08b-16
次世代スーパーコンピュータに向けたグランドチャレンジ・アプリケーションの開発

高橋大介
筑波大学大学院システム情報工学研究科

メンバー

- 筑波大
  - 超高速計算システム分野：佐藤，朴，高橋，多田野
  - 素粒子分野：宇川，吉江，蔵増
  - 物質科学分野：岩田
  - 数値解析分野：櫻井
- 東大：物質科学分野：押山，古家，計算機科学分野：辻
- 広島大：素粒子分野：石川
- 兵庫県立大：物質科学分野：重田
- 理化学研究所
  次世代スーパーコンピュータ開発実施本部
  - システム開発チーム：横川，庄司
  - アプリケーション開発チーム：南，杉原，黒田，井上，長谷川
  - 富士通，NEC情報システムズ，NECソフト，高度情報科学技術研究機構：井上，浅見，守田，飯塚

目的

- 理化学研究所と筑波大学との間で，「大規模シミュレーションによる次世代スーパーコンピュータの性能評価に関する共同研究」を行っている．（H19.9～H23.3）
- 今後利用可能になる次世代スーパーコンピュータに向けたグランドチャレンジ・アプリケーションの開発および性能評価を行う．
- グランドチャレンジ・アプリケーションとしては，
  - 筑波大
    - QCD（素粒子分野）
    - RSDFT（物質科学分野）
    - FFTなど（数値計算ライブラリ）
  - 理研
    - ライフサイエンス，ナノサイエンス，地球科学など

を想定している．

平成20年度後期の成果

- 実空間密度汎関数法（RSDFT）による大規模第一原理計算
  - 次世代スパコンでは，実空間＋バンド並列が必須となる.
  - OpenMPI/MPIハイブリッド並列化を行い，その性能を評価した.
- 次世代スーパーコンピュータに向けた高速フーリエ変換（FFT)のアルゴリズムに関する研究
  - 一次元FFTおよび三次元FFT
  - 大規模並列に適したデータ分割方法の検討
- 地球科学，ナノ分野，流体アプリケーション等
  - 大規模並列における性能測定
  - キャッシュを生かした単体チューニング

次世代スーパーコンピュータに向けた高速フーリエ変換（FFT）のアルゴリズムに関する研究

背景

- 2008年11月のTop500において，2システムがPFlopsの大台を突破している.
  - Roadrunner: 1,105.00 TFlops (129,600 Cores)
  - Jaguar（Cray XT5 QC 2.3GHz）: 1,059.00 TFlops (150,152 Cores)
- 今後出現すると予想される，10PFlops級マシンは，ほぼ確実に10万コアを超える規模のものになると予想される.
  - MPIとOpenMPのハイブリッド実行を行ったとしても，MPIプロセス数が1万個以上になる可能性がある.

目的

- 並列三次元FFTにおける典型的な配列の分散方法
  - 三次元（x，y，z方向）のうちの一次元のみ（例えばz方向）のみを分割して配列を格納.
  - MPIプロセスが1万個の場合，z方向のデータ数が1万点以上でなければならず，三次元FFTの問題サイズに制約.
- x，y，z方向に三次元分割する方法が提案されている[Eleftheriou et al. ’05, Fang et al. ’07].
  - 各方向のFFTを行う都度，全対全通信が必要.
- 本研究では，二次元分割を行うことで全対全通信の回数を減らしつつ，比較的少ないデータ数でも高いスケーラビリティを得ることを目的とする.

