HPC-phys 勉強会

FUJITSU
shaping tomorrow with you

富岳におけるGEMMチューニング

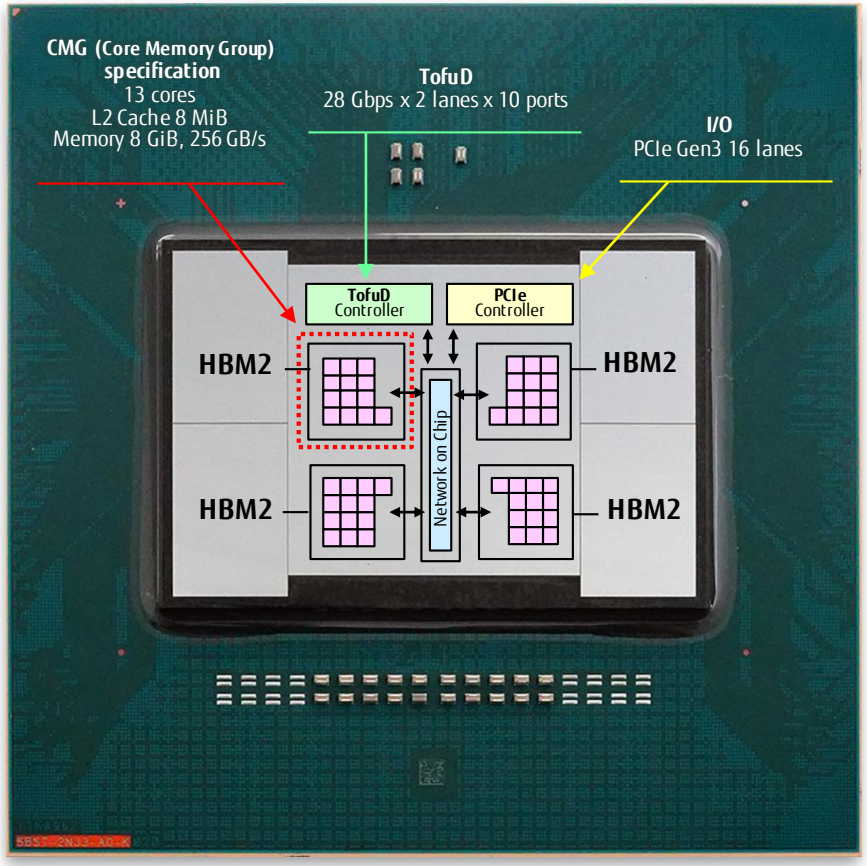
富士通株式会社 プラットフォームソフトウェア事業本部
第三基盤ソフトウェア事業部 第二開発部
竹重和明

2020.12.3

富岳/FX1000/FX700

A64FXプロセッサ

仕様	
命令セット	Armv8.2 SVE (512bit)
演算コア数	48 (12 cores / CMG)
CMG数 (Core Memory Group)	4
ベクトルレジスタ数	32/core
L1Dキャッシュ	64KiB/core
L2キャッシュ	8MiB/CMG (CMG内の12コアで共有)
メモリ	32GiB
その他の特徴	セクターキャッシュ (キャッシュをセクタと呼ばれる部分に分けて更新管理をする) ハードウェアバリア FP16をサポート



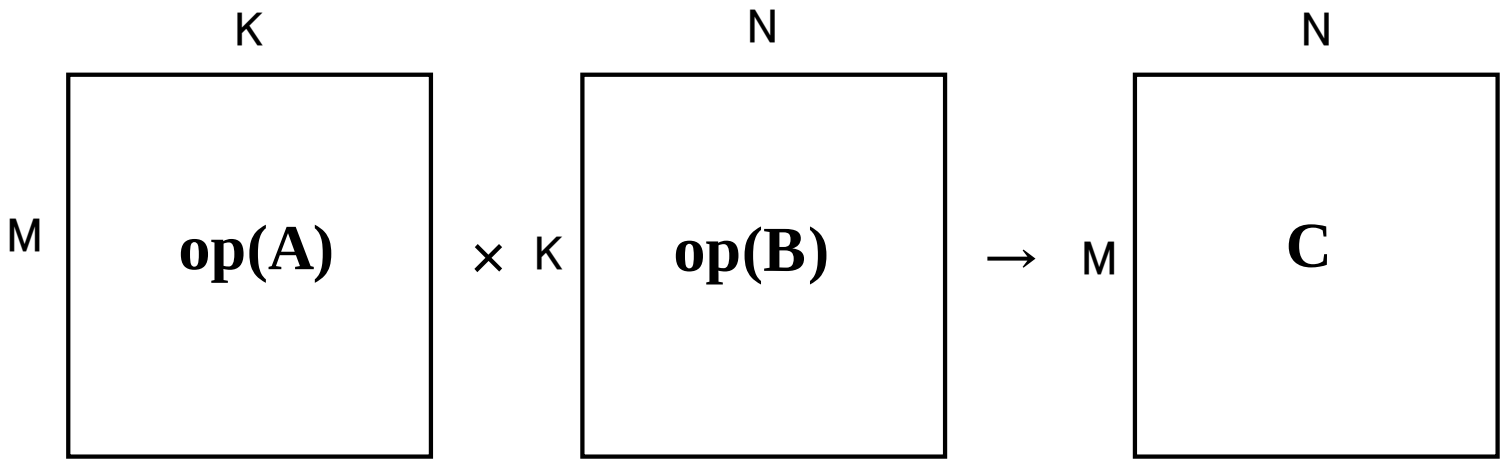
BLASの行列-行列積ルーチン xGEMM

FUJITSU

- 行列-行列積を高速に計算する
  - Linpackベンチマークでは実行時間の大半を占める
  - 多くのBMTやアプリケーションで使われる
  - 近年、AIでも重要となっている
  - 三重ループで書けるが、それでは性能が出ない！

