

# Arrow Lake NPU Introduction and Features

Technical Training Material

WW03, January 2024



# Legal Disclaimer

You may not use or facilitate the use of this document in connection with any infringement or other legal analysis. You may not use or facilitate the use of this document in connection with any infringement or other legal analysis concerning Intel products described herein. You agree to grant Intel a non-exclusive, royalty-free license to any patent claim thereafter drafted which includes subject matter disclosed herein.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps.

All product plans and roadmaps are subject to change without notice.

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at [intel.com](https://www.intel.com).

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

© Intel Corporation. Intel, OpenVINO™, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries.

\*Other names and brands may be claimed as the property of others.

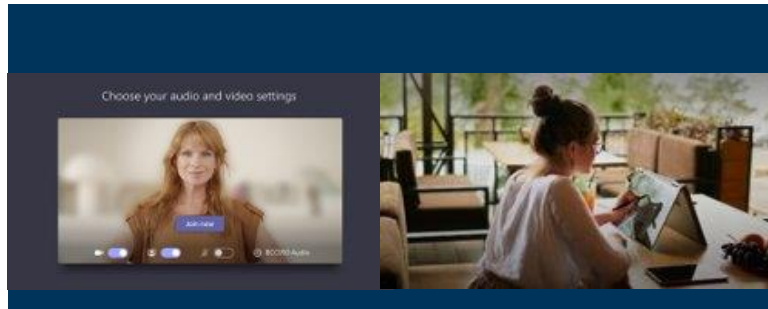
Copyright © 2024, Intel Corporation. All rights reserved.

# Agenda

- Client AI and Roadmap
- What is NPU?
- What is MEP (Windows\* Studio)?
- Arrow Lake (ARL) NPU POR Features
- ARL Audio Processing Object (APO)
- NPU Software Enabling and Experiences
- Q and A

# Client AI and Roadmap

# Transforming the PC Experience



## AI Today Enhancements

Elevated video collaboration & streaming  
Enhanced Audio effects  
Creator and Gaming effects

## Cloud

Massive scalable compute  
High Latency  
Privacy Concerns  
Expensive



## AI Tomorrow Everything

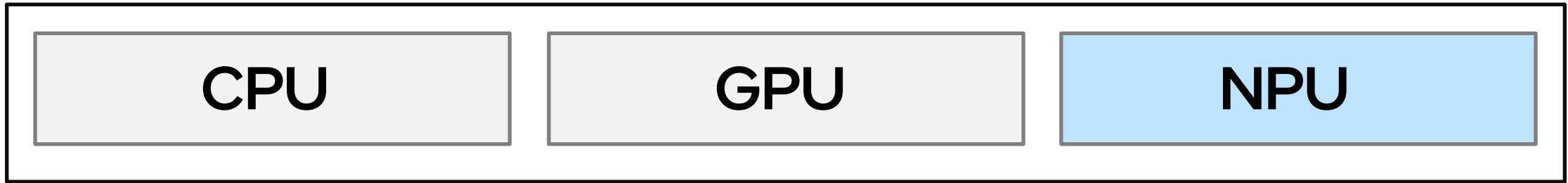
AI Assistants know your daily context  
More creative, productive, & collaborative  
Across everything you do

## Client

Massive distributed scale  
Low Latency  
Improved Privacy  
Lower Cost (to ISV)

# AI Inflection Point

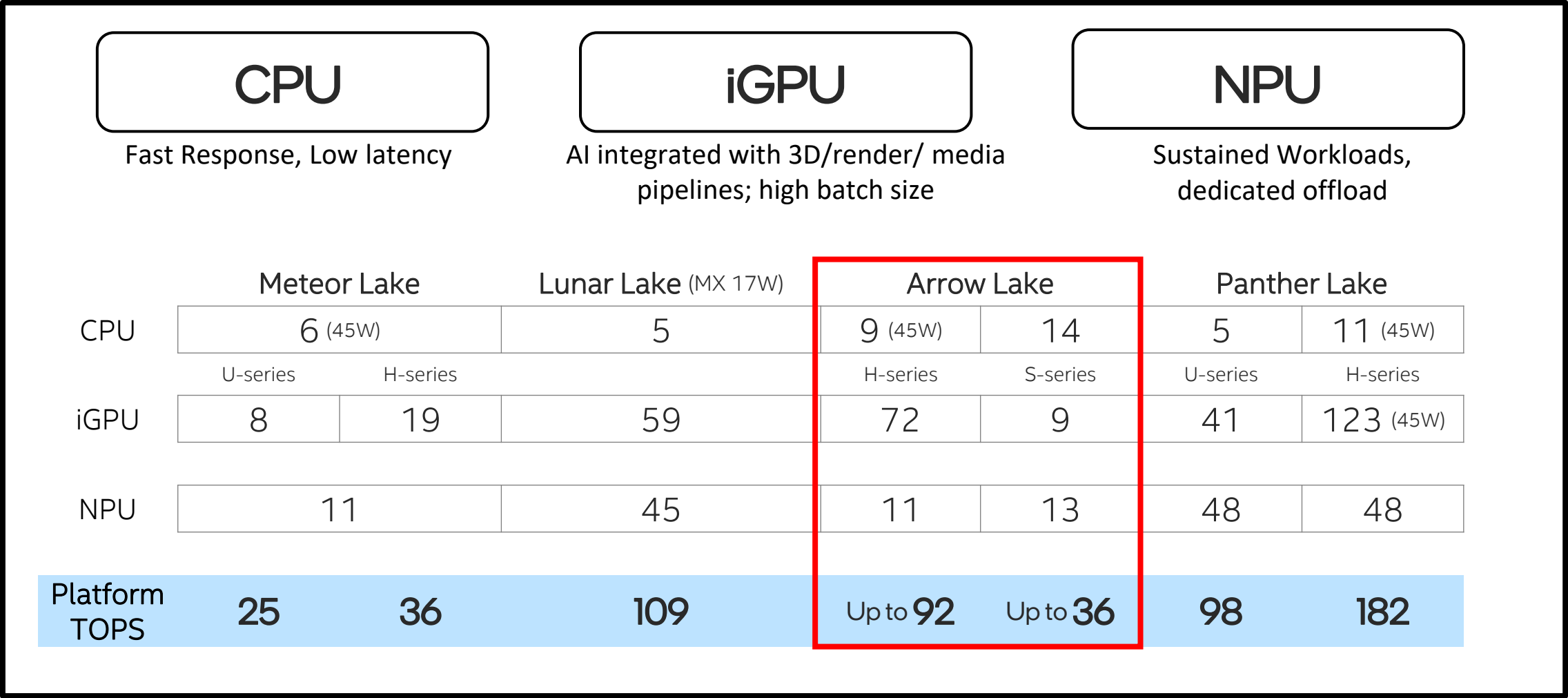
Microsoft\* believes that neural processing units, like Intel's NPU, represent an inflection point in computing and will be key to delivering a whole new range of delightful experiences to Windows\* users on their PCs. These experiences will transform how people use their computers and connect with others.



## Launching New Client Compute Capacity at Scale





# 8Q Client AI Roadmap




# 8Q Client AI Roadmap (Cont.)

Client AI Workloads are Diverse  
No Single Compute Unit Meets All Key Needs

 Bursty, Latency sensitive

 Sustained, Battery life sensitive

 Periodic, Throughput sensitive

HW	Value	RPL	MTL	ARL	LNL MX	PTL
CPU	SW Programmability; low latency, single inference tasks	AVX-256 VNNI H: 4-5 TOPS	AVX-256 VNNI H: ~3-6; U: ~2-3 TOPS	AVX-256 VNNI H: ~7-9.5; S: 14 TOPS	AVX-256 VNNI ~2-5 TOPS	AVX2+ TOPS - H: Up to 11; U: 5
iGPU	AI integrated with 3D/render/media pipelines; high batch size	DP4A H/U: up to 9 TOPS S/HX: 3 TOPS	DP4a(U, H) H: up to 19 TOPS U: up to 8 TOPS	DP4a(U, S, HX) ~9 TOPS  ARL H w/Xe Matrix Extensions (XMX) Up to 72 TOPS	DP4a + Xe Matrix Extensions (XMX) Up to 59 TOPS	DP4a + XMX H: Up to ~123 TOPS U: up to 41 TOPS
iNPU	Dedicated AI Offload, Power efficiency for Battery Life	NA	NPU 2.7 TOPS - H: 11 TOPS; U: 9.5-11; ARL S, HX: 13		NPU 4.0 Up to 45 TOPS	NPU 5.0 Up to 48 TOPS

TOPS will vary slightly based on power & frequency of each sku

## The Right Frameworks for Innovation and Scale:



ONNX



WebAssembly  
WebGPU  
WebNN



# What is NPU?

# Arrow Lake: Neural Processor Unit

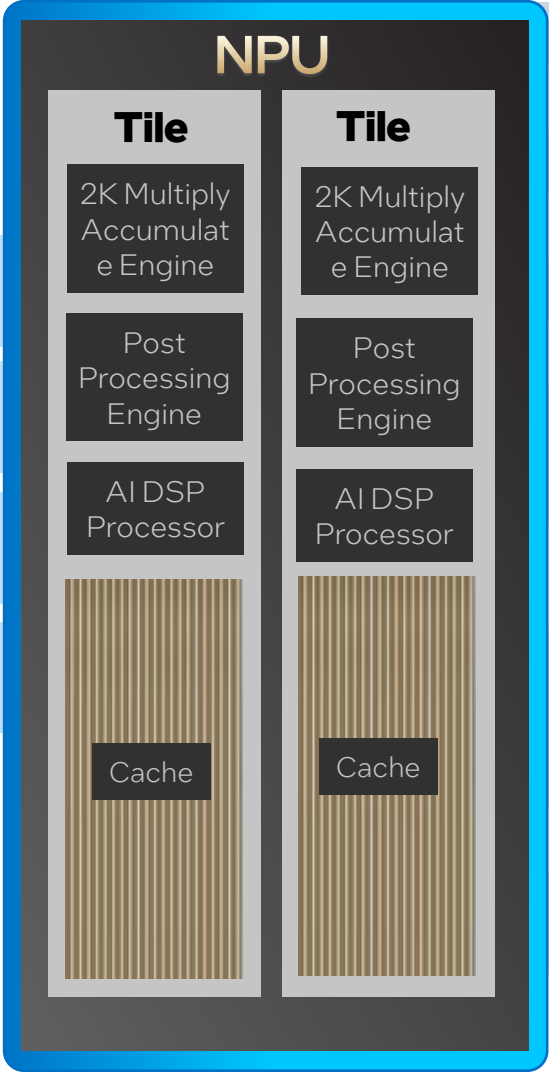
## Power Efficient AI

Fast, Ultra Low Power Inferencing

Improve System and App Responsiveness

Reduce Memory I/O Usage

Drivers for Windows and Linux

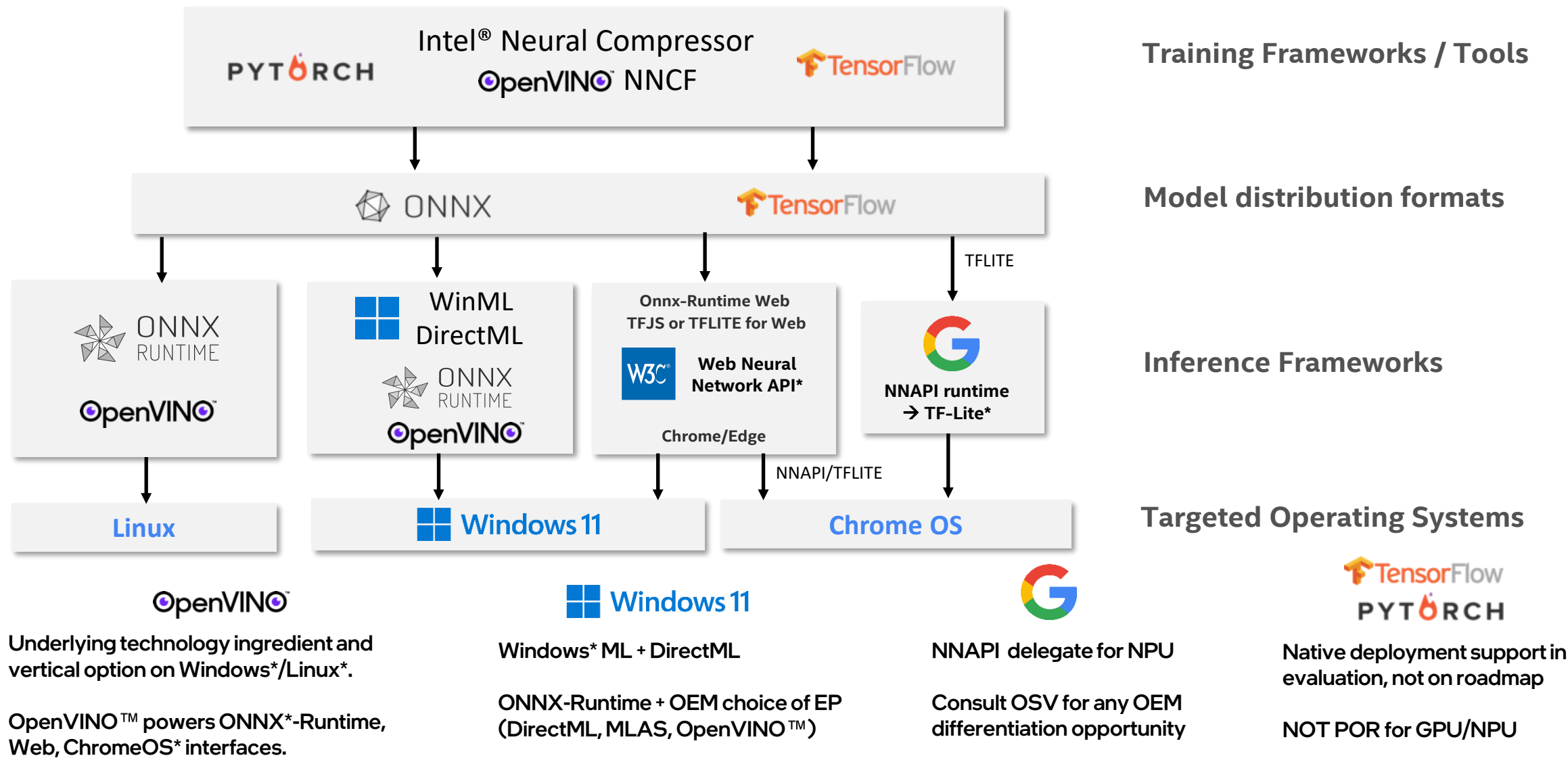


Performance	Up to 13 TOPs (int8)
MAC Engine	4096 (INT8) with FP32 Accumulator
Local Cache	4096 KB Software-Managed
Programmable DSP	VLW supporting Integer, FP, Transcendental
Peak Memory Interface BW	64 GBps, Unified Memory Architecture
Internal Data Type Support	INT8, FP16, BF16, FP32 (emulated)
Hardware Compression	Quantized Data Type Support Fine-Grain Weight and Activation Sparsity Weight Compression
MAC Fixed Function support	General Matrix-Matrix, Matrix-Vector Convolution, Fully Connected, Reshape
Elementwise Fixed Function support	ReLU/PReLU Add/Mul Quantize/Dequantize Reshape
OS Supported	Windows* OS, Chrome*, Linux*
Runtime Framework Support	OpenVINO™ Toolkit, ONNX RT, WinML/DirectML, WebNN

1. At Vmax in 15W MTL/ARL workload. Peak TOPs 13 at 1.6 GHz ResNet50, Int8, BS1, 50% sparsity
2. See backup for workloads and configurations. Results may vary.

# Software Frameworks for Innovation and Scale

## Embracing and Enabling an Open Ecosystem



# NPU Value Prop

## Performance



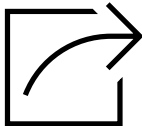
To run advanced, higher quality AI models for Richer Experiences

## Battery Life



Deliver premium AI based experiences without sacrificing battery life

## Responsiveness



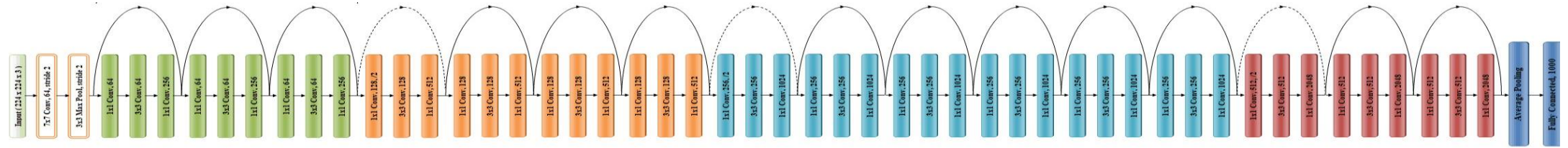
Free the CPU and iGPU for greater responsiveness & productivity

## Workload's Good for the NPU:

Sustained AI Workloads	Offload the CPU, iGPU, dGPU for responsiveness	Require High integrated TOPs
Model Characteristics: small Batch Size, FP 16/Int8, Sparsity enabled models		
Image, Video, Audio		

Client AI - Todd Matsler

# Resnet50 Example Based on MTL



Input 224x224

- Peak TOPS (pTOPS) = Peak Theoretical Max Performance  
 $\text{pTOPS} = \text{max frequency} * (\text{MAC/Clock}) * 2$

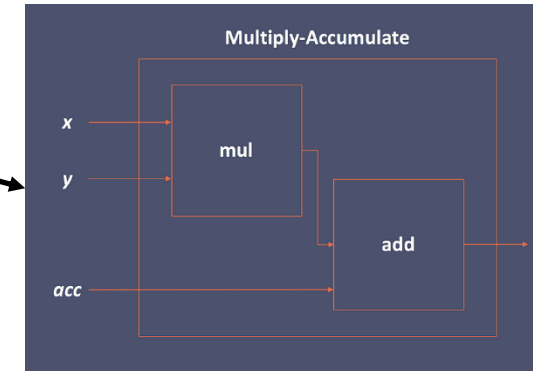
**NPU is 11 pTOPS**

- Effective TOPS (eTOPS) = Real Performance on a given AI Workload (the efficiency of pTOPS)  
 $\text{eTOPS} = (\text{fps} * \text{each frame GOPs}) / 1000$

**NPU 8.2 eTOPS = (1000 \* 8.216) / 1000**

- We use ResNet50: a common network + a good mix of a memory and compute bound network. Is it Perfect? -> No, but it is better than pTOPS as eTOPS shows real workload measured across many HW configs
- AI Benchmark for Client: Not 1 standard Today (UL Procyon Redowa (POR)/MLPerf/GeekBenchML)

Workload  
(MACs operations on NPU)



One multiply-accumulate is two operations

Output (frame)

- Operations per frame: constant value per network, for Resnet50 it is 8.216 GOPs

	pTOPS	ResNet50 <sup>3</sup> fps	eTOPS	Efficiency
Intel MTL iVPU	11	Dense: 715 <sup>1</sup>	5.9 <sup>1</sup>	53%
		Sparse: 895 <sup>1</sup>	7.3 <sup>1</sup>	67%
		Sparse: 1000 <sup>2</sup>	8.2 <sup>2</sup>	75%
QCOM 8cx Gen3	20-25 <sup>4</sup>	825	6.8	27-34%

<sup>1</sup>measured on early MTL Si & SW: B0 Si, pre-beta SW 2/3/2023

<sup>2</sup>with 50% sparsity enabled, estimated target for production Si and SW is ~1000fps & ~8.2 eTOPS

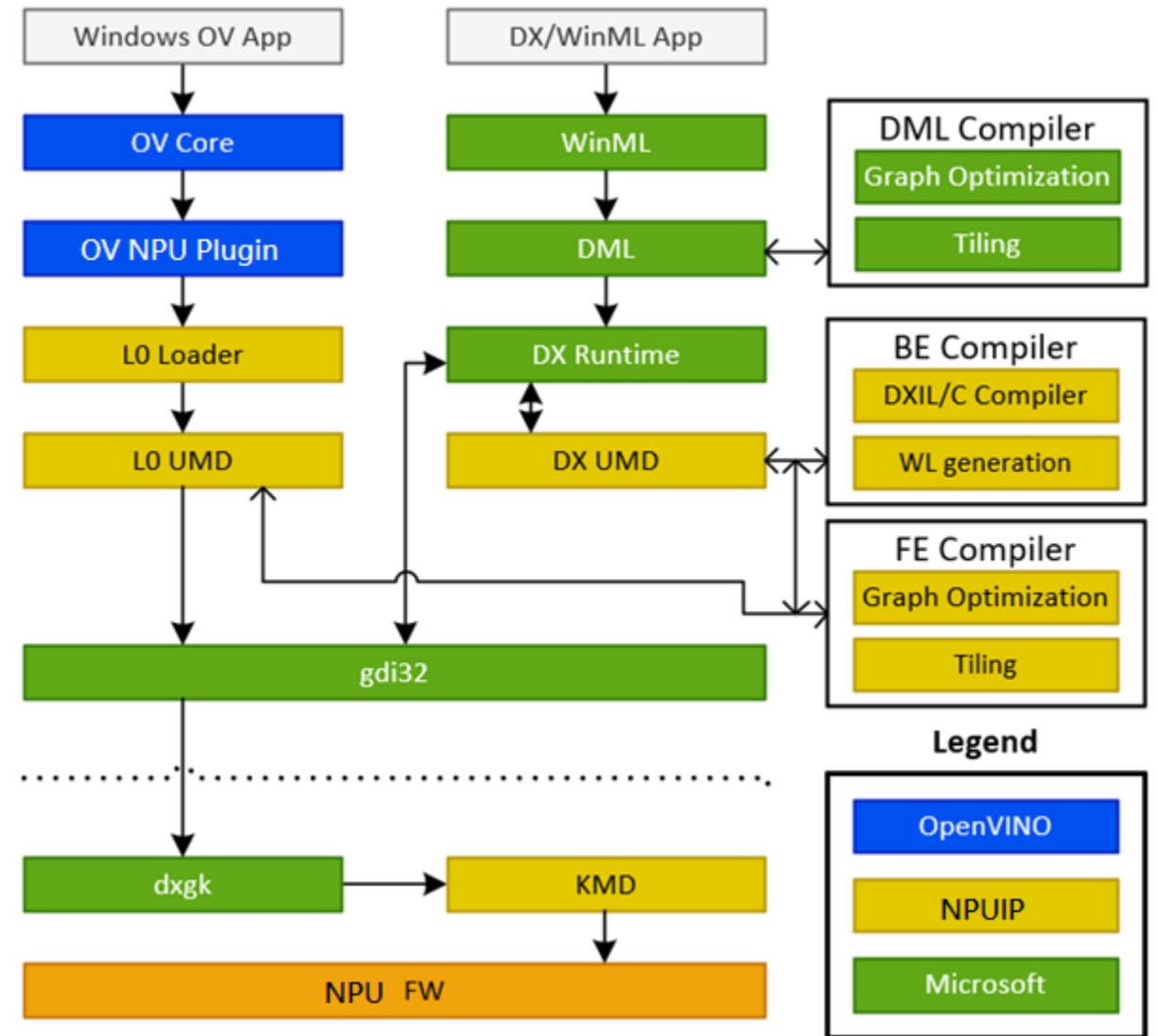
<sup>3</sup>MTL RN50: RN50 1.5 Open Model Zoo; QCOM RN50 version is UL Procyon AI Inference benchmark 2.4.0

