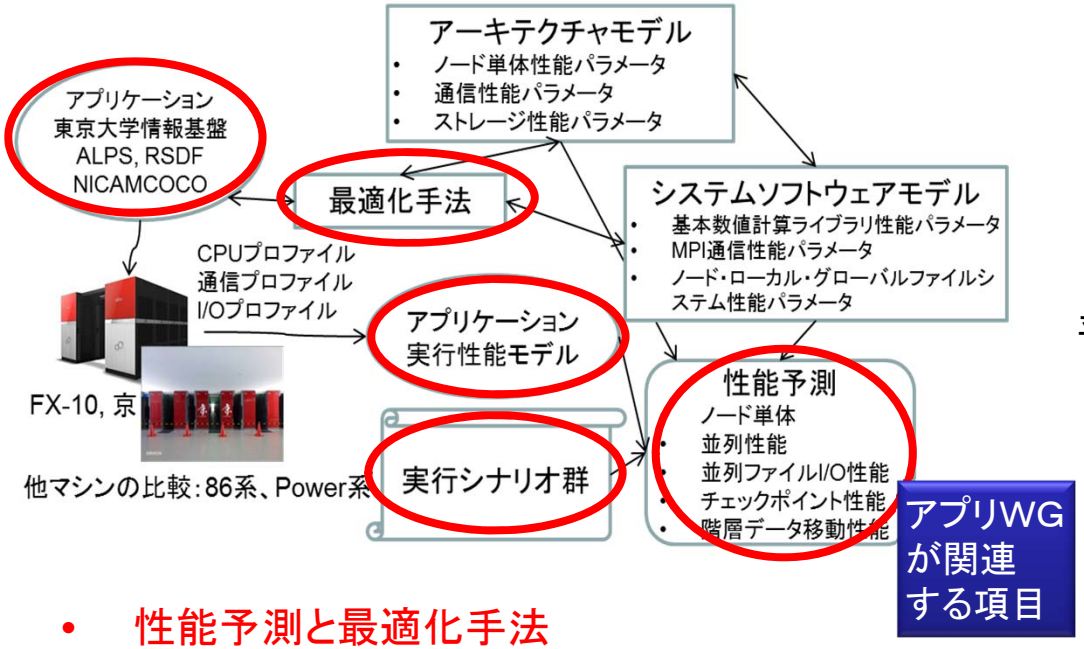
アーキテクチャチーム３：
「レイテンシコアの高度化・高効率化による将来のHPCIシステムに関する調査研究」
アプリケーション性能予測 WG

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（日立情報・通信システム社），高山 恒一（日立中研）
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）

理研FS全体会議資料
日時：2012年11月20日（火）13:00～
場所：東京大学八重洲カンファレンスホール

アプリケーションおよびシステムソフトウェアと性能予測

東京大学情報基盤センター



平成24年度

	7～9	10～12	1～3
システムソフトウェア	Proof of Concept実装＆評価項目精査	Proof of Concept実装	
性能予測と最適化手法	ベンチマーク化、プロファイル測定、プロファイル検討、最適化手法検討		

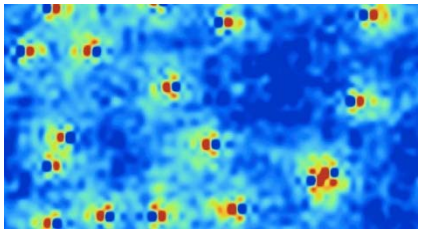
平成25年度

	上半期	下半期
システムソフトウェア	Proof of Concept実装（スパイラル開発モデル）	
性能予測と最適化手法	ターゲットアプリケーション拡大検討 ベンチマーク化、プロファイル測定、プロファイル検討、最適化手法検討	

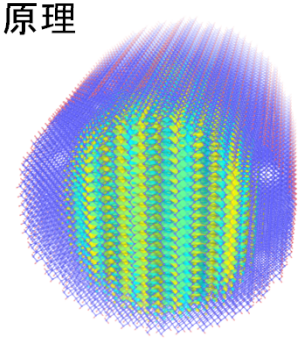
- 性能予測と最適化手法
  - ターゲットアプリケーションから抽出した演算と通信カーネルをベンチマークプログラム化
  - 東京大学情報基盤センターのFX10および京を用いて性能予測のための性能パラメータの抽出
  - 最適化手法の検討
  - 平成25年度ターゲットアプリケーション拡大のために理研AICSと連携
- システムソフトウェア
  - ヘテロOSカーネルおよび低レベル通信機構のproof of concept実装し、proof of concept実装に基づきハードウェア概念設計に反映するとともに、システムレベルの性能パラメータ（通信、ファイルI/O)を示す
    - システムソフトウェア設計では理化学研究所AICSシステムソフトウェア研究チームと連携

ターゲットアプリケーション群

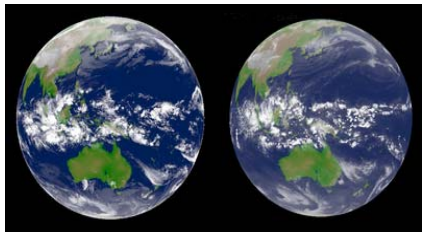
- ALPS/looper
  - 新機能を持った強相関・磁性材料の物性予測・解明。虚時間経路積分にもとづく量子モンテカルロ法と厳密対角化
  - 総メモリ：10～100PB
  - 整数演算、低レイテンシ、高次元のネットワーク
  - 利用シナリオ：1ジョブ当たり24時間、生成ファイル：10GB.
    同時実行1000ジョブ、 合計生成ファイル：10TB.



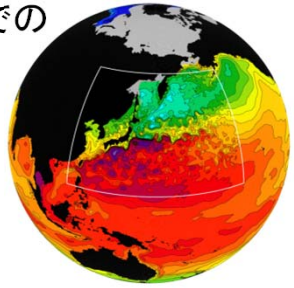
- RSDFT
  - Siナノワイヤ等、次世代デバイスの根幹材料の量子力学的第一原理シミュレーション。実空間差分法
  - 総メモリ：1PB
  - 演算性能：1EFLOPS（B/F =0.1以上）
  - 利用シナリオ：1ジョブ当たり10時間、生成ファイル：500TB.
    同時実行10ジョブ、合計生成ファイル5 PB.



- 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.



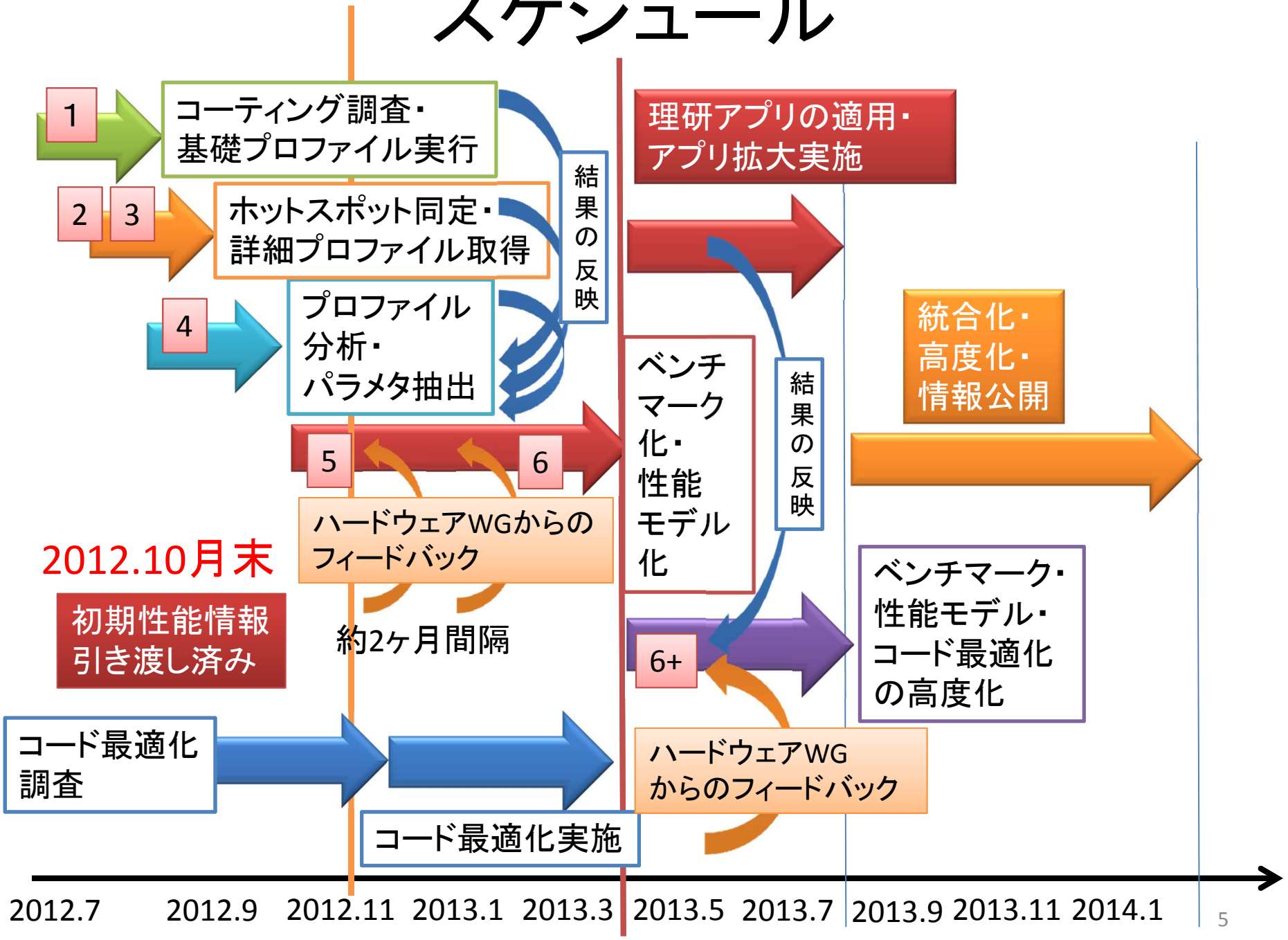
利用シナリオ
アンサンブル型
全系の1/10～1/100資源を用いた1ジョブを、複数同時実行することで、全資源を使い切る形態。

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**：（by TAU？）
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**：ハードウェア因子を用いた数式による実行時間の近似

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
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
ベンチ
マーク
化・
性能
モデル
化
結果の反映
ハードウェアWGからの
フィードバック
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

演算、通信の分類

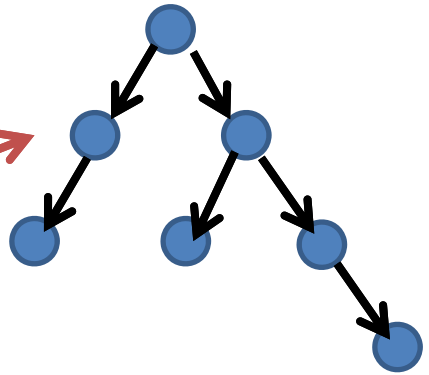
- 演算パターンの分類
  - 主演算の種類
    - 整数演算、浮動小数点演算
  - 配列アクセスパターンの種類
    - 密行列-行列積系
      - 連続アクセス、ブロック化可能
      - 行列サイズが反復ごとに縮小する／しない
    - ステンシル演算系
      - 連続アクセス
      - 等間隔アクセス
    - 疎行列-ベクトル積系
      - 間接参照、非連続アクセス
      - （オーダリングにより）アクセス局所化可能（一部連続アクセス）
      - アクセス局所化不可能（ランダムアクセス）
    - 探索系
      - ランダムアクセス、IF文成立予測不可能
- 通信パターンの分類
  - 隣接通信系
    - 送るべき相手の数の違い
  - 規則通信系
    - バタフライ形状、リング形状、ツリー形状、・・・
  - 同時通信系
    - 2次元プロセッサー・グリッド分割時の行もしくは列グループに対する同時の集団通信
  - 全体全通信系

演算の特徴

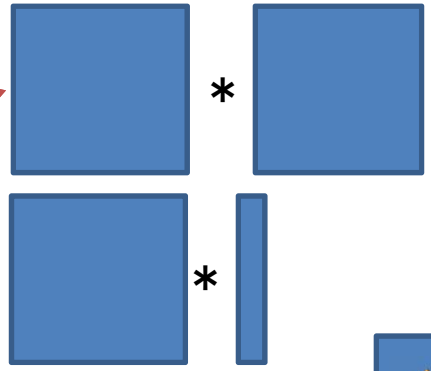
予備評価の結果であり
最終的なベンチマーク性能
を示すものではありません

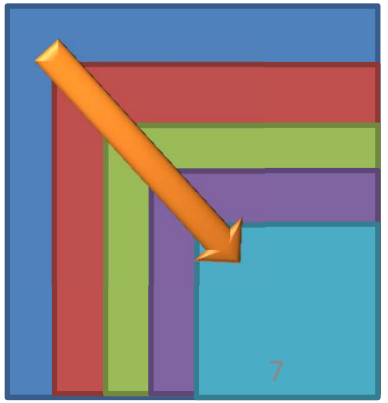
- 探索系
- ALPS/Looper
  - 整数演算
  - 探索処理
    - IF文の成立は入力データ依存、事前に予測はできない（予測できれば探索の必要がない）

●OMP_48
Byte/Inst.=8.36
実行性能：3386[MIPS]
実行性能：2.86%



- 密行列-行列積系
- RSDFT
  - 密行列　行列-行列積（BLAS3）
  - 密行列　行列-ベクトル積（BLAS2）
  - 固有値ソルバー部分
    - BLAS3の問題サイズは、ブロックサイズごとの呼び出しで、反復ごとに縮小していく（Cf. LU分解）
    - BLAS2の比率が高め
  - 直交化部分
    - BLAS3の問題サイズが変化せず





●diag_2d_dgemm
Byte/FLop=0.07
実効性能：42.4 [GFLOPS]
実行効率：71.4%

演算の特徴

予備評価の結果であり最終的なベンチマーク性能を示すものではありません

- ステンシル演算系
- NICAM：以下の2種が存在

- ステンシル演算（力学過程,dynamics）
  - 陽解法（時間ステップ遅いモード）、水平方向
  - Runge Kutta法
  - IF文が多い（トレーサー移流）
    - 質量の厳密保存のため
  - 配列数は1個、以下の2種（modオペレータ）
    - 発散項：**ストライドアクセス(mod_oprt.01)**
    - 移流、フラックス制限：**連続アクセス、最内IF文あり(mod_oprt.03)**
  - **4配列、連続アクセス、最内IF文あり（mod_advlim_thuburn.04）**
  - **3配列、連続アクセス(K、K+1)（mod_src.05）**
  - 隣接通信を含む

- ステンシル演算?（物理過程,physics）
  - 陰解法（時間ステップ速いモード，陽解法の安定化条件満たさず）、鉛直方向
  - 水平陽解法鉛直陰解法（HEVI）、一次元ヘルムホルツ方程式、3重対角行列
  - 鉛直方向に依存関係のある計算がボトルネック
  - 通信を含まない
  - 演算ロードバランスが悪い：雲があるところは計算が重い
  - 雲微物理過程を解く（相変化）：**2配列、連続アクセス、演算多数(mod_mp_nsw6.02)**

●mod_oprt.01（対ピーク：9.2%）

```
do l=1,ADM_lal
 do k=1,ADM_kall
   do n=nstart,nend
     ....
ij=n        ip1j=ij+1
ip1jp1=ij+1+ADM_gall_1d
ijp1=ij+ADM_gall_1d
 scl(n,k,l)=(                    &
+cdiv(0,ij,l,1)*vx(ij    ,k,l) &
+cdiv(1,ij,l,1)*vx(ip1j  ,k,l) &
+cdiv(2,ij,l,1)*vx(ip1jp1,k,l) &
+cdiv(3,ij,l,1)*vx(ijp1  ,k,l) &
+cdiv(4,ij,l,1)*vx(im1j  ,k,l) &
+cdiv(5,ij,l,1)*vx(im1jm1,k,l) &
+cdiv(6,ij,l,1)*vx(ijm1  ,k,l) &
 ...
```

●mod_oprt.03（対ピーク：3.7%）

```
do l=1,ADM_lal
 do k=1,ADM_kall
   do n=nstart,nend
     ....
  s_m1_k_min_n =
min(s_in_min(n,k,l,1),s_in_min(n,k,l,2),s_in
_min(n,k,l,3)...
  if(s_m1_k_min_n==CNST_MAX_REAL)
s_m1_k_min_n = s(n,k,l)
c_out_sum_n &          =
(0.5D0+sign(0.5D0,c(n,k,l,1)))*c(n,k,l,1)&
+ (0.5D0+sign(0.5D0,c(n,k,l,2)))*c(n,k,l,2)&
+ (0.5D0+sign(0.5D0,c(n,k,l,3)))*c(n,k,l,3)&
...
 if(abs(c_out_sum_n)<CNST_EPS_ZERO)
then
...
```

●mod_mp_nsw6.02（対ピーク：8 .3%）

```
do k = kmin, kmax
 do ij = 1, ijdim
    temc = tem(ij,k)-CNST_TEM00
    多数の演算
    V_TERM(ij,k,I_QR) = - (cr * rho_fact * gam_br_dr_1 / gam_br_1 * (olambdr_dr))
    多数の演算
   if( cnst_v_term_qi==cnst_undef ) then
    V_TERM(ij,k,I_QI) = - 3.29D0 * abs( rho_a*tmp ) ** 0.16D0        else
    ...
    if(temc>0.0D0) then
      多数の演算
    else
      ....
```

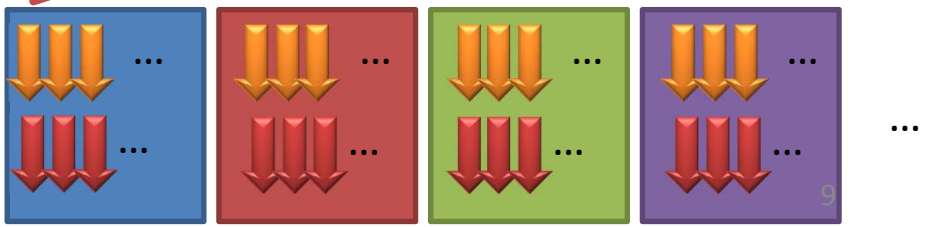
演算の特徴

予備評価の結果であり
最終的なベンチマーク性能
を示すものではありません

- ステンシル演算系
- COCO
  - ステンシル演算
  - ブロック化実装済み
    (ただしflxomp5のみ)
  - 配列は連続アクセス
  - 最内にIF文が存在
    - トレーサー移流
    - 理論上おかしい負の値
      になるのを防ぐ
  - ループ中で同時にアクセスされる配列数が多い
    - 現コードでは、最大で
      11個程度同時参照

●Tflxt.iso-som-flxomp5（対ピーク：2.3%）

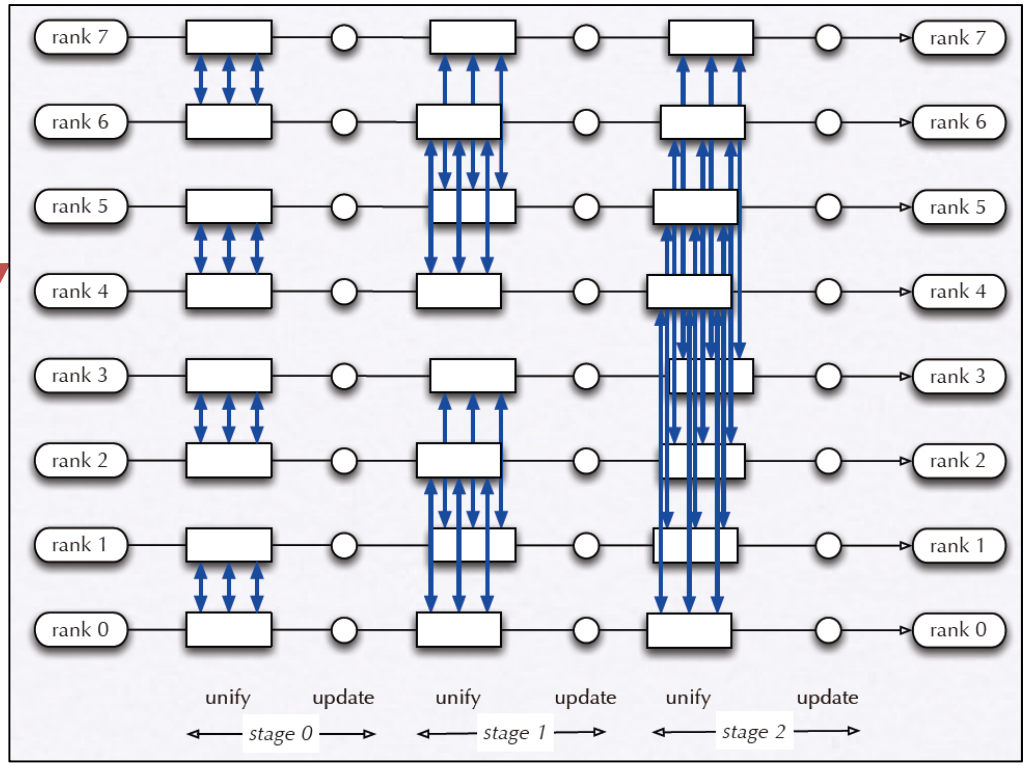
```
DO IJ1 = IJTSTR, IJTEND, IBLOCK
  DO K = KSTR, KEND
    DO IJ = IJ1, IJ2
      SOM = S0(IJ, K, N) - MIN( S0(IJ, KU, N) /
SM(IJ, KU, N), …
  …
      IF ( ABS( SZ(IJ, K, N) ) < 1.5D0 * SOM ) THEN
        SZZ(IJ, K, N) = MIN( SOM + SXP,
MAX( ABS( SZ(IJ, K, N) ) - SOM,   SZZ(IJ, K, N) ) )
      ELSE
        SZZ(IJ, K, N) = MIN( SOM + SXP,
MAX( SOM - SXP, SZZ(IJ, K, N) ) )
      END IF
    …
    ENDDO
  ENDDO
…
ENDDO
```



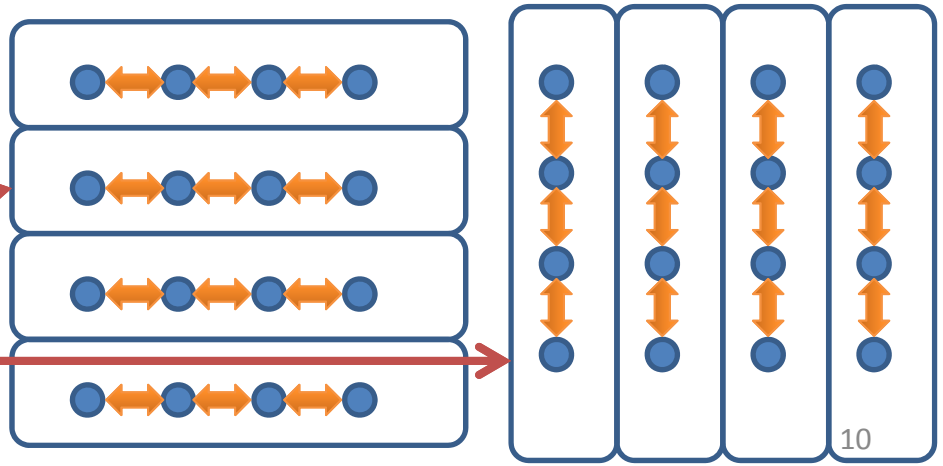
通信の特徴

Copy-Right: Prof. Synge Todo

- 規則通信系
- ALPS/Looper
  - バタフライ形状
    - オーバラップ版

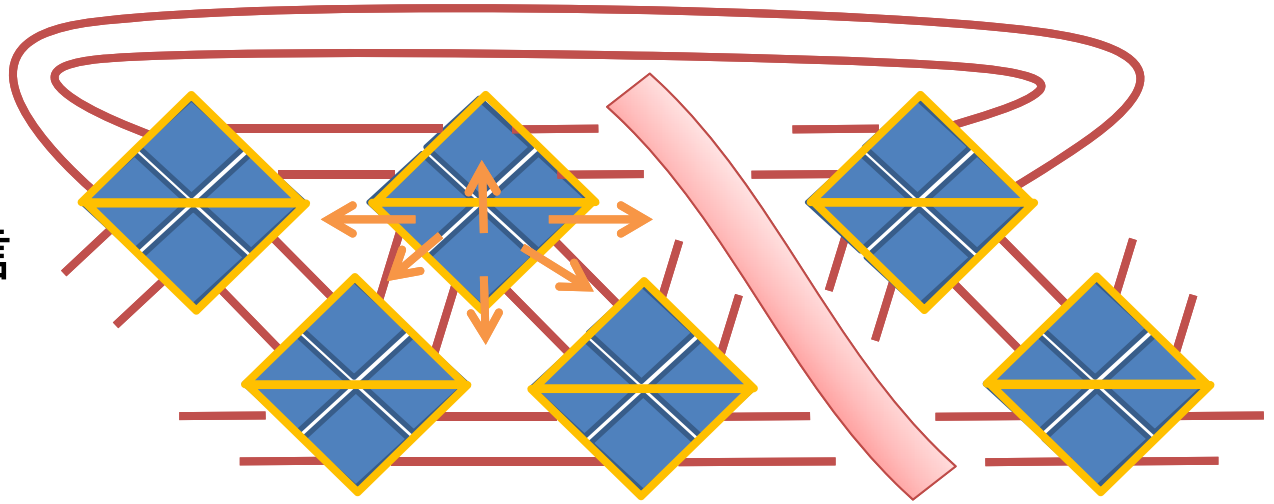


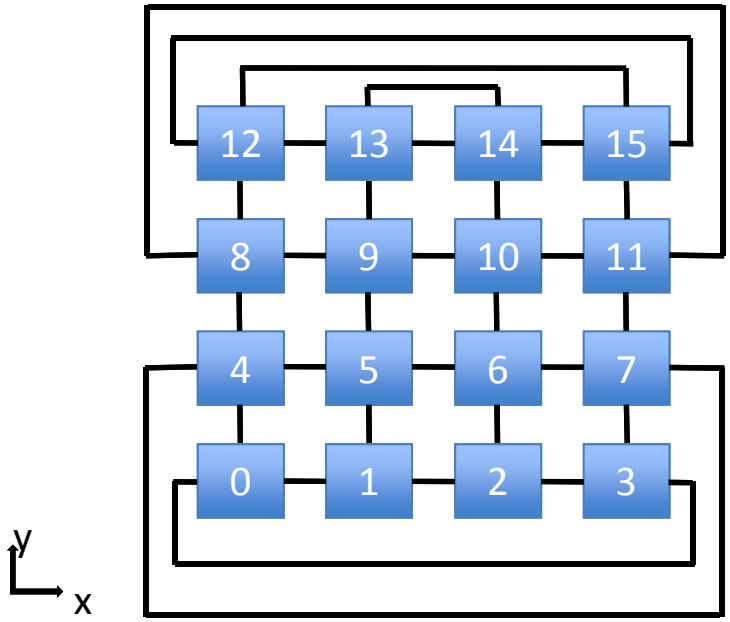
- 同時通信系
- RSDFT
  - 2次元プロセッサー・グリッド配置
  - 行/列方向同時通
    - MPI_Bcast
    - MPI_Allreduce



通信の特徴

- 隣接通信系
- NICAM
  - 隣接通信
  - 主に6方向通信、最大で15方向通信
  - うまくマップすると、一部はノード内通信にできる
- COCO
  - 隣接通信
  - 直交格子
    - 極まわりはひずんでいるが展開して直交格子化
  - モデル座標の東西、南北を分割可能
    - 鉛直方向には分割できない
  - tripolar gridでは、1MPIプロセス当たり
    - ほとんど4方向
    - 場合により3方向





アプリケーション特性のまとめ

アプリ名	演算カーネル			通信カーネル	
	演算種類	配列アクセスパターン	特徴	通信パターン	特徴
ALPS/Looper	整数	探索系	IF文成立事前予測不可能	規則通信系（バタフライ形状）	オーバラップ
RSDFT	浮動小数点	密行列-行列積系	● 固有値ソルバー ● GS直交化	同時通信系	● 行／列方向の別 ● BCASTとAllReduce
NICAM	浮動小数点	ステンシル演算系	● 力学過程（最内IF文あり／なし） ● 物理過程（ループ内演算多数）	隣接通信系	主に6方向、最大で15方向
COCO	浮動小数点	ステンシル演算系	● 最内IF文 ● 同時参照配列多数	隣接通信系	最大で4方向

プロファイル結果

NICAM

- 特徴
  - 大気大循環モデル
  - 差分法，2次元分割
- 計算規模
  - g-level=9（水平格子：2621442）鉛直40層
- 実行形式
  - 40ノード, 160MPIx4omp

演算カーネルの性能(NICAM)

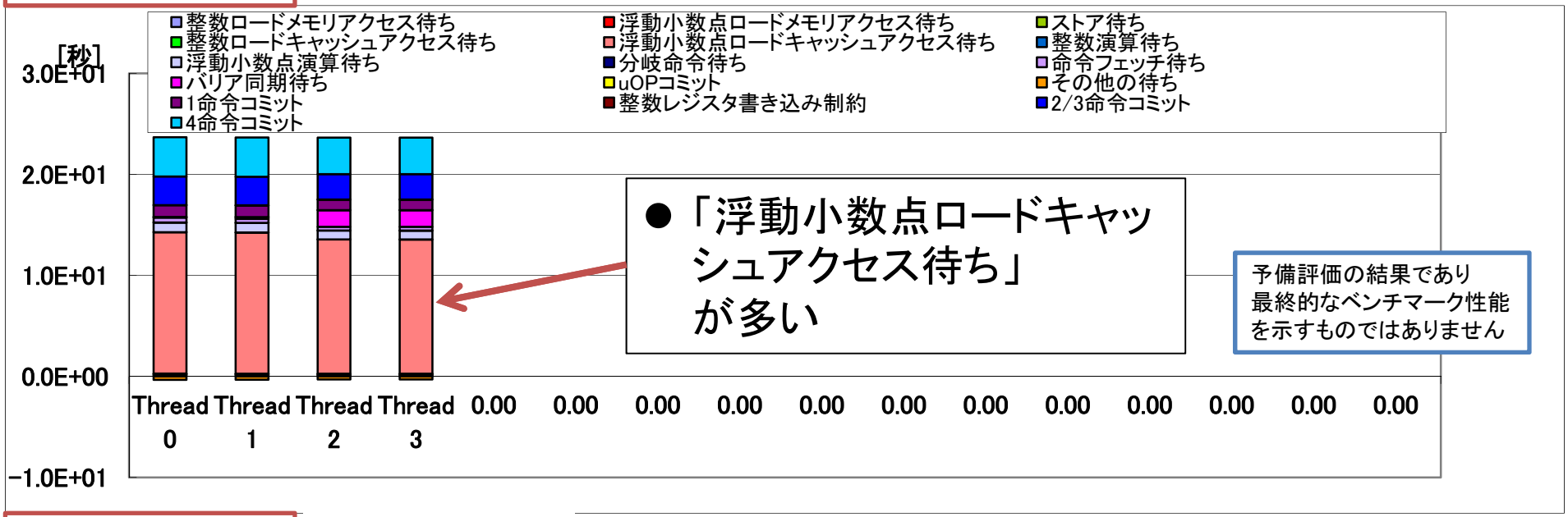
計算種別	カーネル名	スレッド数	GFLOPS (4threads)	% to Peak	Memory Throughput (GB/sec)/chip	% to Peak	B/F	% to Total Time	備考
	全体	4	2.867	4.85	48.80	57.41	4.26	100	
力学過程	mod_oprt_01	4	5.444	9.21	65.20	76.71	2.99	3.72	IF文なし
	mod_oprt_03	4	2.205	3.73	39.35	46.29	4.46	3.56	IF文あり
物理過程	mod_mp_nsw6	4	4.909	8.30	47.77	56.20	2.43	4.52	IF文あり

