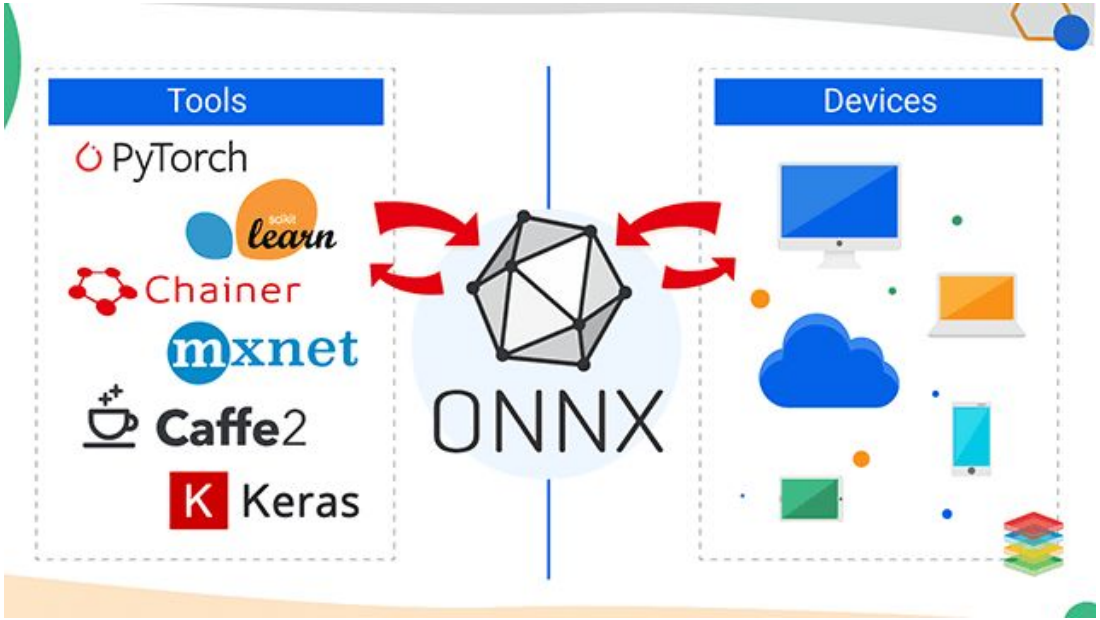
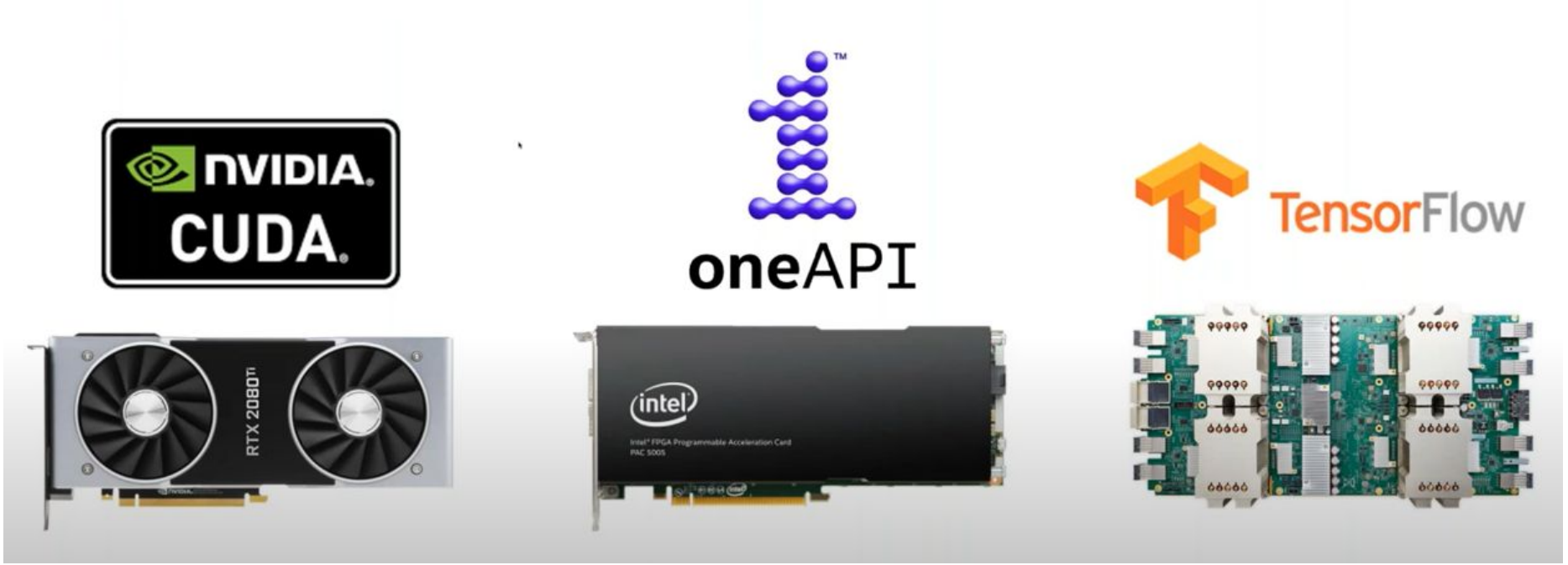
Onnx


