Cuando trabajamos con **Inteligencia Artificial en local, modelos de lenguaje de gran tamaño, generación de imágenes, renderizado 3D o cargas de trabajo científicas**, llega un momento en el que la cantidad de VRAM disponible empieza a ser tan importante como la potencia de cálculo de la propia GPU.

Los habituales 8, 12, 16 o incluso 24 GB de VRAM pueden quedarse cortos cuando queremos cargar modelos grandes completamente dentro de la memoria de la GPU. El siguiente salto realmente interesante son los **48 GB de VRAM en una única tarjeta gráfica**.

Disponer de 48 GB en una sola GPU tiene además una ventaja importante frente a utilizar varias tarjetas gráficas pequeñas: **no necesitamos dividir necesariamente el modelo entre varias GPU**. En una configuración multi-GPU la memoria de las tarjetas no se convierte automáticamente en un único bloque de VRAM, por lo que el software debe soportar explícitamente la distribución del modelo y gestionar las transferencias entre dispositivos.

Aunque NVIDIA domina claramente este segmento, actualmente podemos encontrar soluciones de **AMD, NVIDIA e incluso Intel**.

Qué tarjetas gráficas de 48 GB existen

Una de las sorpresas al revisar el catálogo completo es Intel. Aunque es frecuente asociar las GPU profesionales de gran cantidad de memoria exclusivamente con NVIDIA y AMD, Intel dispone de la **Intel Data Center GPU Max 1100**, basada en la arquitectura Xe-HPC y equipada con nada menos que **48 GB de memoria HBM2e**.

La Intel Max 1100 es especialmente llamativa por su subsistema de memoria: utiliza una interfaz de 1024 bits y alcanza aproximadamente **1.228,8 GB/s de ancho de banda**. Funciona mediante PCI Express 5.0 x16 y tiene un TDP de 300 W. Está orientada principalmente a IA y HPC en servidores, no a una workstation convencional. Intel lanzó este modelo en 2023 y situó su discontinuación prevista en enero de 2026, por lo que actualmente resulta especialmente interesante en el mercado de material profesional usado o reacondicionado.

No encontramos, sin embargo, una GPU Intel de 40 GB. Dentro de la familia Intel Data Center GPU encontramos capacidades de 12, 16, 48 y 128 GB. La conocida GPU de 40 GB pertenece realmente a NVIDIA: la **NVIDIA A100** se comercializó, entre otras configuraciones, con 40 GB de memoria HBM2.

GPU	Arquitectura	Memoria	TDP	Tipo
Intel Data Center GPU Max 1100	Xe-HPC	48 GB HBM2e	300 W	Datacenter / HPC
NVIDIA Quadro RTX 8000	Turing	48 GB GDDR6 ECC	295 W	Workstation
AMD Radeon PRO W7800 48GB	RDNA 3	48 GB GDDR6 ECC	260 W	Workstation
AMD Radeon PRO W7900	RDNA 3	48 GB GDDR6 ECC	295 W	Workstation
NVIDIA RTX A6000	Ampere	48 GB GDDR6 ECC	300 W	Workstation
NVIDIA A40	Ampere	48 GB GDDR6 ECC	300 W	Datacenter
NVIDIA L20	Ada Lovelace	48 GB GDDR6	350 W	Datacenter
NVIDIA L40	Ada Lovelace	48 GB GDDR6 ECC	300 W	Datacenter
NVIDIA L40S	Ada Lovelace	48 GB GDDR6 ECC	350 W	Datacenter
NVIDIA RTX 5880 Ada	Ada Lovelace	48 GB GDDR6 ECC	285 W	Workstation
NVIDIA RTX 6000 Ada	Ada Lovelace	48 GB GDDR6 ECC	300 W	Workstation
NVIDIA RTX PRO 5000 Blackwell	Blackwell	48 GB GDDR7 ECC	300 W	Workstation

Dentro de AMD destacan especialmente la **Radeon PRO W7800 48GB** y la **Radeon PRO W7900**. Ambas utilizan arquitectura RDNA 3, memoria GDDR6 ECC, una interfaz de memoria de 384 bits y alcanzan hasta 864 GB/s de ancho de banda.

La W7800 48GB dispone de 4.480 Stream Processors, 70 Compute Units y un consumo de 260 W, mientras que la W7900 aumenta hasta 6.144 Stream Processors, 96 Compute Units y 295 W. Resulta curioso que, debido a las variaciones del mercado,

podamos encontrar actualmente algunas W7900 incluso más baratas que las W7800 de 48 GB.

NVIDIA dispone de un catálogo mucho más amplio. La **RTX A6000** continúa siendo particularmente interesante para IA debido a sus 48 GB, compatibilidad con CUDA, formato de solamente dos slots y un consumo de 300 W. Aunque pertenece a la generación Ampere, sigue siendo una alternativa atractiva cuando aparece a buen precio en el mercado de segunda mano.

Dentro de Ada Lovelace encontramos las **L20, L40, L40S, RTX 5880 Ada y RTX 6000 Ada**. Las L20, L40 y L40S están fundamentalmente orientadas a servidores, mientras que las RTX 5880 y RTX 6000 Ada están diseñadas para estaciones de trabajo.

