# Actividad extracurricular 07b

## GPUs Hopper vs Blackwell

Isaac González

**Metodos Númericos**

Las arquitecturas **Hopper** y **Blackwell**, desarrolladas por NVIDIA, están diseñadas principalmente para cargas de trabajo de **inteligencia artificial**, **aprendizaje profundo** y **cómputo de alto rendimiento (HPC)**.
Ambas arquitecturas introducen mejoras significativas en eficiencia, rendimiento y soporte para nuevos formatos numéricos, siendo Blackwell una evolución orientada a modelos de IA de gran escala.

---

### ¿Cuál es la diferencia entre FP32 y TF32?

**FP32 (Floating Point 32 bits)** es un formato de punto flotante estándar que ofrece alta precisión numérica y se ha utilizado tradicionalmente en aplicaciones científicas y de ingeniería.

**TF32 (Tensor Float 32)** es un formato introducido por NVIDIA que mantiene el rango dinámico de FP32, pero reduce la precisión de la mantisa.
Esto permite realizar operaciones mucho más rápidas en los Tensor Cores, con una pérdida mínima de precisión aceptable para modelos de aprendizaje profundo.

- FP32 prioriza precisión
- TF32 prioriza rendimiento manteniendo estabilidad numérica

---

### ¿Qué representaciones de datos soportan estas GPUs?

Tanto Hopper como Blackwell soportan múltiples formatos numéricos, entre los más importantes se encuentran:

- FP64: alta precisión, usado en simulaciones científicas
- FP32: precisión estándar
- TF32: optimizado para IA
- FP16 / BF16: entrenamiento eficiente de redes neuronales
- INT8 / INT4: inferencia rápida y eficiente
- FP8 (mejorado en Blackwell): diseñado para modelos de IA a gran escala

Blackwell amplía y optimiza el soporte para precisiones ultrabajas, especialmente FP8, con mayor estabilidad y rendimiento que Hopper.

---

### ¿Por qué la nueva arquitectura prefiere representaciones con menor precisión?

Las arquitecturas modernas priorizan formatos de menor precisión debido a varias razones:

1. **Mayor rendimiento**
   Las operaciones con menos bits requieren menos recursos computacionales y se ejecutan más rápido.

2. **Menor consumo de energía**
   Reducir la precisión disminuye el consumo energético, algo crítico en centros de datos.

3. **Escalabilidad de modelos grandes**
   Los modelos de IA actuales pueden tolerar menor precisión sin afectar significativamente la exactitud.

---