東大FSアプリWG
進捗報告

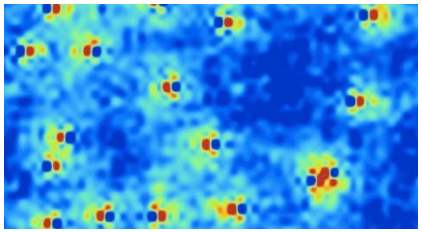
片桐 孝洋, 大島 聡史, 中島 研吾(東大)
米村 崇, 熊洞 宏樹, 樋口 清隆, 橋本 昌人(日立情報・通信システム社),
高山 恒一(日立中研), 藤堂 眞治, 岩田 潤一, 内田 和之, 佐藤正樹,
羽角博康(東大), 黒木聖夫(海洋研究開発機構)

理研FS　第4回全体ミーティング
2013/01/21　10:00 AM – 5:30 PM
TKP東京駅八重洲カンファレンスセンター
10:05－11:00　：　東北・筑波・東大FSチーム状況報告(江川、高橋、片桐)

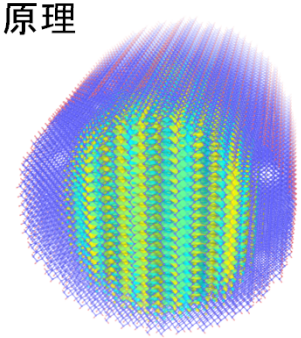
Feasibility Study on
Advanced and Efficient Latency Core-based Architecture for Future HPCI R&D

ターゲットアプリケーション群

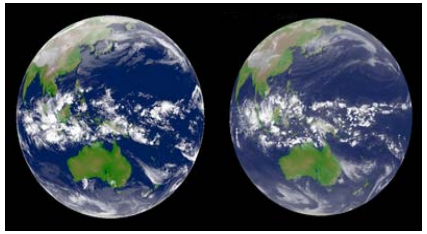
- ALPS/looper
  - 新機能を持った強相関・磁性材料の物性予測・解明。虚時間経路積分にもとづく量子モンテカルロ法と厳密対角化
  - 総メモリ：10～100PB
  - 整数演算、低レイテンシ、高次元のネットワーク
  - 利用シナリオ：1ジョブ当たり24時間、生成ファイル：10GB. 同時実行1000ジョブ、 合計生成ファイル：10TB.



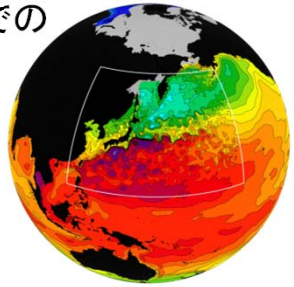
- RSDFT
  - Siナノワイヤ等、次世代デバイスの根幹材料の量子力学的第一原理シミュレーション。実空間差分法
  - 総メモリ：1PB
  - 演算性能：1EFLOPS（B/F =0.1以上）
  - 利用シナリオ：1ジョブ当たり10時間、生成ファイル：500TB. 同時実行10ジョブ、合計生成ファイル5 PB.



- 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.



