1.5 Infraestructura de IA y ejecución de modelos

Cuando trabajamos con inteligencia artificial —especialmente con modelos grandes como redes neuronales o modelos de lenguaje— el hardware se convierte en un elemento fundamental. Los cálculos necesarios para entrenar o ejecutar estos modelos son enormes y requieren procesadores capaces de realizar **millones o incluso billones de operaciones matemáticas por segundo**.

Por esta razón, en los últimos años han aparecido distintos tipos de procesadores especializados para IA. Mientras que los ordenadores tradicionales utilizan principalmente **CPUs**, el desarrollo de la inteligencia artificial ha impulsado el uso de **GPUs, TPUs y NPUs**, cada uno optimizado para distintos tipos de tareas.

De forma sencilla, podemos pensar en ellos como diferentes “motores” de cálculo diseñados para distintos contextos: centros de datos, investigación, ordenadores personales o dispositivos móviles.

CPU (procesador tradicional)
Las CPUs son los procesadores generales de los ordenadores. Son muy versátiles y pueden ejecutar todo tipo de programas, pero no están optimizadas para los cálculos masivos que requieren los modelos de inteligencia artificial.

Una CPU moderna puede realizar solo unas pocas operaciones simultáneas, mientras que los modelos de IA requieren realizar miles o millones de operaciones matemáticas en paralelo.

Por eso, aunque una CPU puede ejecutar modelos pequeños, no suele ser suficiente para entrenar redes neuronales grandes.

GPU (Graphics Processing Unit)
Las **GPUs** fueron diseñadas originalmente para gráficos y videojuegos, pero resultaron ser muy eficaces para inteligencia artificial.

Su gran ventaja es el **procesamiento paralelo**: una GPU puede tener miles de núcleos trabajando al mismo tiempo, lo que permite ejecutar muchas operaciones matemáticas simultáneamente.

Esto las convierte en la herramienta principal para entrenar modelos de deep learning. Hoy en día, la mayoría de los modelos de IA se entrenan utilizando GPUs en centros de datos o clusters de computación.

Ventajas principales de las GPUs

- Gran capacidad de cálculo paralelo
- Compatibilidad con muchos frameworks de IA (PyTorch, TensorFlow, JAX)
- Ecosistema de software muy desarrollado
- Flexibilidad para distintos tipos de tareas

Por esta razón, empresas como NVIDIA dominan gran parte del hardware utilizado para entrenar modelos de IA.

TPU (Tensor Processing Unit)
Las **TPUs** son procesadores diseñados específicamente para inteligencia artificial. Fueron desarrolladas por Google para acelerar operaciones matemáticas típicas de redes neuronales, especialmente multiplicaciones de matrices.

A diferencia de las GPUs, las TPUs utilizan una arquitectura especializada llamada **systolic array**, que permite realizar cálculos de forma extremadamente eficiente en tareas de aprendizaje profundo.

En algunos casos, las TPUs pueden ofrecer una eficiencia energética mucho mayor que CPUs y GPUs para tareas de inferencia y entrenamiento.

Ventajas principales de las TPUs

- Muy eficientes en operaciones de redes neuronales
- Alto rendimiento por consumo energético
- Integración directa con plataformas como Google Cloud

Sin embargo, suelen ser menos flexibles que las GPUs y están más orientadas a ciertos frameworks.

NPU (Neural Processing Unit)
Las **NPUs** son procesadores diseñados específicamente para ejecutar modelos de IA en dispositivos pequeños como móviles, ordenadores portátiles o dispositivos IoT.

Se encuentran, por ejemplo, en chips de teléfonos o en procesadores modernos de laptops (como Apple Silicon o Intel AI Boost). Su objetivo principal es ejecutar modelos de IA de forma eficiente directamente en el dispositivo.

Una de sus principales ventajas es la **eficiencia energética**, ya que consumen mucha menos energía que GPUs o CPUs para tareas de inferencia.

Esto permite ejecutar funciones de inteligencia artificial en tiempo real, como reconocimiento de voz, visión artificial o asistentes inteligentes, sin depender de la nube.

Ventajas principales de las NPUs

- Muy bajo consumo energético
- Ideales para dispositivos móviles o edge computing
- Buen rendimiento para inferencia en tiempo real

Tabla comparativa de tecnologías de hardware para IA

Tecnología	Tipo de dispositivo	Uso principal	Ventajas	Ejemplos
CPU	Ordenadores generales	Tareas generales, control del sistema	Gran flexibilidad	Intel Xeon, AMD EPYC
GPU	Centros de datos, PCs	Entrenamiento de modelos de IA	Gran paralelismo y ecosistema software	NVIDIA H100, A100, AMD MI300
TPU	Infraestructura cloud	Entrenamiento e inferencia de modelos grandes	Muy eficiente para operaciones de tensor	Google TPU v5, v6, v7
NPU	Móviles, laptops, edge devices	Inferencia local de IA	Muy bajo consumo energético	Apple Neural Engine, Intel AI Boost, Qualcomm Hexagon

La infraestructura de hardware es uno de los pilares del desarrollo de la inteligencia artificial. Mientras que las CPUs siguen siendo esenciales para tareas generales, el crecimiento de la IA ha impulsado el desarrollo de aceleradores especializados como GPUs, TPUs y NPUs.

Las GPUs dominan el entrenamiento de modelos grandes, las TPUs ofrecen una gran eficiencia en centros de datos y las NPUs permiten llevar la inteligencia artificial directamente a dispositivos personales.