三次元FFT

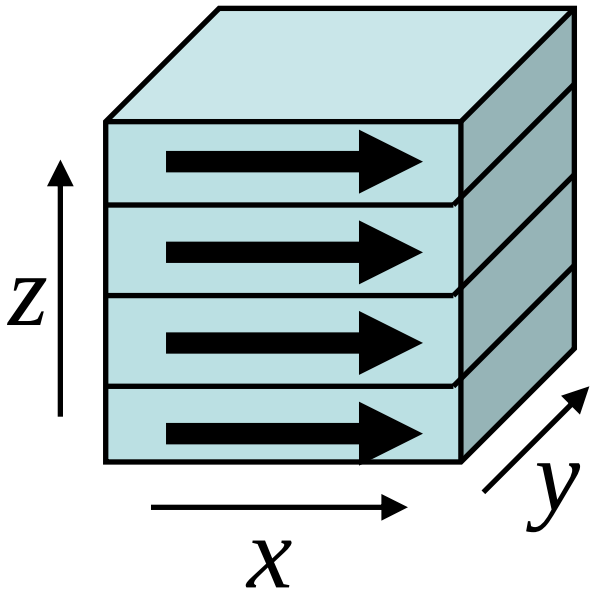
- 三次元離散フーリエ変換（DFT）の定義

$$y(k_1,k_2,k_3)=\sum_{j_1=0}^{n_1-1}\sum_{j_2=0}^{n_2-1}\sum_{j_3=0}^{n_3-1}x(j_1,j_2,j_3)\,\omega_{n_3}^{j_3k_3}\omega_{n_2}^{j_2k_2}\omega_{n_1}^{j_1k_1}$$

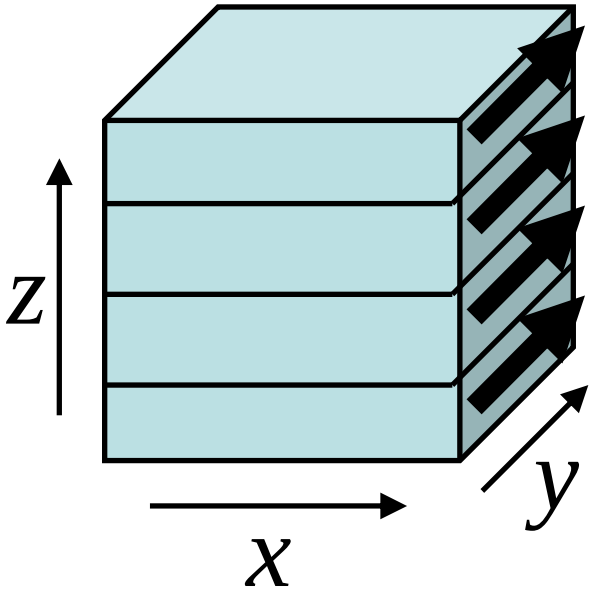
$$0\le k_r\le n_r-1,\ \omega_{n_r}=\exp(-2\pi i/n_r)$$