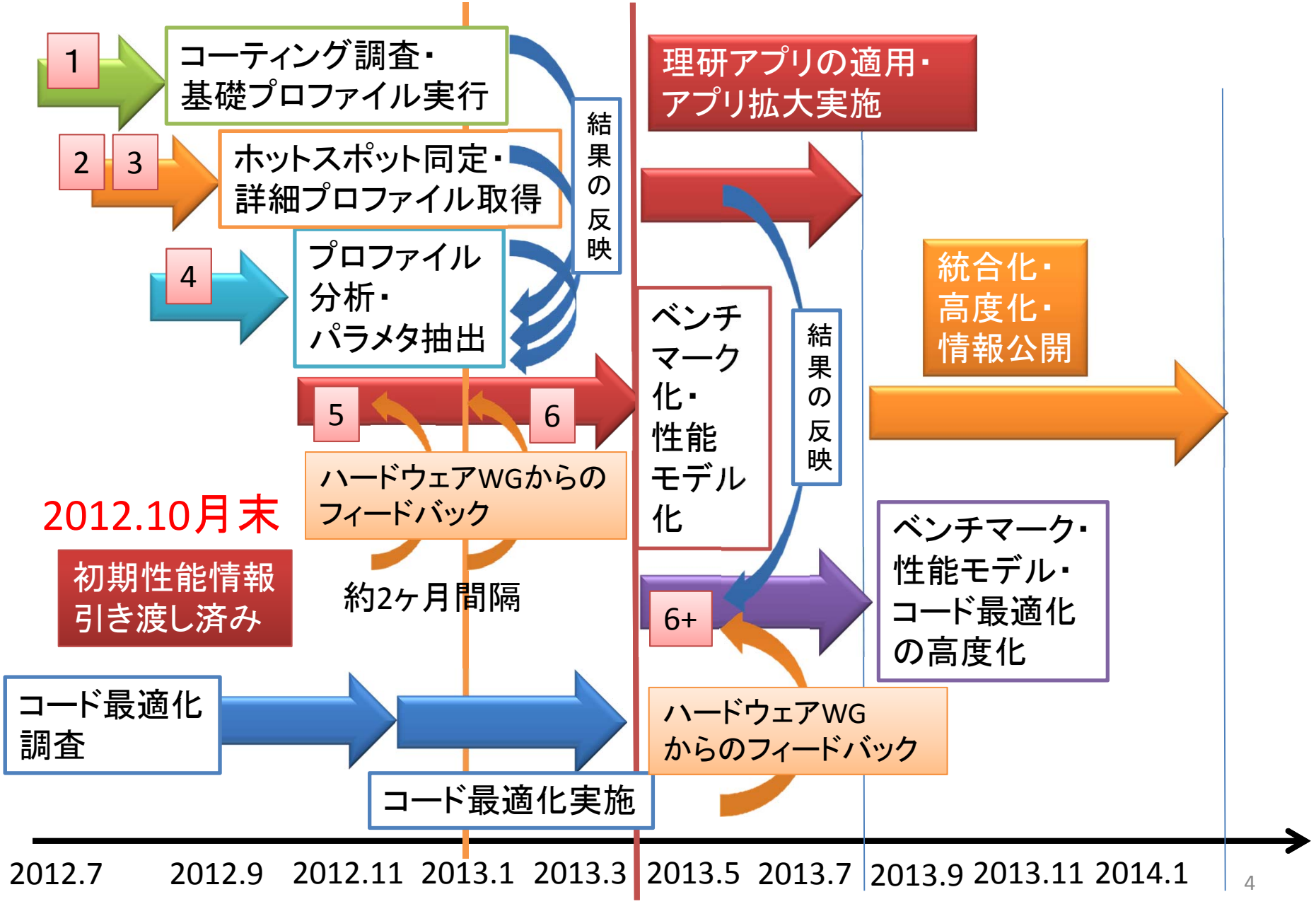
利用シナリオ
アンサンブル型
全系の1/10～1/100資源を用いた1ジョブを、複数同時実行することで、全資源を使い切る形態。

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**：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

現在の進捗のまとめ

- 演算カーネル、通信カーネルの抽出
- 理研FS（運用グループ）との連携
  - 対象となる演算カーネルの選定
  - 京で効率の良い実行環境情報（ジョブ割り当て）の活用
- 演算カーネルに対するチューニング
  - コンパイラオプションの変更
  - コーティング方法の変更
- コデザイン
  - 概念設計中マシンのパラメタの反映
    - メモリ量制約、電力などの制約と演算能力の向上のトレードオフを考慮
- 概念設計中の計算機での実行時間予測
  - 演算カーネルのプロファイルからの演算時間予測
  - 通信パターンなどからの通信時間予測
  - 性能予測手法の検討

