UNIVERSIDAD DEL CEMA

JULIO 2026

# Hardware e IA

*La evolución del silicio para llevar la inferencia del servidor al cliente*

**Facundo Nicolás Eiroa · Nazareno Gastón Ferreyro**

Lenguajes Formales y Teoría de la Computación

Prof. Mario Samuel Moreno

EL PROBLEMA

# La IA no puede depender solo de la nube

Si cada consulta viaja a un servidor remoto, surgen cuatro problemas.

01

## Demora

Cada respuesta viaja lejos y regresa

02

## Privacidad

Los datos del usuario salen del dispositivo

03

## Costo

Mantener esa infraestructura es muy costoso

04

## Internet

Requiere conexión constante y veloz

HIPÓTESIS

# Para que la IA forme parte del uso cotidiano, cada equipo necesita su propio *chip dedicado a IA*.

De este modo no depende de internet ni de una placa de video costosa.

RECORRIDO

# Tres actos

| | | |
|---|---|---|
| 01 | El hardware | Los componentes que realizan los cálculos |
| 02 | Aceleradores de IA | Chips diseñados solo para IA |
| 03 | Ejemplos prácticos | Lo que ya se usa sin notarlo |

01

# El hardware

Los componentes que realizan los cálculos — y por qué la IA los lleva al límite.

# CPU: hace una cosa a la vez

- El procesador de propósito general: resuelve cualquier tarea
- Pero cuenta con pocos núcleos
- Millones de cálculos simultáneos lo saturan

