COMPUTAÇÃO ESCALÁVEL

Frameworks para Computação Paralela - CUDA

ROTEIRO

- Introdução ao CUDA
- Estrutura do CUDA
- Organização das Threads
- Modelo de Memória
- CUDA em Docker e Cloud
- Utilização do CUDA

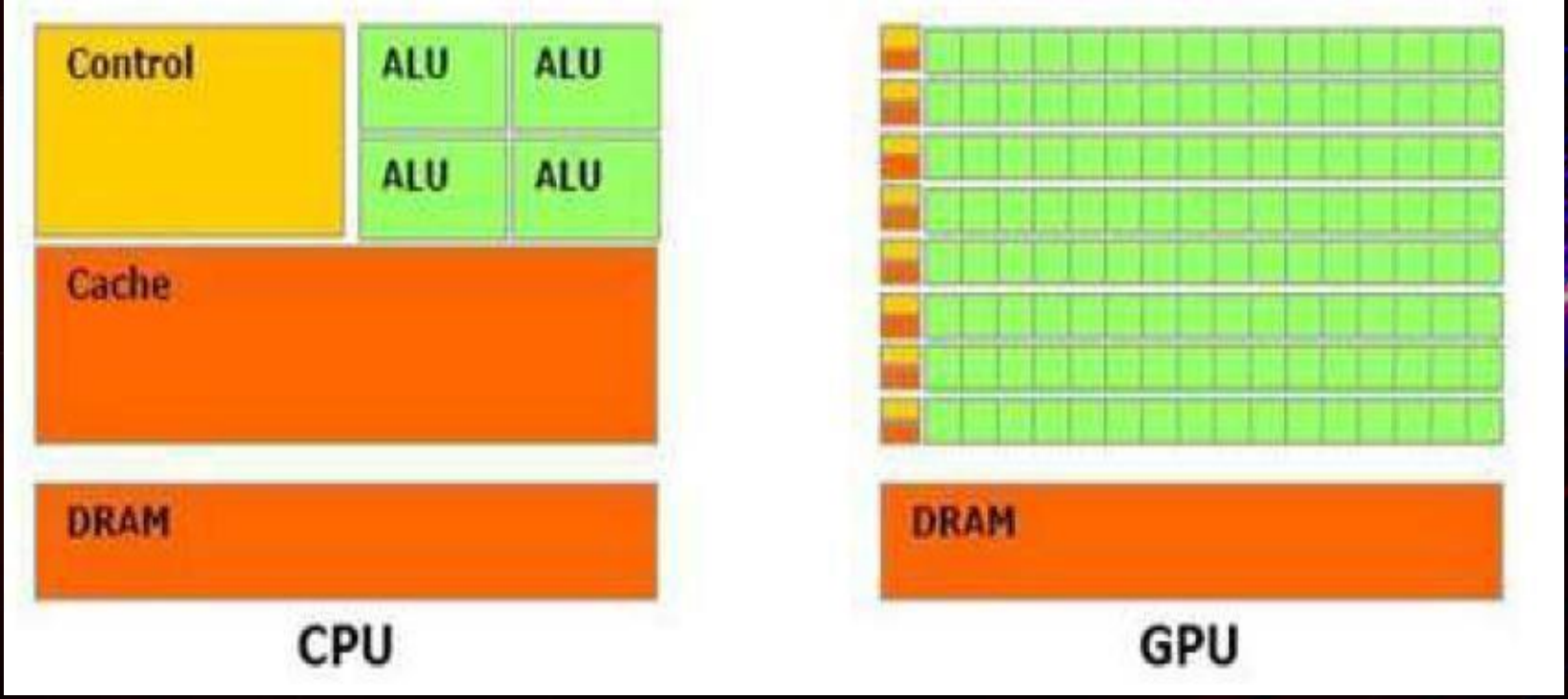
INTRODUÇÃO AO CUDA

- CUDA é uma plataforma de software e hardware para computação paralela de alto desempenho que utiliza o poder de processamento dos núcleos das GPUs da NVIDIA

