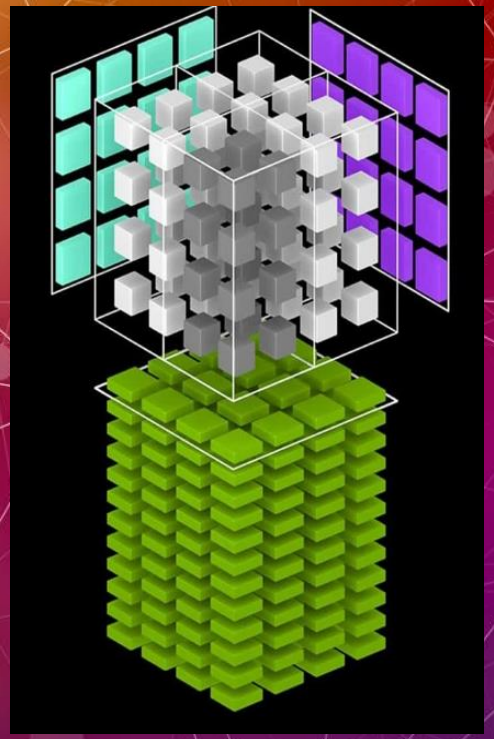


Tensor Cores

Programación en GPUs

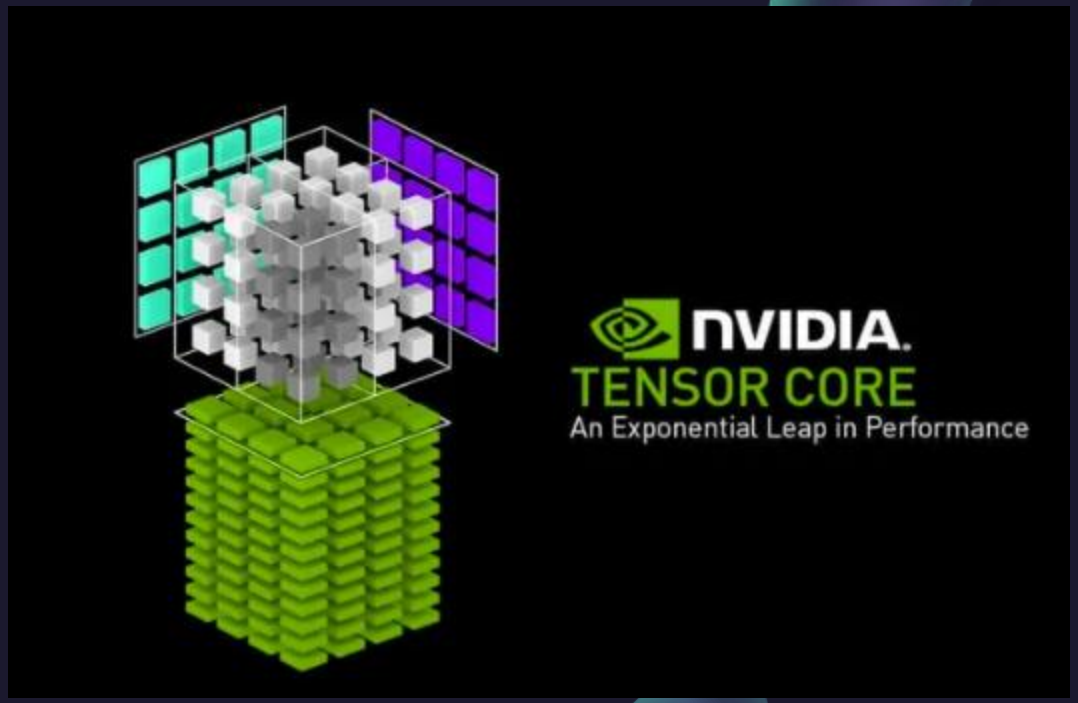
Jose Maureira
Magíster en Informática
Universidad Austral de Chile
23-09-2024

Agenda

- Introducción
- Tecnología a Nivel de Hardware y Programación
- Ventajas de Rendimiento y Limitaciones
- Usos Alternativos de los Tensor Cores
- Impacto en IA y Posible Beneficio para Regresión Simbólica
- Futuro de los Tensor Cores