NÚCLEOS

**decenas**

**Procesa los cálculos en fila**

NÚCLEOS

miles

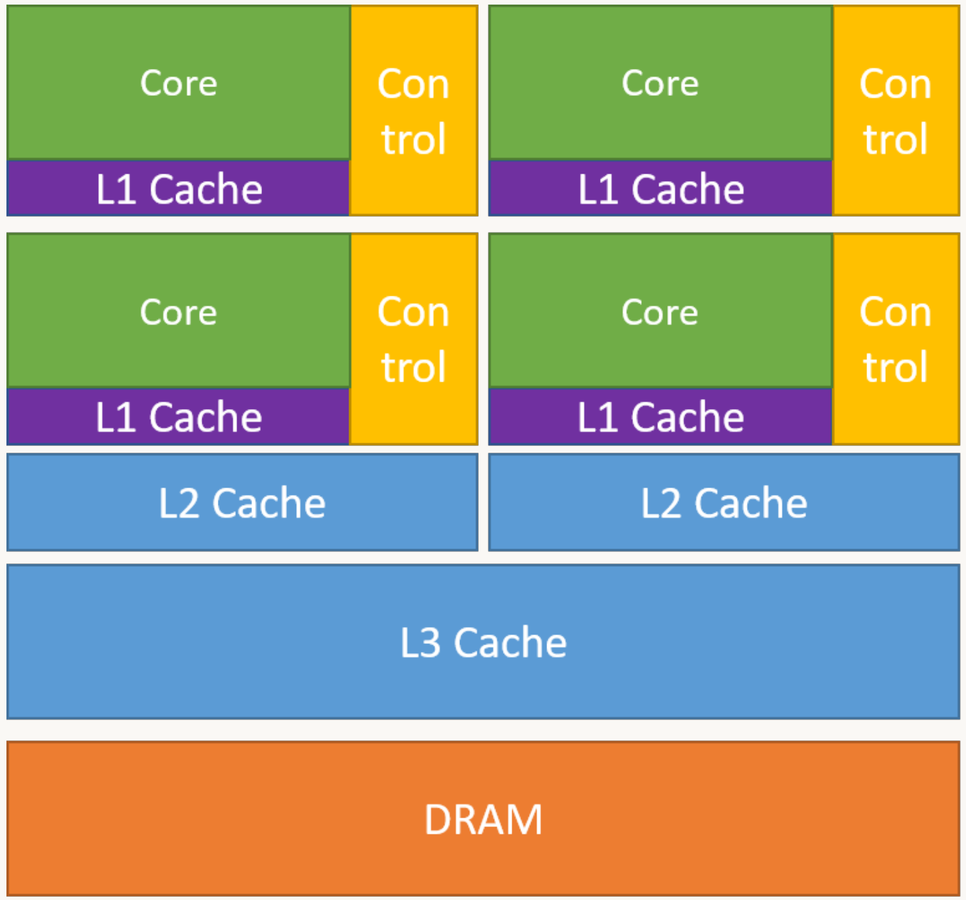
**Todos en simultáneo**

# GPU: miles de cuentas a la vez

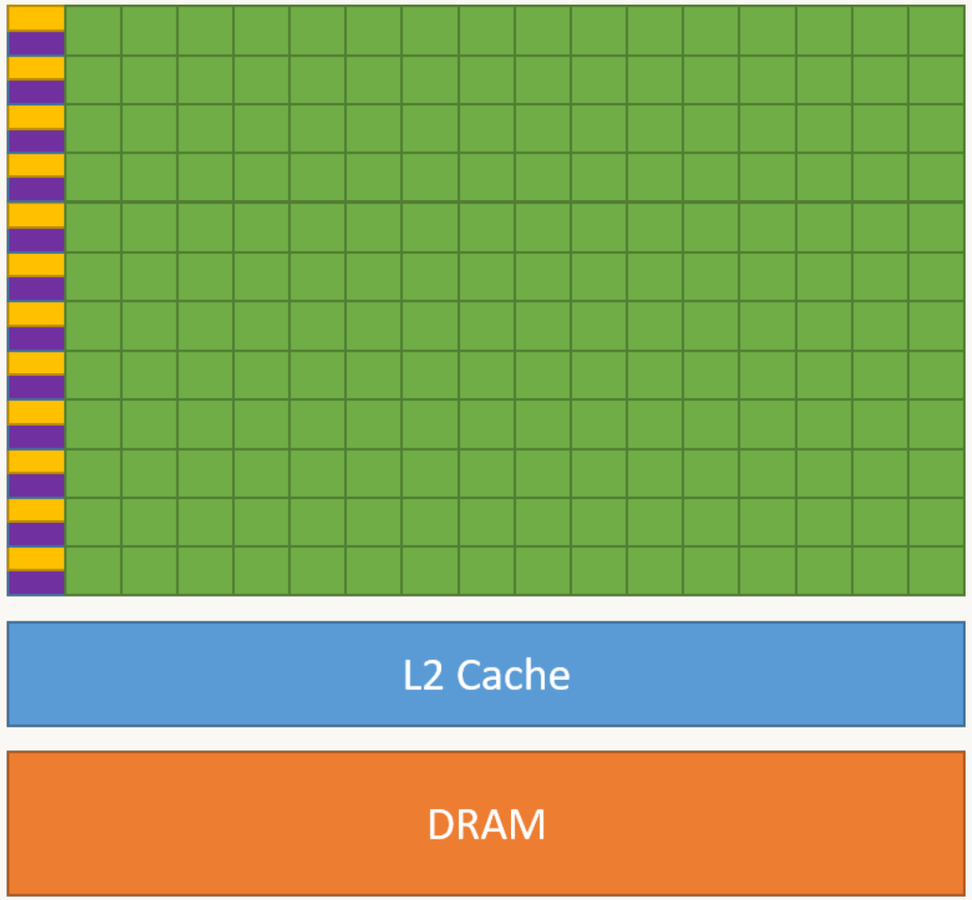
- Diseñada para los videojuegos y los gráficos
- Cuenta con miles de núcleos simples
- Justo lo que la IA necesita: muchos cálculos en paralelo

# Dos formas de repartir el trabajo

Verde: cálculo · Naranja: control Azul: memoria




CPU




GPU

Fuente: NVIDIA / Patterson, Computer Organization and Design

01 · EL HARDWARE

# Cómo NVIDIA ordena esos miles de núcleos

- Cada núcleo es muy simple, pero hay miles
- Operan en **grupos de 32**, en simultáneo
- **CUDA**: las herramientas de NVIDIA que utiliza casi toda la IA