z方向に一次元ブロック分割した場合の並列三次元FFT

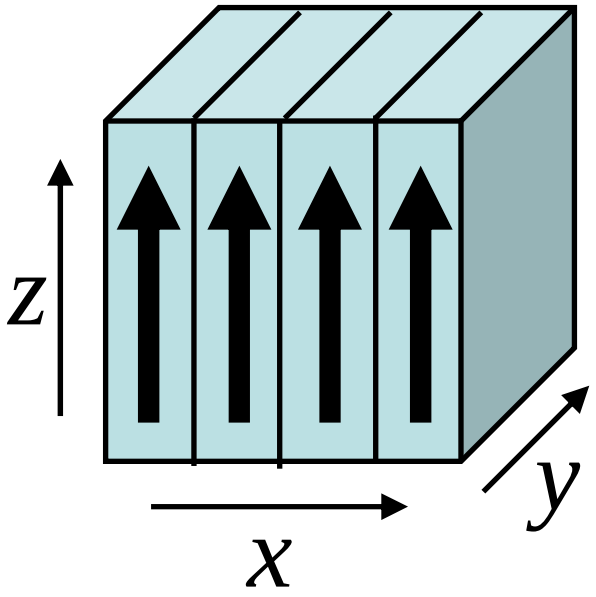
1. x方向FFT



2. y方向FFT



3. z方向FFT



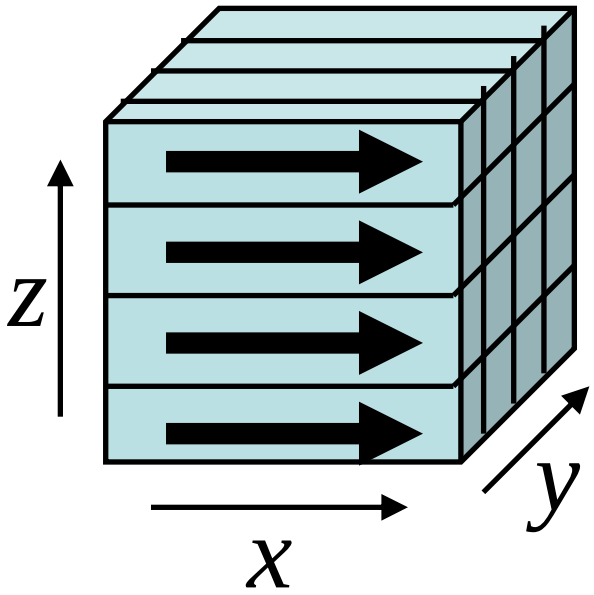
各プロセッサでslab形状に分割

三次元FFTの超並列化

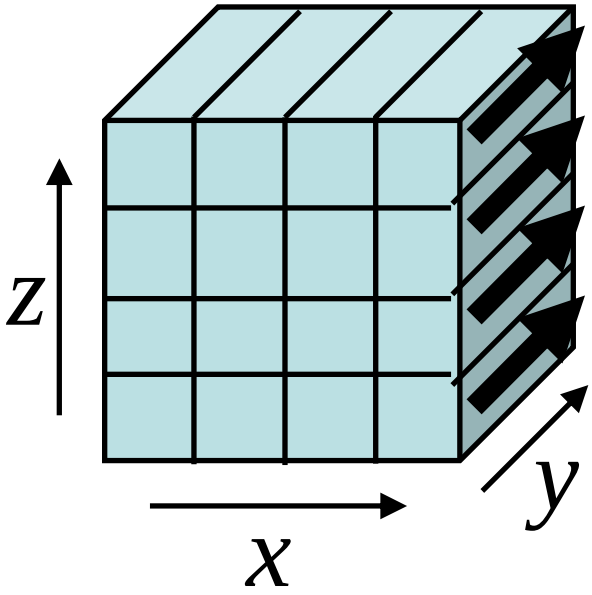
- 並列アプリケーションプログラムのいくつかにおいては，三次元FFTが律速になっている.
- x，y，zのうちz方向のみに一次元分割した場合，超並列化は不可能.
  - 1,024×1,024×1,024点FFTを2,048プロセスで分割できない（1,024プロセスまでは分割可能）
- y，zの二次元分割で対応する.
  - 1,024×1,024×1,024点FFTが1,048,576（=1,024×1,024）プロセスまで分割可能になる.

y, z方向に二次元ブロック分割した場合の並列三次元FFT

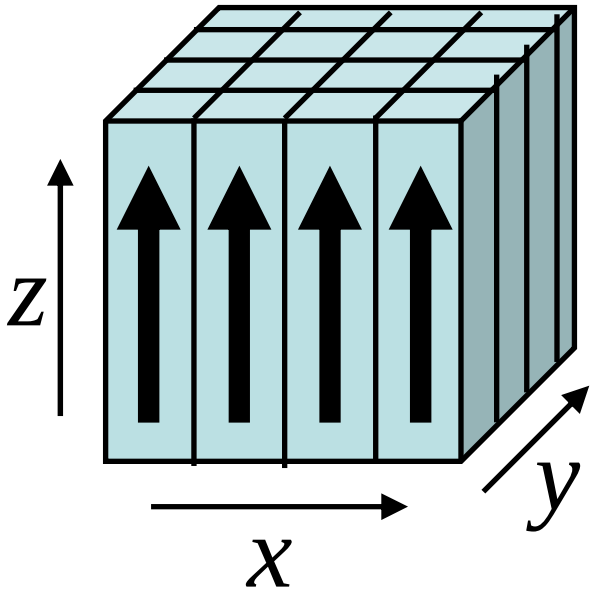
1. x方向FFT



2. y方向FFT



3. z方向FFT



各プロセッサで直方体形状に分割

二次元分割並列三次元FFTの実装

- 二次元分割した場合, $P \times Q$ 個のプロセッサにおいて,
  - P 個のプロセッサ間で全対全通信を Q 組
  - Q 個のプロセッサ間で全対全通信を P 組

