ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA EN SISTEMAS

METODOS NUMERICOS ICCD412

JUAN FRANCISCO PINTO ANDRANGO

Actividad Extracurricular 07b GPUs Hopper vs Blackwell

GR1CC

FECHA DE ENTREGA 28 DE ENERO DEL 2026

Investigue sobre las diferencias entre las arquitecturas Hopper y Blackwell.

Arquitectura Hopper

Arquitectura NVIDIA Hopper

Hopper escala de forma segura diversas cargas de trabajo en cada centro de datos desde pequeñas empresas hasta en la computacion de alto rendimiento. Construido con mas de 80 mil millones de transistores implementando el proceso TSM 4N, Hopper presenta innovaciones revolucionarias que impulsan las GPU *NVIDIA H200 Y H100* las que se conbiana para orfreser aceleraciones en entrenamineto e inferencia de IA generativa.

- la arquitectura Hopper mejora la tecnologia Tensor Core con Trasformer Engine, que esta diseñado para acelerar el entrenamiento de IA
- Los tensor cores aplicacion presiones combianadas de FP8 Y FP16 para acelerar significativamente los calculos de IA
- Hopper Triplica las operaciones de punto flotante por segundo para las precisiones de TF32, FP64, FP16 e INT8
- Los Cores de Hopper impulsan la aceleración de ornden de magnitud

Arquitectura Blackwell

Arquitectura NVIDIA Blackwell

Blackwell aporta a la IA generativa y a la computacion acelerada. apartir de generaciones de tecgologias NVIDIA. Blackwell define un nuevo capitulo para la IA generativa con rendimiento y eficiencia.

Blackwell esta formada por 208 mil millones de transistores, los que se fabrican con el proceso TSMC 4NP. cualquier producto Blackwell cuenta con dos matrices con reticula limitada conectadas con interconexion de chip a chip de 10 Terabytes por segundo (TB/s) en una sola GPU unificada.

- Un transformer Engine utiliza tegnologia Blackwell Tensor Core para la aceleracion de inferencias y el entrenamiento para grandes modelos de lenguaje
- Blackwell Tenso Cores agrega precisiones, como nuevos formatos de microescala para brindar alta precisión
- Blackwell utiliza tecnicas de escalado de microtensor para optimizar el rendimiento y precisión permitiendo
- Duplica el rendimiento para tener una alta precisión
- Blackwell incluye la Computación Confidencial de NVIDIA, para datos confidenciales

Diferencias entre las arquitecturas Hopper y Blackwell.

1. Hopper es un chip tradicional con un limite fisico, Blackwell supera ese limite implementandon dos chips seperados que esta unidos por una conexion ultra rapida de (10TB/s)
2. Presicion Numerica con Hopper (FP8) estandarizo el uso de presion de 8 bits para acelera los entrenamientos de IA, Blackwell maneja el soporte nativo para 4 bits con el Transformer Engine de segunda generacion
3. Cuello de botella Hopper soporta 900GB/s de ancho de banda entre GPUs, Blackwell maneja 1.8TB/s con Bidireccional que es una relevancia en la investigacion
4. Motores Especializados Blackwell añade hardware específico para solucionar problemas que frenan a los investigadores y que Hopper manejaba por software o fuerza bruta

Blackwell ofrece estabilidad y escala masiva. En inferencia o investigación, el soporte de FP4 cambia las reglas del juego, permitiendo ejecutar modelos SOTA (State-of-the-Art) con una fracción de la energía y el coste que requería Hopper

Preguntas de análisis

¿Cuál es la diferencia entre FP32 vs TP32?

FP32 maneja una presición simple mientras que TP32 se usa mas en inteligencia artificaial y es mas eficiente

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

- FP64 destaca en los calculos de ciencias exata como simulaciones y fisica aqui no puede existir fallos
- FP32 Es un estandar antiguo para generar IA
- INT8 se usa en IA para respuestas rapidas cuando una IA ya fue entrenada

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

destaca el rendimiento por segundo cuando hay menor precisión podemos cargar mas operaciones el consumo de energia las operaciones de FP32 o FP64 consumen mucha energia, por eso al haber menor precisión existe mas eficiencias energetica Dar un mejor uso de memoria y ancho de banda con menor precisión como mas datos en caché, menos latencia y menos tráfico

Link del repositorio de Git-Hub

https://github.com/JuanfranPinto/Metodos-Numericos-/blob/main/Actividad-extracurricular-07b-GPUs%20Hopper%20vs%20Blackwell-PJF.ipynb