Reduciendo la huella de carbono en IA: Solapamiento eficiente en GPUs

Nicolás Meseguer, José L. Abellán y Manuel E. Acacio
Programa de Doctorado en Informática

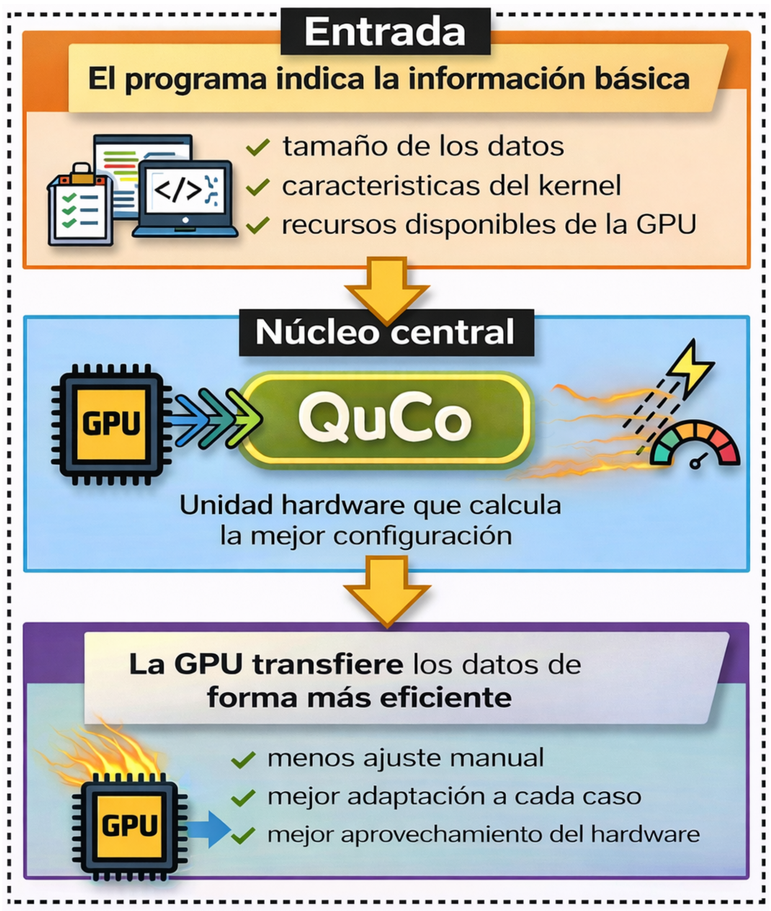
Idea Clave

Las GPUs actuales pueden realizar enormes cantidades de cálculo, pero a menudo se quedan esperando a que lleguen los datos. Poder solapar las transferencias de datos con el cómputo es clave para obtener un alto rendimiento. Sin embargo, configurar correctamente esas transferencias sigue siendo un proceso complejo.

