Project Brainwave

Adesh Shambhu
Chiranjeevi HR

Background

Ubiquitous and Real Time AI

- Conversational agents
- Computer vision
- Natural language processing
- Intelligent Search Engines

Computational demand vs performance growth rate !

GPGPUs and batch-oriented NPUs s are popular for **offline** but not efficient for **online**

What to expect from the Hardware?

- ❖ More computational Power
- ❖ Less Latency
- ❖ Low cost
- ❖ High throughput and efficiency

Project Brainwave

Project Catapult
-the FPGA sits between the datacenter’s top-of-rack (ToR) network switches and the server’s network interface chip (NIC).

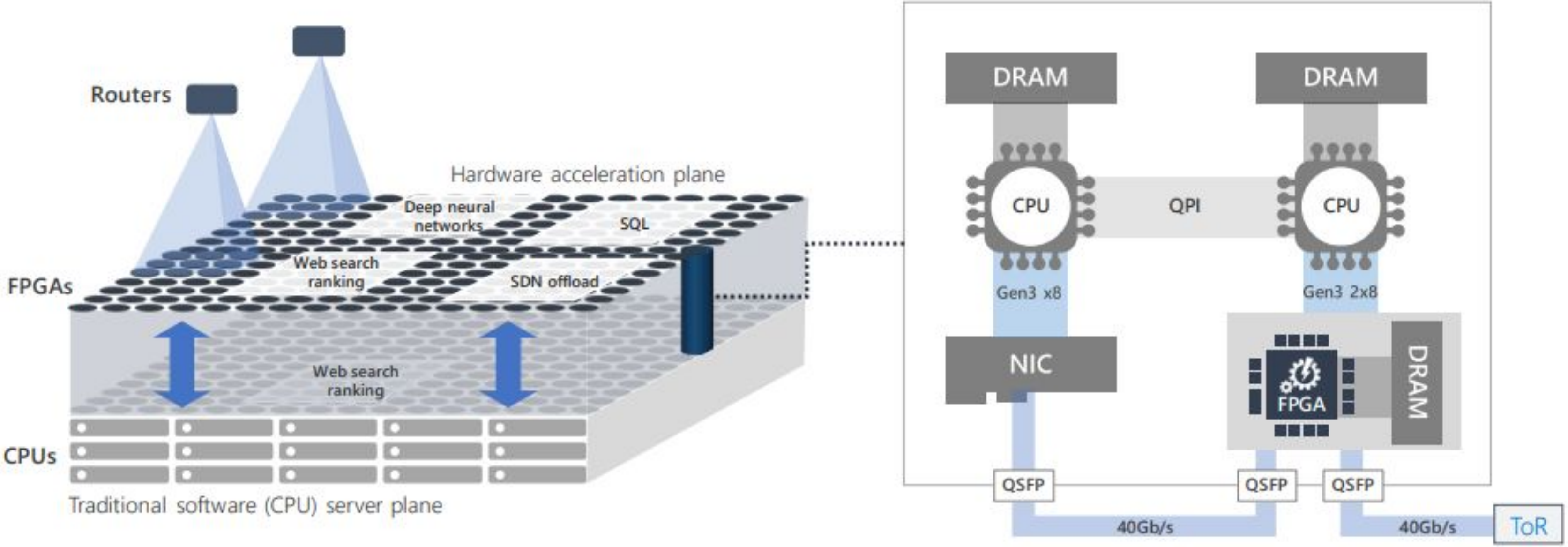
Hyperscale Data Centers at Microsoft.

- Thousands of FPGAs

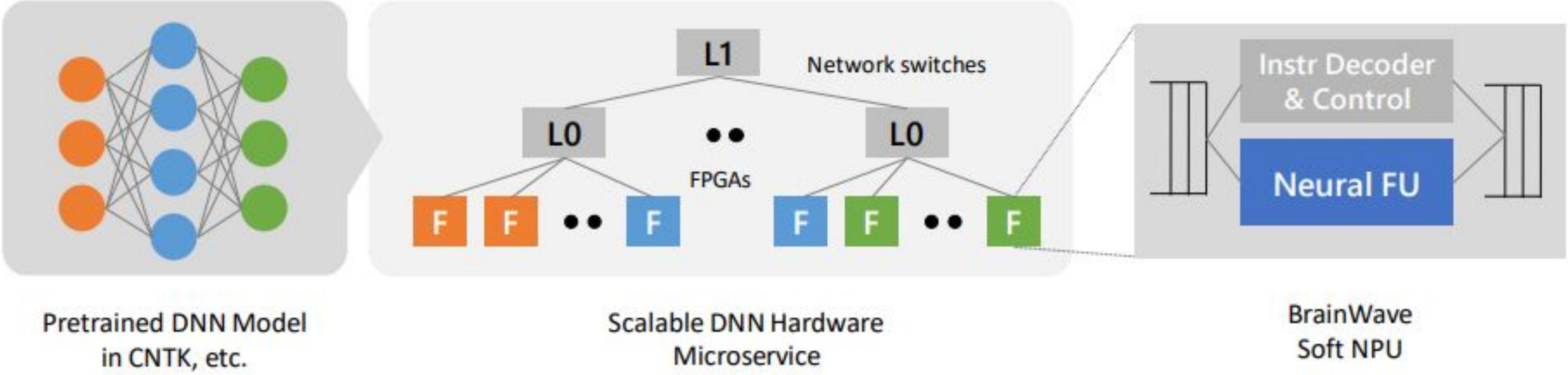
- first to prove the value of FPGAs for cloud computing,

- ➔ is a deep learning platform for real-time AI serving in the cloud
- ➔ High throughput, Ultra-low latency
- ➔ Leverages Intel FPGA infrastructure.
  - ◆ Pools of FPGAs
  - ◆ Callable by any CPU software on the shared network

Microsoft's hyperscale datacenter architecture



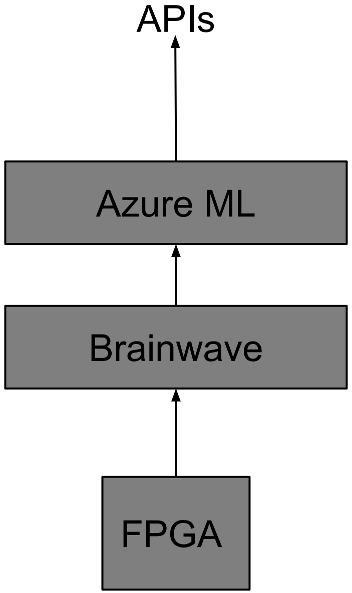
Project Brainwave Architecture



Source : https://www.microsoft.com/en-us/research/publication/serving-dnns-real-time-datacenter-scale-project-brainwave/

Microsoft Azure and Project Brainwave

- Azure provides End-to-End Data science Platform -Azure ML
- Data preparation and Model Training
- Hardware as a service
- Cloud or Edge Deployment
- 1 Azure Box = 24 CPU Cores + 4 Arria 10 FPGAs
- DNNs: ResNet 50, ResNet 152, VGG-16, SSD-VGG, and DenseNet-121



Results

- Brainwave successfully exploits FPGAs on a datacenter-scale fabric for real-time serving of state-of-the-art DNNs.
- Designing a scalable, end-to-end system architecture for deep learning is as critical as optimizing for single chip performance
- Today, Project Brainwave serves DNNs in real time for production services such as Bing Intelligent Search.