スーパーコンピュータ「京」と
アプリケーションの概要

2012年3月9日

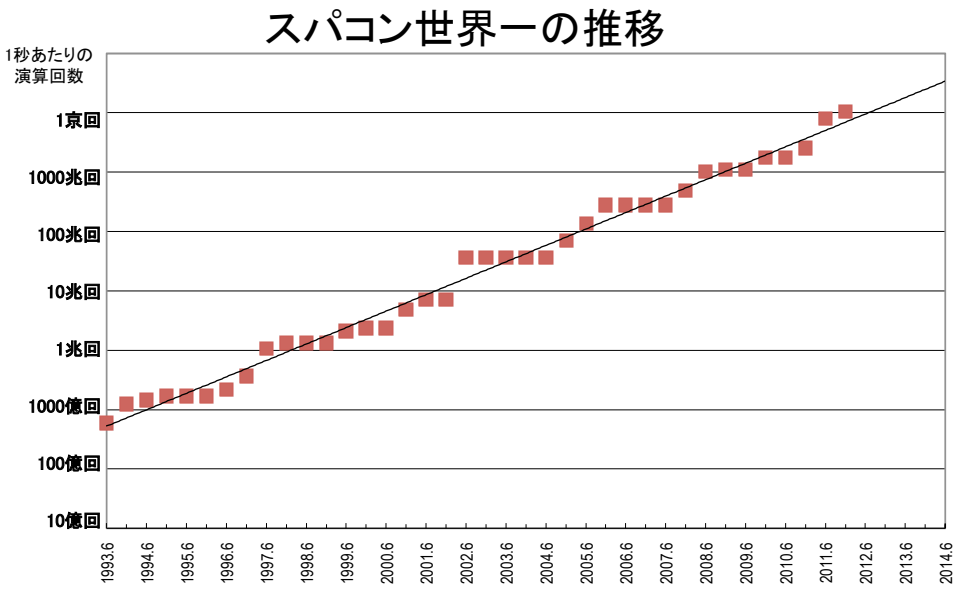
独立行政法人　理化学研究所
次世代スーパーコンピュータ開発実施本部　開発グループ
アプリケーション開発チーム

南　一生
minami_kaz@riken.jp

RIKEN Advanced Institute for Computational Science

プロジェクト概要

スパコン世界一をめぐる競争



平均すると1年に約1.9倍の性能向上！

現在世界最速のスパコンは、世界初のスパコンCRAY-1（1976年）の約7000万倍！

ちなみに

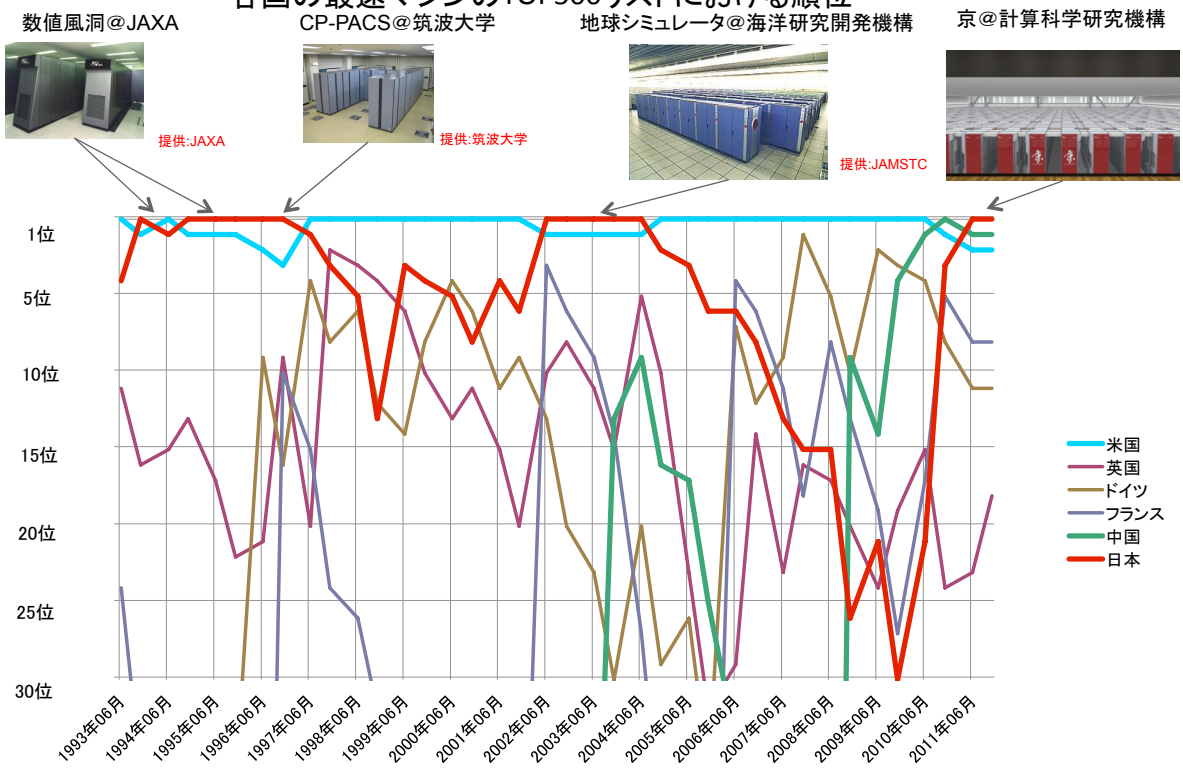
世界初のスパコンCRAY-1の演算性能は、約160メガフロップス
一方、iPhone4Sの演算性能は、約140メガフロップス

