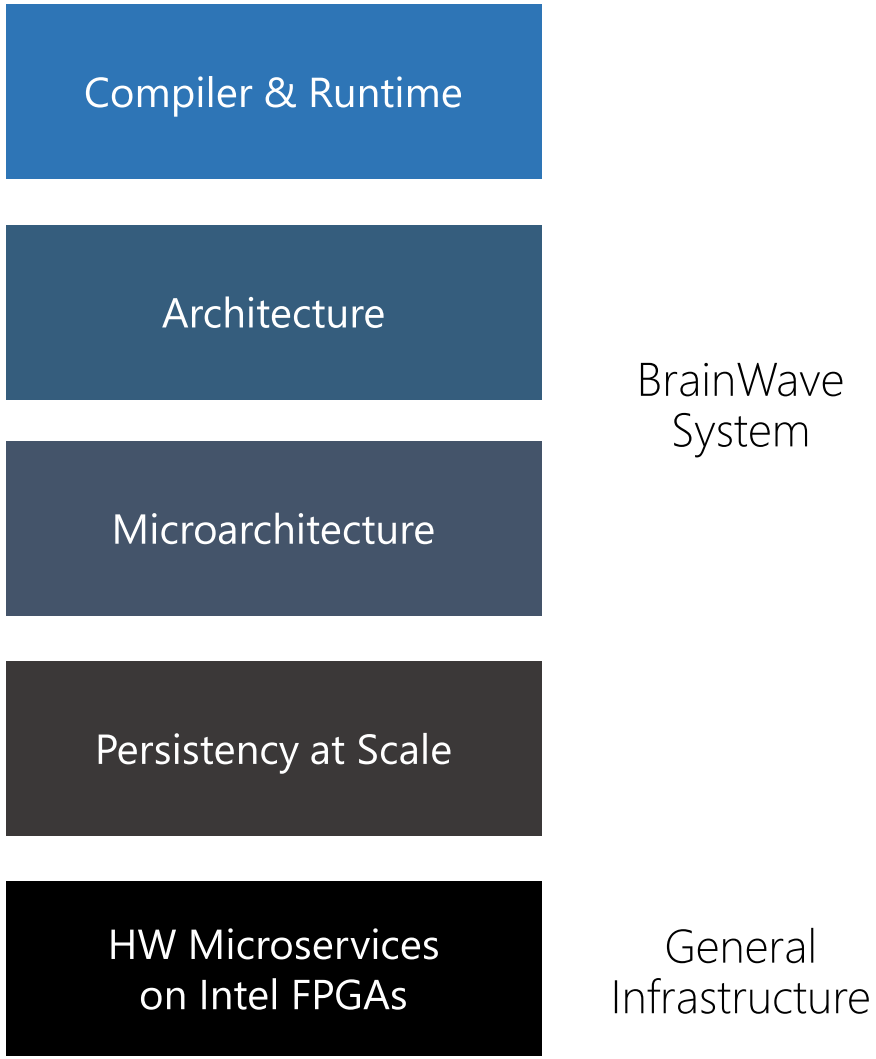


Alternative: “persistent” neural nets

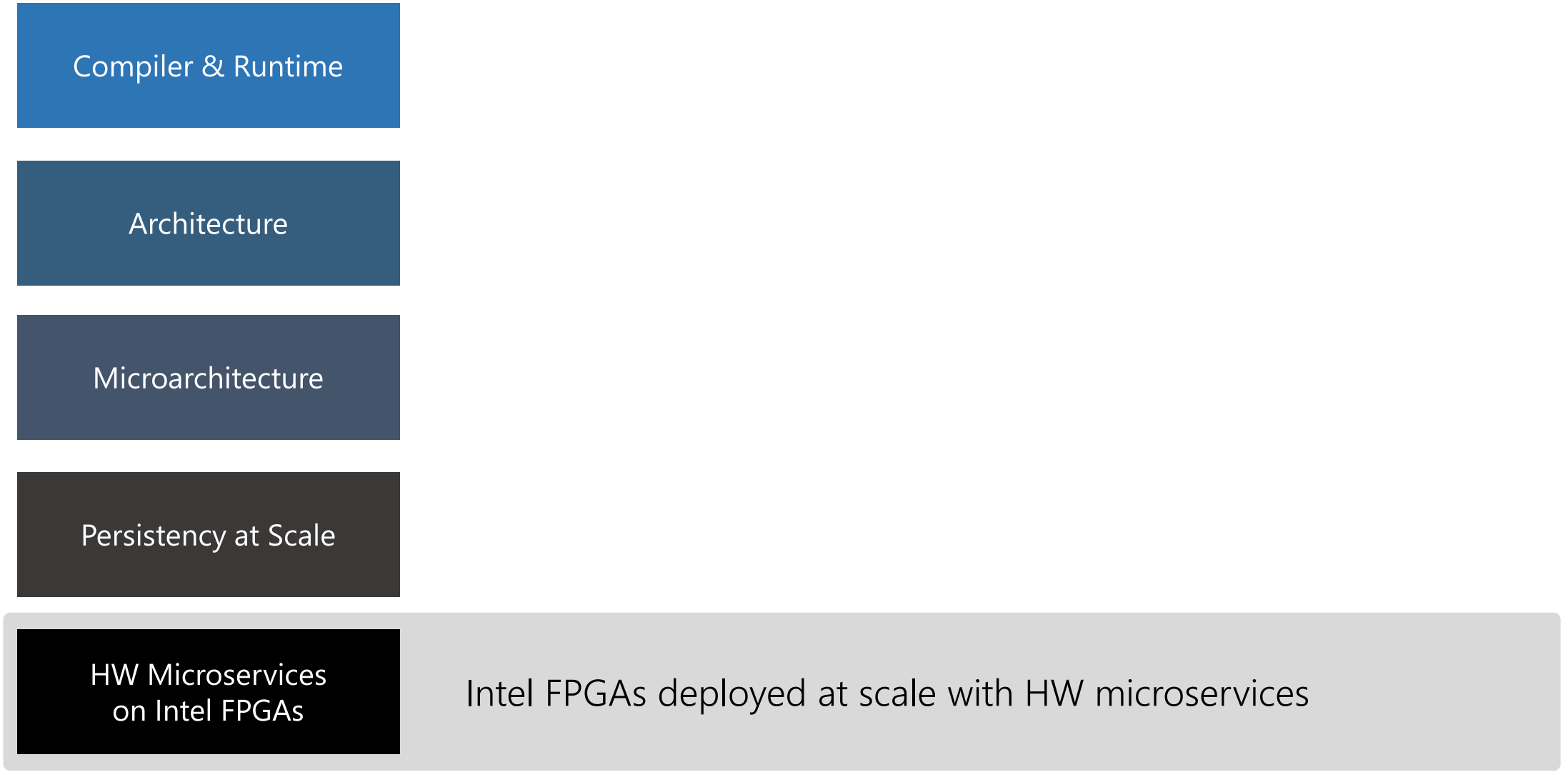


When single request arrives, all chip resources (on-chip memories and compute units) are used to process a single query (no batching required)

