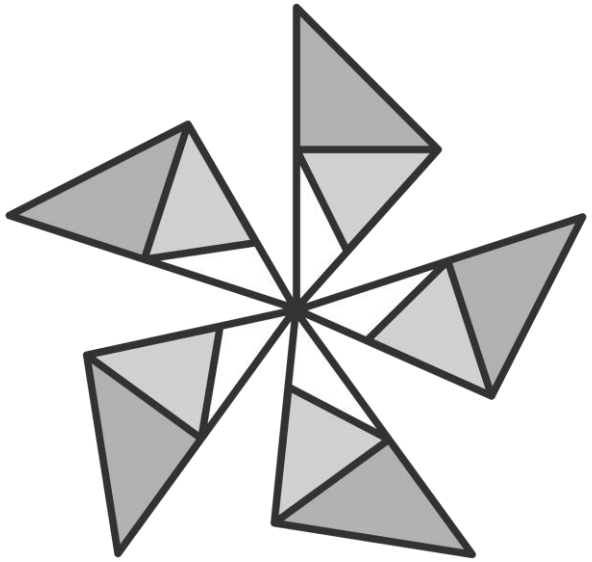


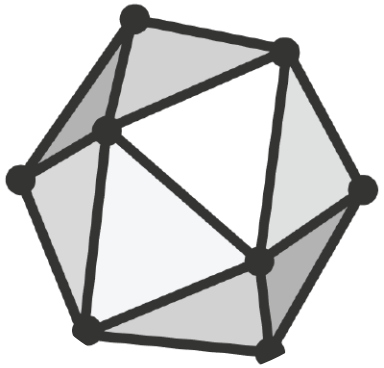
Open Neural Network eXchange (ONNX):