1976年　　2011年

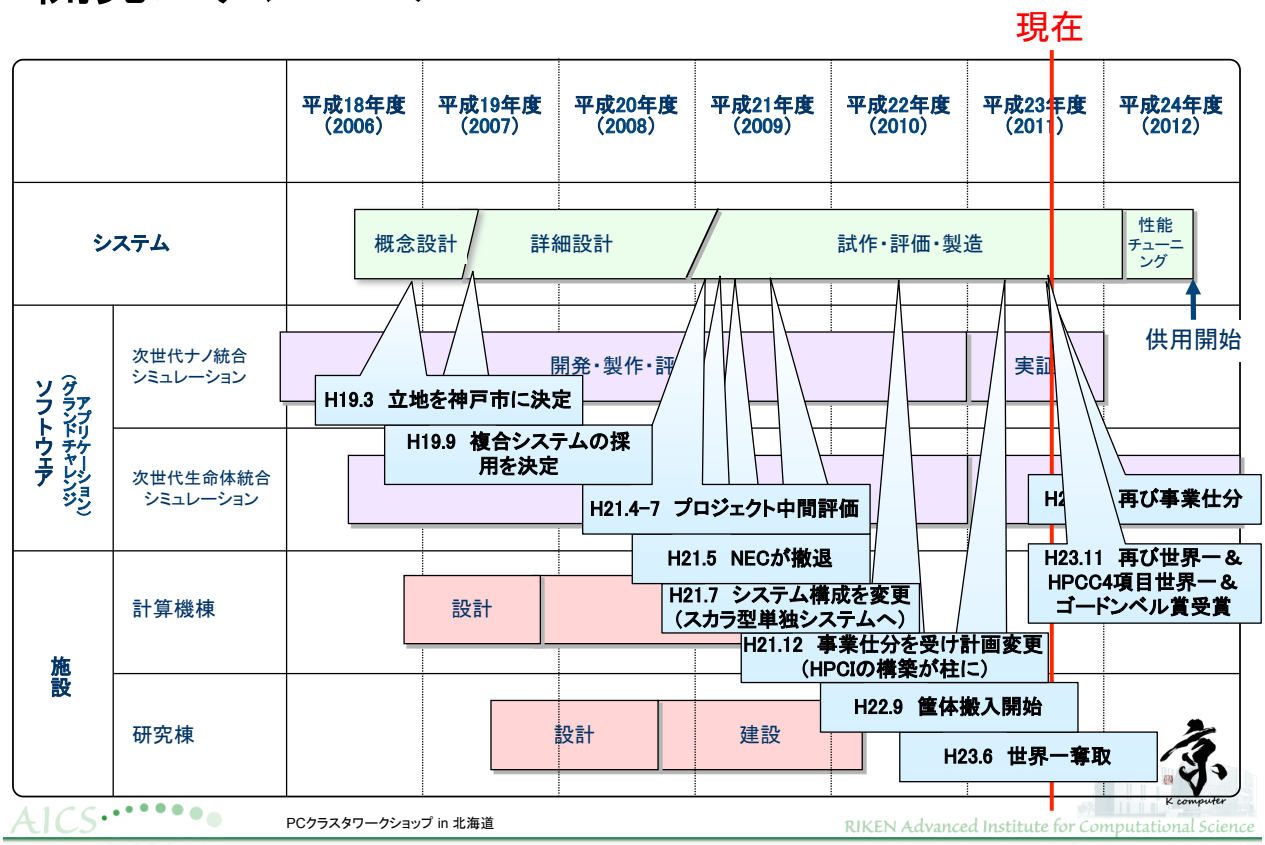

≒


これまでの日本のスパコンプロジェクト

各国の最速マシンのTOP500リストにおける順位

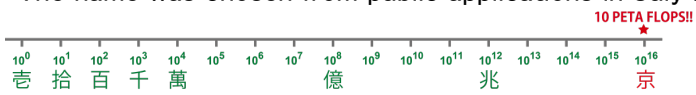


開発スケジュール（平成18年度－平成24年度）



K computer, “京”, is …

- ✓ “京 [kei]” is a nickname of the next-generation supercomputer system
  - ✓ The name was chosen from public applications in July 2010.



  - ✓ “京” stands for a unit corresponding to 10 Peta, which is also the performance target of our project.
  - ✓ The name for non-Japanese people is

“K computer”

~~K Computer~~
~~K-Computer~~
~~Kei comuter~~
……

  - ✓ Another meaning of the “京” is “a big gate.”
    - ✓ A new era of computational science is coming though the gate “京.”

「京」の最近の成果(1/3)

(※)TOP500リストとは？

LINPACKベンチマークの実行性能を指標とした、世界のスパコン上位500位までのランキング。年に2回、6月と11月に更新される

(※)LINPACK(リンパック)とは？

大規模連立一次方程式を解くベンチマークプログラムで、世界のスパコンランキングであるTOP500でスパコンの性能指標として利用されている

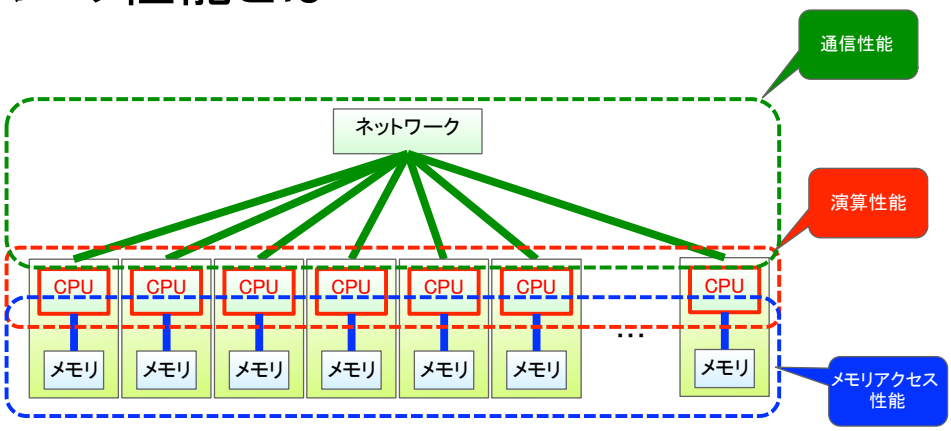
- ✓ 2011/11/15発表の第38回TOP500リストで再び第一位を獲得.

