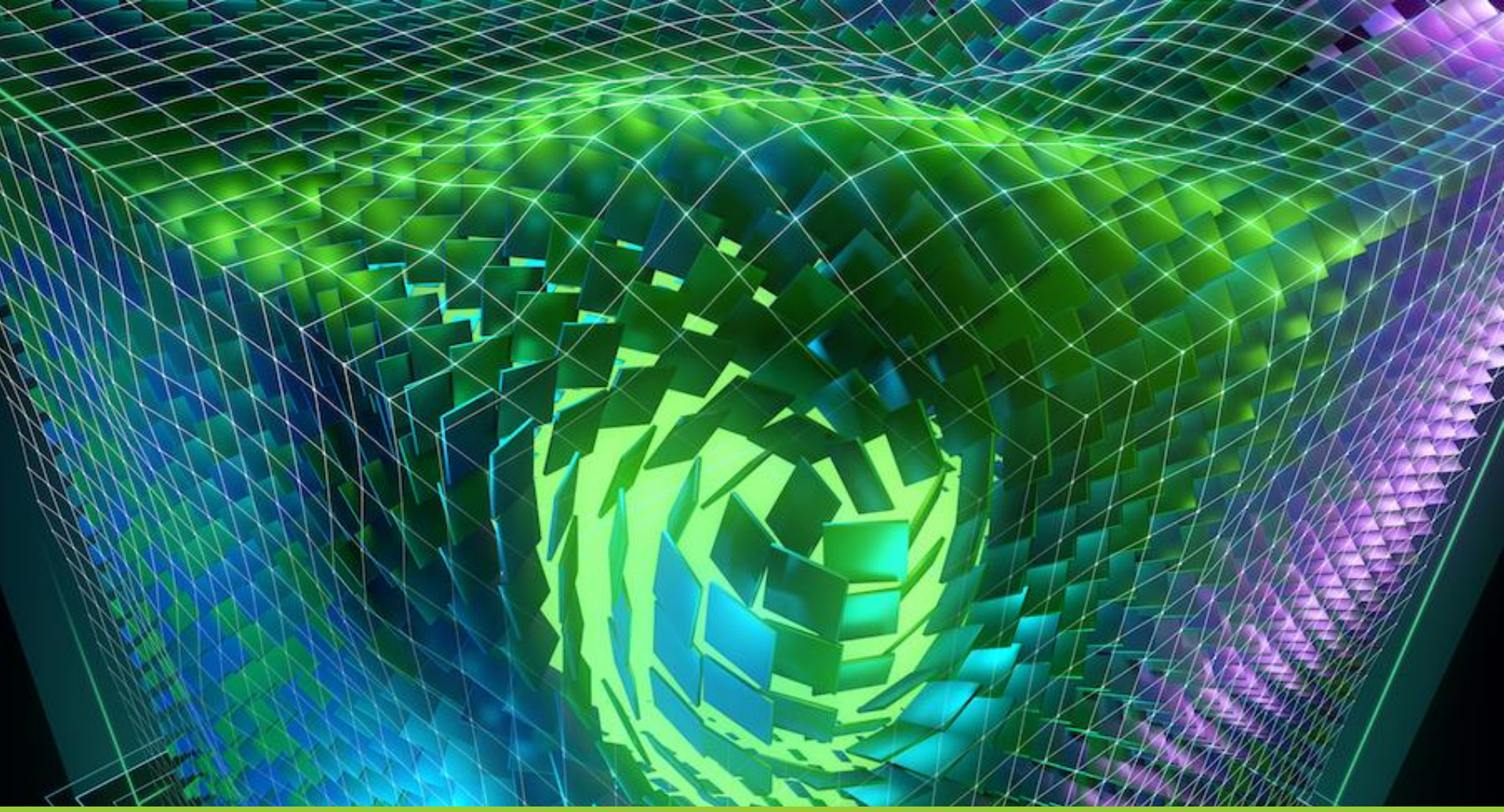


INF351 – Computación de Alto Desempeño

GPU: Unidad de Procesamiento Gráfico

PROF. ÁLVARO SALINAS

¿Qué es la GPU?

La GPU o unidad de procesamiento gráfico (**G**raphics **P**rocessing **U**nit) es el coprocesador presente en las tarjetas gráficas, siendo estas últimas una extensión de la placa madre en un computador.

Están optimizadas para el procesamiento de gráficos y operaciones aritméticas de punto flotante.

