第2章 GPUでもキャッシュが重要

CUDAのAPIを使って行列積演算を高速化する

松岡 洋

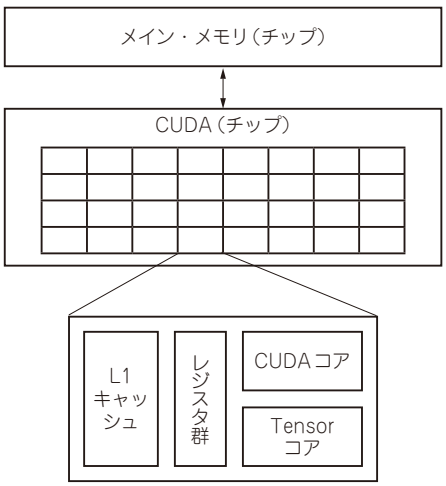

図1　エヌビディアのGPUの構造

● GPUプログラミングの最適化余地

深層学習においてGPUを使って非常に高速な演算が行われていますが，GPUにもCPUと同等にメモリ・アクセスによるストールの問題は残っています．キャッシュをうまく使いこなすことで，演算を高速化できるのはCPUと同じです．

エヌビディアのGPUはCUDAと呼ばれるフレームワークを使ってプログラミングします．キャッシュを意識してプログラムできるようなAPIがCUDAにも用意されています．

リスト1　行列積を行うCUDAのプログラム

```
cublasStatus_t cublasHgemm(
  cublasHandle_t handle,
                    // cuBlasライブラリコンテキストハンドル
  cublasOperation_t transa,    // 行列Aの転置指示
  cublasOperation_t transb,    // 行列Bの転置指示
  int m, int n, int k,         // 行列サイズ
  const __half *alpha,
  const __half *A, int lda,    // 行列Aデータ
  const __half *B, int ldb,    // 行列Bデータ
  const __half *beta,
__half *C, int ldc             // 結果C
);
```

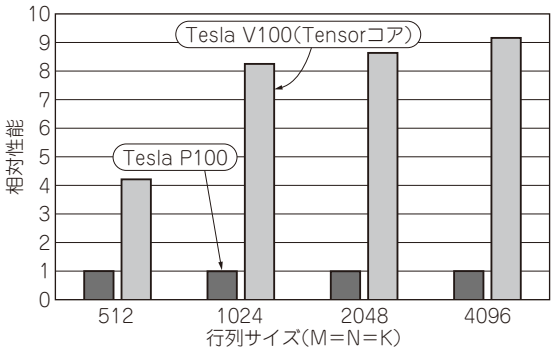

図2　CUDAコアの世代別に行列積の性能を比較した

基礎知識：GPUの構成とCUDA関数

● 推論処理に適した半精度の演算ユニットもある

図1にエヌビディアのGPUの構造を示します．深層学習で多用される行列積演算をCUDAコアで行いますが，これを高速に行うために，Tensorコアと呼ばれる演算ユニットも使われます（Voltaと呼ばれる世代から追加された）．

Tensorコアは部分行列間の積和演算ユニットです．16ビット浮動小数点数（半精度とも呼ばれる）に対応しています．CUDAコアによる演算に比べて10倍以上高速に演算を行うことができます．深層学習モデルの学習処理はCUDAコアで行い，多少精度を落としてもよい推論処理は，高速なTensorコアで行えるような構成になっています（図2）．CUDAでの半精度の行列積はリスト1の関数で行います．

● CUDA関数を呼べばTensorコアで行列積を求められる

Tensorコアでの行列積はリスト2に示すcublasGemmEx関数で行います．CUDAでの行列積は半精度の他にも単精度，および倍精度など精度ごとに別々の関数が用意されています．関数に指定する引数によってCUDAコアを使うのかTensorコアを使