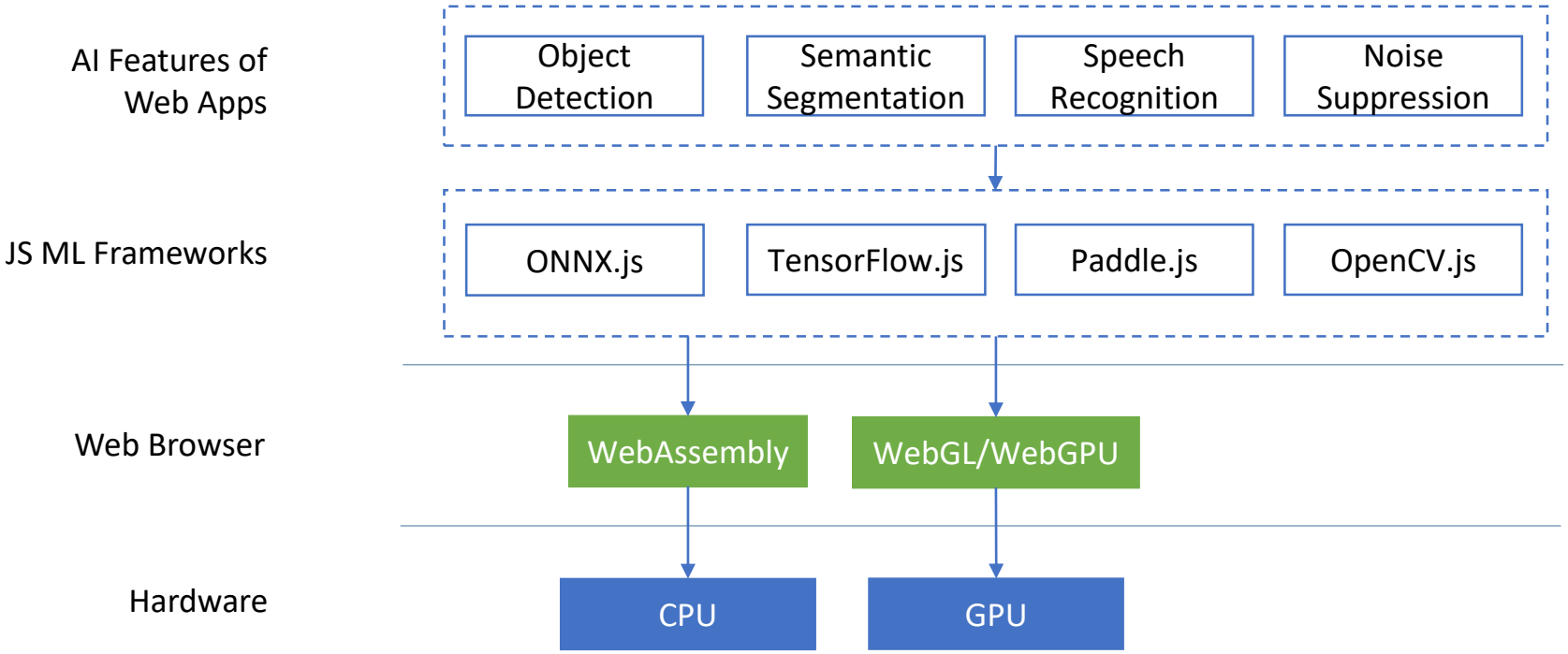
Access purpose-built ML hardware with Web Neural Network API

