気象・気候・惑星科学共通ライブラリー化の提案 ---今後のHPCを見据えて

理化学研究所・計算科学研究機構
富田浩文

私、誰ですか？

- 東京大学大学院工学系研究科航空宇宙工学専攻
  - 1999：博士（工学）取得
  - 学位論文：ベナール対流のパターンフォーメーションに関する数値的研究
    - レイリーベナール対流、ベナールマランゴニ対流
- 1999: 地球フロンティア研究システム　参加
- 2004: （独）海洋研究開発機構　研究員、主任研究員
- 2011～（独）理化学研究所・計算科学研究機構
  複合系気候科学研究チーム　TL

高解像度大気モデルの開発と
それを用いた全球大気運動の研究

背景

- おのおののグループで、おのおのモデルを作っている。
  - モデルを作ること自体はとてもいいこと。
    - モデルを作ることで、いろいろ原理的なことが分かる。
  - でも、力は、分散。
    - 分散するのもいいこと？
- そもそも、地球科学のモデルは、一人で作れるものではない。
  - コンポーネントが多く、また多様化している。
- 多くのコンポーネントの組み合わせによって、結果が異なる。
  - 追試必要だが、やってるひまない。
  - 不健全。
- コンピュータのアーキテクチャーが複雑化しようとしている。
  - ヘテロジニアス、多階層メモリ構造
  - 普通に書いていたのでは、性能がでない。

一つのコンポーネントでも腐るほどある

- 流体力学部分：
  - 水平離散化：
    - スペクトル法、ダブルフーリエ、
    - 正20面体（各種）格子、
    - 立方体格子、陰陽格子、スペクトルエレメントなど、
  - 鉛直スキーム
    - 静力学方程式系
      - ほぼ、プリミティブ方程式できまり。
    - 各種非静力方程式系
      - 非弾性系、弾性系
    - 山の取り扱いなど
  - 時間スキーム
    - Explicit
      - Time split : Leapfrog, RK2,3,4
    - Implicit
      - 鉛直方向のみ？　3Dで？

物理過程の複雑化、多様化

- 各種パラメタリゼーション
  - 対流スキーム
    - Kuo, AS,　Tiedike,　K-F,
- 各種雲微物理
  - 大規模凝結スキームも多々
  - 1モーメント→ 2モーメント→ ビン法
- 放射：
  - 平衡平板モデル→ 3D放射
- エアロゾル、化学パッケージ
  - 種類がどんどん増えている。
- 陸面過程、などなど
  - 評価方法よくわからない。（私の勉強不足）
- 海洋モデルの渦パラメタリゼーション
  - GM他
- 生態系、氷床、、、う～ん、