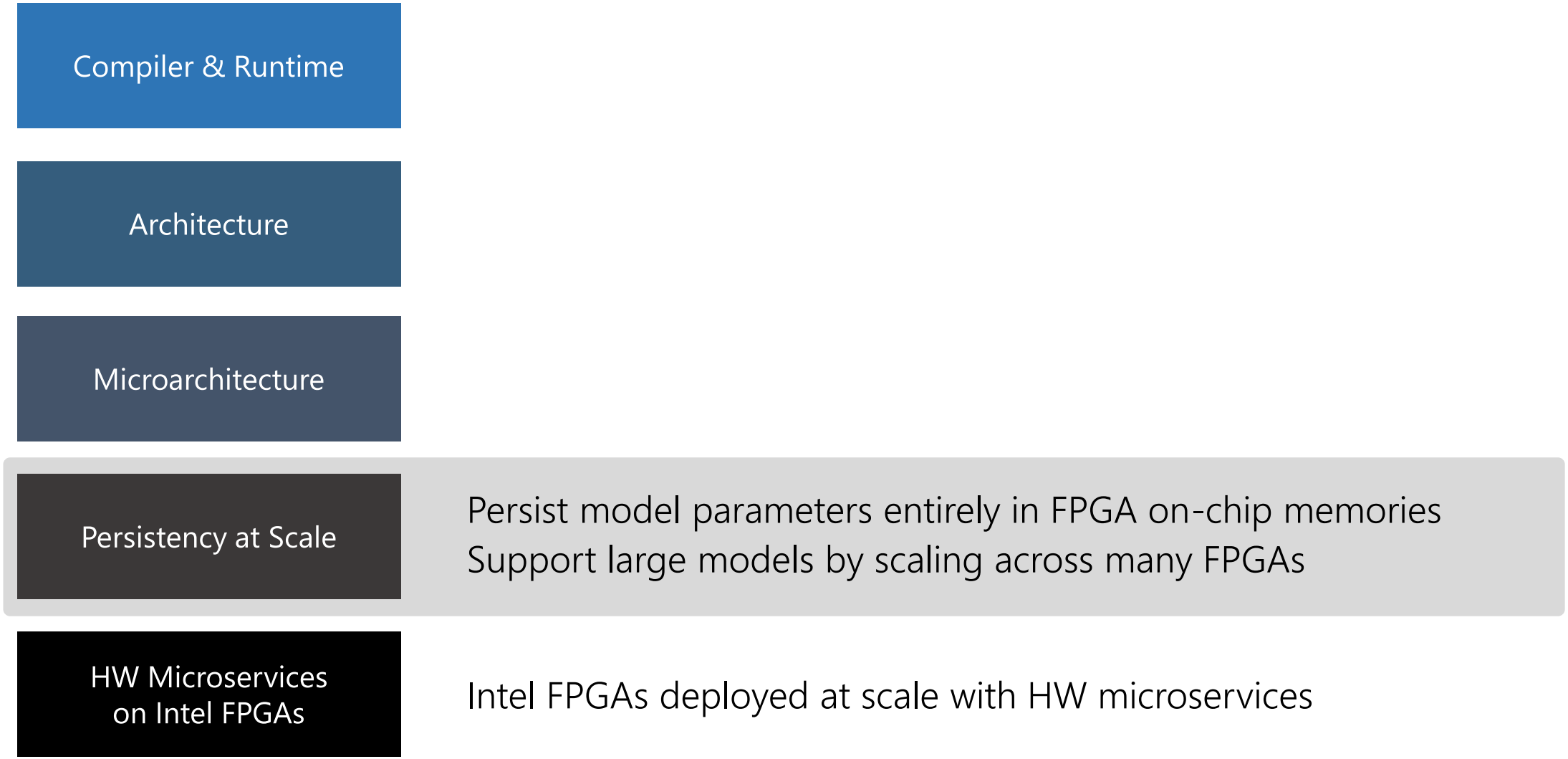
The Brainwave stack



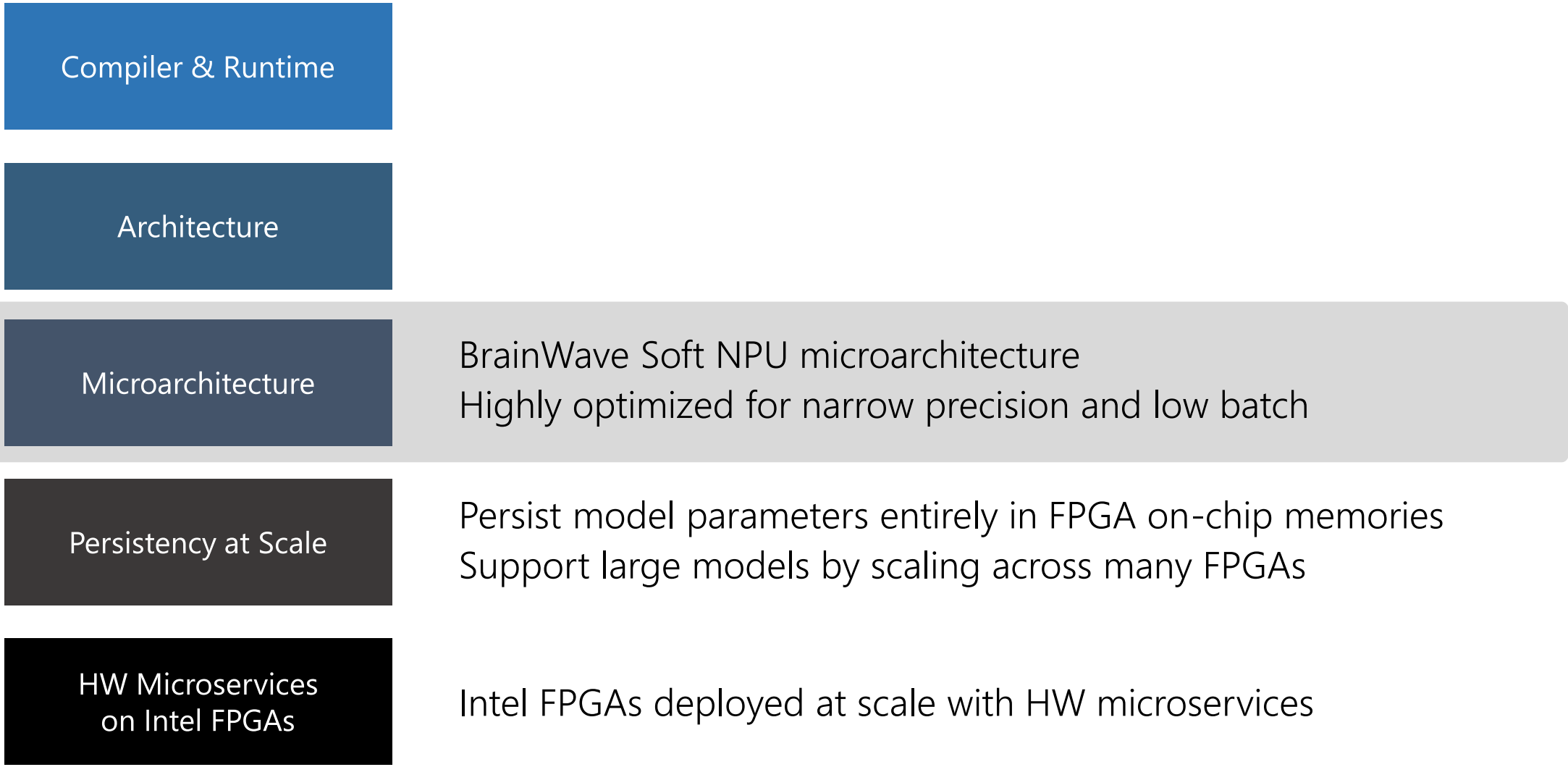
The Brainwave stack



The Brainwave stack



The Brainwave stack



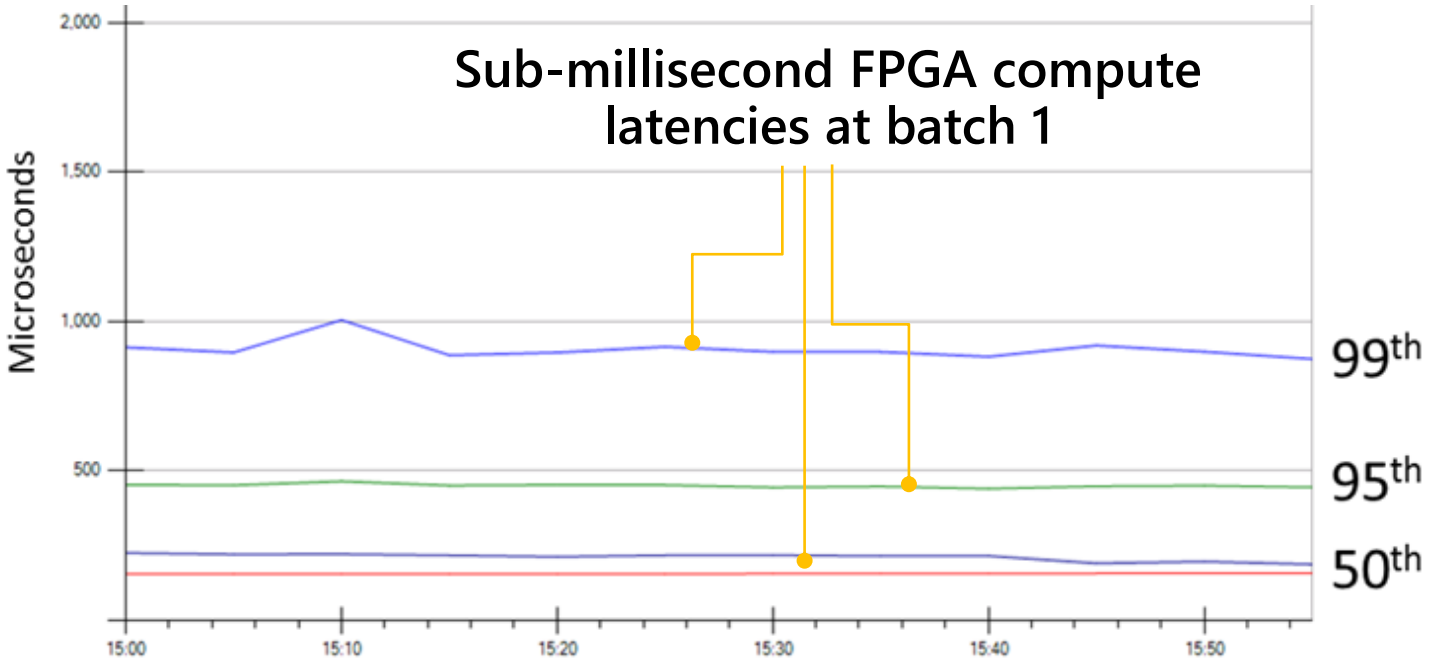
The Brainwave stack

Compiler & Runtime	
Architecture	Adaptive ISA for narrow precision DNN inference Flexible and extensible to support fast-changing AI algorithms
Microarchitecture	BrainWave Soft NPU microarchitecture Highly optimized for narrow precision and low batch
Persistency at Scale	Persist model parameters entirely in FPGA on-chip memories Support large models by scaling across many FPGAs
HW Microservices on Intel FPGAs	Intel FPGAs deployed at scale with HW microservices

The Brainwave stack

Compiler & Runtime	A framework-neutral federated compiler and runtime for compiling pretrained DNN models to soft NPUs
Architecture	Adaptive ISA for narrow precision DNN inference Flexible and extensible to support fast-changing AI algorithms
Microarchitecture	BrainWave Soft NPU microarchitecture Highly optimized for narrow precision and low batch
Persistency at Scale	Persist model parameters entirely in FPGA on-chip memories Support large models by scaling across many FPGAs
HW Microservices on Intel FPGAs	Intel FPGAs deployed at scale with HW microservices

