シミュレーションが 未来をひらく

ペタフロップス・アプリケーションを目指して
- 第一原理電子状態計算プログラムRSDFT -

理化学研究所　計算科学研究機構
運用技術部門　ソフトウェア技術チーム
長谷川　幸弘

HPCI分野5全体シンポジウム

K computer

ゴードン・ベル賞/最高性能賞

受賞内容

- First-principles calculations of electron states of a silicon nanowire with 100,000 atoms on the K computer
- RSDFT: Real Space Density Functional Theory code
- 筑波大学，東京大学，理研の三極で協力
- 現実規模のシリコンナノワイヤーのシミュレーション
- 1万原子～4万原子まで計算
- 10万原子の性能評価を実施
  - 実効性能：3.08PFLOPS
    (7.06PFLOPS構成)
  - 実行効率：43.6%
  - 現実的時間で可能

SC11@Seatle,USA

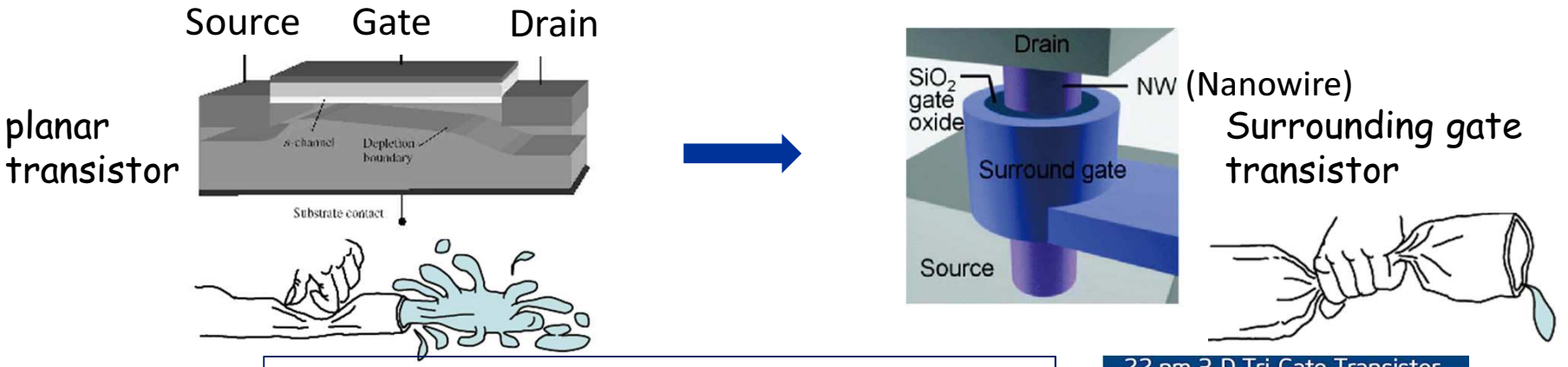
Outline

- ターゲット
- RSDFTの概要
- 京の概要
- 京向けのチューニング
- 性能ベンチマーク
- 科学的成果

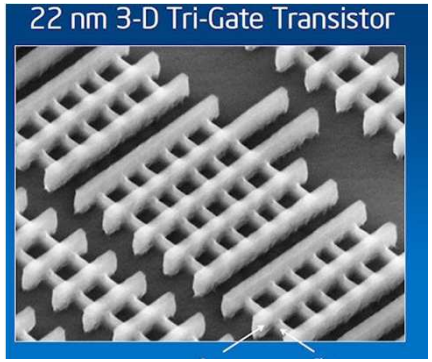
ターゲット: シリコンナノワイヤの電子状態の予測

Si nanowire (SiNW), a booster
in the next-generation semiconductor technology
More Moore → More than Moore

http://www.itrs.net/



Gate controllability
→ suppress short-channel effects
suppress leaks at off state
→ save energy



Actually tri-gate by Intel in 2011

Number of atoms in SiNW channels
→ 10,000 - 100,000 atoms !

Outline

- ターゲット
- **RSDFTの概要**
- 京の概要
- 京向けのチューニング
- 性能ベンチマーク
- 科学的成果

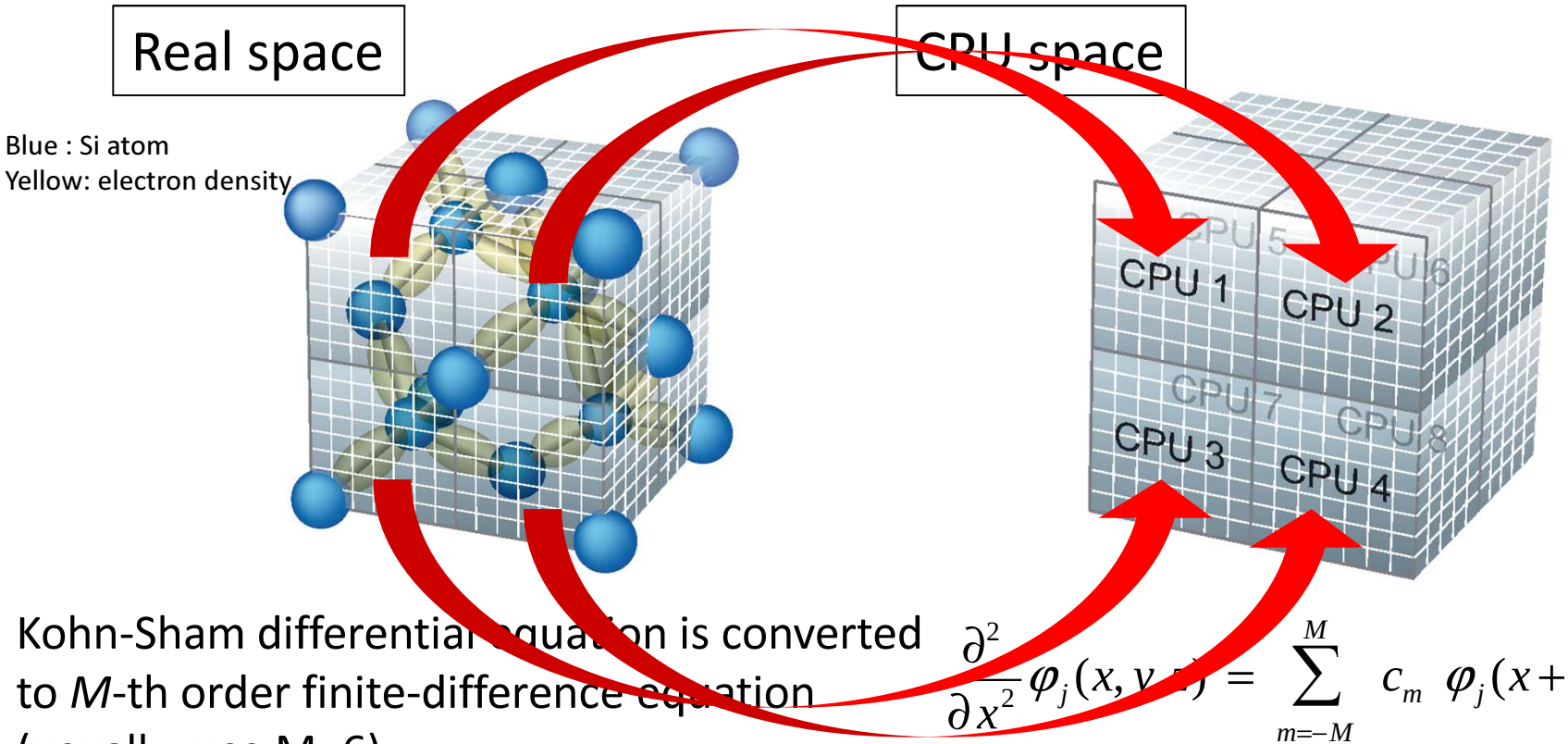
RSDFTのスペック

RSDFT: Real Space Density Functional Theory code

- 筑波大学（現在、東京大学）岩田先生が開発
- 量子力学に基づく第一原理電子状態計算プログラム
- 基本原理は密度汎関数理論
- 擬ポテンシャル法を採用
- 実空間差分法を採用
  - FFT(Fast Fourier Transformation)を必要としない、超並列計算向き
- 高い実行効率を実現
  - BLAS level.3の行列-行列積での実装
- ハイブリッド並列（MPI + OpenMP）を実装
- Fortran90，5万行
- シミュレーション対象：金属、半導体、表面・界面、欠陥、ナノ構造

実空間差分法と並列化

J.-I. Iwata *et al*., J. Comp. Phys. (2010)



Kohn-Sham differential equation is converted to *M*-th order finite-difference equation (usually use M=6)

$$\frac{\partial^2}{\partial x^2}\varphi_j(x,y,z) = \sum_{m=-M}^{M} c_m \, \varphi_j(x+mH,y,z)$$

Advantages

- Almost free from FFT, reducing communication burden ⇒ high efficiency
- Flexible boundary condition to wave-functions ⇒ molecules, clusters, surfaces, etc.

