畳み込み実装5

shun sato

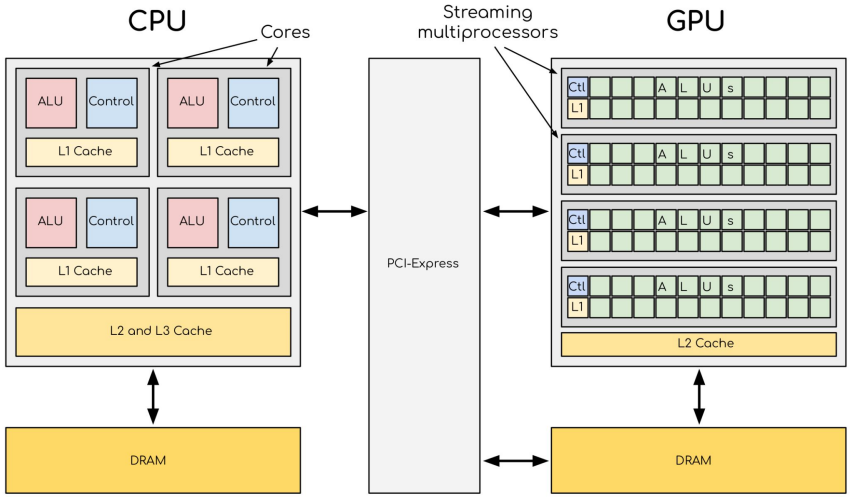
GPUを使ってみよう

- 今や機械学習はGPUを使わないと実行速度が出せない
- GPUは何が得意ですか？

NVIDIAのGPU

GPUとは

- GPUは並列計算が得意なハードウェア
- 大量の計算ユニットでまとめて計算
  - 深層学習モデル(行列計算)と相性がいい！
- RTX 4090
  - 16,384個の並列計算ユニット
  - 24GBのメモリ
  - 2.5GHzで動作



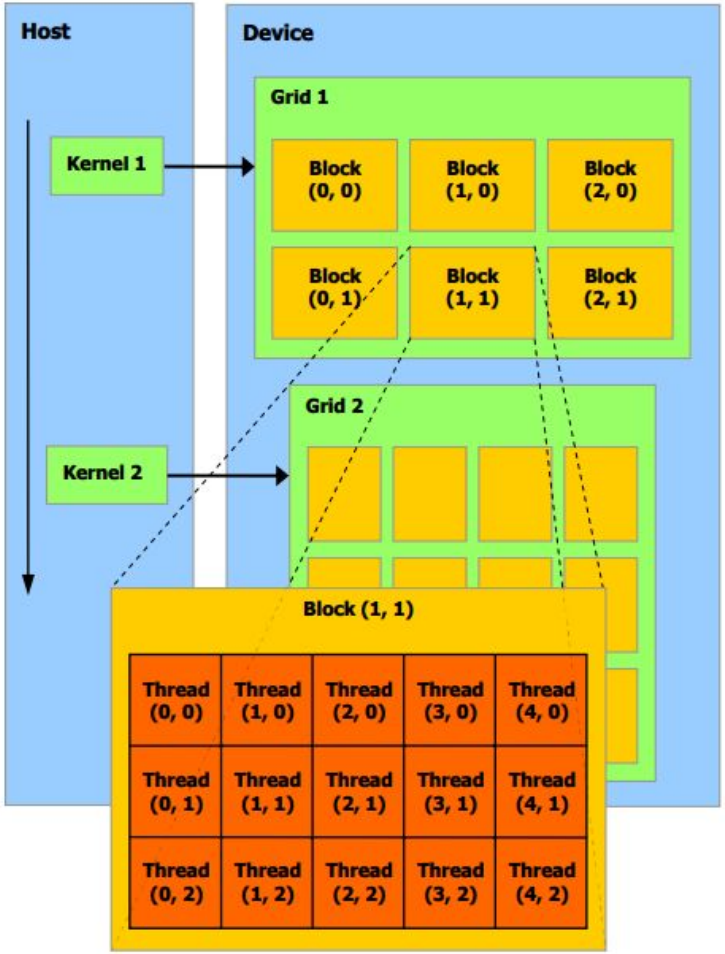
https://enccs.github.io/gpu-programming/2-gpu-ecosystem/

