

Conversación con Gemini

Exported: 12/6/2026 12:00:13

Jordi H.

Has dicho Cuantos tops tiene una L40S

Gemini

La NVIDIA L40S es una tarjeta diseñada específicamente para cargas de trabajo de IA y gráficos profesionales. Dependiendo de la precisión del cálculo (si es con o sin aceleración de núcleos Tensor), sus capacidades varían significativamente. Aquí tienes el desglose de su rendimiento en **TFLOPS** (Teraflops) y **TOPS** (Tera Operations Per Second)

Rendimiento de Cómputo (Precisiones estándar)

Tipo de Cómputo	Rendimiento (Suma-Producto)
FP32 (Punto flotante de 32 bits)	91.6 TFLOPS
TF32 (Tensor Float 32)	366 TFLOPS*
BFLOAT16 / FP16 (Tensor Core)	733 TFLOPS*
FP8 (Tensor Core)	1,466 TFLOPS*

Rendimiento en Números Enteros (TOPS)

Para tareas de inferencia que utilizan cuantización de enteros:

- INT8 (Tensor Core): 1,466 TOPS*

Nota: Los valores marcados con un asterisco (*) incluyen la aceleración mediante **Sparsity** (dispersión), una técnica que permite duplicar el rendimiento teórico en modelos de IA compatibles

Puntos clave de la arquitectura

- Arquitectura Ada Lovelace: A diferencia de la H100 (arquitectura Hopper), la L40S destaca por su versatilidad, ya que incluye núcleos RT de 3ª generación, lo que la hace excelente no solo para inferencia de LLMs, sino también para renderizado y Omnivers
- Memoria: Cuenta con 48 GB de memoria GDDR6 es con ECC, lo que ofrece un ancho de banda de 864 GB/s
- Potencia: Tiene un consumo (TDP) de 350W y utiliza refrigeración pasiva, diseñada para su integración en servidores de centros de datos

Jordi H.

Has dicho hazme una pequeña tabla en TOPS de las tarjetas que se pueden comprar para pc de torre y su precio. hazme otra tabla TOPS con tarjetas gpu de pequeño formato para pcs sli

Python

CODE

```
import os

# Create HTML content for the tables using a clean, professional design suitable for a technical
report.
html_content = """<!DOCTYPE html>
<html lang="es">
<head>
<meta charset="UTF-8">
<title>Comparativa de GPUs: Rendimiento en TOPS y Precios (2026)</title>
<style>
@page {
size: A4;
margin: 18mm 15mm;
background-color: #f8fafc;
}
*, *::before, *::after {
box-sizing: border-box;
}
body {
font-family: 'Segoe UI', -apple-system, BlinkMacSystemFont, Roboto, Helvetica, Arial, sans-serif;
color: #1e293b;
margin: 0;
padding: 0;
font-size: 10pt;
line-height: 1.5;
}
.header-banner {
background-color: #0f172a;
color: #ffffff;
margin: -18mm -15mm 25px -15mm;
padding: 25px 15mm;
border-bottom: 4px solid #10b981;
}
h1 {
font-size: 18pt;
margin: 0 0 5px 0;
font-weight: 700;
letter-spacing: -0.5px;
}
.subtitle {
font-size: 11pt;
color: #94a3b8;
margin: 0;
}
h2 {
font-size: 13pt;
color: #0f172a;
margin: 25px 0 12px 0;
padding-left: 8px;
border-left: 4px solid #10b981;
page-break-after: avoid;
}
p {
margin: 0 0 12px 0;
color: #475569;
}
table {
width: 100%;
border-collapse: collapse;
margin-bottom: 20px;
```

```

background-color: #ffffff;
box-shadow: 0 1px 3px rgba(0,0,0,0.05);
page-break-inside: avoid;
}
th {
background-color: #1e293b;
color: #ffffff;
font-weight: 600;
text-align: left;
padding: 8px 10px;
font-size: 9.5pt;
border: 1px solid #1e293b;
}
td {
padding: 8px 10px;
border: 1px solid #e2e8f0;
font-size: 9.5pt;
color: #334155;
}
tr:nth-child(even) td {
background-color: #f1f5f9;
}
.text-right {
text-align: right;
}
.text-center {
text-align: center;
}
.highlight {
font-weight: 600;
color: #0f172a;
}
.note-box {
background-color: #f0fdf4;
border-left: 4px solid #16a34a;
padding: 10px 12px;
margin: 15px 0;
font-size: 9pt;
color: #14532d;
}
.footer {
margin-top: 30px;
font-size: 8.5pt;
color: #64748b;
text-align: center;
border-top: 1px solid #e2e8f0;
padding-top: 10px;
}
</style>
</head>
<body>