  行う必要がある.
- MPI_Comm_Split()を用いてMPI_COMM_WORLDをy方向（P プロセッサ）とz方向（Q プロセッサ）でコミュニケータを分割した.
  - 各コミュニケータ内でMPI_Alltoall()を行う.
- 入力データがy, z方向に, 出力データはx, y方向に二次元ブロック分割されている.
  - 全対全通信はy方向で1回, z方向で1回の合計2回で済む.

一次元分割の場合の通信時間

- 全データ数を N, プロセッサ数を $P \times Q$, プロセッサ間通信性能を W（Byte/s）, 通信レイテンシを L （sec）とする.
- 各プロセッサは $N/(PQ)^2$ 個の倍精度複素数データを自分以外の $PQ-1$ 個のプロセッサに送ることになる.
- 一次元分割の場合の通信時間は

$$T_{1\mathrm{dim}} = (PQ-1)\left(L + \frac{16N}{(PQ)^2 \cdot W} \right)$$

$$\approx PQ \cdot L + \frac{16N}{PQ \cdot W} \quad \text{（sec）}$$

二次元分割の場合の通信時間

- y方向の P 個のプロセッサ間で全対全通信を Q 組行う.
  - y方向の各プロセッサは $N/(P^2Q)$ 個の倍精度複素数データを，y方向の $P-1$ 個のプロセッサに送る.
- z方向の Q 個のプロセッサ間で全対全通信を P 組行う.
  - z方向の各プロセッサは $N/(PQ^2)$個の倍精度複素数データを，z方向の $Q-1$ 個のプロセッサに送る.
- 二次元分割の場合の通信時間は

$$T_{2\dim} = (P-1)\left(L + \frac{16N}{P^2Q \cdot W}\right) + (Q-1)\left(L + \frac{16N}{PQ^2 \cdot W}\right)$$