Ningxin Hu

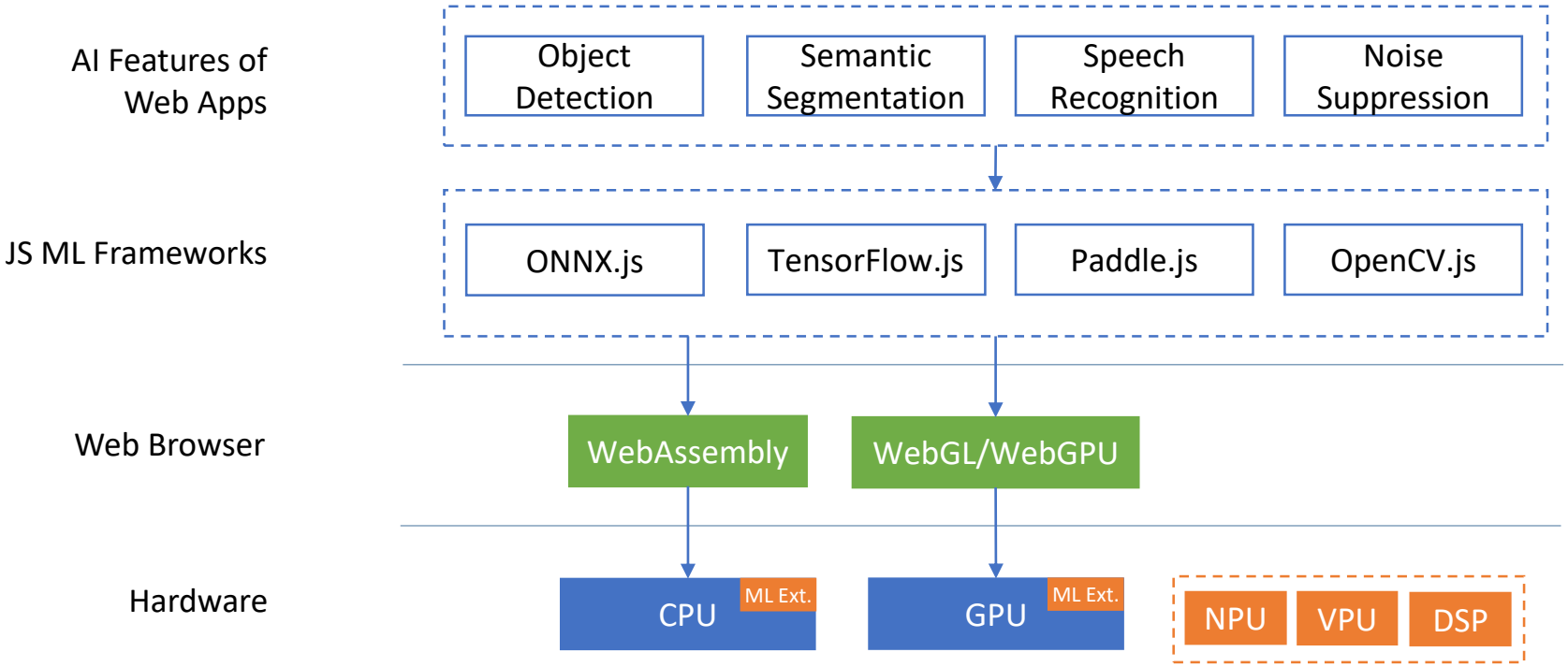
Intel Corporation

July 2020

The JS ML frameworks and AI Web apps

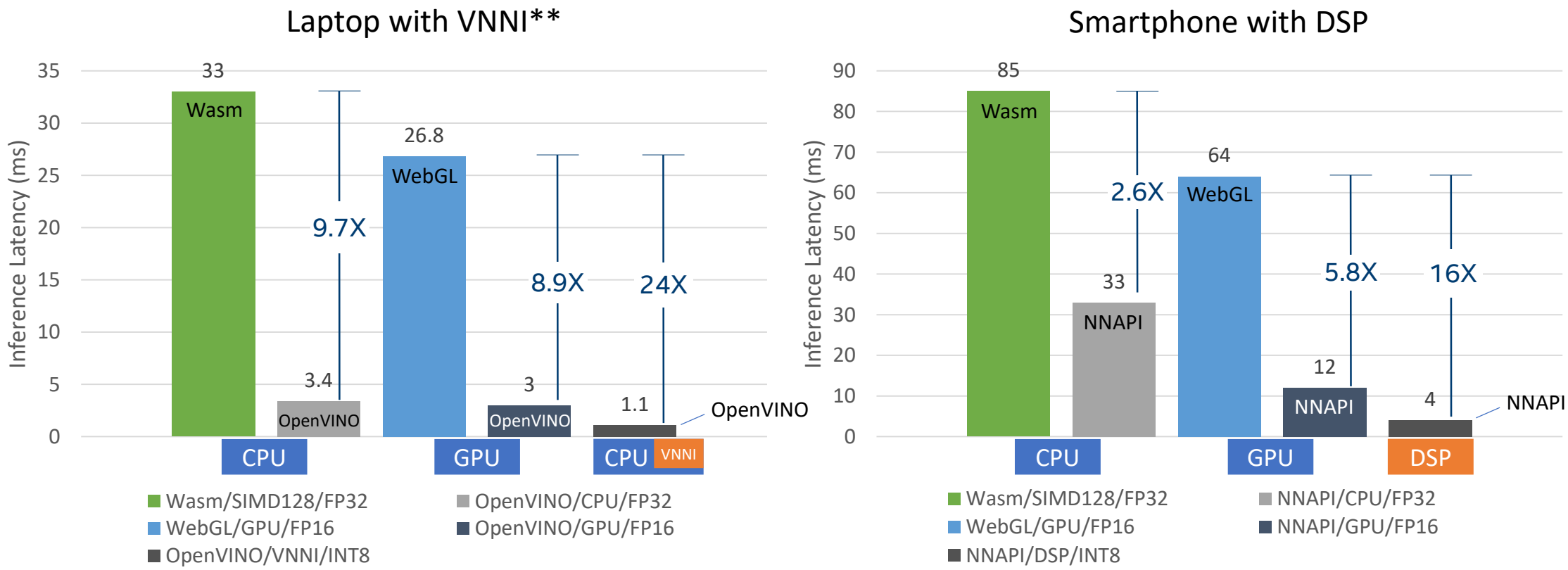


