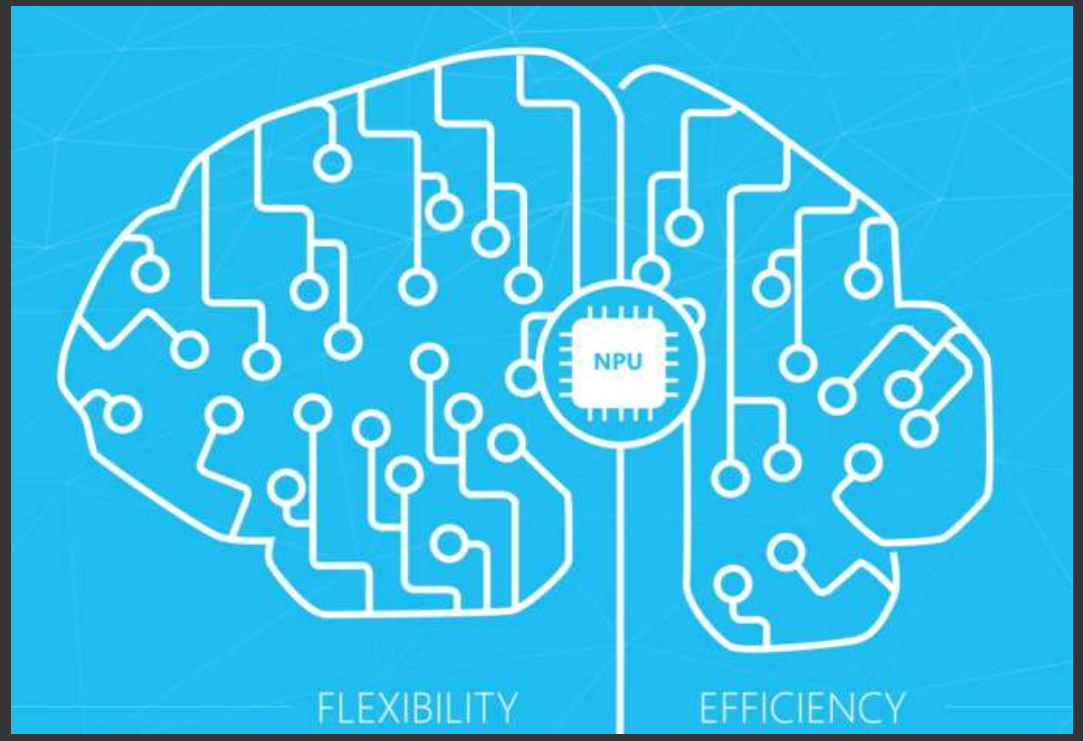
# Global-Scale FPGA-Accelerated Deep Learning Inference with Microsoft's Project Brainwave

Gabriel Weisz

Bing Engineering

Microsoft



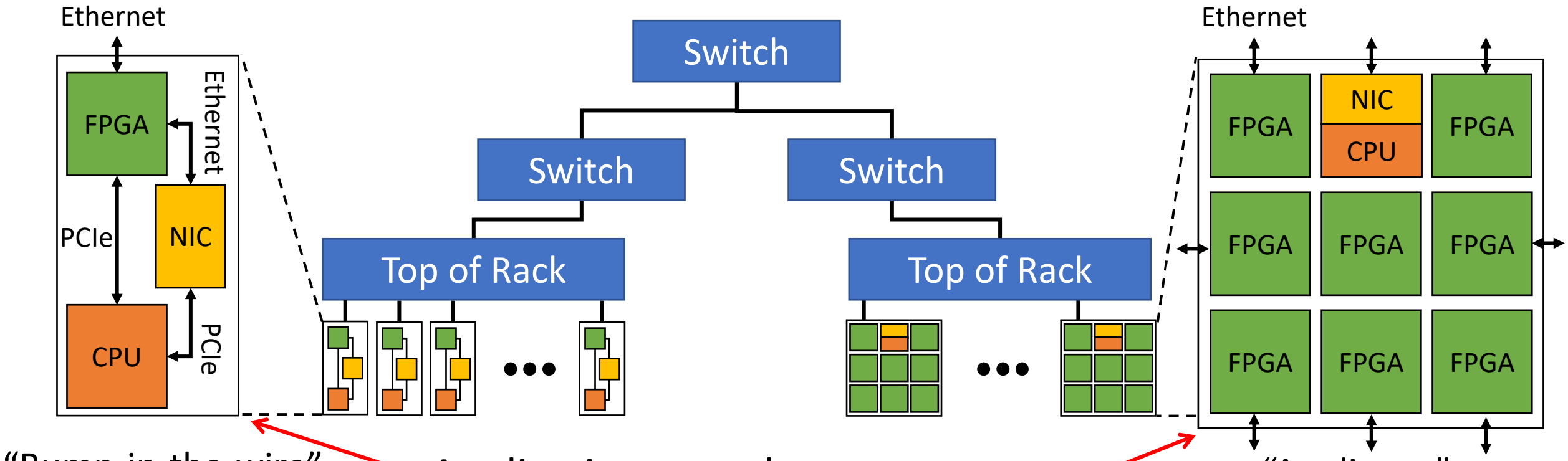

Microsoft

# Uses of Catapult FPGAs in Our Data Centers


Machine Learning


Accelerated Networking

# Catapult Network Architecture




"Bump-in-the-wire" configuration

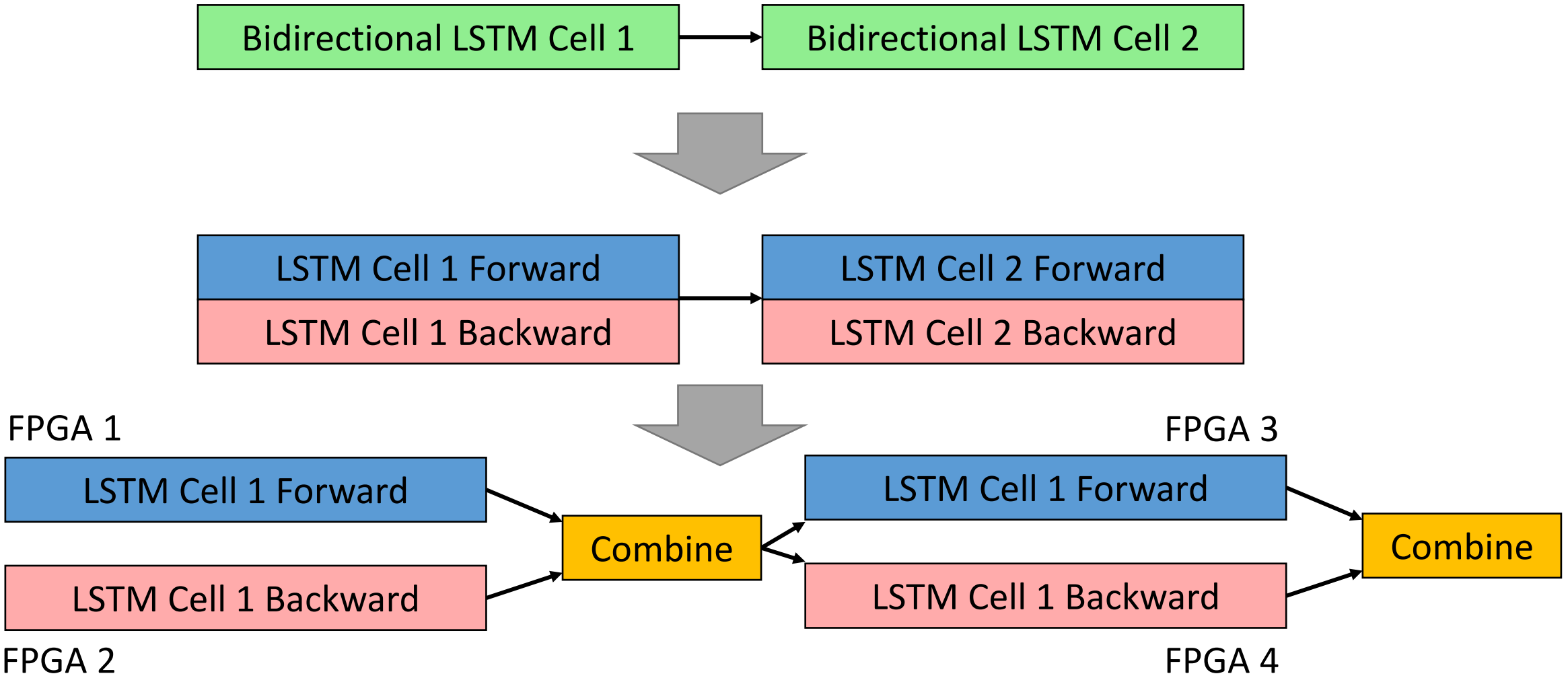
Applications see the same stream-based interface in either configuration

