JS Bangkok 2.0.0

ONNX RUNTIME

web neural network

AI web development with **Web Neural Network (WebNN) API**

Enable web app executing AI utilizing CPU, GPU, NPU on client

Surasuk Oakkharaamonphong

Microsoft MVP AI Platform & Developer Technologies

Technical Coach at Arise & INFINITAS by Krungthai

Microsoft MVP Most Valuable Professional

Arise by INFINITAS

INFINITAS by Krungthai

Date: 18 Oct 2024 1:55 pm #JSBangkok #JSBKK

CodeBangkok

@CodeBangkok · 4.48K subscribers · 71 videos

Microsoft Developer Thailand >

Subscribe Join

Go Programming - Basic Syntax

19K views • 2 years ago

Go Programming - GORM

10K views • 2 years ago

Go Programming - Fiber Web Framework

12K views • 2 years ago

Go Programming - Hexagonal Architecture EP.2 (log, error)

7.1K views • 2 years ago

idity Programming (1)

views • 2 years ago

Semantic Kernel - Global AI Bootcamp 2024

330 views • 2 weeks ago

Prompt Engineering - The Art and Science of Querying AI

610 views • 3 weeks ago

AWS re:Invent 2023 Recap

206 views • 1 month ago

Orchestrate your AI with Semantic Kernel

391 views • 2 months ago

ARISE CONNEXT - Make a better world with AI

668 views • 5 months ago

GitHub Copilot Day

1.2K views • 6 months ago

Semantic Kernel - AI Orchestration in Copilot Stack

573 views • 6 months ago

Azure OpenAI Service Development with Semantic Kernel

394 views • 6 months ago

Swift Programming - Struct, Class, Protocol, Error Handling

818 views • 1 year ago

Swift Programming - Basic Syntax

1.1K views • 1 year ago

Metaverse Development with Web3.js | Three.js | VR | NFT

2.9K views • 1 year ago

.NET MAUI Release

1.2K views • 1 year ago

podman & docker containers technology

2.8K views • 8 months ago

ChatGPT OpenAI Technology

1.1K views • 9 months ago

Azure OpenAI Service

622 views • 10 months ago

git Workflow and Branching Strategies

1.8K views • 1 year ago

Go Programming - gRPC

16K views • 1 year ago

Go Programming - Kafka

13K views • 2 years ago

Solidity Programming: ERC-721 NFT

2.2K views • 2 years ago

Solidity Programming: ERC-20 Token Standard

1.5K views • 2 years ago

git - Distributed Version Control

2.8K views • 1 year ago

Metaverse Development with WebXR, Augmented Reality (AR)

547 views • 1 year ago

React Web Development EP.2 API, Async Test, Mocks

1.4K views • 1 year ago

React Web Development - Get Started

2.9K views • 1 year ago

Rust Programming EP.3 HTTP Server

2K views • 2 years ago

Rust Programming EP.2

2.3K views • 2 years ago

Flutter BLoC Update 8.x

4.3K views • 2 years ago

Rust Programming

6.4K views • 2 years ago

Flutter: Input & Form

1.9K views • 2 years ago

สัมภาษณ์ในรายการ Digital Thailand คำเงิน รายการโดย เอ็น ปานระพี ช่อง 3 Family 28 เรื่อ...

1.3K views • 5 years ago

Xamarin Test Cloud

55 views • 6 years ago

Elasticsearch: Get Started

6.4K views • 2 years ago

Back to sessions

The Web is AI Ready—maximize your AI web development with WebNN

Wednesday, May 22 | 5:23 AM - 5:38 AM Indochina Time Duration 15 minutes

StudioFP126

On Demand | Microsoft Build Stage | In Seattle + Online

Speaker:  Moh Haghighat | Intel Corporation

Resources

Download Video

Download Transcript

View slide deck

Session tags

Session type Microsoft Build Stage

Topic AI Development

Topic Developer Tools

Level Advanced (300)

Delivery Type In Seattle

Delivery Type Online

Tag AI Infrastructure

Tag Client Development

Recording Availability Recorded

Developing Web-Based AI Apps for the Client

An overview of emerging web capabilities to enable exciting AI experiences on the open web platform

Moh Haghighat, Chai Chaoweeraprasit

Moh Haghighat

Intel Fellow, Chief Web Technologist

Location:

Santa Clara, California

Education:

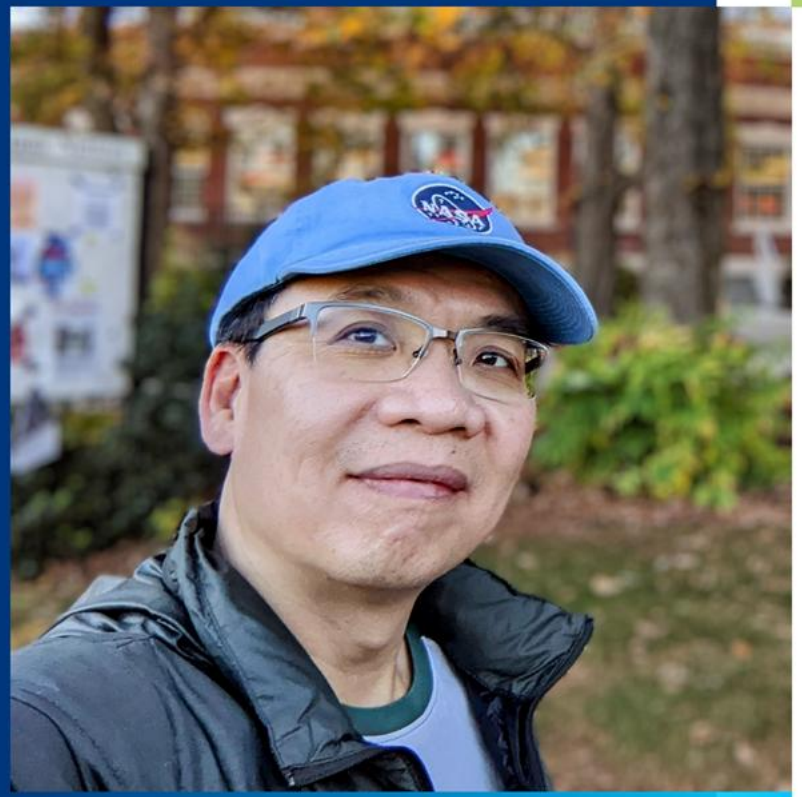
Ph.D., Computer Science, University of Illinois at Urbana-Champaign