前回からの演算カーネルの差分

- コードチューニング
  - コンパイラによるSIMD化の促進 (-ksimd=2等)
  - ALPS/looperコードチューニング
    - 非連続アクセスによる高スレッド実行時の効率悪化対策
    - スレッド毎に計算した結果を集約する処理
      - 1～15スレッドまでの結果を0番スレッドに足しこむ処理
      - 各スレッドがスレッド数分の箇所にアクセスするストライドアクセス
      - 3.40[s]　から　1.53[s]に短縮　（約1/2.2）
  - COCOコードチューニング
    - タイリングサイズ調整(flaxomp5)
      - IBLKを536に変更することにより24.414[s]→14.717[s]，39.7%の短縮
    - flxtrc_omp2，flxtrc_omp3の条件分岐最適化
      - 条件分岐中のストアをスカラ変数に変更
      - flxtrc_omp2: 22.640[s] → 19.269[s]，14.9%の短縮
      - flxtrc_omp5: 13.893[s] → 13.557[s]，2.4%の短縮
- 最新コードに差し換え
  - NICAM
    - 力学過程の演算カーネルでIF文を含まないカーネルになった
    - 効率の良い実装：　物理過程:対ピーク8%→22%に向上

演算カーネルの性能(ALPS)

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

カーネル名	スレッド数	GIPS(16threads)	% to Peak	SIMD %	Memory Throughput (GB/sec)/chip	% to Peak	B/I	% to Total Time	備考
全体	16	8.178	6.91	0.12	13.77	16.20	1.68	100	
OMP_47	16	3.531	2.98	1.03	23.96	28.19	6.79	19.20	探索処理
OMP_43	16	11.08	9.36	0	0.78	0.92	0.07	9.55	探索処理

合計：28.7%

演算カーネルの性能(RSDFT)

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

割合は、プロセス0、スレッド0の値

カーネル名	スレッド数	GFLOPS (16threads)	% to Peak	SIMD %	Memory Throughput (GB/sec)/chip	% to Peak	B/F	% to Total Time	備考
全体	16	91.48	38.68		5.81	6.84	0.06	100	
diag_2d_	16	202.57	85.66	81.36	8.42	9.91	0.04	28.91	固有値ソルバー
gram_schmidt_sub_blkcyc_	16	159.07	67.27	76.79	5.61	6.60	0.04	33.33	直交化

合計：62.2%

演算カーネルの性能(NICAM)

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

計算種別	カーネル名	スレッド数	GFLOPS(16threads)	% to Peak	SIMD%	Memory Throughput (GB/sec)/chip	% to Peak	B/F	% to Total Time	備考
	全体	16	13.01	5.50	47.20	40.12	47.20	3.08	100	
力学過程	02_mod_oprt (oprt_divergence)	16	25.62	10.83	41.30	27.19	31.99	1.06	3.21	IF文なし、ステンシル
	04_mod_oprt3d (oprt3d_divdamp)	16	16.5	6.98	78.85	39.1	46.00	2.37	2.86	IF文なし、ステンシル、複数配列参照
	05_mod_src (src_flux_convergence_PRL_8_)	16	9.29	3.93	84.17	54.98	64.68	5.92	2.16	IFなし、ステンシル、配列初期化あり
	07_mod_src (src_flux_convergence_PRL_17_)	16	2.97	1.26	80.06	54.91	64.60	18.49	1.64	IF文なし、内部演算少ないループが複数
	08_mod_oprt (oprt_divergence2_rev)	16	13.48	5.70	60.59	40.19	47.28	2.98	1.35	IFなし、複数配列参照
物理過程	09_mod_mp_nsw6	16	52.18	22.07	71.64	14.28	16.80	0.27	1.34	**IF文あり**、最適化済み

合計：12.56%

演算カーネルの性能(COCO)