"Appliance" configuration

The appliance configuration is great for applications using >1 FPGA

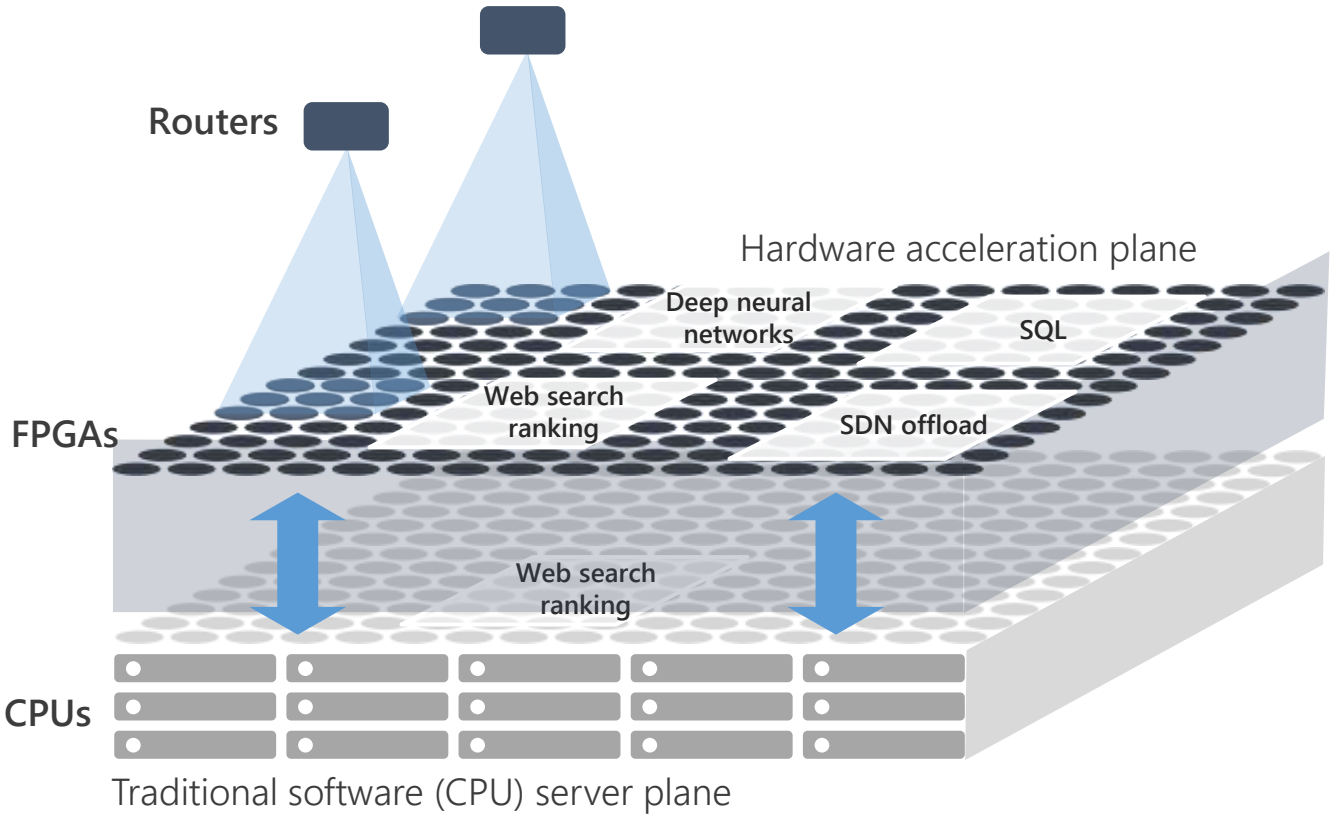
Microsoft

# Mapping a DNN Model to Multiple FPGAs



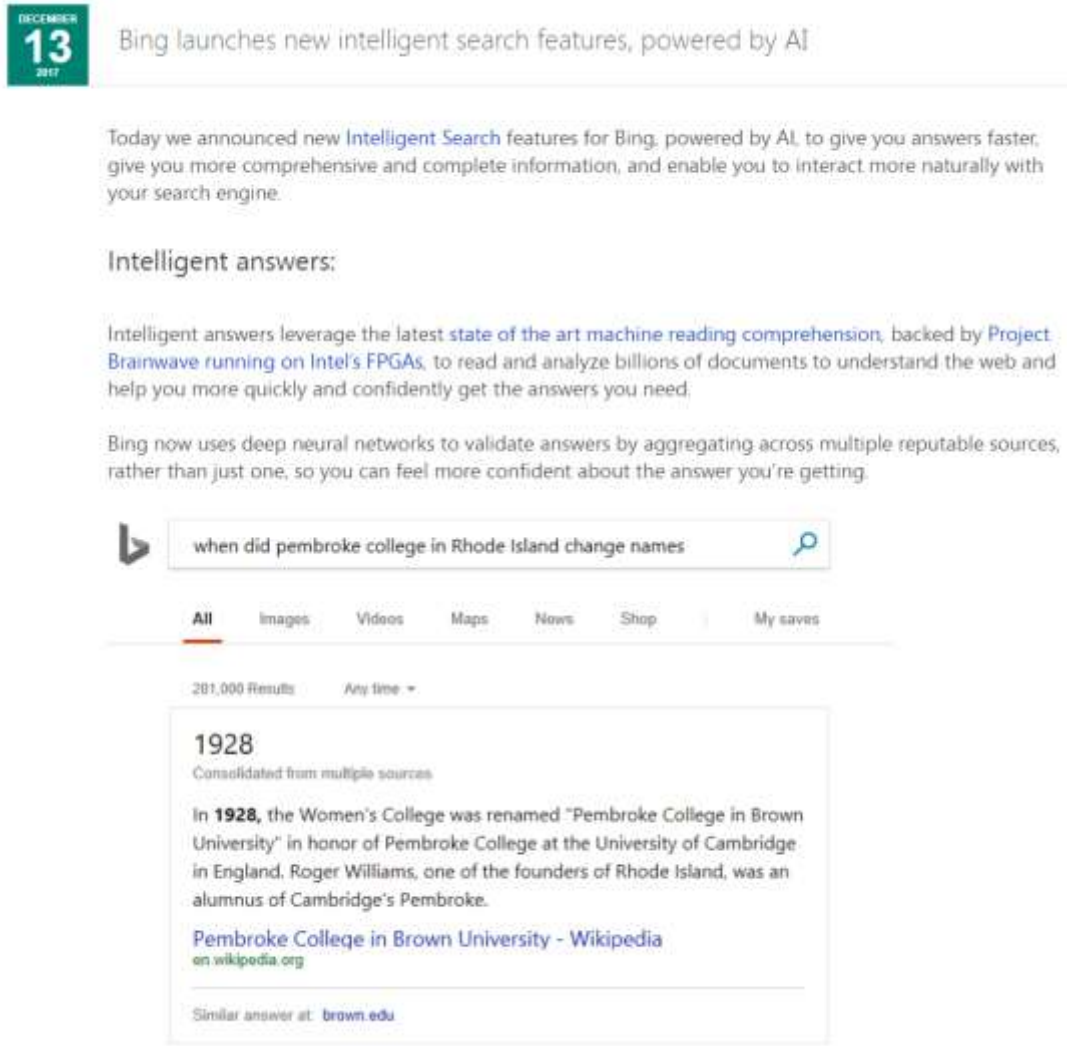

Microsoft

# Catapult + Software = Hardware Microservices




- **Interconnected FPGAs form a separate plane of computation**
- **FPGAs are used and managed independently from the CPU**
- **Applications are mapped across multiple FPGAs and CPUs**

# Bing Intelligent Search Powered By Brainwave



DECEMBER 13 2017

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.



FPGA-Accelerated model is **much** faster even though it is more complicated

| 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 | |

CPU vs Stratix V performance on production models

Microsoft

[Slide courtesy Eric Chung]

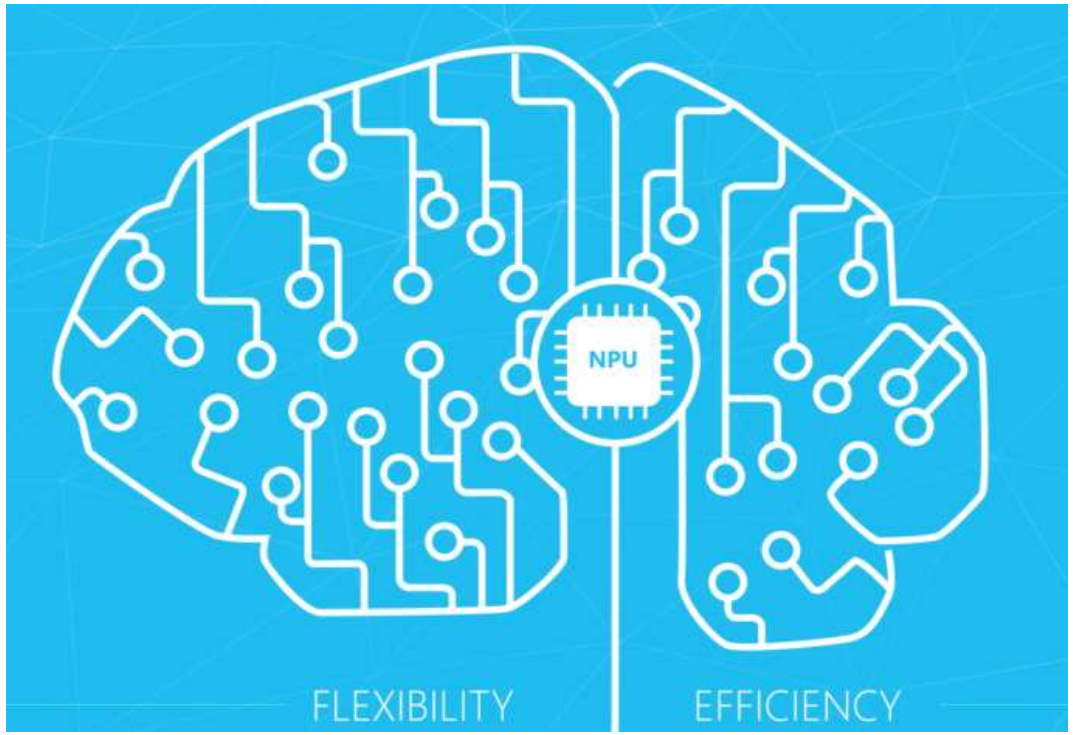
# Brainwave Components

FPGA-based overlay ("NPU")

- Highly parameterized
- Run-time programmable
- Supports multiple FPGA device generations

Enterprise-grade software stack

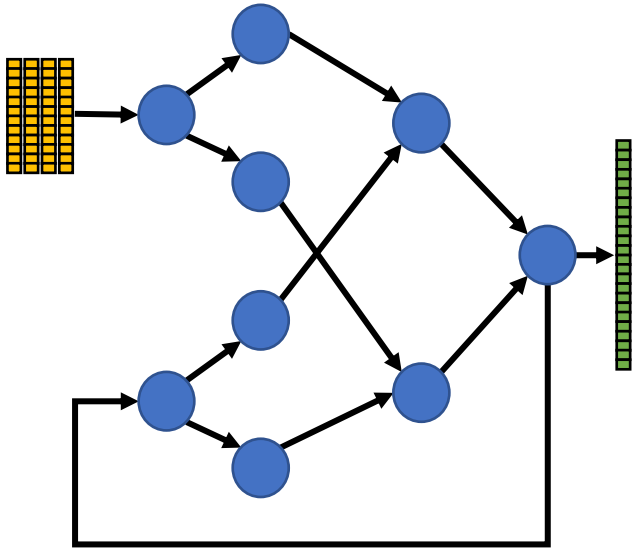
- FPGA management
- Orchestration of computations
- Model compiler