	K computer(Oct 2011)	K computer(Jun 2011)
R_{max}	10.51PFLOPS	8.162PFLOPS
R_{peak}	11.28PFLOPS	8.774PFLOPS
efficiency	93.2%	93.0%
N_{max}	11,870,208	10,725,120
Execution time	29h28m	27h59m

ランキング	国	システム名	演算性能 (ペタフロップス)	実行効率 (%)	1ワットあたりの演算性能	実行時間 (時間)
1	Japan	K computer	10.510	93	824.56	29.47
2	China	Tianhe-1A	2.566	55	635.15	3.37
3	US	Jaguar	1.759	75	253.07	17.27
4	China	Nebulae	1.271	43	492.64	1.91
5	Japan	TSUBAME2.0	1.192	52	852.27	2.40

演算性能だけでなく、高効率、低消費電力、高信頼性を同時に実証

スパコンの性能とは？



演算性能を上げるのは、比較的容易
メモリアクセス性能と通信性能を上げるのは困難

TOP500の指標となるLINPACKベンチマークは
演算性能のみで決まる
（メモリアクセス性能、通信性能はほとんど考慮されない）

「京」の最近の成果（2/3）

HPCチャレンジアウォード：科学技術計算で多用される計算パターンから抽出した28項目の処理性能によって、スパコンの総合的な性能を評価するHPCチャレンジベンチマークプログラムから、特に重要な4つのベンチマークをHPCチャレンジアウォードとして、毎年11月のSCにて表彰

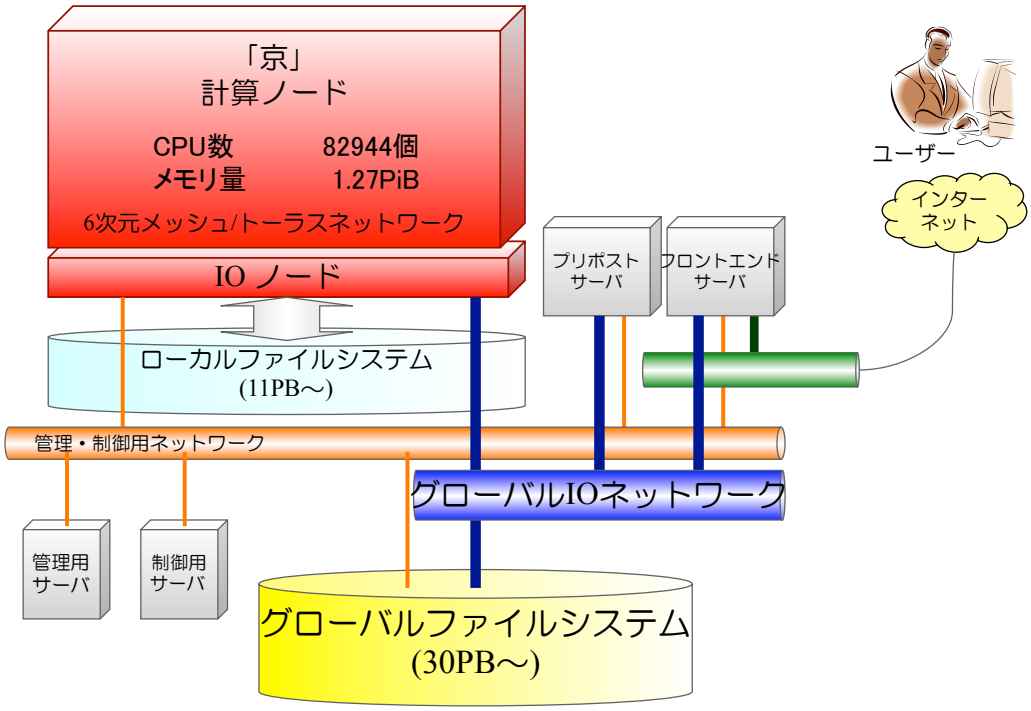
1. Global HPL（大規模な連立1次方程式の求解における演算速度） → 演算性能
2. Global RandomAccess（並列プロセス間でのランダムメモリアクセス性能） → 通信性能
3. EP STREAM(Triad) per system（多重負荷時のメモリアクセス速度） → メモリアクセス性能
4. Global FFT（高速フーリエ変換の総合性能） → 演算性能 通信性能

2011/11/16発表の2011年HPCチャレンジアウォードの4部門すべてで第一位を獲得

Global HPL	Performance (TFLOP/s)	System	Institutional Facility
1st place	2,118	K computer	RIKEN
1st runner up	1,533	Cray XT5	ORNL
2nd runner up	736	Cray XT5	UTK
Global RandomAccess	Performance (GUPS)	System	Institutional Facility
1st place	121	K computer	RIKEN
1st runner up	117	IBM BG/P	LLNL
2nd runner up	103	IBM BG/P	ANL
EP STREAM (Triad) per system	Performance (TB/s)	System	Institutional Facility
1st place	812	K computer	RIKEN
1st runner up	398	Cray XT5	ORNL
2nd runner up	267	IBM BG/P	LLNL
Global FFT	Performance (TFLOP/s)	System	Institutional Facility
1st place	34.7	K computer	RIKEN
1st runner up	11.9	NEC SX-9	JAMSTEC
2nd runner up	10.7	Cray XT5	ORNL

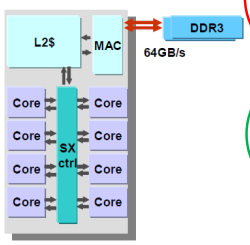
システム概要

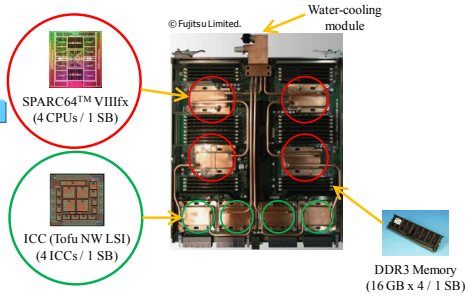
「京」の概要


「京」
計算ノード
CPU数 82944個
メモリ量 1.27PiB
6次元メッシュ/トーラスネットワーク
IO ノード
ローカルファイルシステム
(11PB～)
プリポスト
サーバ
フロントエンド
サーバ
ユーザー
インター
ネット
管理・制御用ネットワーク
グローバルIOネットワーク
管理用
サーバ
制御用
サーバ
グローバルファイルシステム
(30PB～)

CPUの詳細

	諸元
演算性能（ピーク）	128 GFLOPS (16 GFLOPS x 8 cores)
コア数	8
クロック周波数	2.0 GHz
浮動小数点演算器	乗加算ユニット x 4 (2 SIMD) 除算器 x 2
レジスタ数	浮動小数点レジスタ (64bit) : 256 汎用レジスタ (64bit) : 188
キャッシュ	L1I$: 32 KB (2way) L1D$: 32 KB (2way) L2$: Shared 6 MB (12way)
メモリ帯域	64 GB/s (0.5B/F)





