ESCUELA POLITÉCNICA NACIONAL
FACULTAD DE INGENIERÍA EN SISTEMAS
MÉTODOS NUMÉRICOS

**Nombre:** Camila Caicedo

**Curso:** GR1CC

**Fecha de entrega:** 28/01/2026

# Actividad extracurricular 07b: GPUs Hopper vs Blackwell

## Instrucciones

Investigue sobre las diferencias entre las arquitecturas Hopper y Blackwell.

| Característica | Hopper | Blackwell |
|---|---|---|
| Generación | Anterior a Blackwell | Más reciente y avanzada |
| Enfoque principal | IA y HPC de alto rendimiento | IA a gran escala y máxima eficiencia |
| Eficiencia energética | Alta | Aún mayor que Hopper |
| Tensor Cores | Muy avanzados | Mejorados y optimizados |
| Soporte de formatos numéricos | FP64, FP32, TF32, FP16, BF16, INT8 | Todos los de Hopper + mejor rendimiento en formatos de baja precisión |
| Rendimiento en IA | Muy alto | Superior, especialmente en modelos grandes |
| Escalabilidad | Alta | Más optimizada para sistemas multi-GPU |
| Consumo energético por operación | Mayor que Blackwell | Menor, más eficiente |

## Preguntas de análisis

- ¿Cuál es la diferencia entre FP32 vs TP32?

| Característica | FP32 | TF32 |
|---|---|---|
| Nombre completo | Floating Point 32 bits | Tensor Float 32 |
| Precisión | Alta (23 bits de mantisa) | Menor (10 bits de mantisa) |
| Rango de valores | Amplio | Igual que FP32 |
| Velocidad | Más lento | Mucho más rápido en Tensor Cores |
| Uso principal | Cálculos científicos, entrenamiento tradicional | Entrenamiento acelerado en GPUs modernas |
| Exactitud de resultados | Muy alta | Suficientemente alta para redes neuronales |

| | | |
|---|---|---|
| Impacto en rendimiento | Menor rendimiento | Mayor rendimiento y eficiencia |

- **¿Qué representaciones de datos soportan estas GPUs (FP64, FP32, INT8)?**

| Formato | Bits | Precisión | Rango | Uso principal | Ventaja principal |
|---|---|---|---|---|---|
| FP64 | 64 | Muy alta | Muy amplio | Simulaciones científicas, HPC | Máxima precisión |
| FP32 | 32 | Alta | Amplio | Entrenamiento tradicional | Buen balance entre precisión y rendimiento |
| TF32 | 32 | Media (mantisa reducida) | Igual a FP32 | Entrenamiento acelerado en GPUs modernas | Mucha mayor velocidad que FP32 |
| FP16 | 16 | Media | Limitado | Entrenamiento e inferencia en IA | Reduce memoria y aumenta velocidad |
| BF16 | 16 | Media | Amplio (como FP32) | Entrenamiento de redes neuronales | Mejor estabilidad que FP16 |
| INT8 | 8 | Baja | Limitado | Inferencia rápida | Muy rápido y bajo consumo |
| INT4 | 4 | Muy baja | Muy limitado | Inferencia ultra rápida | Máxima eficiencia y menor uso de memoria |

Las GPUs de las arquitecturas Hopper y Blackwell soportan diversas representaciones de datos, entre ellas FP64 para cálculos científicos de alta precisión, FP32 para entrenamiento estándar de redes neuronales, y formatos de menor precisión como FP16, BF16, INT8 e INT4, que se usan principalmente para acelerar la inferencia y reducir el uso de memoria.

- **¿Por qué la nueva arquitectura prefiere representaciones numéricas con menor precisión?**

La nueva arquitectura prefiere representaciones numéricas con menor precisión porque permiten realizar operaciones más rápidas, consumir menos energía y almacenar más datos en menos espacio.

Además, en tareas de inteligencia artificial, estas reducciones de precisión no suelen afectar de forma significativa el rendimiento final de los modelos, lo que hace el proceso más eficiente y económico.