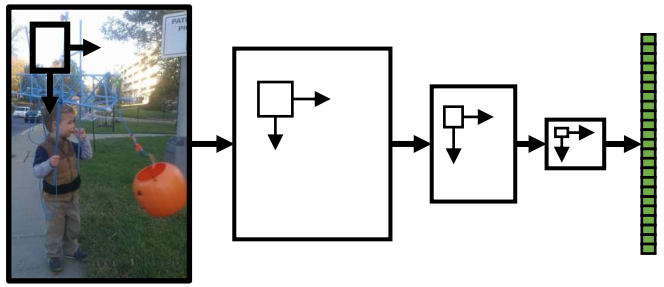

**This talk focuses on the overlay**

Microsoft

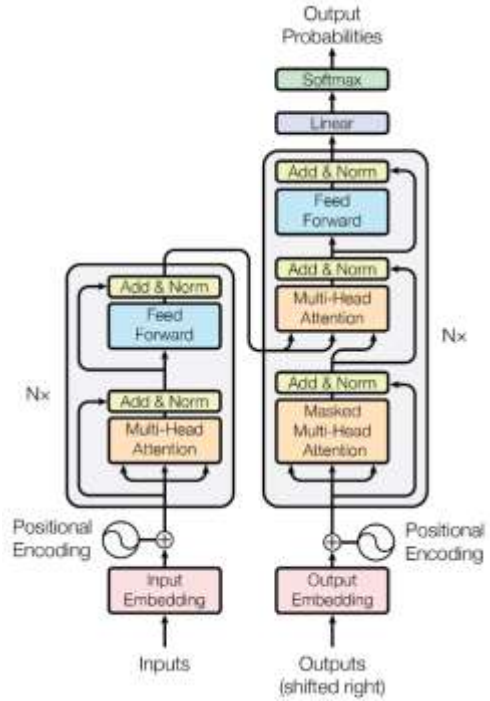
# Deep Learning Network Topologies



Recurrent Networks



Convolutional Networks



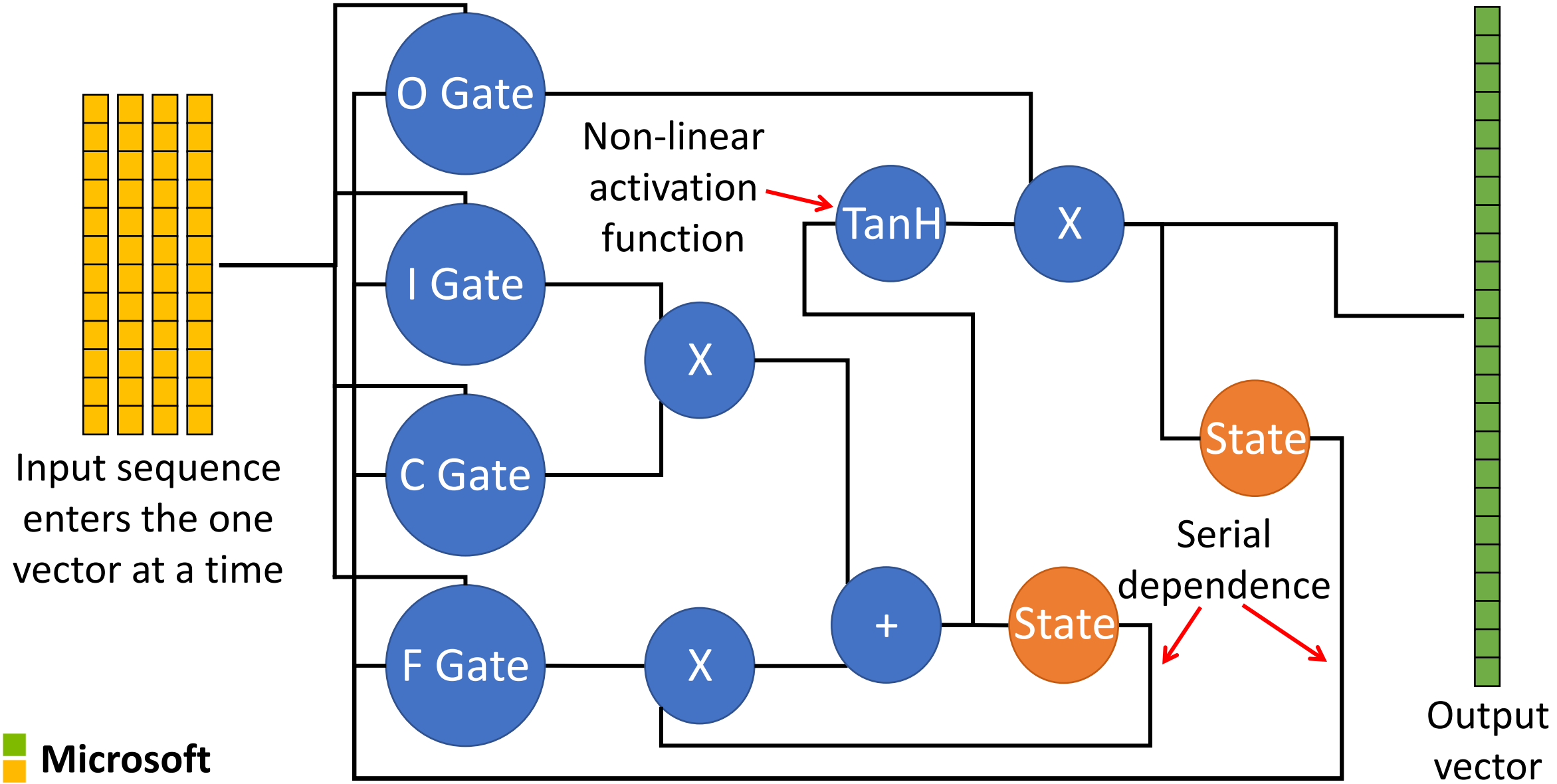

[Vaswani+, "Attention is all You Need", arXiv]

Transformer Networks

**What computations do we need to support?**

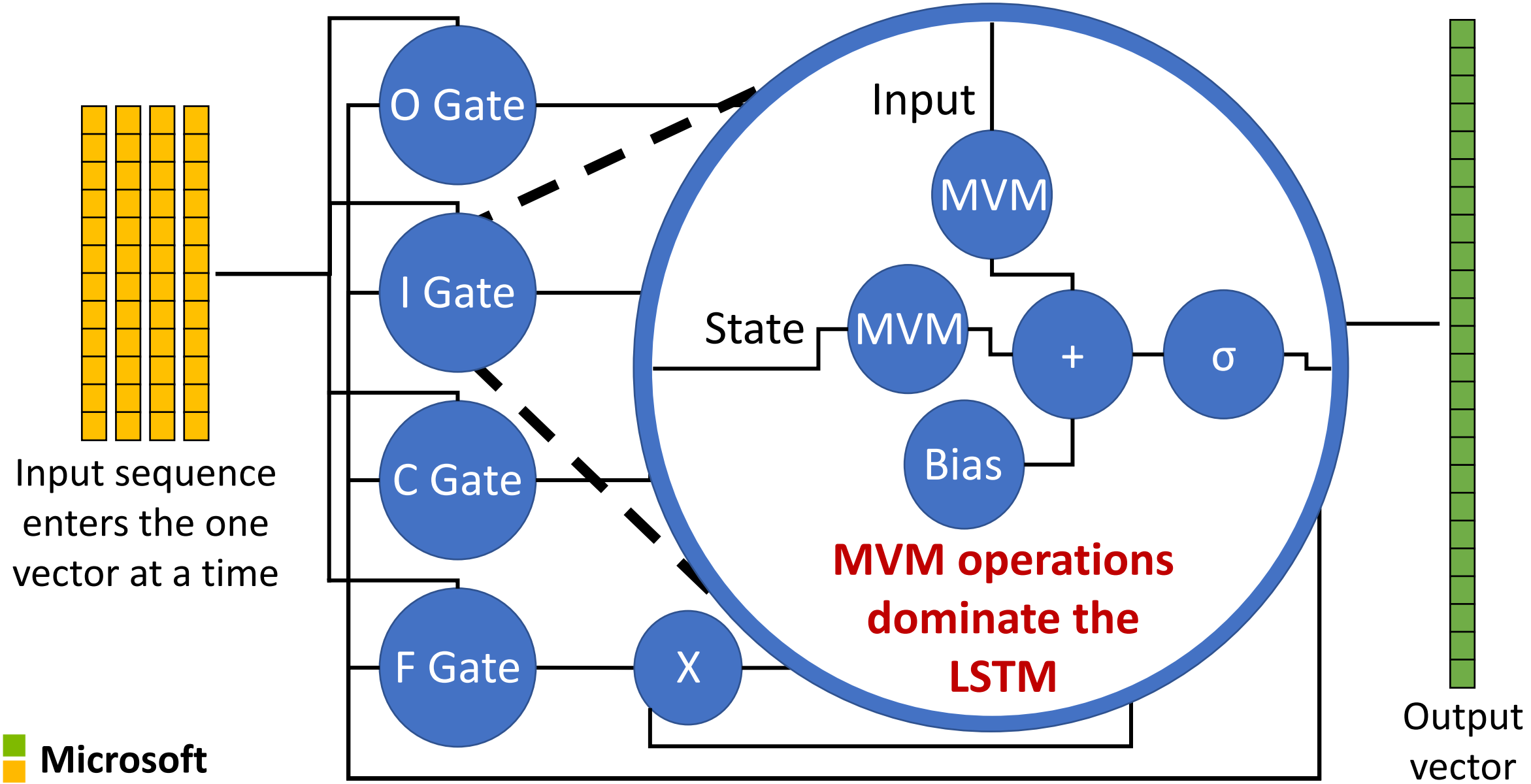
Microsoft

# Example RNN: Long Short-Term Memory (LSTM)



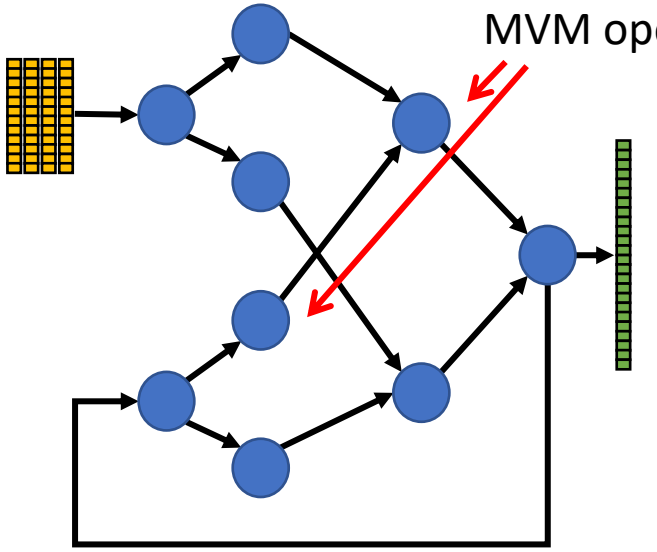

Microsoft

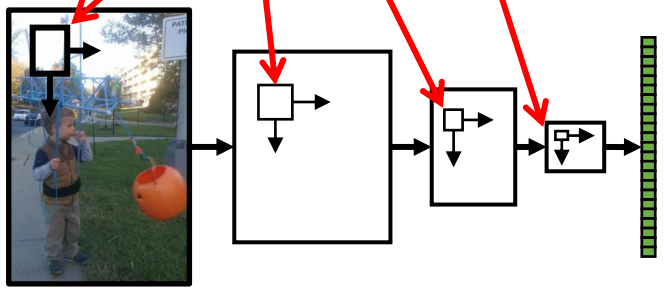
# Example RNN: Long-Short Term Memory (LSTM)

