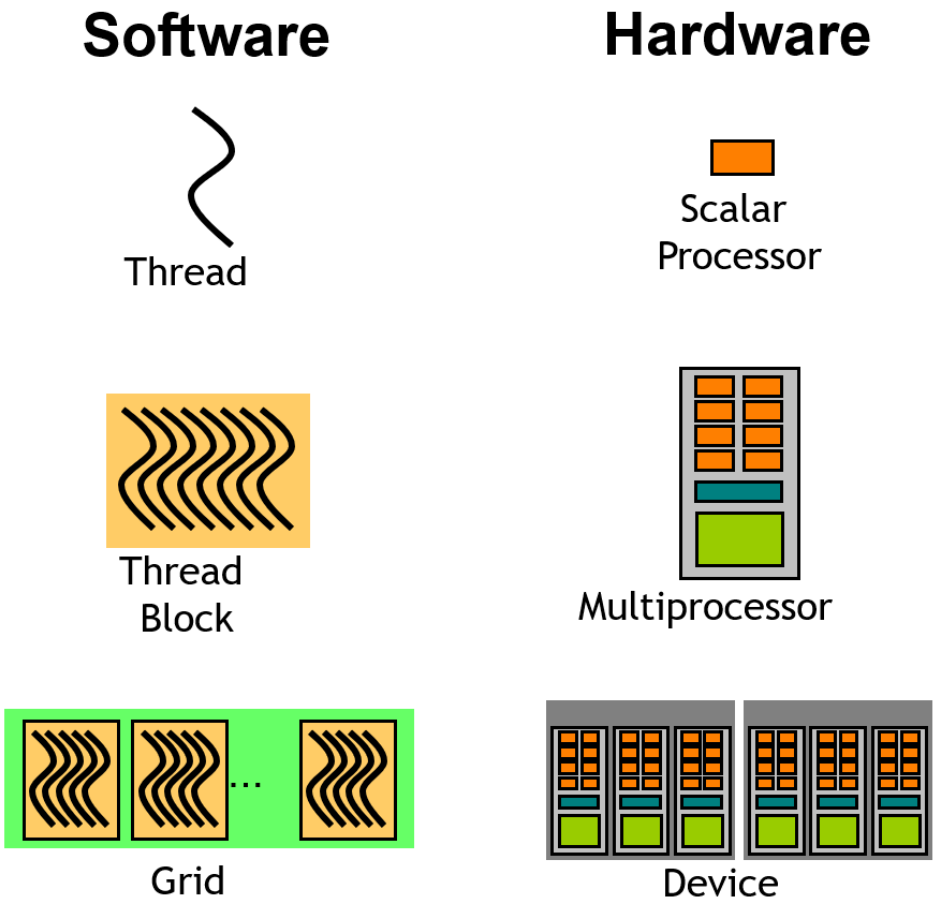
Tema 3: Paralelismo de datos a gran escala con GPU

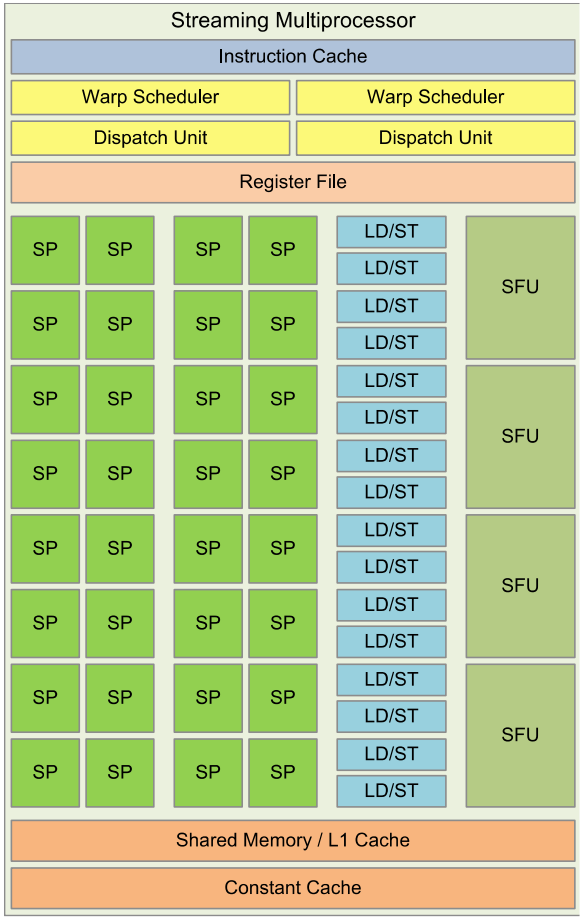
Un resumen rápido…

Hilo, warp, bloque y grid



- Warp: conjunto de 32 hilos que se ejecutan a la vez.
- Bloque: conjunto de hilos que se ejecuta en el mismo SM.
- Grid: conjunto de hilos que se ejecuta en el mismo dispositivo.

¿Qué es un SM?