GPGPU Software

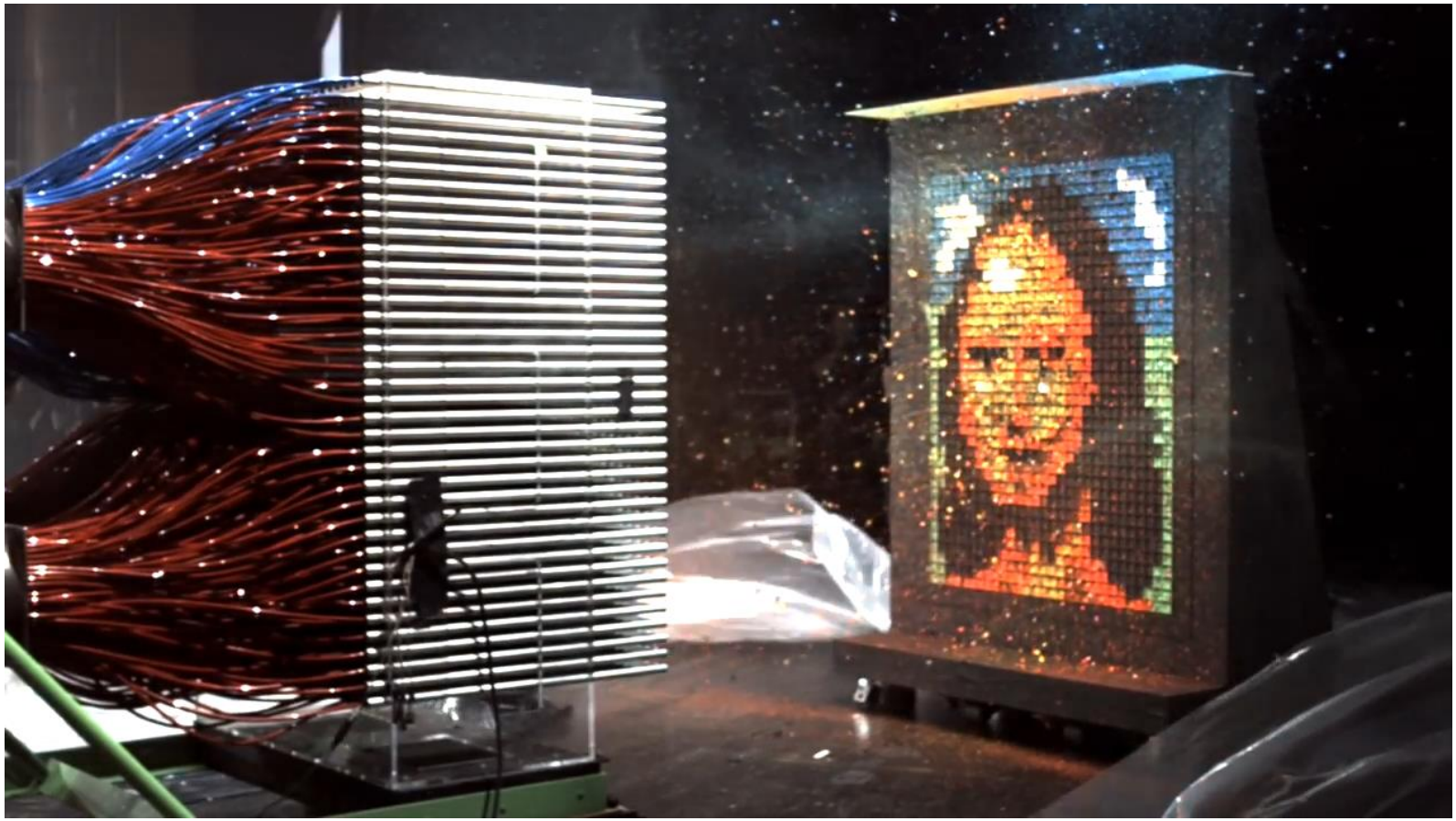
GPGPU (**G**erenal-**P**urpose **GPU**) es un concepto que hace referencia a la realización de cálculos no especializados que normalmente estarían a cargo de una CPU.

Debido a que las GPU normalmente se dedican al procesamiento de gráficos, cuando se aprovecha la potencia de una GPU para otro tipo de aplicaciones, e.g. científicas, estamos hablando de GPGPU.

GPU vs CPU – Aspectos en común

En general, una GPU es lo mismo que una CPU: muchos transistores en un circuito integrado capaces de realizar cálculos matemáticos a través de operadores binarios.