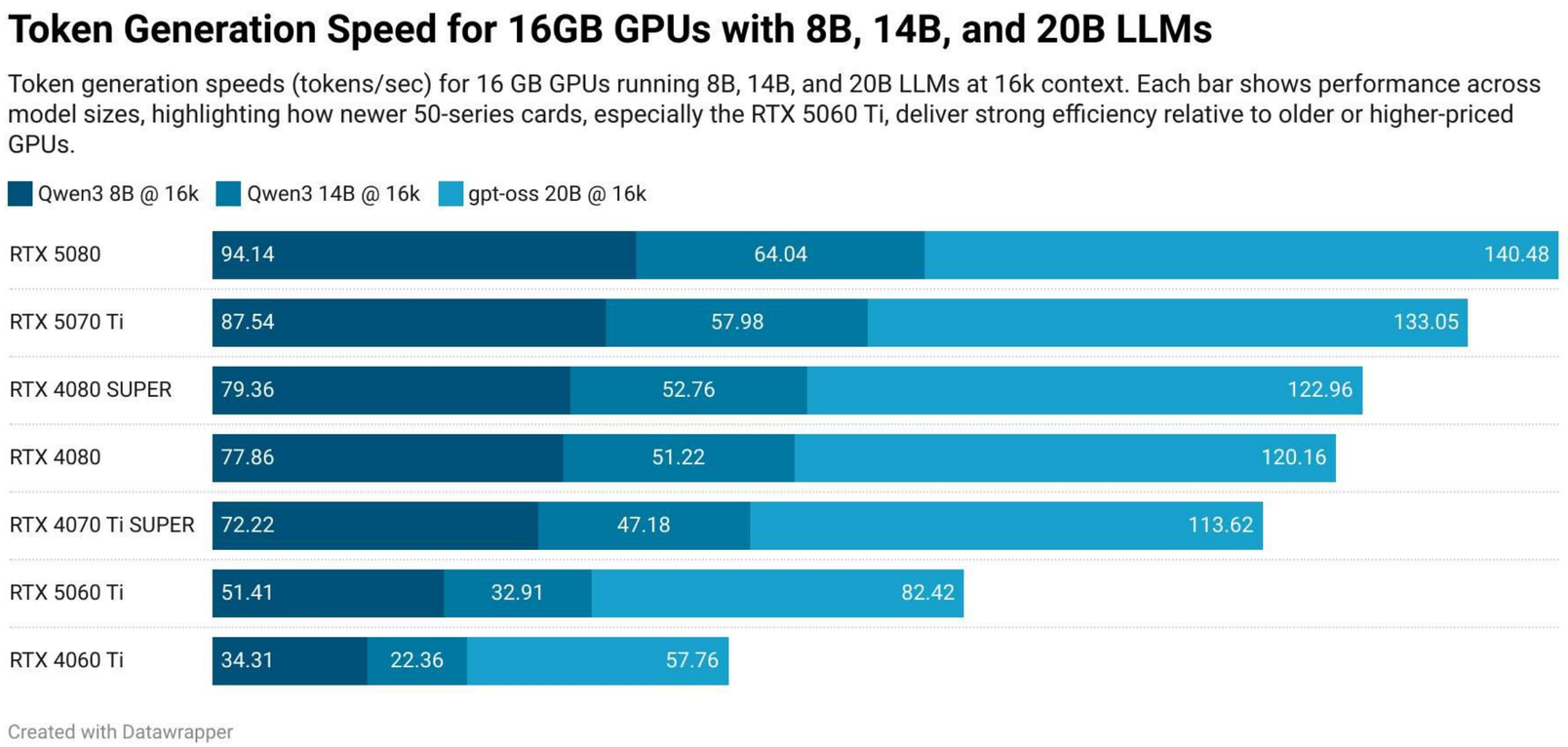
OPERAN EN GRUPO

## 32 a la vez = 1 grupo

Limitación: solo funciona con placas *NVIDIA*

# Una placa común ya ejecuta IA en casa

Palabras por segundo que genera cada modelo de IA




Fuente: Hardware Corner — Qwen 8B/14B y gpt-oss 20B @ 16k

# Pero esa placa tiene su costo

### Energía

Consume mucha electricidad

### Calor

Necesita refrigeración mas robusta

### Precio

Cara para el usuario común

*Se necesita algo más compacto y eficiente.*

02

# Aceleradores de IA

Chips diseñados para una sola cosa: la IA.

# NPU: el chip de IA de bajo consumo

- Un chip dedicado exclusivamente a tareas de IA
- Funciona en tiempo real con **muy poco consumo**
- Lleva la IA a equipos sin placa de video