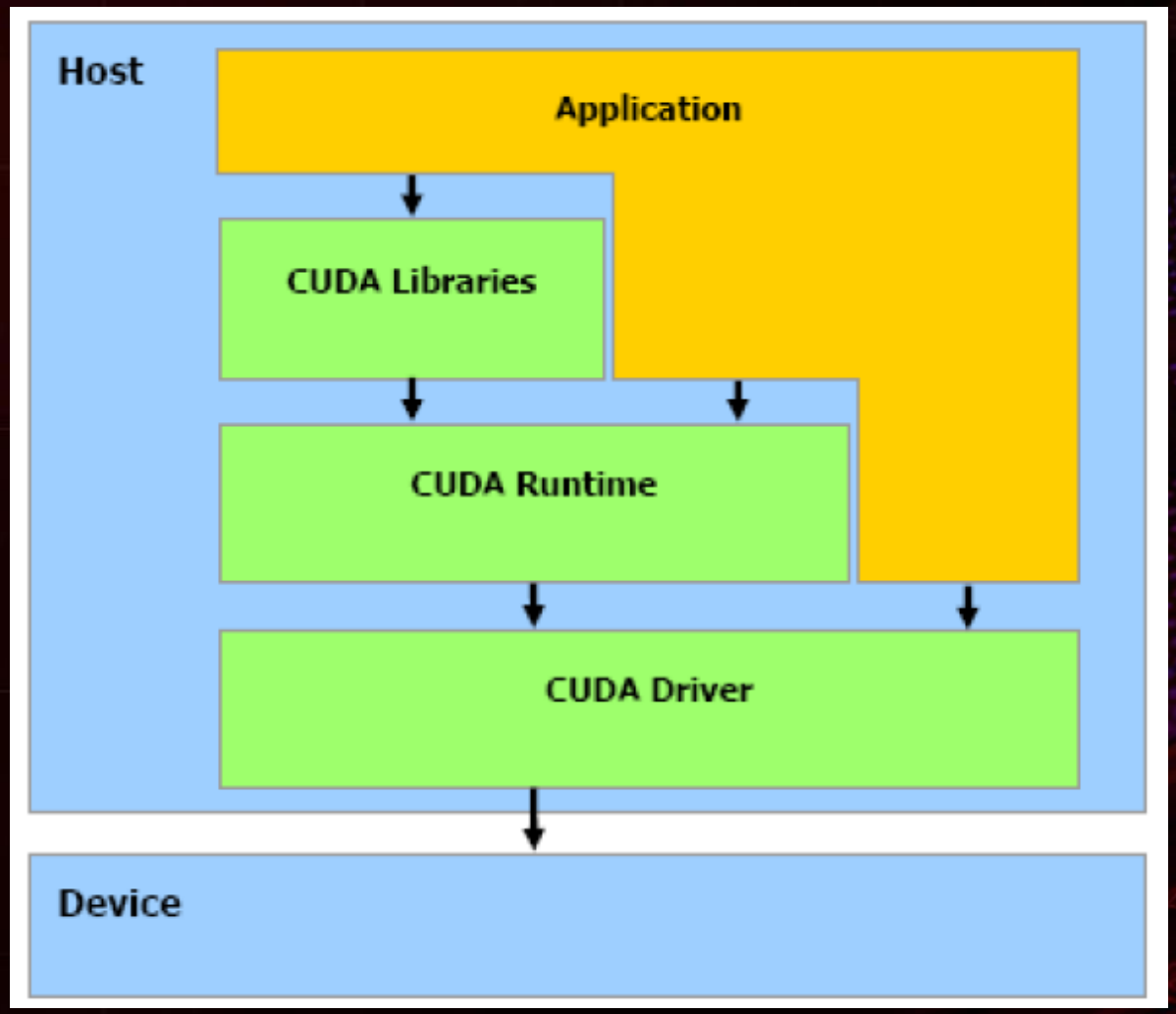
- Possui demanda nos campos matemático, científico e biomédico, além das demandas na computação e na engenharia devido às características das aplicações nesses campos, que são altamente paralelizáveis

INTRODUÇÃO AO CUDA

- O motivo para as GPUS serem tão eficientes é que elas são exclusivamente dedicadas a processar dados não tendo que guardar informações de cache ou controlar fluxos de dados