Accelerate and operationalize models for deployment

VINITRA SWAMY| Microsoft

CECILIA LIU| Microsoft

WiDS 2020



ONNX

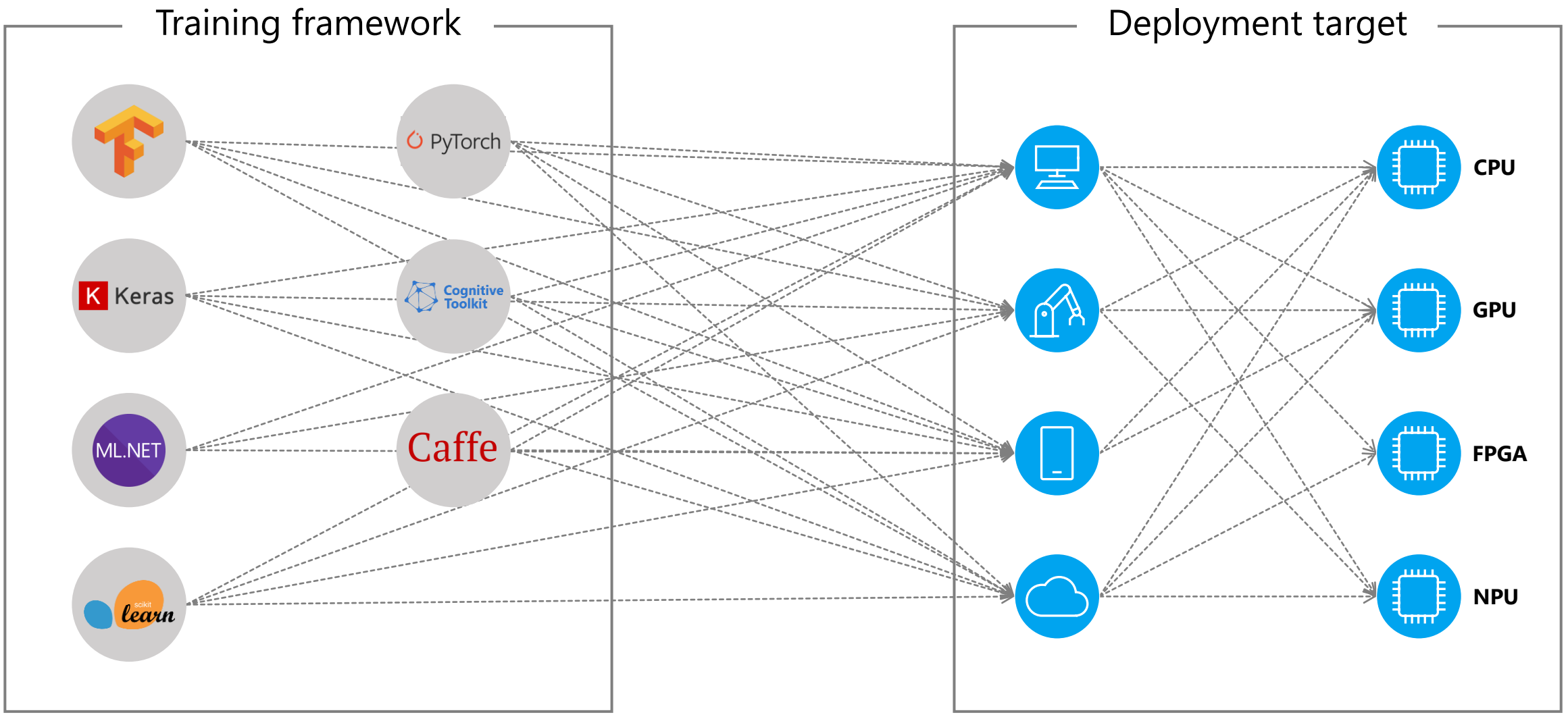
github.com/onnx

onnx.ai

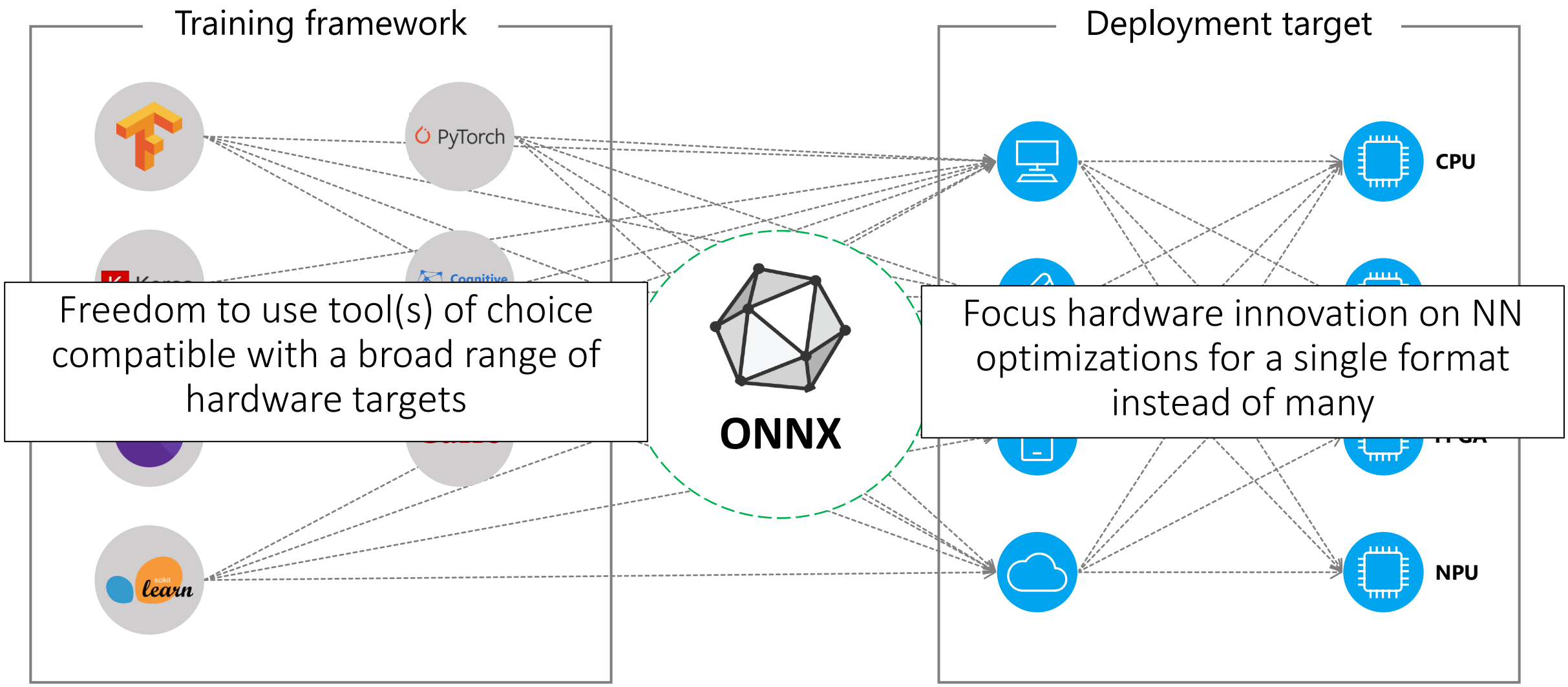
OPEN **N**EURAL **N**ETWORK E**X**CHANGE