プロセッサの性能を引き出す最大限の最適化が必須

$$C \leftarrow \alpha op(A) op(B) + \beta C$$



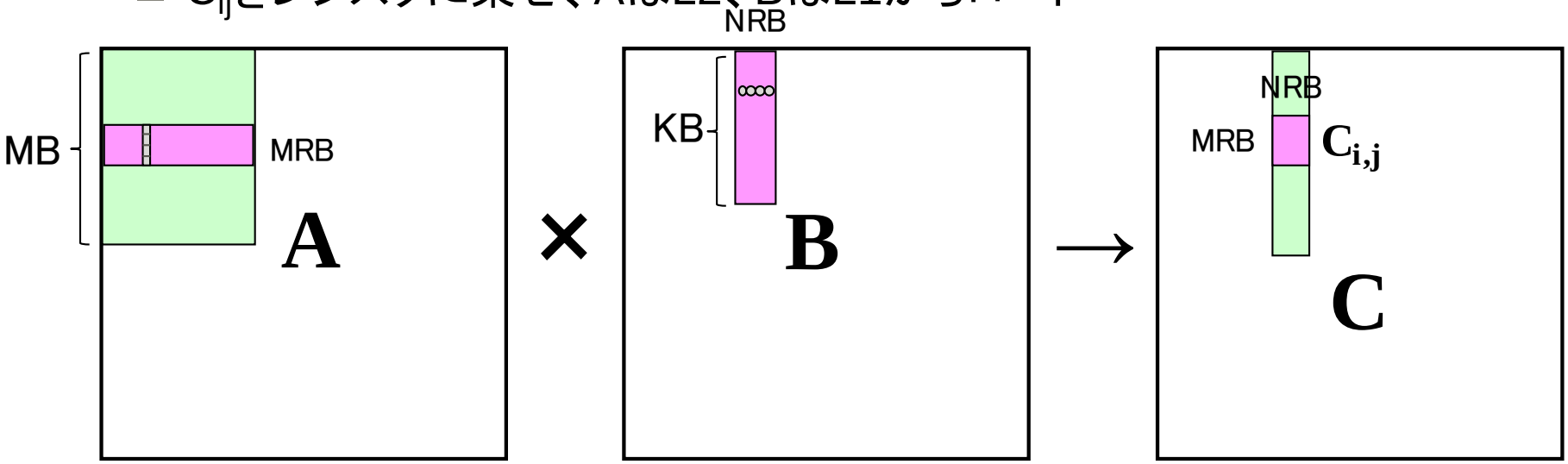
富岳向けGEMM

FUJITSU

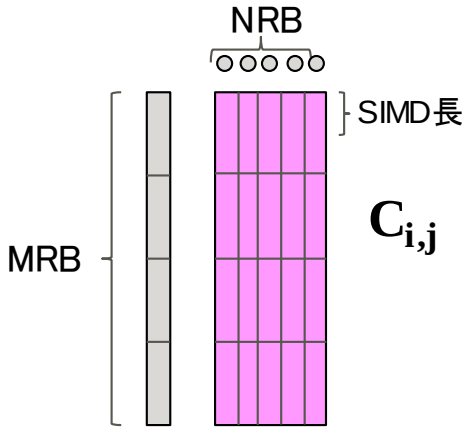
- 富岳プロセッサの特性と命令を活かしたチューニング
  - SVE命令の活用
    - fma系の演算命令
    - 多くの命令でpredicateによるマスク処理が可能
  - ベクトル長512bit
    - 倍精度 8要素、単精度16要素、半精度(FP16) 32要素
  - レジスタブロッキング
  - キャッシュブロッキング
  - セクタキャッシュを活用
- スレッド並列化
  - 共有キャッシュの活用
  - 分割方法の工夫
  - 複数CMGへの対応