SCF計算のフロー

Kohn-Sham 方程式

ハミルトニアン

$$\left[-\frac{1}{2}\nabla^2 + v_{\mathrm{nucl}}(\mathbf{r}) + \int \frac{n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}' + \frac{\delta E_{\mathrm{xc}}[n]}{\delta n(\mathbf{r})}\right] \varphi_i(\mathbf{r}) = \varepsilon_i \varphi_i(\mathbf{r})$$

波動関数

電子密度 $n(\mathbf{r}) = \sum_i |\varphi_i(r)|^2$

ϕ_i ：電子軌道（=波動関数）
i ：電子準位（=エネルギーバンド）
r ：空間離散点（=空間格子）

Self-Consistent Field procedure

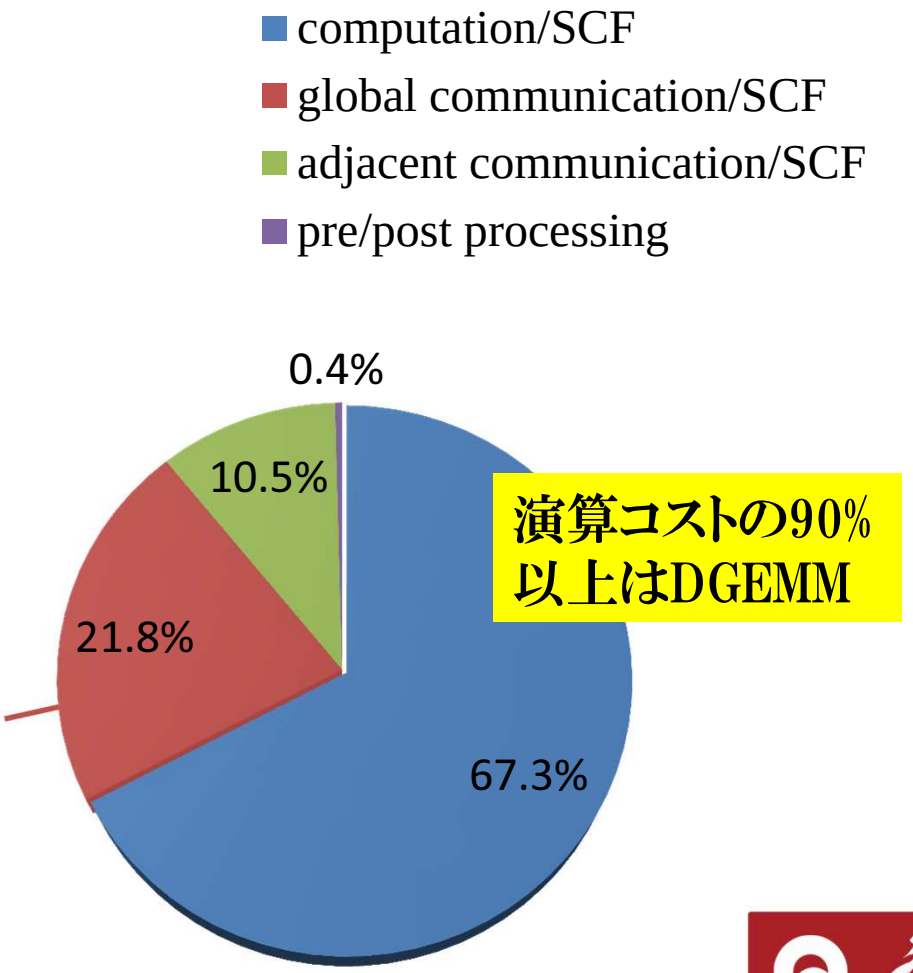
		計算量（原子数: N）
1	（CG）共役勾配法	$O(N^2)$
2	（GS）グラムシュミット正規直交化	$O(N^3)$
3	密度とポテンシャルの更新	$O(N)$
4	（SD）部分対角化	$O(N^3)$

計算コストの分析

- 8,000 原子のシリコンナノワイヤ(SiNW)
  - ✓ 反復回数：100
  - ✓ 格子数 ：120×120×20
  - ✓ バンド数 ：16,000
- 並列プロセス数：256
- 256 ノード (2,048 コア)

MPI通信のバリエーションが多い
（関数の種類、転送サイズ）

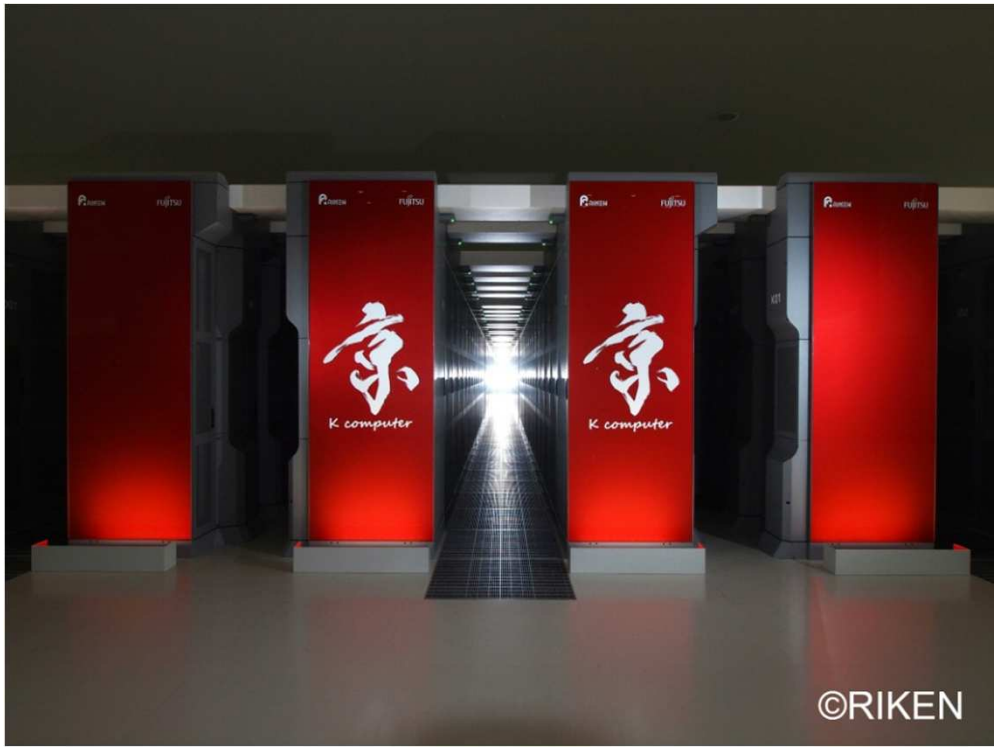
- MPI_Allreduce
- MPI_Reduce
- MPI_Bcast
- MPI_Allgatherv