<sup>4</sup>QCOM reports 29 pTOPS for 8cx Gen3 full SOC (CPU/GPU/NP); NPU only pTOPS estimated by Intel, based on current Intel internal analysis of available information

# NPU OpenVINO™ Plus DirectML Stack

- Unified driver architecture using Microsoft\* Compute Driver Model (MCDM)
- OpenVINO™ via Level 0 interface, WindowsML/DirectML via DX12
- DX12 UMD in NPU SW stack supports WindowsML/DirectML
- OpenVINO™ tools used to quantize/lower any ONNX\* model for NPU execution
- OpenVINO™ Apps compile & execute models using NPU Compiler tool chain & L0 NPU driver
- DirectML use DML compiler plus NPU compiler tool chain and DX12 UMD
- The NPU Driver package includes NPU FW and Compilers to support JIT Compile

## Windows\* SW Stack

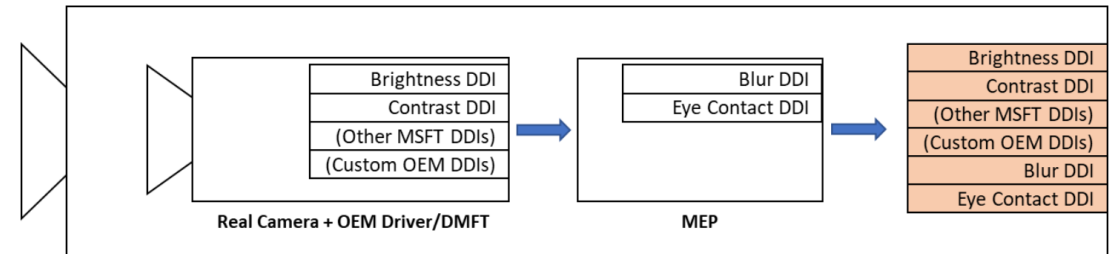


# What is MEP (Windows® Studio)?

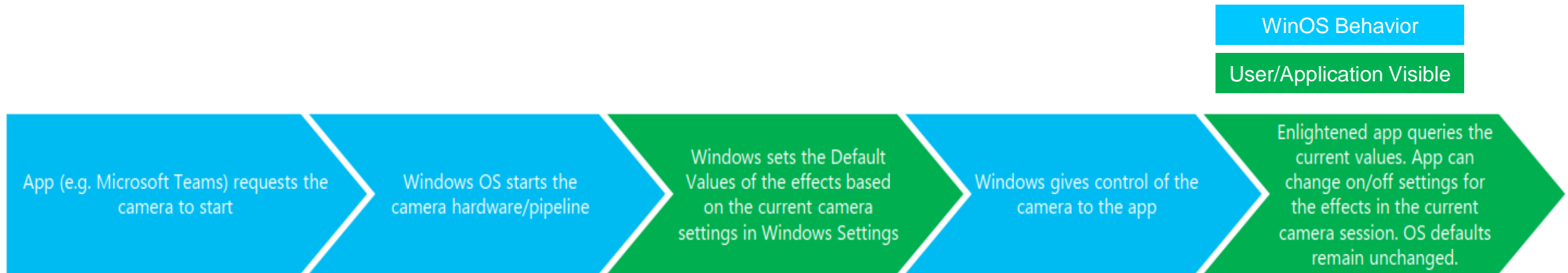
# What is MEP?

## Why Microsoft\* Effect Pack?

- MEP standardized control interfaces (Camera DDIs and APIs)
- MSFT provide consistent AI Models:
  - Optimized Algorithm for NPU
  - OEM/ISV apps can apply effects to any camera



### ■ MEP DDI Interface



- Example: Configurable MEP effects by Application

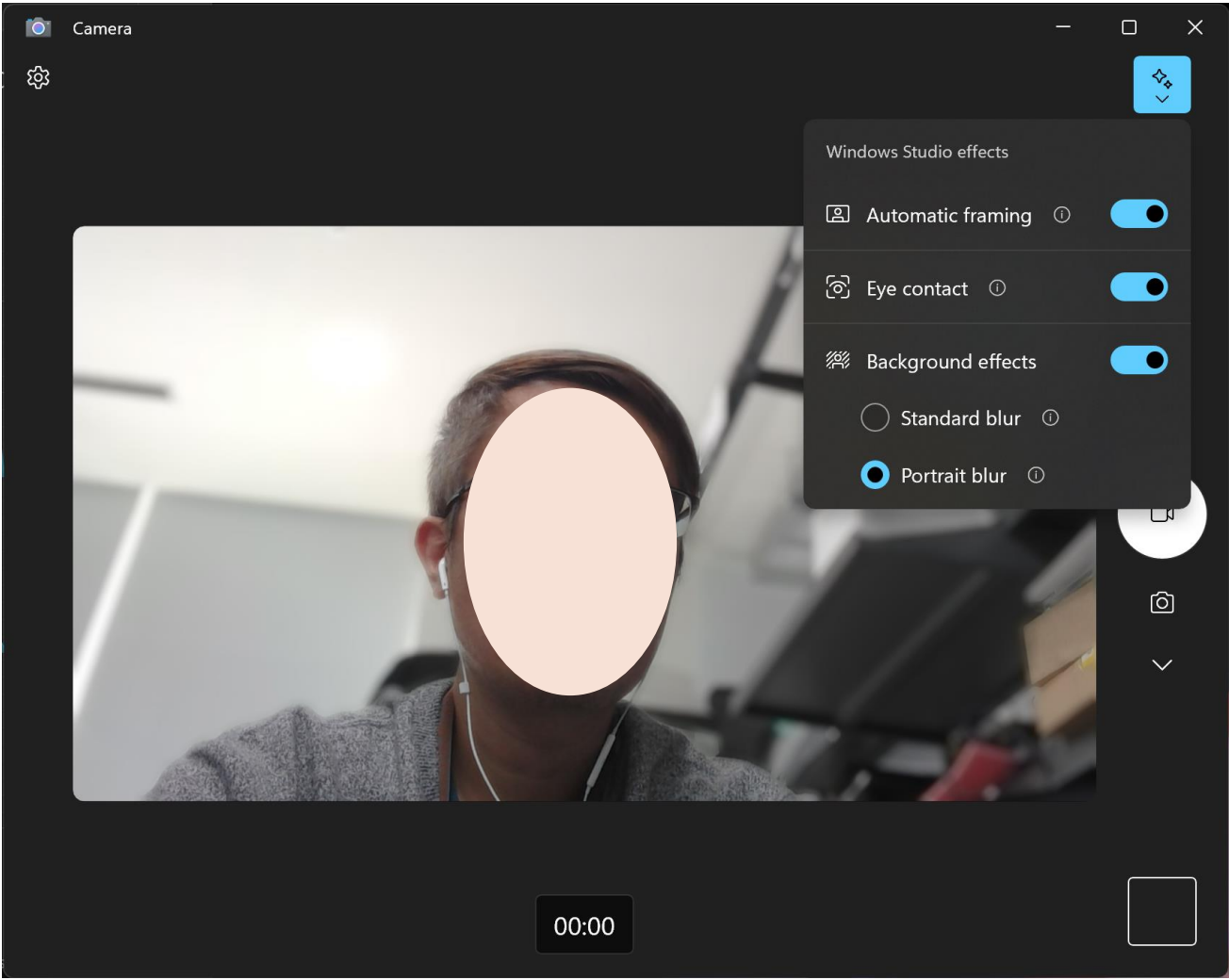
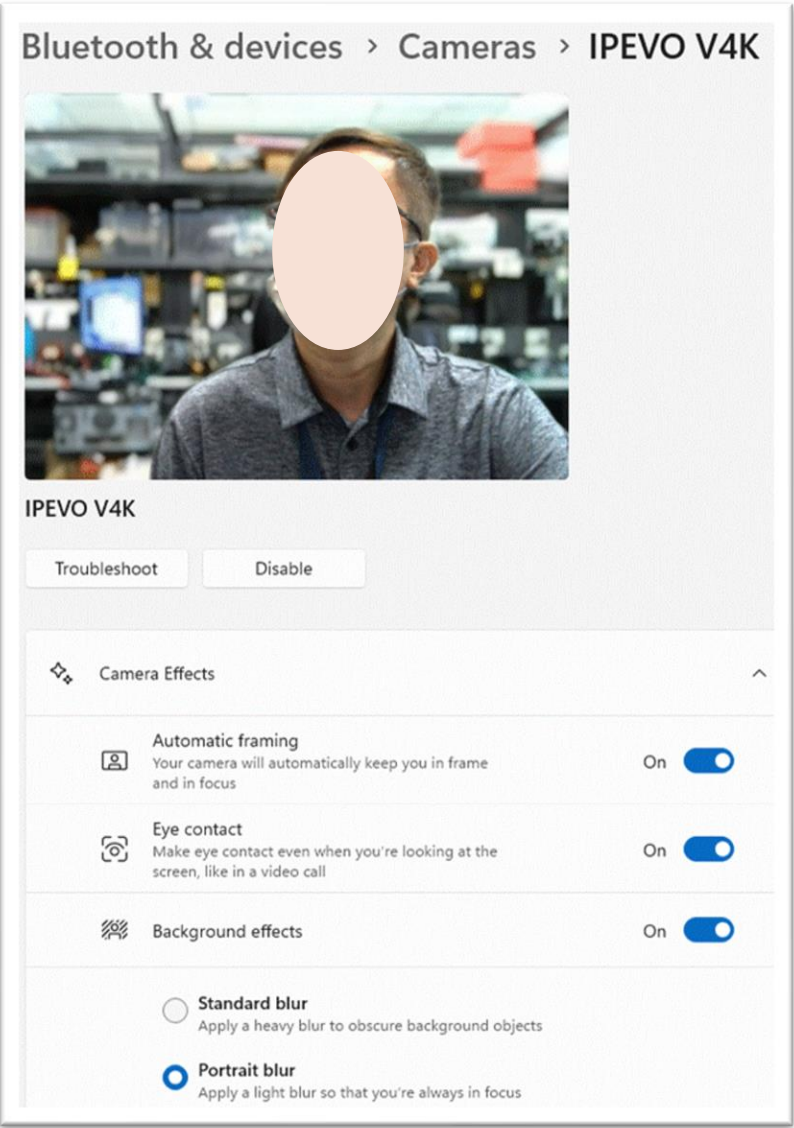