Tools
PyTorch
learn
Chainer
mxnet
Caffe2
Keras
ONNX
Devices

1 - Challenge with Deep Learning

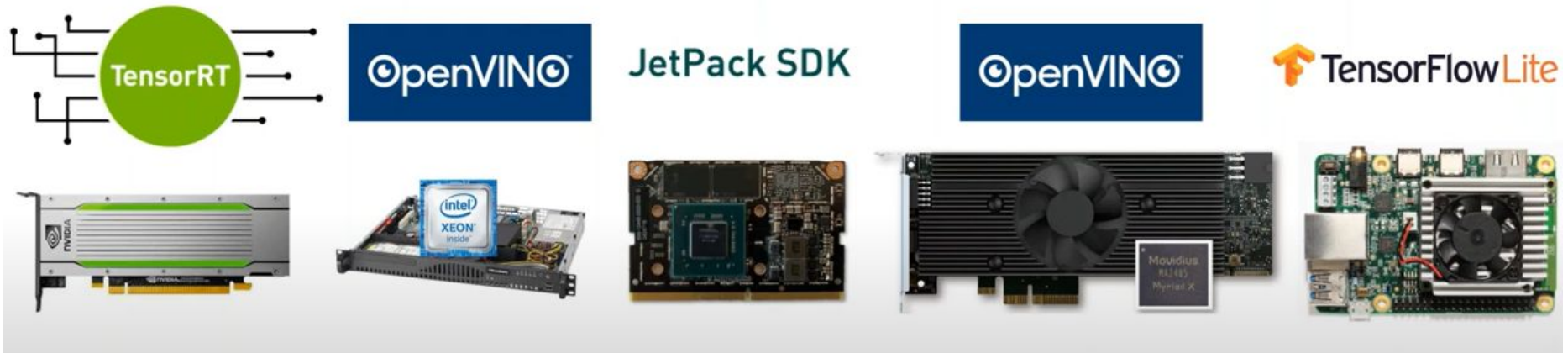
Cần một "cầu nối" để các framework có thể giao tiếp với nhau