Introducción

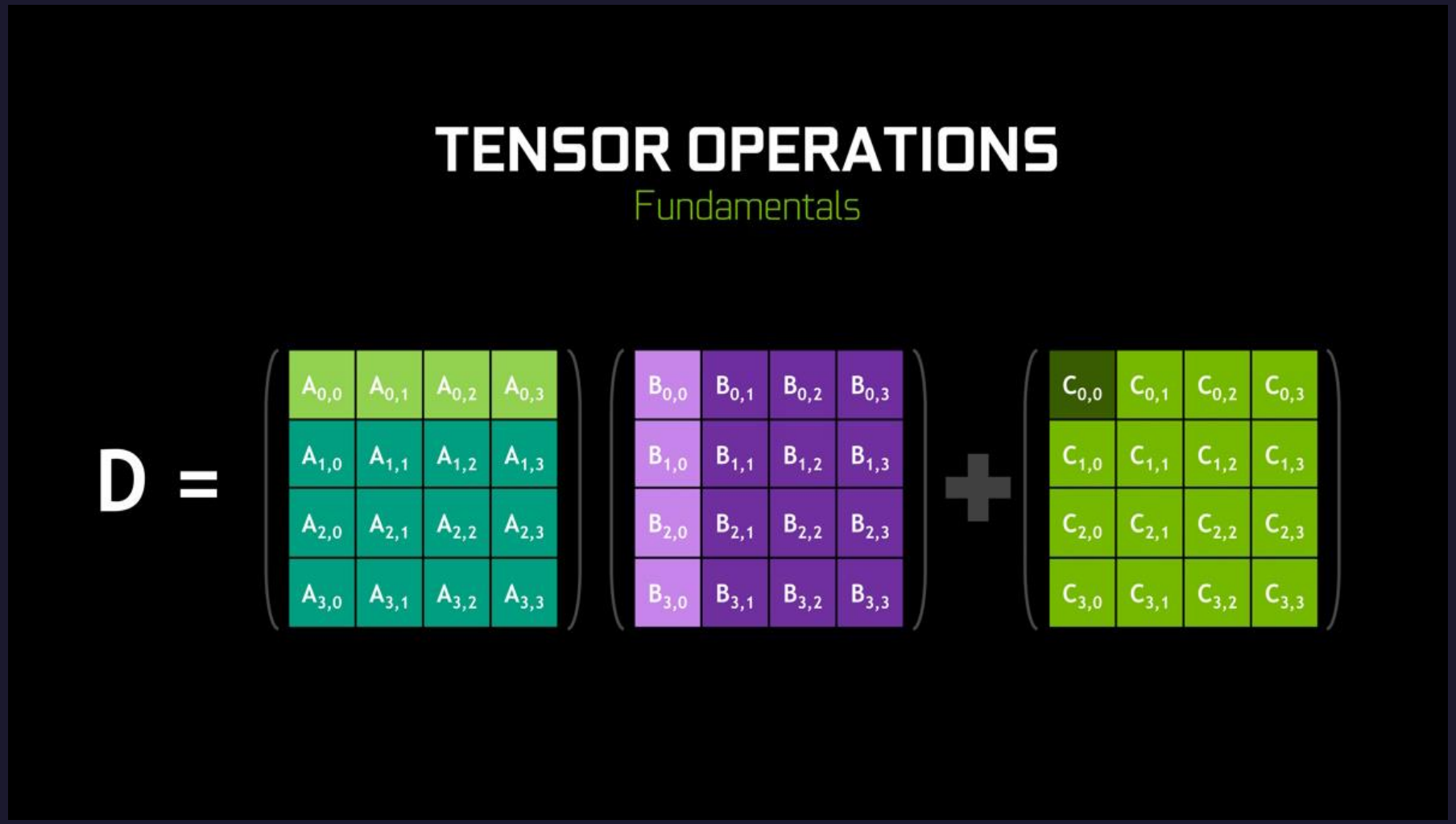
- **¿Qué son los Tensor Cores?**
  - Unidades especializadas en GPUs de NVIDIA.
  - Optimizadas para operaciones de matrices.
- **Relevancia de los Tensor Cores**
  - Aceleración en tareas de IA, especialmente deep learning.
  - Reducción de tiempos de entrenamiento de modelos.
- **Temas a cubrir en la presentación**
  - Tecnología a nivel de hardware y programación.
  - Ventajas y limitaciones.
  - Aplicaciones fuera de la IA.
  - Impacto en mi investigación.
  - Futuro de los Tensor Cores.