45nm CMOS process
チップサイズ: 22.7mm x 22.6mm
トランジスタ数 : 760M
Power : 58W (TYP, 駆動温度30℃), 水冷

ex.
Sandy Bridge
YMM reg.(256bit) x 16 = 64要素

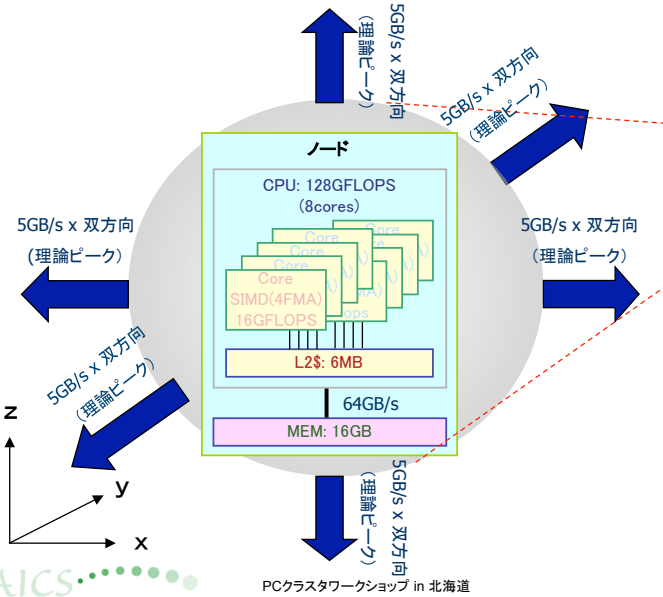
他のチップとの比較

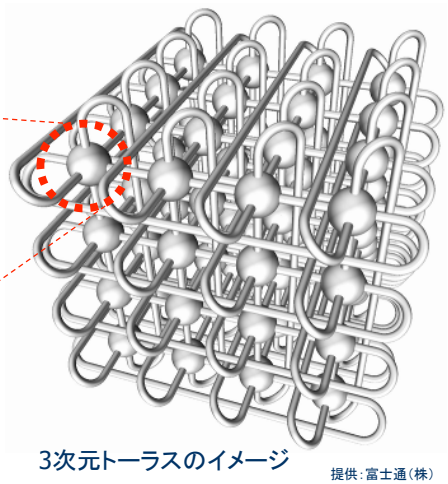
Vendor	Name	Core	Process rule (nm)	Peak performance (GFLOPS)	Cache (MB)	Power (W)	GF/W	System (w/planned)
IBM	PowerPC A2	16	45	204.80	32	55	3.72	Sequoia (BlueGene/Q)
Intel	E3-1260L	4	32	105.60	8	45	2.35	
Fujitsu	SPARC64VIIIfx	8	45	128.00	6	58	2.21	K computer
IBM	Power7	8	45	256.00	32	200	1.28	
AMD	Opteron 6172	12	45	100.80	12	80	1.26	XE6,etc.
Intel	Xeon X5670	6	32	79.92	12	95	0.84	TSUBAME2.0,etc.

高性能かつ低消費電力

計算ノードとインターコネクトの構成

- 計算ノードの構成
  - CPU（8コア） : 1個
  - ICC（インターコネクト用LSI） : 1個
  - メモリ : 16GB
- インターコネクトの構成
  - ユーザービューは3次元トーラス
  - 帯域：3次元の正負各方向にそれぞれ 5GB/s x 2（双方向）【理論ピーク】
  - ケーブル：約200,000本，約1000km