The purpose-built ML hardware



The performance gap: Web and native

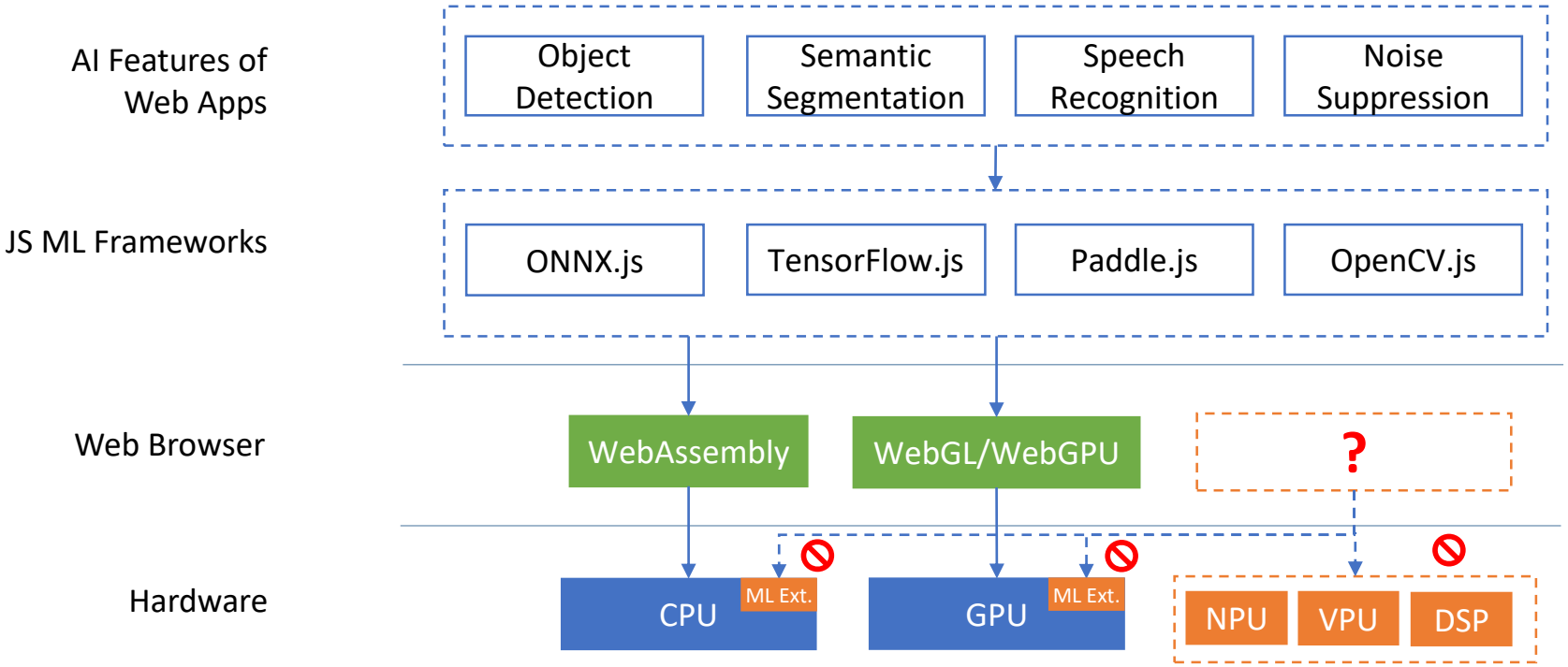
MobileNet* Inference Latency (smaller is better)



