Interoperable AI:

ONNX & ONNXRuntime in C++

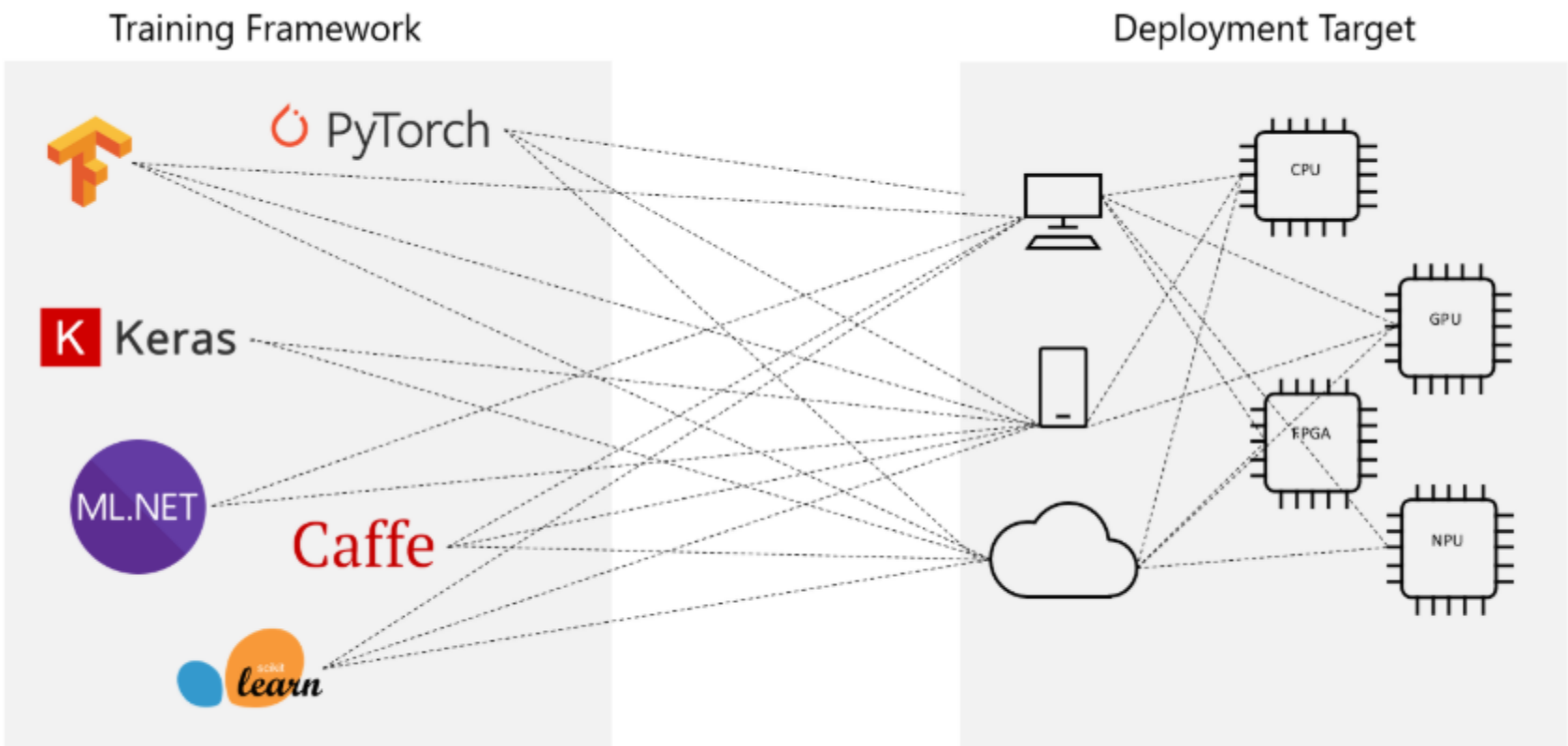
Marco Arena

Mattia Verasani

C++ DAY

2020

