Microsoft Surface

Surface and AI

AI has **already arrived**

AI adoption is no longer an emerging trend. Data and analytics decision-makers are already building AI technologies and 74% are seeing a positive impact in their organizations.[1]

It's no longer about whether to adopt AI— It's about *preparing* for its impact

"(AI is) deeply relevant to how work gets done. Leaders see a long tail of potential benefits from AI, nearly half of which accrue to internal processes for how work gets done or to the workforce itself."[2]
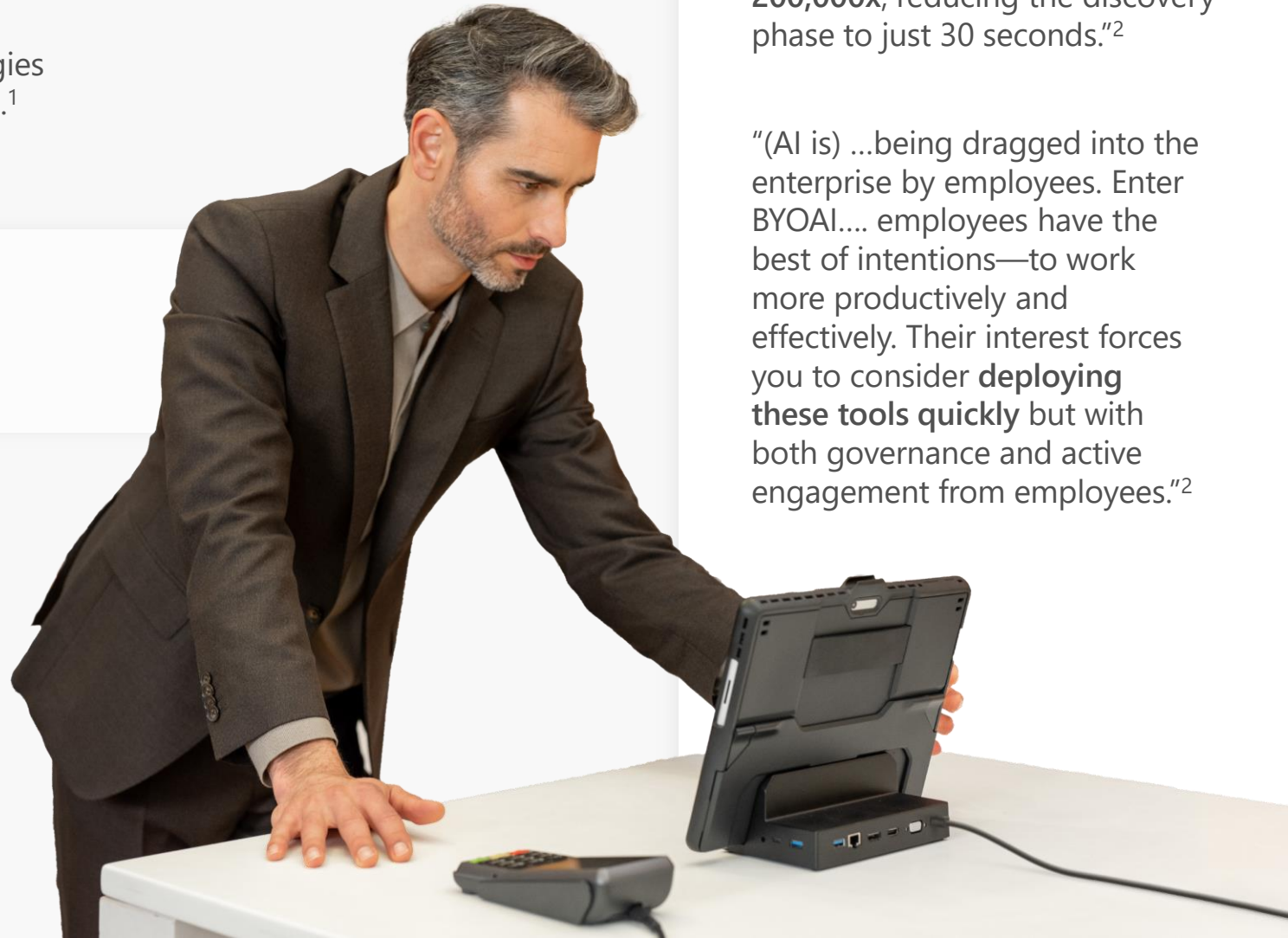
"(Using AI,) Dow reduced its **two- to three- month-long** product development process for polyurethane formulations **by 200,000x**, reducing the discovery phase to just 30 seconds."[2]