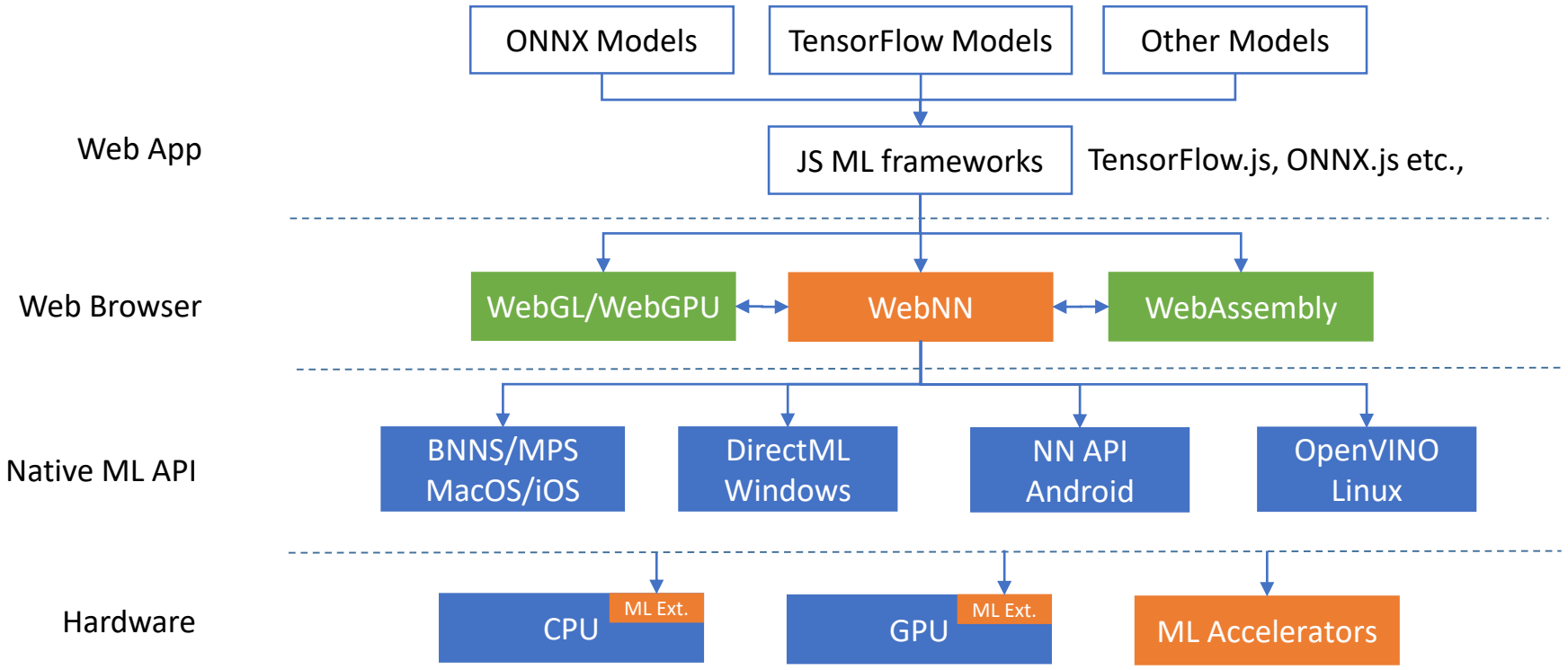
* Batch size: 1, input size: 224x224, width multiplier: 1.0