# Intel® NPU (NPU) Running Windows\* Studio (MEP)

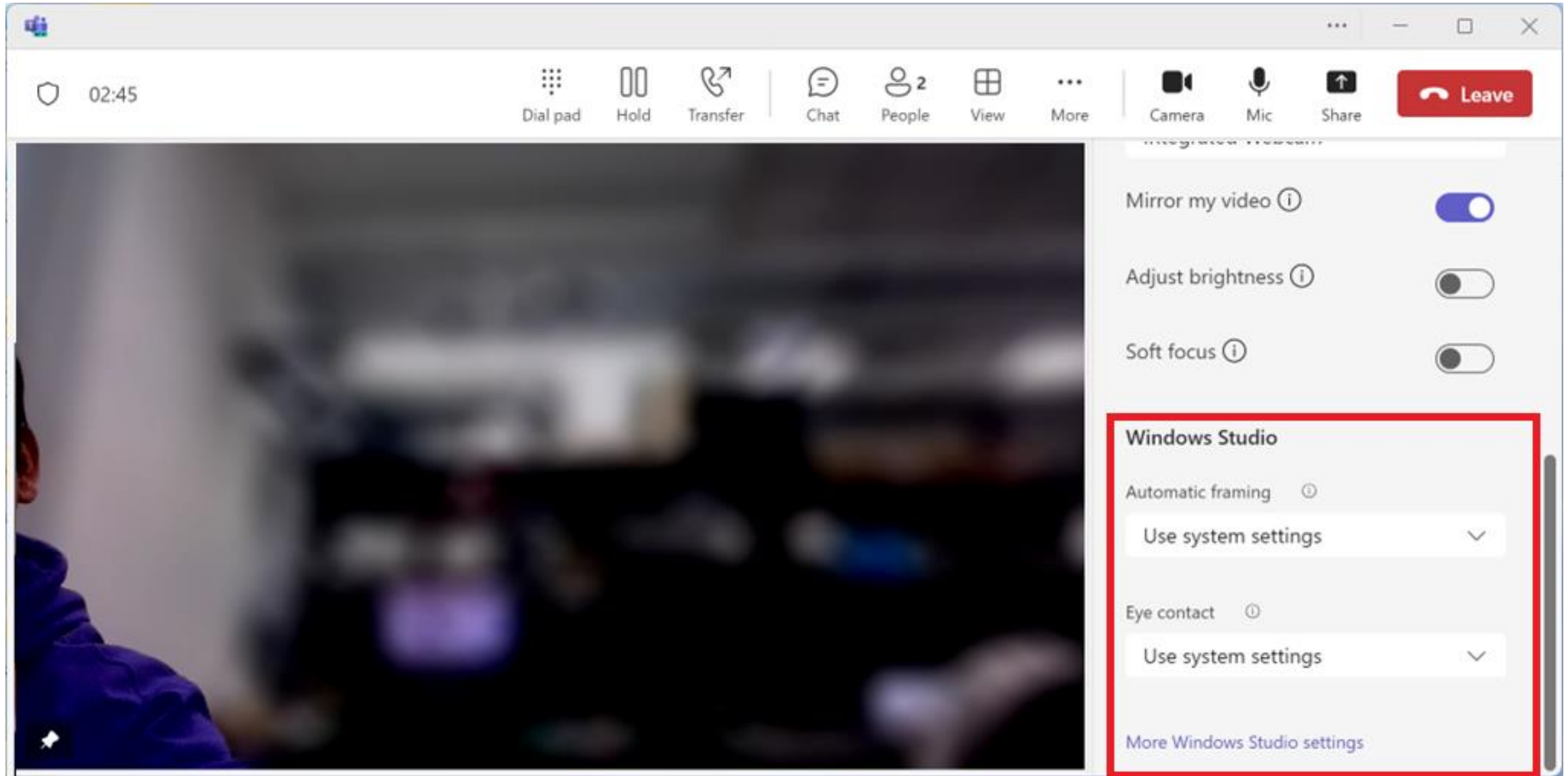
The screenshot displays three overlapping windows from a Windows 11 desktop environment:

- Device Manager:** The leftmost window shows the 'DESKTOP-K2PONTN' system. Under the 'Neural processors' category, 'Intel(R) AI Boost' is highlighted.
- Performance Tab:** The middle window shows the 'Performance' section. The 'NPU' section is active, displaying 'Intel(R) AI Boost' with a utilization of 15%. A graph shows the NPU usage over time. Below the graph, a table shows 'Shared memory usage' with columns for 'Utilization', 'Shared memory', and 'Driver version'. The utilization is 15%, shared memory is 0.0/3.8 GB, and the driver version is 31.0.0.0. Other system metrics like CPU (10% 2.06 GHz), Memory (6.7/7.6 GB 88%), Disk 0 (C:) (SSD 0%), Wi-Fi (S: 0 R: 0 Kbps), GPU 0 (Intel(R) Arc(TM) 13%), and NPU 0 (Intel(R) AI Boost 15%) are also visible.
- Windows Studio Effects:** The rightmost window shows the 'Integrated Webcam' settings. It includes a 'Troubleshoot' button and a 'Disable' button. Below, the 'Windows Studio Effects' section is expanded, showing settings for 'Automatic framing' (Off), 'Eye contact' (Off), and 'Background effects' (On).

