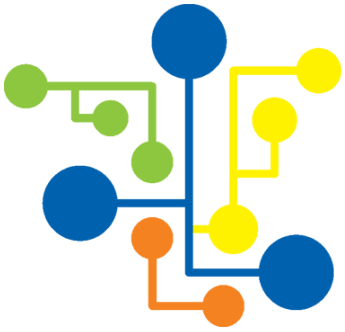


Data Science Academy

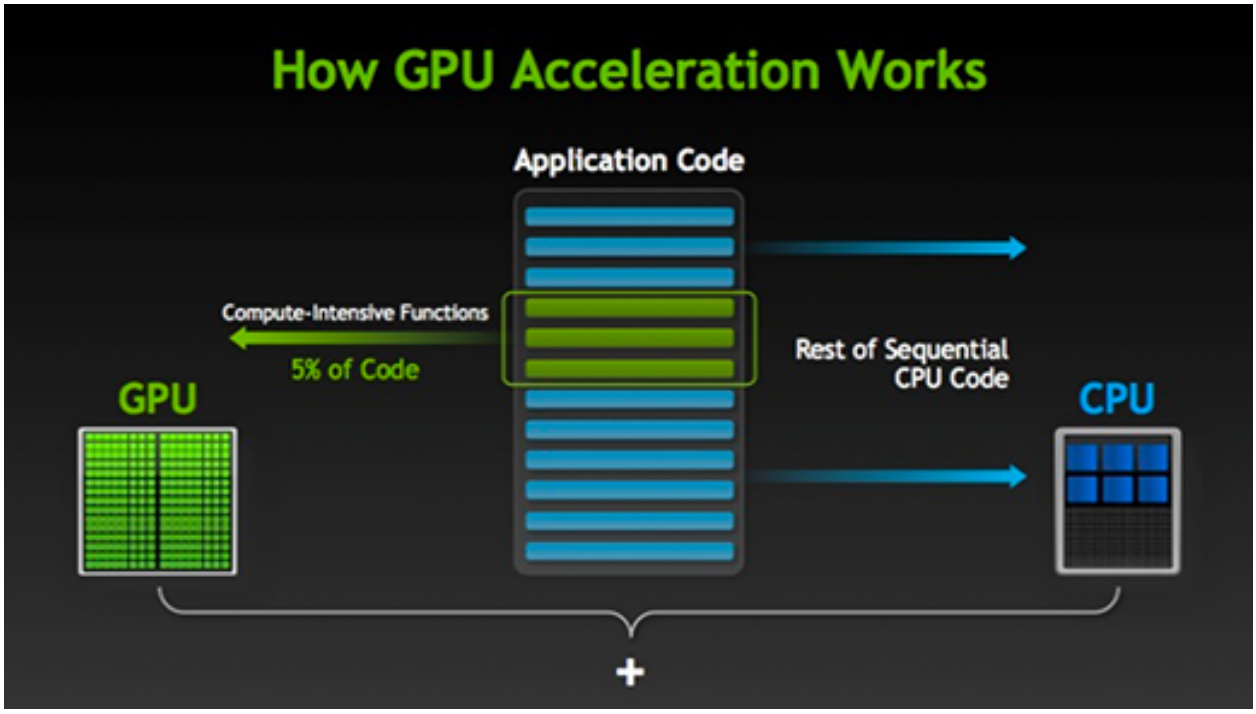
www.datascienceacademy.com.br

Programação Paralela em GPU

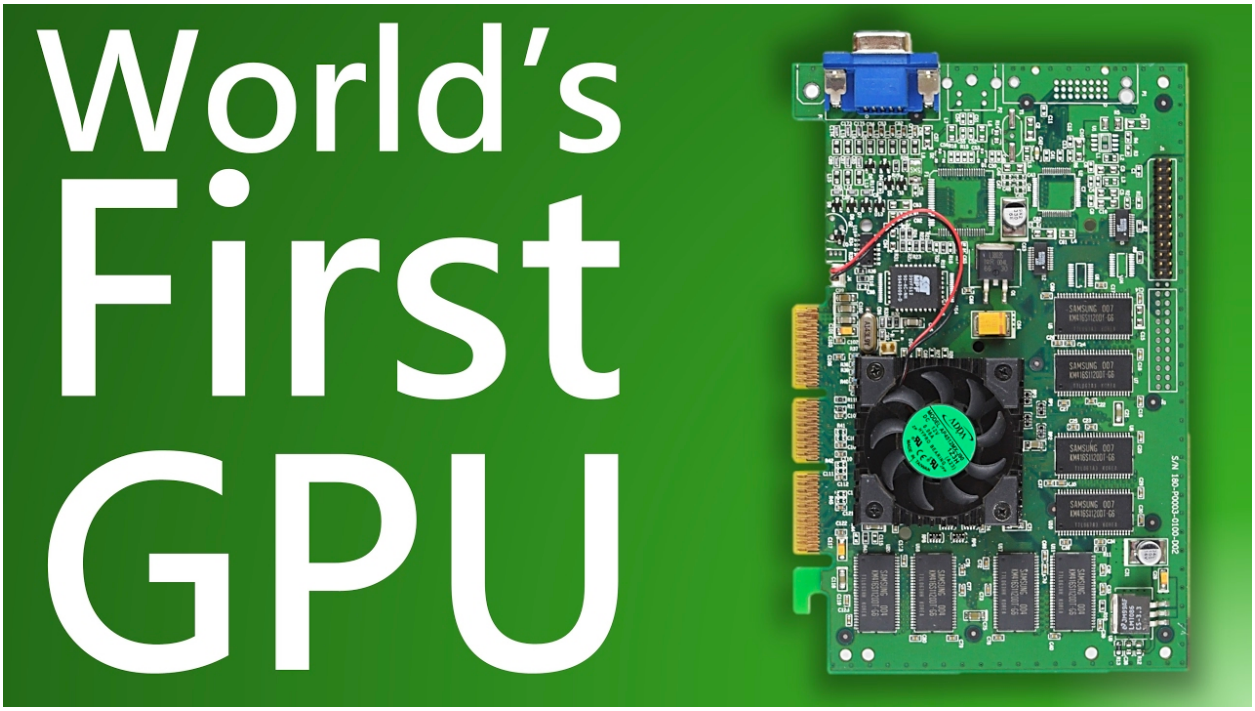
Computação em GPU Nvidia

Atualmente já existe um consenso em relação a importância das Graphical Processing Units (GPUs) em sistemas computacionais. O crescimento de sua capacidade de processamento e quantidade de aplicações que tiram proveito de sua computação de alto desempenho vem aumentando desde a sua criação em 1999. Sua arquitetura se baseia em um grande numero de núcleos para processamento massivo paralelo com foco na ciência energética. Essas placas também podem ser utilizadas para aplicações de propósito geral (técnica denominada GPGPU) atraves da programação em CUDA (Compute Unified Device Architecture).

Desde a criação da computaçãao, os processadores operavam como um único nucleo programável, contendo códigos executados sequencialmente. Porem, a demanda por maior desempenho fez com que empresas investissem muito dinheiro no aumento de sua velocidade, exigindo cada vez mais do silício. Para contornar o limite físico do silício, foram criados os mecanismos de pipeline, threads entre outros. Adicionar processadores em paralelo ou vários núcleos em um único chip foram outras alternativas para aumentar seu desempenho. Porem, para aproveitar-se destes tipos de processamento é preciso mudar os códigos sequenciais e buscar por técnicas de software que possam aproveitar esses novos recursos.



Ao perceber a alta demanda por um tipo de processamento específico, a empresa americana NVIDIA, conhecida pela fabricação de placas de vídeo, lançou, em 1999, sua primeira GPU: a GeForce 256 e logo mais a frente, em 2000, criou as GPUs de propósito geral (GPGPU - General-Purpose Computing on Graphics Processing Units). Hoje a NVIDIA possui mais de 1 bilhão de GPUs vendidas.



Inicialmente, a GPU era um processador de funcão fixa, construído sobre um pipeline gráfico que se sobressaía apenas em processamento de gráficos em três dimensões. Desde então, a GPU tem melhorado seu hardware, com foco no aspecto programável da GPU. O resultado disso é um processador com grande capacidade aritmética, não mais restrito as operações gráficas.