"(AI is) ...being dragged into the enterprise by employees. Enter BYOAI.... employees have the best of intentions—to work more productively and effectively. Their interest forces you to consider **deploying these tools quickly** but with both governance and active engagement from employees."[2]

AI is more than cloud computing

AI analytics and modeling require vast amounts of data, which are best suited for cloud, but performing some workloads at the device level can deliver more efficient processing, or *inferencing,* by being offloaded to the device. Efficiencies could be applied to:

- **Visual inferencing:** AI applied directly to the camera feed
- **Audio inferencing:** AI applied to audio inputs
- **Live transcription:** AI applied to language processing

Modern devices are being engineered with specialized processors to support executing those models – and others - locally, in real time.

AI shines on the endpoint in a couple ways:

- **Hardware-accelerated AI**
  AI workloads like Windows Studio Effects are processed through specialized processors like the NPU (Neural Processing Unit) for better performance
- **Cloud-delivered AI**
  AI workloads are processed in the cloud but utilize device features like touchscreens for enhanced experiences

AI is the capability of a computer program or a machine to **think, learn, and take actions** without being explicitly encoded with commands.

AI on device

Impacts end-user experiences like Windows Studio Effects or enhances device performance by executing AI models, locally, on real-world data
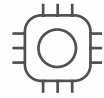
AI on cloud

Delivers scale advantages for data mining, analytics, and complex problem-solving for large, abstract data sets

A brief look at processors
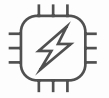
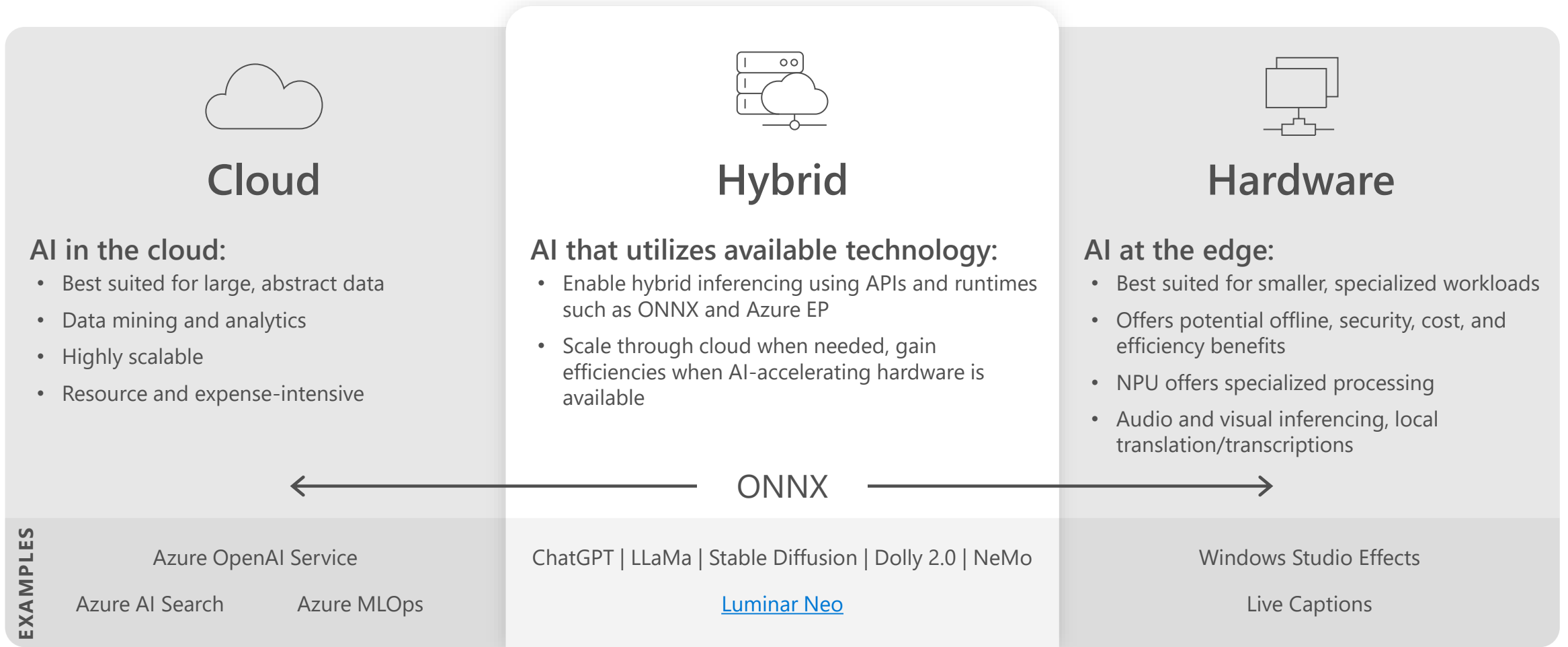|	**CPU** Central Processing Unit	**GPU** Graphics Processing Unit	**NPU** Neural Processing Unit
History	• Foundational to computing • Iterative updates in recent years	• Provided significant boosts in graphical and data processing • Similar iterative trajectory now turning interest towards AI	• New frontier in processing innovation • Exponential improvements in just a few years
Purpose	• Brain of the computer • Performs basic operations from software instructions, loaded from memory	• Specialized to render 2D and 3D objects • Can perform operations in parallel, processing vectors of data simultaneously	• Specific architecture for deep learning • Integrated as an element of the SoC • Hardwired matrix without need for memory access, reduced precision • Trained specifics, Inference operations
Pros	Can execute any line of software	Efficient for repeatable calculations	Audio, Video, Data inspection
Neg	Not efficient at specific operations	Floating Point, not needed for AI	Generic code execution