GPU vs CPU – Big picture



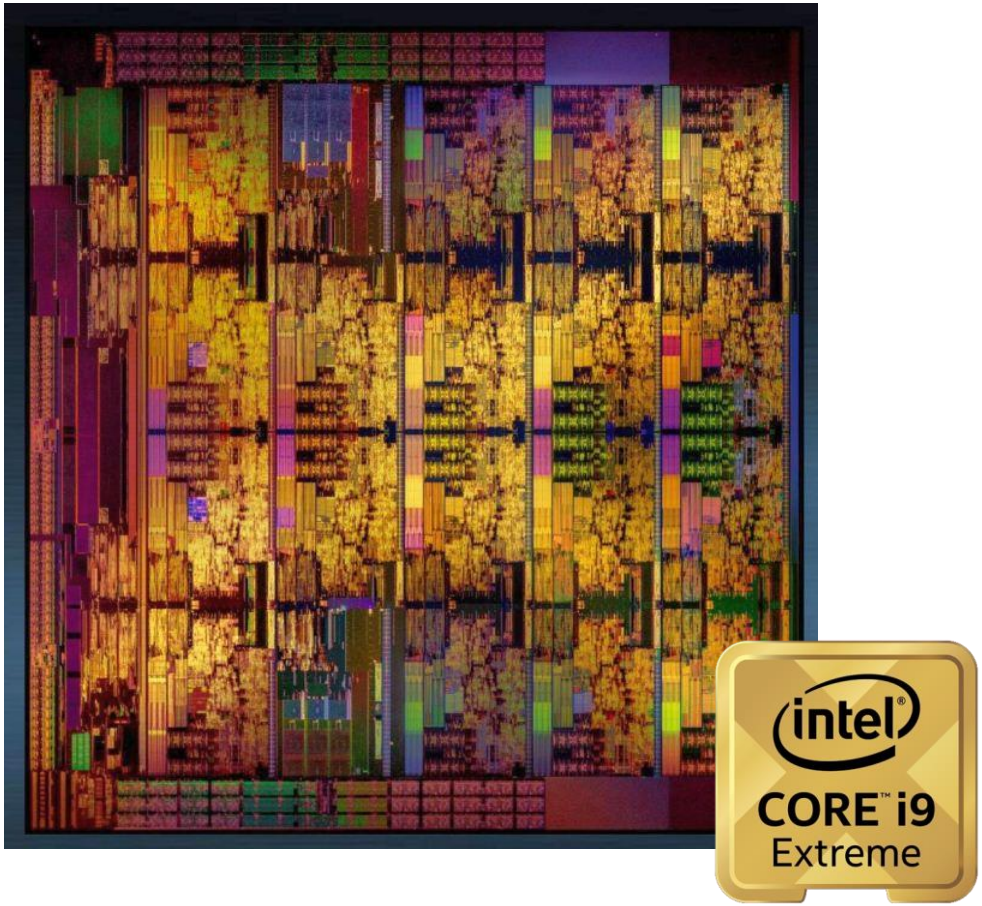
GPU vs CPU – Cantidad de núcleos

Una GPU posee una cantidad mayor de núcleos o cores que una CPU. Esto es lo que permite un paralelismo masivo de instrucciones.