```

```

<div class="header-banner">

```

```

<h1>Guía de Rendimiento en IA: TOPS y Precios de GPUs</h1>

```

```

<div class="subtitle">Análisis comparativo para PCs de Torre Estándar y Equipos de Pequeño Formato
(SFF/Slim)</div>
</div>

```

El rendimiento en Inteligencia Artificial y computación de enteros para GPUs comerciales y profesionales se mide principalmente en **TOPS** (Tera Operations Per Second) utilizando cuantización **INT8**. A continuación se presentan los datos actualizados de las opciones disponibles en el mercado, optimizadas para el despliegue local de modelos.

1. Tarjetas GPU para PC de Torre Estándar (Full-Height / PCIe Estándar)

Estas tarjetas requieren un chasis convencional ATX/Micro-ATX, espacio de 2 a 4 slots y fuentes de alimentación potentes (conectores de 12VHPWR o PCIe de 8 pines).

Modelo de GPU	VRAM (GB)	Rendimiento INT8 (TOPS)	Consumo (TDP)	Precio Aprox. ()
NVIDIA GeForce RTX 4090	24 GB GDDR6X	1,321 TOPS	450W	1,850 - 2,100
NVIDIA GeForce RTX 4080 Super	16 GB GDDR6X	836 TOPS	320W	1,050 - 1,150
NVIDIA GeForce RTX 4070 Ti Super	16 GB GDDR6X	706 TOPS	285W	850 - 950
NVIDIA GeForce RTX 4070 Super	12 GB GDDR6X	568 TOPS	220W	620 - 690
AMD Radeon RX 7900 XTX	24 GB GDDR6	613 TOPS	355W	950 - 1,050
NVIDIA GeForce RTX 4060 Ti (16GB)	16 GB GDDR6	353 TOPS	165W	460 - 520

Nota de Arquitectura (Torre): Los valores indicados para NVIDIA incluyen la aceleración mediante núcleos Tensor de 4ª generación con la característica de *Sparsity* activada. La RTX 4060 Ti de 16GB se mantiene como una de las opciones más eficientes en coste por GB de VRAM para LLMs locales, a pesar de su menor tasa de TOPS brutas.

2. Tarjetas GPU de Bajo Perfil / Pequeño Formato (Low Profile / Slim PCs)

Diseñadas específicamente para servidores compactos de 2U o ordenadores de escritorio tipo SFF

(Small Form Factor). Tienen restricciones severas de altura (Low Profile) y no suelen requerir alimentación externa o se limitan a un slot.</p>

Modelo de GPU	VRAM (GB)	Rendimiento INT8 (TOPS)	Consumo (TDP)	Precio Aprox. ()
NVIDIA RTX 4000 Ada Generation (SFF)	20 GB GDDR6 ECC	385 TOPS	70W	1,450 - 1,600
NVIDIA RTX 2000 Ada Generation (SFF)	16 GB GDDR6 ECC	241 TOPS	70W	690 - 750
NVIDIA GeForce RTX 4060 Low Profile	8 GB GDDR6	242 TOPS	115W	330 - 370
NVIDIA RTX A2000 (Ampere)	12 GB GDDR6 ECC	145 TOPS	70W	450 - 550 (Mercado/Stock)
AMD Radeon RX 6400 Low Profile	4 GB GDDR6	~30 TOPS	53W	120 - 140

Nota de Arquitectura (Slim/SFF): Las tarjetas de la serie profesional "NVIDIA RTX Ada SFF" son excepcionales para entornos de laboratorio compacto (homelabs). Al no requerir conectores de alimentación adicionales (se alimentan del slot PCIe, máx. 75W) y contar con memoria con soporte ECC, ofrecen una estabilidad corporativa ideal para cargas continuadas de inferencia ligera y desarrollo de arquitecturas RAG.