Outline

- ターゲット
- RSDFTの概要
- **京の概要**
- 京向けのチューニング
- 性能ベンチマーク
- 科学的成果

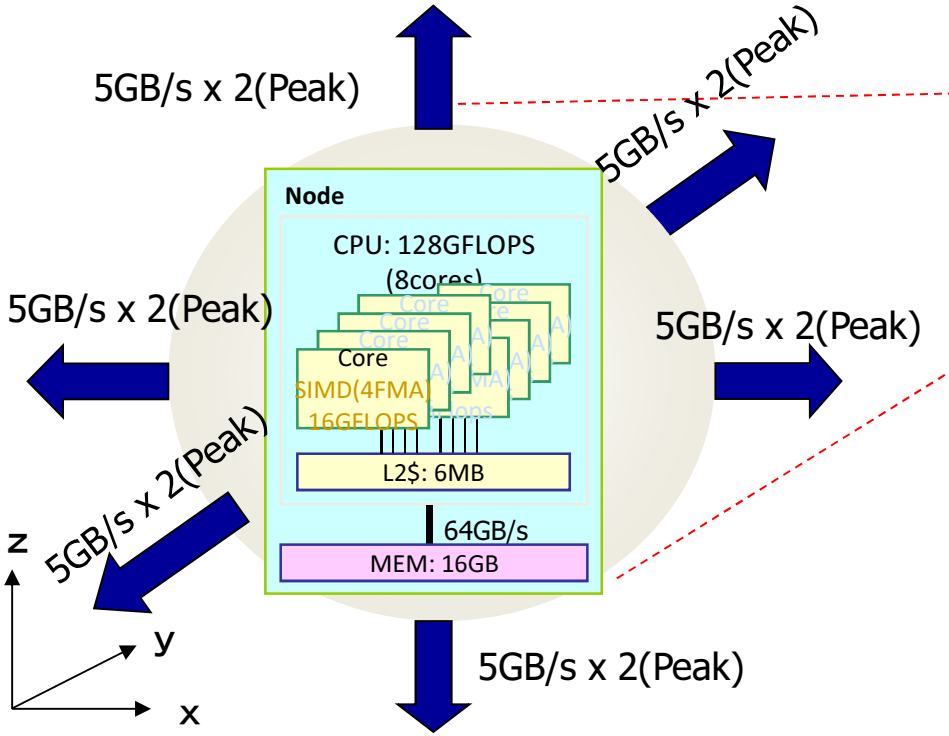
スーパーコンピュータ「京」の特徴

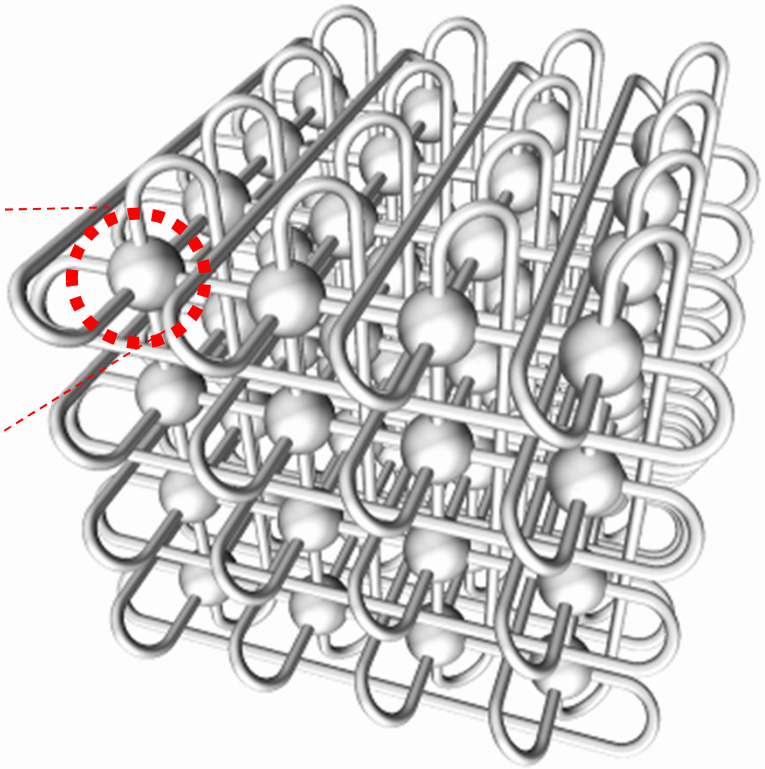


Total Compute Node(CPU) : 82,944
Total CPU Core : 663,552
Total Peak Performance : 10.62 PFLOPS
Total Memory : 1.27PiB

計算ノードとインターコネクトネットワーク

- Compute Node (128GFLOPS)
  - CPU(SPARC64™ VIIIfx) : 1 (8 cores)
  - Memory capacity : 16GB
  - L1D cache: 32KB
  - L2 cache : 6MB (Share)
  - ICC (Interconnect Controller) : 1
- Tofu(Torus fusion) interconnect
  - 6 dimensional mesh/torus network
  - User's programing point of view : Logical 3 dimensional torus network
  - Peak bandwidth: 5GB/s (each direction)





(courtesy of Fujitsu Ltd.)

Outline

- ターゲット
- RSDFTの概要
- 京の概要
- **京向けのチューニング**
- 性能ベンチマーク
- 科学的成果

京向けのチューニング –演算部–

- 演算部
  - 様々なチューニングが既に実装済み
  - 行列積の計算が演算処理の90%以上を占める.
  - チューニングされたDGEMMが必要.

→ 効率96.6%

RSDFT

- 実空間差分法
- ベクトルの内積計算が基本
- 空間並列

計算コアの最適化
- 行列積化

ターゲット計算機：PACS-CS, T2K-Tsukuba

スレッド並列の実装

ターゲット計算機：PACS-CS, T2K-Tsukuba

計算コアの行列積化 -Gram-Schmidt -

ベクトル積を行列積に変換

再帰分割法

$$\varphi_1 = \psi_1$$

$$\varphi_2 = \psi_2 - \varphi_1\langle\varphi_1{}^*|\psi_2\rangle$$

$$\varphi_3 = \psi_3 - \varphi_1\langle\varphi_1{}^*|\psi_3\rangle - \varphi_2\langle\varphi_2{}^*|\psi_3\rangle$$

$$\varphi_4 = \psi_4 - \varphi_1\langle\varphi_1{}^*|\psi_4\rangle - \varphi_2\langle\varphi_2{}^*|\psi_4\rangle - \varphi_3\langle\varphi_3{}^*|\psi_4\rangle$$

$$\varphi_5 = \psi_5 - \varphi_1\langle\varphi_1{}^*|\psi_5\rangle - \varphi_2\langle\varphi_2{}^*|\psi_5\rangle - \varphi_3\langle\varphi_3{}^*|\psi_5\rangle - \varphi_4\langle\varphi_4{}^*|\psi_5\rangle$$

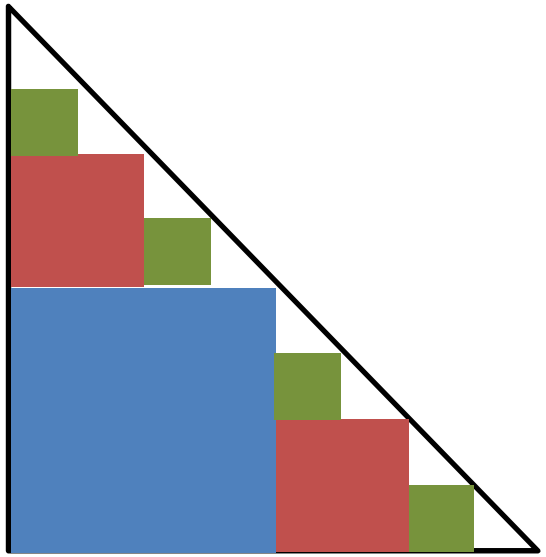
$$\varphi_6 = \psi_6 - \varphi_1\langle\varphi_1{}^*|\psi_6\rangle - \varphi_2\langle\varphi_2{}^*|\psi_6\rangle - \varphi_3\langle\varphi_3{}^*|\psi_6\rangle - \varphi_4\langle\varphi_4{}^*|\psi_6\rangle - \varphi_5\langle\varphi_5{}^*|\psi_6\rangle$$

$$\varphi_7 = \psi_7 - \varphi_1\langle\varphi_1{}^*|\psi_7\rangle - \varphi_2\langle\varphi_2{}^*|\psi_7\rangle - \varphi_3\langle\varphi_3{}^*|\psi_7\rangle - \varphi_4\langle\varphi_4{}^*|\psi_7\rangle - \varphi_5\langle\varphi_5{}^*|\psi_7\rangle - \varphi_6\langle\varphi_6{}^*|\psi_7\rangle$$

$$\varphi_8 = \psi_8 - \varphi_1\langle\varphi_1{}^*|\psi_8\rangle - \varphi_2\langle\varphi_2{}^*|\psi_8\rangle - \varphi_3\langle\varphi_3{}^*|\psi_8\rangle - \varphi_4\langle\varphi_4{}^*|\psi_8\rangle - \varphi_5\langle\varphi_5{}^*|\psi_8\rangle - \varphi_6\langle\varphi_6{}^*|\psi_8\rangle - \varphi_7\langle\varphi_7{}^*|\psi_8\rangle$$

三角部(DGEMV)

四角部(DGEMM)



- 依存関係のある三角部とない四角部にブロック化して計算
- 再帰的にブロック化することで四角部を多く確保

※SDも同様に行列積化が可能

京向けのチューニング　-通信部-

- 通信部
  - 並列軸の拡張
  - ロードバランスの最適化
  - Tofuネットワークへの最適マッピング
    - MPI通信へのTofu向けアルゴリズムの適用
      - 中、長メッセージ向け
      - 直接網向けの集団通信アルゴリズム
      - Bcast, Reduce, Allreduce ： Trinaryx3
      - 複数の通信路を同時に用いるパイプライン転送
      - 隣接プロセスへのRDMA通信のみで構成
      - Tofuライブラリを直接使用する実装のため低レイテンシ

並列軸の拡張 – バンド並列の導入 –

- 従来の空間並列にバンド並列を拡張
  - 通信コストを削減
  - 8万を超える並列性の確保

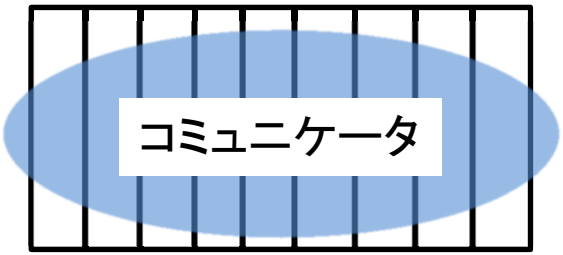
Kohn-Sham 方程式

$$\left[-\frac{1}{2}\nabla^2 + v_{\mathrm{nucl}}(\mathbf{r}) + \int \frac{n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}' + \frac{\delta E_{\mathrm{XC}}[n]}{\delta n(\mathbf{r})}\right] \varphi_i(\mathbf{r}) = \varepsilon_i \varphi_i(\mathbf{r})$$

バンド間の依存性なし

φ_i: 電子軌道(=波動関数)
i : 電子準位(=エネルギーバンド)
r : 空間離散点(=空間格子)