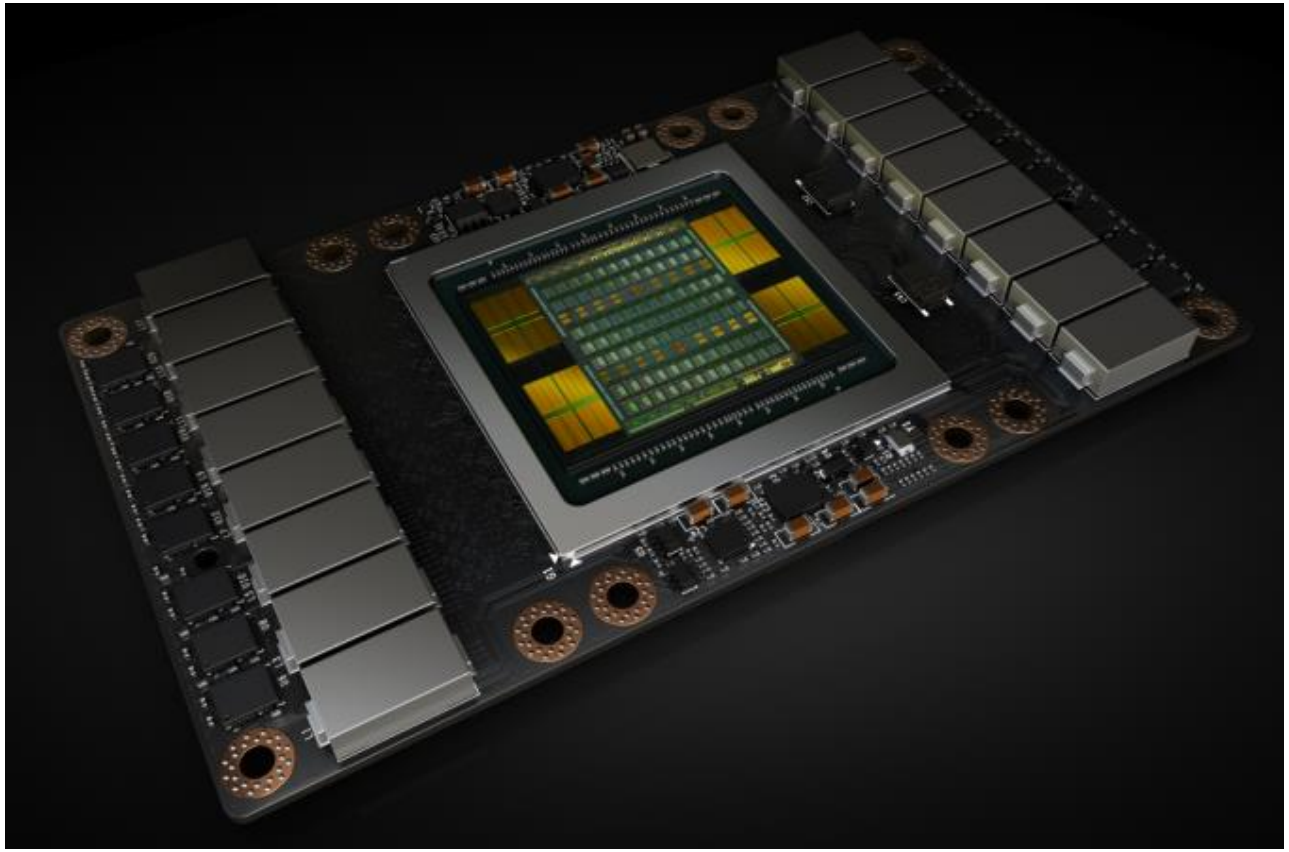
Ejemplo:

- Intel i9-7980XE posee 18 núcleos capaces de ejecutar 2 subprocesos cada uno, i.e. 36 núcleos lógicos.
- NVIDIA GV100 posee 5376 CUDA cores capaces de ejecutar 1 subproceso cada uno, i.e. 5376 cálculos en paralelo.

GPU vs CPU – i9-7980XE



GPU vs CPU – NVIDIA GV100



GPU vs CPU – NVIDIA GV100

GPU vs CPU – NVIDIA GV100

GPU vs CPU – Tipos de trabajos

Si bien las GPU están optimizadas para realizar cálculos de aritmética de punto flotante (en eso se basan los gráficos 3D), pueden realizar otros tipos de tareas también sencillas, pero siempre obtienen el mejor rendimiento posible cuando todos los subprocesos realizan el mismo trabajo.

Siguiendo otra arquitectura de trabajo, las CPU realizan de forma más eficiente tareas diversas y de mayor complejidad, e.g. la carga de programas.

GPU vs CPU – Frecuencia

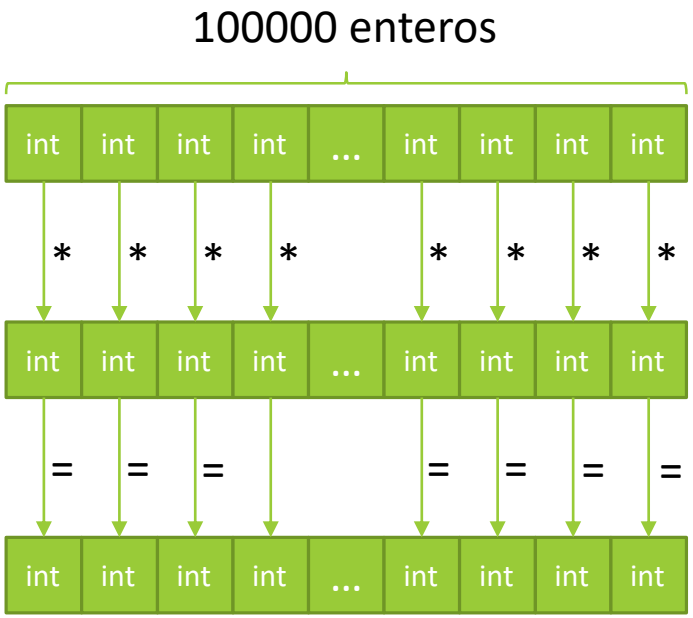
Las CPU modernas rondan los 4 GHz mientras que las GPU presentan frecuencias cercanas a los 1.5 GHz.