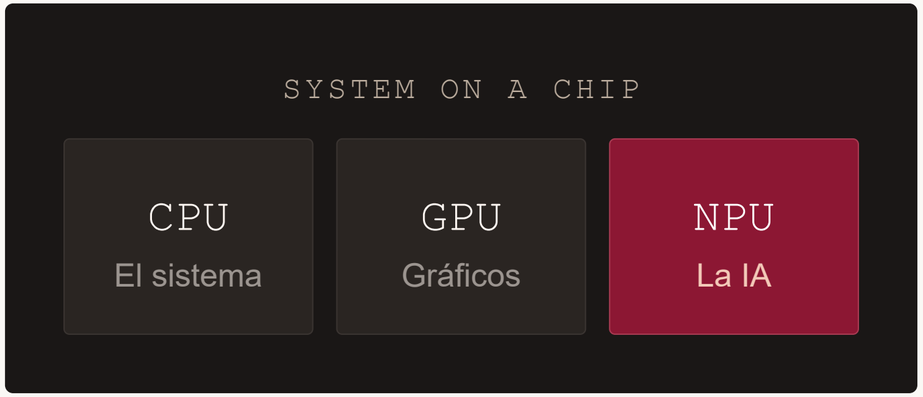
TAMBIÉN CONOCIDA COMO

*Acelerador de IA*

El chip especialista en IA

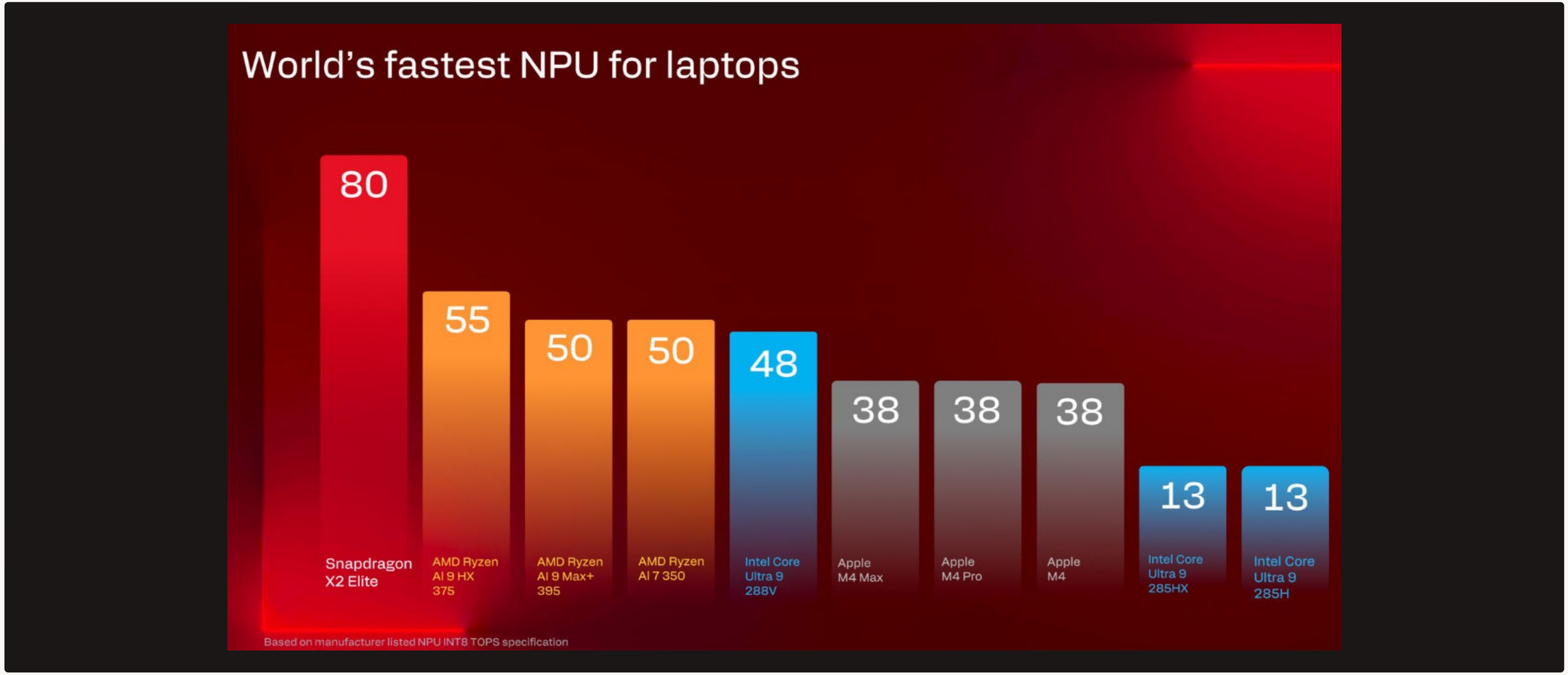
# Todo en un solo chip

- La NPU se integra junto a la CPU, en el mismo chip
- El sistema le delega el trabajo pesado de IA
- Evita encender la placa de video → más autonomía