Objetivos

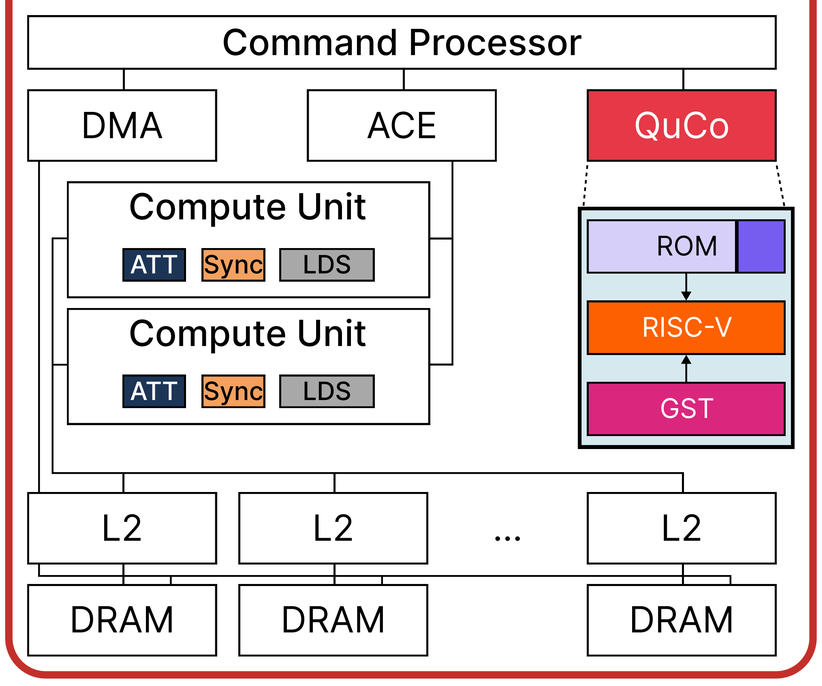
1. Configurar **automáticamente** las transferencias a cada arquitectura y aplicación.
2. Aprovechar mejor los **recursos** para lograr una computación **rápida** y **eficiente**.
3. Reducir el **consumo energético**.

La Solución: QuCo



QuCo (**Qu**eue **Co**nfigurator) [1] es una pequeña unidad hardware integrada en la GPU que **configura de manera óptima y automática** las transferencias de memoria. De este modo, evita búsquedas manuales costosas, adapta la ejecución a cada aplicación/arquitectura y ayuda a mantener un flujo constante de datos, aprovechando así mejor los recursos del sistema. Esto permite alcanzar un **alto rendimiento** de forma más sencilla y con mayor **portabilidad** entre distintas GPUs.

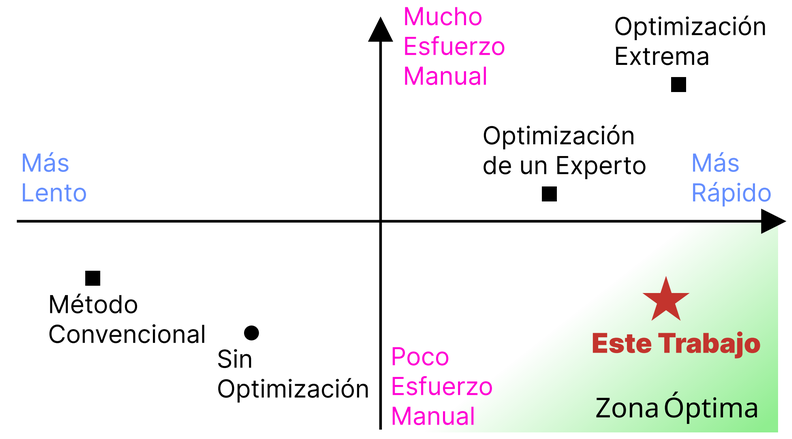
- Automatiza la configuración
- Se adapta a cada *kernel*/arquitectura
- Reduce la complejidad



El Problema: Mover Datos Cuesta

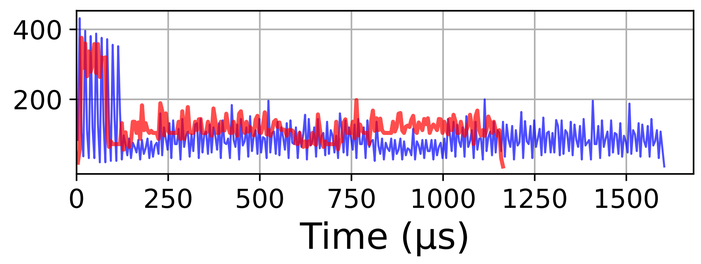
Configurar las transferencias de datos de forma eficiente es una tarea muy **compleja** y **dependiente de la aplicación y de la arquitectura**. Una búsqueda manual es tedioso y casi-imposible para aplicaciones reales. Cuando la configuración no es la adecuada, se pierde rendimiento y se desaprovechan recursos.

