Infuse AI in your Windows apps with .NET

Alexandre Chohfi @AlexandreChohfi
Nikola Metulev @metulev

AI in Windows apps








Demo

Windows Copilot Library

APIs backed by on device ML models

Phi Silica and **Text Recognition** will be available in an upcoming WinAppSDK experimental release

Vector Embeddings, **RAG**, **Text Summarization** and more will follow

```
if (!LanguageModel.IsAvailable())
{
    await LanguageModel.MakeAvailableAsync();
}

var languageModel = await LanguageModel.CreateAsync();

var response = await languageModel.GenerateResponseAsync("what's the meaning of life?");
```

DirectML

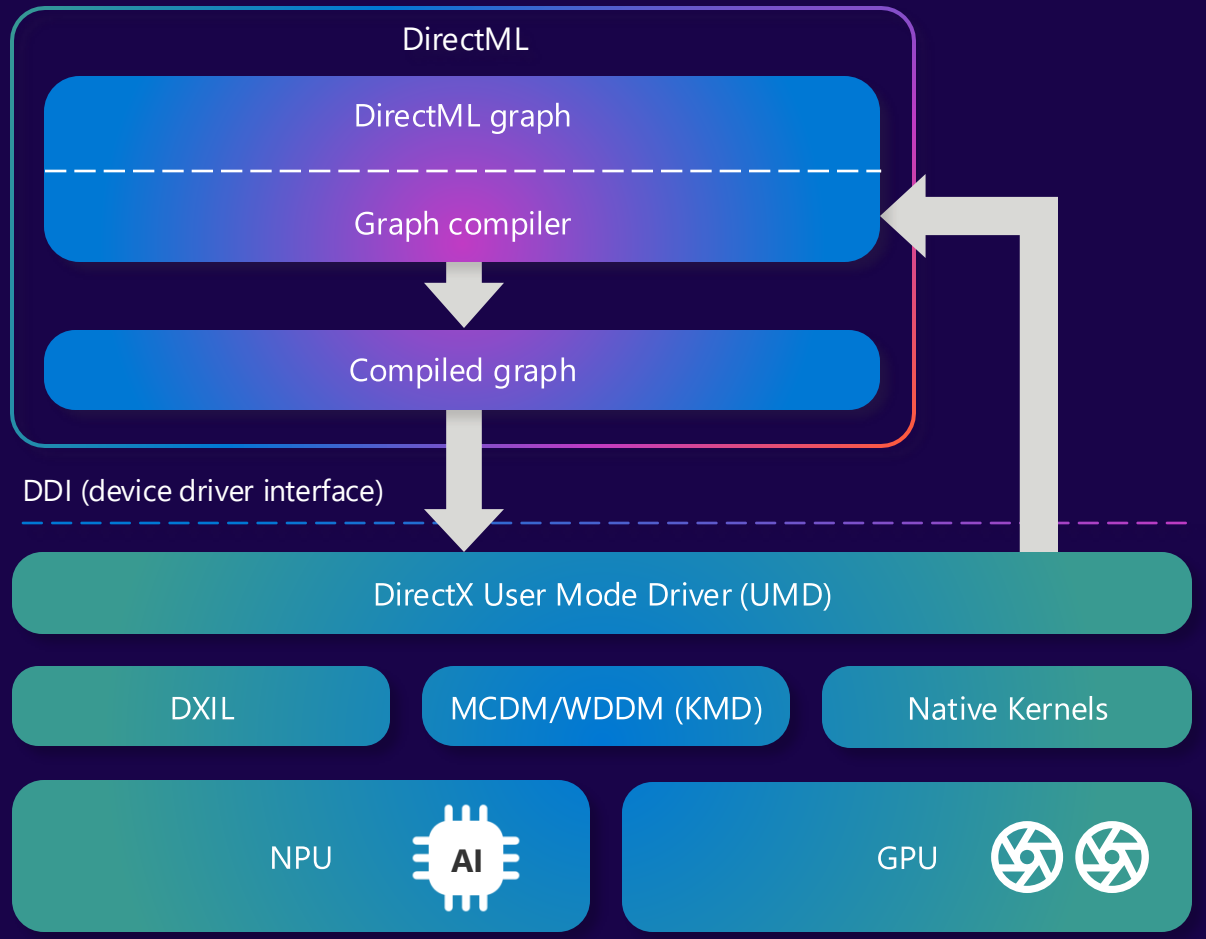
High-performance, hardware-accelerated DirectX 12 library for machine learning on Windows

A single, hardware-abstraction API for optimization and deployment, that scales across hardware