ESTRUTURA DO CUDA

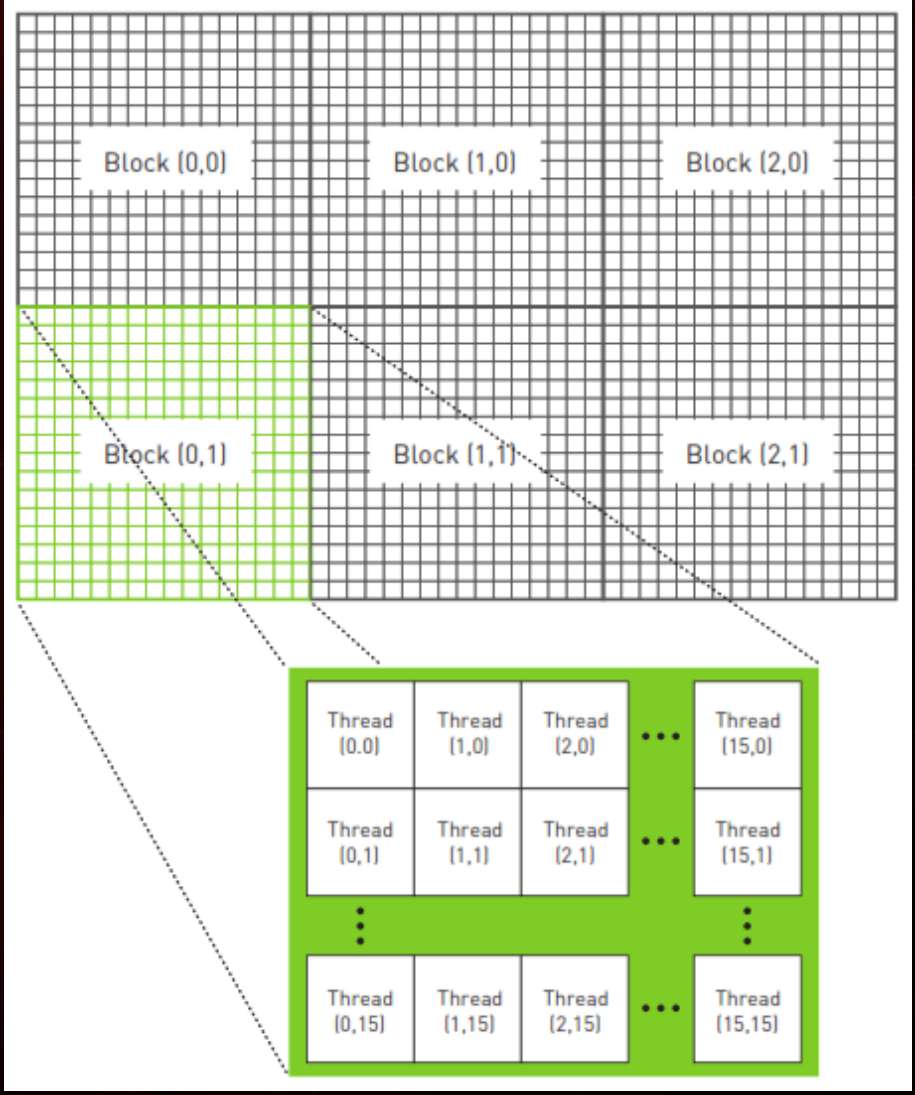


ORGANIZAÇÃO DAS THREADS DO CUDA

- O escalonamento das threads da plataforma CUDA utiliza dois conceitos: bloco e grid. Através deles se organiza a repartição dos dados entre as threads, a organização e a distribuição dos dados ao hardware

- Em geral, cada bloco dispõe de duas ou três coordenadas dimensionais dadas por palavras chave: **blockIdx.x, blockIdx.y e blockIdx.z**