Experience:

28 years as Software Architect at Intel

Areas of interest:

Web Technologies, Programming Models, Compilers, Parallelism



Chai Chaoweeraprasit

Partner Engineering Manager, Microsoft

Location:

Mountain View, California

Education:

Bachelor of Computer Engineering

Experience:

Two decades of core operating system engineering

Areas of interest:

Artificial Intelligence, Deep Neural Network, Web Standard Development, Core Operating System, Computer Graphics, GPU Hardware Acceleration, Digital Typography, Software Internationalization and Unicode, Coding Education

เสวนาถอดรหัสเอไอ

ผ่านมุมมอง F1

ลงทะเบียนฟรี! SAT 6 July 13:00 at Microsoft Thailand

Dr. Komes C
Panelist

F1 เอ็นจิเนียร์คนไทยที่คุมทีมแพลทฟอร์ม AI ของไมโครซอฟต์ บอกเล่าประสบการณ์กว่า 20 ปี ของคนไทยที่มีโอกาสได้คุมงานเดฟสำคัญของวินโดวส์ จนมาถึง AI ..โปรเจกต์ล่าสุดที่คุมคือ Phi 3 เอไอตัวเล็ก เขามองอนาคตของเอไออย่างไร ชีวิตในซิลิคัลวัลเลย์โอเคไหม จังหวะชีวิตที่ทำให้ไปอยู่อเมริกามากกว่าอยู่ไทยคืออะไร มาร่วมพูดคุยเสวนาสไตล์จิบกาแฟคุยกับตัวเป็นๆกันได้