レジスタブロッキング

- DGEMMの最内ループではCの小片C_{ij}を計算
  - C_{ij}をレジスタに乗せ、AはL2、BはL1からロード

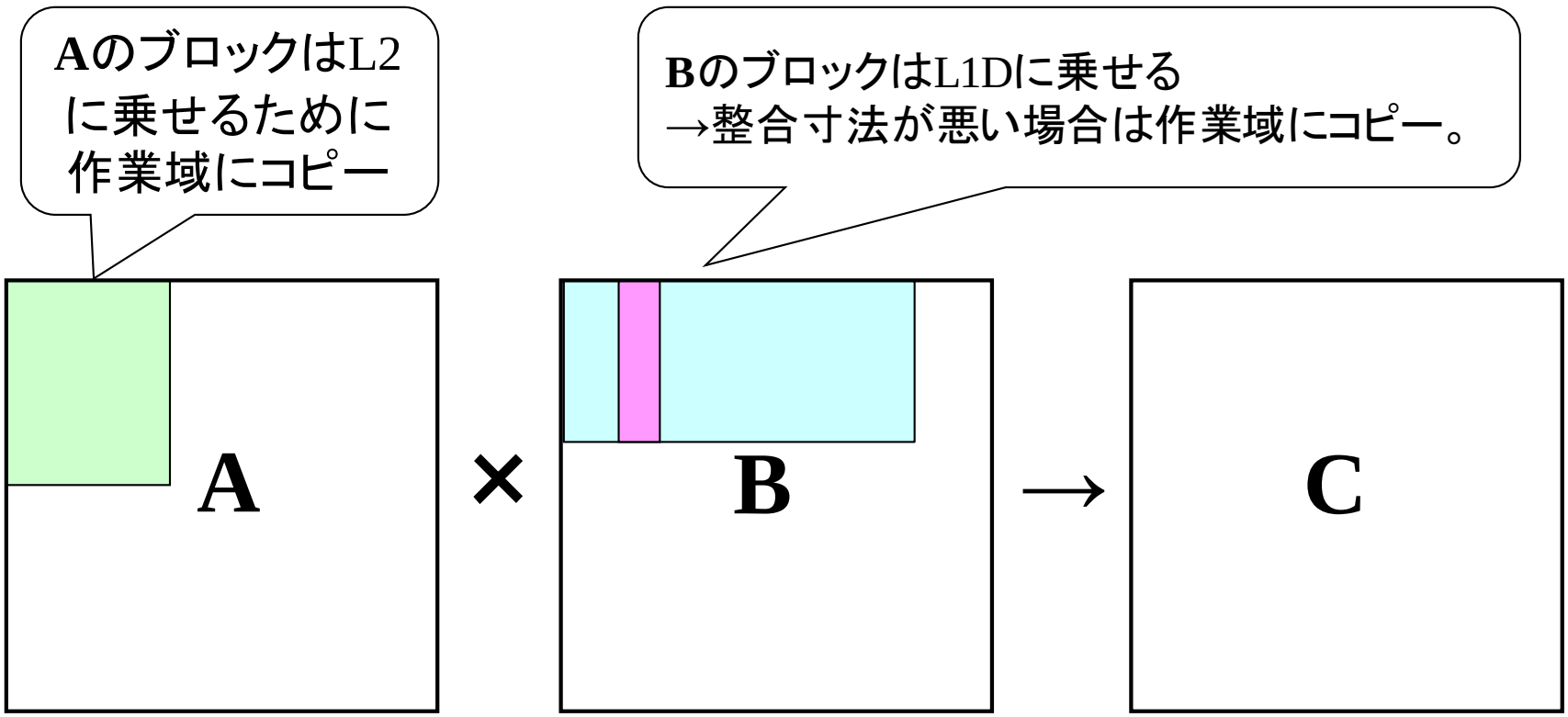


- 最内ループの中は100%の効率で実行
- L1、L2キャッシュのスループットおよび演算・ロードレイテンシの隠ぺいを考慮してMRB、NRBを決定
- 富岳・FX1000ではMRB=32(SIMD長×4)、NRB=5、MB=640、KB=640
- NRBはAの要素のL2からのロードと、演算のバランスで決定
- MのSIMD長の余り部分はpredicateを使ってマスク演算