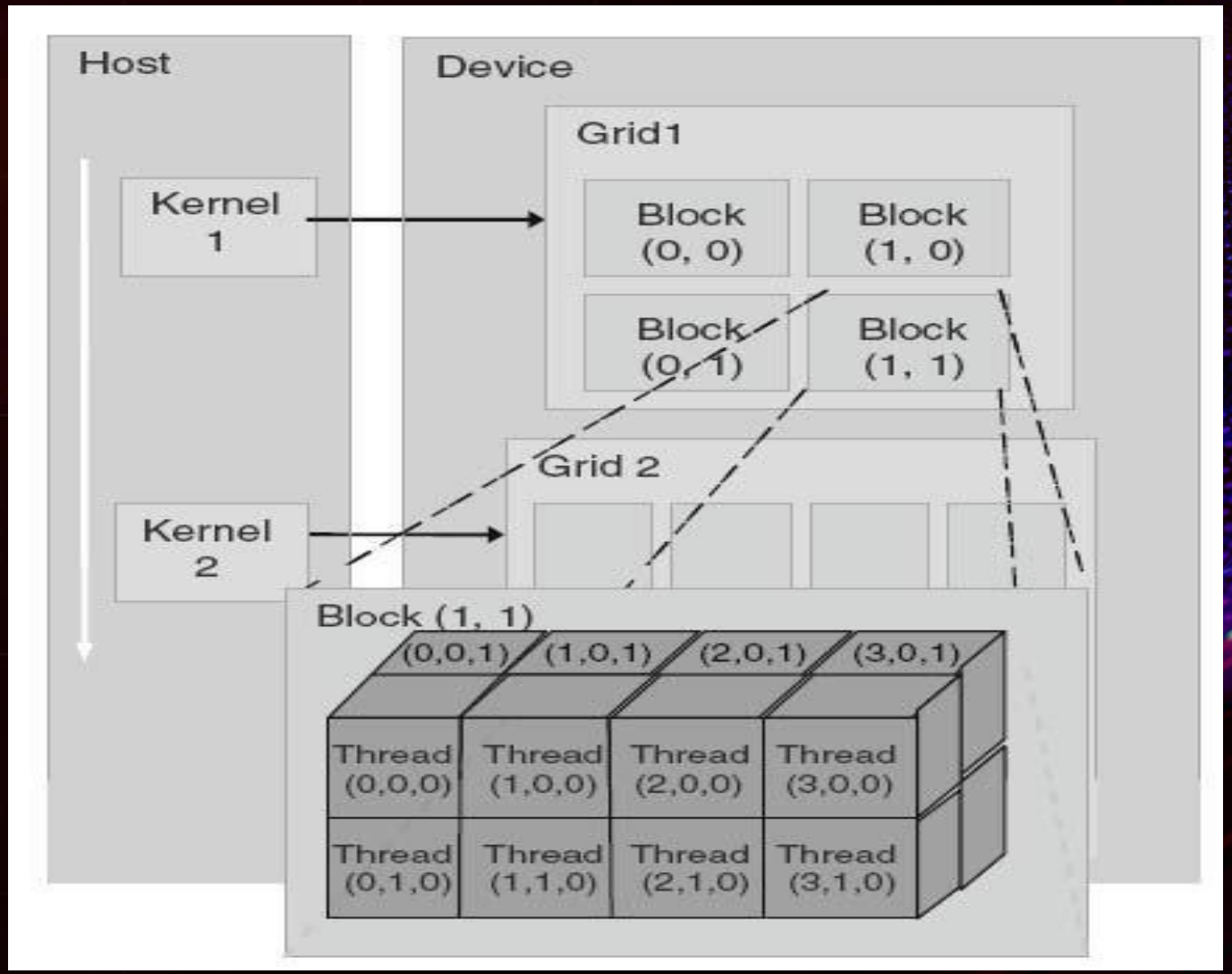
ORGANIZAÇÃO DAS THREADS DO CUDA



Block 0	Thread 0	Thread 1	Thread 2	Thread 3
Block 1	Thread 0	Thread 1	Thread 2	Thread 3
Block 2	Thread 0	Thread 1	Thread 2	Thread 3
Block 3	Thread 0	Thread 1	Thread 2	Thread 3

SANDERS, J.; KANDROT, E. **CUDA by example : an introduction to general-purpose GPU programming**. [S.l.]: Addison-Wesley, 2010. p. 311