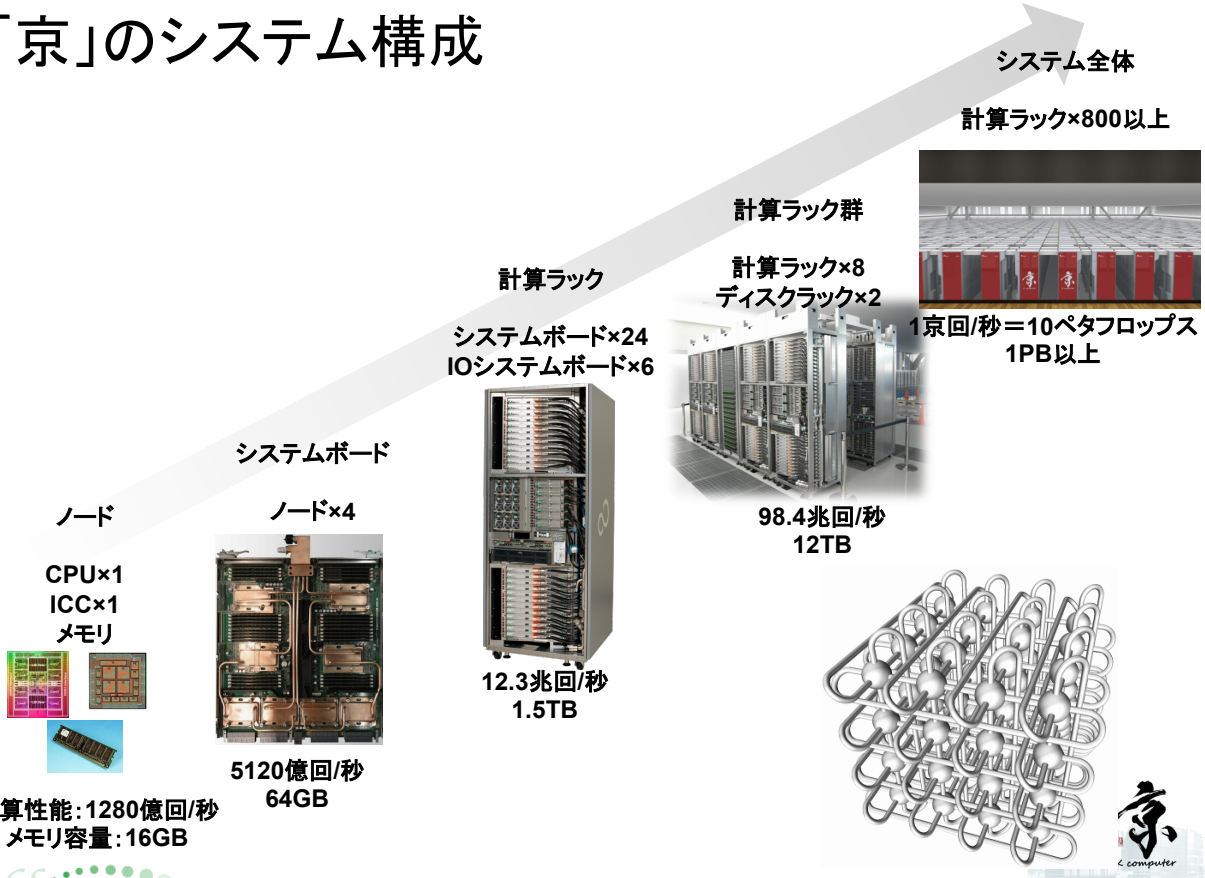




3次元トーラスのイメージ

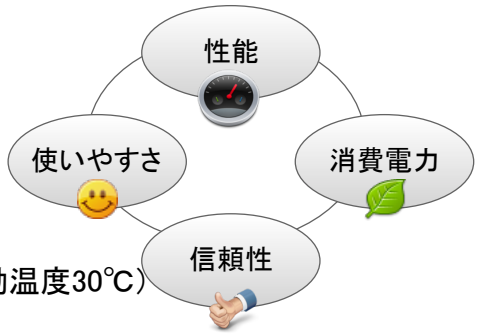
提供：富士通（株）

「京」のシステム構成



システムの特長

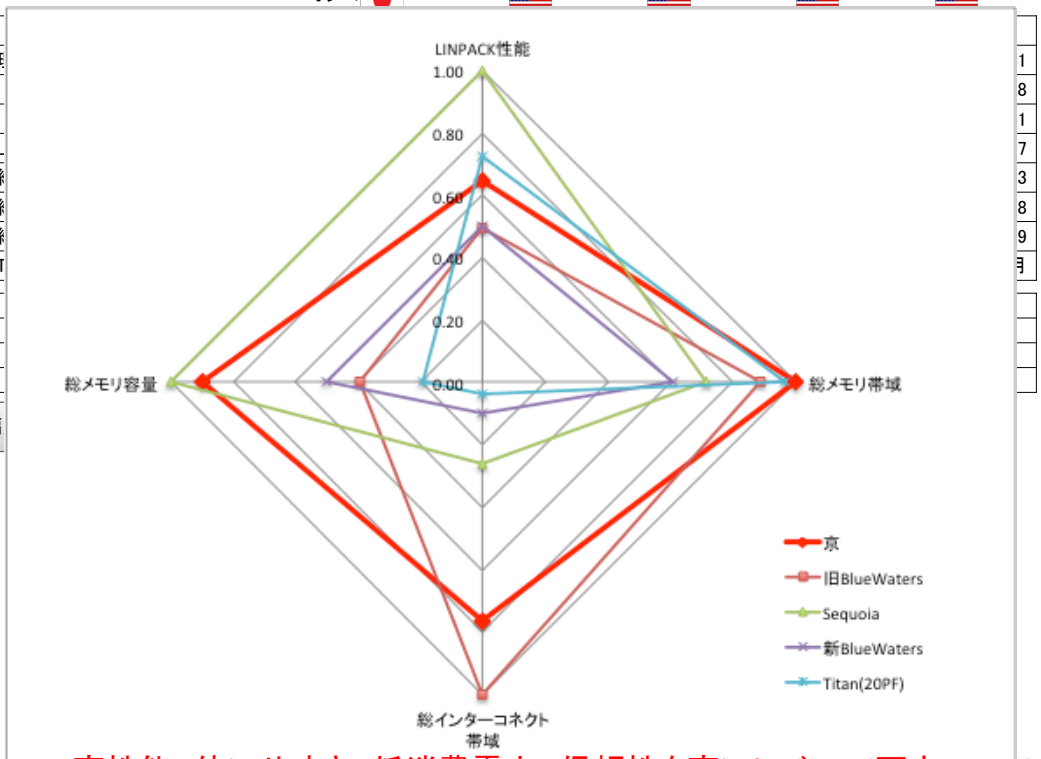
- ✓ 世界トップクラスの演算性能と汎用性（使いやすさ）の両立
  - ✓ LINPACK(※) 10ペタフロップス（1秒間に1京回）
  - ✓ ペタフロップス級のアプリケーション実効性能
  - ✓ 広範囲のアプリケーションに対応可能
    - ✓ 高帯域のメモリアクセスとインターコネクト



- ✓ 高性能と低消費電力の両立
  - ✓ CPU：128ギガフロップス，58W（LINPACK時、駆動温度30℃）
  - ✓ システム全体：824.56メガフロップス/ワット
    - ✓ 省エネスパコンランキング（GREEN500）2011年6月版で第6位
    - ✓ 大規模システムとしてはトップクラス

- ✓ 高い信頼性の確保
  - ✓ 「壊れない」，「壊れても全てが止まらない」．「壊れた部分はすぐ直せる」
  - ✓ ネットワークの高信頼性化： 自動代替経路，自動再構成機能
  - ✓ 約30時間の高負荷連続運転（LINPACK計測）を実証