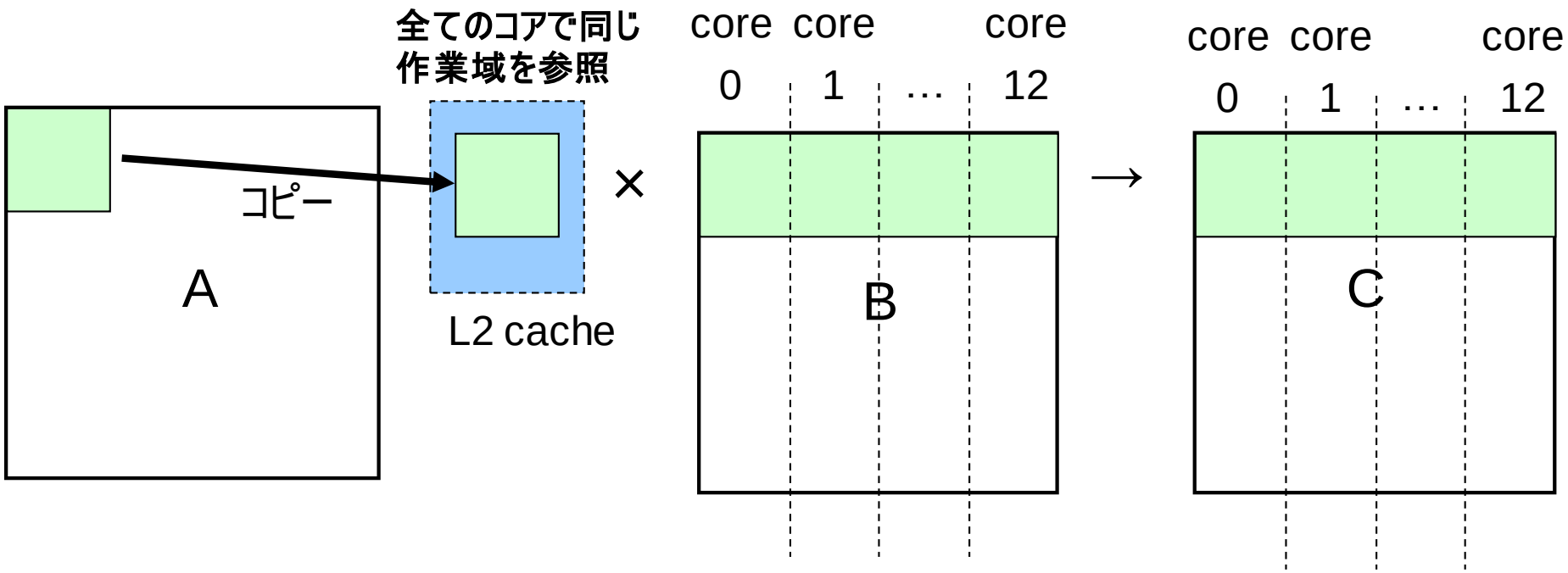
キャッシュブロッキング

- 行列積(C=AB)は行列をブロックに分け，L1D，L2キャッシュに乗せて計算
- キャッシュに乗せておくため、行列の一部を作業域にコピーして使用
  整合寸法の値によっては、コピーを回避
- A,B,Cはキャッシュに確実に乗せるために、L1のセクタキャッシュを使用



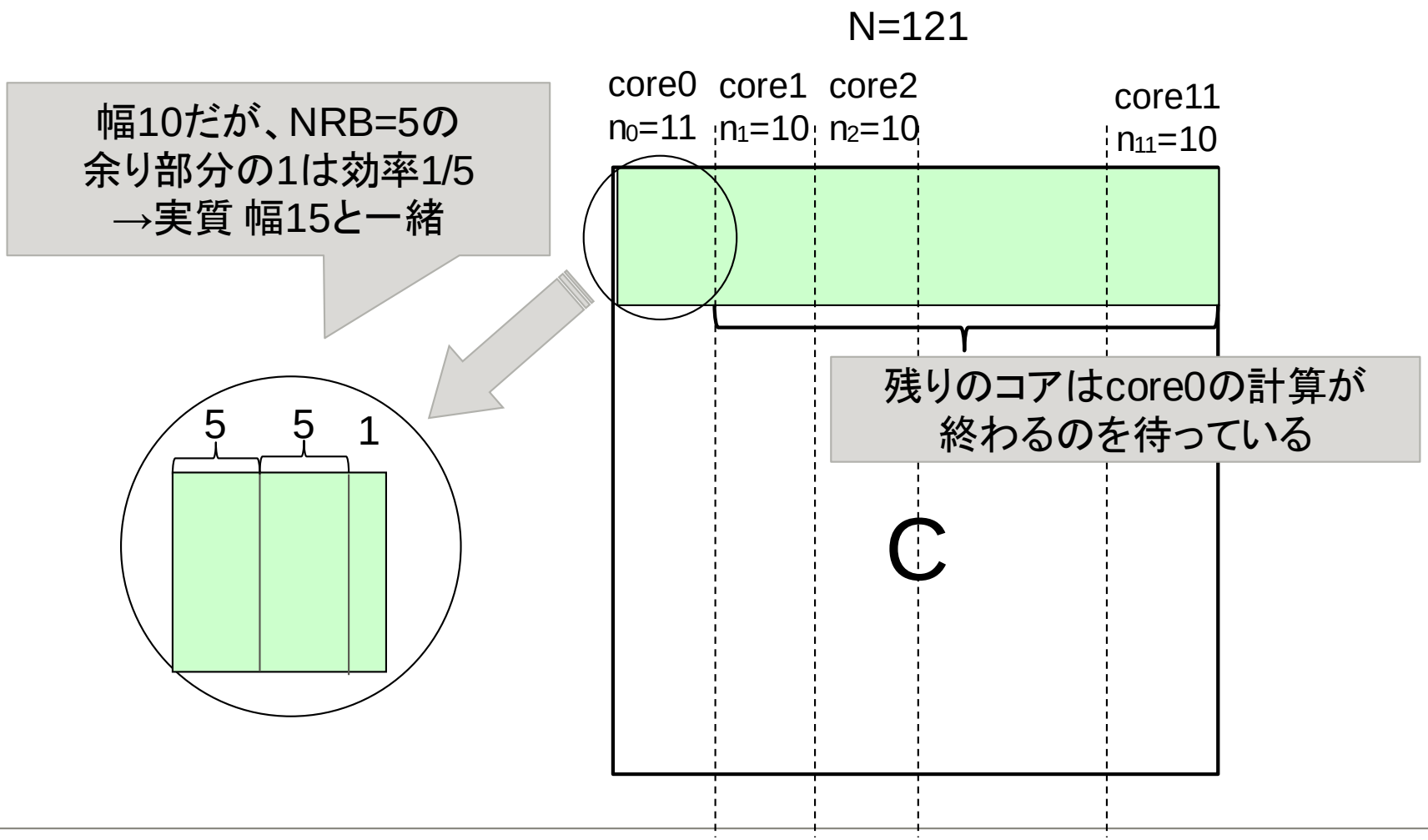
スレッド並列化

- 共有キャッシュを使ったスレッド並列計算
  - 行列Aのブロックを L2に乗せる
  - ブロックが大きいほうが高効率
  - 共有キャッシュを活用して，Aのブロックを共有し効率Up
  - ただし、Nが12で割り切れないとき、ロードバランスが悪くなる