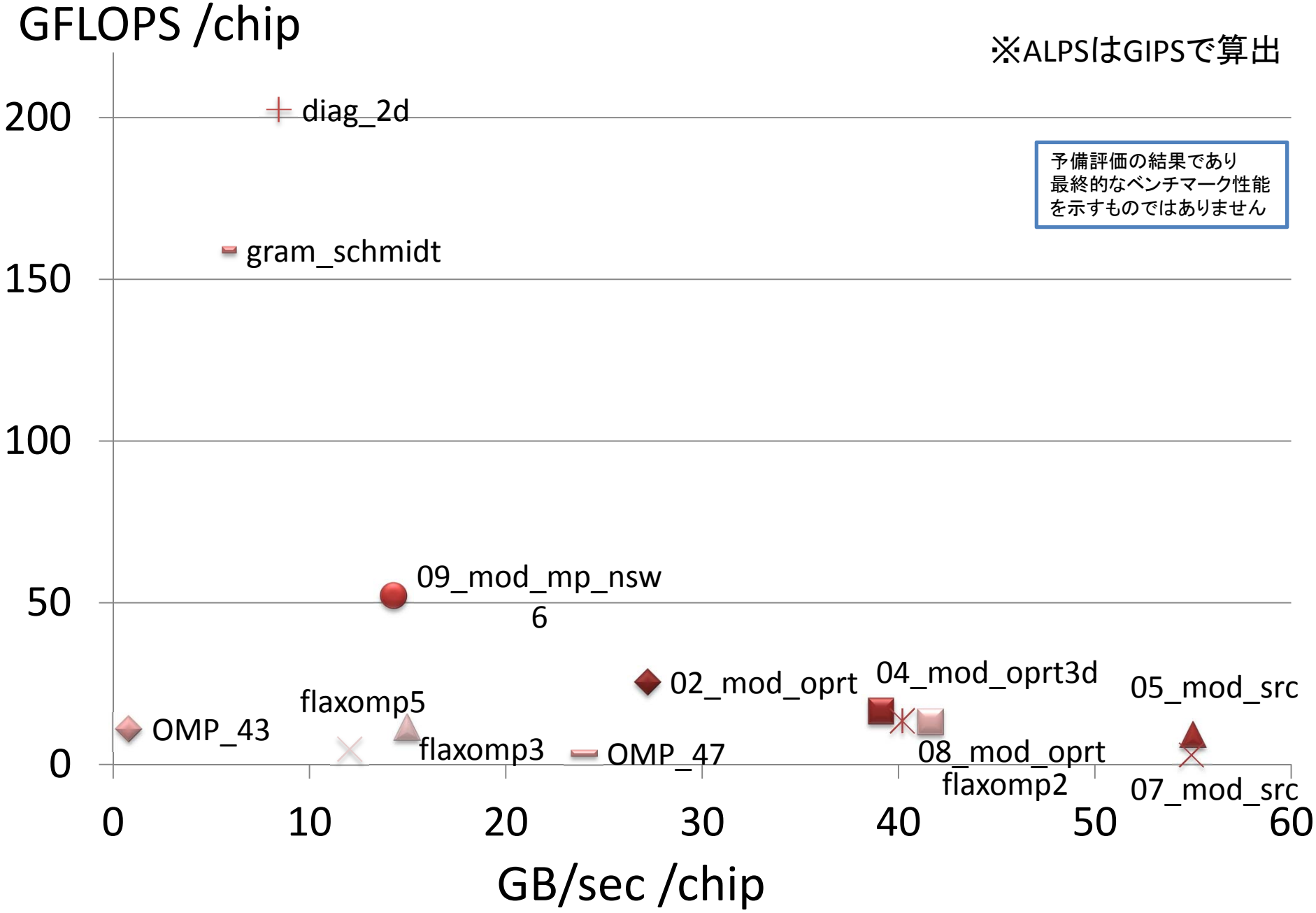
予備評価の結果であり
最終的なベンチマーク性能
を示すものではありません

プロセス0、
gathervに関する時間を除く時間の割合

カーネル名	スレッド数	GFLOPS(16threads)	% to Peak	SIMD %	Memory Throughput (GB/sec)/chip	% to Peak	B/F	% to Total Time	備考
全体	16	0.932	0.39	28.70	4.64	5.46	4.98	100	
flxomp2	16	13.21	5.59	43.04	41.6	48.94	3.15	11.85	IF文あり
flxomp3	16	11.53	4.88	50.24	14.95	17.59	1.30	8.17	IF文あり
flxomp5	16	4.86	2.06	66.42	12.03	14.15	2.48	10.07	IF文あり、ブロック化