ライバルとの比較



高性能、使いやすさ、低消費電力、信頼性を高いレベルで両立

計算科学研究機構

敷地面積：20，000m^2

甲子園球場のグラウンドの約1．5倍

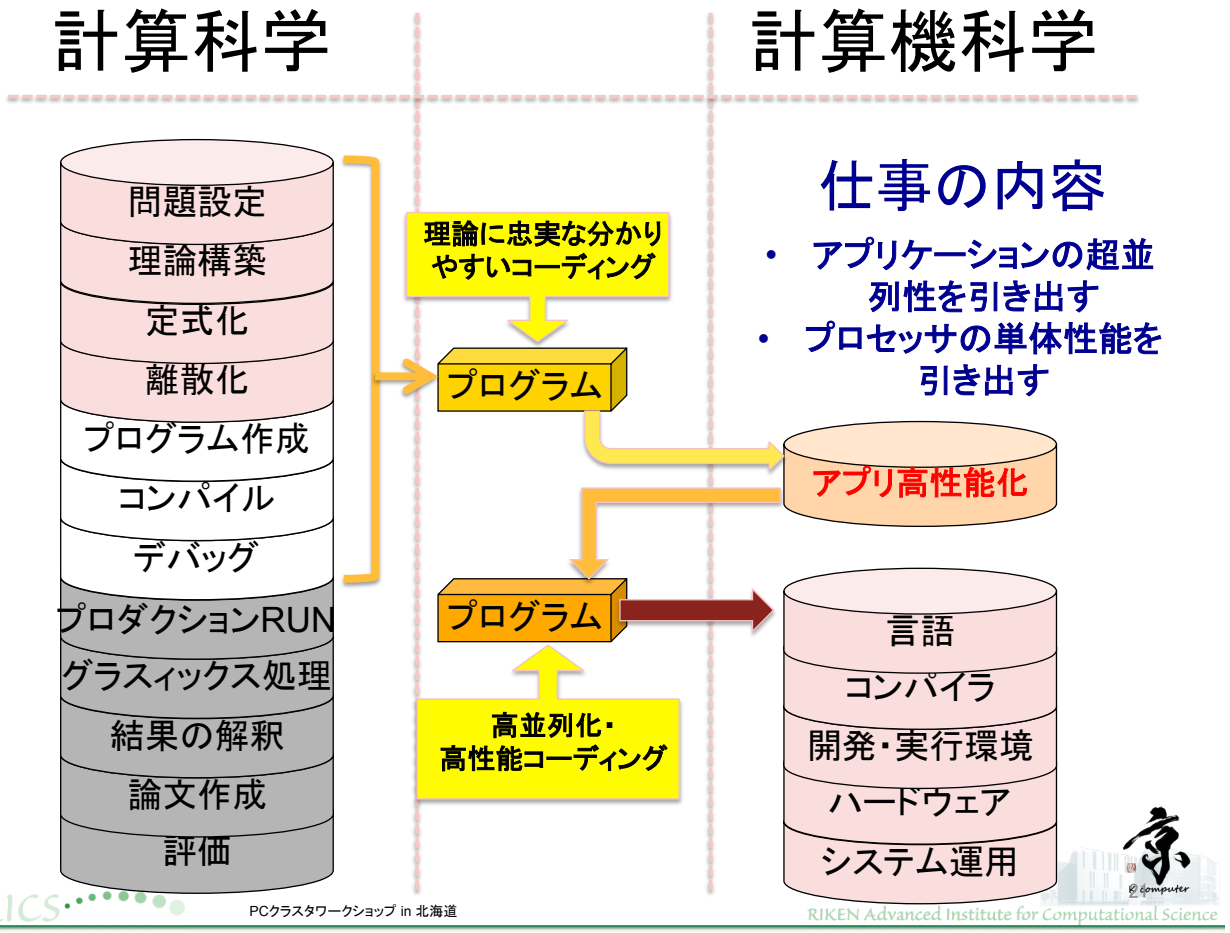
「京」の搬入・設置（ムービー）

理研で実施している「京」のアプリケーション性能最適化

目的

✓次世代スパコンの運用に先立ち、システム性能を実証する

- 並列性能が見込めるコード複数本を対象
- 次世代スパコンの汎用性を踏まえ分野バランスを考慮
- 次世代スパコンの汎用性を踏まえ計算特性のバランスを考慮
  - 並列化手法が比較的理解しやすい・かなり複雑
  - アプリケーションの要求B/F値が高い・低い



重点化アプリケーション

プログラム名	分野	アプリケーション概要	期待される成果	手法
NICAM	地球科学	全球雲解像大気大循環モデル	大気大循環のエンジンとなる熱帯積雲対流活動を精緻に表現することでシミュレーションを飛躍的に進化させ，現時点では再現が難しい大気現象の解明が可能となる.	FDM (大気)
Seism3D	地球科学	地震波伝播・強震動シミュレーション	既存の計算機では不可能な短い周期の地震波動の解析・予測が可能となり，木造建築およびコンクリート構造物の耐震評価などに応用できる.	FDM (波動)
PHASE	ナノ	平面波展開第一原理分子動力学解析	第一原理計算により，ポスト35nm世代ナノデバイス，非シリコン系デバイスの探索を行う.	平面波 DFT
FrontFlow/Blue	工学	Large Eddy Simulation (LES)に基づく非定常流体解析	LES解析により，エンジニアリング上重要な乱流境界層の挙動予測を含めた高精度な流れの予測が実現できる.	FEM (流体)
RSDFT	ナノ	実空間第一原理分子動力学計算	大規模第一原理計算により，10nm以下の基本ナノ素子(量子細線，分子，電極，ゲート，基盤など)の特性解析およびデバイス開発を行う.	実空間 DFT
LatticeQCD	物理	格子QCDシミュレーションによる素粒子・原子核研究	モンテカルロ法およびCG法により，物質と宇宙の起源を解明する.	QCD

重点化アプリケーションの特徴

ES：地球シミュレータ1

プログラム名	分野	アプリケーション概要	コードの計算科学的な特性	手法
NICAM	地球科学	全球雲解像大気大循環モデル	ESではピーク性能比40%程だがプログラミング上は**B/F性能を要求する**ように見える。次世代スパコンではキャッシュの有効利用を始め高速演算機構の活用が必須。	FDM (大気)
Seism3D	地球科学	地震波伝播・強震動シミュレーション	ESではピーク比40%程だがプログラミング上は**B/F性能を要求する**ように見える。次世代スパコンではキャッシュの有効利用を始め高速演算機構の活用が必須。	FDM (波動)
PHASE	ナノ	平面波展開第一原理分子動力学解析	**単体性能向上は主要処理の行列・行列積化により可能**との予測が立っている。しかし原子数を相当増やさないと**超高並列は難**のため高並列化の検討要。	平面波 DFT
FrontFlow/Blue	工学	Large Eddy Simulation (LES)に基づく非定常流体解析	ESではピーク比25%程だがプログラミング上は**B/F性能を要求するように見え、またリストアクセス**のため次世代スパコンではキャッシュの有効利用を始め高速演算機構の活用、データアクセスの効率化が必須。	FEM (流体)
RSDFT	ナノ	実空間第一原理分子動力学計算	**単体性能は主要処理の行列・行列積化により可能**との予測が立っている。しかしメッシュ分割数を相当増やさないと**超高並列は難**のため高並列化の検討要。	実空間 DFT
LatticeQCD	物理	格子QCDシミュレーションによる素粒子・原子核研究	通信トポロジーを意識した**高度な並列化チューニング**が必要。最内ループのリカレンスのため高い**単体性能が得にくい**。	QCD