Atualmente, as GPUs são processadores dedicados para processamento gráfico da classe SIMD (Single Instruction Multiple Data). GPUs são desenvolvidas especificamente para cálculos de ponto flutuante, essenciais para renderização de imagens. Suas principais características são sua alta capacidade de processamento massivo paralelo e sua total programabilidade e desempenho em cálculos que exigem um volume grande de dados, resultando em um grande throughput. A partir de APIs como DirectX, OpenGL e Cg, foi possível criar algoritmos paralelos para GPUs. Entretanto, isso requeria ao programador um grande conhecimento das APIs e da arquitetura das placas gráficas, alem de ser necessario que o problema seja representado atraves de coordenadas de vértices, texturas e shaders, aumentando drasticamente a complexidade do programa.

Para facilitar a interface entre o programador e as aplicações GPU, a NVIDIA apresentou a Compute Unified Device Architecture (CUDA) em 2006. Trata-se de uma plataforma de computaçãao paralela e modelo de programação que disponibiliza um aumento significativo de desempenho ao aproveitar o poder da GPU. Ao fornecer abstrações simples com respeito a organização hierárquica de threads, memória e sincronização, o modelo de programação CUDA permite aos programadores escreverem programas escaláveis sem a necessidade de aprender a

multiplicidade de novos componentes de programação. A arquitetura CUDA suportas diversas linguagens e ambientes de programação, incluindo C, Fortran, OpenCL, e DirectX Compute. Ela tambem tem sido amplamente utilizada por aplicações e trabalhos de pesquisa publicados. Hoje ela esta implantada em notebooks, estações de trabalho, clusters de computaçãao e supercomputadores

As GPUs estão rapidamente se aperfeiçoando, tanto em programabilidade quanto em capacidade. A taxa de crescimento computacional anual das GPUs esta aproximadamente em 2.3x, em contraste a isso, o das CPUs esta em 1.4x. Ao mesmo tempo, as GPUs estão ficando cada vez mais acessíveis. Com isso, a comunidade de pesquisa tem mapeado com sucesso uma rápida demanda por soluções computacionais complexas com um surpreendente resultado. Este fato tem posicionado a GPU como uma alternativa aos microprocessadores nos futuros sistemas de alto desempenho.