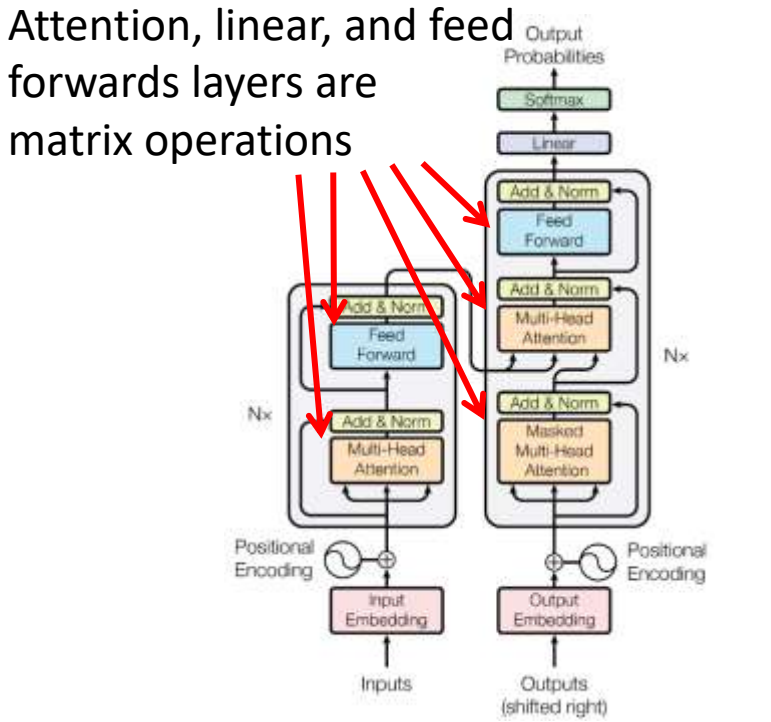



Microsoft

# (Almost) Everything is a Matrix Operation




[Vaswani+, "Attention is all You Need", arXiv]

Recurrent Networks

Convolutional Networks

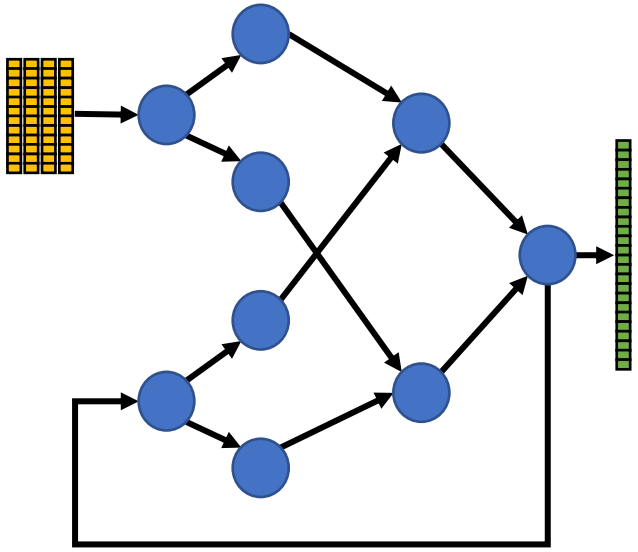
Transformer Networks

**What other computations are important?**

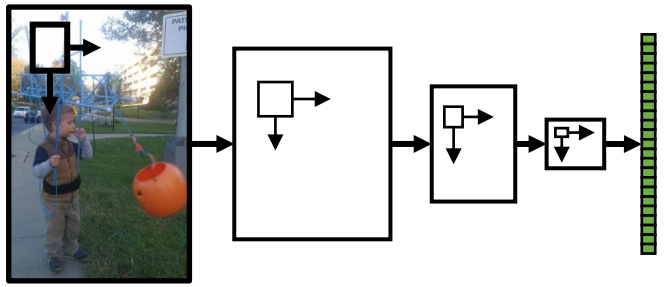
Microsoft

# What Isn’t a Matrix Operation?

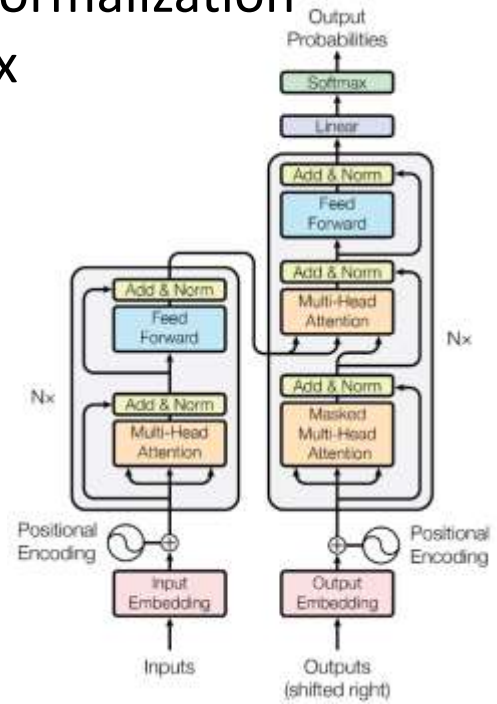
Non-linear Activation Functions
Vector-vector operations



Pooling operations
Batch Normalization



Layer Normalization
SoftMax




[Vaswani+, “Attention is all You Need”, arXiv]

Recurrent Networks

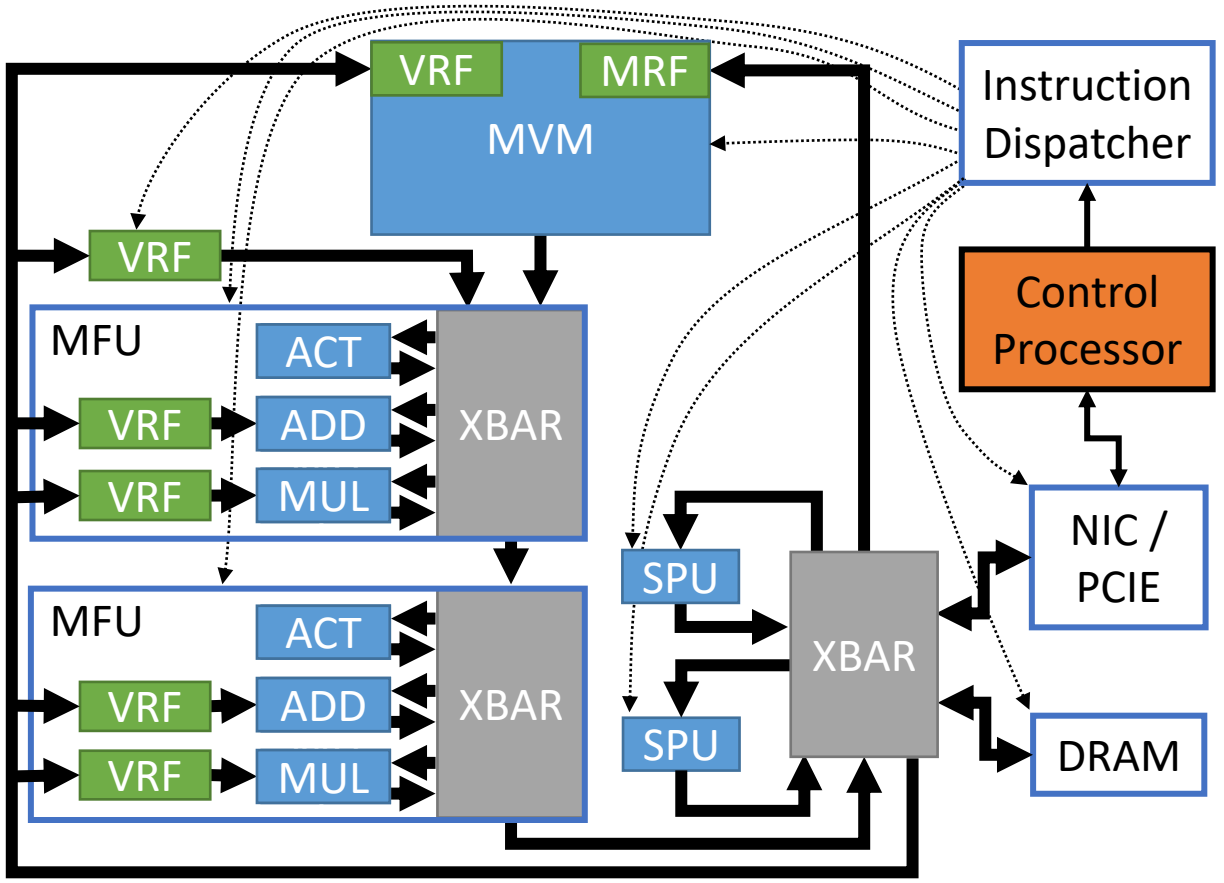
Convolutional Networks

Transformer Networks

**These operations must also run on the FPGA**

Microsoft

# Brainwave Overlay Architecture




| | | | |
|---|---|---|---|
| VRF | Vector Register File | MFU | Vector Multi-Function Unit |
| MRF | Matrix Register File | ACT | Activation Unit |
| MVM | Matrix Vector Multiply | ADD | Vector Elementwise Add |
| SPU | Special Purpose Unit | MUL | Vector Hadamard Product |