RSDFTコード –計算手順–

1. 波動関数Φの初期値を与える
2. CG法($O(N^2)$)により、波動関数Φを更新する
3. 波動関数Φを規格直交化（グラムーシュミット($O(N^3)$)）する
4. 局所ポテンシャルV_{LOC}を更新する。更新前後で変化が無ければ計算終了
5. ハミルトニアンを更新
6. 部分対角化($O(N^3)$)（MB×MB空間で）を行い、1.に戻る

空間＋バンド並列の実装

最適マッピング

バンド並列の実装

- 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

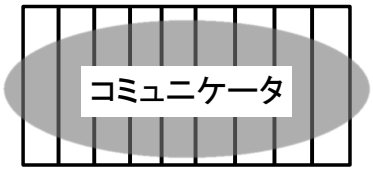
Kohn–Sham 方程式

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

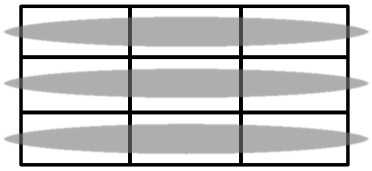
バンド間の依存性なし

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



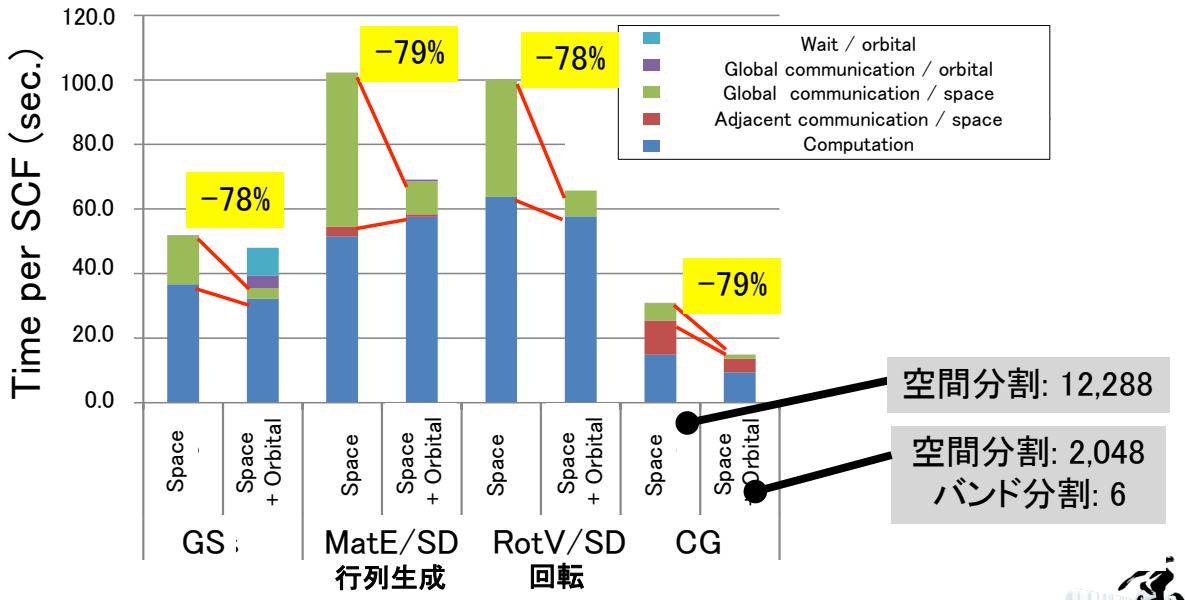
並列軸=2



通信対象が減る
分割粒度が大きくなる

バンド並列の効果

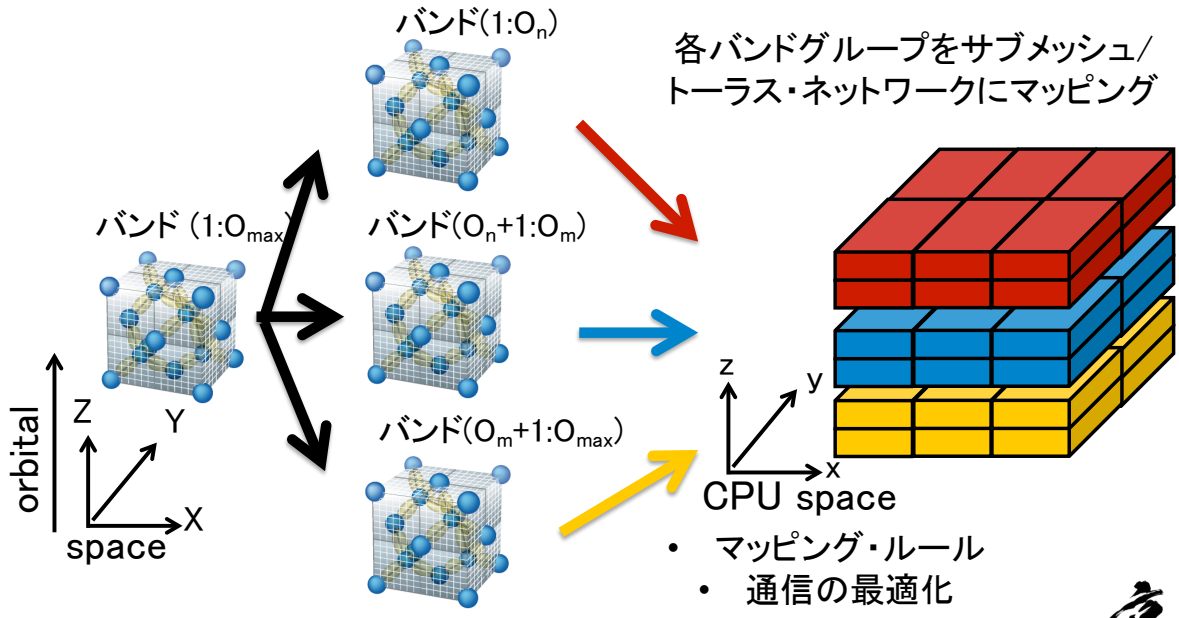
SiNW, 19,848 原子, 格子数:320x320x120, バンド数:41,472
トータル並列プロセス数は12,288で固定