スレッドで等分割したときの問題点

- N=120→121で計算時間が増加
- N=120を12スレッドで割ったとき、各コアの担当する列数は10 or 11
  この1列の違いが実行時間に大きく影響する。



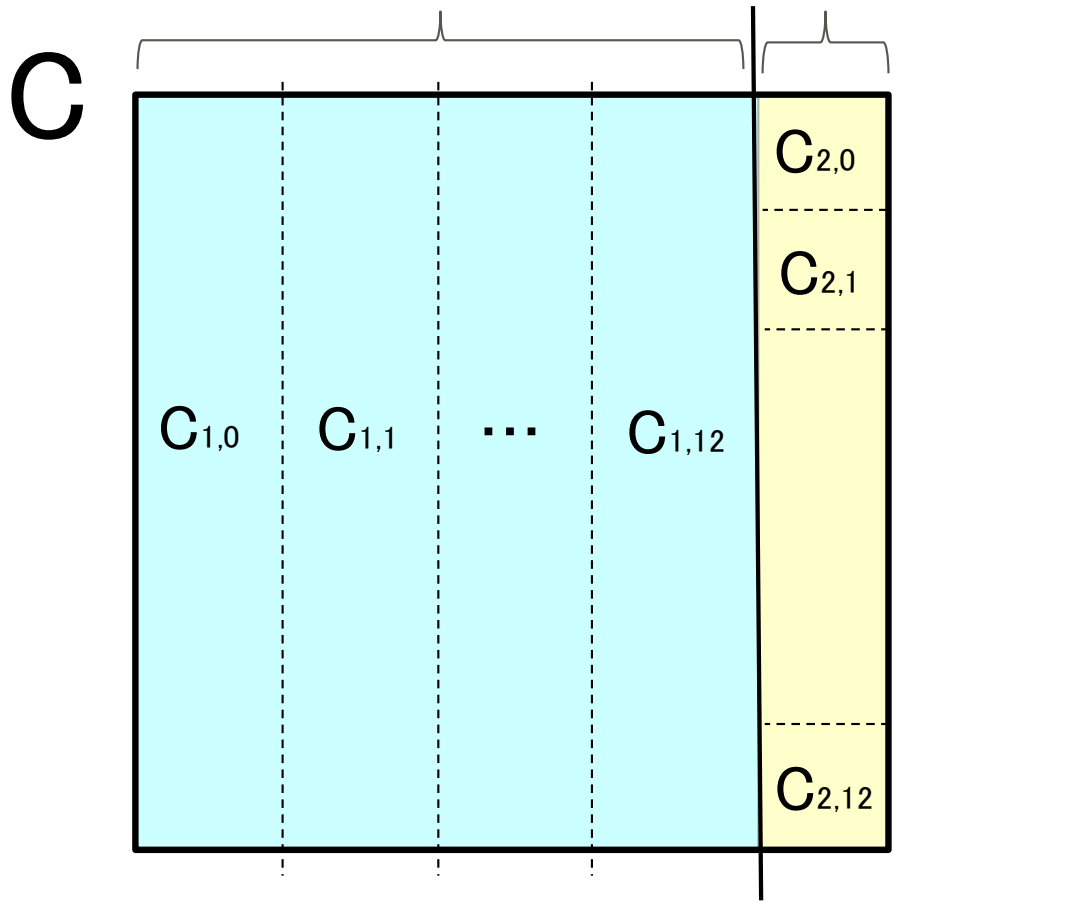
等分割の問題点の対策

FUJITSU