Microsoft

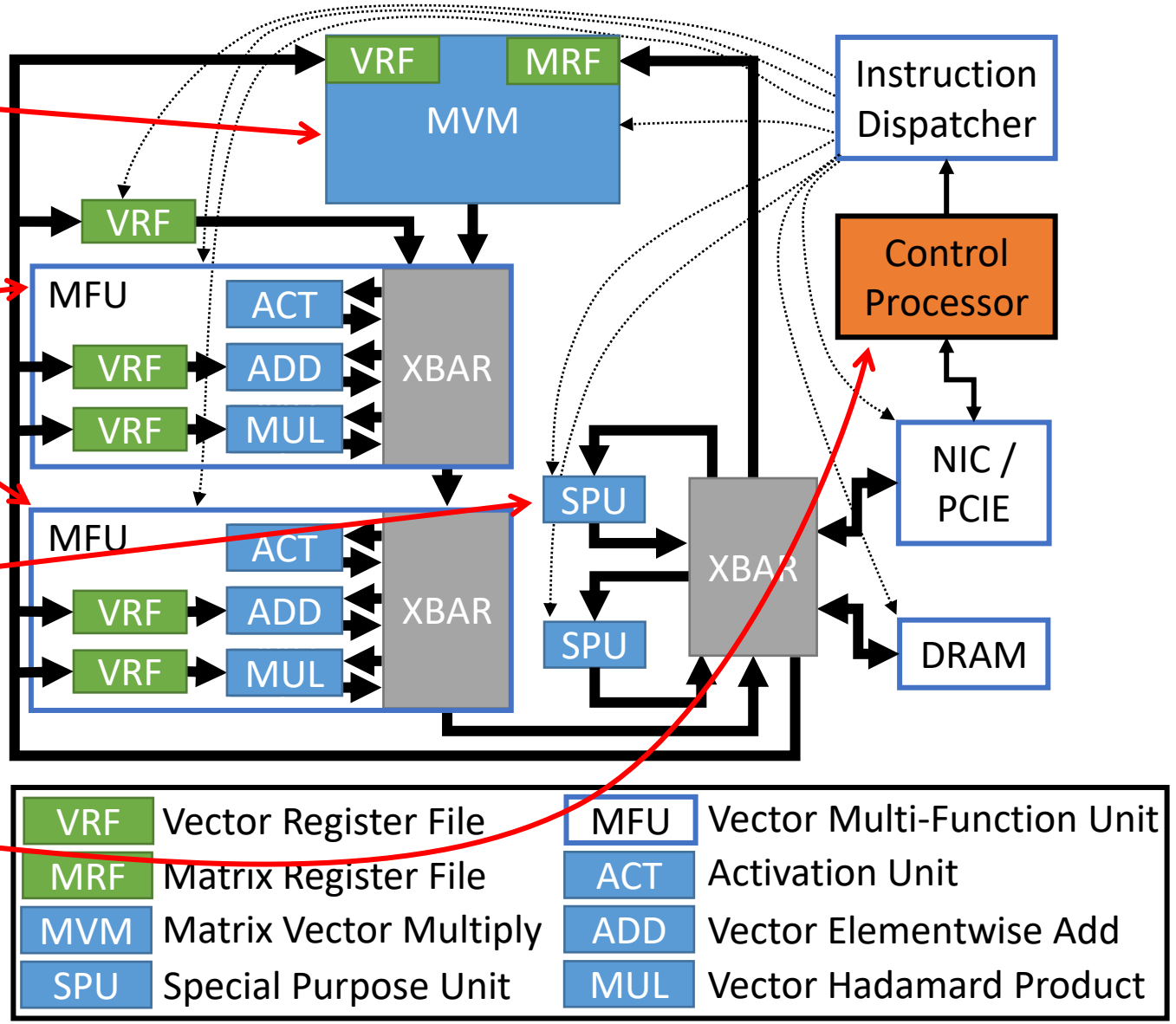
# Brainwave Overlay Architecture

Most of the computation is in the MVM

Data flows from the MVM to MFUs containing vector function units in a software-defined pipeline

Special processing units allow sub-vector operations

Control processor runs a program that maps the DNN model to Brainwave




Microsoft

# Mapping DNN Operators to Brainwave

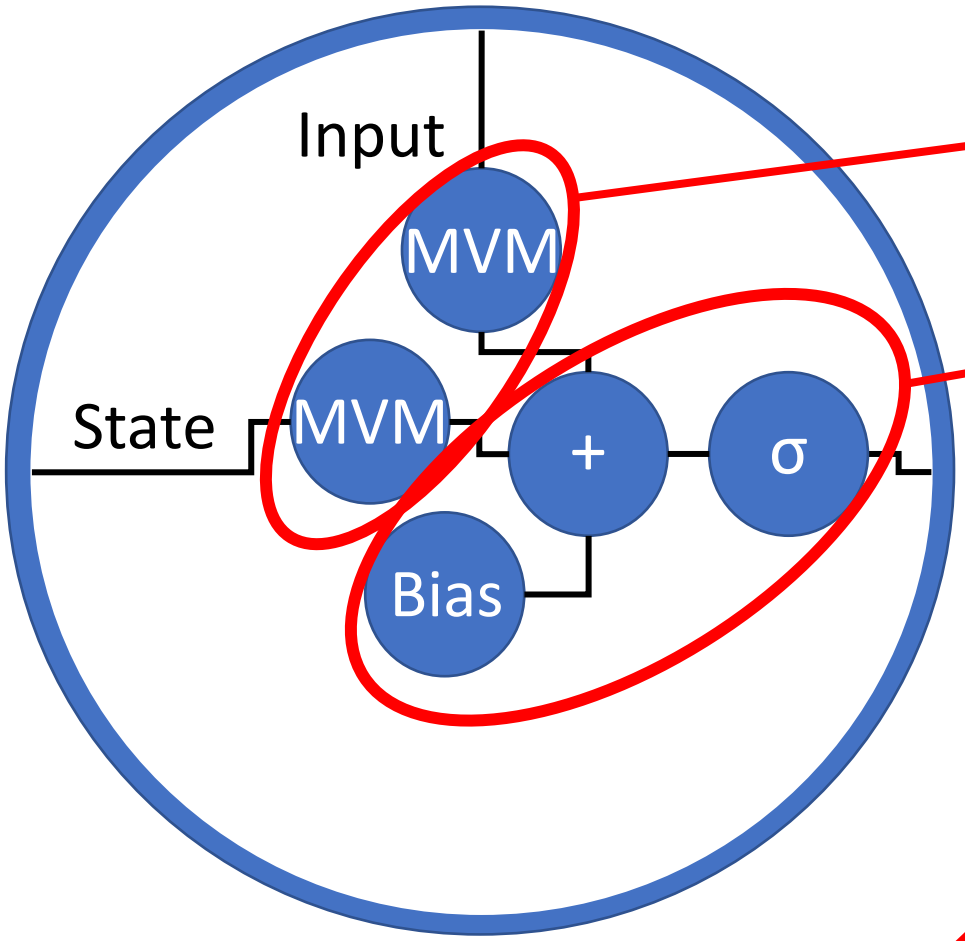
**Operations common in Deep Learning Networks**

LSTM
GRU
Convolution
SoftMax
Bias
Scale
Max Pool
Batch Norm
Sigmoid
TanH

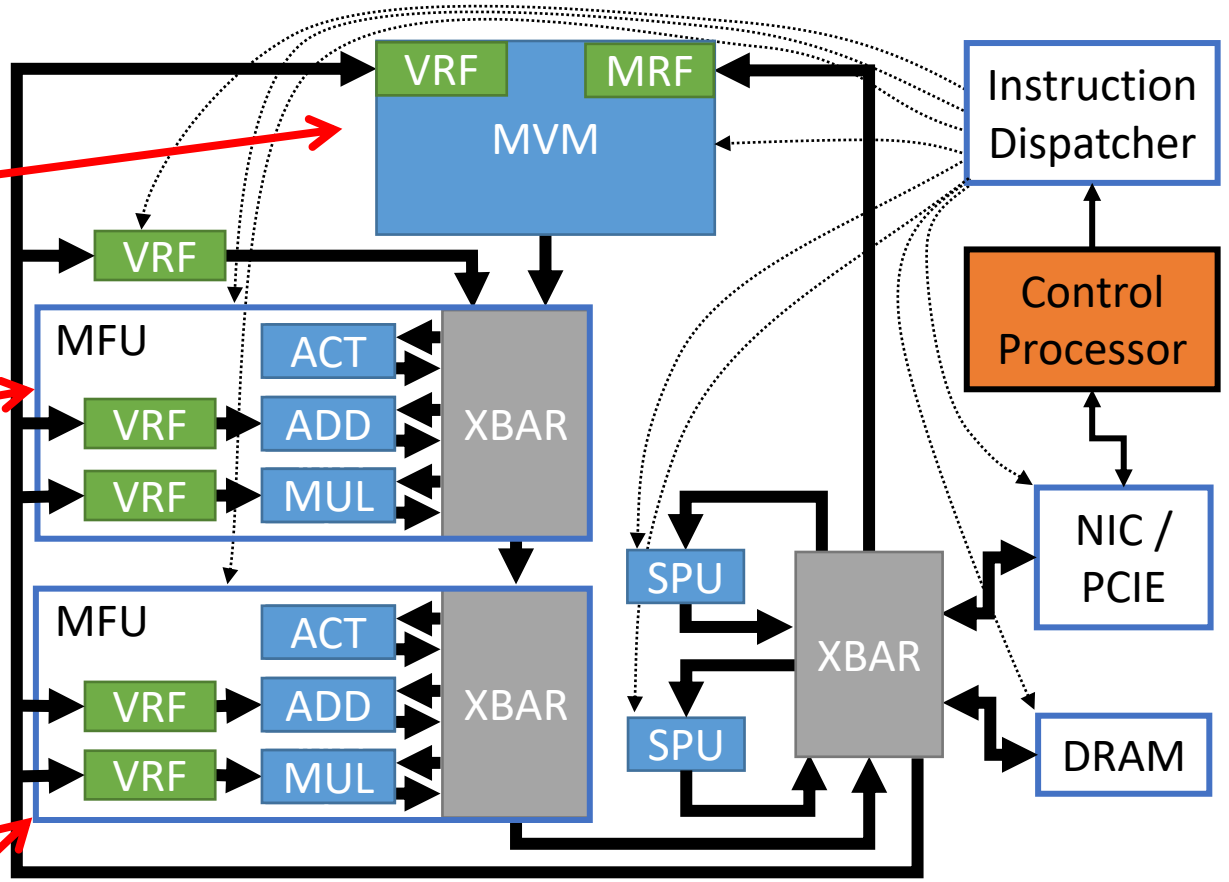
**Operations supported by the Brainwave accelerator**