$$\approx (P+Q)\cdot L + \frac{32N}{PQ \cdot W} \text{（sec）}$$

一次元分割と二次元分割の場合の通信時間の比較

- 一次元分割の通信時間

$$T_{1\mathrm{dim}} \approx PQ \cdot L + \frac{16N}{PQ \cdot W}$$

- 二次元分割の通信時間

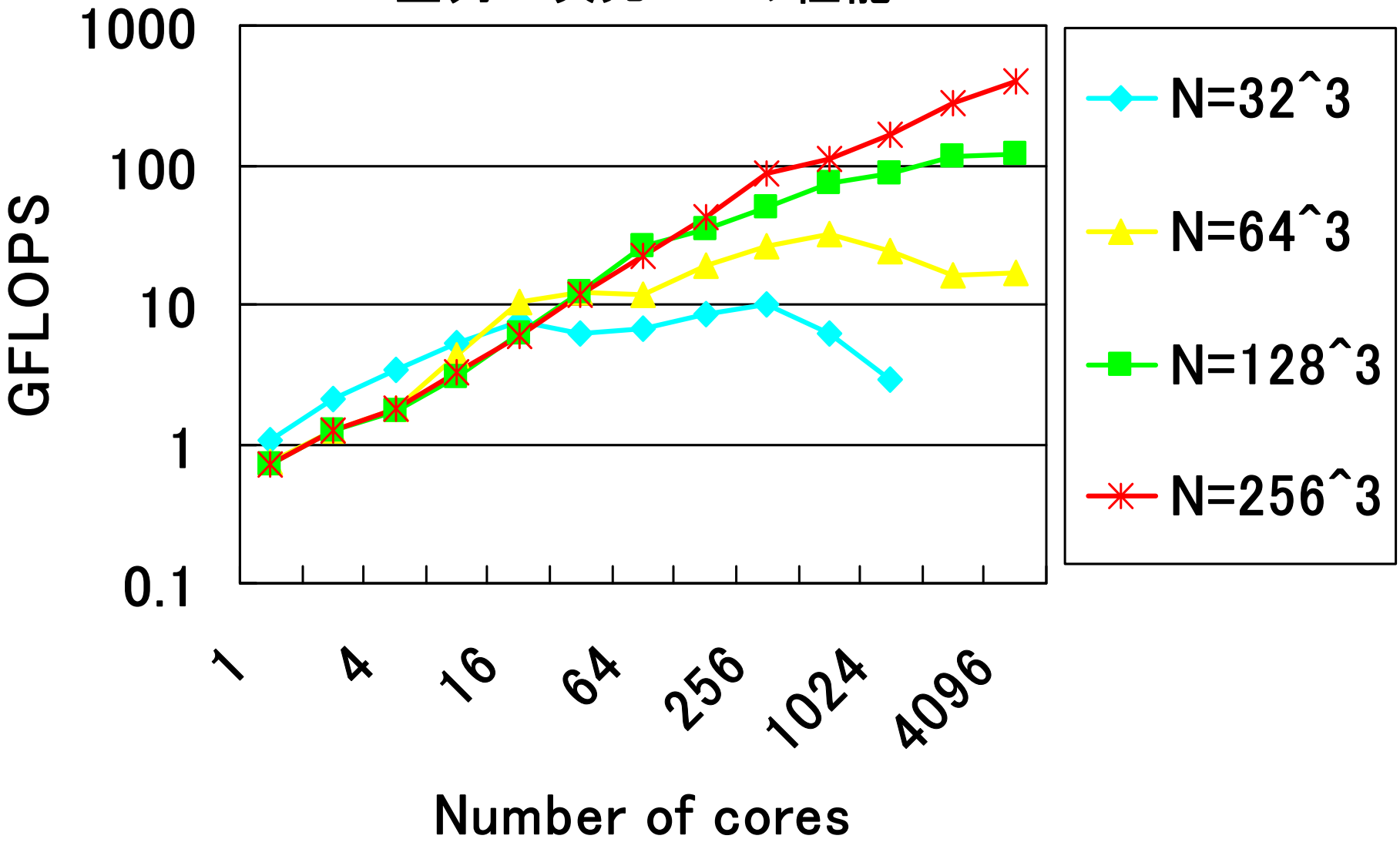
$$T_{2\mathrm{dim}} \approx (P+Q) \cdot L + \frac{32N}{PQ \cdot W}$$