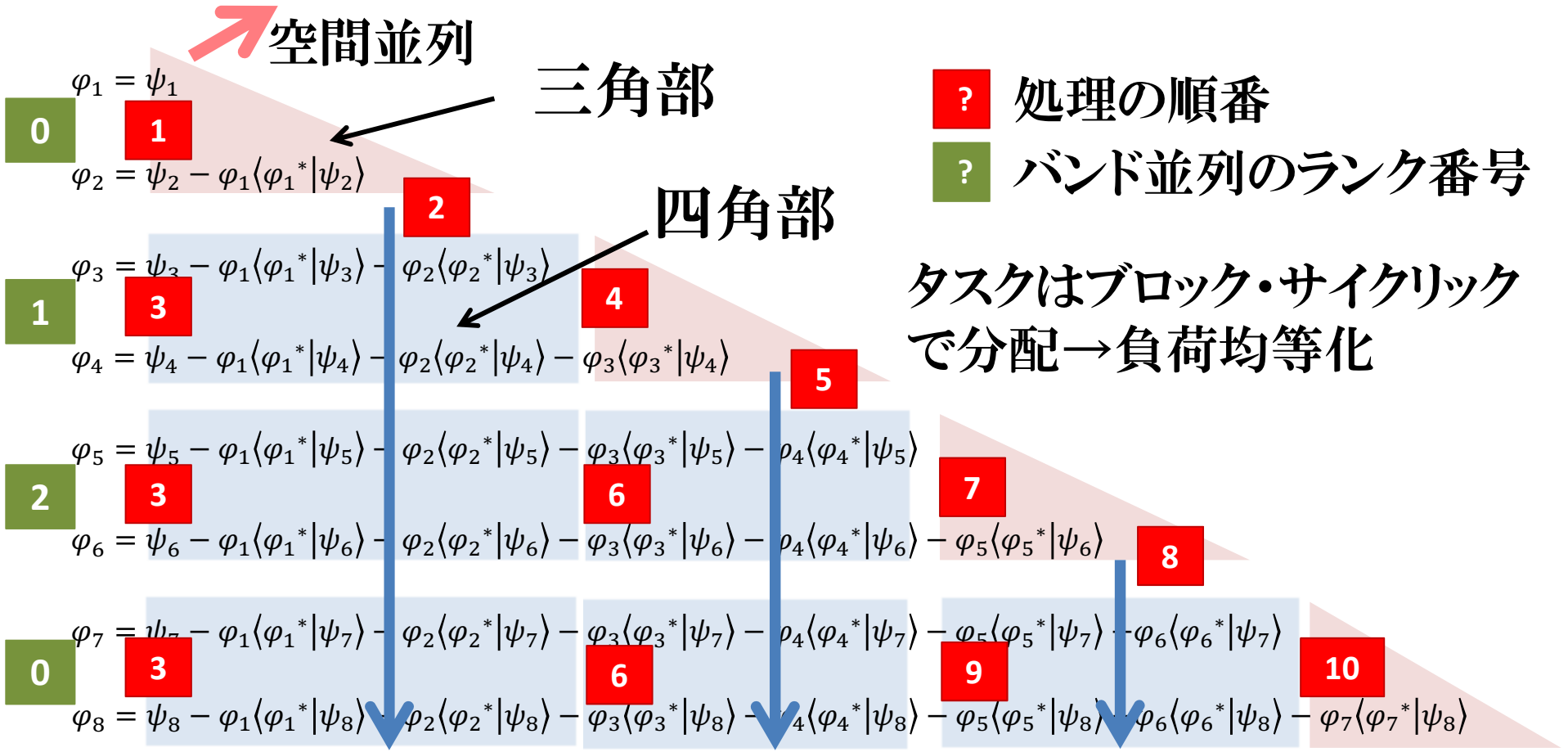
並列軸=1



並列軸=2

通信対象が減る
分割粒度が大きくなる

バンド並列の実装 - Gram-Schmidtの場合 -



? 処理の順番

? バンド並列のランク番号

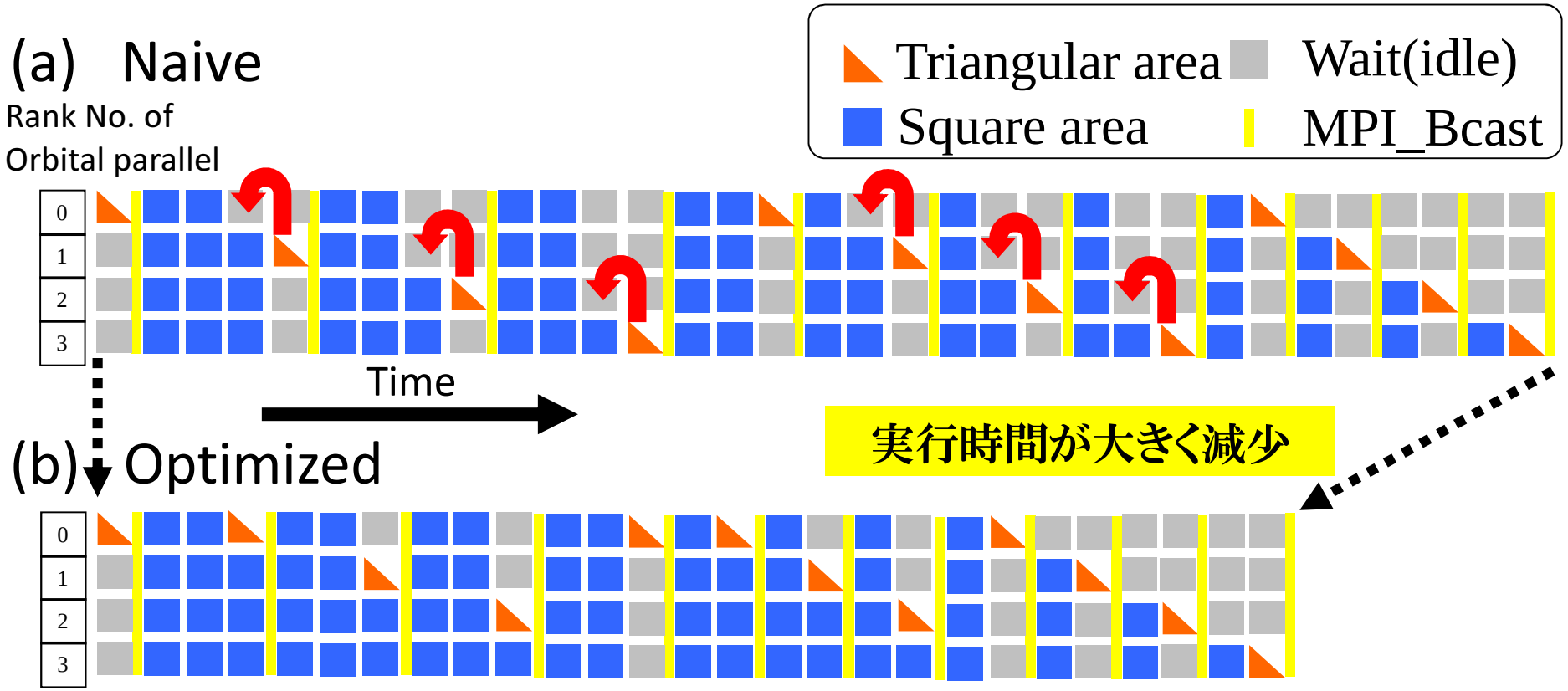
タスクはブロック・サイクリックで分配→負荷均等化

(1)三角部の計算

(2)計算した値を四角部に転送(バンド方向の各プロセッサに分配)