Documento técnico de referencia para el dimensionamiento de hardware local orientado a IA. Precios estimados basados en canales de distribución estándar europeos.

```
</html>
"""

with open("comparativa_gpus_tops.html", "w", encoding="utf-8") as f:
    f.write(html_content)

from weasyprint import HTML
HTML("comparativa_gpus_tops.html").write_pdf("comparativa_gpus_tops.pdf")
print("PDF generado con éxito.")
```

Salida de código

```
CODE

PDF generado con éxito.
```

Su archivo PDF está listo comparativ..._gpus_tops PDF He preparado un documento técnico en PDF limpio y bien estructurado que recoge de forma visual ambas tablas. A continuación, te detallo los datos clave d e rendimiento en **INT8 TOPS** (la precisión utilizada de forma nativa para inferencia y cuantización de LLMs) jun to con los precios estimados de merca

1. Tarjetas para PC de Torre Estándar (Full-Height / Slots Múltiples)

Estas tarjetas requieren cajas ATX estándar, un espacio considerable de disipación (2 a 4 slots) y conectores de alimentación dedicados de alta potencia

Modelo de GPU	VRAM (GB)	Rendimiento INT8	Consumo (TDP)	Precio Aprox. (€)
NVIDIA GeForce	24 GB GDDR6X	1,321 TOPS	450W	1.850€ - 2.100€
NVIDIA GeForce	16 GB GDDR6X	836 TOPS	320W	1.050€ - 1.150€
NVIDIA GeForce	16 GB GDDR6X	706 TOPS	285W	850€ - 950€
AMD Radeon RX	24 GB GDDR6	613 TOPS	355W	950€ - 1.050€
NVIDIA GeForce	12 GB GDDR6X	568 TOPS	220W	620€ - 690€
NVIDIA GeForce	16 GB GDDR6	353 TOPS	165W	460€ - 520€

Nota: La RTX 4060 Ti de 16GB sigue siendo muy popular en laboratorios locales por su bajo coste por GB de VRAM, fundamental para cargar modelos grandes sin hacer "offloading" a la CPU, a pesar de tener meno s TOPS bruto

2. Tarjetas de Pequeño Formato (Low Profile / Slim PCs)

Diseñadas para cajas SFF (Small Form Factor) o servidores compactos (2U). Muchas de ellas tienen la gran ventaja arquitectónica de tener un **TDP limitado a 70W-75W**, alimentándose directamente de la ranura PCIe si n cables adicionales de la fuent

Modelo de GPU	VRAM (GB)	Rendimiento INT8	Consumo (TDP)	Precio Aprox. (€)
NVIDIA RTX 4000	20 GB GDDR6 ECC	385 TOPS	70W	1.450€ - 1.600€
NVIDIA GeForce	8 GB GDDR6	242 TOPS	115W	330€ - 370€
NVIDIA RTX 2000	16 GB GDDR6 ECC	241 TOPS	70W	690€ - 750€
NVIDIA RTX A2000	12 GB GDDR6 ECC	145 TOPS	70W	450€ - 550€ (Stock)
AMD Radeon RX	4 GB GDDR6	~30 TOPS	53W	120€ - 140€

Aspectos técnicos a tener en cuenta para Homelabs o PCs Compactos:

1. Memoria ECC: Las gamas profesionales (RTX 4000 SFF / 2000 SFF / A2000) incorporan memoria con corrección de errores (ECC). Para la consistencia a largo plazo y la estabilidad en cargas continuas de inferencia o embeddings dentro de un servidor casero, esto es un gran punto a favor sobre las de consumo.
 2. El "Truco" de la RTX 4060 LP: Es la única de bajo perfil doméstica con alta potencia en TOPS para su precio, pero ten cuidado con el chasis: a diferencia de las de gama profesional (que se ciñen estrictamente a los 75W del slot y ocupan poco espacio), la versión LP de consumo consume 115W y requiere un conector de alimentación externo de 8 pines, además de usar un disipador ligeramente más grueso que puede obstaculizar otros slots de la placa.
-