Introducción a la Programación Paralela con CUDA

Introducción

La programación es el acto de escribir código fuente para crear un programa para un computador. Cómo sabemos los computadores son máquinas que ejecutan instrucciones de forma secuencial, es decir, una instrucción a la vez. Por lo tanto, la programación secuencial es la forma natural de programar para un computador. Sin embargo, los computadores modernos tienen múltiples núcleos de procesamiento y pueden ejecutar múltiples instrucciones al mismo tiempo. La programación paralela es la forma de programar para aprovechar estos múltiples núcleos de procesamiento y ejecutar múltiples instrucciones al mismo tiempo.

Cuando queremos empezar a hablar de la programación tenemos que tener en claro conceptos clave, siendo el primero de ellos el de algoritmo. Un algoritmo es un conjunto de instrucciones que resuelven un problema. Por ejemplo, el algoritmo para sumar dos números es el siguiente:

1. Leer el primer número
2. Leer el segundo número
3. Sumar los dos números
4. Imprimir el resultado
5. Fin

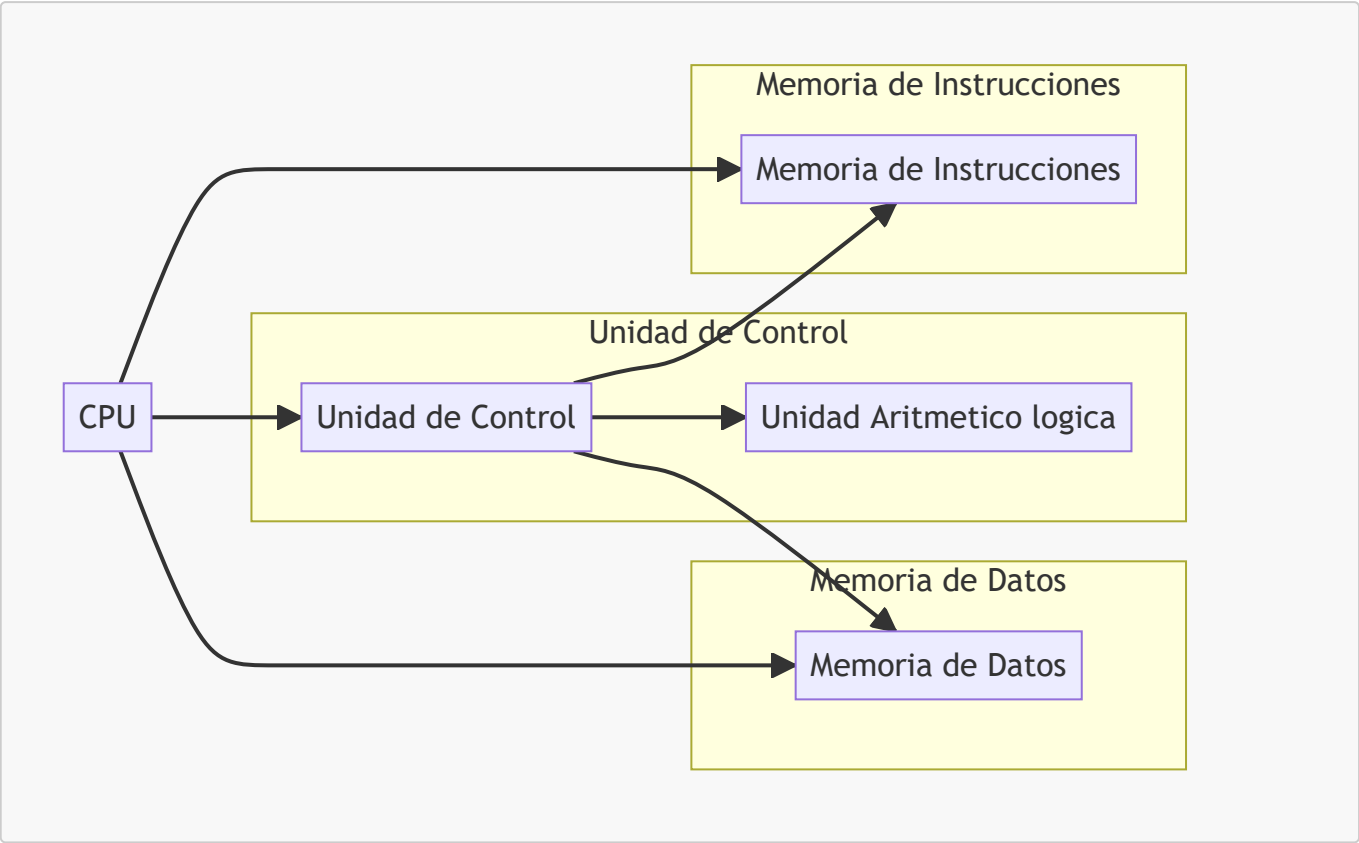
Esto es un algoritmo, pero no es un programa. Un programa es la implementación de un algoritmo en un lenguaje de programación. Por ejemplo, el algoritmo anterior implementado en C es el siguiente:

```c
#include <stdio.h>

int main() {
    int a, b, c;
    scanf("%d", &a);
    scanf("%d", &b);
    c = a + b;
    printf("%d\n", c);
    return 0;
}
```

Una vez que esto queda claro, podemos explicar porque siempre la programación se ha visto de manera secuencial. Esto viene dado por un concepto conocido como arquitectura de Von Neumann. Esta arquitectura es la que tienen todos los computadores modernos y se basa en la idea de que el programa y los datos están en la misma memoria. Por lo tanto, el programa se ejecuta de manera secuencial, leyendo instrucción por instrucción y ejecutándolas una a la vez. Esto se puede ver en la siguiente figura:

Tambien existe la arquitectura Harvard, en la cual el programa y los datos están en memorias separadas. Esto permite que el programa se ejecute de manera paralela, leyendo y ejecutando múltiples instrucciones al mismo tiempo. Esto se puede ver en la siguiente figura:



La parte más importante de la computación de alto rendimiento es la Unidad de Procesamiento Central, o CPU. Esta es la parte del computador que ejecuta las instrucciones. La CPU tiene una serie de registros, que son pequeñas memorias que se encuentran dentro de la CPU. Estos registros son muy rápidos, pero también son muy pocos. Por lo tanto, la CPU tiene una memoria externa, que es la memoria RAM. Esta memoria es más lenta que los registros, pero tiene mucha más capacidad. Por lo tanto, la CPU tiene que leer los datos de la memoria RAM y guardarlos en los registros para poder operar con ellos. Además de esto el procesador cuenta con las Unidades Aritmético Lógicas, o ALU. Estas son las que realizan las operaciones aritméticas y lógicas. La CPU también cuenta con la Unidad de Control, que es la que se encarga de leer las instrucciones y decirle a la ALU que operación tiene que realizar. Por último, la CPU cuenta con la Unidad de Punto Flotante, o FPU. Esta es la que se encarga de realizar las operaciones aritméticas con números de punto flotante.

Los procesadores actuales dada la arquitectura con la que están construidos pueden contar con múltiples núcleos de procesamiento. Esto quiere decir que pueden ejecutar múltiples instrucciones al mismo tiempo. Esto se puede ver en la siguiente figura: