SIMD (GPU) - CUDA
Repaso

Arquitectura y rendimiento

+Tarjeta: 14 SM de 128 cores / warp: 32 hilos / tamaño máximo de bloque: 1024 hilos
+Memoria: global, constante, compartida
+Rendimiento:
- Importancia de un uso eficiente de la memoria
- Tamaño de bloque múltiplo de warp (256-1024) + número suficiente de bloques

Kernels unidimensionales

+threadIdx.x / blockIdx.x / blockDim.x / gridDim.x
+Calcular idx
+kernel con un número de hilos = número de elementos a tratar
+kernel con un número de hilos < número de elementos a tratar

Kernels bidimensionales

+Tipo de datos dim3
+Entender la estructura (no aprender la fórmula para el cálculo de idx)

Sincronización y reducciones

+Todos los threads de un bloque (memoria compartida): __syncthreads
+Acceso atómico a una variable: atomicAdd, atomicSub, atomicMax, atomicMin
+Reducciones: entender los 2 algoritmos

Programación

+Kernel unidimensional (número de hilos / número de elementos a tratar)
+Variables estáticas o dinámicas, constantes en el device / compartidas en el kernel
+Pasos en el programa principal:
- gestionar la memoria en el device: cudaMalloc / memoria estática
- copiar la información host - device: cudaMemcpy / cudaMemcpyToSymbol
- gestionar el tamaño del bloque y el número de bloques
- llamar al kernel
- copiar la información device - host: cudaMemcpy
- eliminar memoria dinámica: cudaFree
+NO funciones de cálculo de tiempo de ejecución