GPU の 基礎

NII
五島 正裕

概 要

- 現在 GPU は, 特にその高い演算性能から, AI 分野における最も有力な実行プラットフォームとなっている.
- GPU の高い演算性能は, AI 分野で特徴的に現れるテンソル計算に強く適応した結果, 得られるものである. GPU が利用するテンソル計算の性質には, 1. データ並列性・規則性の高さ, 2. 計算精度に対する要求の低さ, 3. 演算強度の高さなどがある.
- 本講演では, GPU の基礎として, GPU がそれらの性質をどのように利用しているかを, CPU との比較という観点から概説する.

ピーク性能 と チップ面積

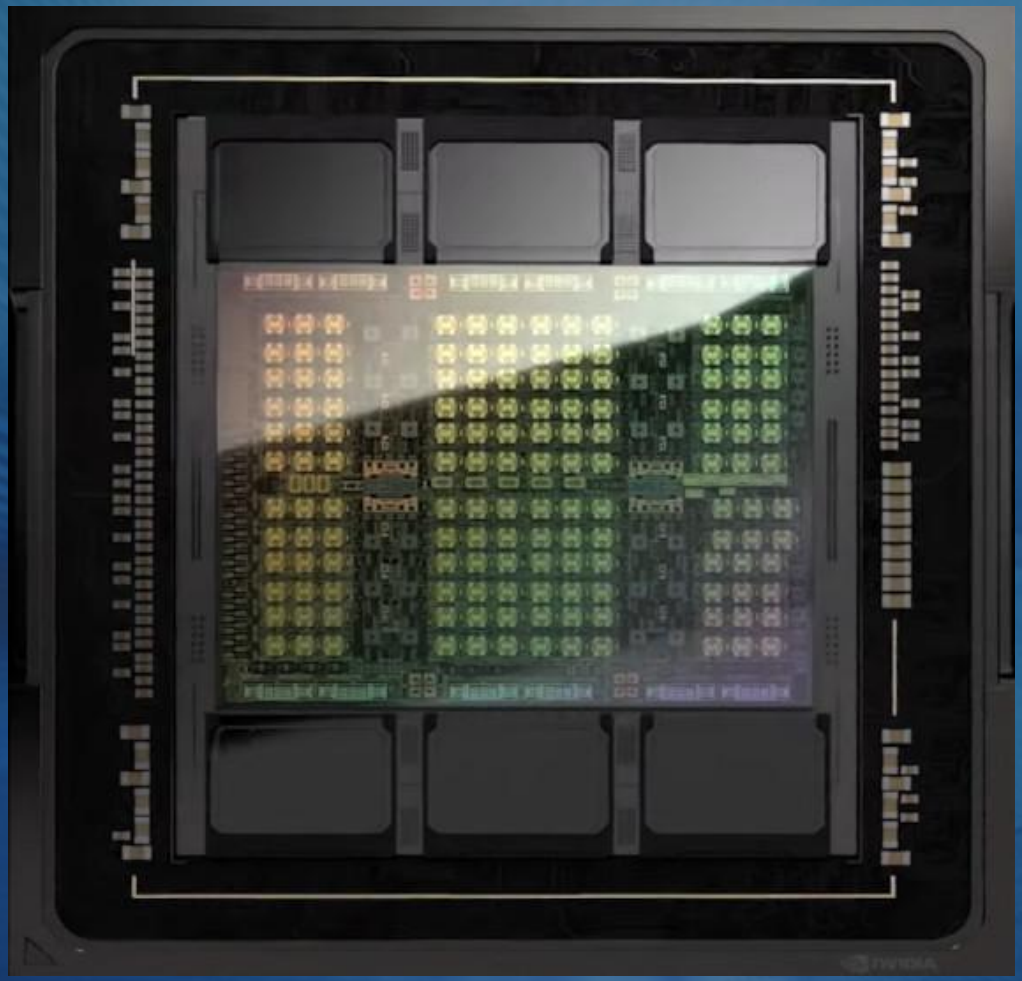
GPU vs. CPU ピーク性能比

- 現在 GPU は, 特にその高い（ピーク）演算性能から,
  AI 分野における最も有力な実行プラットフォームとなっている
- GPU: NVIDIA H100 Tensor Core GPU
  - 4 PFLOPS = 4,000 TFLOPS (FP8)
- CPU: Intel Meteor Lake
  - ＜ 1 TFLOPS (AVX-512, FP64)
- ＞ 4,000倍 !