Finalmente encontramos la **NVIDIA RTX PRO 5000 Blackwell**, una de las opciones tecnológicamente más modernas. Incorpora 48 GB de memoria **GDDR7 ECC**, alcanza aproximadamente **1.344 GB/s de ancho de banda**, utiliza PCI Express 5.0 y mantiene un consumo máximo de 300 W.

Cuánto cuestan 48 GB de VRAM

Los precios en este segmento son especialmente difíciles de comparar porque mezclamos tarjetas actuales, hardware profesional de generaciones anteriores y aceleradoras de datacenter que normalmente se venden integradas dentro de servidores. Además, el mercado de segunda mano puede modificar completamente el orden.

Si nos centramos primero en **tarjetas nuevas que podemos encontrar en el mercado europeo**, los precios aproximados observados en 2026 son los siguientes:

GPU	**Precio aproximado desde**
AMD Radeon PRO W7900 48GB	3.300 €
AMD Radeon PRO W7800 48GB	3.500 €
NVIDIA RTX A6000 48GB	6.700 €
NVIDIA L40S 48GB	7.700 €
NVIDIA RTX PRO 5000 Blackwell 48GB	7.900 €
NVIDIA L40 48GB	8.000 € aproximadamente
NVIDIA RTX 6000 Ada 48GB	8.500 €
NVIDIA RTX 5880 Ada 48GB	8.700 €
NVIDIA A40 48GB	9.000 € aproximadamente

Estos precios deben considerarse únicamente orientativos. En hardware profesional podemos encontrar diferencias de varios miles de euros entre distribuidores para exactamente la misma GPU.

El mercado de segunda mano resulta bastante más interesante. Las antiguas **Quadro RTX 8000 de 48 GB** pueden aparecer alrededor de los 2.000-2.500 €, mientras que las RTX A6000 usadas empiezan a entrar en una franja mucho más atractiva que su precio nuevo.

La **Intel Data Center GPU Max 1100** merece una mención especial. Al haber salido del ciclo comercial habitual podemos encontrar unidades procedentes de servidores por precios radicalmente inferiores a los que tenía originalmente este hardware. Se han visto unidades usadas o reacondicionadas alrededor de los 2.000-2.400 dólares. No podemos comparar directamente este precio con una GPU nueva comprada en España porque debemos tener en cuenta procedencia, IVA, importación, garantía y compatibilidad del servidor.

Qué GPU de 48 GB nos interesa para Inteligencia Artificial

Si nuestro objetivo principal es ejecutar **LLM en local**, no debemos fijarnos únicamente en la cantidad de memoria. El ecosistema de software es igualmente importante.

NVIDIA continúa teniendo una ventaja práctica considerable gracias a **CUDA** y al enorme número de frameworks y herramientas optimizados para sus GPU. Por ello, una RTX A6000 usada puede resultar mucho más interesante para determinados proyectos de IA que una GPU teóricamente más moderna pero con peor soporte por parte del software que vayamos a utilizar.

AMD utiliza **ROCm** como plataforma de computación y actualmente tanto la Radeon PRO W7800 48GB como la W7900 aparecen dentro del hardware profesional soportado por ROCm. Su principal atractivo es evidente: podemos conseguir **48 GB de VRAM nuevos por poco más de 3.000 €**, aproximadamente la mitad de lo que cuesta entrar en determinadas soluciones NVIDIA profesionales nuevas.

Intel juega en un terreno diferente con la Max 1100. Sus **48 GB de HBM2e y 1.228,8 GB/s de ancho de banda** resultan técnicamente muy atractivos, pero debemos comprobar cuidadosamente la compatibilidad de nuestro software con el ecosistema Intel oneAPI antes de adquirir una tarjeta de este tipo únicamente por su cantidad de memoria o por su precio de segunda mano.

También debemos distinguir entre una **GPU de workstation y una aceleradora de datacenter**. Tarjetas como la RTX A6000, RTX 6000 Ada, RTX PRO 5000 Blackwell, W7800 o W7900 llevan refrigeración activa y podemos instalarlas mucho más fácilmente en una workstation convencional.

En cambio, modelos como las NVIDIA A40, L20, L40 o L40S utilizan diseños destinados principalmente a servidores y habitualmente dependen del enorme flujo de aire generado por los ventiladores del propio chasis. Instalar una GPU pasiva de datacenter dentro de una torre convencional sin diseñar correctamente su refrigeración puede provocar temperaturas excesivas incluso aunque dispongamos de una fuente de alimentación suficientemente potente.

Para montar sistemas multi-GPU también debemos considerar el espacio físico. Una configuración con tres tarjetas de doble slot necesita una placa base con los PCI Express suficientemente separados y un chasis capaz de alojarlas. Tres GPU de 48 GB nos proporcionarían **144 GB de VRAM física total**, pero seguiremos dependiendo del software para repartir el modelo y las cargas de trabajo entre las diferentes GPU.

Por precio puro, las Radeon PRO W7800 48GB y W7900 son actualmente las formas más económicas de comprar **48 GB nuevos en una única GPU**. Si necesitamos CUDA, resulta especialmente interesante vigilar el mercado de **RTX A6000 usadas**. Si queremos la arquitectura profesional NVIDIA más moderna disponible con exactamente 48 GB, la **RTX PRO 5000 Blackwell** es la opción de referencia, aunque su precio se acerca actualmente a los 8.000 €.