ONNX Community

Training frameworks **x** Deployment targets

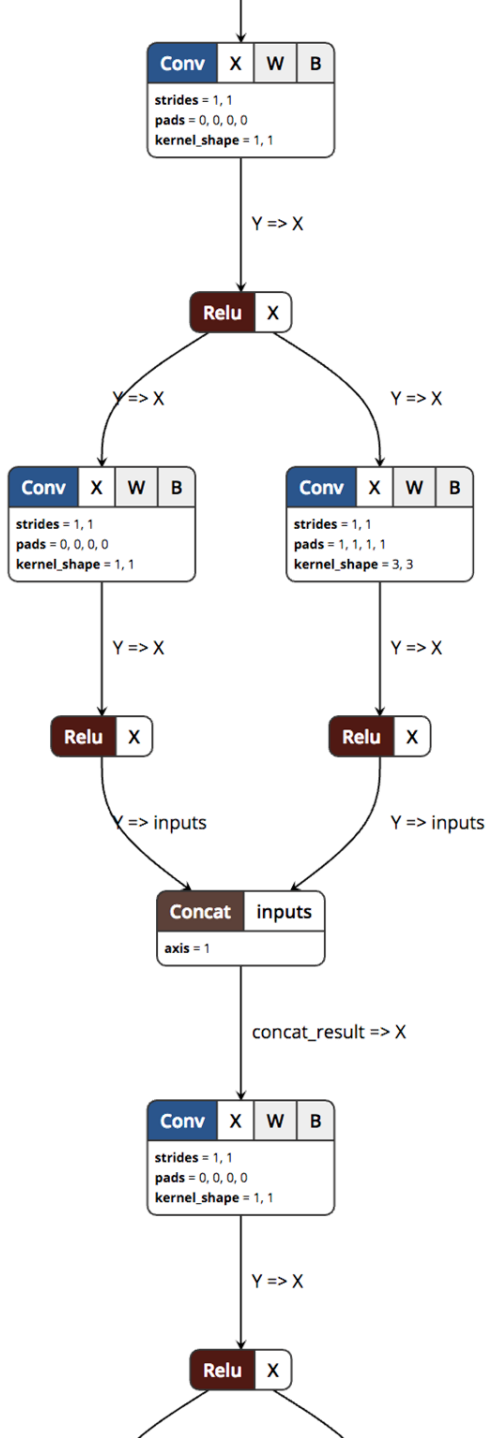


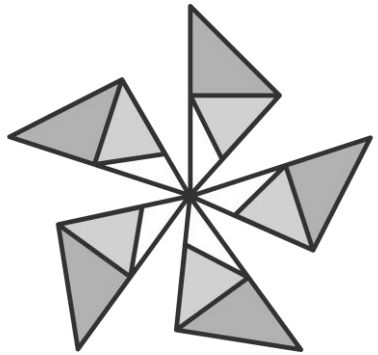
ONNX: an interoperable format for ML models



ONNX

- Standard format for ML models consisting of:
  - common Intermediate Representation
  - full operator spec
- Model = graph composed of computational nodes
- Supports both DNN and traditional ML
- Backward compatible with comprehensive versioning