Tecnología a Nivel de Hardware y Programación

- ¿Qué son los Tensor Cores?
  - Unidades en GPUs NVIDIA optimizadas para cálculos matriciales.
  - Realizan operaciones en precisión mixta (FP16/INT8 → FP32).
- Funcionamiento de los Tensor Cores
  - Operan en precisión mixta: mayor eficiencia, menos consumo.
  - Operaciones clave: GEMM (Multiplicación de Matrices General), FMA (Fused Multiply-Add).
- **Programación con Tensor Cores**
  - Librerías: **cuBLAS**, **cuDNN** (alta optimización sin necesidad de bajo nivel).
  - Instrucciones de bajo nivel en CUDA: **wmma** (Warp Matrix Multiply Accumulate).

Tecnología a Nivel de Hardware y Programación



Ventajas de Rendimiento y Limitaciones

Ventajas

- Aceleración en operaciones matriciales (**hasta 8x más rápido**).
- Reducción de tiempos en entrenamiento de modelos de IA.
- **Eficiencia energética**: Menos consumo de energía.

Limitaciones

- Optimización enfocada en operaciones de **precisión mixta**.
- Uso limitado fuera de deep learning y multiplicaciones de matrices.
- Requiere ajustes en los algoritmos para aprovechar su potencial.

Usos Alternativos de los Tensor Core

.

•**Simulación de ondas elásticas**: Uso de Tensor Cores INT8 en la simulación de propagación de ondas en mallas estructuradas *"Low-Ordered Orthogonal Voxel Finite Element with INT8 Tensor Cores for GPU-Based Explicit Elastic Wave Propagation Analysis"(T. Ichimura et al., 2024).*

•**Optimización matemática en ingeniería eléctrica**: Aceleración de descomposición QR para fitting vectorial *"TC-GVF: Tensor Core GPU based Vector Fitting via Accelerated Tall-Skinny QR Solvers" (V. Kukutla et al., 2024).*

Simulaciones científicas

- Dinámica molecular y sistemas complejos.

Gráficos por computadora

- Aceleración de técnicas como ray tracing.

Criptografía

- Cálculos matriciales en algoritmos de cifrado.

Procesamiento de señales

- Mejora del rendimiento en transformadas de Fourier (FFT).

Impacto en IA y Beneficio en mi Investigación (Tesis)

Impacto en IA

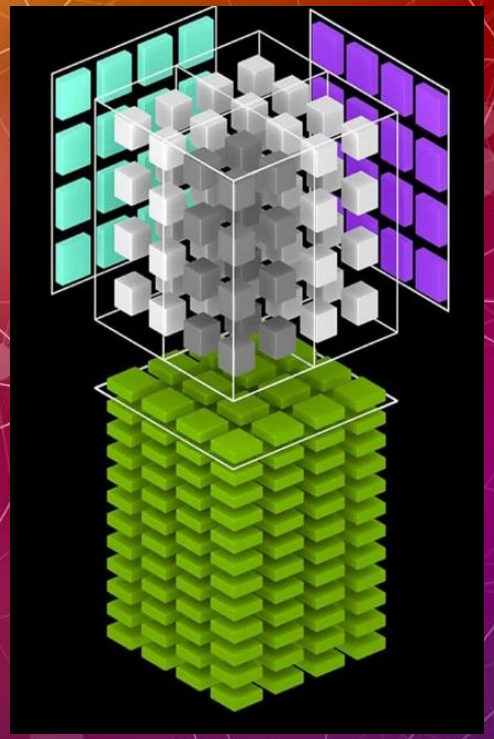
- Aceleración del entrenamiento e inferencia de redes neuronales.
- Aplicaciones en **transformers**, **CNNs**, vehículos autónomos y más.

Beneficio para mi investigación

- Aceleración en la regresión simbólica utilizando deep learning.
- Reducción del tiempo de entrenamiento de modelos complejos.
- Evaluación de la eficiencia en el uso de GPUs y paralelización.

Futuro de los Tensor Cores

- Nuevas generaciones: Soporte para TF32, FP64, ampliando su uso a ciencia e ingeniería.
- Separación de núcleos de IA: Núcleos dedicados en futuras arquitecturas (RTX 50).
- Expansión más allá de IA: Aplicaciones en medicina, biotecnología, energías renovables.
- Optimización en la comunicación: Eliminación de cachés entre núcleos IA.



Tensor Cores

Programación en GPUs

Jose Maureira
Magíster en Informática
Universidad Austral de Chile
23-09-2024