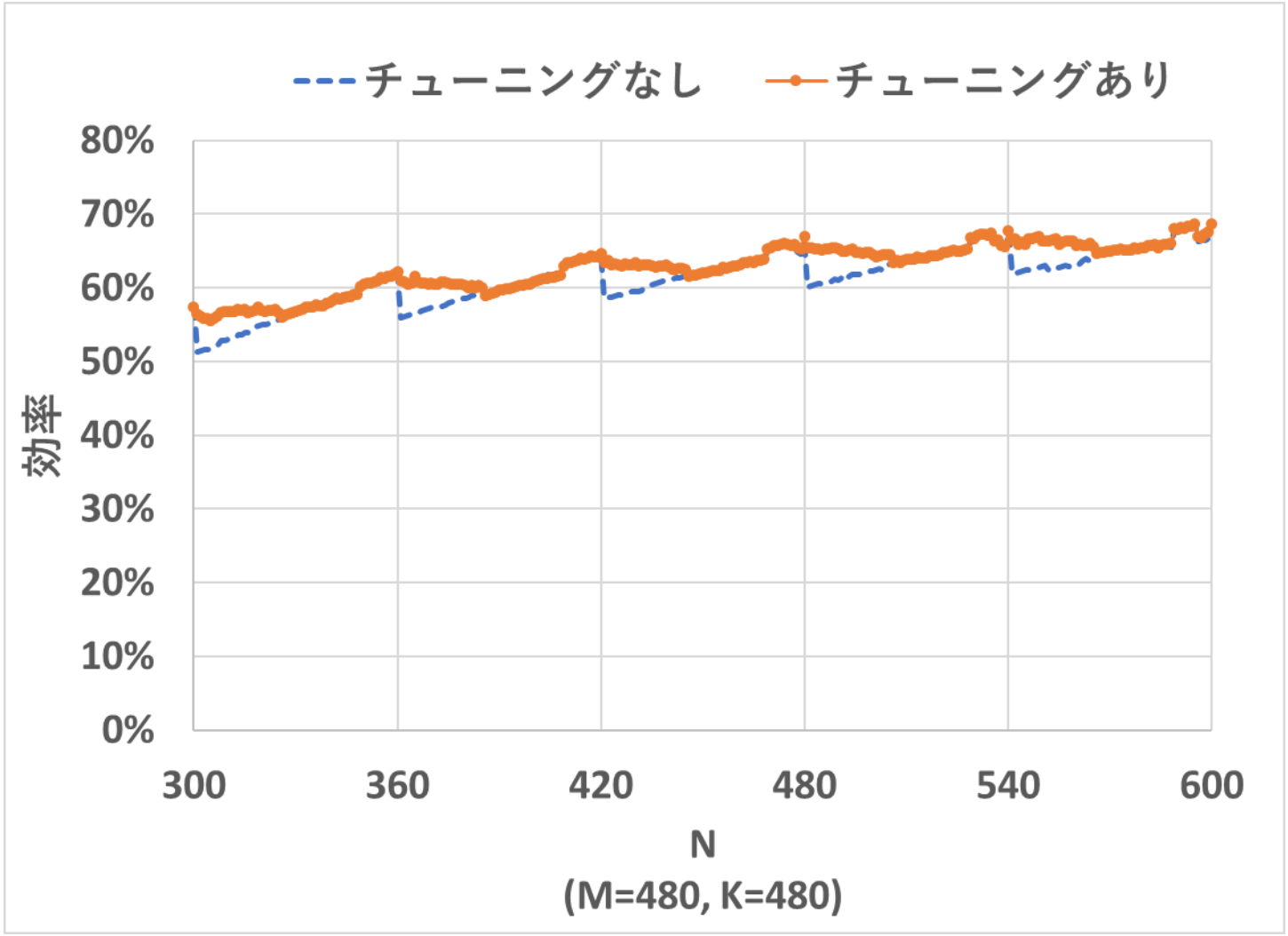
- 行列を2つに分けてそれぞれ別の次元で並列化

分割-1
2次元目をスレッドで等分割。
1スレッドあたりの幅はNRBの倍数

分割-2
1次元目をスレッドで
等分割



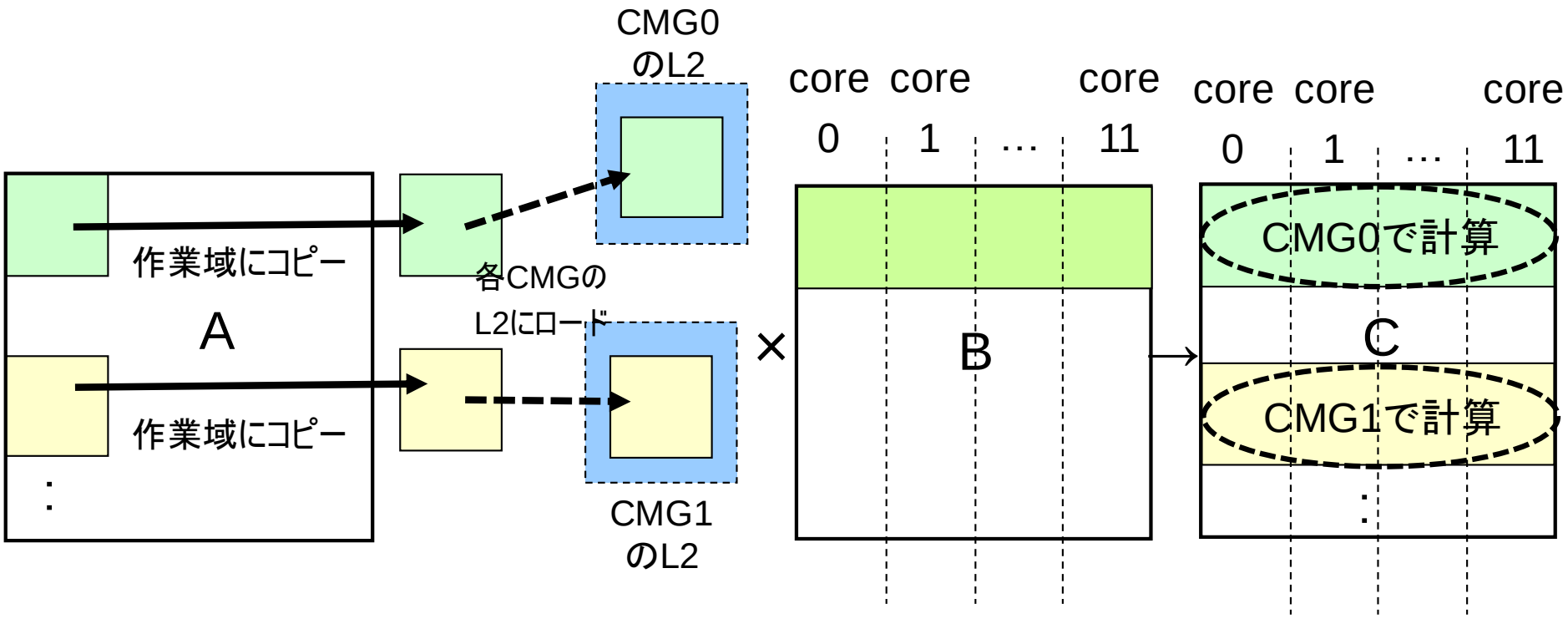
分割を改善した効果