(3)四角部を並列に計算

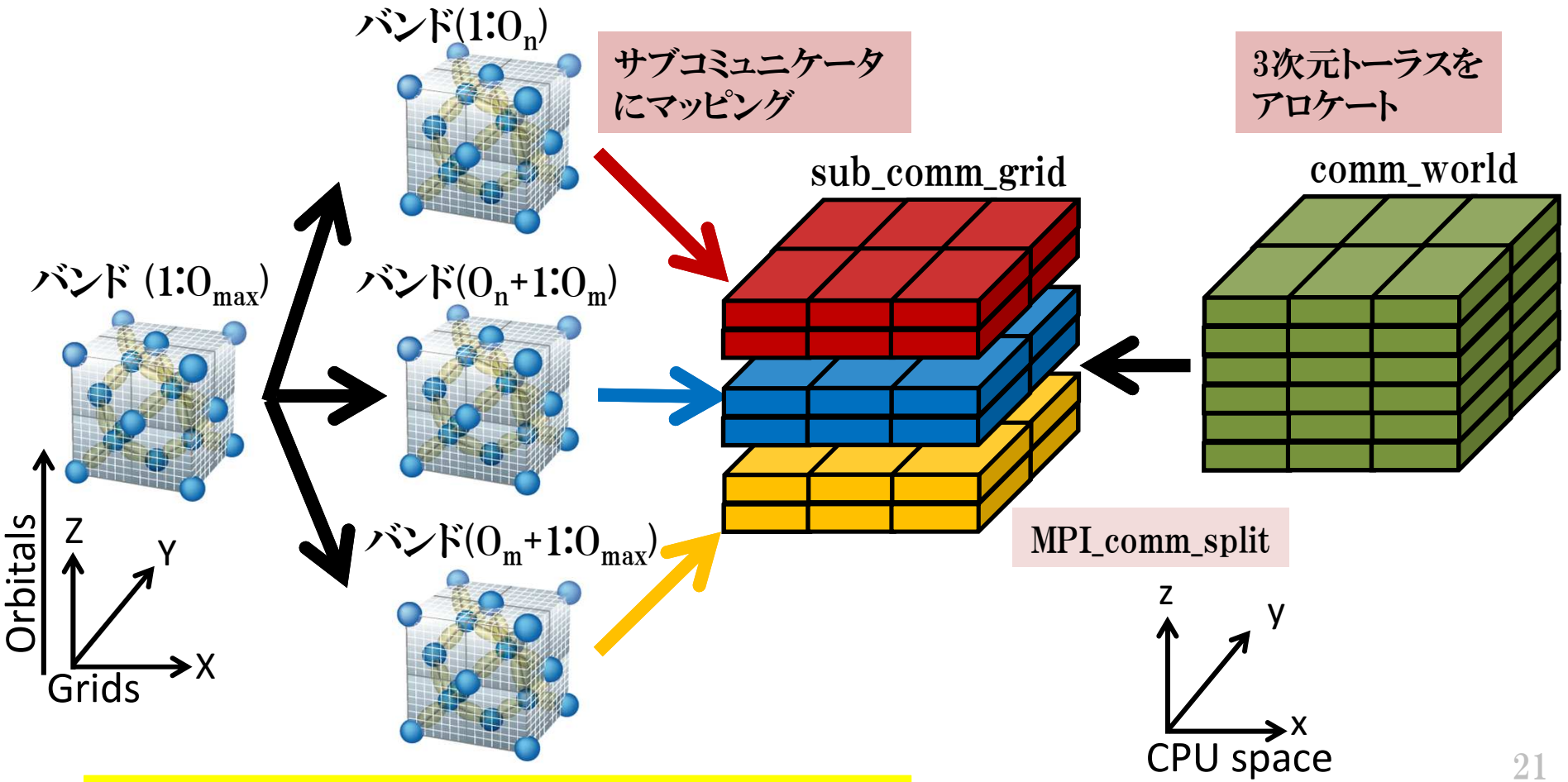
ロードバランス - Gram-Schmidtの場合 -



三角部と四角部の計算をオーバーラップする

Tofuネットワークへのマッピング

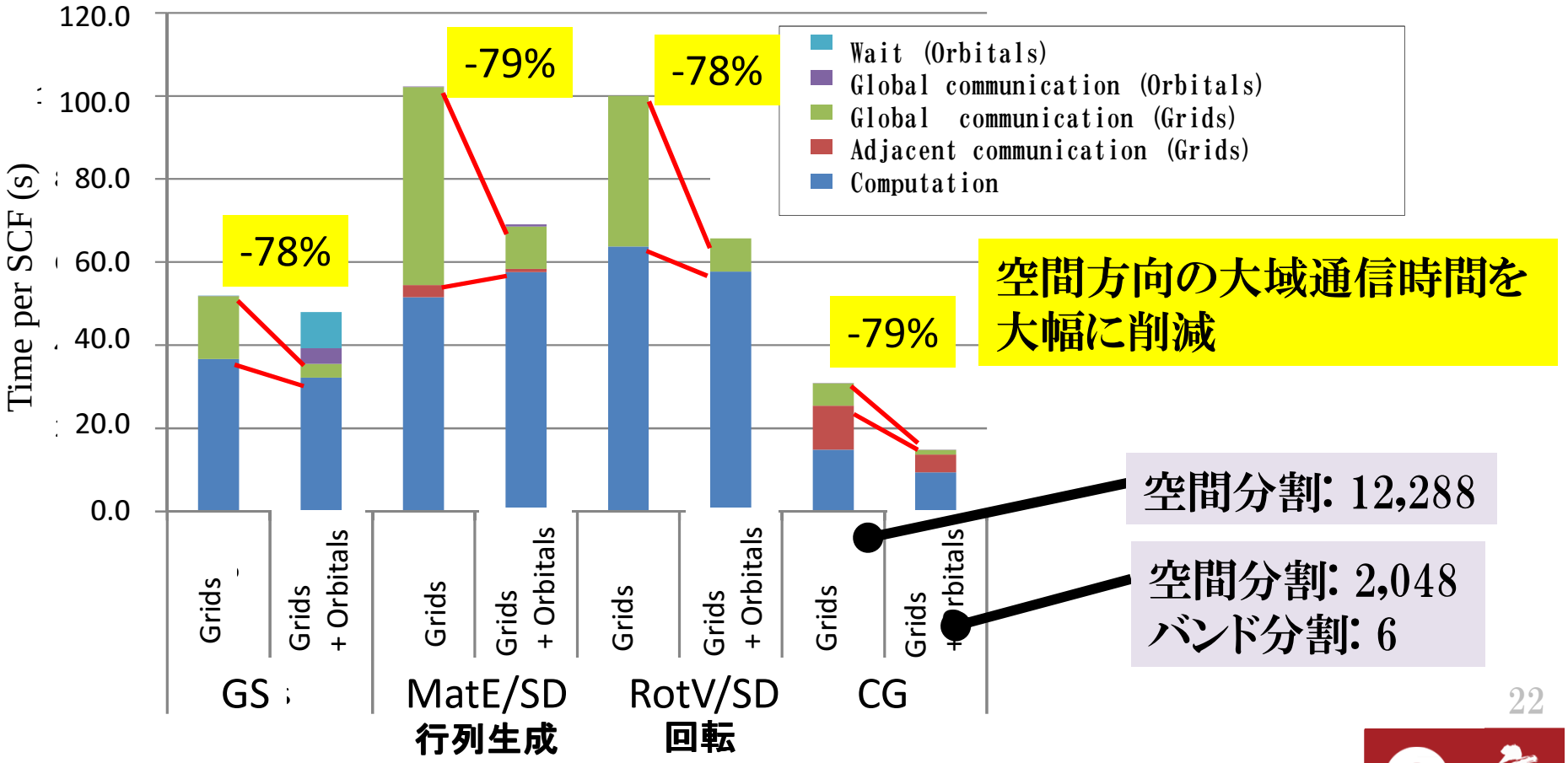
空間並列 | 空間並列＋バンド並列 | Tofuネットワークへのマッピング



サブメッシュ/トーラス内で通信が閉じられる

バンド並列の効果

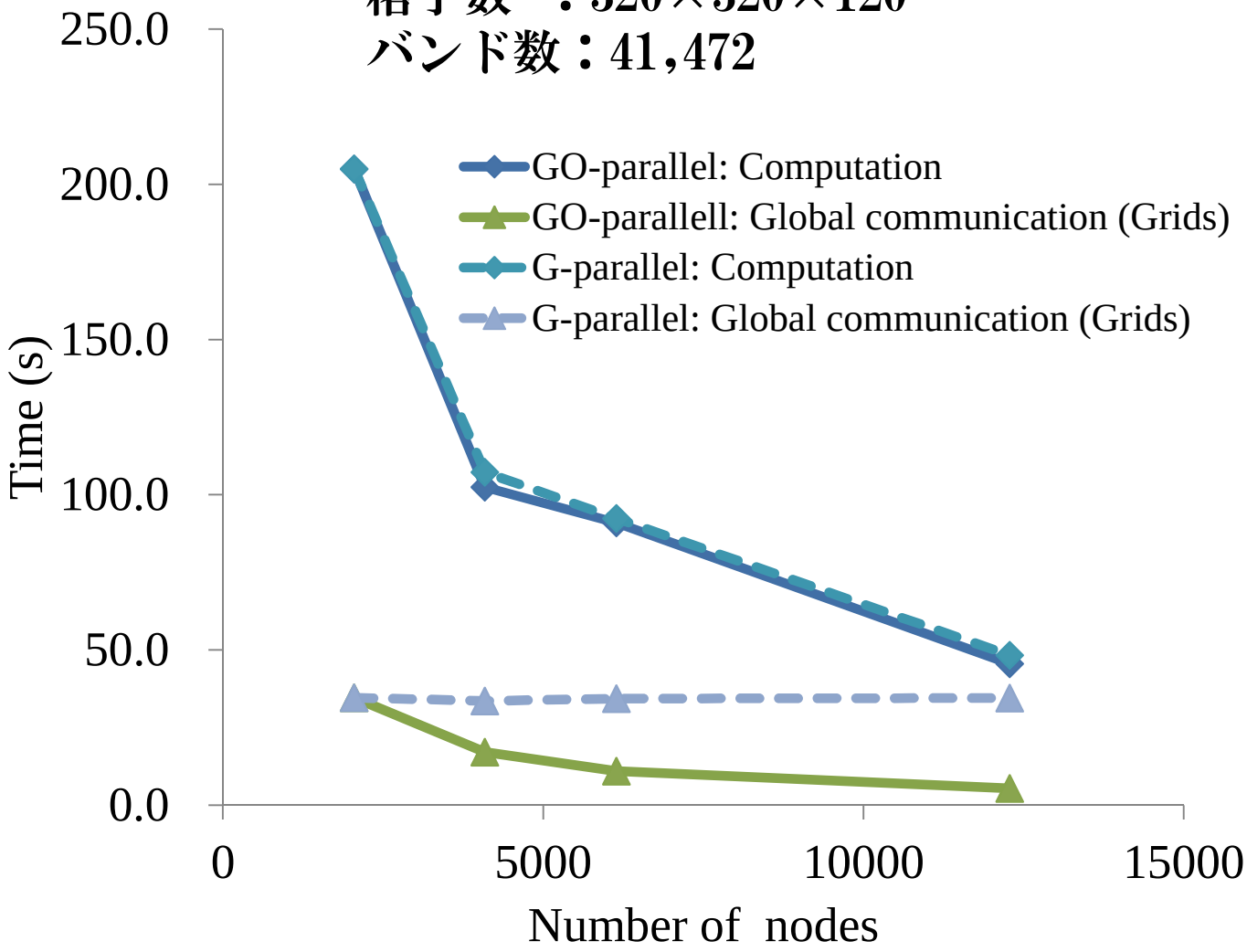
SiNW, 19,848 原子, 格子数:320×320×120, バンド数:41,472
トータル並列プロセス数は12,288で固定