ORGANIZAÇÃO DAS THREADS DO CUDA

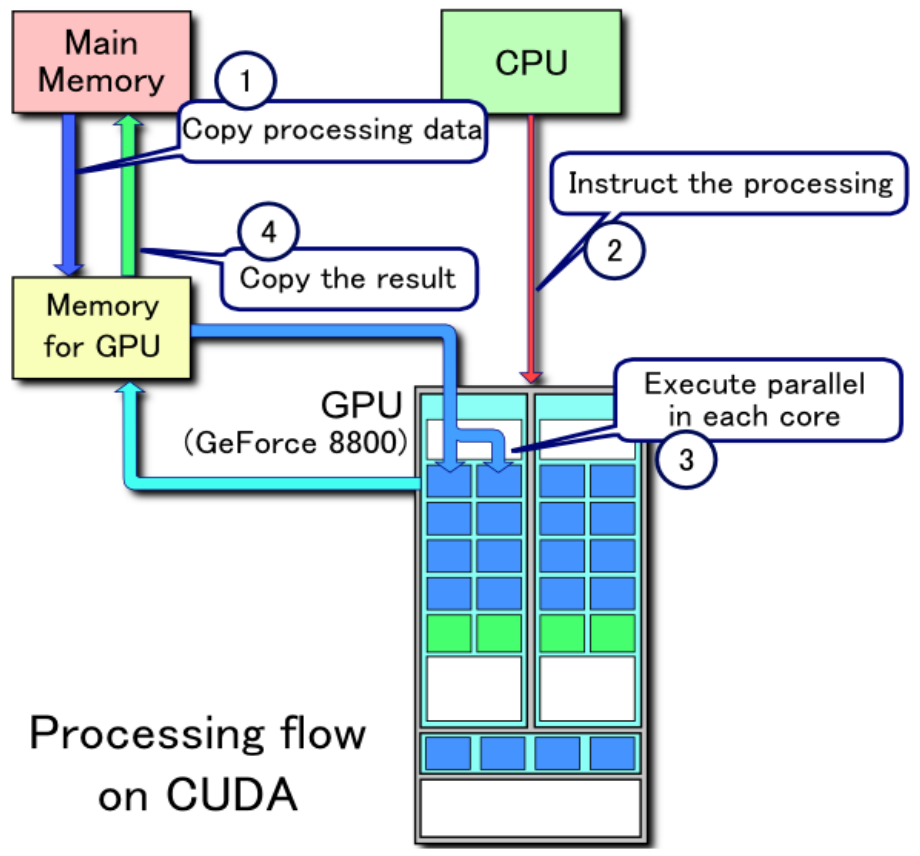


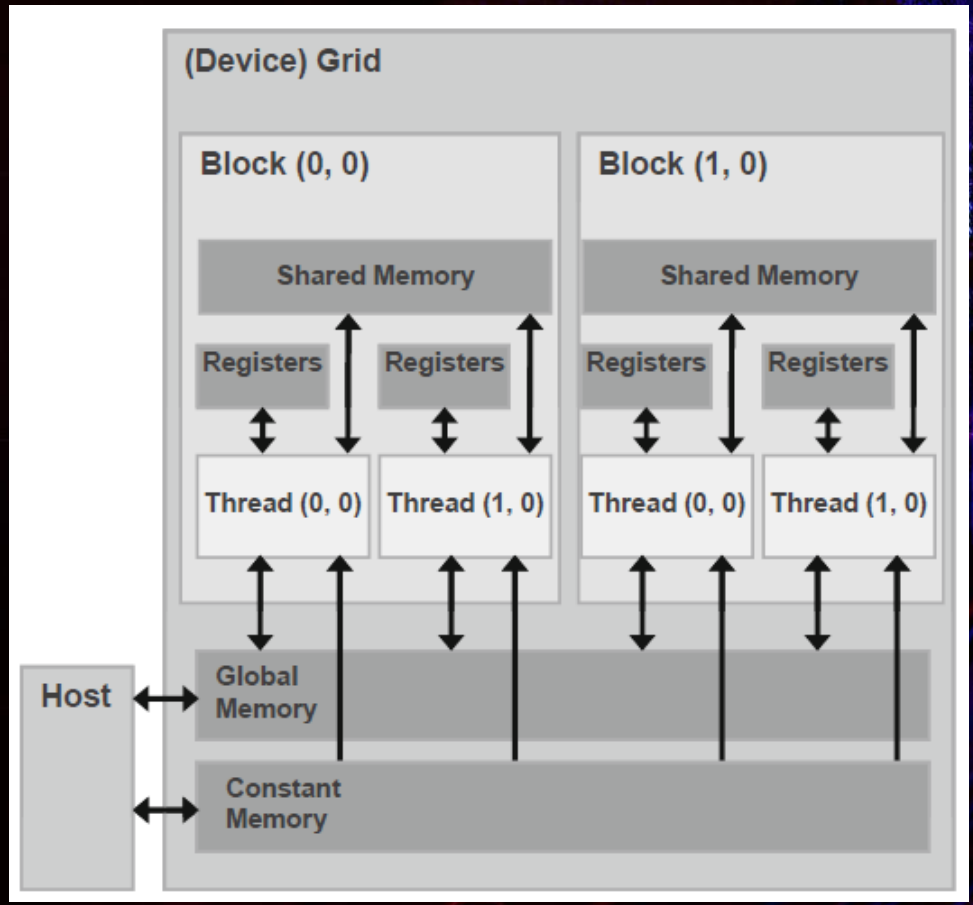
MODELO DE MEMÓRIA CUDA

- Memória global e memória constante podem ser escritas e lidas pelo host chamando funções da API