# Windows\* Settings Camera Page and Inbox Camera Application
















# New TEAMS\* – Settings Page



# ARL NPU POR Features

# ARL NPU POR Features – MEP + APO

MEP (Windows* Studio Effects)	ARL NPU		APO Vendor
Background Blur	NPU		Realtek NPU 
Bokeh	NPU		Waves NPU 
Eye Contact Correction	NPU		Elevoc NPU 
Voice Focus*	CPU		Dolby NPU 
Auto Framing	NPU		Fortemedia NPU 
Voice access/ Live caption	NPU		Intelligo NPU 
More.. (TBD)	NPU		



**Notes:**

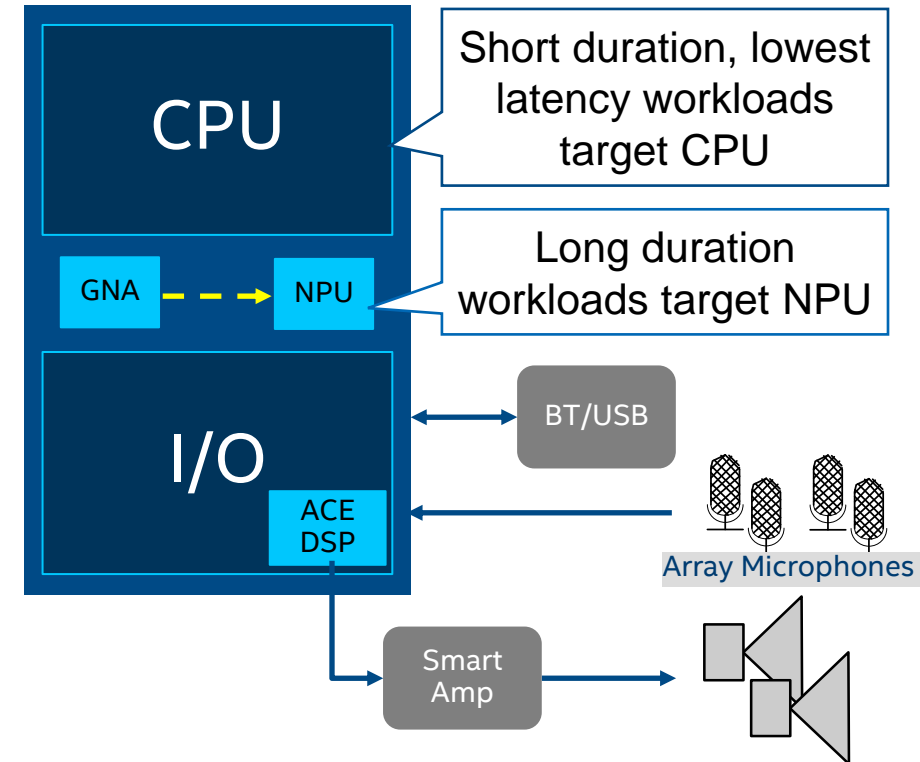
- Intel will work with Microsoft\* to evaluate future capabilities for ARL and beyond
- Customers should contact Microsoft\* to discuss MEP feature roadmap
- APO depends on OEM choice and optimization with Intel
- For “Voice Focus”, it is CPU only for now. NPU is TBD.

# ARL APO

# Audio AI Offload Transitions to NPU

TGL-MTL platforms include GNA for offloading Dynamic Noise Suppression (DNS) from CPU.

- DNS and other audio AI workloads (example: ASR) will migrate to NPU
- Migration starts on MTL, completes on LNL (no GNA)
- Most SO audio processing will run on either NPU or CPU
- Post processing of audio playback runs on a DSP

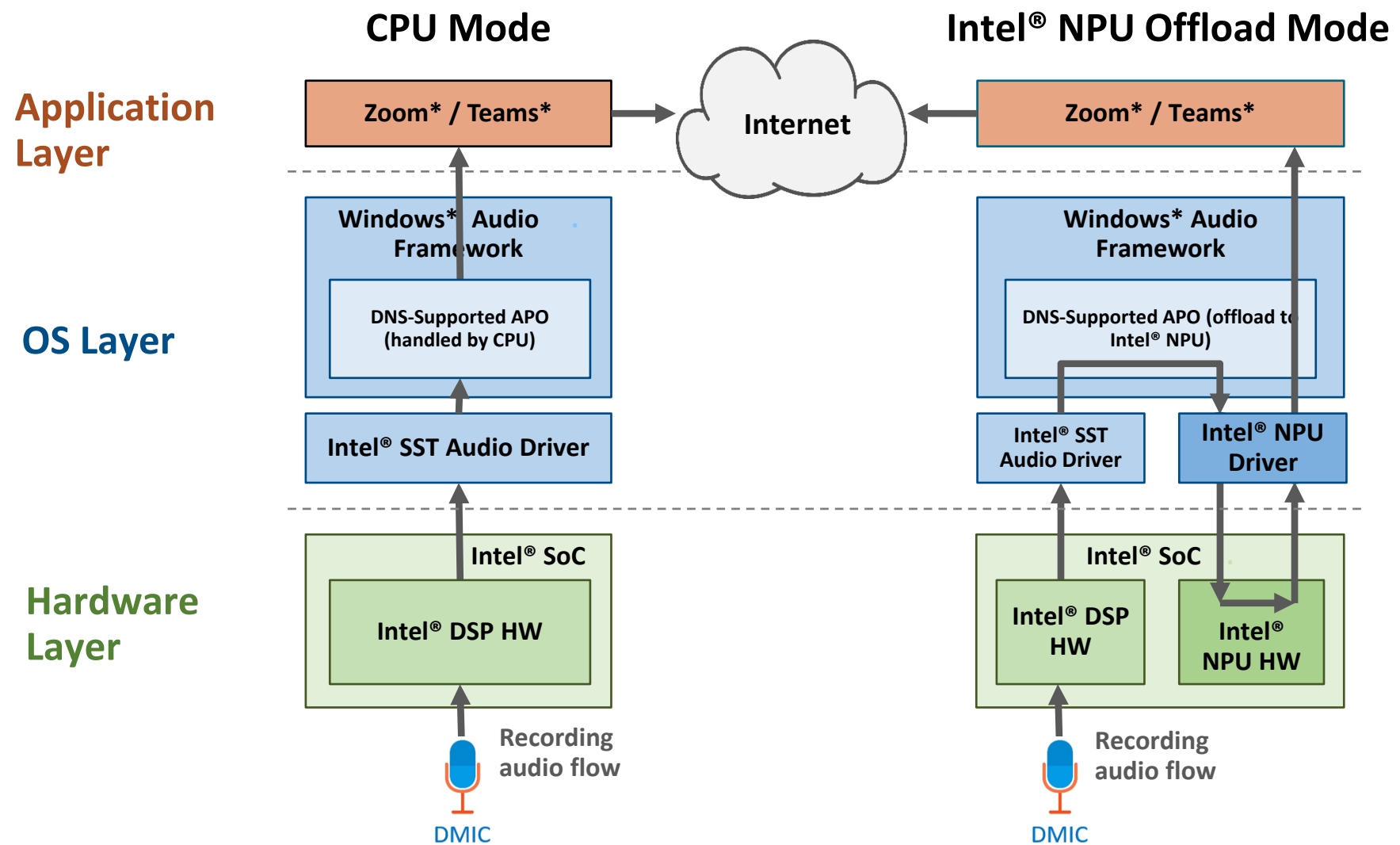


# APO – Audio Processing Object

- Windows\* allow OEMs and third-party audio hardware manufacturers to include custom digital signal processing effects as part of their audio driver's value-added features. These effects are packaged as user-mode system effect Audio Processing Objects (APOs).
- Audio processing objects (APOs), provide software based digital signal processing for Windows\* audio streams. An APO is a COM host object that contains an algorithm that is written to provide a specific Digital Signal Processing (DSP) effect.
- Examples of APOs include graphic equalizers, reverb, tremolo, Acoustic Echo Cancellation (AEC) and Automatic Gain Control (AGC). APOs are COM-based, real-time, in-process objects.