ONNX Runtime

aka.ms/onnxruntime

github.com/microsoft/onnxruntime

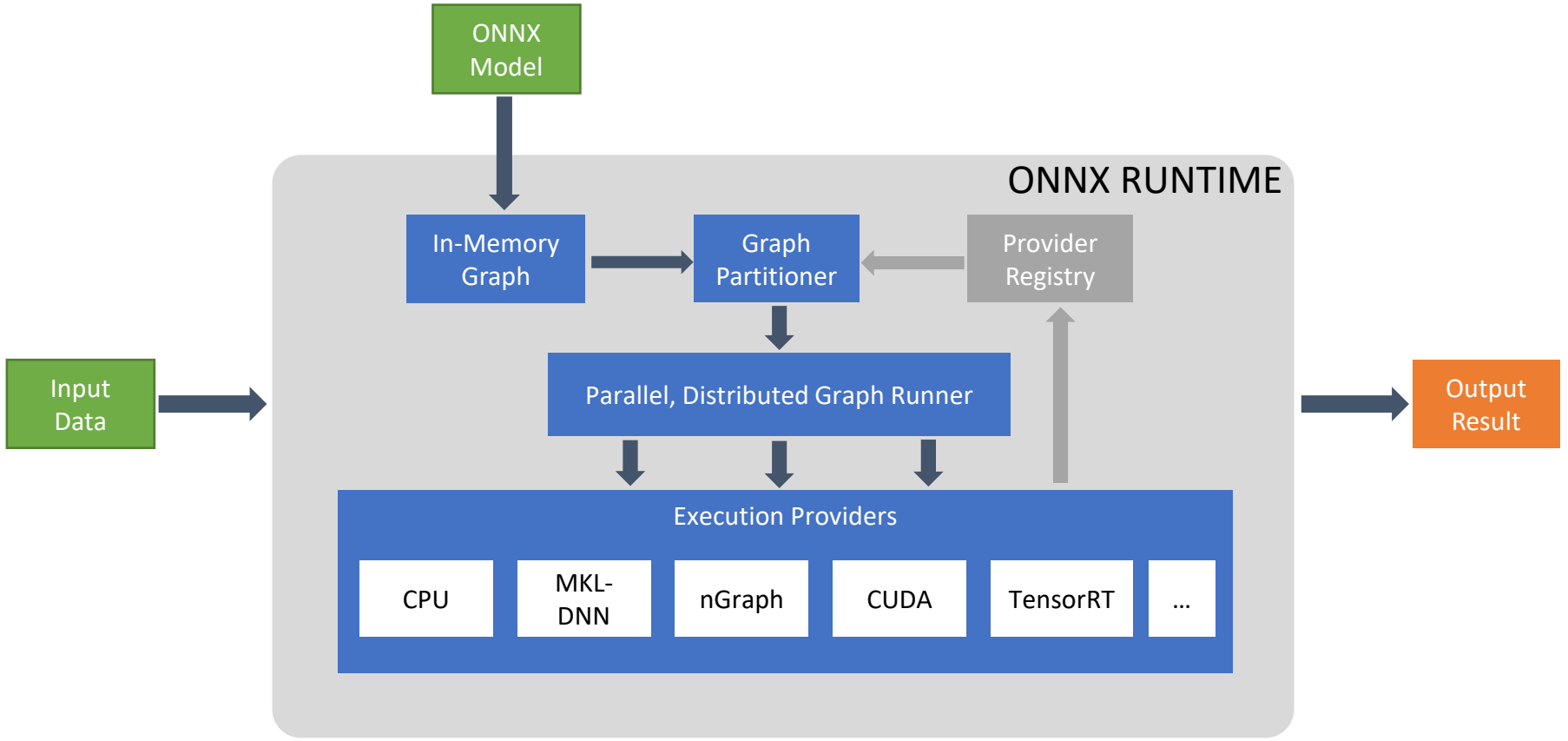
ONNX Runtime: open source high performance Inference Engine for ONNX models

- Performance focused design
- Full ONNX operator support
- Flexibility for custom operators not in the spec
- Backwards and forwards compatible to minimize versioning issues with software or model updates