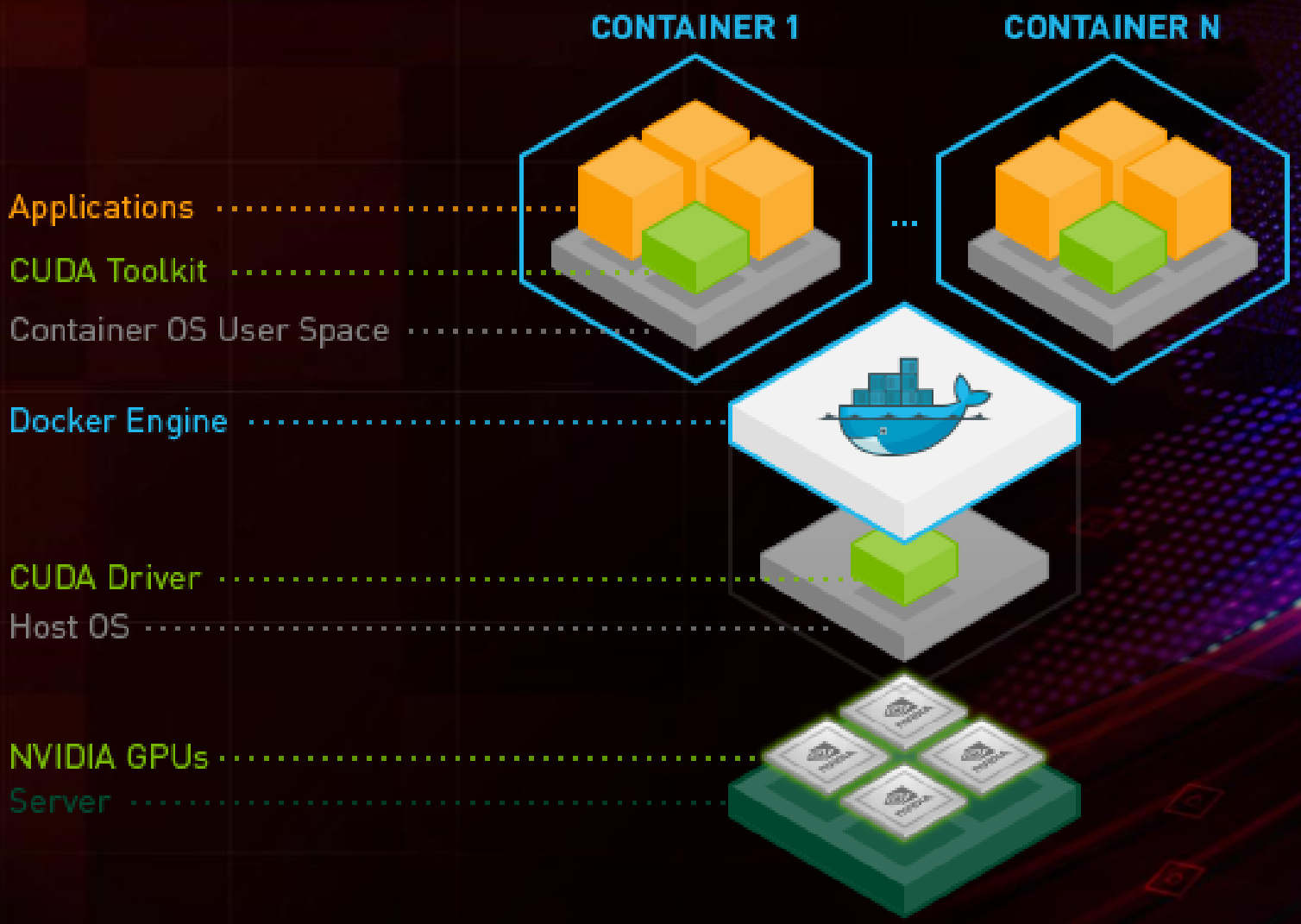
  - **Memória global:** maior e também a mais lenta
  - **Memória constante:** é pequena e mais rápida Suporta, basicamente, apenas só leitura