MVM
Vector add/sub/max
Hadamard product
Sigmoid
TanH
Square root
Inverse

# Mapping LSTM Gates




Ops combining gate outputs



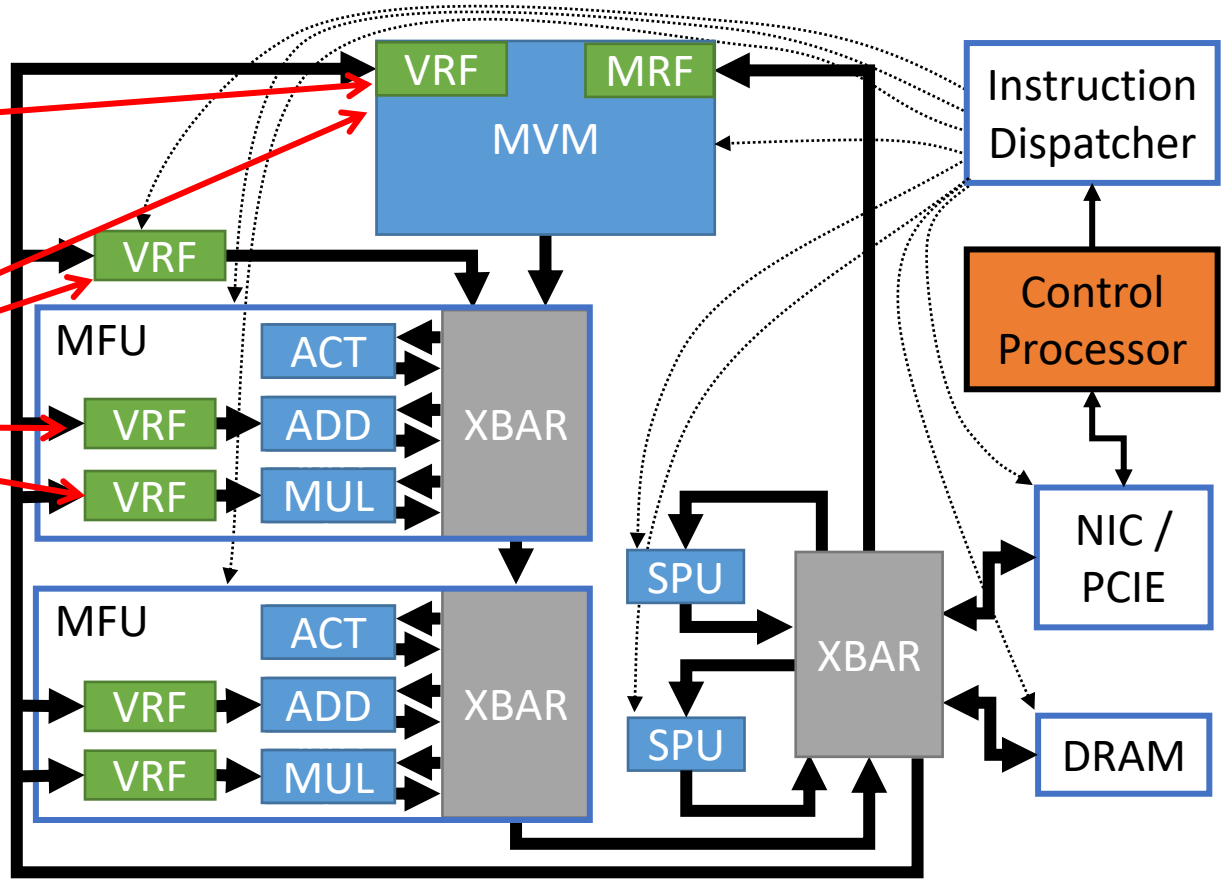

| | | | |
|---|---|---|---|
| VRF | Vector Register File | MFU | Vector Multi-Function Unit |
| MRF | Matrix Register File | ACT | Activation Unit |
| MVM | Matrix Vector Multiply | ADD | Vector Elementwise Add |
| SPU | Special Purpose Unit | MUL | Vector Hadamard Product |

Microsoft

# Specializing the Overlay for the DNN

Configurable parallelism – one large MVM or several parallel MVMs

Vector sizes, element sizes, in-vector parallelization, VRF sizes are configurable

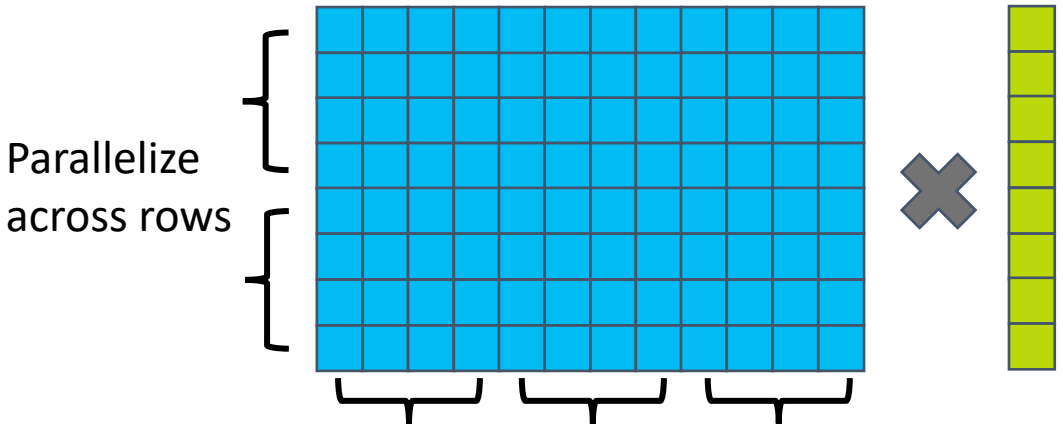



| | | | |
|---|---|---|---|
| VRF | Vector Register File | MFU | Vector Multi-Function Unit |
| MRF | Matrix Register File | ACT | Activation Unit |
| MVM | Matrix Vector Multiply | ADD | Vector Elementwise Add |
| SPU | Special Purpose Unit | MUL | Vector Hadamard Product |

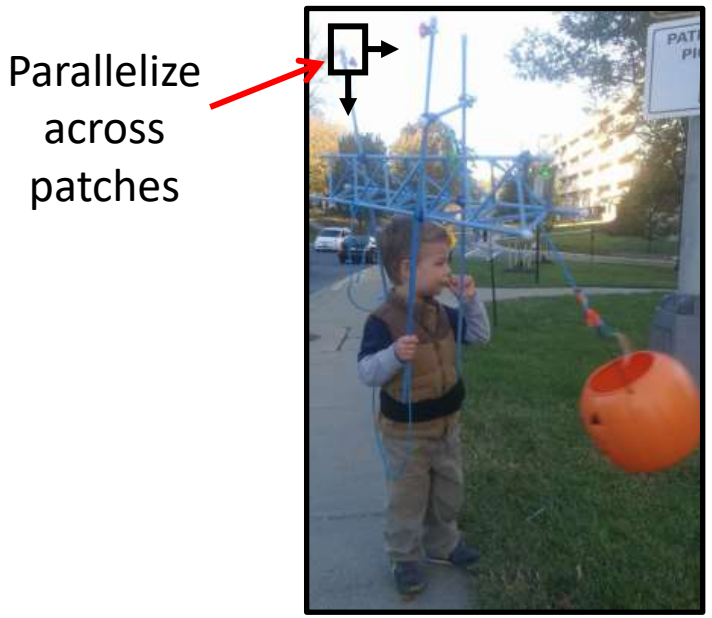
Microsoft

# Specializing the Matrix-Vector Multiplier

## Recurrent network with large matrices




## Convolutional network with many small filter operations



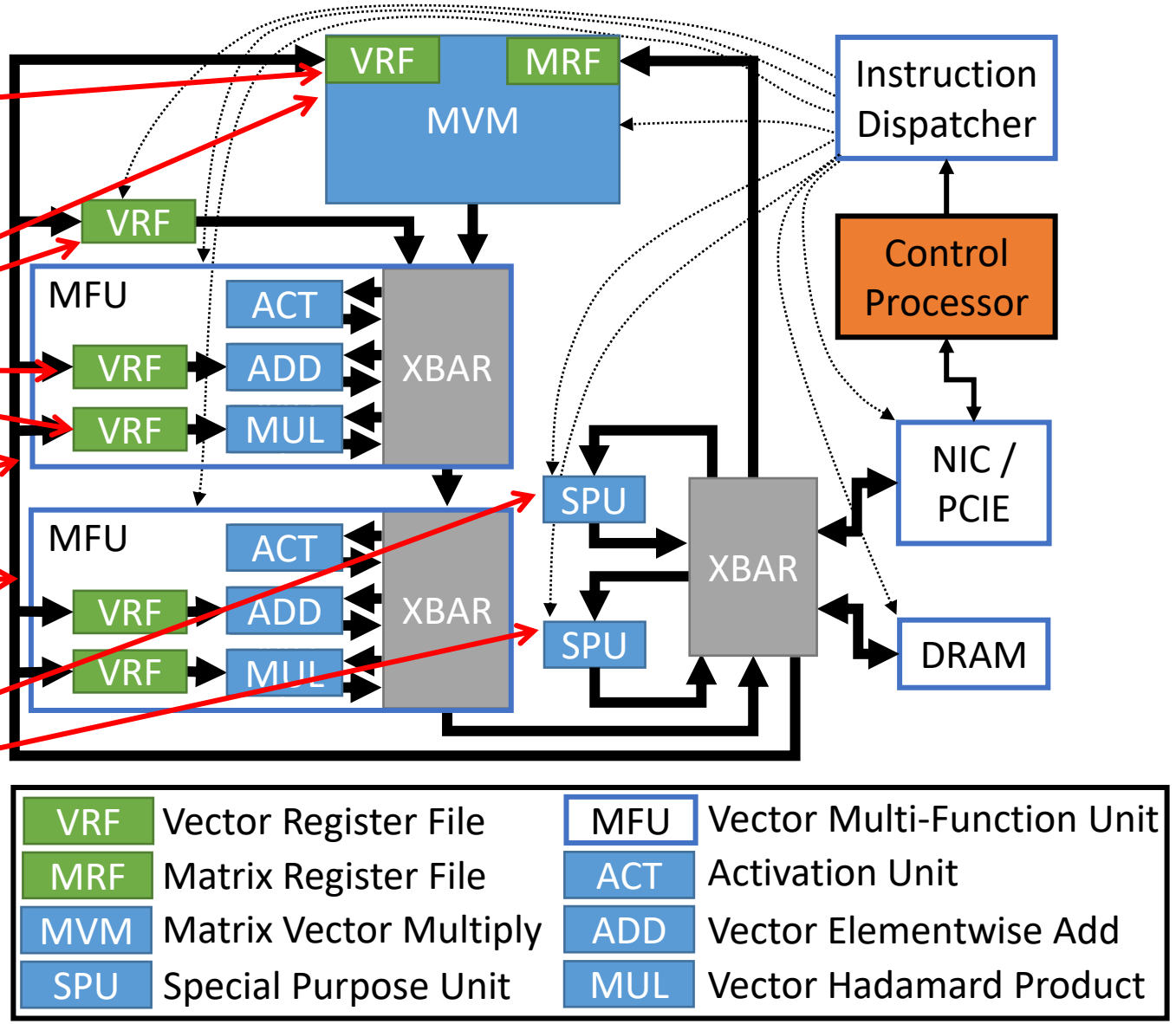
# Specializing the Overlay for the Network

Configurable parallelism – one large MVM or several parallel MVMs

Vector sizes, element sizes, in-vector parallelization, VRF sizes are configurable

# of MFUs and their composition is configurable

# of SPUs and composition is configurable




Microsoft

# Convolutional Neural Networks

- Convolutions:
  - Convolutions slide a window over the image
  - The set of input data at each location is called a “patch”
  - The convolution computes a dot product between each patch and a set of filters.
  - The output of the convolution operation is a 2D array of vectors each containing one element per filter
- Batch normalization reduces the range of the activation values, reducing covariate shift
- Pooling operations reduce the size of the feature maps

# ResNet-152: A Convolutional Neural Network

Input: 224 x 224 image
50k input vectors of 3 elements

Intermediate feature maps range from 112X112 vectors of depth 64 to 7X7 vectors of depth 2048