ピーク 演算性能

- ピーク性能（理論最大性能）
  - ＝ 演算器の個数 × 動作周波数
  - ＝ チップ面積 × チップ面積に占める 演算器の割合 × 動作周波数
- チップ面積 同じ，動作周波数 同じ とすると…
  - ≒ チップ面積に占める 演算器の割合

NVIDIA H100 Tensor Core GPU



- チップ面積に占める 演算器の割合 が 非常に 高い！

https://www.comptoir-hardware.com/actus/cartes-graphiques/45736-nvidia-gtc-hopper-h100-te-voila-.html

Intel Meteor Lake Compute Die

- 演算器はすごく小さい！

CPU は 何をやっているのか？

CPU は, 何をやっているのか？

- 演算器 の チップ面積占める 演算器の割合 は 極小
  - 逆に, 演算器以外の大部分は, いったい何をやっているのか？
- CPU にとって 大事なこと:
  - ×: 演算それ自体
  - ○: スケジューリング:
    あるサイクルに, 演算器の前に, 必要な 命令 と データ(複数) が「会う」
- (チップ面積 や 消費電力 などの 資源 の観点からは)
  「CPU の お仕事は スケジューリング」!
  - スケジューリングの たいへんさ は みなさん, ご存じのとおり
    - CPU だと,「遅刻」が許されないので,「流会 → リスケ」は頻繁
  - CPU は, スケジューリング技術 の 30年分の積み重ね

GPU は, 何をやっているのか？

- 現在の AI の 計算 にとって 大事なこと
  - ○: テンソル計算（行列積）
  - ×: スケジューリング
    - 「いつ・どこで演算するか十分前に分かっている」
- CPU と GPU の桁違いの性能差は,
  CPU がスケジューリングに充てている回路面積を,
  GPU は演算器に充てている ことによる

GPU の基本技

テンソル計算 の性質 と GPU

- 現在の AI の計算（の主要な部分）はテンソル計算（行列積）
- GPU が利用している テンソル計算の性質：
  1. データ並列性・規則性の高さ
  2. 計算精度に対する要求の低さ
  3. 演算強度の高さ