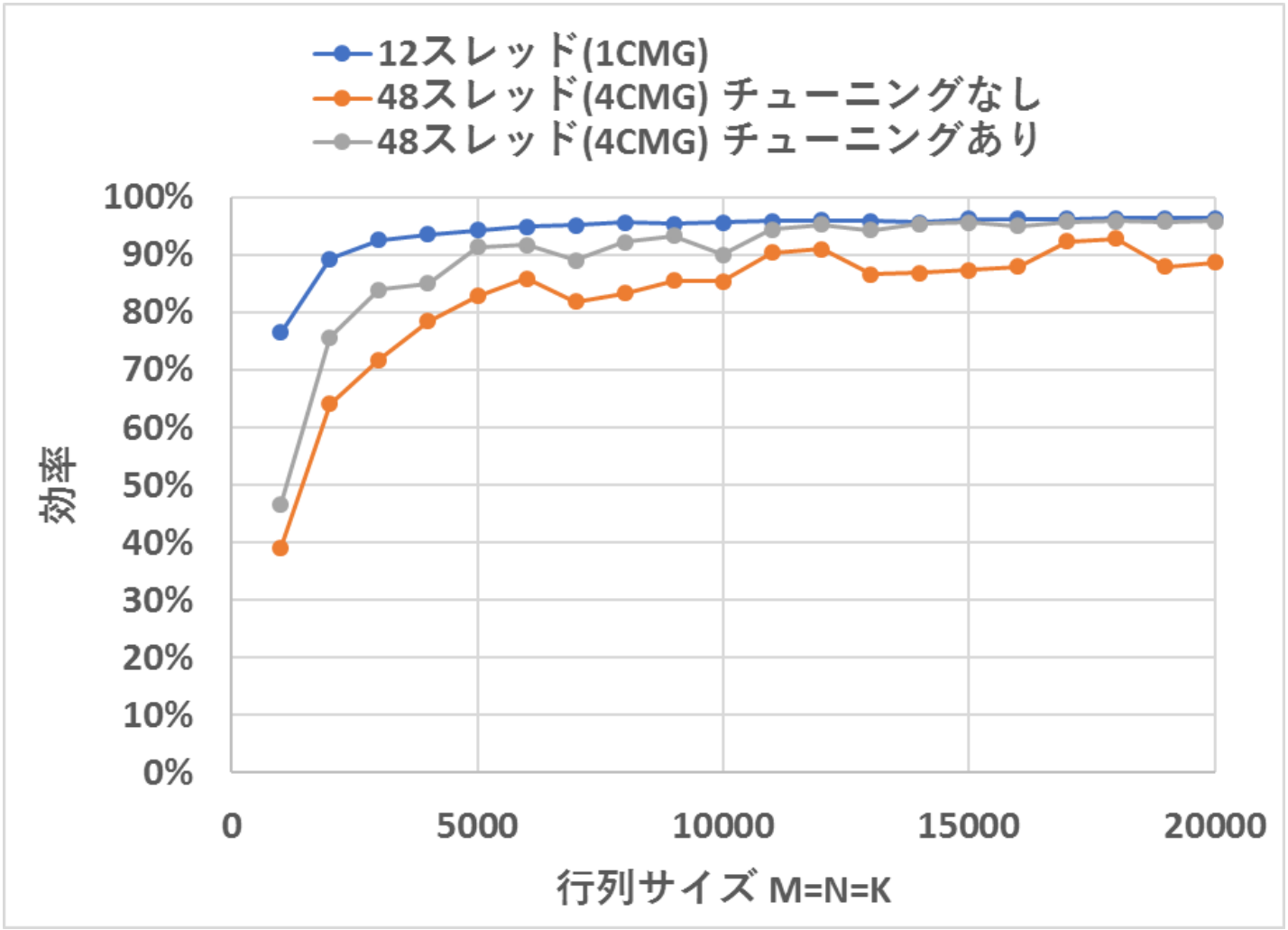
複数CMG向けチューニング

FUJITSU

- 多数コアのための分割
  - 単純に等分割すると、作業域にコピーしたAの要素をCMGを跨いでL2にロードするためのロスが発生する
  - 行列の分割次元を、CMG単位とコア単位で別にする
  - 全スレッドをCMG単位の処理とコア単位の処理の2段階で管理