Cross platform, multi language API

OS	Windows	Linux	Mac		
Language	Python (3.5-3.7)	C++	C#	C	Java
Architecture	X64	X86	ARM64	ARM32	
Hardware Acceleration	Default CPU	CUDA	TensorRT	DirectML	MKL-DNN
	MKL-ML	nGraph	NUPHAR	OpenVINO	
Installation Instructions	pip install onnxruntime				

Leverages and abstracts hardware accelerators



ONNX @ Microsoft

PLATFORMS


AzureML


WinML

ML.Net

PRODUCTS

CPU GPU Cloud Edge

Up to 18.7x

Performance gains seen by Microsoft services

100s of Millions

# of devices where ONNX Runtime is running

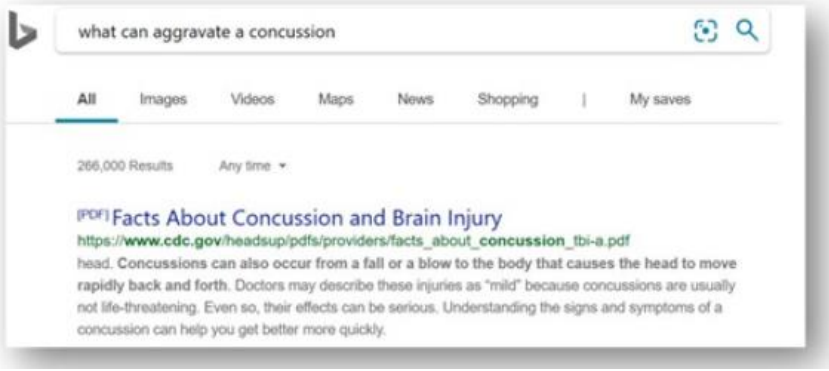
Billions

# of requests handled by ONNX Runtime across Microsoft services

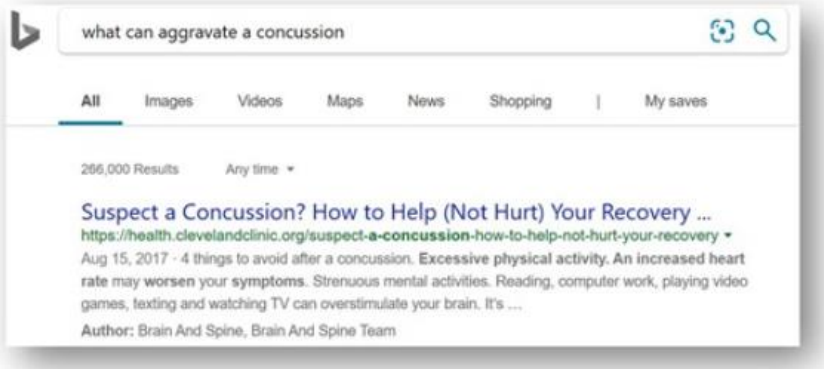
Bing With ONNX Runtime

Apply **BERT model** to every Bing search query globally making Bing results more relevant and intelligent

BEFORE



AFTER



Inference 3-layer BERT with 128 sequence length with ONNX Runtime

- On CPU, 17x latency speed up with ~100 queries per second throughput.
- On NVIDIA GPUs, more than 3x latency speed up with ~10,000 queries per second throughput on batch size of 64