** VNNI: Vector Neural Network Instruction

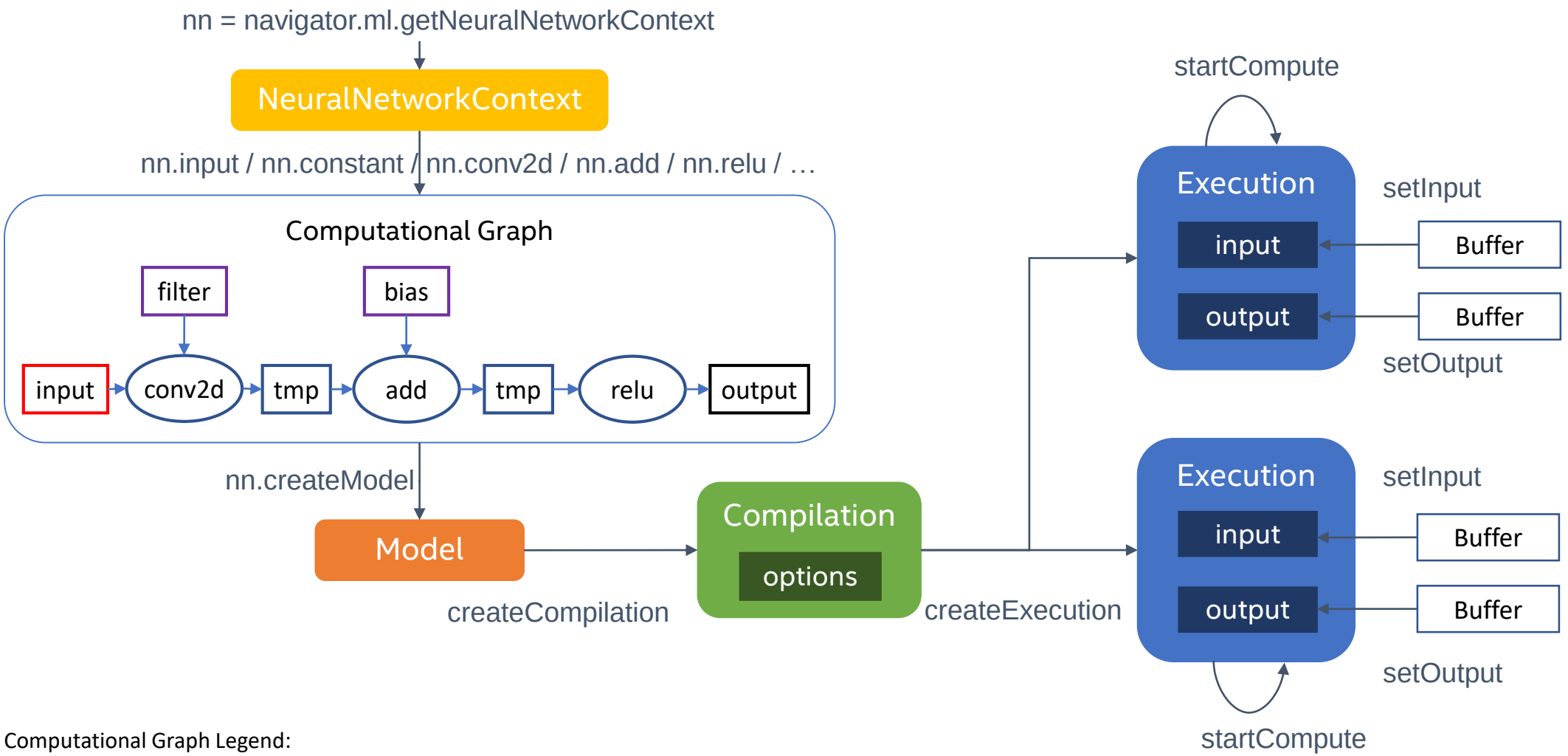
The Web is disconnected from ML hardware