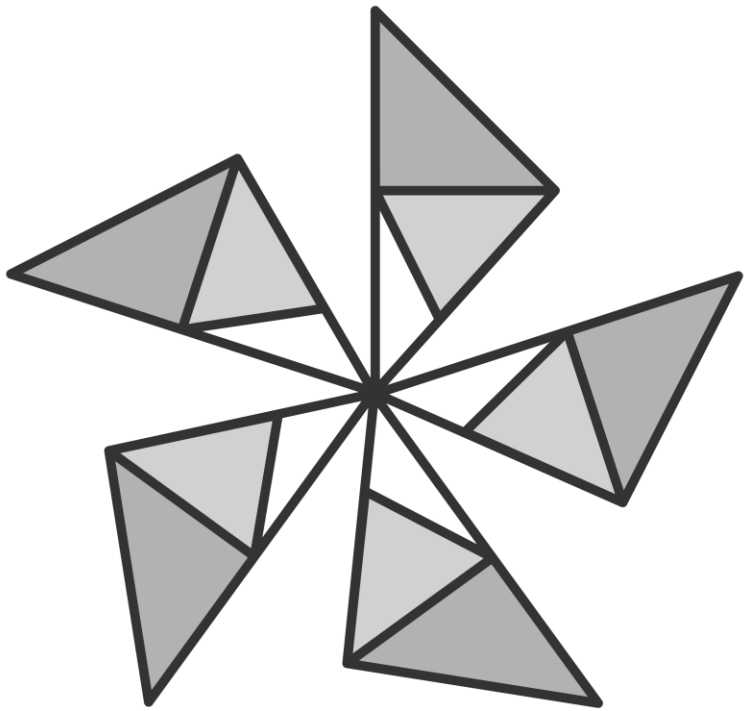
++it Italian C++ Community



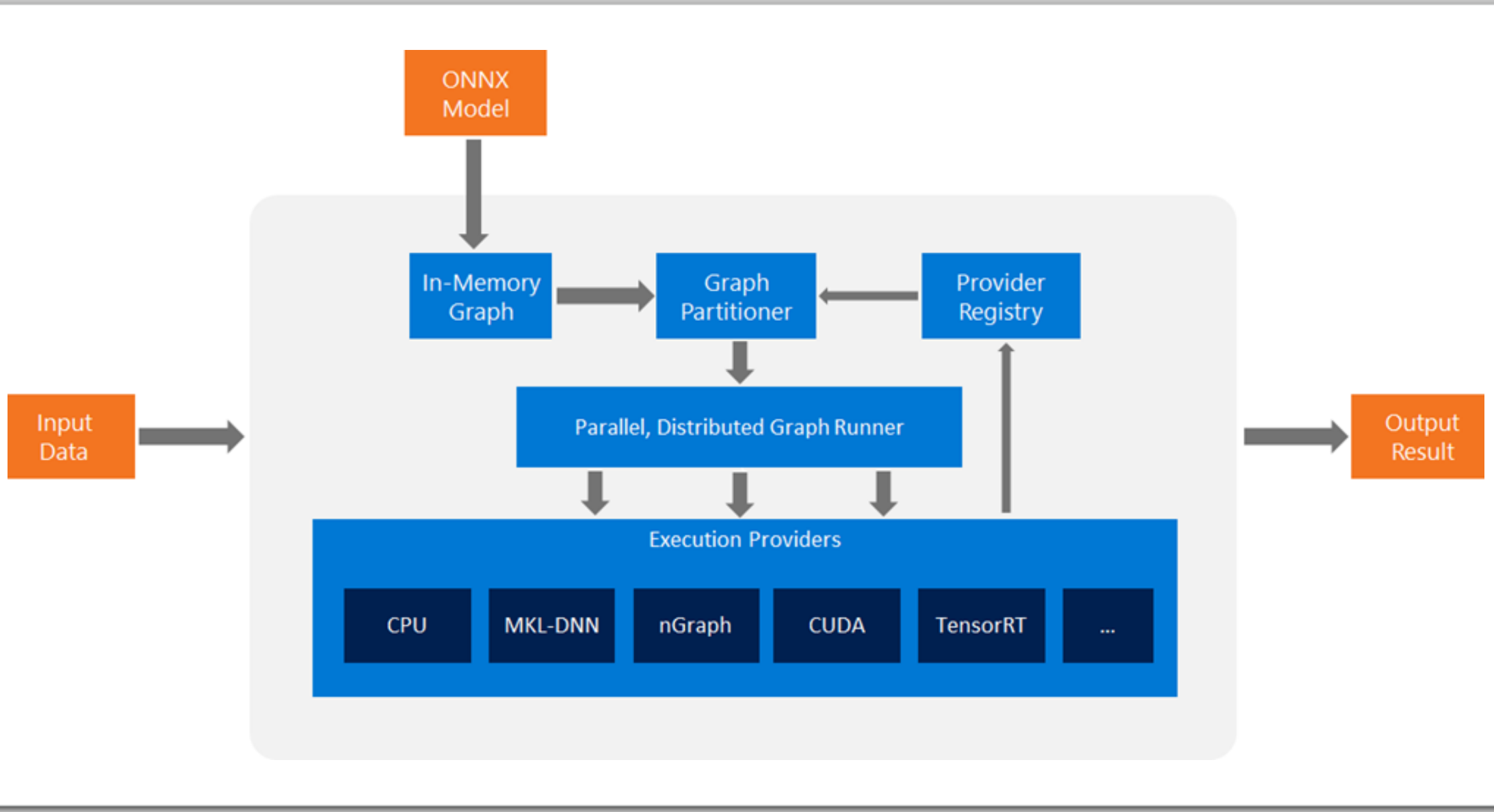
Plethora of possible training and deployment combinations