Output: 1000 floats each corresponding to a score for that category

Won the 2015 ILSVRC challenge and achieved human-level accuracy

151 convolutional layers

60 million model parameters

11 billion FLOPS

ResNet-152 is procedurally generated using blocks of network layers that repeat

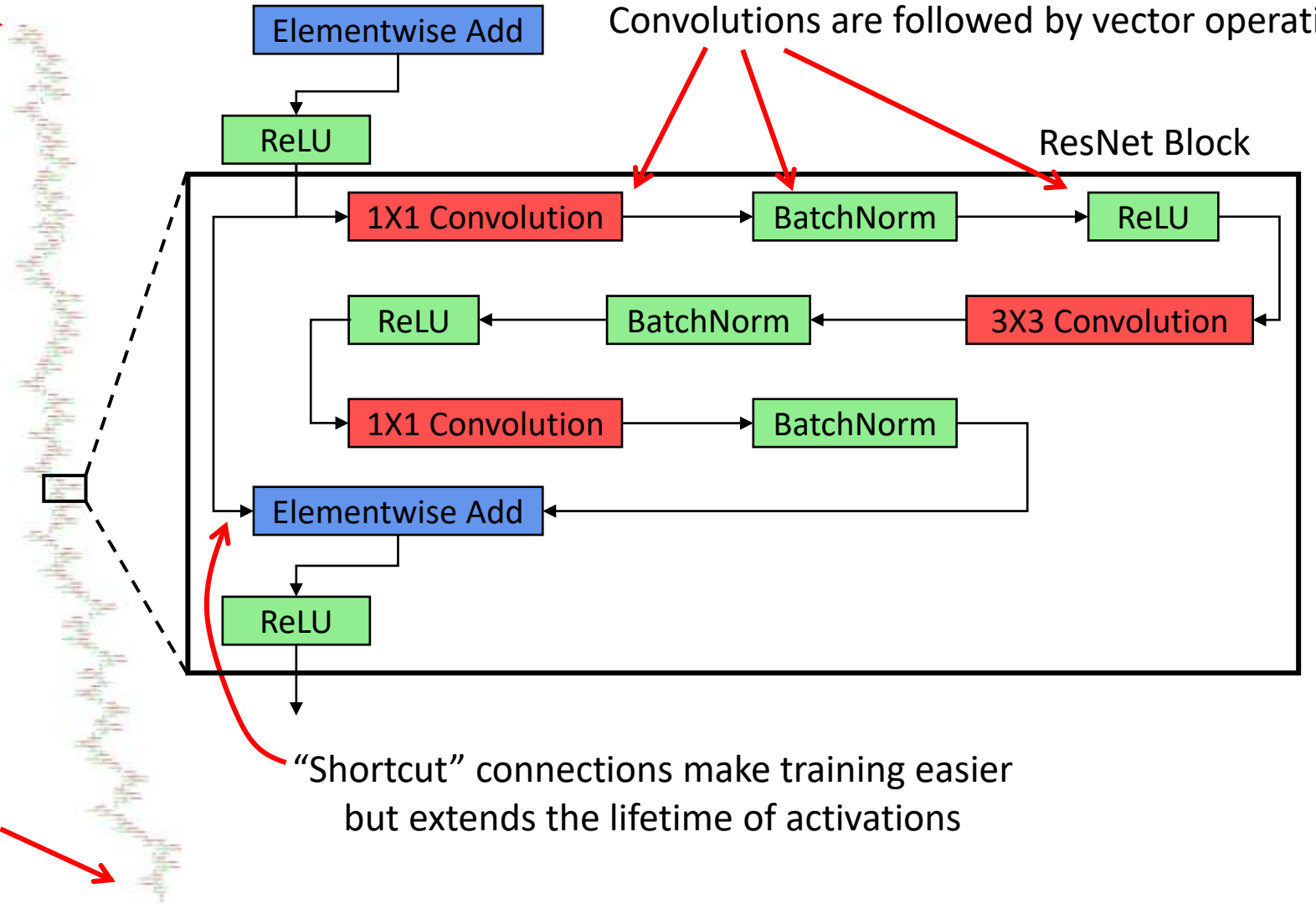
Microsoft

# “Res” = “Residual” Learning with Shortcuts

Input: 224 x 224 image
50k input vectors of 3 elements

Intermediate feature maps range from 112X112 vectors of depth 64 to 7X7 vectors of depth 2048

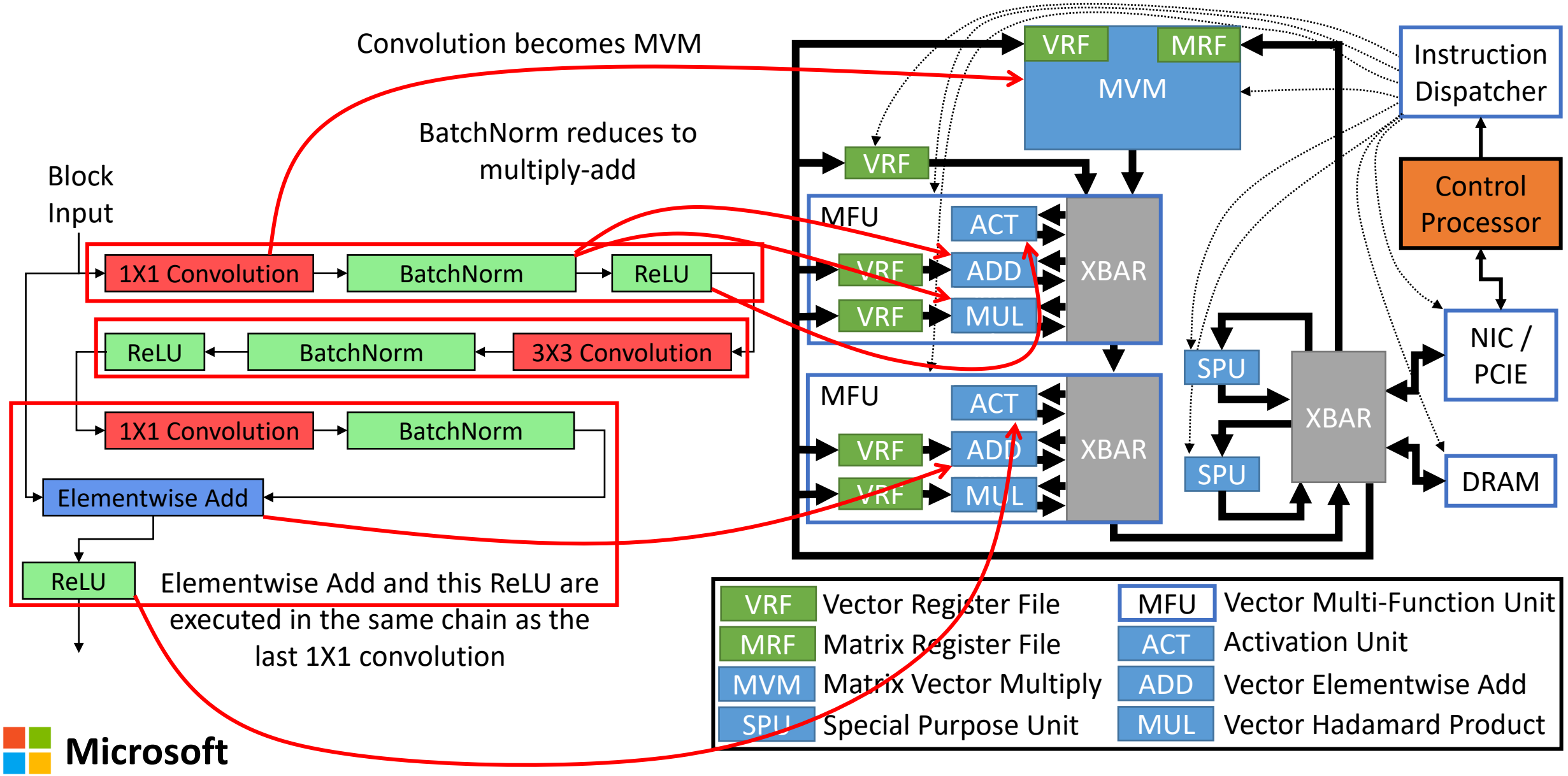
Output: 1000 floats each corresponding to a score for that category




“Shortcut” connections make training easier but extends the lifetime of activations

Microsoft

# Mapping ResNet-152 to Brainwave




Microsoft

# ResNet-152 and ResNet-50 Performance

- All convolution layers run on the FPGA
- Experiments use a batch size of 1 and a single FPGA
- Classifier runs on host computer
- Results are for the layers running on the FPGA and include data transfers
- Results on Arria 10 GX 1150 running at 300 MHZ

| ResNet variant | ResNet-152 | ResNet-50 |
|---|---|---|
| Convolution Layers | 151 | 49 |
| Inference Latency (ms) | 4 | 1.65 |
| Top-1 Accuracy (%) | 75.4 | 73.3 |
| Reference Top-1 (%)* | 77 | 75.3 |
| Top-5 Accuracy (%) | 92.4 | 91.1 |
| Reference Top-5 (%)* | 93.3 | 92.2 |

* [github.com/KaimingHe/deep-residual-networks]

ResNet-152 reference results:

[Ma+ ISCAS 17]: 72 ms on the Arria 10 GX 1150

[Aziz+ HPCA 2019]: 35 ms on the Virtex-7 485T

ResNet-50 reference results:

[Chen+ FPGA 2019]: 8 ms on VU9P

Our experiments: 25% faster than an NVIDIA P40

# FPGA-Accelerated CNNs in Azure

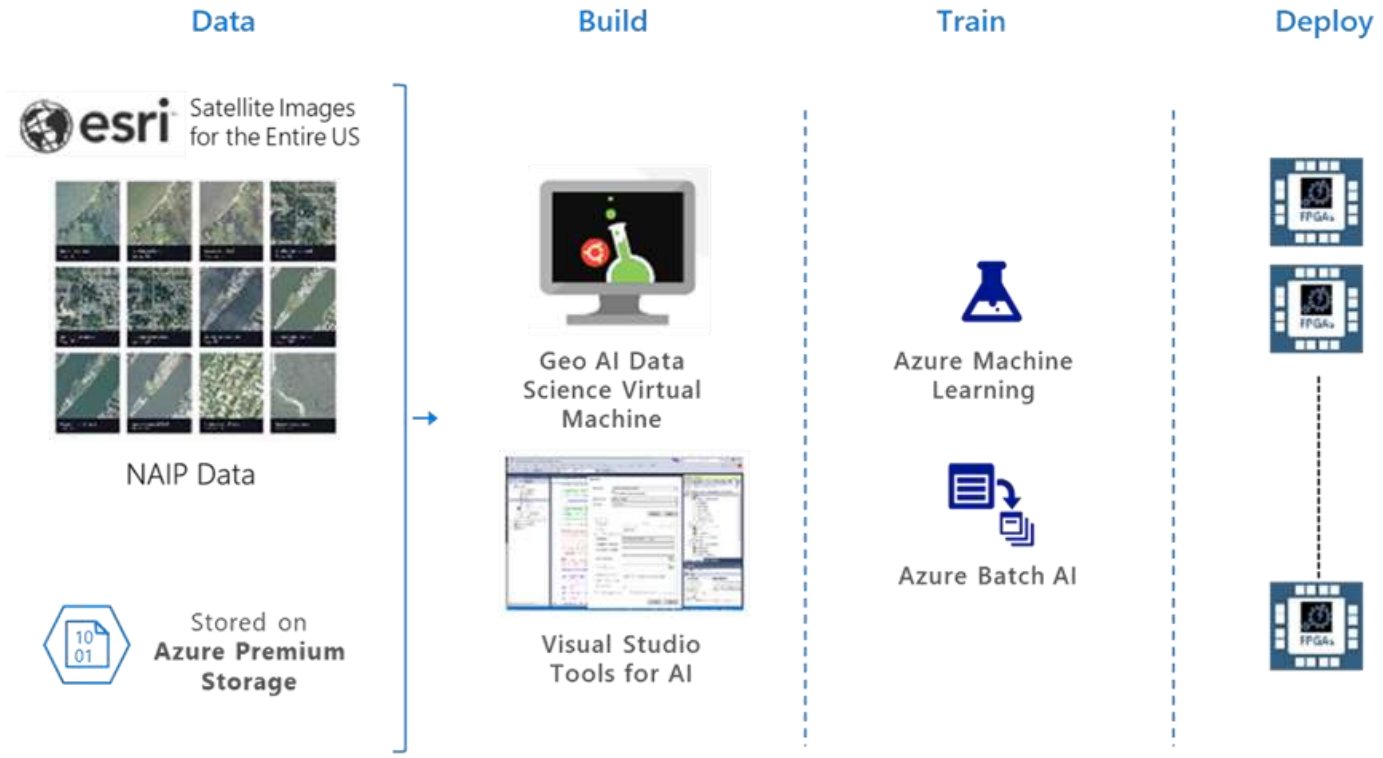
5 well-known convolutional neural networks