CUDAアーキテクチャ

- CUDAはNVIDIAのGPUが採用しているアーキテクチャ

- 計算ユニットを階層構造でまとめている

1. Grid
   - カーネルごとに1つ発行
2. Block
   - Threadをまとめる単位
3. Thread
   - 命令を実行する単位



https://www.nvidia.co.jp/docs/io/51174/nvidia_cuda_programming_guide_1.1_jpn.pdf

CUDAアーキテクチャ

色々な種類のメモリ

Registers
- 最速:Threadごとに割り当て

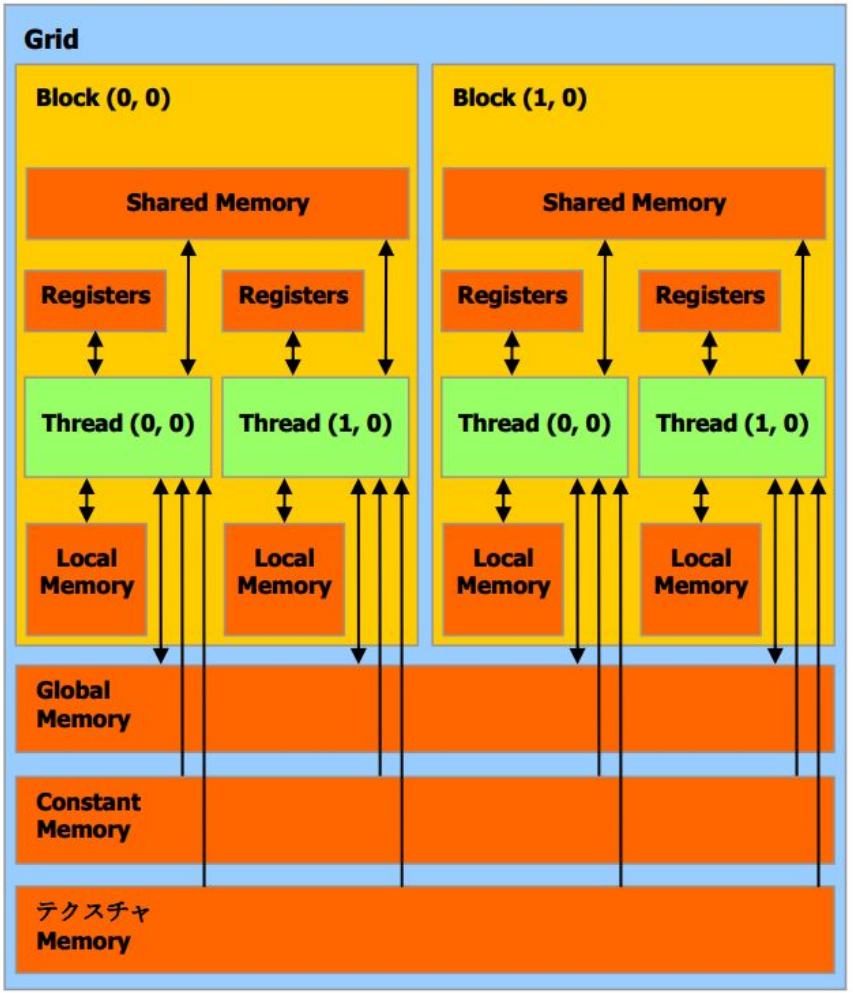
Local Memory
- 遅い:Registerが足りないときに割り当て

Shared Memory
- 速い:Blockごとに割り当て(共有)

Constant Memory
- 普通:読み取り専用で全体共通

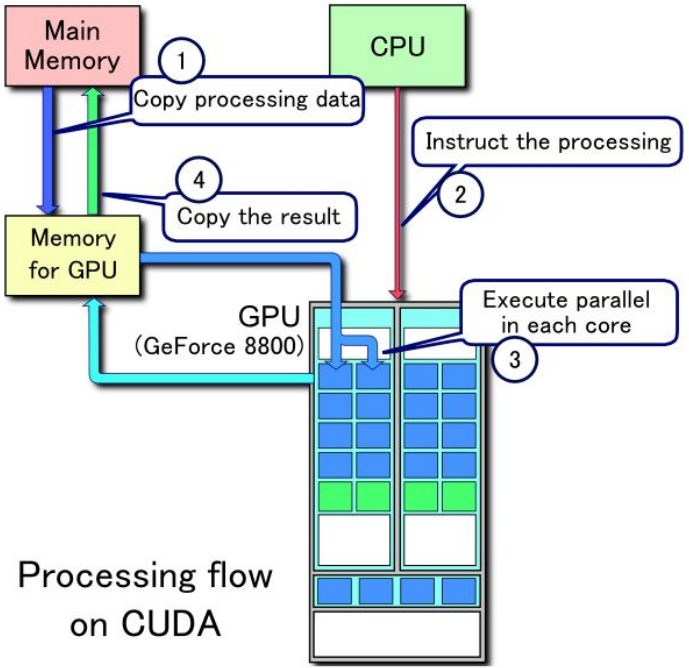
Texture Mmeory
- 普通:2Dの参照に特化

Global Memory
- 非常に遅い:全体でなんでも共通



GPUで命令を実行するまで

1. ホスト(CPU側)からデバイス(GPU側)に必要なデータを転送
2. ホストがカーネル(GPUの命令)を発行 ➡ 転送
3. デバイスがカーネルを実行
4. デバイスからホストに結果を転送(完了)