Idea clave: Si no se consigue un flujo de datos constante, no se aprovechan todos los recursos.

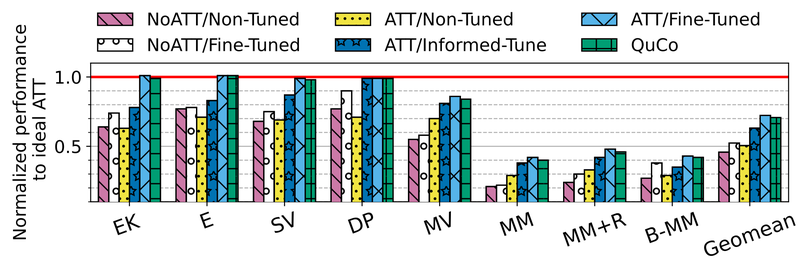


Idea clave: Buscamos obtener la mejor configuración con el menor esfuerzo posible.

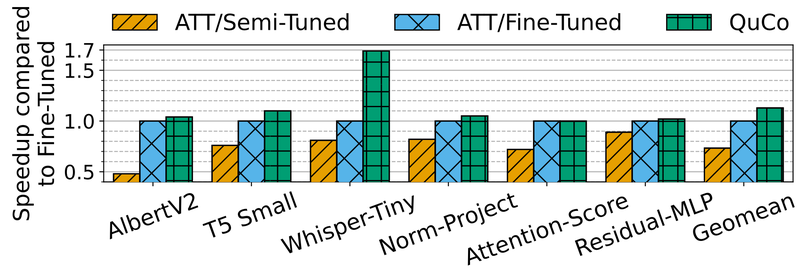
Resultados



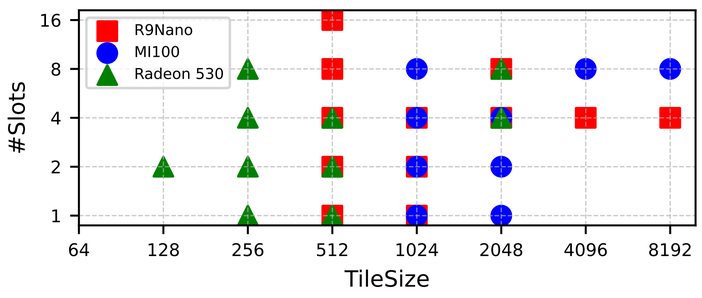
Actividad de la memoria principal (DRAM) usando QuCo (rojo) vs tradicional (azul) en una MxM.



Ejecuciones de diferentes versiones de *kernels* normalizados a un escenario ideal (más es mejor).



Aceleración lograda por QuCo en distintos modelos de IA del estado del arte (más es mejor).



Variabilidad de los parámetros seleccionados por QuCo según la GPU.

Conclusiones

1. Mejor rendimiento

QuCo acelera de forma consistente *kernels* y modelos actuales de **IA**, obteniendo aceleraciones de hasta **2**× sobre las versiones tradicionales.

2. Cercano al ideal

Alcanza un rendimiento muy próximo al **óptimo teórico** sin necesidad de costosas búsquedas ni ajustes manuales por parte del programador.

3. Se adapta solo

Ajusta su configuración según la **aplicación** y la **GPU**.

4. Más eficiente y sostenible

Aprovecha mejor el hardware y reduce el **consumo energético**.

Referencias

[1] N. Meseguer *et al.*, "Quco: Efficient and flexible hardware-driven automatic configuration of tile transfers in gpus," in *2026 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2026, pp. 1–14.

Agradecimientos

Proyecto de investigación financiado por el MCIN/AEI y el MICIU/AEI, con cofinanciación de fondos FEDER y NextGenerationEU. Nicolás Meseguer fue financiado bajo la beca FPI 21803/FPI/22 de la Fundación Séneca.