Unidad de cómputo de la GPU. Contiene:

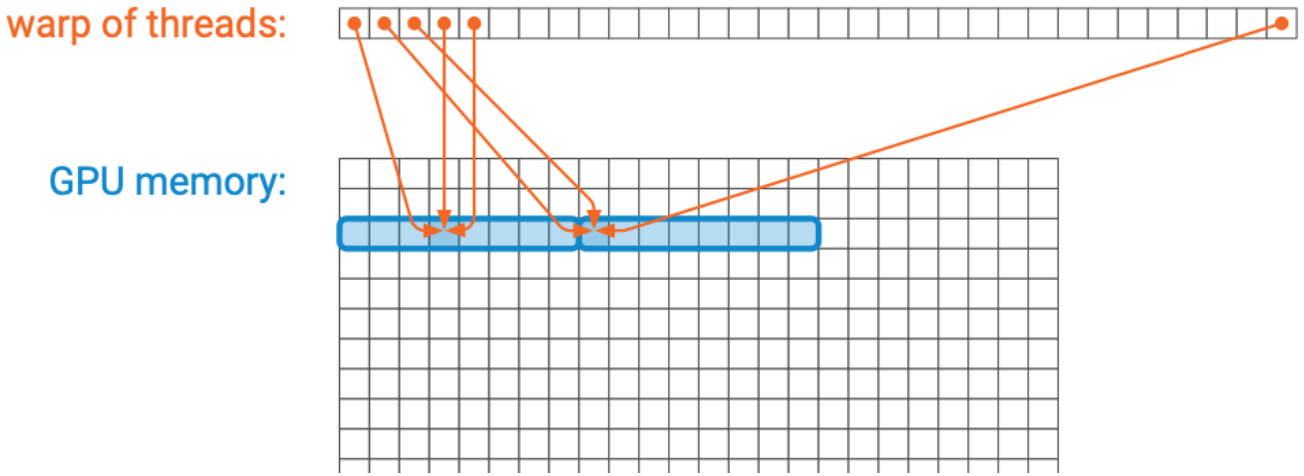
- Procesadores escalares (SPs).
- Unidades de funciones especiales (SFUs).
- Registros.
- Memoria compartida.

Jerarquia de memoria de la GPU

- Registros
  - Privada para cada hilo.
  - `float a=0.1;`
- Compartida
  - Privada para cada bloque, compartida para hilos del mismo bloque.
  - `__shared__ float a=0.1;`
- Global
  - Compartida para todo el grid.
  - `__device__ float a=0.1;`

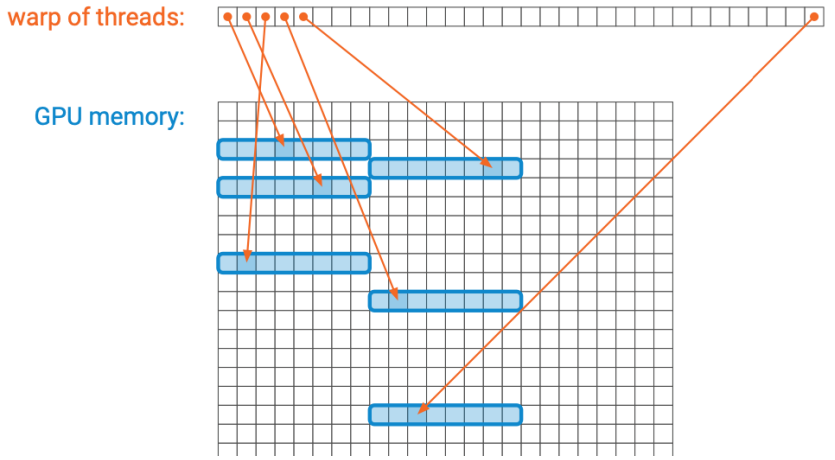
Coalescing

La GPU detecta accesos a memoria y los agrupa (coalesce) en accesos conjuntos para maximizar el ancho de banda



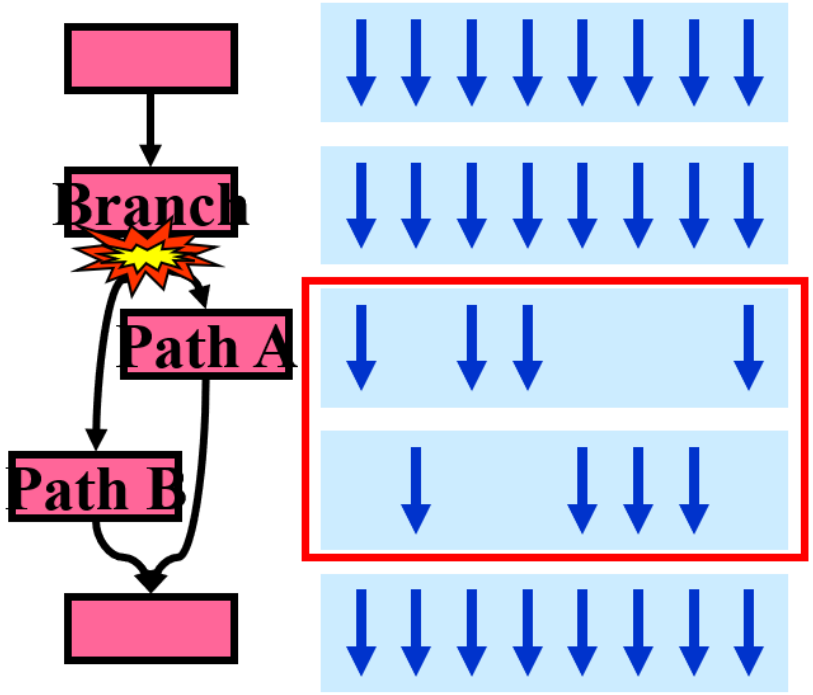
Problemas de memoria

- *Random access:* si no se accede a los datos de manera secuencial, gran parte de las lecturas/escrituras malgastan ancho de bando.
- *Unaligned access:* si las direcciones no están alineadas, puede que no usemos el ancho de banda al complete o que necesitemos más transferencias.



Divergencia

Si dentro de un mismo warp, dos hilos "toman" flujos de ejecución diferentes, se produce divergencia.



Soluciones a la divergencia

- Cambiar el algoritmo: juntamos en el mismo bloque hilos que no difieran mucho.
  - No siempre es posible.
- Dynamic warp forming
  - Juntamos hilos con mascaras compatibles.