東大FSアプリWG
進捗報告

片桐 孝洋，大島 聡史，中島 研吾（東大）
米村 崇，熊洞 宏樹，樋口 清隆，橋本 昌人，高山 恒一（日立情報・通信システム社），
藤堂 眞治，岩田 潤一，内田 和之，佐藤正樹，羽角博康（東大），
黒木聖夫（海洋研究開発機構）

代理発表： 西澤 誠也（理研）

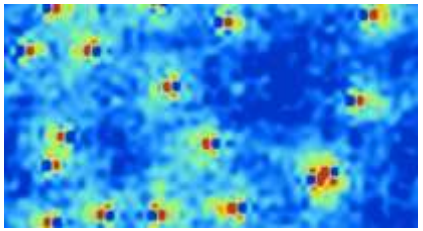
【将来HPCI調査研究「アプリ分野」第8回全体ミーティング】
日時：8月6日（火） システム設計チーム報告（進行：杉田） 14:10－15:10
場所：TKP大手町ビジネスセンター 7階ホール7A

Feasibility Study on
Advanced and Efficient Latency Core-based Architecture for Future HPCI R&D

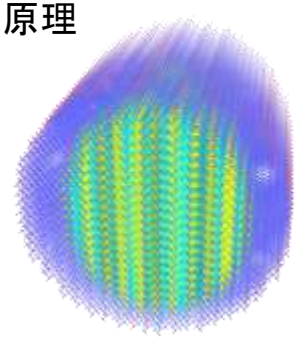
スケジュール確認

ターゲットアプリケーション群

- **ALPS/looper**
  - 新機能を持った強相関・磁性材料の物性予測・解明。虚時間経路積分にもとづく量子モンテカルロ法と厳密対角化
  - 総メモリ：10～100PB
  - 整数演算、低レイテンシ、高次元のネットワーク
  - 利用シナリオ：1ジョブ当たり24時間、生成ファイル：10GB.
    同時実行1000ジョブ、合計生成ファイル：10TB.

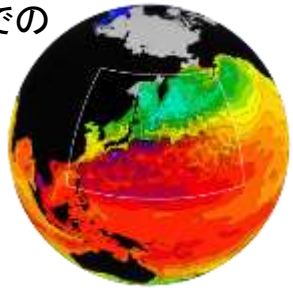


- **RSDFT**
  - Siナノワイヤ等、次世代デバイスの根幹材料の量子力学的第一原理シミュレーション。実空間差分法
  - 総メモリ：1PB
  - 演算性能：1EFLOPS（B/F =0.1以上）
  - 利用シナリオ：1ジョブ当たり10時間、生成ファイル：500TB.
    同時実行10ジョブ、合計生成ファイル5 PB.



- **NICAM**
  - 長期天気予報の実現、温暖化時の台風・豪雨等の予測
  - 正20面体分割格子非静力学大気モデル。水平格子数kmで全球を覆い、積雲群の挙動までを直接シミュレーション
  - 総メモリ: 1PB、メモリ帯域: 300 PB/sec
  - 演算性能: 100 PFLOPS （B/F = 3）
  - 利用シナリオ：1ジョブ当たり240時間、生成ファイル：8PB.
    同時実行10ジョブ、合計生成ファイル：80 PB.

- **COCO**
  - 海況変動予測、水産環境予測
  - 外洋から沿岸域までの海洋現象を高精度に再現し、気候変動下での海洋変動を詳細にシミュレーション
  - 総メモリ: 320 TB、メモリ帯域: 150 PB/sec
  - 演算性能: 50 PFLOPS（B/F = 3）
  - 利用シナリオ：1ジョブ当たり720時間、生成ファイル：10TB.
    同時実行100ジョブ、合計生成ファイル：1 PB.