En la práctica, los sistemas modernos suelen combinar varios de estos componentes, creando arquitecturas heterogéneas capaces de aprovechar lo mejor de cada tipo de procesador.

Dónde ejecutar modelos: nube, local y edge

Cuando trabajamos con modelos de inteligencia artificial, una de las decisiones importantes es **dónde se van a ejecutar**. En la práctica existen tres opciones principales: usar infraestructura en la nube, ejecutarlos en equipos propios o hacer que funcionen directamente en dispositivos. Cada opción tiene ventajas y limitaciones, y suele elegirse en función del tipo de proyecto, los recursos disponibles y el nivel de control que se necesita sobre los datos.

Computación en la nube
La nube es probablemente la forma más sencilla de empezar a trabajar con inteligencia artificial. Plataformas como Google Colab, AWS o Azure permiten ejecutar modelos sin instalar nada en el ordenador. El usuario simplemente abre un entorno en línea y puede utilizar recursos potentes, como GPUs o grandes cantidades de memoria.

La principal ventaja de la nube es que **permite escalar fácilmente el hardware según la necesidad**. Si un proyecto necesita mucha potencia de cálculo, se pueden utilizar servidores muy potentes durante unas horas o días. Esto es especialmente útil para entrenar modelos grandes o para proyectos que requieren GPUs avanzadas.

Otra ventaja es la facilidad de uso: muchas plataformas ya incluyen entornos de programación, bibliotecas y datasets preparados para trabajar.

Sin embargo, también tiene algunas limitaciones. La nube depende de una conexión a internet y normalmente implica costes asociados al tiempo de cálculo o al uso de recursos. Además, enviar datos a servidores externos puede generar preocupaciones relacionadas con la privacidad.

En educación, la nube es muy útil para **demos en clase, experimentos o proyectos puntuales que necesitan hardware potente**.

Infraestructura local (on-premises)
Otra posibilidad es ejecutar los modelos directamente en ordenadores propios, como PCs, portátiles o servidores del centro educativo o de la empresa. A esto se le llama infraestructura **local o on-premises**.

La principal ventaja de este enfoque es el **control total sobre los datos**. Como la información no sale del equipo o del servidor local, es más fácil mantener la privacidad y cumplir políticas de seguridad.

También se evita la latencia de red, es decir, el tiempo que tarda la información en viajar a servidores remotos.

El inconveniente principal es que los recursos de hardware suelen ser más limitados. Un ordenador personal normalmente dispone de menos memoria y menos potencia de cálculo que un servidor en la nube. Por eso, este enfoque suele utilizarse para **inferencias de modelos medianos o experimentos de entrenamiento ligero**.

En muchos entornos educativos o de investigación se utilizan PCs con GPUs de escritorio (por ejemplo tarjetas RTX) para ejecutar modelos open-source o experimentar con proyectos de IA.

Computación en el edge (dispositivos)
La tercera opción es ejecutar los modelos directamente en dispositivos como móviles, sensores,

cámaras inteligentes o microcontroladores. Este enfoque se conoce como **edge computing**.

La idea es que el procesamiento se realice **cerca del lugar donde se generan los datos**, en lugar de enviarlos a un servidor remoto. Esto tiene varias ventajas importantes.

La primera es la **latencia extremadamente baja**. Al procesar la información localmente, las respuestas pueden generarse casi en tiempo real, algo fundamental para aplicaciones que requieren decisiones rápidas.

Otra ventaja es la **privacidad**, ya que los datos sensibles pueden procesarse directamente en el dispositivo sin enviarse a la nube.

Además, los sistemas edge pueden funcionar incluso sin conexión a internet, lo que permite operar en entornos remotos o con conectividad limitada.

Sin embargo, los dispositivos edge tienen recursos limitados. Suelen disponer de menos memoria, menos almacenamiento y menor capacidad de cálculo que los servidores cloud.

Por esta razón, los modelos que se ejecutan en estos dispositivos suelen estar **comprimidos u optimizados** para funcionar con menos recursos.

En educación, este enfoque aparece en proyectos que utilizan teléfonos móviles, cámaras inteligentes o microcontroladores (como Arduino o Raspberry Pi) para ejecutar modelos de reconocimiento de imágenes o sonido directamente en el dispositivo.

Tabla comparativa: nube vs local vs edge

Entorno	Dónde se ejecuta	Ventajas	Limitaciones	Ejemplos de uso
Nube	Centros de datos remotos	Gran potencia de cálculo, escalabilidad, fácil acceso	Costes, dependencia de internet, privacidad	entrenamiento de modelos grandes, demos con GPUs
Local (on-premises)	PCs o servidores propios	Control de datos, sin latencia de red	hardware limitado	investigación, ejecución de modelos open-source
Edge	dispositivos y sensores	latencia muy baja, privacidad, funciona offline	recursos muy limitados	móviles, IoT, reconocimiento de imágenes en dispositivos

No existe una única solución válida para todos los casos. La nube ofrece potencia y escalabilidad, los sistemas locales ofrecen control y privacidad, y el edge permite ejecutar inteligencia artificial directamente en dispositivos con respuestas inmediatas.

Por eso, muchos sistemas actuales utilizan **arquitecturas híbridas**, donde el entrenamiento se realiza en la nube, mientras que la inferencia o el uso final del modelo se ejecuta en equipos locales o dispositivos edge.

Este enfoque combinado permite aprovechar lo mejor de cada entorno y es una de las tendencias más importantes en la infraestructura moderna de inteligencia artificial.

Revision #5
Created 2026-03-04 17:45:57 CET by Luis Hueso
Updated 2026-03-16 20:18:45 CET by Luis Hueso