Deployed in production datacenters



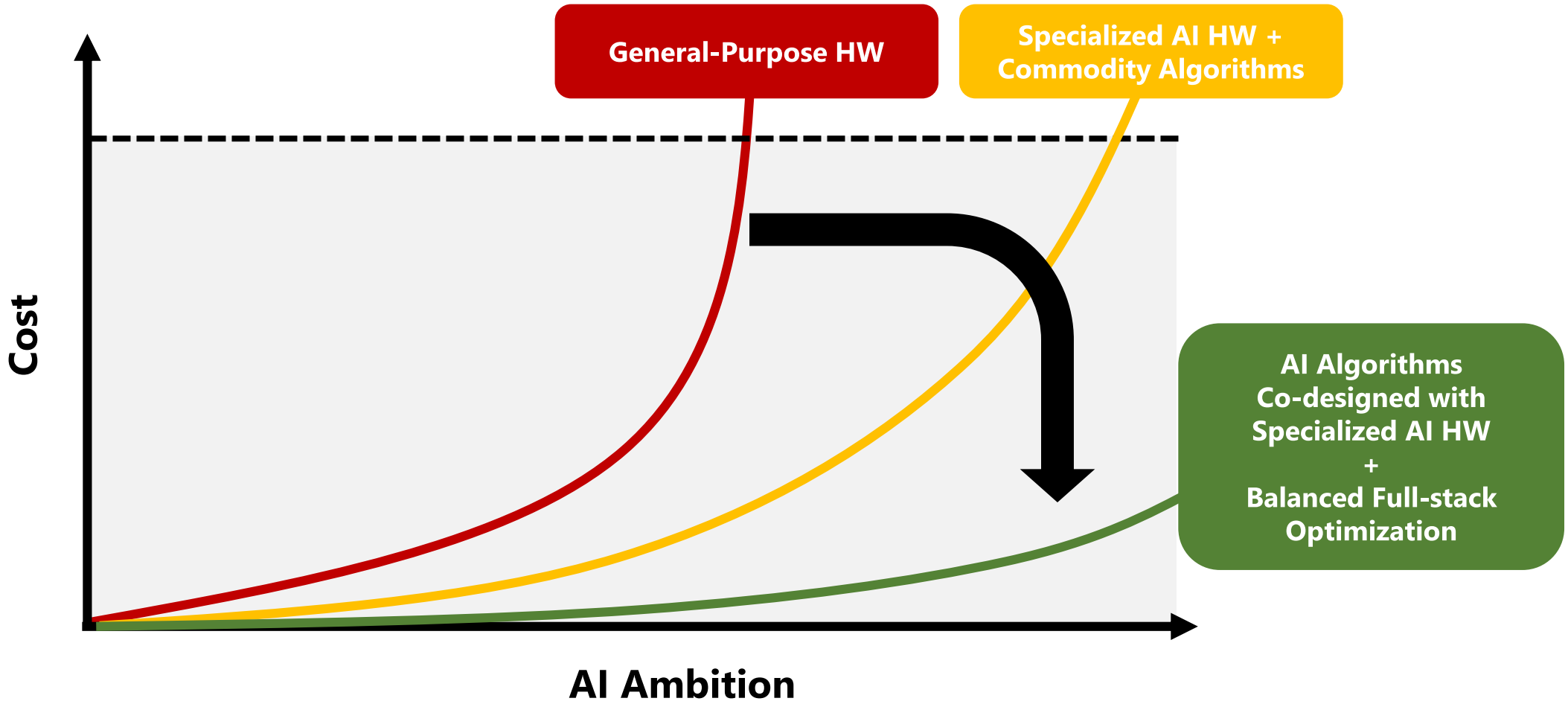
Deployment of LSTM-based NLP model (tens of millions of parameters)

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

Tail latencies in BrainWave-powered DNN models appear negligible in E2E software pipelines

Closing Thoughts …

Bending the AI ambition-cost curve



We are hiring …

Check out Azure AIArch for our open positions:

- Data & Applied Scientist
- Software Engineering
- Hardware Engineering

Send Resumes To:

hiring4azurehardware@microsoft.com

bita.rouhani@microsoft.com