スケーラビリティ　-並列軸の拡張-

RotV/SD

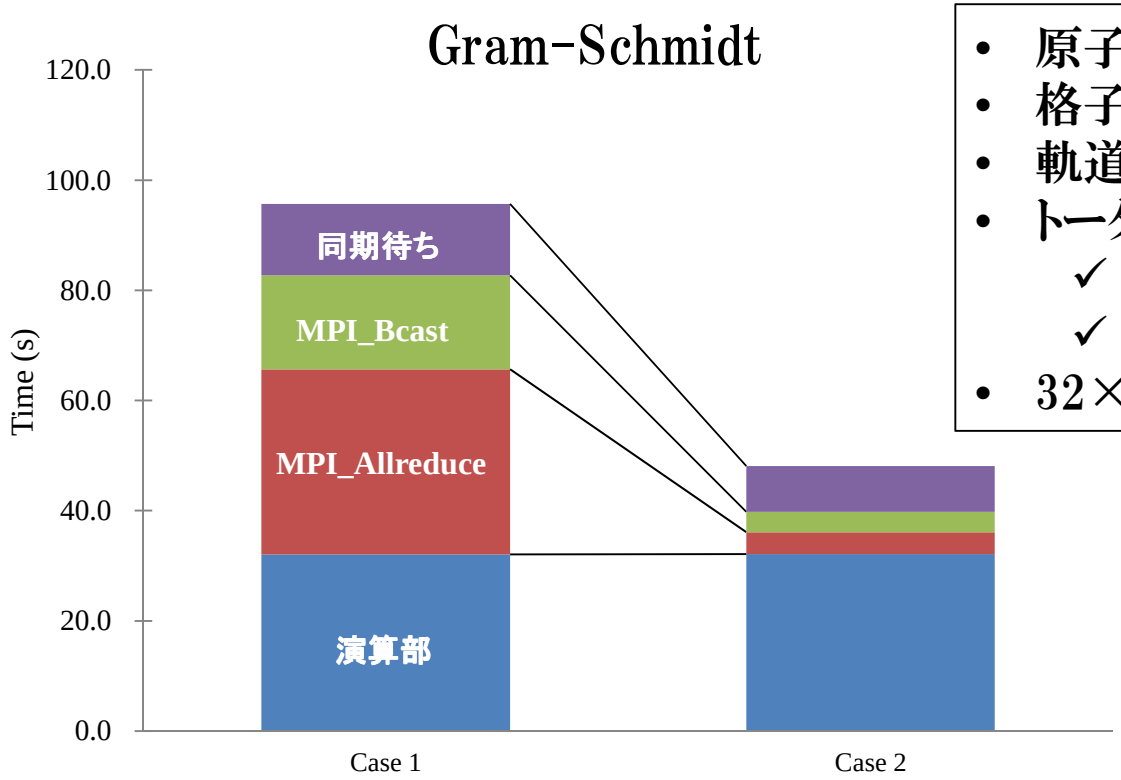
原子数　：19,848
格子数　：320×320×120
バンド数：41,472



最適マッピングの効果

Case 1: 1次元マッピング
Case 2: 3次元マッピング(最適)

- サブコミュニケータ間の通信コンフリクトが発生しない
- MPI通信でTofu向けアルゴリズム(Trinaryx3)が選択される



- 原子数: 19,848
- 格子数: 320×320×120
- 軌道数: 41,472
- トータルプロセス数: 12,288
  - ✓ 空間並列: 2,048(32×32×2)
  - ✓ バンド並列: 6
- 32×32×12のトーラスにマッピング

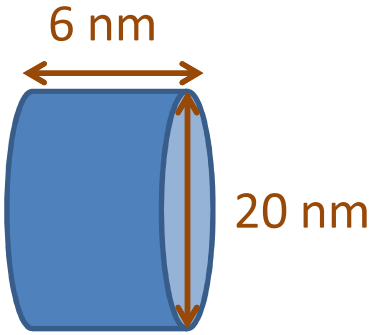
Outline

- ターゲット
- RSDFTの概要
- 京の概要
- 京向けのチューニング
- **性能ベンチマーク**
- 科学的成果

性能ベンチマーク (1/3)

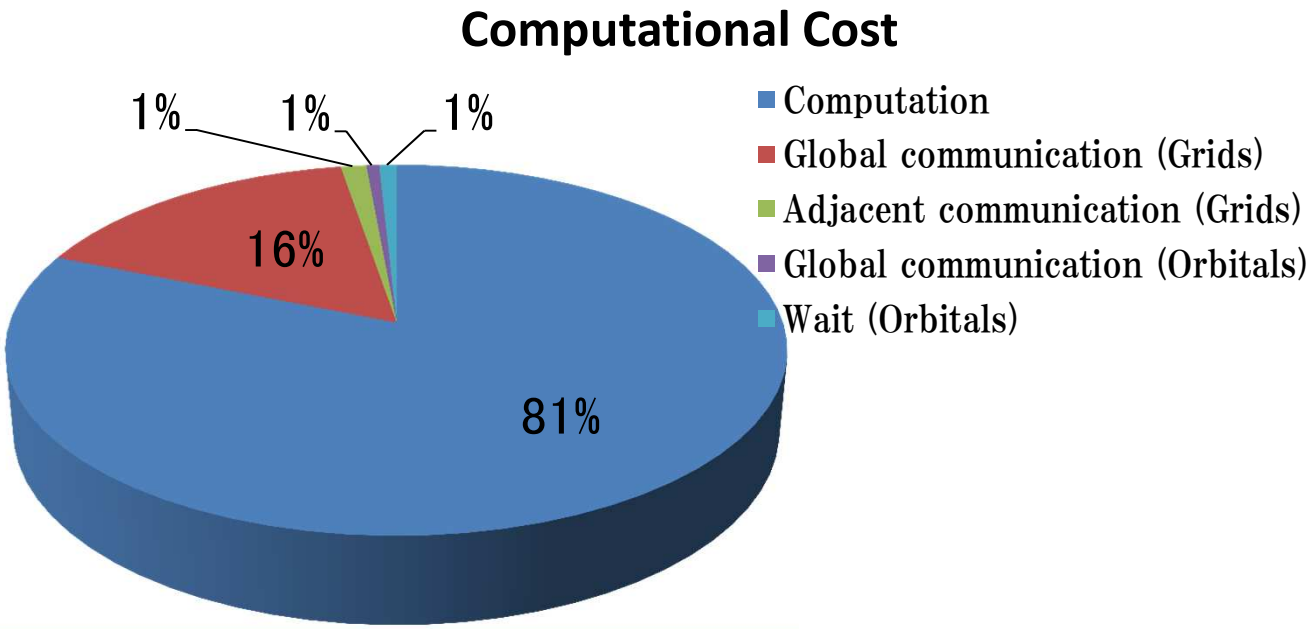
測定条件

- SCF計算の1反復を測定
- SiNW 107,292原子
  - 格子数: 576×576×192
  - バンド数: 230,400
- 並列プロセス数:55,296
  - 空間分割:18,432　×　バンド分割数:3
- 使用ノードのピーク性能: 7.06PFLOPS
  - 55,296ノード(442,368コア)



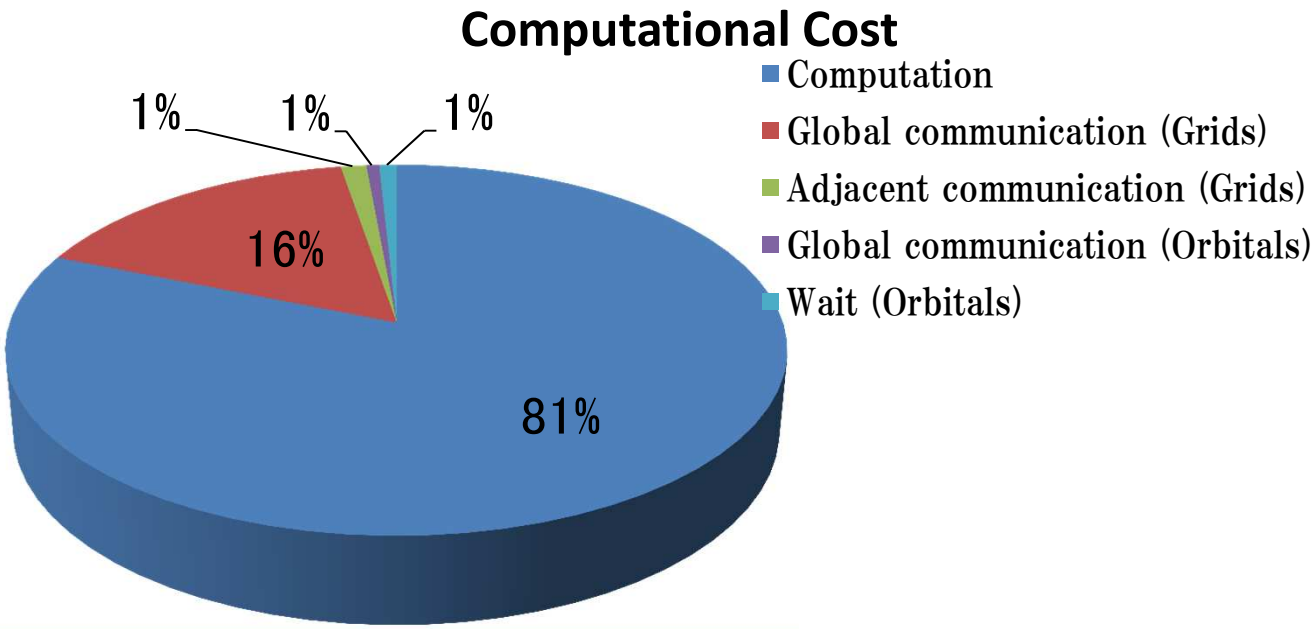
性能ベンチマーク (2/3)