1 - Challenge with Deep Learning



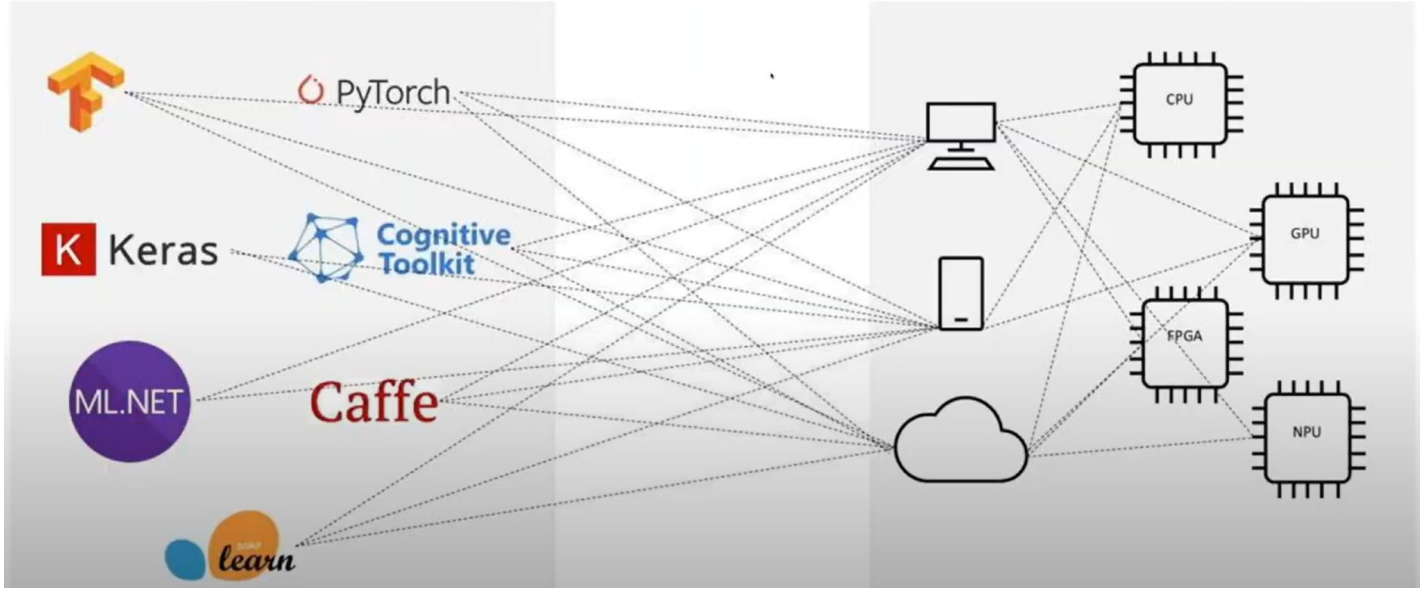
Giao tiếp giữa phần cứng và framework

1 - Challenge with Deep Learning



Sự ổn định giữa framework trên nhiều thiết bị khác nhau

1 - Challenge with Deep Learning



Không chỉ trên các thiết bị nhỏ, mà còn là sự tương tác với cloud, windows device, OS.

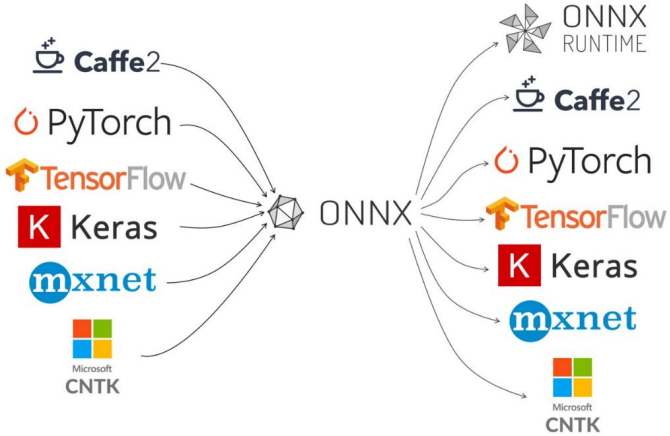
2 - Open Neural Network Exchange

ONNX là một định dạng open standard cho việc diễn giải và trao đổi mô hình neural network giữa các nền tảng và framework khác nhau. Điều này giúp các nhà phát triển và nghiên cứu có thể dễ dàng chia sẻ và sử dụng các mô hình neural network trên nhiều nền tảng khác nhau mà không cần phải chuyển đổi lại từng mô hình.