大域通信時間を大幅に削減

最適マッピング -Tofuネットワーク-

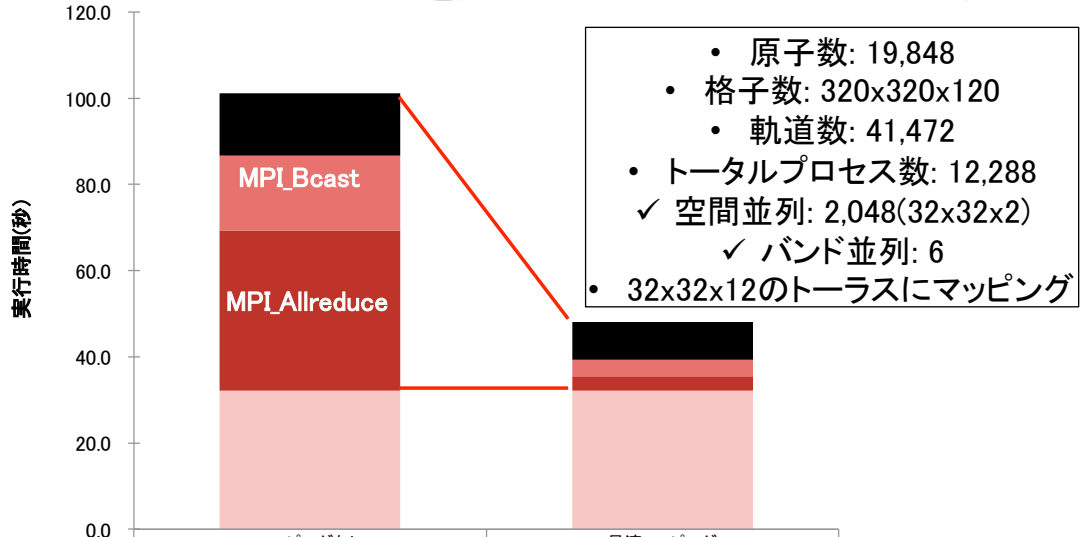
空間並列 | 空間並列＋バンド並列 | Tofuネットワークへのマッピング



サブメッシュ/トーラス内で通信が閉じられる

最適マッピングの効果- Gram-Schmidt -

最適マッピング → サブコミュニケータ間のコンフリクトが発生しない
MPI通信でTofu向けアルゴリズムが選択される

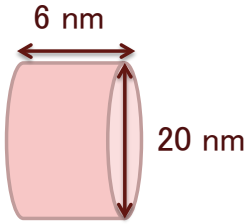


	マッピングなし	最適マッピング
■通信 同期/軌道	14.473	8.658
■通信 大域/軌道	17.364	3.827
■通信 大域/空間	37.246	3.336
■演算	32.095	32.191

トータル性能評価

測定条件

- SCF計算の1反復を測定(4万原子までは収束計算実施完了)
- 107,292 原子のSiNW
  - ✓ 格子数: 576x576x192
  - ✓ バンド数: 229,824

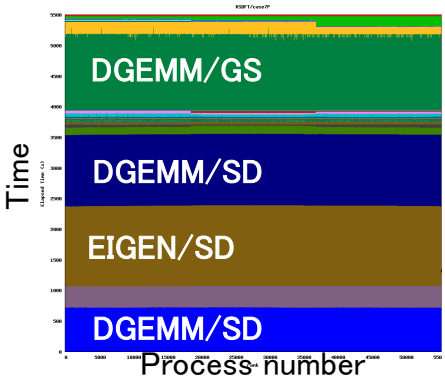


- 並列プロセス数:55,296
  - ✓ 空間分割:18,432 x バンド分割: 3
- 使用ノードのピーク性能: 7.07PFLOPS
  - ✓ 55,296 ノード(442,368 コア)

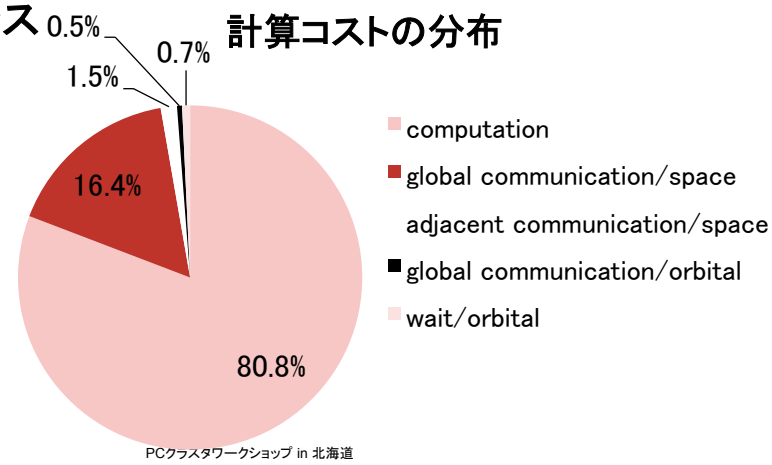
トータル性能評価

- 実効性能は 3.08 PFLOPS .
- ピーク性能に対する実行効率は43.6 % .
- 全実行時間に対する通信時間の割合は 19.0% .
- SCF計算の反復1回の実行時間は 5,500 秒(1.5 時間).

プロセス間のロードバランス

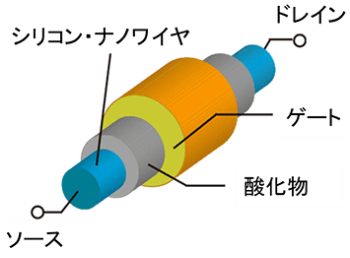


計算コストの分布



SC11にてゴードンベル賞受賞

- ✓ 2011/11/18発表の2011年ゴードン・ベル賞の最高性能賞(Winner)を受賞
  - ✓ ゴードン・ベル賞:アプリケーションの実性能と計算科学の成果に対してアメリカ計算機学会が授与する賞、毎年11月のSCにて表彰
  - ✓ 受賞論文:「「京」による100,000原子シリコン・ナノワイヤの電子状態の第一原理計算」
  - ✓ ファイナリスト5件すべてが受賞(Special Achievement:1、Honorable Mention:3)



おわりに

今後とも「京」に対するご理解とご協力をお願い致します