Big Pongrapee
Moderator

F1

ชัย เชาว์วีระประสิทธิ์

หัวหน้าทีมพัฒนาแพลตฟอร์ม AI

จัดโดยชมรมผู้ใช้เอไอไทยและพันธมิตร

DaVinci AI สถาบันศิลป์ปัญญาประดิษฐ์ดาวินชี

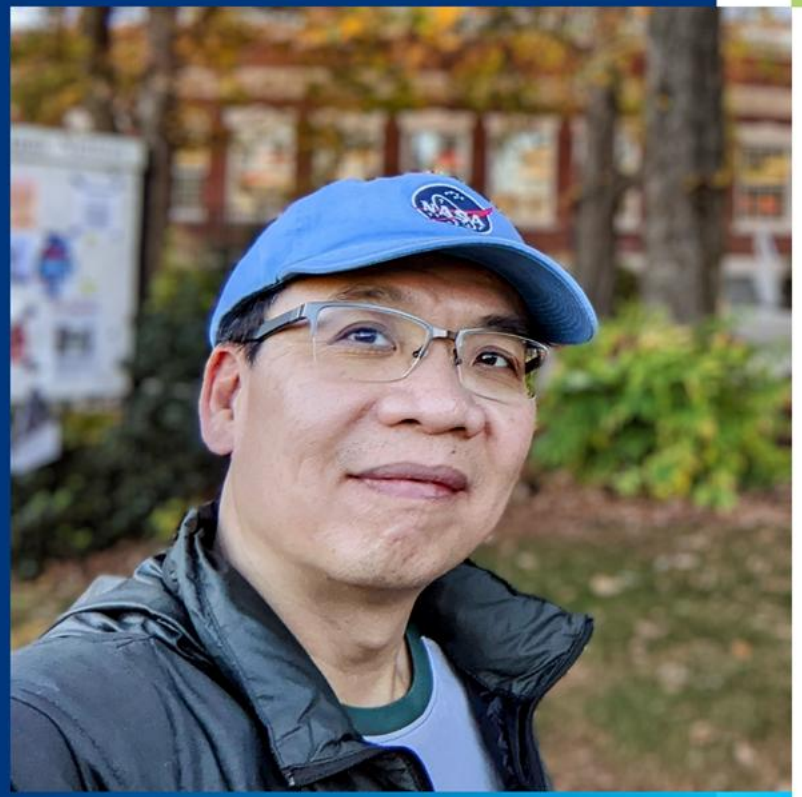
KBTG

Microsoft

AI

รัน AI บนเว็บเบราเซอร์

30 สิงหาคม 2024

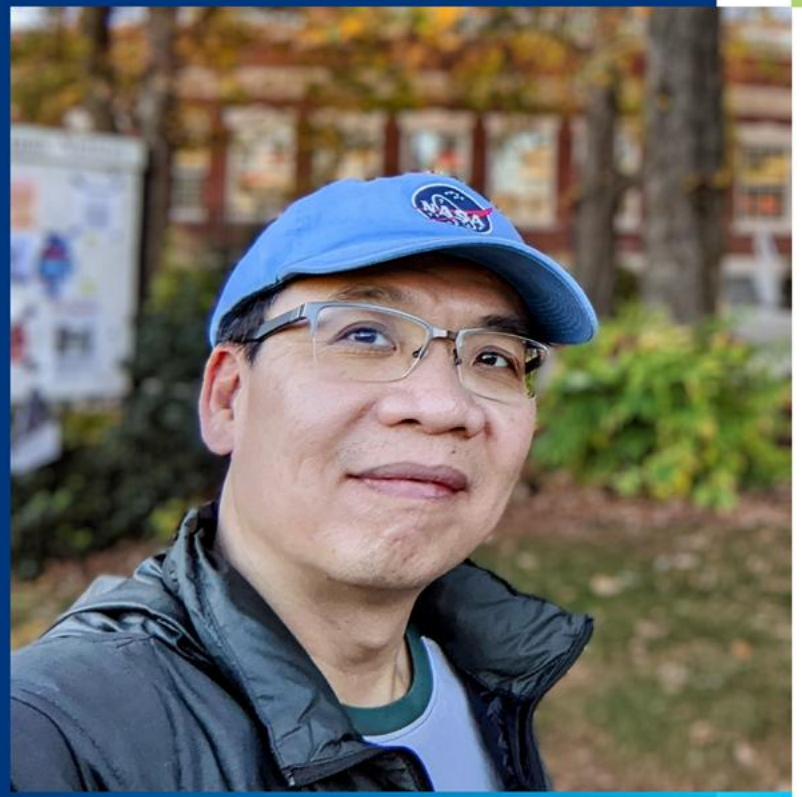


Chai Chaoweeraprasit

การจะเอา AI มารันในเว็บแอพต้องทำในเว็บเบราเซอร์ ถ้าใครเคยลองจะรู้ว่ามันพอทำได้ แต่ช้าาาามาก อันนี้พูดถึงรัน AI บนเครื่องโดยตรงนะครับ ไม่ผ่านคลาวด์

เว็บแอพใช้ JavaScript เป็นหลัก สามารถเข้าถึง GPU ได้ผ่านทาง **WebGPU** มาตรฐานใหม่ที่มาแทน **WebGL** ทั้งสองตัวนี้เป็น Web API ซึ่งเป็นสเป็กที่รับรองโดย W3C

ตัว **WebGPU** สามารถใช้รัน AI บน GPU สปีดพอใช้ได้ แต่ยังไม่ดีมาก และยังค่อนข้างจำกัดโดยเฉพาะกับโมเดลแบบ Transformers

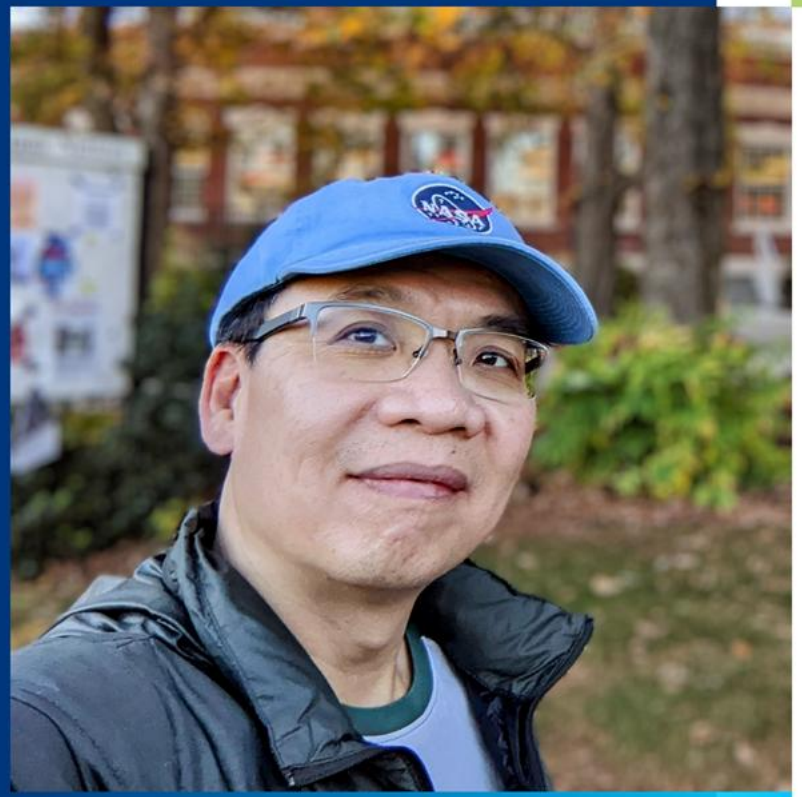


Chai Chaoweeraprasit

AI เป็น Workload ที่ค่อนข้างเฉพาะทาง จะเร็วได้จริงๆ ต้องรันแบบเนทีฟเท่านั้น ที่ว่าเนทีฟนี่หมายถึงรัน AI โดยเรียกใช้ฟังก์ชันเฉพาะทางที่อยู่ในตัว OS คือถ้าบน Windows ก็ต้อง DirectML หรือถ้าบน MacOS ก็ CoreML

เลยมาคิดว่าคงต้องสร้างมาตรฐานบนเว็บตัวใหม่ ที่เป็น Web API แบบ **WebGPU** แต่ใช้สำหรับ AI โดยเฉพาะ เจาะช่องให้เว็บแอพสามารถต่อตรงเข้าถึงฟังก์ชันเนทีฟในตัว OS ไม่ผ่านตัวกลาง

บังเอิญเพื่อนอีกคนซึ่งเป็นวิศวกรอยู่ที่ Intel คิดตรงกัน เลยช่วยกันร่างมาตรฐานใหม่ ใช้ชื่อว่า **WebNN** ชื่อเต็มว่า Web Neural Network API มี W3C หนุนหลัง แบ่งงานกันคือผมเขียนสเป็กส่วนเขาเป็นคนทดสอบ เริ่มทำกันจริงจังช่วงโควิดล็อคดาวน์เมื่อหลายปีที่ผ่านมา



Chai Chaoweeraprasit

จุดเด่นของ **WebNN** อีกอย่างหนึ่งคือไม่ขึ้นกับฮาร์ดแวร์ แอพสามารถเลือกได้ว่าจะรันบนซีพียู GPU หรือแม้แต่ NPU และไม่ขึ้นกับฟอร์แมตหรือเฟรมเวิร์ก จะใช้กับโมเดลแบบ ONNX หรือ TFLite ก็ได้ตามใจ

เมื่อหลายเดือนก่อนสัตยากล่าวเปิดตัว **WebNN** Developer Preview ในงานเดฟประจำปีของ Microsoft สร้างความฮือฮาพอสมควรโดยเฉพาะกับเหล่าเดฟที่ทำแอพทั้งหลาย