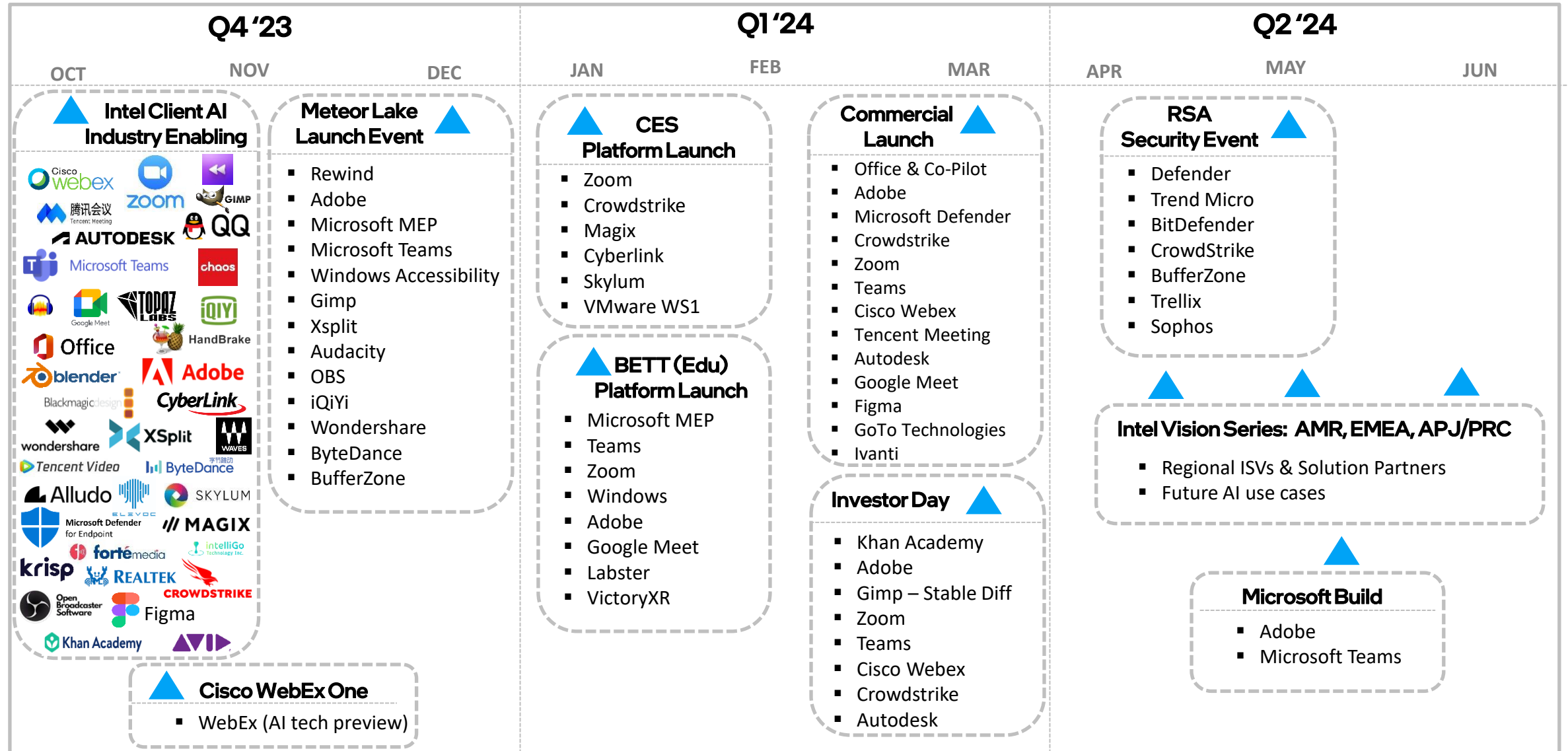


# Audio Flow with Intel® NPU DNS



# NPU Software Enabling and Experience

# Meteor Lake ISV AI Moments



# NPU Software Enabling

- Microsoft\* Collaboration:

- Windows\* Studio Effects, OS Accessibility, and New OS Experiences
- 1st party App AI experiences: Office and Teams
- Co-engineering DirectML for NPU – for broader scale
  - Supports ONNX Runtime DML-EP

- Industry-standard Software Framework Support for Broad, Open ISV Application Ecosystem

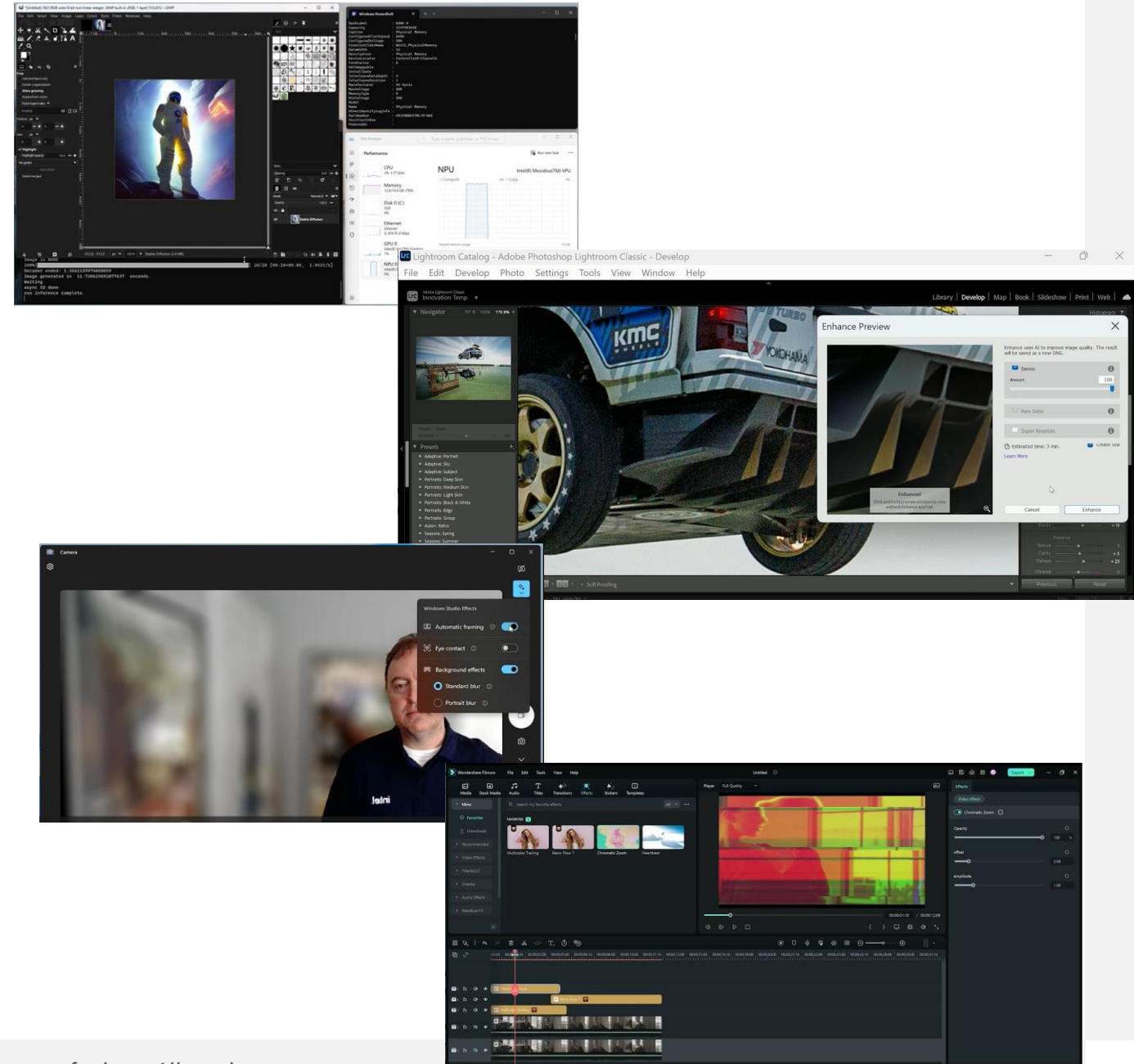
- Enablement of Select OEM Proprietary Models