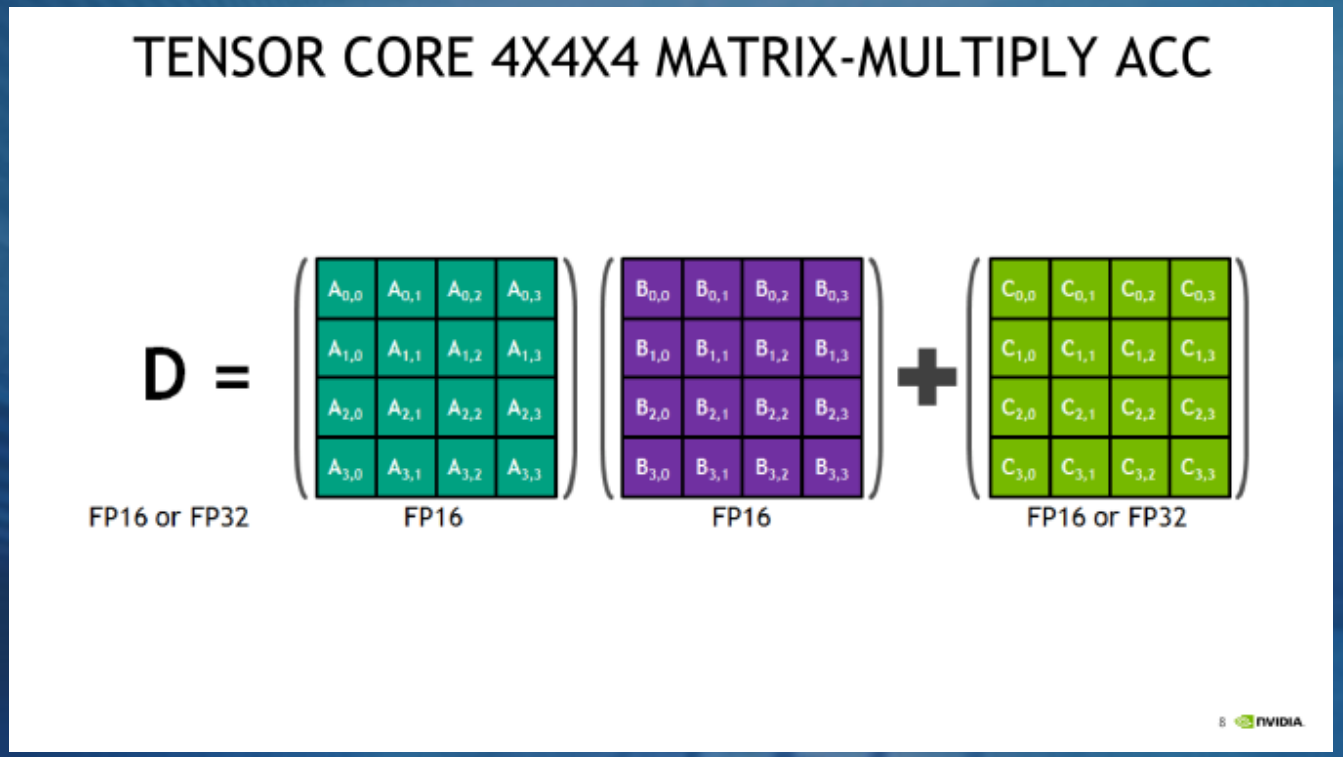
1. データ並列性・規則性の高さ → SIMD

- SIMD (Single Instruction/Multiple Data stream)
  - 1つのレジスタ内に, v 個のデータをパックし,
  - 1命令で, 同じの演算を v 個 同時に行う
- 演算器以外の（制御に関わる）チップ面積を $1/v$ に削減できる

2. 計算精度に対する要求の低さ → FP16, FP8

- 低精度 浮動小数点 フォーマット の サポート
  - FP64 : 倍精度. 高性能計算 (High-Performance Comp.) 分野 では 標準
  - FP32 : 単精度
  - FP16 : 半精度
  - FP8 : (?)
- SIMD と合わせて, 2～4倍の性能向上
  - ただし, 演算数で数えた名目上のこと

3. 演算強度の高さ → Tensor Unit



- 1 行列積和命令
  - 3x4x4 = 48 要素のアクセスで, 2x4x4x4 = 128 FLOPs
  - アクセスに関わる電力を削減
    - ピーク演算性能が上がる訳ではない

https://www.anandtech.com/show/12673/titan-v-deep-learning-deep-dive/3

NVIDIA H100 Tensor Core GPU

- GPU が利用している テンソル計算の性質:
  1. データ並列性・規則性の高さ → SIMD
  2. 計算精度に対する要求の低さ → FP16, FP8
  3. 演算強度の高さ → Tensor Unit
- それぞれ, 何割～倍程度の改善で, 桁違いの性能差は生まない

まとめ

- GPU の高いピーク性能は（一次近似的には）
  スケジューリングが要らない ことによる
- 現在の AI 用アクセラレータ は「Tensor Unit の塊」
- 他社製品も大同小異
  - AMD Instinct や Google TPU で NVIDIA を置き換えられない
    技術的・商業的理由はほとんどない
  - NVIDIA の市場支配（H100 は 471万円！）は,
    「NVIDIA でなければ」という市場の思い込みが大きい

今後の方向性

- AI アクセラレータ が Tensor Unit の塊から変わるかは, AI アプリ次第
  - 疎行列？ GNN？
    - 現状は,「4要素のうち2要素が0」のような,
      密行列の枠組み内での小変更による対応は見られる
    - 本格的にやると, スケジューリング が必要になる