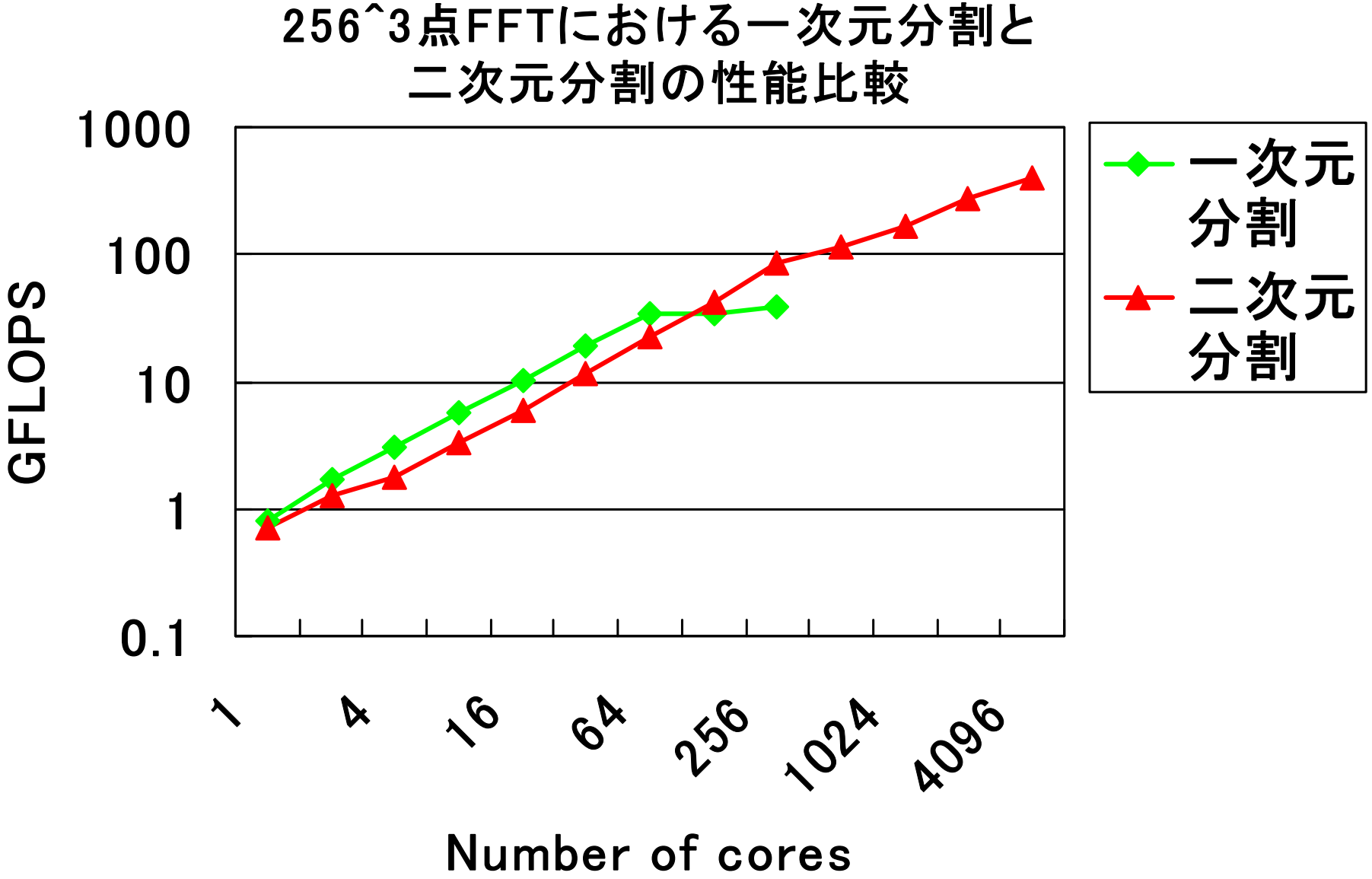
- 二つの式を比較すると，全プロセッサ数 $P \times Q$ が大きく，かつレイテンシ L が大きい場合には，二次元分割の方が通信時間が短くなることが分かる．