2 - Open Neural Network Exchange

Một số đặc điểm của ONNX :

- **Khả năng tương thích nhiều nền tảng**: ONNX cho phép build và train model trên một framework như PyTorch, TensorFlow, hoặc Caffe, sau đó chuyển đổi model này sang ONNX để sử dụng trên các framework khác mà không cần training lại.



2 - Open Neural Network Exchange

Một số đặc điểm của ONNX :

- **Khả năng mở rộng**: ONNX hỗ trợ nhiều loại model khác nhau bao gồm deep neural networks cũng như các machine learning model khác như linear regression, logistic regression…

2 - Open Neural Network Exchange

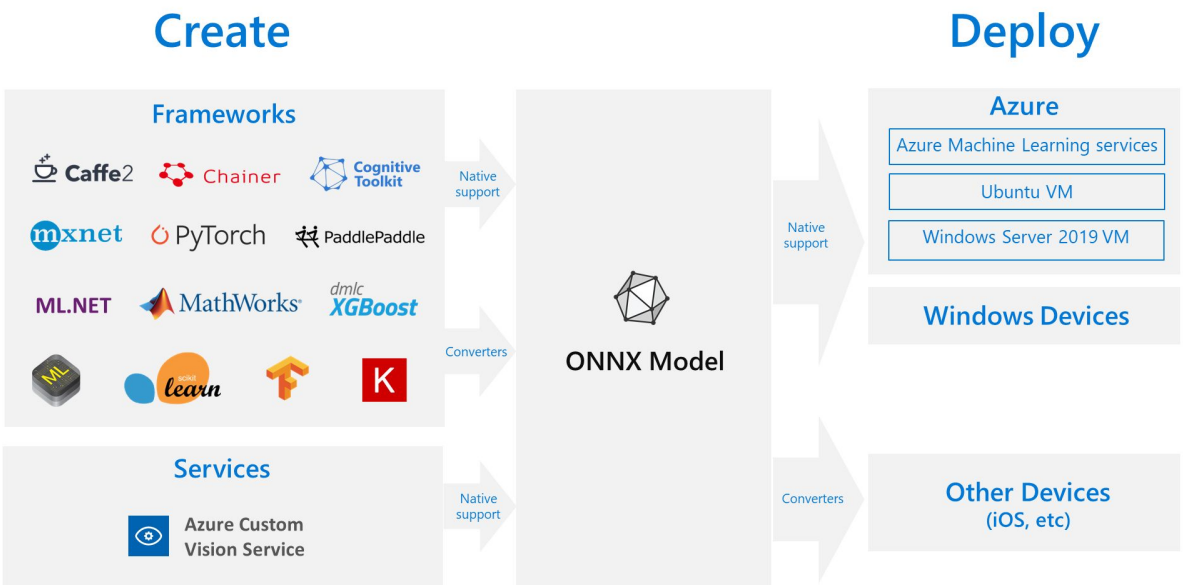
Một số đặc điểm của ONNX :

- **Cộng đồng**: Cộng đồng phát triển mạnh, đảm bảo luôn cập nhật xu hướng nghiên cứu và phát triển AI

- **Dễ triển khai**: ONNX giúp đơn giản hóa việc triển khai mô hình trên các nền tảng như di động, thiết bị nhúng hay edge computing, nên có thể chuyển đổi mô hình ONNX sang các dạng khác nhau như TensorFlow Lite, Core ML của Apple, hoặc các dạng tương tự.

- **Tích hợp với nhiều tool**: Có nhiều công cụ hỗ trợ chuyển đổi, kiểm tra và tối ưu hóa mô hình ONNX, giúp dễ dàng kiểm tra tính đúng đắn và hiệu suất của mô hình.