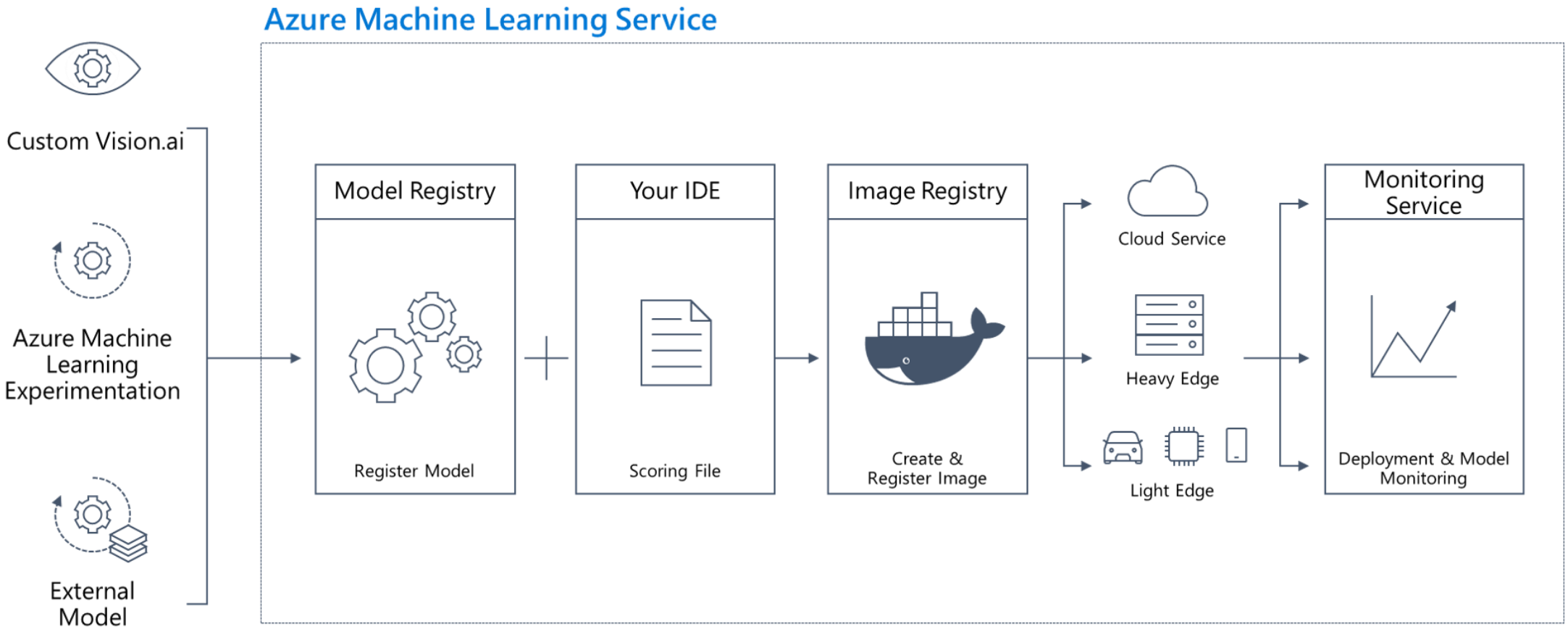
		Batch size	Inference on	Throughput (Query per second)	Latency (milliseconds)
CPU	Original 3-layer BERT	1	Azure Standard F16s_v2 (CPU)	6	157
	ONNX Model	1	Azure Standard F16s_v2 (CPU) **with ONNX Runtime**	111	9
GPU	Original 3-layer BERT	4	Azure NV6 GPU VM	200	20
	ONNX Model	4	Azure NV6 GPU VM **with ONNX Runtime**	500	8
	ONNX Model	64	Azure NC6S_v3 GPU VM **with ONNX Runtime + System Optimization** (Tensor Core with mixed precision, Same Accuracy)	10667	6

DEMO

https://aka.ms/wids2020

Accelerators for a range of hardware

Deploying ONNX Runtime to Azure ML



https://aka.ms/wids2020

Resources

- ONNX: https://github.com/onnx/onnx
- ONNX Converters: https://github.com/onnx/onnxmltools/tree/master/onnxmltools
- ONNX Tutorials: https://github.com/onnx/tutorials
- ONNX Runtime: https://github.com/microsoft/onnxruntime
- ONNX Runtime Tutorials: https://github.com/microsoft/onnxruntime#examples-and-tutorials
- Performance Tuning with ONNX Runtime: https://github.com/microsoft/onnxruntime/blob/master/docs/ONNX_Runtime_Perf_Tuning.md
- Training, Inferencing, and deployment in AzureML with ONNX models: https://aka.ms/onnxnotebooks
- Deploying to Edge and IoT devices: https://github.com/Azure-Samples/onnxruntime-iot-edge
- Windows ML: https://docs.microsoft.com/en-us/windows/ai/windows-ml/

VINITRA SWAMY |
viswamy@microsoft.com

CECILIA LIU |
ziyue.liu@microsoft.com