WebNN: the architecture view



WebNN: the programming model

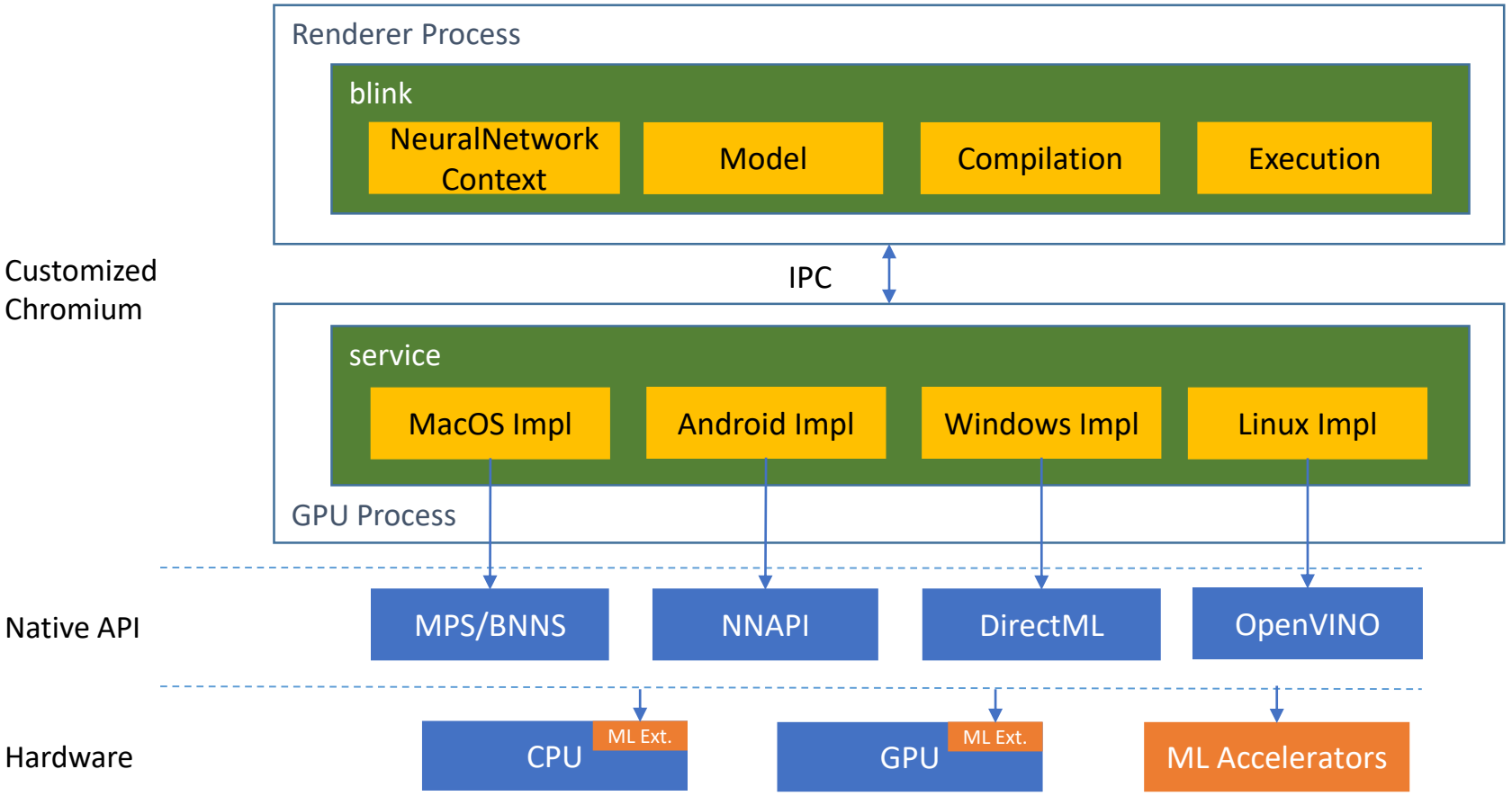


Computational Graph Legend:

input | constant | output | operand | operation

https://webmachinelearning.github.io/webnn/

WebNN: the proof-of-concept implementation



https://github.com/otcshare/chromium-src

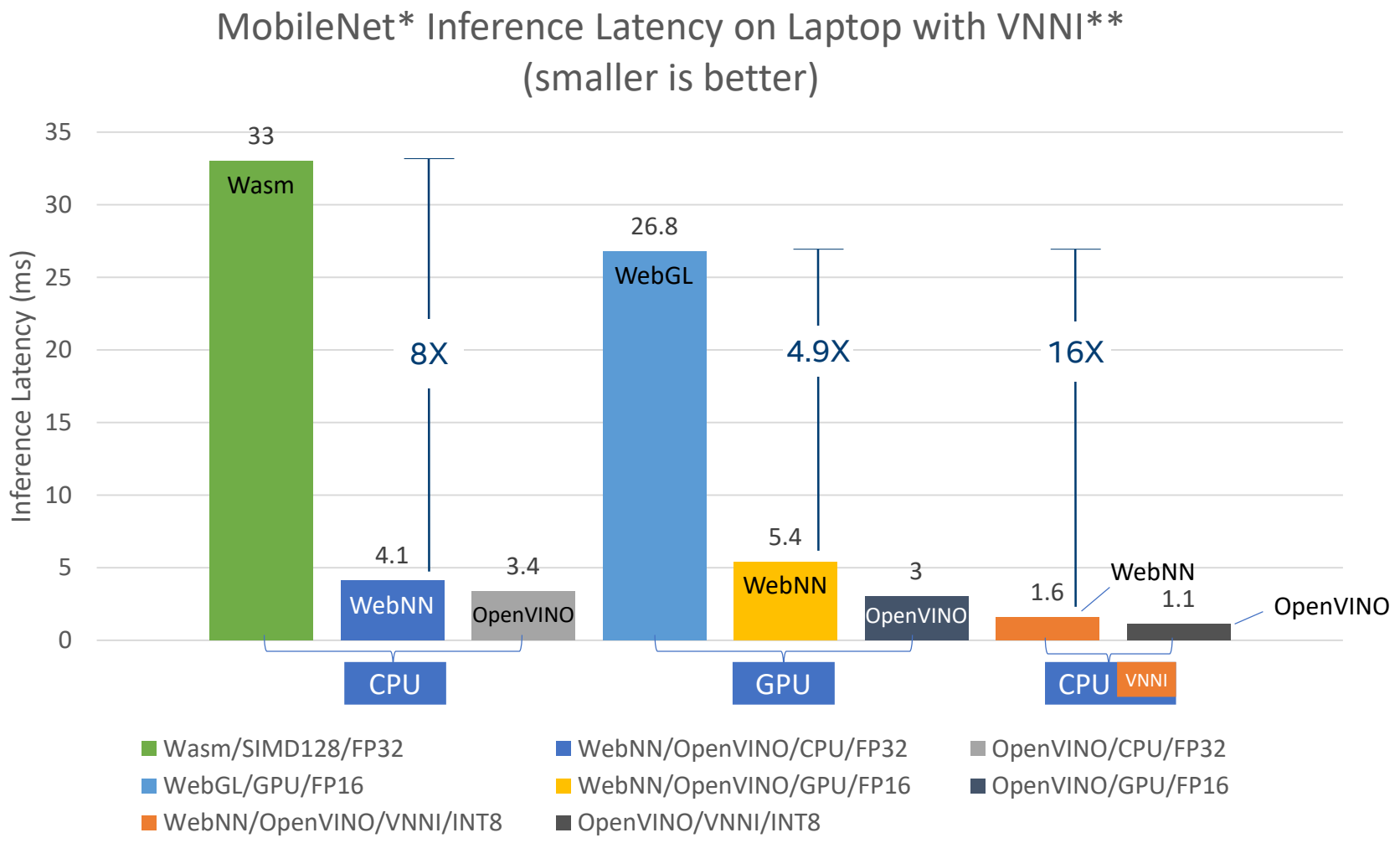
WebNN: the demos

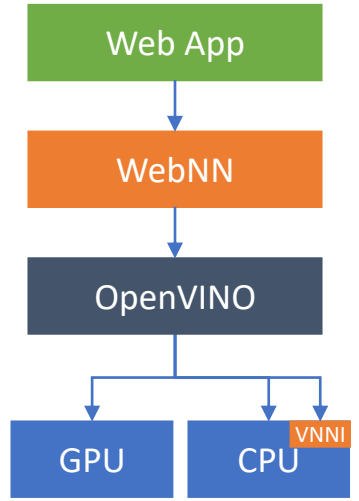
WebNN image classification on a laptop with VNNI