Tomemos un ejemplo y hagamos algunos cálculos básicos bajo algunos supuestos:

- Consideremos una operación que toma 4 clock cycles, e.g. una multiplicación de enteros.
- Asumamos que la cantidad de ciclos por operación es la misma en la GPU y la CPU.
- No consideremos otros aspectos que podrían interferir, e.g. accesos a memoria o inicializaciones.
- La CPU a utilizar será un Intel i9-7980XE, la cual alcanza los 4.2 GHz y, según dijimos, puede ejecutar 36 subprocesos en paralelo.
- La GPU a utilizar será una NVIDIA GV100, la cual alcanza los 1.53 GHz y, según dijimos, puede ejecutar 5376 subprocesos en paralelo.

GPU vs CPU – Frecuencia

Primer caso:



CPU:

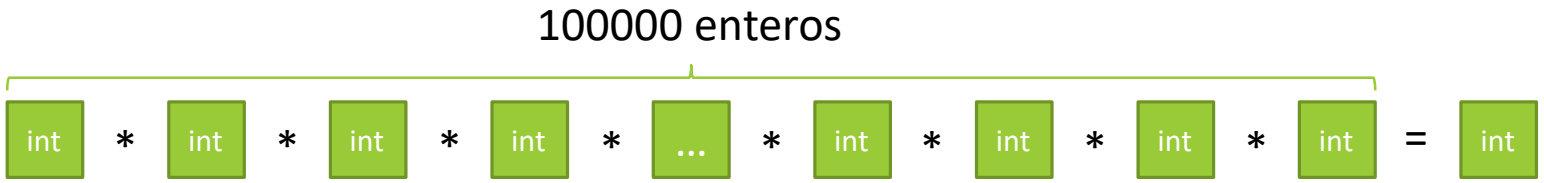
- Número de subprocesos secuenciales: $[100000/36] = 2778$
- Ciclos totales: $2778 \times 4 = 11112[ciclos]$
- Tiempo: $\frac{11112[ciclos]}{4.2\times10^9[Hz]} = 2.65 \times 10^{-6}[s]$

GPU:

- Número de subprocesos secuenciales: $[100000/5376] = 19$
- Ciclos totales: $19 \times 4 = 76[ciclos]$
- Tiempo: $\frac{76[ciclos]}{1.53\times10^9[Hz]} = 4.97 \times 10^{-8}[s]$

GPU vs CPU – Frecuencia

Segundo caso:



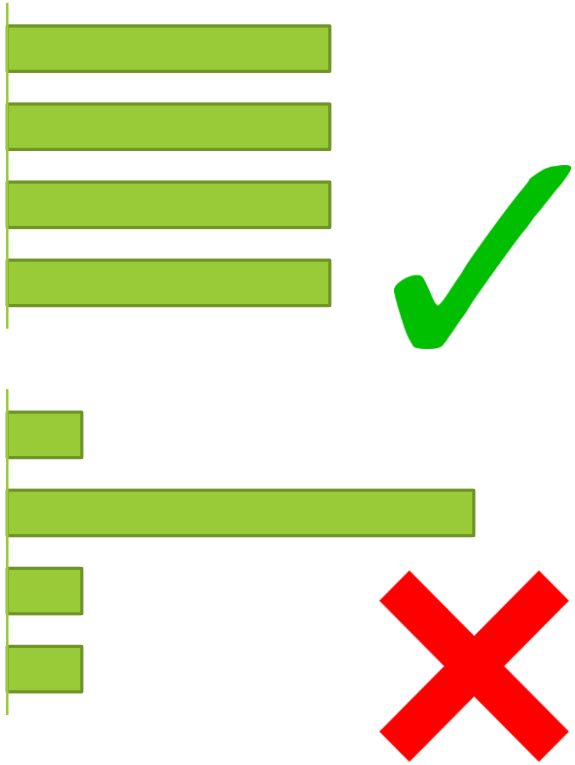
CPU:

- Número de subprocesos secuenciales: 10^5
- Ciclos totales: $10^5 \times 4 \ = 4 \times 10^5 [ciclos]$
- Tiempo: $\frac{4 \times 10^5 [ciclos]}{4.2 \times 10^9 [Hz]} = 9.52 \times 10^{-5} [s]$

GPU:

- Número de subprocesos secuenciales: 10^5
- Ciclos totales: $10^5 \times 4 \ = \ 4 \times 10^5 [ciclos]$
- Tiempo: $\frac{4 \times 10^5 [ciclos]}{1.53 \times 10^9 [Hz]} = 2.61 \times 10^{-4} [s]$

Principales problemas: Regularización



Principales problemas: Bandwidth

Principales problemas: Serialización