Isaac Lima – CUDA

O que é CUDA:

CUDA® é uma plataforma de computação paralela e um modelo de programação desenvolvido pela NVIDIA para computação geral em unidades de processamento gráfico (GPUs). Com o CUDA, os desenvolvedores podem acelerar drasticamente os aplicativos de computação, aproveitando o poder das GPUs.

Em aplicativos acelerados por GPU, a parte sequencial da carga de trabalho é executada na CPU - que é otimizada para desempenho de encadeamento único - enquanto a parte de cálculo intensivo do aplicativo é executada em milhares de núcleos de GPU em paralelo. Ao usar o CUDA, os desenvolvedores programam em linguagens populares como C, C ++, Fortran, Python e MATLAB e expressam o paralelismo através de extensões na forma de algumas palavras-chave básicas.

Como funciona:

O CUDA funciona como uma plataforma comum em todas as famílias de GPUs NVIDIA, para que você possa distribuir e dimensionar seu aplicativo nas configurações da GPU.

As primeiras GPUs foram projetadas como aceleradores gráficos, tornando-se mais programáveis ​​nos anos 90, culminando na primeira GPU da NVIDIA em 1999. Pesquisadores e cientistas rapidamente começaram a aplicar o excelente desempenho de ponto flutuante dessa GPU para computação de propósito geral. Em 2003, uma equipe de pesquisadores liderada por Ian Buck revelou o Brook, o primeiro modelo de programação amplamente adotado para estender o C com construções de dados paralelos. Mais tarde, Ian Buck ingressou na NVIDIA e liderou o lançamento do CUDA em 2006, a primeira solução do mundo para computação geral em GPUs.

Desde a sua criação, o ecossistema CUDA cresceu rapidamente para incluir ferramentas de desenvolvimento de software, serviços e soluções baseadas em parceiros.

O CUDA acelera os aplicativos em uma ampla variedade de domínios, desde o processamento de imagens até o aprendizado profundo, análise numérica e ciência computacional.

Referencias Bibliograficas

<https://developer.nvidia.com/about-cuda>

<https://developer.nvidia.com/cuda-zone>