2 - Open Neural Network Exchange



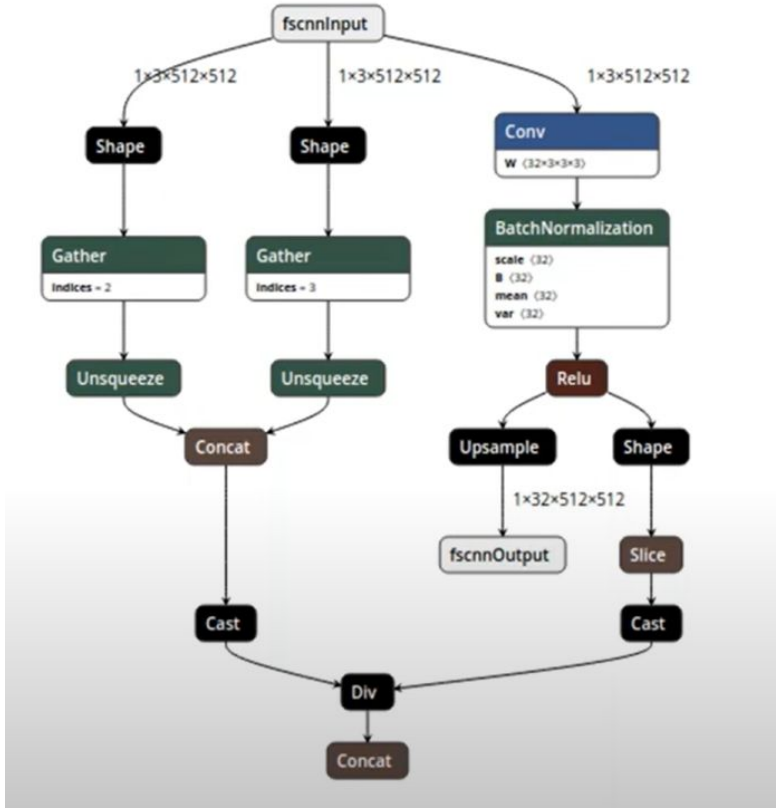
Onnx hỗ trợ sự giao tiếp giữa các framework với nhiều nơi triển khai (model/service,...) khác nhau

3 - Onnx Ecosystem

Hệ sinh thái của Onnx

4 - Onnx File Format

1. Model
   - Version info
   - Metadata
   - Acyclic computation data flow graph
2. Graph
   - Inputs & outputs
   - List of computation nodes
   - Graph name
3. Computation node
   - Zero or more inputs of defined types
   - One or more outputs of defined types
   - Operator
   - Operator parameters



5 - Onnx Data types

1. Tensor type
   - int8, int16, int32, int64
   - uint8, uint16, uint32, uint64
   - float16, float, double
   - bool
   - string
2. Non-tensor types in Onnx-ML
   - Sequence
   - Map

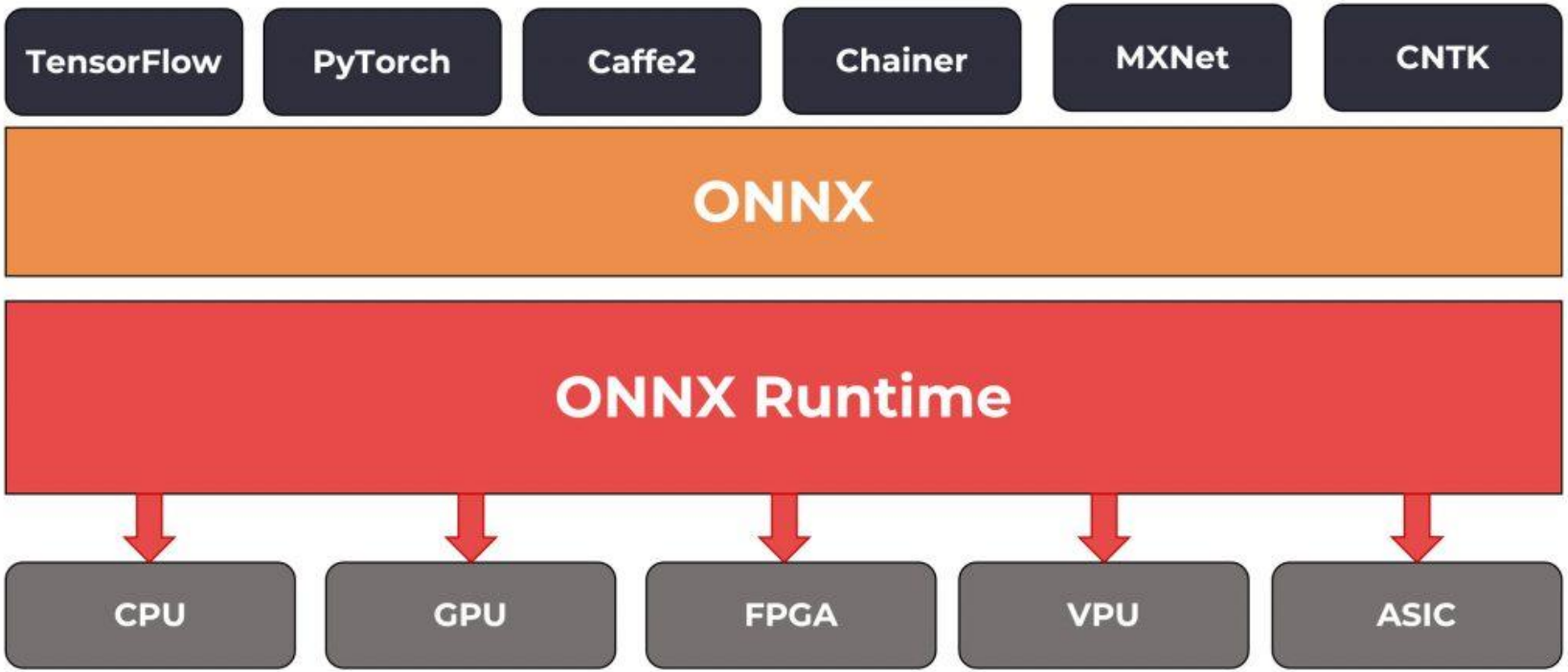
6 - Onnx operators

1. Tensor type
   - int8, int16, int32, int64
   - uint8, uint16, uint32, uint64
   - float16, float, double
   - bool
   - string
2. Non-tensor types in Onnx-ML
   - Sequence
   - Map