性能評価

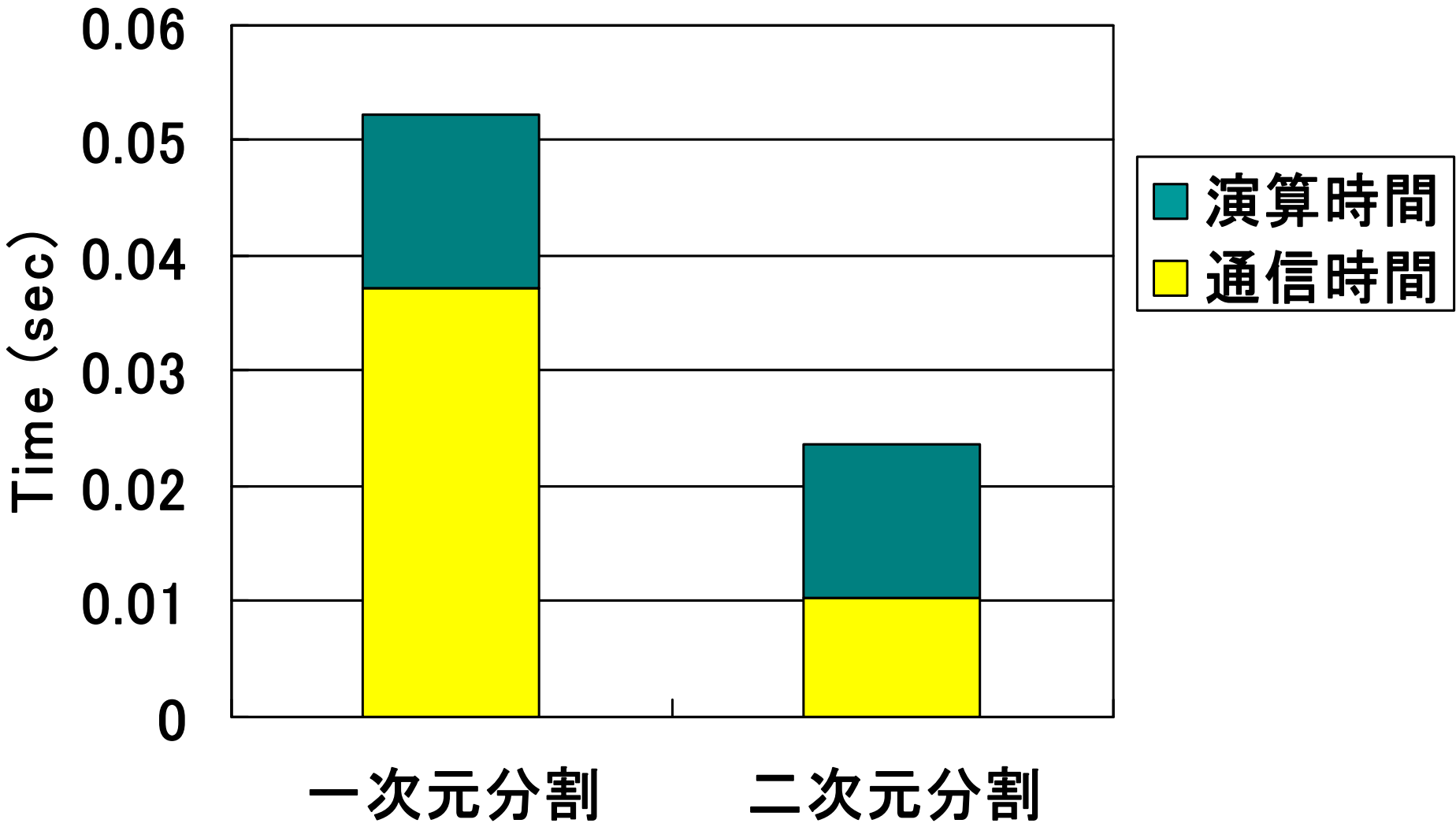
- 性能評価にあたっては，提案する二次元分割を行ったvolumetric並列三次元FFTと，一次元分割を行った並列三次元FFTの性能比較を行った.
- Strong Scalingとして$N = 32^3, 64^3, 128^3, 256^3$ 点の順方向FFTを1～4,096MPIプロセスで連続10回実行し，その平均の経過時間を測定した.
- 評価環境
  - T2K筑波システムの256ノード（4,096コア）を使用
  - flat MPI（1core当たり1MPIプロセス）
  - MPIライブラリ：MVAPICH 1.2.0
  - Intel Fortran Compiler 10.1
  - コンパイルオプション：”ifort –O3 –xO”（SSE3ベクトル命令）

二次元分割を行ったvolumetric 並列三次元FFTの性能




256^3点FFTにおける一次元分割と
二次元分割の性能比較
GFLOPS
1000
100
10
1
0.1
1
4
16
64
256
1024
4096
Number of cores
一次元分割
二次元分割

並列三次元FFTの実行時間の内訳
(256cores, 256^3点FFT)



まとめ

- ナノ分野グランドチャレンジアプリケーションの一つである3D-RISMにおいて，並列三次元FFTが律速となっている.
- 並列三次元FFTにおいて，二次元分割により通信時間を削減することで，MPIプロセス数が多い場合に性能を改善した.
- T2K筑波システムの4,096コアを用いて性能評価を行った結果，N=256^3点FFTにおいて401GFlopsを超える性能が得られた.