ตอนนี้ทีมแพลตฟอร์มเราเพิ่งออกอัพเดตตัวใหม่สำหรับรองรับ NPU ทั้งจากค่าย Intel และค่าย ARM บนชิพ Snapdragon X ของ Qualcomm รัน AI แบบเนทีฟบนเว็บเบราเซอร์ผ่าน **WebNN** ที่ขี่อยู่บนแพลตฟอร์ม DirectML สำหรับคอ MacOS คงต้องรอไปอีกสักพัก เพิ่งเห็นเริ่มๆ ทำกันอยู่

WebNN

Announcing Developer Preview

Web Neural Network API

W3C Candidate Recommendation Draft, 11 May 2024

▼ More details about this document

This version:
https://www.w3.org/TR/2024/CRD-webnn-20240511/

Latest published version:
https://www.w3.org/TR/webnn/

Editor's Draft:
https://webmachinelearning.github.io/webnn/

Previous Versions:
https://www.w3.org/TR/2024/CRD-webnn-20240510/

History:
https://www.w3.org/standards/history/webnn/

Implementation Report:
https://wpt.fyi/results/webnn?label=master&label=experimental&aligned&q=webnn

Test Suite:
https://github.com/web-platform-tests/wpt/tree/master/webnn

Feedback:
GitHub
Inline In Spec

Editors:
Ningxin Hu (Intel Corporation)
Dwayne Robinson (Microsoft Corporation)

Former Editor:
Chai Chaoweeraprasit (Microsoft Corporation)

Explainer:
explainer.md

Polyfill:
webnn-polyfill / webnn-samples

web neural network

Standard
W3C API

Unified Abstraction

Hetero
HW Exec

CPU, GPU, NPU

Integrated
ML Frameworks

ONNX RT Web, …

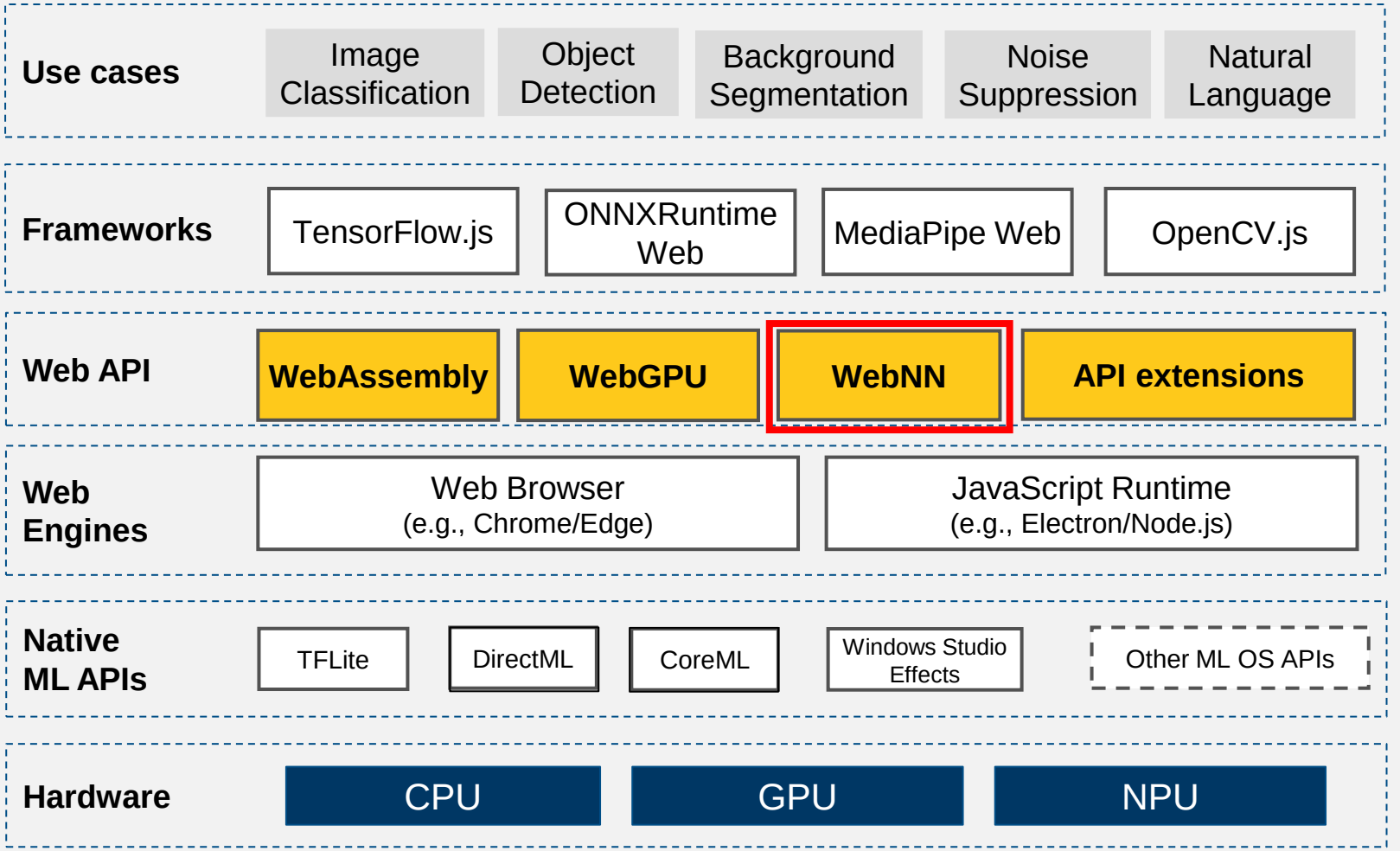
Near Native
Execution Characteristics

Perf. & Power

General
Computational Graph

BYOM

Hardware-Accelerated Web AI Overview



Updated

ONNX Runtime Web

Inference AI models in the web browser

Accelerates in-browser inferencing across CPUs, GPUs and NPUs, leveraging WebAssembly, WebGPU and WebNN

Supports generative AI models including Stable Diffusion, Tinyllama, Phi and more