MODELO DE MEMÓRIA CUDA

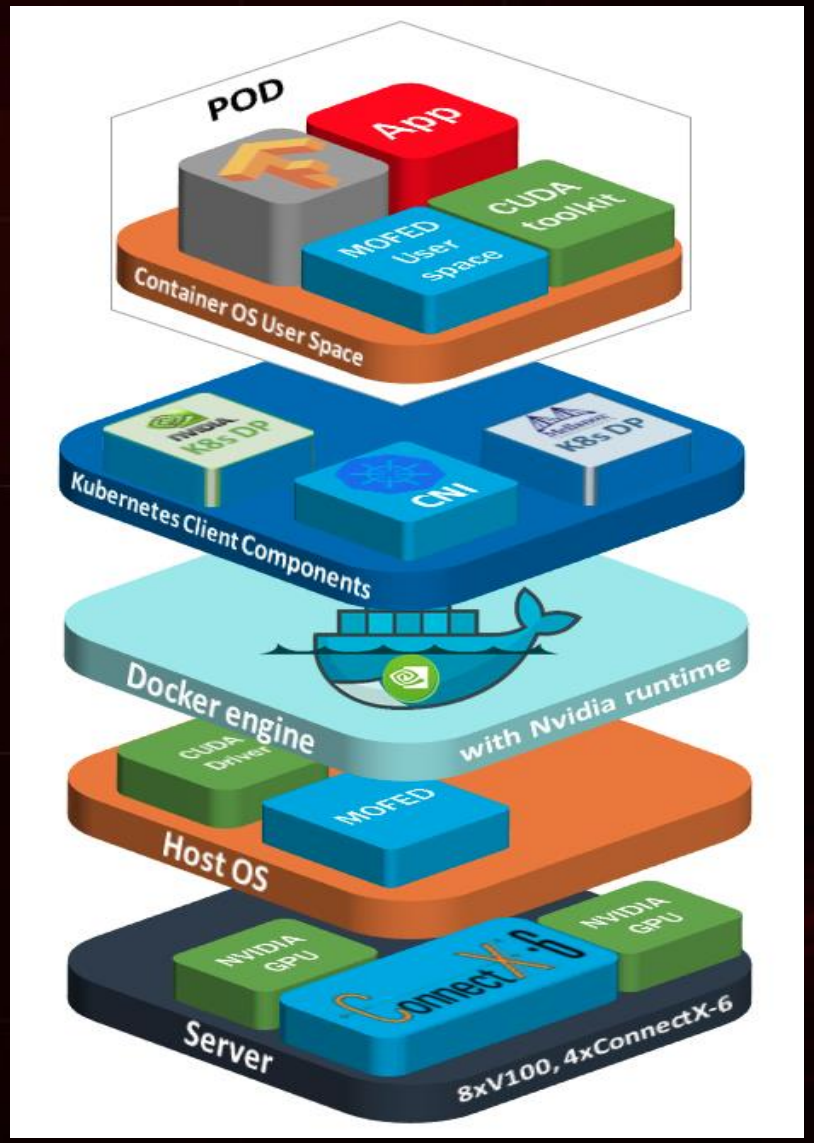




CUDA EM DOCKER E CLOUD



CUDA EM DOCKER E CLOUD



UTILIZAÇÃO DO CUDA

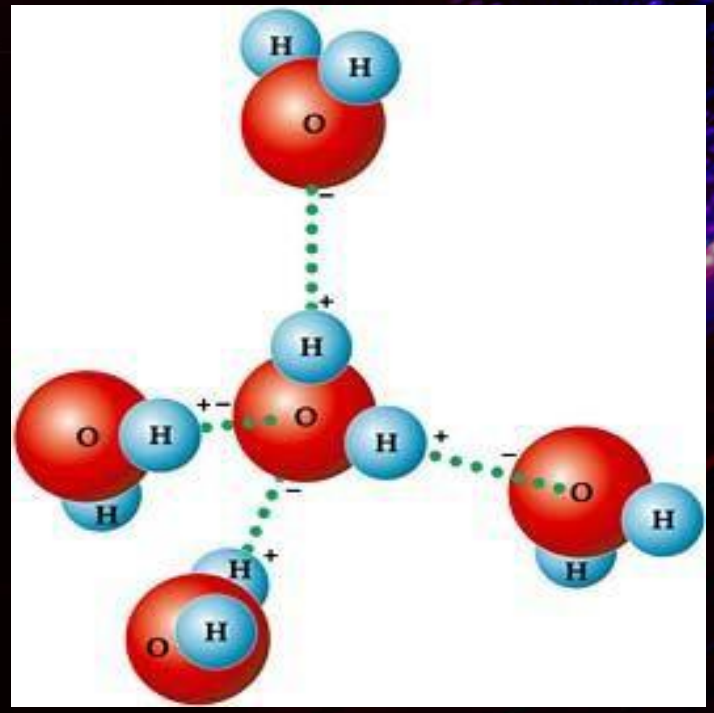
- Para utilizar o CUDA é preciso:
  - Ter uma placa gráfica da Nvidia
  - Ter o driver da Nvidia devidamente configurado no host
  - Ter um toolkit contendo compilador e ferramentas adicionais
  - Para saber os requisitos, o site da Nvidia detalha tais necessidades
    - *https://developer.nvidia.com/cuda-downloads*

APLICAÇÕES

- Ataques cardíacos são a maior causa de mortes no mundo todo
- Utilização de GPUs com o objetivo de simular o fluxo sanguíneo e identificar placas arteriais ocultas sem fazer uso de técnicas de imageamento invasivas ou cirurgias exploratórias

APLICAÇÕES

- Simulação de moléculas com melhor desempenho com a utilização de GPUs
- A aceleração do desempenho é resultado da arquitetura paralela das GPUs



BIBLIOGRAFIA

1. https://docs.nvidia.com/ai-enterprise/deployment-guide/dg-docker.html
2. https://thenewstack.io/nvidia-opens-gpus-for-ai-work-with-containers-kubernetes/
3. https://medium.com/unicoidtech/usando-nvidia-gpus-no-google-kubernetes-engine-ee4261713d86
4. https://edisciplinas.usp.br/pluginfile.php/4146828/mod_resource/content/1/MaterialCUDA.pdf
5. https://www.oficinadanet.com.br/post/14818-o-que-e-cuda
6. https://eradsp2010.files.wordpress.com/2010/10/curso2_cuda_camargo.pdf
7. https://deinfo.uepg.br/~alunoso/2020/SO/Programacao-CUDA/
8. https://lief.if.ufrgs.br/pub/Cursos/Cuda/aula06.pdf
9. http://ubiq.inf.ufpel.edu.br/ippd/lib/exe/fetch.php?media=cuda-lucas_pires_-_nvidia_cuda.pdf