Full GPU support and expanding to include NPUs for full breadth of hardware support with AMD, Intel, Nvidia, and Qualcomm

Support for 4-bit Activation-Aware Quantization (AWQ) for minimal impact on accuracy to enable more models across Windows hardware

Available as NuGet package or as part of Windows 10 SDK or newer



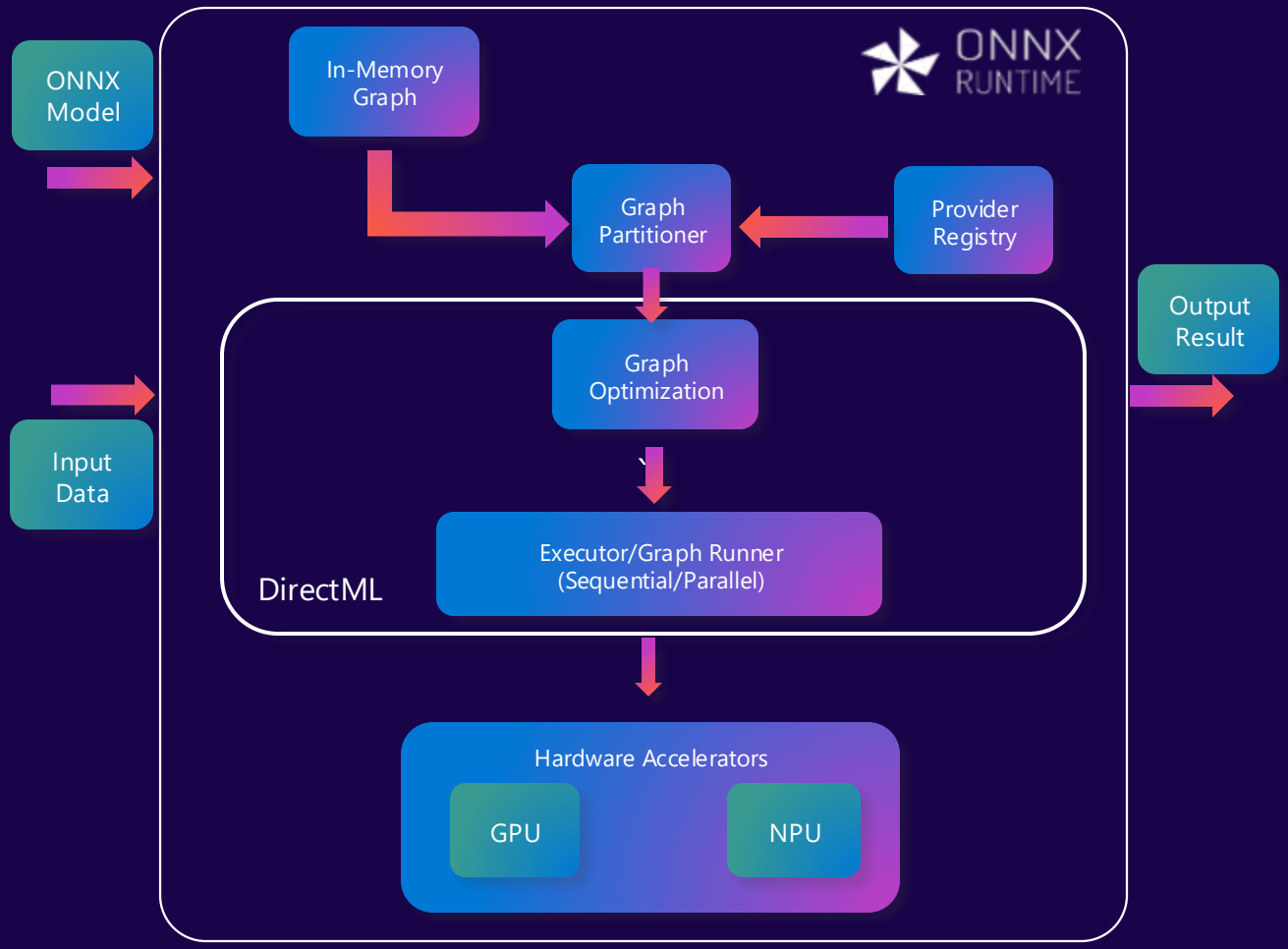
ONNX, ONNX Runtime (ORT), ORT GenAI, Olive

ONNX = Open and interoperable file format for ML and DNN models.

ONNX Runtime = Fast and efficient model inference and training engine that works across a diverse range of hardware accelerators.

ORT Generate API (ORT GenAI) = High performance, easy-to-use API for GenAI models

Olive = A toolkit for hardware-aware AI model optimization.



Thank you