- 実効性能は **3.08PFLOPS** /SCF
- 効率は **43.6 %**
- 通信コストは 19.0%
- SCF計算の反復1回の実行時間は　5,500 秒



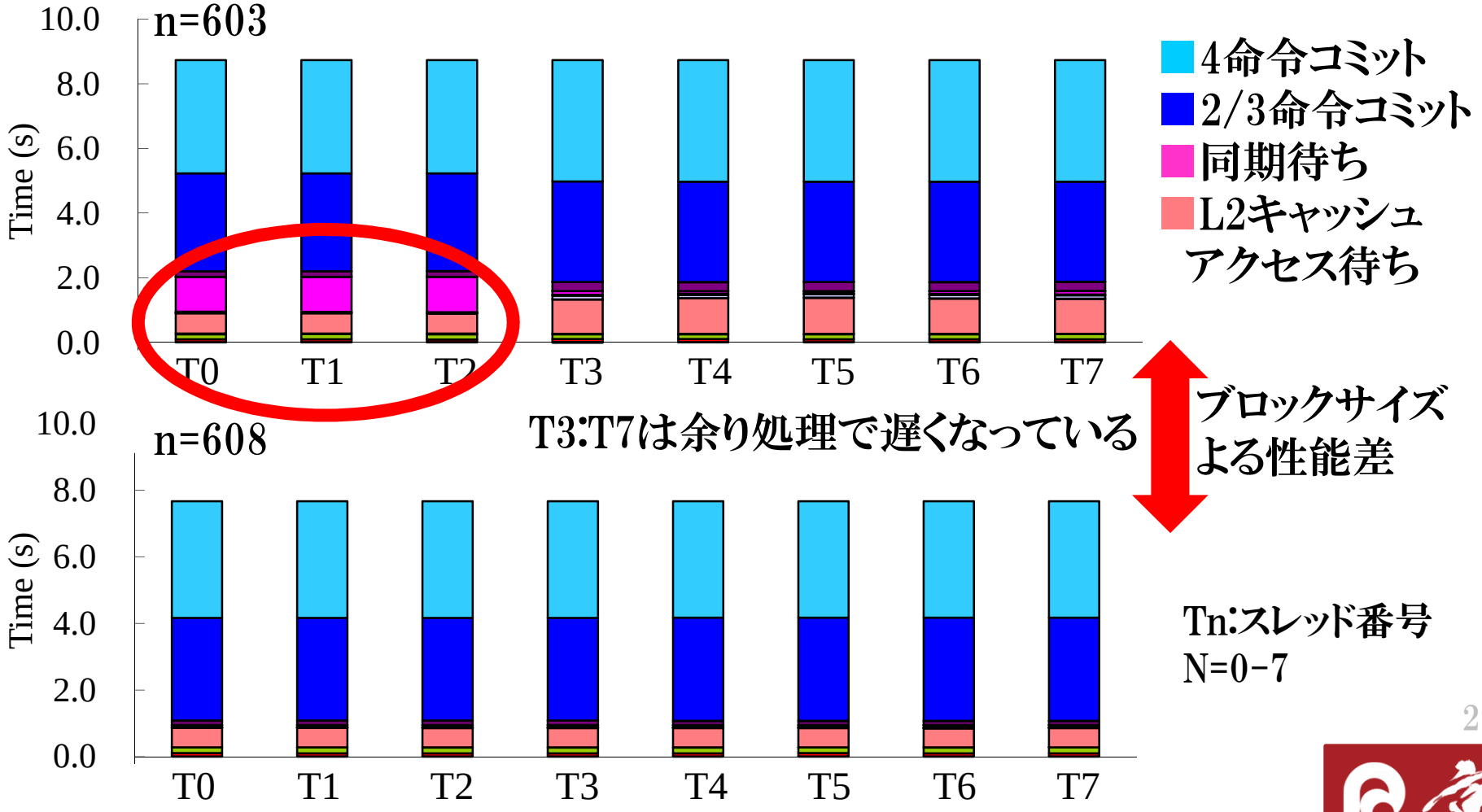
SC11後 性能ベンチマーク (3/3)

- 実効性能は **4.18 PFLOPS** /SCF
- 効率は **59.02 %**
- 通信コストは 19.0%
- SCF計算の反復1回の実行時間は　4,053 秒



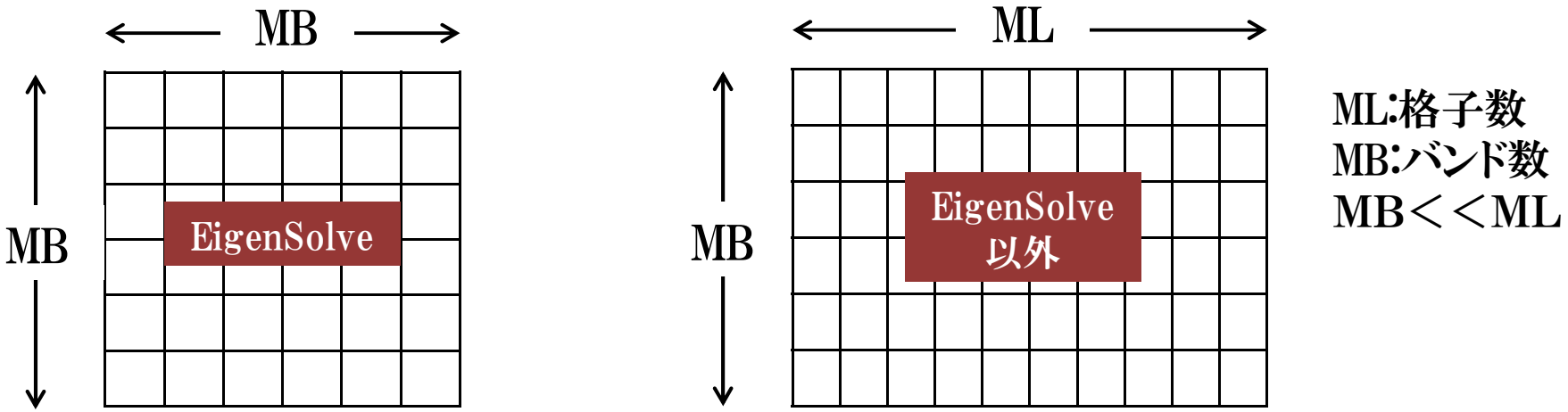
DGEMMのブロックサイズ

call dgemm(transa, transb, m, n, k, alpha, lda, ldb, beta, ldc)
[C] ＝ alpha･[A]×[B] ＋ beta･[C]