- ResNet-152
- ResNet-50
- DenseNet-121
- VGG-16
- SSD-VGG

The system includes an SDK, web-based GUI,

and tutorials [https://aka.ms/aml-real-time-ai]

# Azure-Hosted ResNet-50 Based Land Classification




**200M Images, 20TB**

Land cover mapping for the whole of US in

**10+ minutes**

Created a national land cover map in about 10 minutes using $42 of compute time

[https://blogs.microsoft.com/green/2018/05/23/achievement-unlocked-nearly-200-million-images-into-a-national-land-cover-map-in-about-10-minutes/]

Microsoft

# Azure-Hosted ResNet-50 for Particle Physics

## FPGA-accelerated machine learning inference as a service for particle physics computing

Javier Duarte · Philip Harris · Scott Hauck · Burt Holzman · Shih-Chieh Hsu · Sergo Jindariani · Suffian Khan · Benjamin Kreis · Brian Lee · Mia Liu · Vladimir Lončar · Jennifer Ngadiuba · Kevin Pedro · Brandon Perez · Maurizio Pierini · Dylan Rankin · Nhan Tran · Matthew Trahms · Aristeidis Tsaris · Colin Versteeg · Ted W. Way · Dustin Werran · Zhenbin Wu

Received: - / Accepted: -



**Abstract** Large-scale particle physics experiments face challenging demands for high-throughput computing resources both now and in the future. New heterogeneous computing paradigms on dedicated hardware with increased parallelization, such as Field Programmable Gate Arrays (FPGAs), offer exciting solutions with large potential gains. The growing applications of machine learning algorithms in particle physics

J.D., B.H., S.J., B.K., M.L., K.P., N.T., and A.T. are supported by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy.

[https://arxiv.org/pdf/1904.08986.pdf]

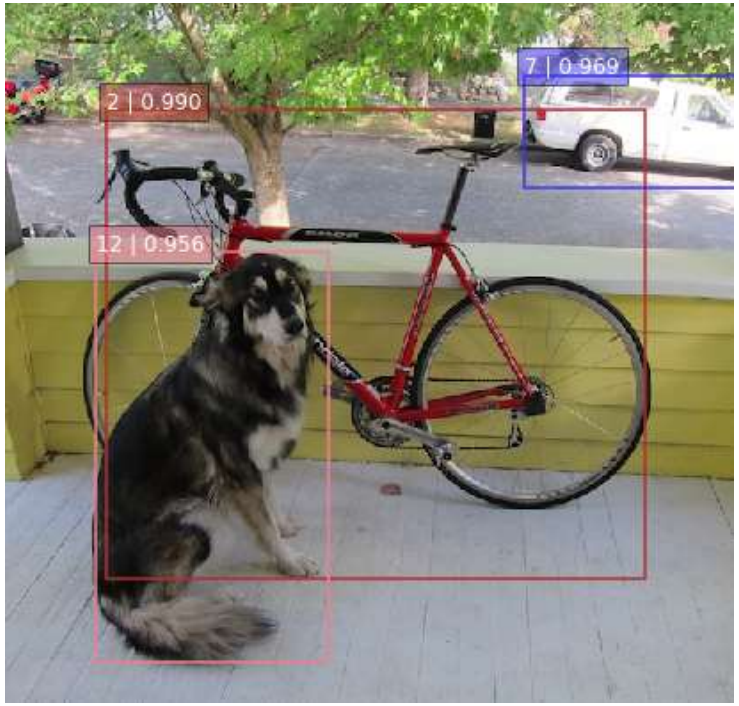
Microsoft

# FPGA-Accelerated CNNs in Azure

5 well-known convolutional neural networks

- ResNet-152
- ResNet-50
- DenseNet-121
- VGG-16
- SSD-VGG

This one localizes objects in the image




The system includes an SDK, web-based GUI, and tutorials [https://aka.ms/aml-real-time-ai]

Microsoft

# SSD-VGG For Empty Shelf Detection at the Edge


KROGER CORPORATE > INVESTOR RELATIONS > PRESS RELEASES > PRESS RELEASE

## Kroger and Microsoft Partner to Redefine the Customer Experience and Introduce Digital Solutions for the Retail Industry

- America's largest grocery retailer and global technology company partnering to pilot two connected experience stores
- Companies will jointly bring to market Retail as a Service product for retailers and present the solution at NRF 2019: Retail's Big Show

Company Release - 1/7/2019 6:30 AM ET

CINCINNATI and REDMOND, Wash., Jan. 7, 2019 /PRNewswire/ — The Kroger Co. (NYSE: KR) and Microsoft Corp. (Nasdaq: MSFT) today announced a collaboration to redefine the customer experience using Kroger Technology products powered by Microsoft Azure, the retailer's preferred cloud platform for Retail as a Service (RaaS). Through this innovative partnership, Kroger will pilot a connected store experience and together with Microsoft, jointly market a commercial RaaS product to the industry.

[http://ir.kroger.com/file/Index?KeyFile=396285733]

Microsoft

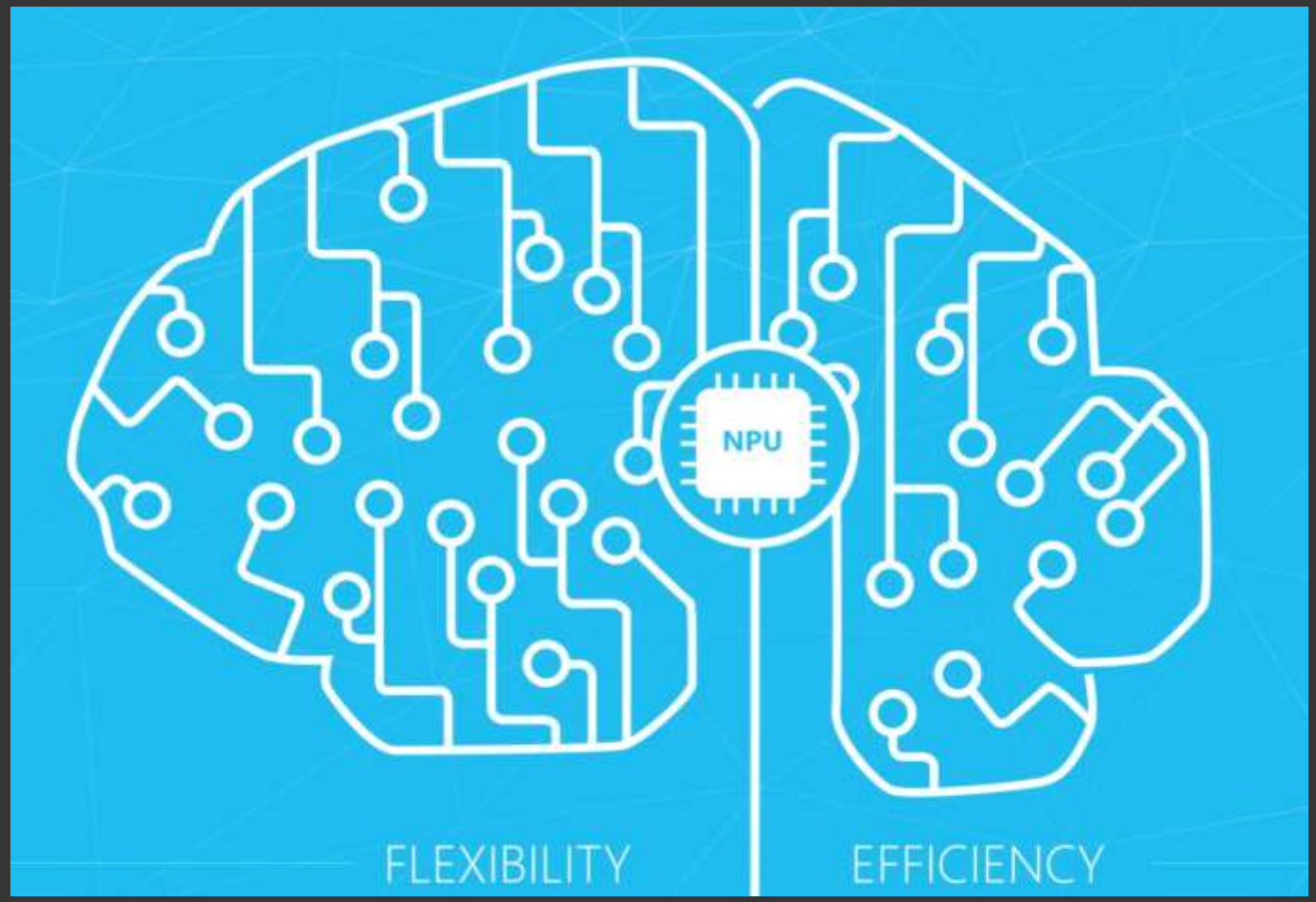
# Takeaways

FPGAs are great for neural networks because we can specialize for the network and update the overlay in place

- Can switch any FPGA to a different configuration for load balancing
- Neural networks keep changing and FPGAs allow us to keep up

Brainwave is co-designed across hardware and software to take advantage of this flexibility to perform neural network inference at a massive scale for 1st party models on Bing and 3rd party models on Azure

Brainwave is still under development and we're scaling it to better FPGA hardware and bigger models




https://www.microsoft.com/en-us/research/project/project-brainwave/
https://www.microsoft.com/en-us/research/project/project-catapult/
https://aka.ms/aml-real-time-ai

Microsoft