Adopted widely by both 1Ps and 3Ps including Hugging Face Transformer.js

GA + update

ONNX Runtime Mobile

Inference AI models on mobile and embedded devices

A comprehensive solution offering minimized model and runtime sizes for on-device inferencing

Accelerates inference on Android and iOS with MLAS, NNAPI, CoreML, XNNPACK accelerators

Java / Kotlin, Objective-C, JavaScript, C#, and C++ APIs

Updated

ONNX Runtime Generate API

Inference Generative AI models in the cloud or at the edge

Easy-to-use API for cross-platform GenAI applications

High performance

Python, C#, C / C++ Language bindings. Java / Objective-C
Go coming soon

aka.ms/ortgenapi

Copilot+PC

intel CORE ULTRA 200V

AMD RYZEN AI 9 HX

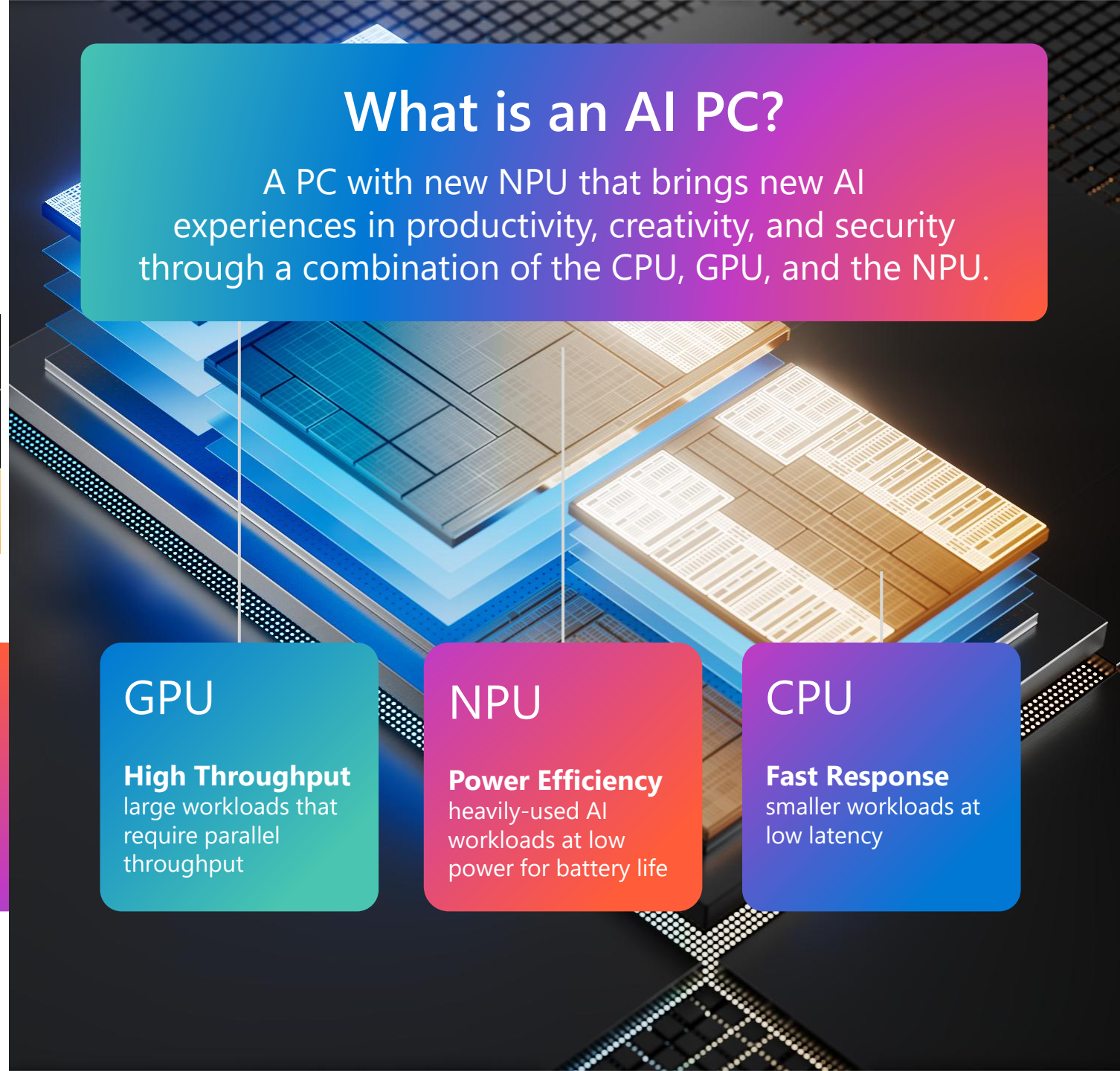
Snapdragon X Elite

40+ Trillion

Operations per second (TOPS)

What is an AI PC?

A PC with new NPU that brings new AI experiences in productivity, creativity, and security through a combination of the CPU, GPU, and the NPU.



GPU

High Throughput
large workloads that require parallel throughput

NPU

Power Efficiency
heavily-used AI workloads at low power for battery life

CPU

Fast Response
smaller workloads at low latency

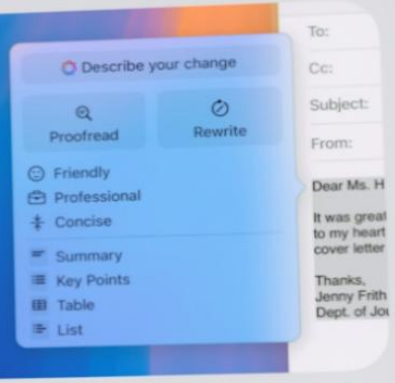
A more personal Siri

Private Cloud Compute

Clean Up in Photos

Summaries in Messages

Writing Tools



Reduce Interruptions in Focus

Apple Intelligence

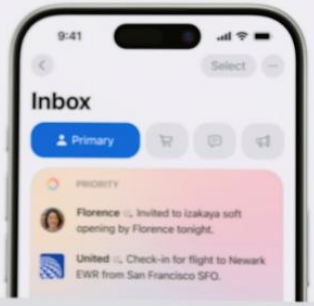
Image Playground

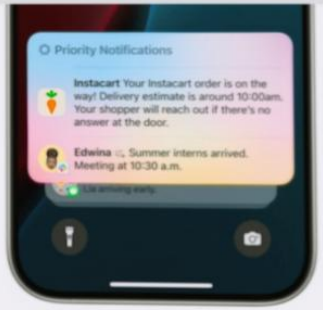
Genmoji

ate a Memory Movie

Describe a Memory...