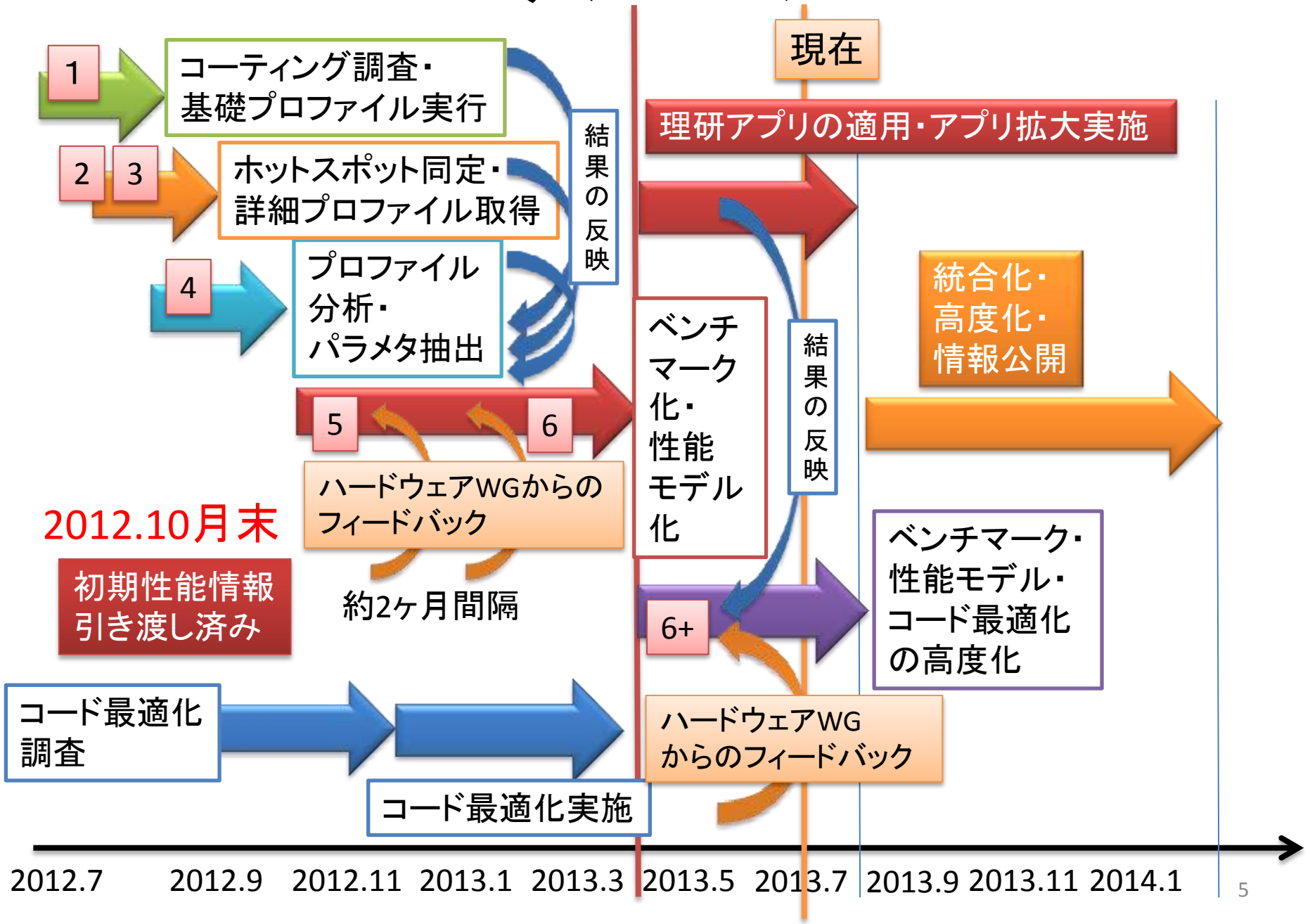
利用シナリオ
アンサンブル型
全系の1/10～1/100資源を用いた1ジョブを、複数同時実行することで、全資源を使い切る形態。

要求性能は「計算科学ロードマップ白書」(2012年3月)の見積値からの抜粋、および、開発者による新見積値である

性能モデル化手法

1. **ホットスポット同定：富士通社の**基本プロファイラで複数のホットスポット（ループレベル）を同定、全体性能の予測をホットスポットのみで行う
   - ホットスポットの部品化
   - 数理レベル（支配方程式、離散化方法）の処理ブロックとの対応を検討
2. **カーネル分離：**（目視により）計算部分、通信部分、I/O部分の分離
   - 計算部分：演算カーネル
   - 通信部分：通信カーネル
   - I/O部分：I/Oカーネル
3. **通信パターン確認**
4. **詳細プロファイルと分析：　富士通社の**詳細プロファイラを用い、ホットスポットごとにハードウェア性能情報（＝性能パラメタ）を取得し分析
   - 演算カーネルの　演算効率／命令発行量／キャッシュ利用効率　など
   - 通信カーネルの　通信回数／量／通信待ち時間　など
   - I/Oカーネルの　データ読み書き　量／頻度　など
5. **ベンチマーク化：**ホットスポットのみで動作するようにコードを再構成
   - マシン特化の書き方、および、汎用的な書き方、の2種を区別
   - 演算カーネル、通信カーネル、I/Oカーネルの分類
6. **詳細モデル化と予測：**ハードウェア因子を用いた数式による実行時間を近似。
   - 富士通社の性能予測ツールにより、概念設計マシンの実行時間を予測

スケジュール


1
コーティング調査・
基礎プロファイル実行
2
3
ホットスポット同定・
詳細プロファイル取得
4
プロファイル
分析・
パラメタ抽出
結果の反映
現在
理研アプリの適用・アプリ拡大実施
統合化・
高度化・
情報公開
5
6
ハードウェアWGからの
フィードバック
ベンチ
マーク
化・
性能
モデル
化
結果の反映
2012.10月末
初期性能情報
引き渡し済み
約2ヶ月間隔
6+
ベンチマーク・
性能モデル・
コード最適化
の高度化
コード最適化
調査
ハードウェアWG
からのフィードバック
コード最適化実施
2012.7
2012.9
2012.11
2013.1
2013.3
2013.5
2013.7
2013.9
2013.11
2014.1

今後の予定

H25年度　理研FS選定ミニアプリ評価

- QCD
  - FX10で評価を開始済み
- Modylas
  - 東大、日立、富士通、九州大、ごとに個別にライセンス契約を結ぶ
  - 日立と東大は契約済み。理研から、コード受け渡し済み。
- Front Flow / Blue
  - 理研運営グループと連携。コード引き渡し済み。
- NGS Analyzer（予定）
  - ゲノム系、理研FSでミニアプリ準備中。（8月中引き渡しを目標）

H25年度予定（1／3）

- H24年度残務作業
  - –RSDFTの通信時間解析と性能予測
    - 性能予測中（9月中に初期評価終了を目標）
- カーネルベンチの公開
  - – NICAMとCOCOについて、カーネルベンチを開発済み。
  - – 入力データによらず動くように、ベンチマークを再構築中。

H25年度予定（2／3）

- 超並列向きアルゴリズム採用と性能評価
  - RSDFTの直交化処理
    - 通信回数が少ない新アルゴリズムを適用
    - CAQR (Communication Avoidance QR)
    - 米澤CREST採択課題の筑波大櫻井グループと連携
      （開発中のコードを利用させてもらう）
      - CAQRプログラムをRSDFTに組み込み
        性能評価を予定

H25年度予定（3／3）

- コード最適化
  1. 新規4アプリのコードチューニング
     を予定
  2. 複数の異機種環境での評価
     - FX10上の性能が妥当であるか検証するため
     - Intel、AMD、Power7での性能評価
     - Intel MIC (Xeon Phi)での性能評価
     - 上記のコードチューニングも、一部実施予定

新規コードの評価状況

- QCD

QCDの性能評価： アプリケーションの概要

QCDのプログラム構成とFS向けターゲット性能

(1) 問題サイズ（ FS target ）
* Class 4 : 160x160x160x160 (default MPI config: 20x20x20)
* Class 5 : 256x256x256x256 (default MPI config: 32x32x32)

(2)評価対象はcloverとBiCGStab

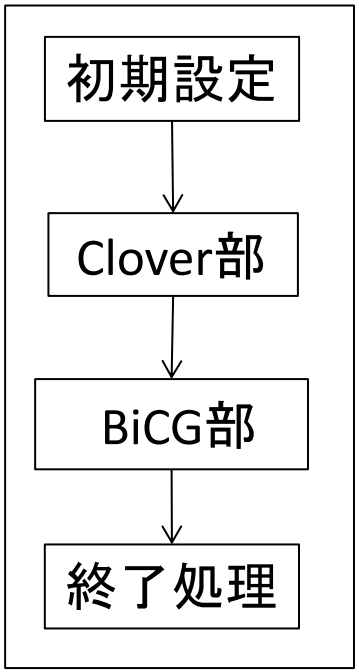
(3)要求性能

- BiCGStabの実行時間が3.1[ms]/ステップ以下
- clover部(clover, clover_inv)の実効性能（Flops値）がBiCGstabの20%以上

まず，入手したミニアプリに対して以下を実施

- cloverの一部に OpenMP directive を追加
- default MPI configでは最外側ループ長が8以下になるケースが多く、メニィコアプロセッサ向けSMP化

上記SMP対策を施してからプロファイルを取得



QCDの性能評価： Clover部のSMP化(1)

(1)clover関数のシリアル部へのOMP並列化の適用

【変更前】clover.h255行目

```
do ix=1,NX
do iy=1,NY
do iz=1,NZ
```

- 32x32x32の格子を4x4x4MPI分割するため各MPIプロセスはNX x Ny x Nz = ８ x ８ x ８ を計算
- 16OMPの演算負荷均等化のため，ループを融合してループ長64にしてから分割

【変更後】

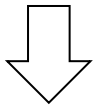
```
!$OMP PARALLEL DO COLLAPSE(2)
PRIVATE(ix,iy,iz,ieoxyz,ix1,iy1,iz1,ix2,iy2,iz2,ix3,iy3,iz3,&
!$OMP&                    itb,itb0,itb1,itb2,itb3,jc,ic,ve,vo)
do ix=1,NX
do iy=1,NY
do iz=1,NZ
```

QCDの性能評価： Clover部のSMP化(2)

(2)clover, full2linear_clv関数のOMPプロセスの演算負荷均等化

1.【変更前】clover.h 67行目

!$OMP PARALLEL DO PRIVATE(ix,iy,iz,ieoxyz,itb,ic,jc,ix2,iy2,iz2,itb2,ix4,iy4,iz4,itb4)

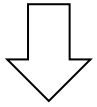


【変更後】

!$OMP PARALLEL DO COLLAPSE(2)
PRIVATE(ix,iy,iz,ieoxyz,itb,ic,jc,ix2,iy2,iz2,itb2,ix4,iy4,iz4,itb4)

2.【変更前】clover.h 162行目

!$OMP PARALLEL DO PRIVATE(ix,iy,iz,ieoxyz,itb,ic,jc,ix5,iy5,iz5,itb5,ix6,iy6,iz6,itb6)

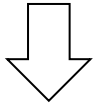


【変更後】

!$OMP PARALLEL DO COLLAPSE(2) SCHEDULE(STATIC,2)
PRIVATE(ix,iy,iz,ieoxyz,itb,ic,jc,ix5,iy5,iz5,itb5,ix6,iy6,iz6,itb6)

3.【変更前】full2linear_clv.h90 43行目

!$OMP PARALLEL DO PRIVATE(iy,iz,ieoxyz,itb,itb0)



【変更後】

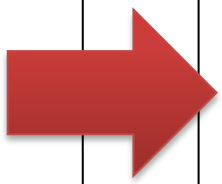
!$OMP PARALLEL DO COLLAPSE(2) PRIVATE(iy,iz,ieoxyz,itb,itb0)

QCDの性能評価： Clover部のSMP化(3)

(3)clvinv_ldl関数(clvinv_ldl.h90)のOMPプロセスの負荷均等化

【変更前】 clvinv_ldl.h90 31行目

```
!$OMP PARALLEL PRIVATE(ics,jcs)
!$OMP DO
 do jcs=1,CLSP/2
 do ics=1,CLSP/2
  zunit(ics,jcs)=(0.0d0,0.0d0)
 enddo
 enddo
!$OMP END DO
!$OMP DO
 do ics=1,CLSP/2
  zunit(ics,ics)=(1.0d0,0.0d0)
 enddo
!$OMP END DO
!$OMP END PARALLEL
```



【変更後】

```
!$OMP PARALLEL PRIVATE(ics,jcs)
!$OMP DO COLLAPSE(2)
 do jcs=1,CLSP/2
 do ics=1,CLSP/2
  if (ics .eq. jcs) then
   zunit(ics,jcs)=(1.0d0,0.0d0)
  else
   zunit(ics,jcs)=(0.0d0,0.0d0)
  end if
 enddo
 enddo
!$OMP END DO
!$OMP END PARALLEL
```

QCDの性能評価： Clover部の基本プロファイル

32x32x32x160 (4x4x4並列)プロシージャープロファイル

対策前

Cost	%	Operation (S)	Start	End	
1386	61	14.0	1	345	clover_
276	12	2.8	--	--	__brk
154	7	1.6	119	155	ldlbksb._OMP_17_
84	4	0.8	175	203	ldldcmp._OMP_18_
69	3	0.7	162	238	clover._OMP_8_
51	2	0.5	43	98	full2linear_clv._OMP_26_
49	2	0.5	49	61	clvinv_ldl._OMP_14_
44	2	0.4	67	141	clover._OMP_7_
33	1	0.3	642	656	comlib.comlib_sumcast_r8_

対策後

Cost	%	Operation (S)	Start	End	
271	24	2.8	--	--	_brk
223	20	2.3	642	656	comlib.comlib_sumcast_r8_
144	13	1.5	118	154	ldlbksb._OMP_17_
89	8	0.9	174	202	ldldcmp._OMP_18_
72	6	0.7	256	339	clover(OMP化)
57	5	0.6	162	256	clover._OMP_8_
52	5	0.5	67	141	clover._OMP_7_
45	4	0.5	48	60	clvinv_ldl._OMP_14_
33	3	0.3	43	98	full2linear_clv._OMP_26_
30	3	0.3	99	118	clover_f1f2._OMP_12_

メモリ確保時間が大きい
原因は現在解析中

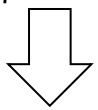
※ __brkはメモリ確保のライブラリ

QCDの性能評価： BiCG部のSMP化

(1)BiCGstab_hmc関数のOMPプロセスの演算負荷均等化

1.【変更前】bicgstab_hmc.h90 138行目

!$OMP PARALLEL DO PRIVATE(ix,iy,iz,ieoxyz,itb,ic,is) REDUCTION(+:ctmp0)



【変更後】

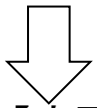
!$OMP PARALLEL DO COLLAPSE(2) PRIVATE(ix,iy,iz,ieoxyz,itb,ic,is) REDUCTION(+:ctmp0)

2. bicgstab_hmc.h90 166行目，221行目，254行目，305行目も同様

(2)copy_y関数のOMPプロセスの演算負荷均等化

1.【変更前】copy_y. h90 26行目

!$OMP PARALLEL DO PRIVATE(ix,iy,iz,ieoxyz,itb0,itb1,ic,is)



【変更後】

!$OMP PARALLEL DO COLLAPSE(2) PRIVATE(ix,iy,iz,ieoxyz,itb0,itb1,ic,is)

2.【変更前】copy_y.h90 46行目，61行目，85行目，104行目，118行目，
140行目，158行目，172行目， 194行目も同様

QCDの性能評価： BiCG部の基本プロファイル

32x32x32x160 (4x4x4並列)プロシージャープロファイル

対策前

Cost	%	Operation (S)	Start	End	
7901	31	79.2	34	691	mult_eo_tzyx._OMP_31_
6770	27	67.8	463	477	comlib.comlib_sendrecv_c16_
1432	6	14.3	254	275	bicgstab_hmc._OMP_5_
1199	5	12.0	1	214	copy_y_
986	4	9.9	138	153	bicgstab_hmc._OMP_2_
794	3	7.9	166	185	bicgstab_hmc._OMP_3_
771	3	7.7	221	238	bicgstab_hmc._OMP_4_
648	4	6.5	52	67	mult_mb_pre._OMP_33_
576	2	5.8	85	118	copy_y._OMP_21_

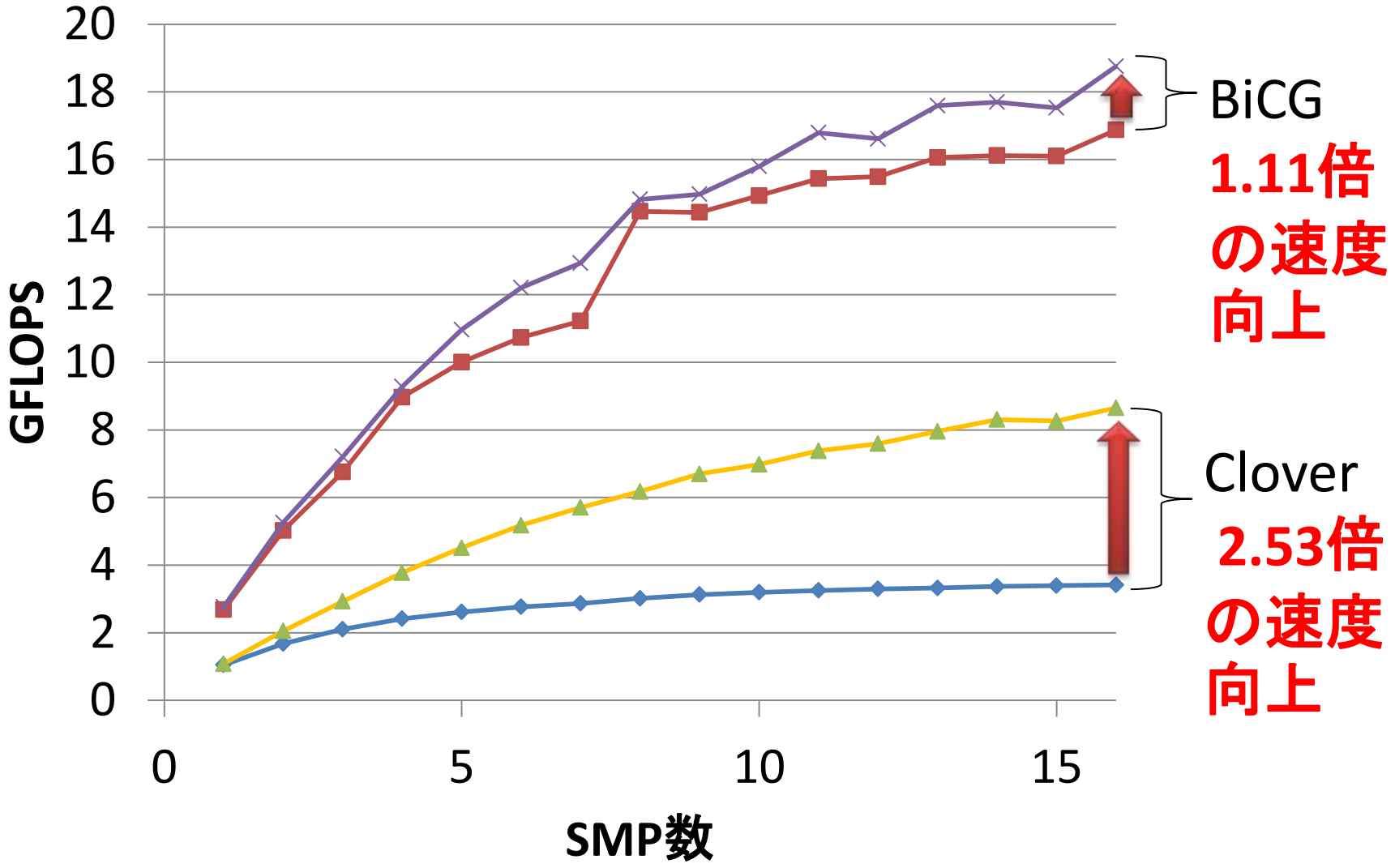
対策後

Cost	%	Operation (S)	Start	End	
7876	35	79.0	34	691	mult_eo_tzyx._OMP_31_
6856	30	68.7	463	477	comlib.comlib_sendrecv_c16_
810	4	8.1	254	275	bicgstab_hmc._OMP_5_
799	4	8.0	166	185	bicgstab_hmc._OMP_3_
706	3	7.1	305	321	bicgstab_hmc._OMP_6_
681	3	6.8	52	67	mult_mb_pre._OMP_33_
571	3	5.7	138	153	bicgstab_hmc._OMP_2_
553	2	5.5	528	539	comlib.comlib_barrier_
529	2	5.3	642	656	comlib.comlib_sumcast_r8_

← カーネル

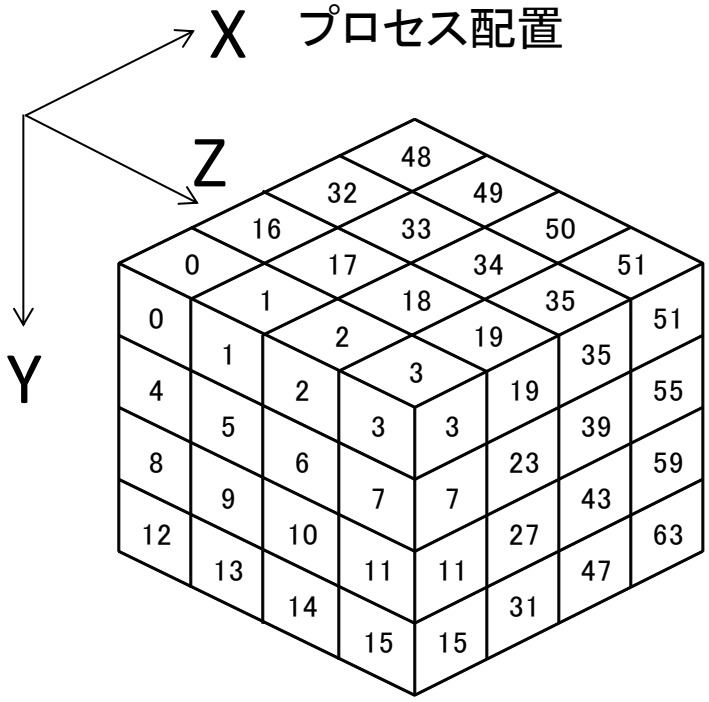
その他

QCDの性能評価： チューニング結果（FX10）



QCDの性能評価： 通信分析(1/2)

QCDのデータ分割方法と通信パターン



図中の数字はMPIプロセス番号を示す

(1)データ分割

- 4次元空間の内，xyzの空間3次元を分割してMPIプロセスにマッピング

(2)通信パターン

- 隣接空間を担当するMPIプロセス間で境界データを交換
- xyz格子の端は周期境界となるようにMPIプロセス間で境界データを交換
- 1要素のMPI_allreduce

QCDの性能評価： 通信分析(2/2)

Bicgstab内部の通信の種類，通信長と回数の解析結果
解析条件：格子数32x32x32x160(4x4x4分割)
反復回数 Iter= 359(=180ステップ＊2-1)

通信関数	実行関数	データ種類	通信長[B]	回数	回数の式
MPI_Allreduce	bicgstab_hmc	real(8)	8	539	1.5*Iter+初期1回
MPI_Allreduce	bicgstab_hmc	complex(8)	16	538	1.5回*Iter
MPI_Barrier	copy_y	----	----	718	2回*Iter
MPI_Sendrecv	copy_y	complex(8)	995328	4308	12回*Iter

995328[B]=3(カラー)×4(スピン)×81×8×8×16[B]
81=T方向160の半分の80 に 1を追加
8=32要素/4並列

clover部の通信の種類，通信長と回数の解析結果

通信関数	実行関数	データ種類	通信長[B]	回数
MPI_Allreduce	clvinv_ld	real(8)	8	4