全体時間に対しMPI通信時間の占める割合は5.3%
＝上記カーネル1つ分

予備評価の結果であり
最終的なベンチマーク性能
を示すものではありません

詳細プロファイル結果(NICAM、力学過程)

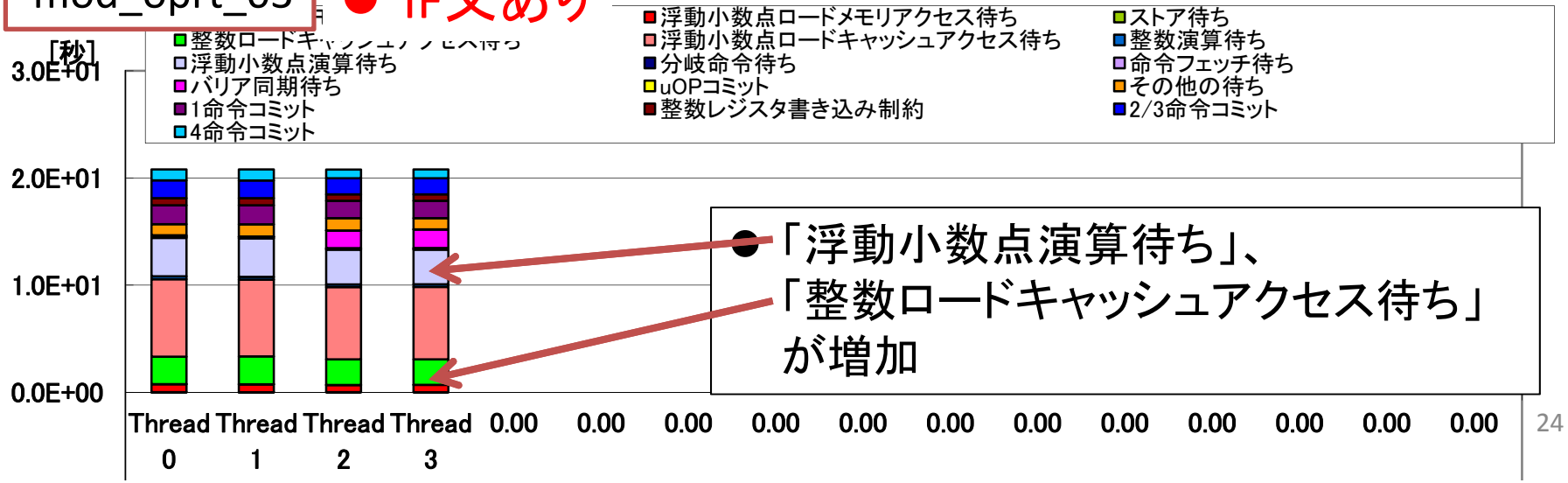
mod_oprt_01 ● IF文なし



● 「浮動小数点ロードキャッシュアクセス待ち」が多い

予備評価の結果であり
最終的なベンチマーク性能
を示すものではありません

mod_oprt_03 ● IF文あり



● 「浮動小数点演算待ち」、「整数ロードキャッシュアクセス待ち」が増加

NICAMの通信

(a) mpi_irecv / mpi_isend						
Kind	Byte	Call	0-4K	4K-64K	64K-1024K	1024K-
AVG	161,944	28,806	7,006	2,932	17,926	943
MAX	194,332	72,015	47,296	7,329	28,717	1,006
MIN	65,677	24,005	4,316	2,299	16,384	503

予備評価の結果であり
最終的なベンチマーク性能
を示すものではありません

おわりに

- 本発表では、東京大学で実施している「レイテンシコアの高度化・高効率化による将来のHPCIシステムに関する調査研究」における、アプリケーション評価の概要を紹介した
- コンピュータ・サイエンスの見地から、アプリケーションにおいて、以下の分類をして性能特性を評価すべきである
  - メモリアクセスのパターン
  - 通信のパターン
- 実アプリのコードから抽出した従来の分類では考慮されていない事項の評価が必要
  - 整数演算、探索処理の性能指標
  - ステンシル演算において最内にIF文があるループ構成
  - ループ本体が大きい（多数の演算）時のコンパイラ最適化
  - 集団通信の2次元プロセス・グリッド上での同時発行時の性能予測方法

今後の予定

- カーネルのチューニング
  - 抽出した実コードカーネルのチューニング、および、チューニング方法論の確立
- ベンチマークの作成
  - カーネルレベルのベンチマーク（カーネルベンチ）の作成と公開
- 性能パラメータの明確化
  - 特に、通信とI/Oに関する性能パラメータの確定
- Co-designの確立
  - 計算機ハードウェアの概念設計に、アプリケーション特性を反映
- 性能評価手法の汎用化
  - 理研FSと連携し、我が国における戦略分野のアプリケーション全般をカバーする評価手法の確立を目指す