# ¿Qué tan potente es cada NPU?

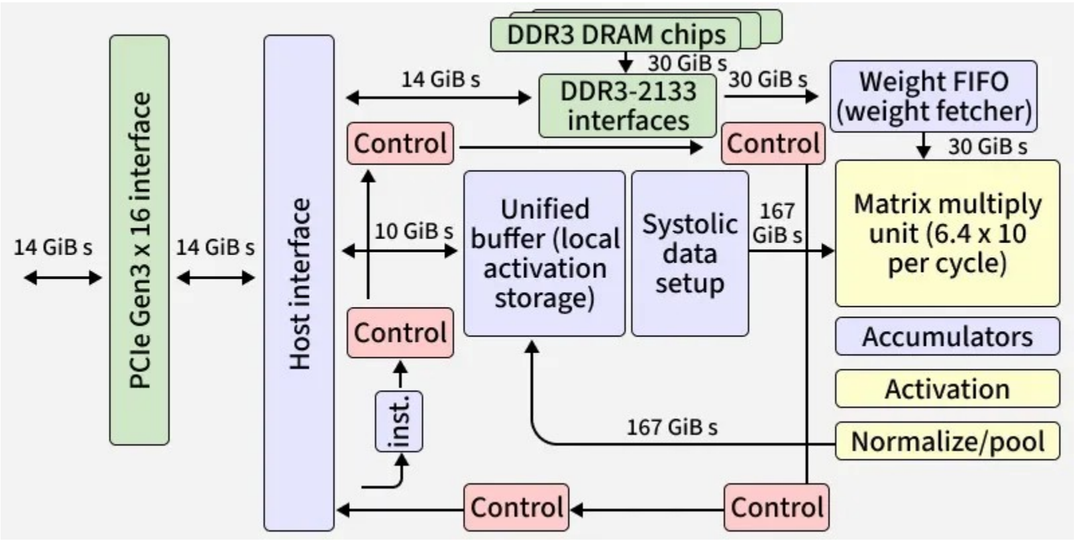
**TOPS** = operaciones por segundo. Más = mejor




Fuente: Qualcomm — Snapdragon X2 Elite, 80 TOPS

# TPU: el chip de IA de Google

- Diseñado a medida para los servidores de Google
- Su núcleo: una enorme grilla de unidades de cálculo
- Usa números más livianos: **duplica la velocidad**
- Memoria de alta velocidad, apilada junto al chip




Fuente: Geeks for Geeks — TPU v1

03

# Ejemplos prácticos

Aplicaciones que tu equipo ya realiza con IA, hoy.

03 · EJEMPLOS PRÁCTICOS

# Generación de frames

La IA genera cuadros adicionales → el juego se ve más fluido.

NVIDIA · DLSS

## Con red neuronal

Una IA entrenada genera los cuadros faltantes. Mayor calidad, casi sin errores.

AMD · FSR / AFMF

## Sin red neuronal

Compara el movimiento entre cuadros. Funciona en cualquier juego, pero con más errores.

03 · EJEMPLOS PRÁCTICOS

# La evolución de DLSS

RTX 40 · Ada

## DLSS 3

×2

Genera 1 cuadro por cada cuadro real.

→

RTX 50 · Blackwell

## DLSS 4

×4

Genera hasta 3 cuadros. Más rápido y con menos memoria.

→

2026 · CES

## DLSS 4.5

×6

Se ajusta automáticamente al monitor.

03 · EJEMPLOS PRÁCTICOS

# Fondo borroso en videollamadas

El fondo borroso de las videollamadas, resuelto cuadro por cuadro.

Paso 1

Distinguir la persona del fondo

+

Paso 2

Desenfocar solo el fondo

×

Cada cuadro

30+ veces por segundo, sin cortes

*En la CPU resulta exigente; en gráficos integrados se entrecorta y consume batería.*

# Por qué la NPU lo resuelve mejor

- Una IA ligera detecta la silueta con consumo mínimo
- Lo procesa **dentro de la cámara**, antes de llegar a la app
- Zoom o Teams reciben la imagen **ya procesada**

EN RESUMEN

# ¿Qué concluimos?

01 La IA local requiere un chip diseñado para ello

02 La NPU lo logra sin placas costosas ni dependencia de internet

03 Ya es una realidad: en los juegos y las videollamadas

PARA PENSAR JUNTOS

# Abrimos el debate

01 ¿Tendrían en cuenta los TOPS de la NPU al comprar una laptop, como hoy se observa la RAM o el procesador?

02 Si el equipo ejecuta la IA en local, ¿le confiarían datos que hoy no subirían a la nube?

03 ¿Qué otra tarea cotidiana convendría delegar a un acelerador local?

# Gracias

**Facundo Nicolás Eiroa · Nazareno Gastón Ferreyro**

Lenguajes Formales y Teoría de la Computación · UCEMA · 2026