https://en.wikipedia.org/wiki/CUDA

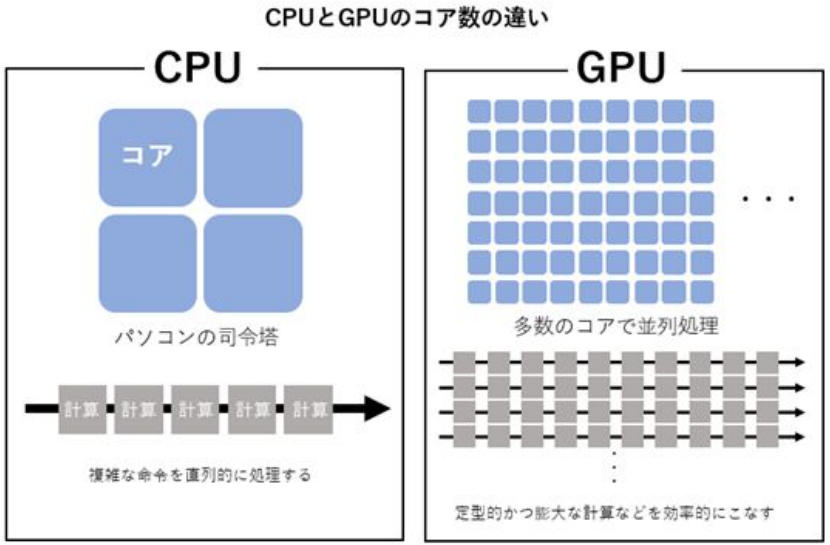
CPU vs GPU

CPU
- 計算は超高速
- 複雑な命令も高速に実行
- コアが4~24個

GPU
- 各コアの計算は遅い
- シンプルな命令の実行に特化
- コアが1000個以上

- GPUが得意な計算
  - 計算する回数が決まっている
  - 同時に計算ができる

➡ 画像処理や行列演算が得意



https://allai.jp/what-is-gpu/

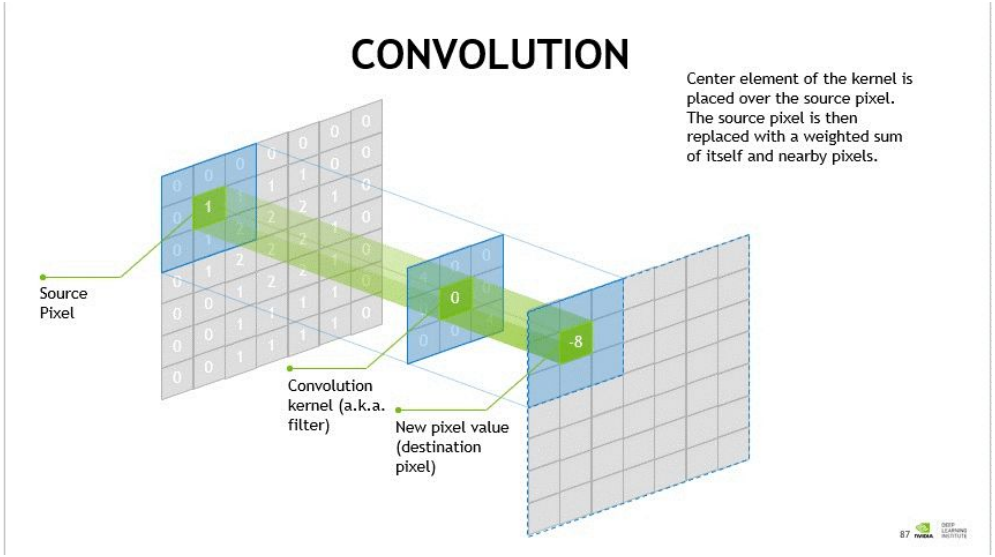
畳み込みをGPUで処理する

畳み込み演算

- 各ピクセルは同時に計算してもいい
  - GPUと相性抜群

- CPUはループで頑張ってた
  - 1024×1024=1,048,576

- GPUのボトルネック
  - 画像の転送の方が時間かかる
  - 処理は一瞬



https://www.nvidia.com/en-us/glossary/convolutional-neural-network/