AI across the enterprise

Cloud

AI in the cloud:

- Best suited for large, abstract data
- Data mining and analytics
- Highly scalable
- Resource and expense-intensive

Hybrid

AI that utilizes available technology:

- Enable hybrid inferencing using APIs and runtimes such as ONNX and Azure EP
- Scale through cloud when needed, gain efficiencies when AI-accelerating hardware is available

Hardware

AI at the edge:

- Best suited for smaller, specialized workloads
- Offers potential offline, security, cost, and efficiency benefits
- NPU offers specialized processing
- Audio and visual inferencing, local translation/transcriptions

← ONNX →

EXAMPLES

Azure OpenAI Service

Azure AI Search   Azure MLOps

ChatGPT | LLaMa | Stable Diffusion | Dolly 2.0 | NeMo

Luminar Neo

Windows Studio Effects

Live Captions

Advantages of running AI models on Surface devices with GPUs and NPUs

Greater privacy: Running AI models locally on a device can provide greater privacy as data is processed on the device itself, without the need to send it to a third-party service for processing

Less latency: By running AI models locally, the latency associated with sending data to and receiving results from a third-party service may be eliminated, potentially resulting in faster processing times
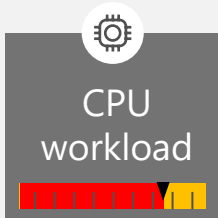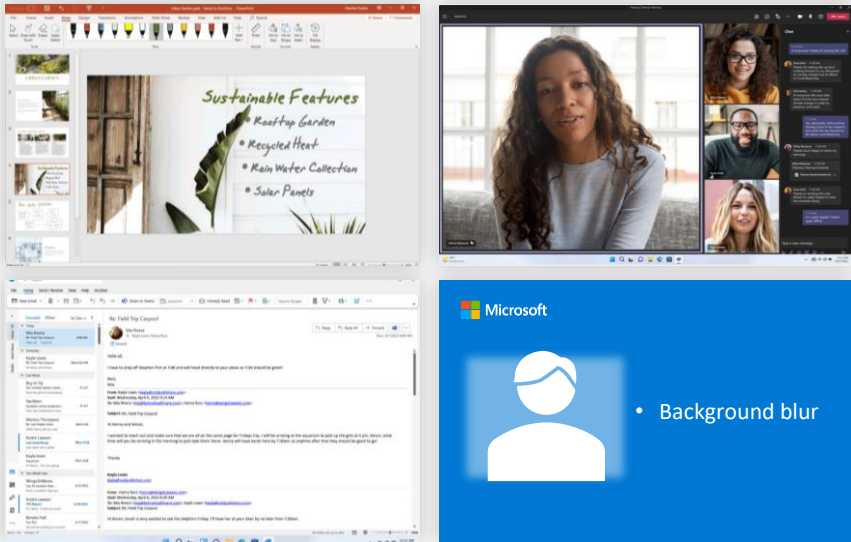
Cost savings: By running AI models locally, may reduce need to pay third-party services for their compute power, potentially resulting in cost savings