合計：30.0%

演算カーネルの分類（実測B/F値）



実行時間予測方法

- FX10の実行時間

$$T_{FX10} = \sum_i K_i + \sum_i C_i$$

演算カーネル
および、通信カーネル
以外の時間は削除

K_i : 演算カーネルiの
プロファイラによる実測時間

C_i : 通信カーネルiの
プロファイラによる実測時間

- 概念設計マシンType(j)の実行時間予測

$$T_{Type(j)} = \sum_i \hat{K}_i + \sum_i \hat{C}_i$$

$\hat{K}_i$: 演算カーネルiの
性能予測ツールによる
予測時間

$\hat{C}_i$: 通信カーネルiの
予測時間

全体時間予測方法

アプリ名	スケーリングモデル	予測方法	理想的なエクサでの実行時間	備考
ALPS/Looper	弱スケーリング	主要カーネル2種の時間＋通信時間	プロセス数に依存せず一定時間	予測可能。
RSDFT	弱スケーリング（★） ★エクサ環境での実行時間を考慮	主要カーネル2種の時間＋通信時間	基準ノード数に対し理想的な演算性能（FLOPS値）の向上	領域量O(n^2)に対し、計算量O(n^3)、通信と計算のオーバラッピングの考慮が困難
NICAM	弱スケーリング	主要カーネル6種の実行時間＋通信時間	プロセス数に依存せず一定時間	予測可能。通信時間を悲観的にみる
COCO	弱スケーリング	主要カーネル3種の実行時間＋通信時間	プロセス数に依存せず一定時間	予測可能。通信時間を悲観的にみる

通信時間予測方法

アプリ名	対象箇所	パターン	メッセージ長	備考
ALPS/Looper	Union/Find部分1箇所	バタフライ	不変。	予測容易
RSDFT	三重対角化部分	2次元マルチキャストBcast	可変（反復ごとに減少してゆく）	予測困難、反復数はO(N)
	直交化部分	2次元マルチキャスト Allreduce	不変？ 反復ごとに通信が発生。	予測困難、反復数はO(N)
NICAM	袖領域のデータ交換に関する1箇所	隣接通信	メッセージ総量は固定。プロセス依存。	予測可能、ただし最大の通信回数になる箇所で代表
COCO	袖領域のデータ交換に関する1箇所	隣接通信	メッセージ総量は固定。ほぼ一定長？	予測可能、ただし最大の通信回数になる箇所で代表

問題点（1/2）

- RSDFTの性能予測方法
  - 単純な弱スケーリングは使えない
    - ノードあたりの問題サイズを固定すると、実行時間が増加し現実的な時間で終了しなくなる
  - 強スケーリングでは通信律速に
    - 実用上は強スケーリングになるはず
    - しかし現実的な時間で計算が終了する、できるだけ大きなサイズを1チップのサイズにしないと、台数効果が出ない
  - 現在取得しているノードあたりの問題サイズを弱スケーリングし、現実的な時間で終了するかを検証して推定
    - FX10での問題サイズと実行時間から、概念設計マシンの予測時間は、問題サイズの3乗に比例する係数で近似して予測
    - 科学的に意味のある大きさか、実用に耐えられる実行時間かどうかも、開発者側と検討する
  - 通信予測
    - 1回の通信時間を予測した上で、モデル上で近似

問題点(2/2)

- プロセスとジョブの割り当て方法
  - 通信パターンから、妥当と思われる割り当てを行う
  - 結果（性能）を評価し、問題があれば、割り当てを最適化する
    - NICAMでは、開発者の知見を採用する

今後の予定

- 演算カーネルのさらなるチューニング
  - 配列確保のしかたを変更し、ループ中で連続アクセスにするなど、データ構造の最適化
  - その他
- コデザインを考慮した実行環境での再プロファイル
  - 概念設計中のマシンのメモリ量制約を考慮
  - 状況により、実行シナリオ（同時実行数など）の変更
- 異機種環境での性能評価(HPCIマシンの利用)
  - 京、ただし大規模ノード実行での性能を見るために利用
  - NEC SX-9
  - HITACHI SR16000/M1
  - FUJITSU PRIMERGY CX400 (Sandy Bridge EP)
  - Cray XE6 (AMD Opteron6000, Bulldozer)
- 概念設計マシンでの性能予測
  - 特に、RSDFTの予測手法の確立