**COS’ E’ CUDA**

CUDA ( Compute Unfied Device Architecture) è un architettura hardware creata da NVIDIA pensata per l’elaborazione parallela che si basa dunque sulla suddivisione di un problema in sottoproblemi e lo svolgimento di ognuno di questi concorrentemente.

Tutto ciò è possibile sfruttando la potenza di calcolo della GPU (unità di elaborazione grafica) che assieme alla CPU permette di rendere molto efficiente in termini di tempo l’esecuzione di molte applicazioni.

Il calcolo parallelo su GPU offre prestazioni senza precedenti caricando porzioni compute-intensive dell'applicazione sulle GPU, mentre la porzione rimanente di codice continua ed eseguire su CPU. Pertanto, le GPU devono operare congiuntamente con le CPU mediante bus PCI-Express

cpu-gpu.png

Le applicazioni in cui CPU e GPU lavorano insieme consistono di due parti:

* **codice Host**
* **codice Device**

Il codice Host esegue su CPUs e il codice device esegue su GPUs. Il codice CPU è responsabile della gestione dell'ambiente, dell'IO e della gestione dei dati per il device stesso, prima di caricare task intensivi sul device. La GPU è usata per accelerare l'esecuzione di questa porzione di codice basandosi sul parallelismo dei dati. La CPU è ottimizzata per sequenze di operazioni in cui il controllo del flusso è impredicibile; le GPU lo sono per carichi dominati da semplice flusso di controllo.

Performancecpu-gpu.png

**Il modello di programmazione CUDA**