ONNX provides interoperability between frameworks

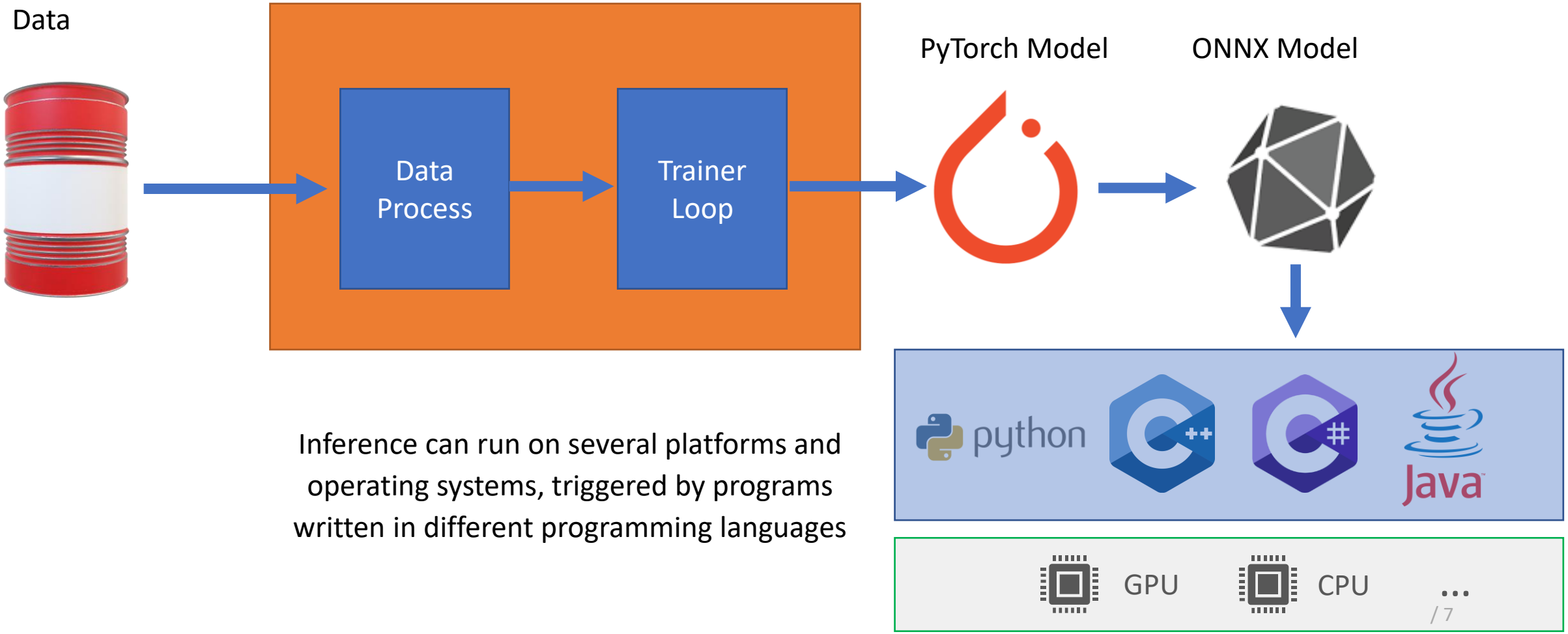

ABBYY
Alibaba Group
阿里巴巴集团
AMD
arm
aws
Baidu 百度
BECKHOFF
BITMAIN
cadence
CEVA
Facebook
Open Source
GRAPHCORE
habana
Hewlett Packard
Enterprise
HUAWEI
IBM
Idein Inc
intel AI
MathWorks
MAXAR
MEDIATEK
Microsoft
NVIDIA
NXP
OctoML
OPEN AI LAB
开 放 智 能
Preferred
Networks
SIEMENS
SONY
Qualcomm
sas
商汤
sensetime
skymizer
SYNOPSYS
Tencent
unity
verizon
media
WOLFRAM
Yandex



ONNX RUNTIME

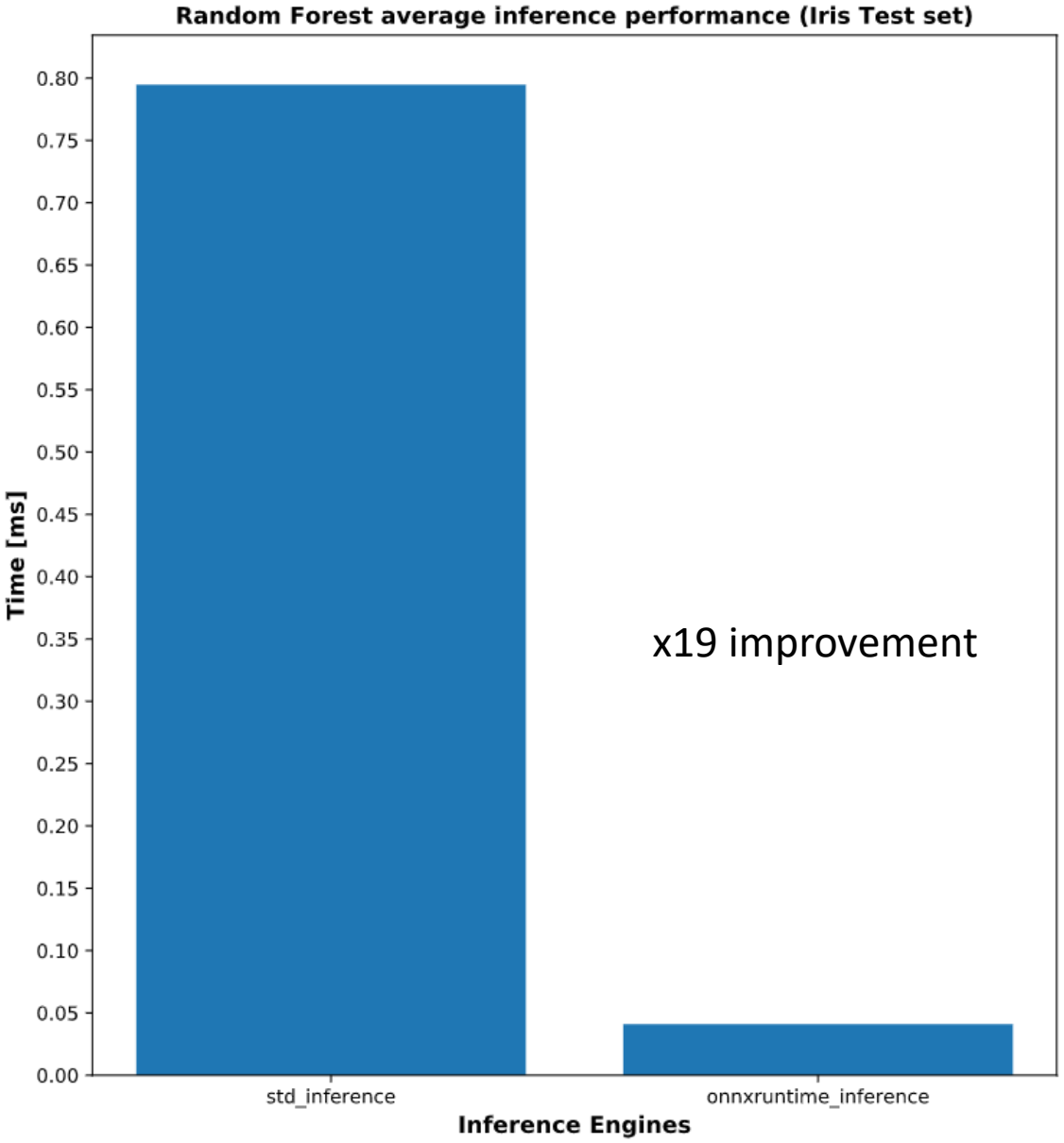

ONNX Model
In-Memory Graph
Graph Partitioner
Provider Registry
Input Data
Parallel, Distributed Graph Runner
Output Result
Execution Providers
CPU
MKL-DNN
nGraph
CUDA
TensorRT
...