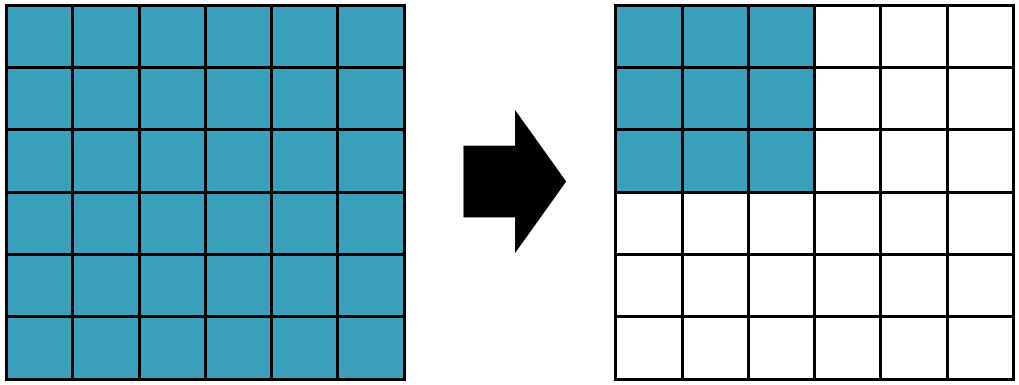
固有値求解部のチューニング

EigenSolveの通信がボトルネック ← 行列サイズが小さすぎる



ML:格子数
MB:バンド数
MB＜＜ML

一部のプロセスで固有値求解部を実行する



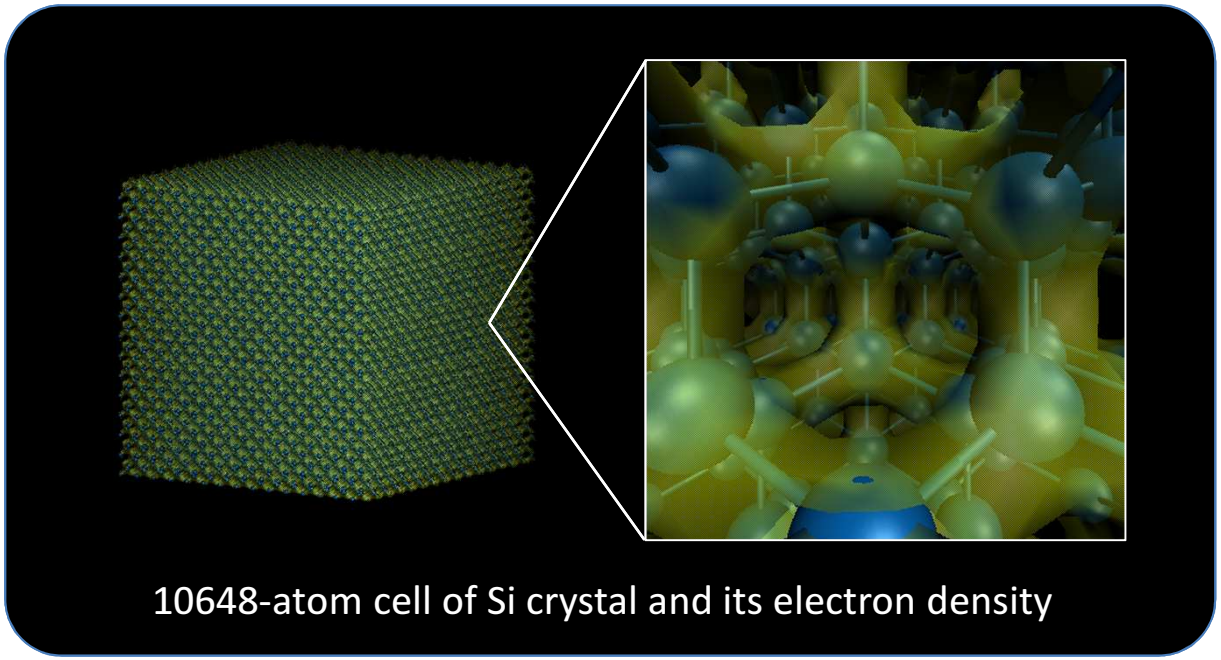
水色部: アクティブ・プロセス

Outline

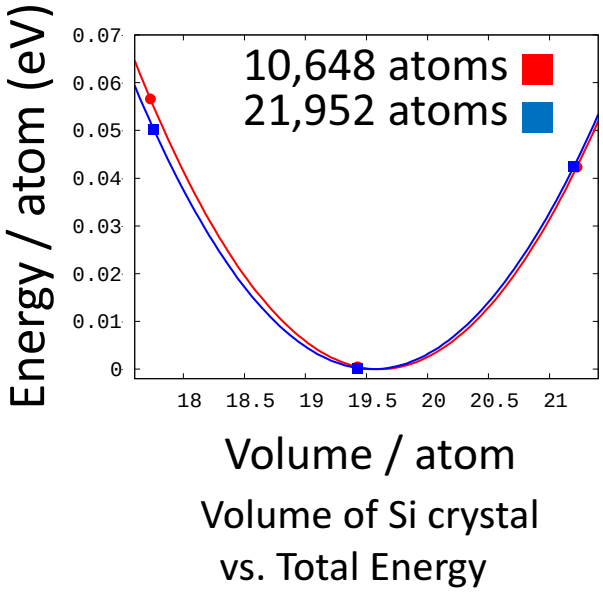
- ターゲット
- RSDFTの概要
- 京の概要
- 京向けのチューニング
- 性能ベンチマーク
- **科学的成果**

プログラムの機能検証

We have successfully obtained the self-consistent electron density and total energy for extremely large systems. Structural stability can also be described precisely.



10648-atom cell of Si crystal and its electron density



Volume of Si crystal vs. Total Energy

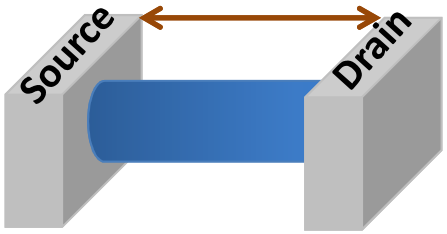
LDA = Local Density Approximation (exchange-correlation functional)

	LDA (21,952 atoms)	LDA (10,648 atoms)	LDA (2 atoms)	Expt.
Lattice constant	5.39Å	5.39Å	5.39Å	5.43Å

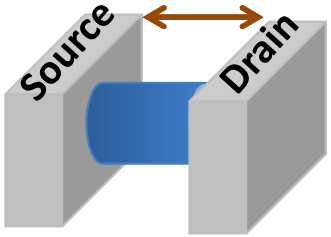
DOS of SiNW near CBM

DOS : Density of Sates
CBM: Conduction Band Minimum

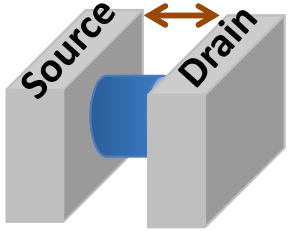
10 nm (39,696 atoms)



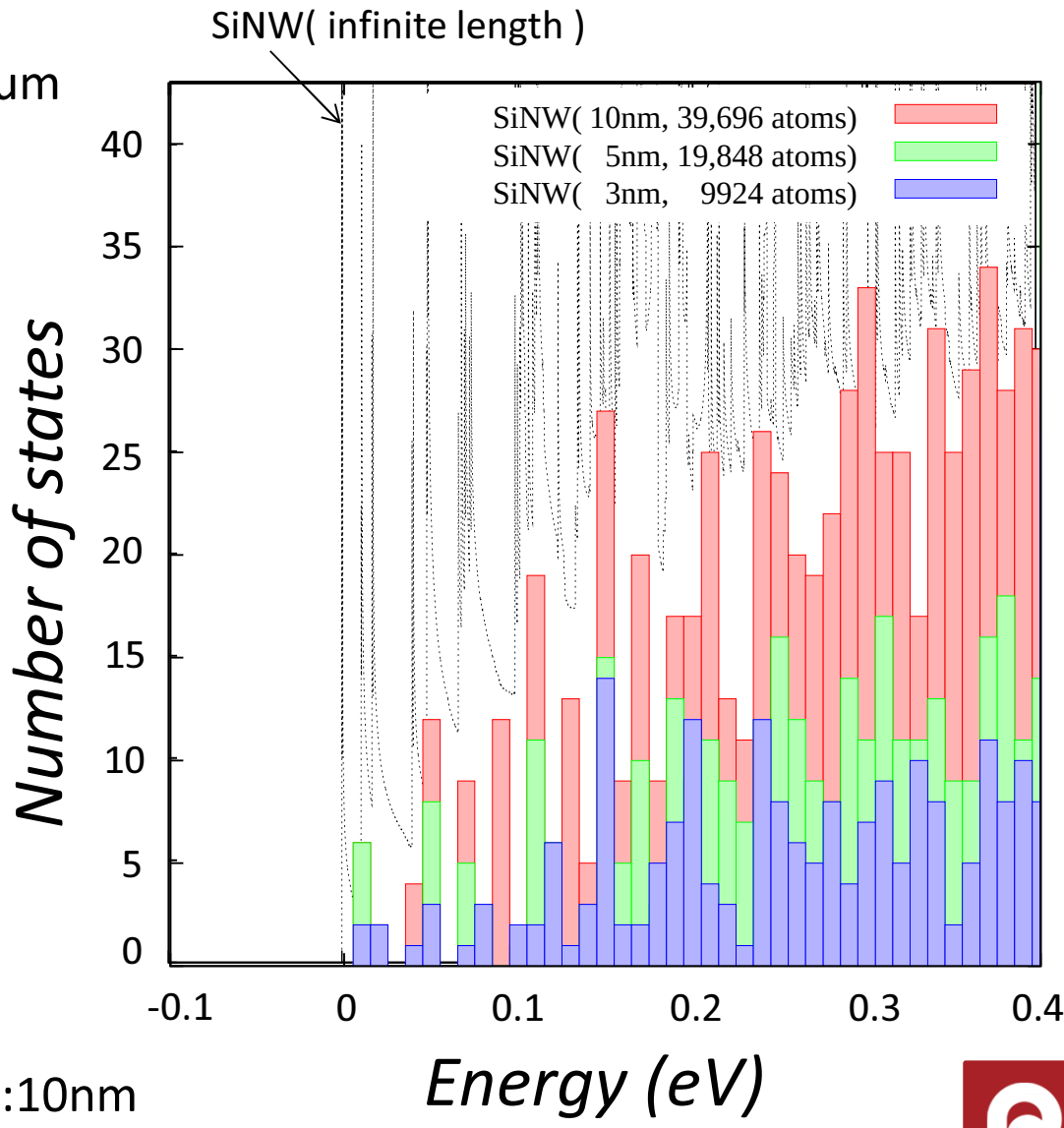
5 nm (19,848 atoms)



3 nm (9924 atoms)



Diameter:10nm



まとめ

- RSDFTの超並列対応には通信部のチューニングが必要不可欠
  - 並列軸を拡張
  - Tofuネットワークへの最適化

- 性能ベンチマークを実施
  - シリコンナノワイヤ10万原子
    - 「京」の性能を限界まで引き出すことに成功、実行効率59.02%
    - 10万原子の計算が7P構成を用いれば5日程度で可能

- プログラムの正しさを検証
  - 結果の妥当性を確認（実験値、理論値との比較）

- 科学的成果の創出
  - シリコンナノワイヤ1万原子～4万原子の計算を実施
  - **実サイズのシリコンナノワイヤ**をシミュレート

ご清聴ありがとうございました