“Microsoft believes that neural processing units, like Intel’s NPU, represent an inflection point in computing and will be key to delivering a whole new range of delightful experiences to Windows users on their PCs. These experiences will transform how people use their computers and connect with others. We are closely partnering with Intel on NPU and are excited to share more soon.” - Vivek Pradeep, Partner Research Manager, Microsoft

# AI Workload Briefcase

- Visit [AI Workload Briefcase](#),
- Videos Reference:
  - [GIMP with Stable Diffusion](#)
  - [Adobe Lightroom Classic AI Photo Editing](#)
  - [AI Enhanced Collaboration with Windows Studio Effects](#)
  - [Wondershare Filmora: AI Video Editing](#)
- Workload Assets:
  - [GIMP with Stable Diffusion](#)
  - [Adobe Lightroom Classic AI Photo Editing](#)
  - [XSplit VCam NPU Background Segmentation](#)
- [Etc.](#)



# Enhanced Collaboration Experiences

## New and improved features

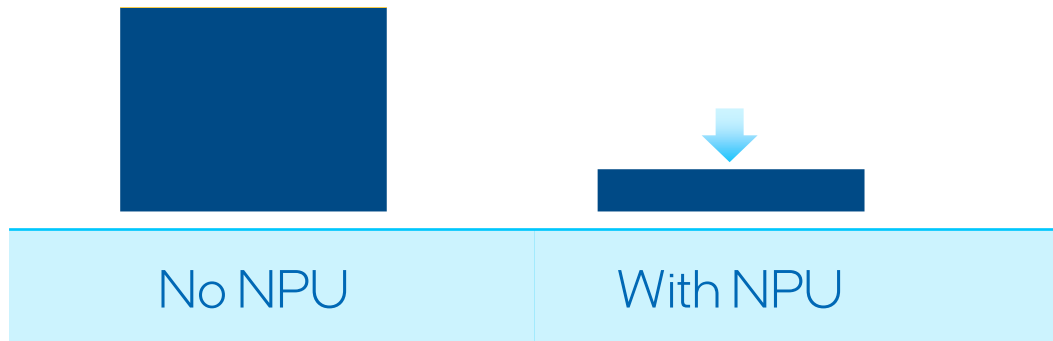
### New!

Auto-framing  
Eye contact  
Avatar representation  
Gesture recognition

### Improved!

Background Concealment  
Dynamic Noise Suppression

## CPU Workload



Advanced Blur

## Basic Blur



## Advanced Blur





# Generative AI Experiences

## Open-source GIMP plug-in for Stable Diffusion at Performance

5

10

20

30

40

50



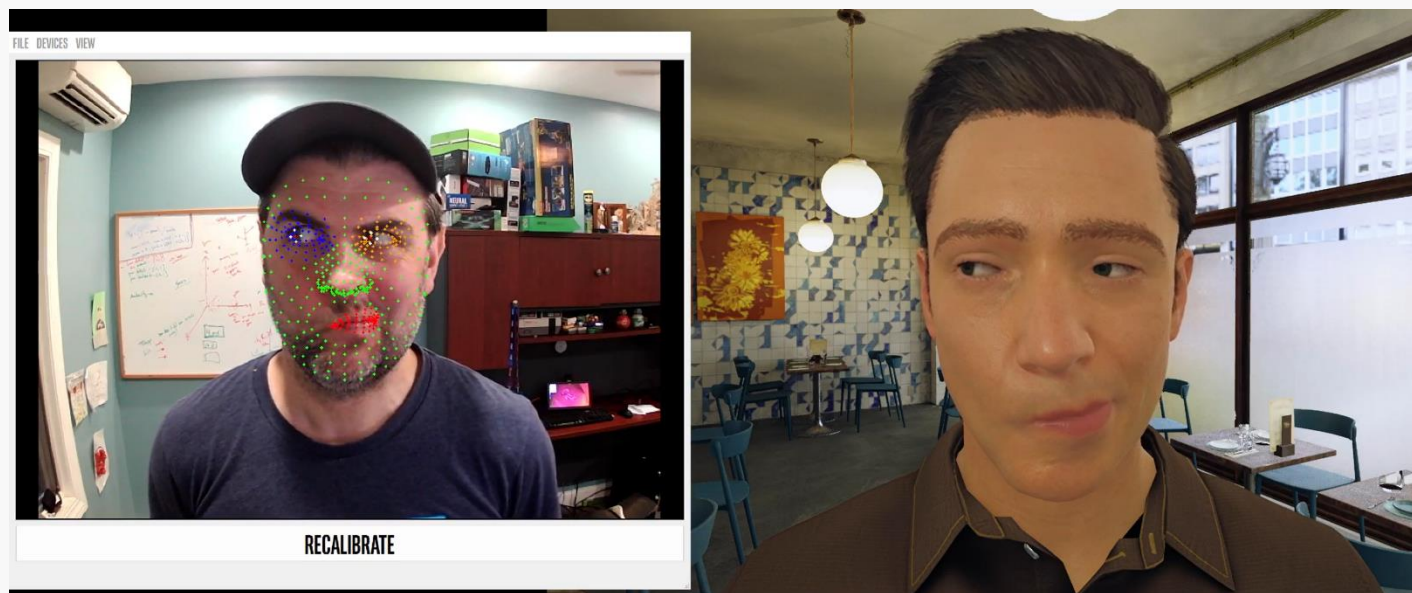
Text Prompt: cyborg man with a highly detailed, intricate details, carved by Michelangelo

# Seeding Open-Source Projects



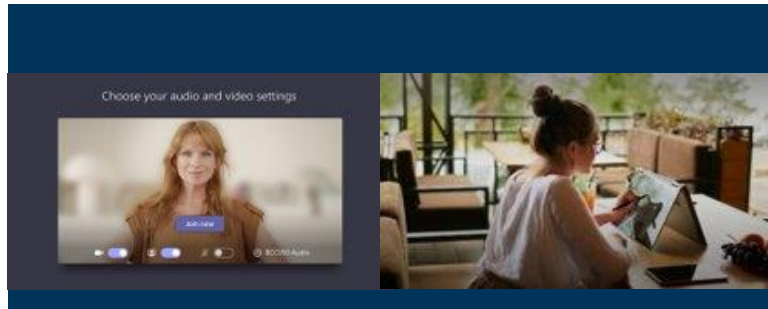
Realtime motion  
capture for  
Unreal Engine

Maintain/improve  
render  
performance on the  
GPU by offloading  
the AI to the NPU





# Transforming the PC Experience



## AI Today Enhancements

Elevated video collaboration & streaming  
Enhanced Audio effects  
Creator and Gaming effects

## Cloud

Massive scalable compute  
High Latency  
Privacy Concerns  
Expensive



## AI Tomorrow Everything

AI Assistants know your daily context  
More creative, productive, & collaborative  
Across everything you do

## Client

Massive distributed scale  
Low Latency  
Improved Privacy  
Lower Cost (to ISV)