Priority messages in Mail



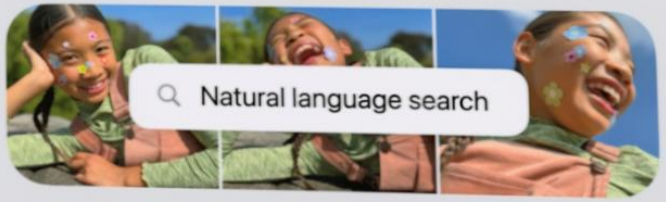


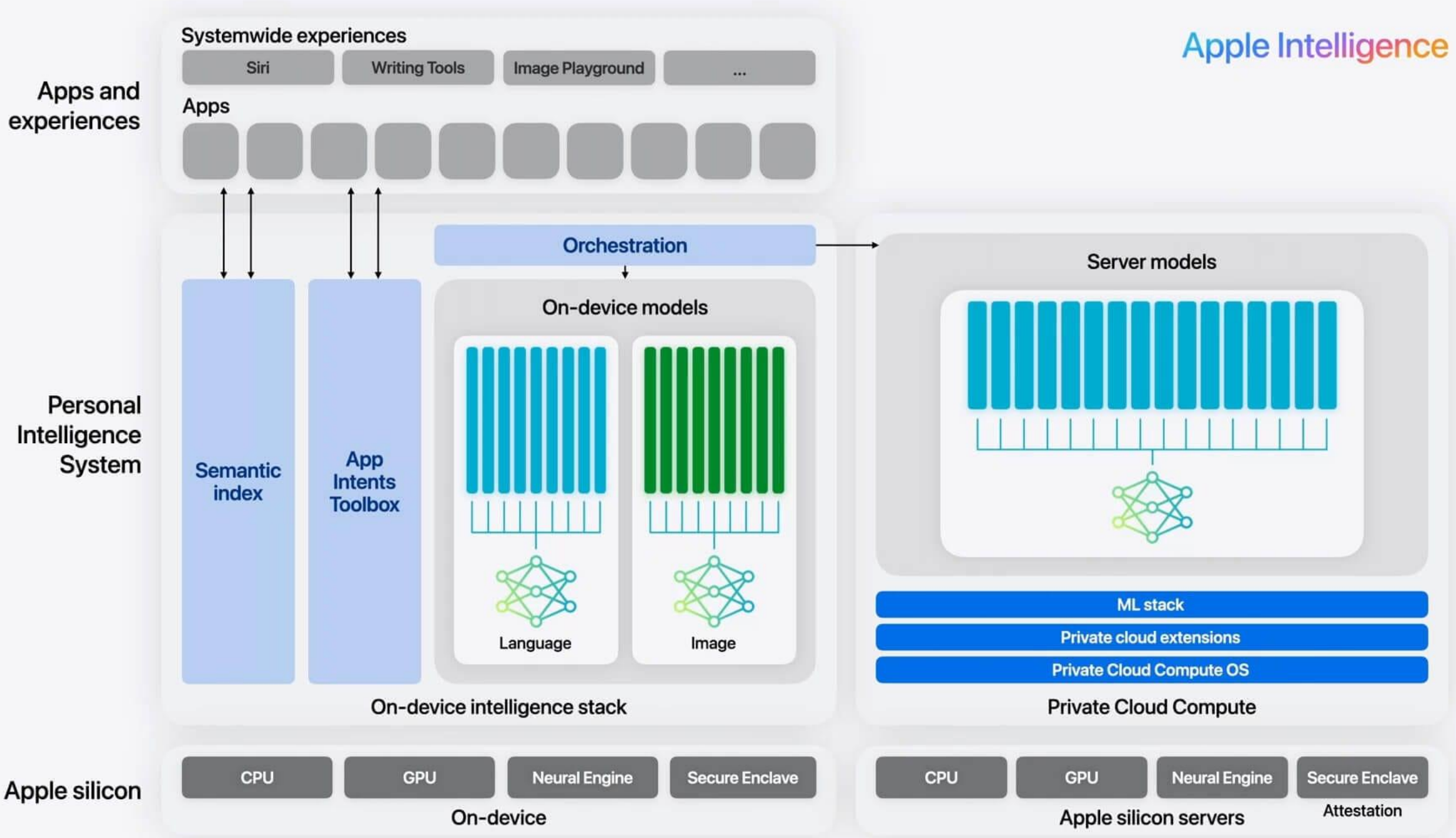
Priority notifications

Image Wand

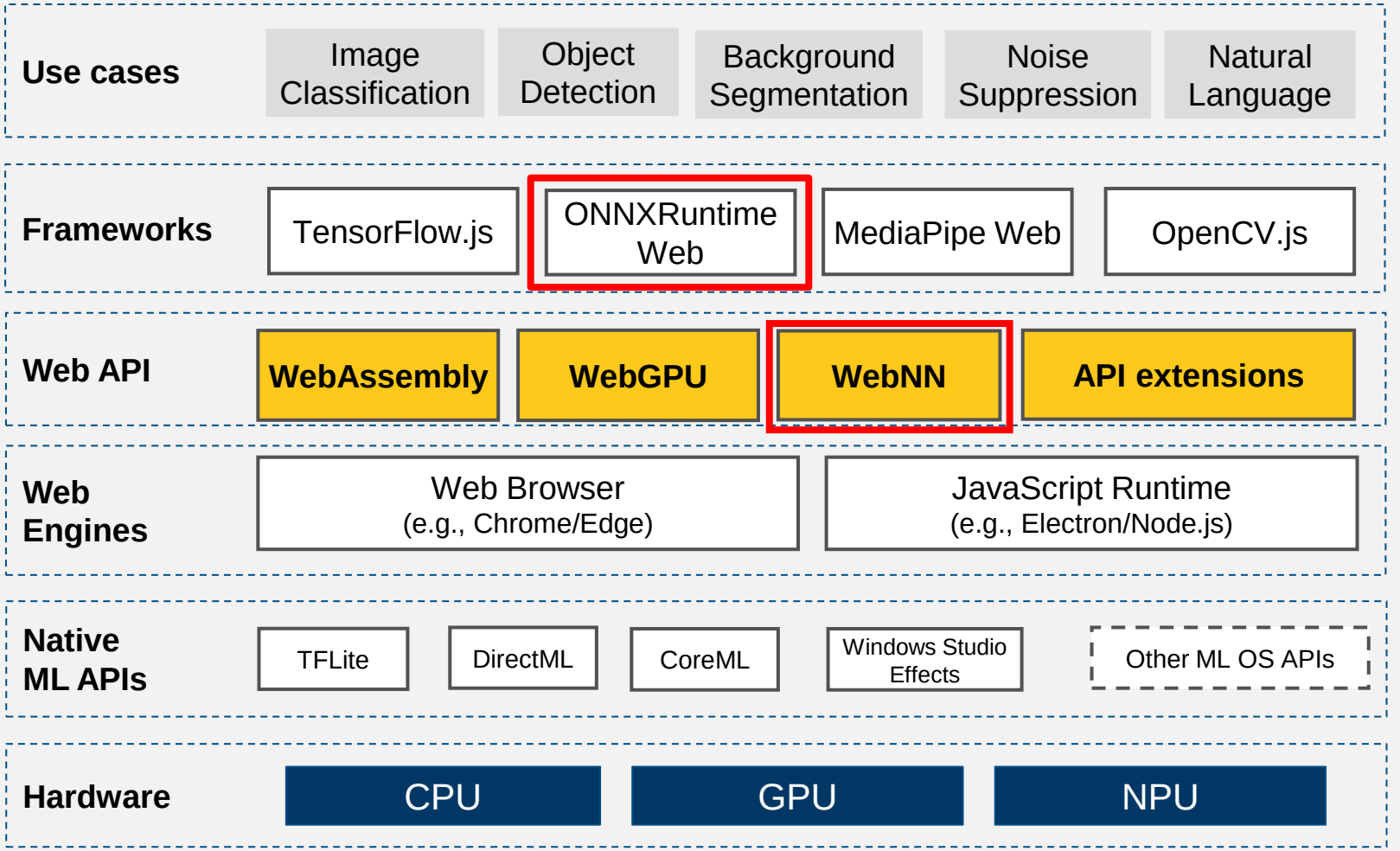
Audio recording

summaries



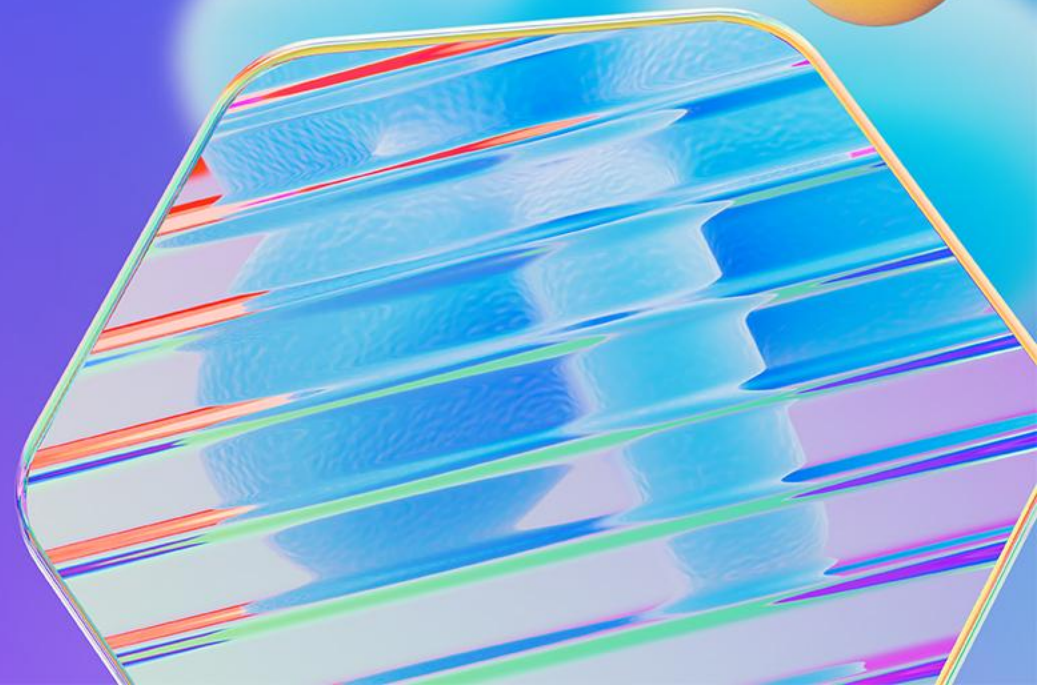

Apple Intelligence
Apps and experiences
Systemwide experiences
Siri
Writing Tools
Image Playground
...
Apps
Personal Intelligence System
Semantic index
App Intents Toolbox
Orchestration
On-device models
Language
Image
On-device intelligence stack
Server models
ML stack
Private cloud extensions
Private Cloud Compute OS
Private Cloud Compute
Apple silicon
CPU
GPU
Neural Engine
Secure Enclave
On-device
CPU
GPU
Neural Engine
Secure Enclave
Attestation
Apple silicon servers

Hardware-Accelerated Web AI Overview



ONNX - Open Neural Network Exchange

The open standard for machine learning interoperability



ONNX

Common format for ML Models

Can be exported from PyTorch, Tensor Flow, etc

Cross-platform

Desktop | IoT | Mobile | Cloud

Fast!