WebNN image classification on a smartphone with DSP

https://intel.github.io/webml-polyfill/examples/image_classification

WebNN: the PoC performance

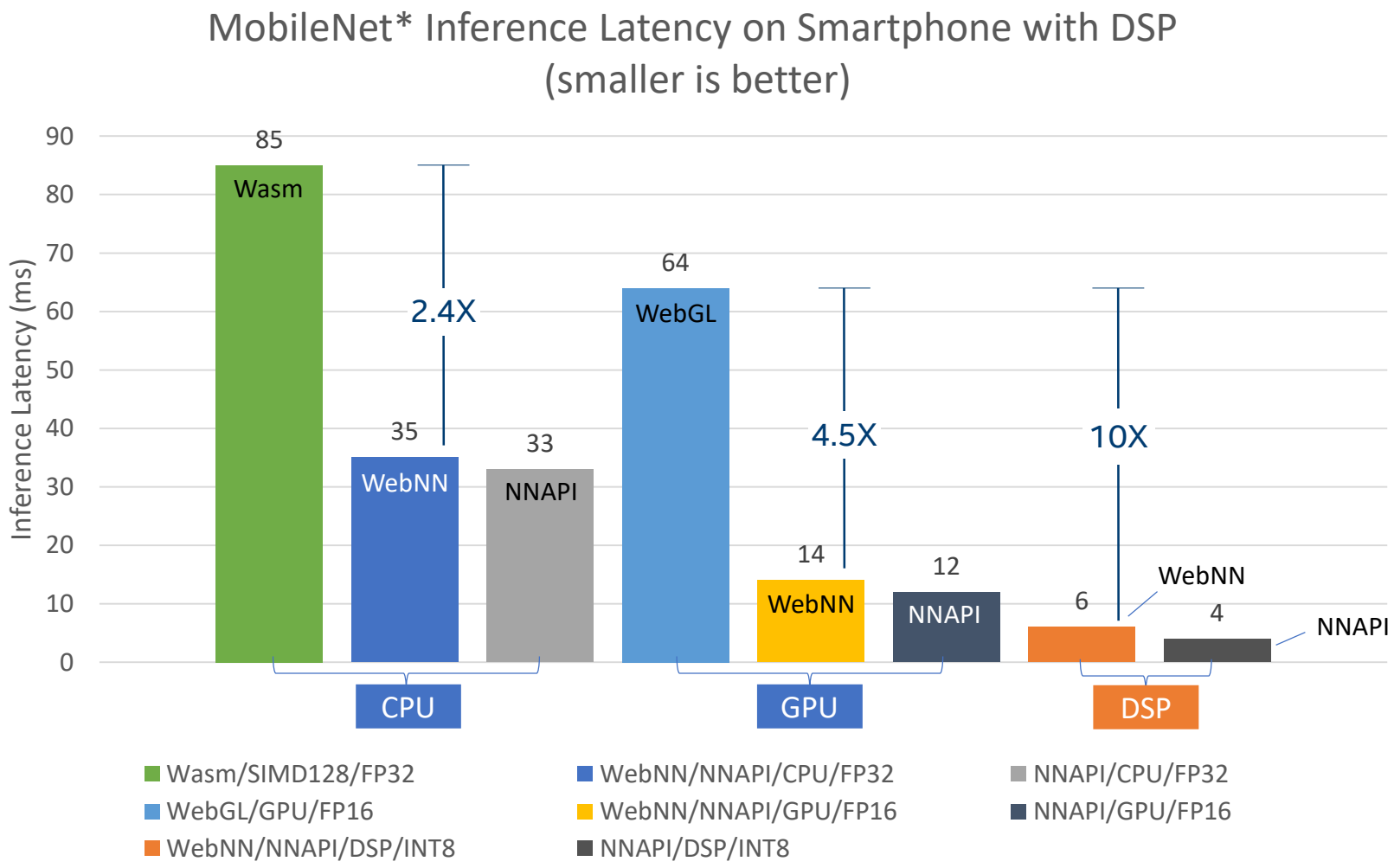


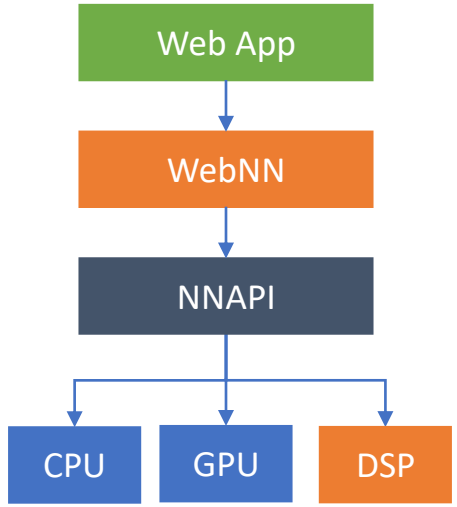


* Batch size: 1, input size: 224x224, width multiplier: 1.0

** VNNI: Vector Neural Network Instruction

WebNN: the PoC performance – cont'd





* Batch size: 1, input size: 224x224, width multiplier: 1.0

Call for Participation

https://www.w3.org/community/webmachinelearning/

https://webmachinelearning.github.io/webnn/

TABLE OF CONTENTS

Web Neural Network API

Draft Community Group Report, 1 July 2020

This version:
https://webmachinelearning.github.io/webnn/

Issue Tracking:
GitHub

Editors:
Ningxin Hu (Intel Corporation)
Chai Chaoweeraprasit (Microsoft Corporation)

Explainer:
explainer.md

Abstract

This document describes a dedicated low-level API for neural network inference hardware acceleration.

Status of this document

This specification was published by the Machine Learning for the Web Community Group. It is not a W3C Standard nor is it on the W3C Standards Track. Please note that under the W3C Community Contributor License Agreement (CLA) there is a limited opt-out and other conditions apply. Learn more about W3C Community and Business Groups.

§ 1. Introduction

We're working on this section. Meanwhile, please take a look at the explainer.

§ 2. Use cases

§ 2.1. Application Use Cases

This section illustrates application-level use cases for neural network inference hardware acceleration. All

Thanks

Appendix

- WebNN spec: https://webmachinelearning.github.io/webnn/
- WebML CG: https://www.w3.org/community/webmachinelearning/
- NNAPI: https://developer.android.com/ndk/guides/neuralnetworks
- DirectML: https://docs.microsoft.com/en-us/windows/win32/direct3d12/dml-intro
- MPS: https://developer.apple.com/documentation/metalperformanceshaders
- OpenVINO: https://docs.openvinotoolkit.org/
- TensorFlow.js: https://js.tensorflow.org/
- ONNX.js: https://github.com/microsoft/onnxjs
- Paddle.js: https://github.com/PaddlePaddle/Paddle-Lite/tree/develop/web
- OpenCV.js: https://docs.opencv.org/3.4.10/d5/d10/tutorial_js_root.html
- AI-benchmark: http://ai-benchmark.com/
- TensorFlow.js benchmark: https://tensorflow.github.io/tfjs/e2e/benchmarks/
- Wasm SIMD128: https://github.com/WebAssembly/simd
- AVX512-VNNI: https://en.wikichip.org/wiki/x86/avx512_vnni