Efficient processing: Running AI models locally on a device can deliver more efficient processing by offloading workloads to the device. This can be applied to visual inferencing, audio inferencing, live translation, and transcription
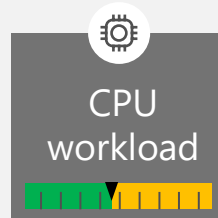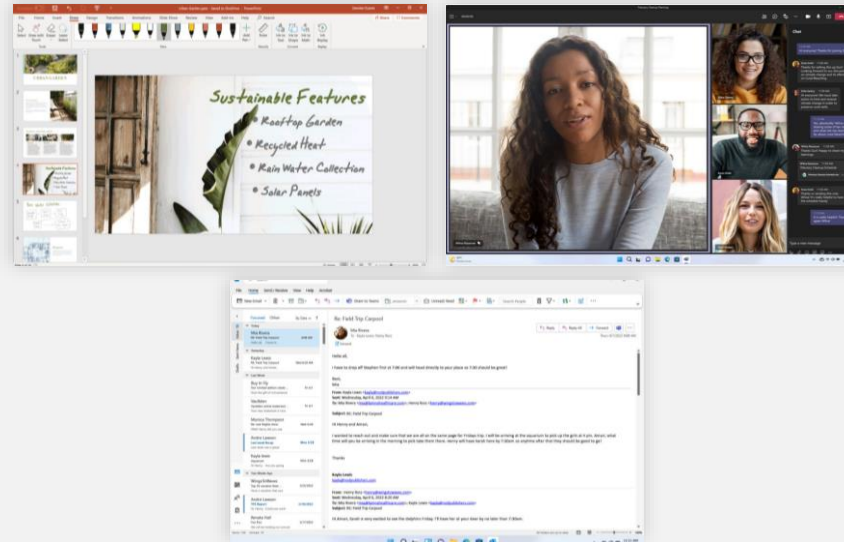
Free your CPU and gain powerful NPU-enabled AI experiences

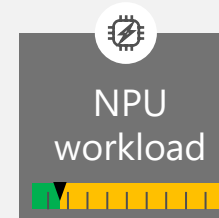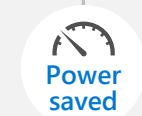# AI models will scale on the edge

As the cost of running AI increases, running smaller, specialized models locally on a device helps organizations realize economic scalability.

"We're literally going to have lots and lots of applications which will have local models and we'll have hybrid models, and I think that's the future of AI going forward."

-Satya Nadella

Leon Mallett • 2nd   2h •••
Chief Operating Officer at IAM Cloud

Why? LLMs have a fundamentally similar problem to proof of

"...Efficiency matters...the obvious solution to this is using appropriately-scaled AI for whatever problems you want it to help solve."

Clem Delangue 😊 • 2nd   + Follow •••
Co-founder & CEO at Hugging Face
23h • 🌐

"...most companies will realize that smaller, cheaper, more specialized models make more sense for 99% of AI use-cases..."

AI_Skeptic ✓   •••
@The_AI_Skeptic

The cost of running AI is becoming an issue, even in domains where people are prepared to pay for it:

Surface and the AI benefit

Employees

Today, AI on Surface Pro 9 with 5G and Surface Laptop Studio 2, enhances human-to-human communication with Windows Studio Effects for videoconferencing, and reduces latency for complex AI workloads while keeping data secure.
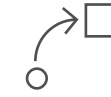
IT

Information processed through AI takes advantage of chip-to-cloud security, enabling organizations to maintain privacy and control of their data.

Business

Improve employee experiences now and be ready to take advantage of innovations to come. Microsoft's approach to Surface and AI ensures a commitment to performance and security.