Supports DirectML, a cross-hardware ML acceleration API

The ONNX Trilogy

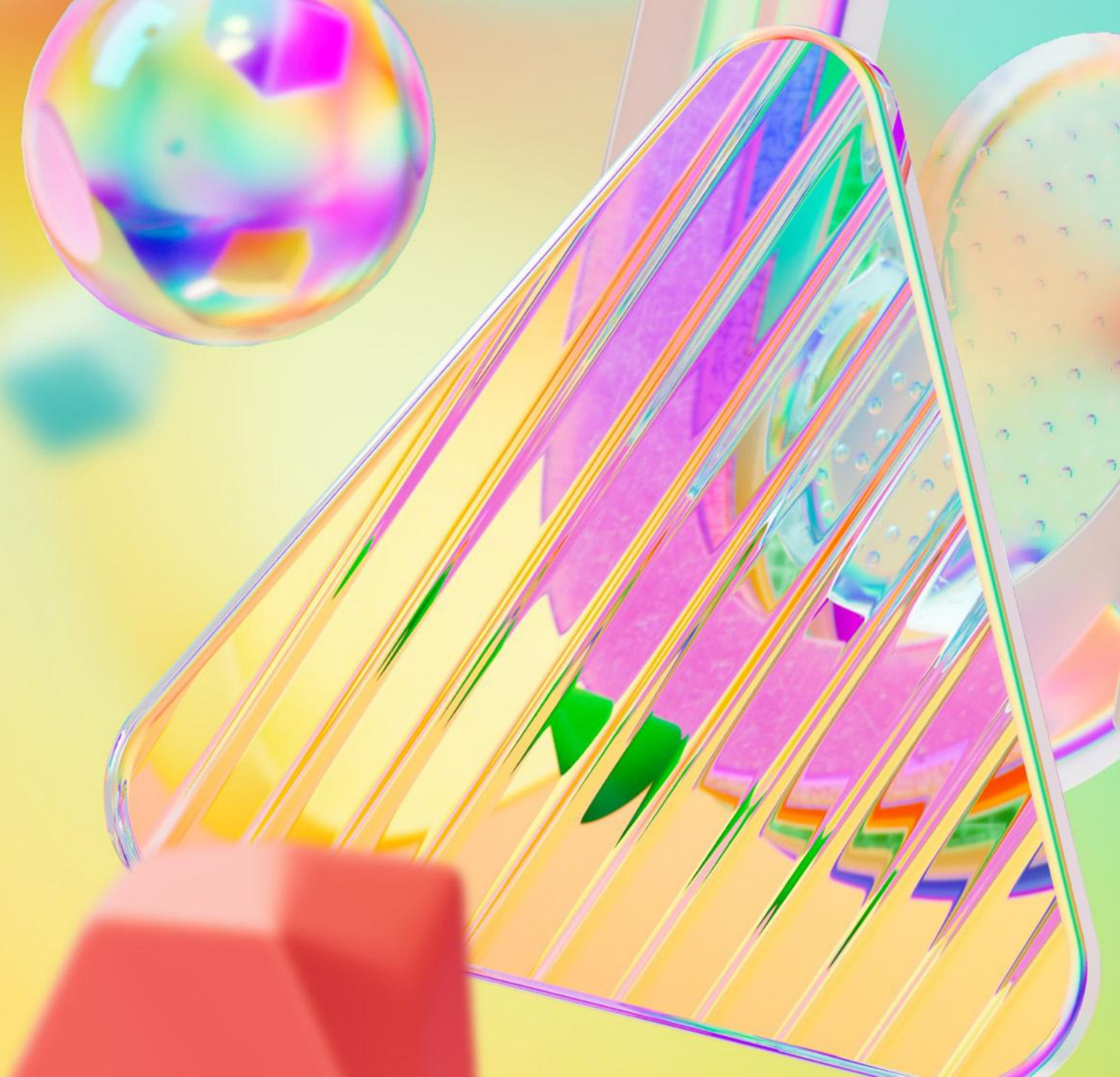
Open and interoperable file format for ML and DNN models.

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

A toolkit for *hardware-aware AI model optimization*. Output is an ONNX formatted model to run with quality and efficiency on the ONNX runtime.

Microsoft

Demo



AI Models

- Stable Diffusion 1.5 is a latent text-to-image diffusion model capable of generating photo-realistic images given any text input.
- SD-Turbo is a fast generative text-to-image model.
- Segment Anything is a new AI model from Meta AI that can "cut out" any object. You can segment any object from your uploaded images.
- Whisper Base is a pre-trained model for automatic speech recognition (ASR) and speech translation.
- Image Classification MobileNet and ResNet models. Take images as input and classify the major object in the image into a set of pre-defined classes.

References

- **WebNN Spec:** https://www.w3.org/TR/webnn/
- **WebNN Explainer:** https://github.com/webmachinelearning/webnn/blob/main/explainer.md
- **WebNN Implementation Status:** https://webmachinelearning.github.io/webnn-status/
- **Awesome WebNN:** https://github.com/webmachinelearning/awesome-webnn
- **WebNN Samples:** https://microsoft.github.io/webnn-developer-preview/ & https://webmachinelearning.github.io/webnn-samples/
- **WebNN Image Classification:** https://webmachinelearning.github.io/webnn-samples/image_classification/
- **WebNN Semantic Segmentation:** https://webmachinelearning.github.io/webnn-samples/semantic_segmentation/index.html
- **ONNX Runtime WebNN Execution Provider:** https://github.com/microsoft/onnxruntime/tree/main/onnxruntime/core/providers/webnn
- **Developing Web-Based AI Apps for the Client:** https://static.rainfocus.com/intel/innv2023/sess/1689805601141001Xyjy/supmat/DevelopingWebBasedAIApps_Final_1694887505835001jjpL.pdf