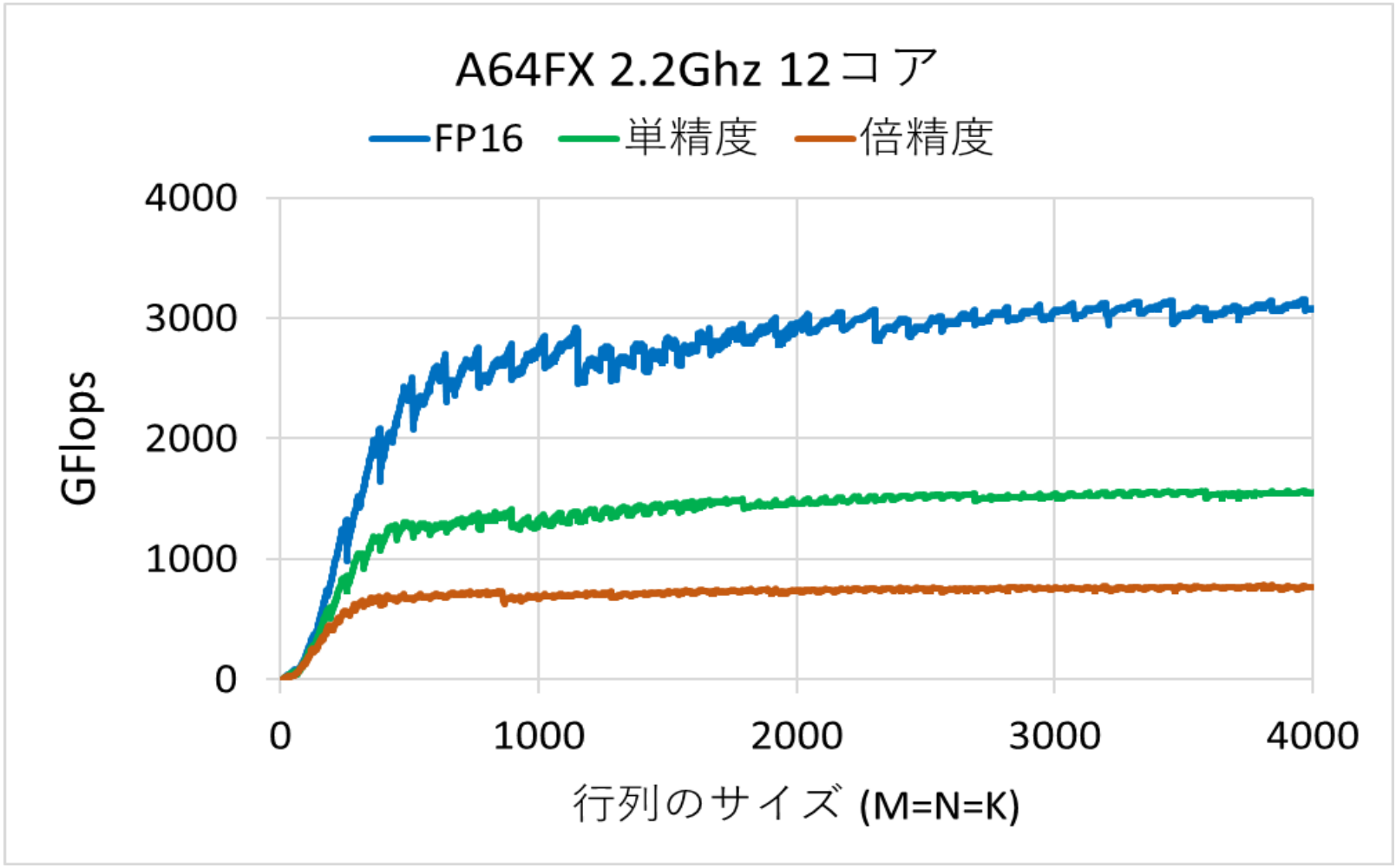


複数CMG向けチューニングの効果



FP16、単精度、倍精度の性能

- 倍精度は1命令で8要素計算するが、単精度は16要素、FP16は32要素を1命令で計算できる → 性能は倍精度の2倍、4倍



まとめ

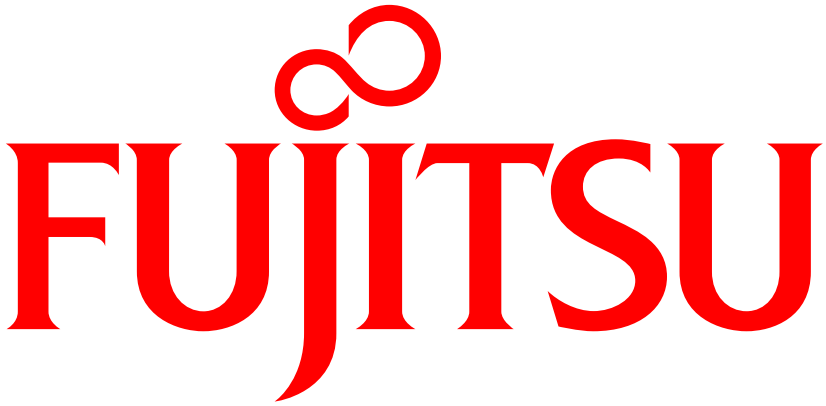
FUJITSU

- A64FX向けにGEMMをチューニング
  - レジスタブロッキング、キャッシュブロッキング
  - CMG共有の8MiBのL2を有効に利用

最大96%の高効率を実現

→ 富岳Top500世界一獲得に貢献

- 並列性能の改善
  - 等分割の余り部分の性能ダウンを、2段階の分割することで回避
  - 4つのCMGに対応してCMG単位で分割して、それぞれをコア単位で並列化
- 単精度は倍精度の倍、FP16はさらに倍の性能
  - Netlib版BLASにないFP16ルーチンは独自にサポート



shaping tomorrow with you