A PyTorch environment example

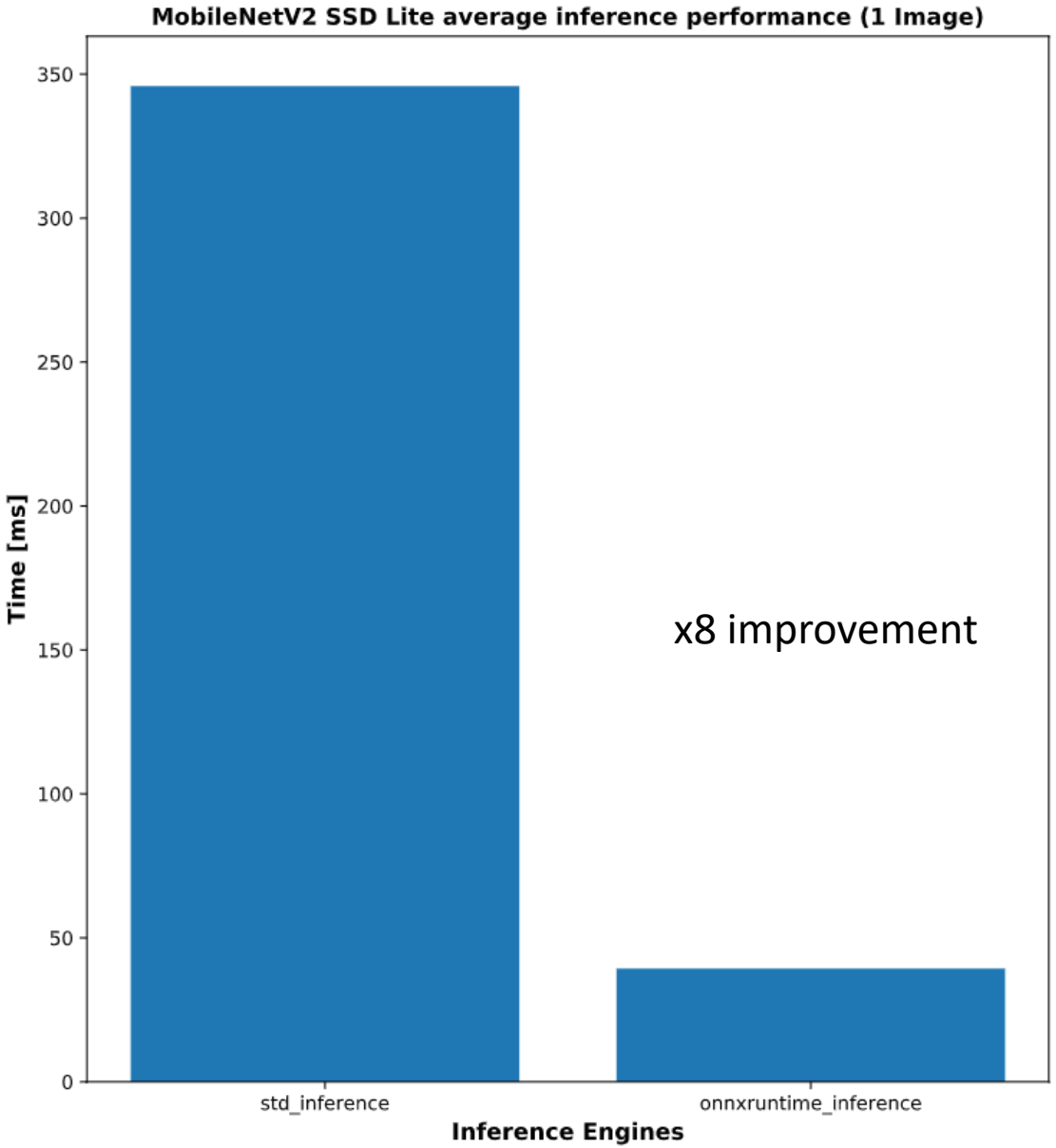


Inference can run on several platforms and operating systems, triggered by programs written in different programming languages

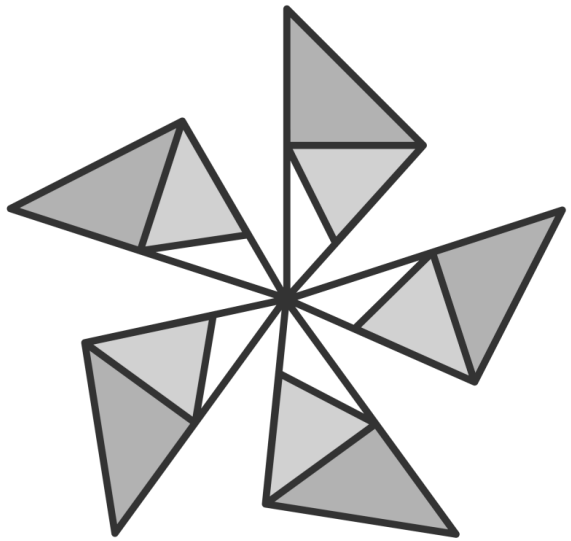
ONNX Runtime improves inference performance



x19 improvement



x8 improvement



ONNX RUNTIME

Inference Session

Model inspection

Creating Tensors (CPU)

Running inference
(partial outputs supported)

Custom Loggers

Execution Providers

Profiling support

Graph Optimizations

Custom Operators

Custom Allocators

Threading settings

Hands-on!

https://github.com/ilpropheta/onnxruntime-demo

https://github.com/MatRazor/ONNXRuntime_tutorial_collection

A Systematic Assessment of Embedded Neural Networks for Object Detection

Micaela Verucchi*, Gianluca Brilli*, Davide Sapienza*, Mattia Verasani†, Marco Arena†,
Francesco Gatti*‡, Alessandro Capotondi*, Roberto Cavicchioli*, Marko Bertogna* and Marco Solieri*
*Università di Modena e Reggio Emilia, Italy - name.surname@unimore.it
‡Consorzio Nazionale Interuniversitario per le Telecomunicazioni (CNIT), Italy - 189382@studenti.unimore.it
†Tetra Pak, Italy - name.surname@tetrapak.com

***Abstract*—Object detection is arguably one of the most important and complex tasks to enable the advent of next-generation autonomous systems. Recent advancements in deep learning techniques allowed a significant improvement in detection accuracy and latency of modern neural networks, allowing their adoption in automotive, avionics and industrial embedded systems, where performances are required to meet size, weight and power constraints.**

detect objects to be manipulated, or defects to be signalled. Many are the other applicative domains, ranging from robotics, to avionics or simply surveillance. Many also are the more complex vision tasks that can be built upon object detection, such as instance segmentation, image captioning, or object tracking.

Resources

- Graph Optimizations
- Performance Tuning (and post about chrome://tracing)
- Execution Providers
- Custom Execution Provider
- Custom Operators
- C API sum-up
- OnnxRuntime for mobile
- OnnxRuntime Server (same idea as *tfserving*)
- OnnxRuntime high level design