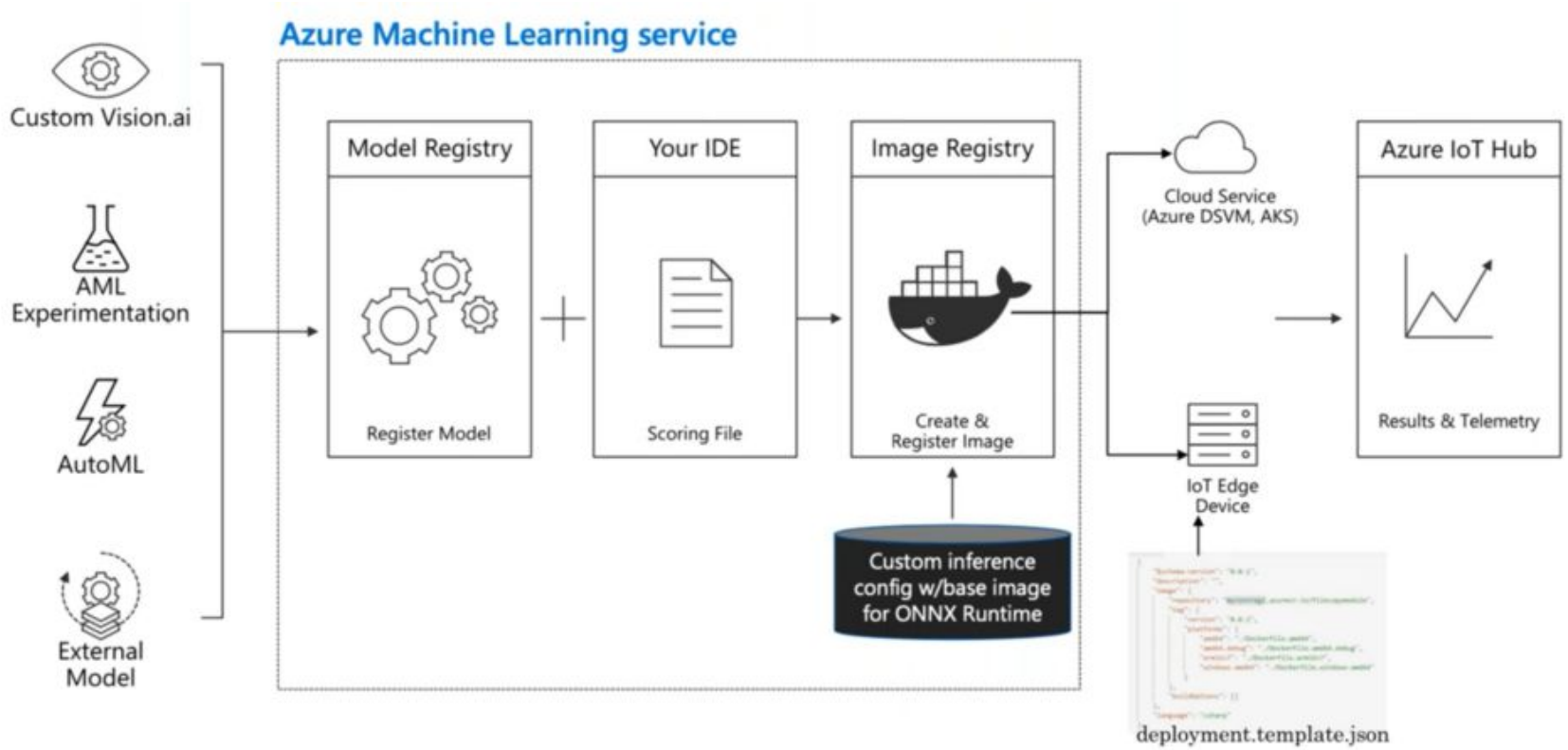
7 - Onnx Runtime

- **Tốc độ cao**: ONNX Runtime được thiết kế để cung cấp hiệu suất thực thi mô hình neural network nhanh chóng. Sử dụng các kỹ thuật tối ưu hóa để tận dụng tốt nhất khả năng của phần cứng, bao gồm cả CPU và GPU
- **Khả năng tương thích**: ONNX Runtime hỗ trợ nhiều phiên bản của ONNX Specification và duy trì khả năng tương thích ngược với các phiên bản cũ hơn, giúp cho việc thực thi các mô hình ONNX trên các phiên bản ONNX khác nhau trở nên thuận tiện.
- **Hỗ trợ nhiều nền tảng**: ONNX Runtime có thể thực thi các mô hình ONNX trên nhiều nền tảng khác nhau bao gồm Windows, Linux, MacOS và cả các thiết bị nhúng.
- **Tích hợp với nhiều ngôn ngữ lập trình**: ONNX Runtime cung cấp API cho nhiều ngôn ngữ lập trình khác nhau như C++, C#, Python, và Java, giúp dễ dàng tích hợp mô hình ONNX vào các ứng dụng một cách dễ dàng.

7 - Onnx Runtime



8 - Deploying onnx models - Azure ML



9 - Onnx Model Interop