提案

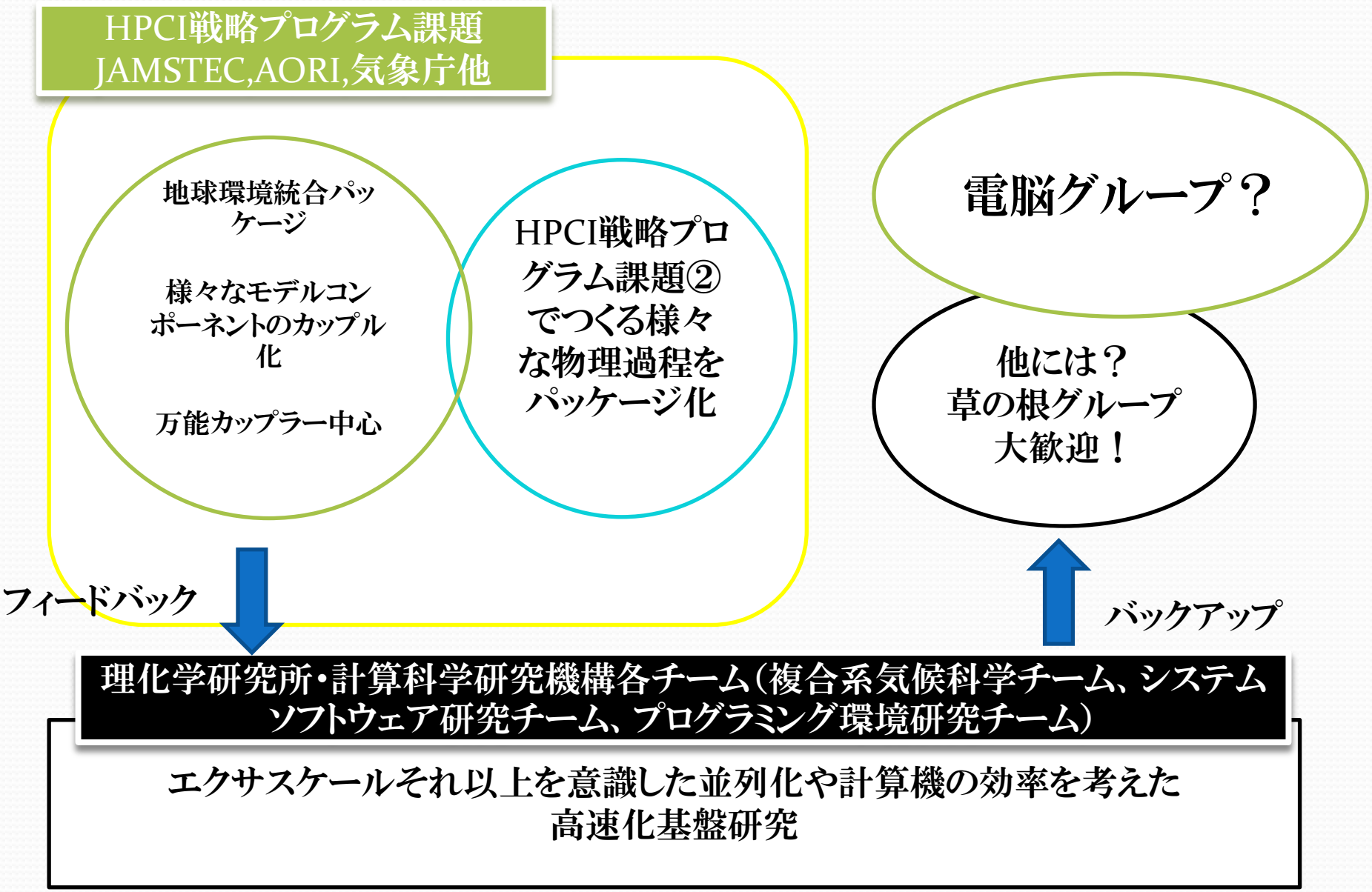
- 多くの主要なスキームを網羅し、
  - 誰でも自由に使えて、
  - 性能が出せるライブラリーを提供したい。
    - モデラーの育成
    - エクサスケールそれ以上を意識。
  - ドキュメント整備も大事。
- 一つのグループでは無理。
  - 国内外多くの機関、グループを巻き込みたい。
  - 難しい？？？

具体的には、実務者レベルから

- インターフェースを統一することは重要。
- コーディングスタイルの一貫性

- 一度、国内で主要モデルグループの実務者レベルでの会合が必要。
- どこがあげられますか？？？

構想図


HPCI戦略プログラム課題
JAMSTEC,AORI,気象庁他
地球環境統合パッケージ
様々なモデルコンポーネントのカップル化
万能カップラー中心
HPCI戦略プログラム課題②でつくる様々な物理過程をパッケージ化
電脳グループ？
他には？
草の根グループ
大歓迎！
フィードバック
バックアップ
理化学研究所・計算科学研究機構各チーム（複合系気候科学チーム、システムソフトウェア研究チーム、プログラミング環境研究チーム）
エクサスケールそれ以上を意識した並列化や計算機の効率を考えた
高速化基盤研究

- ライブラリー構築はいいが、、、
- HPCの動向を考えとかないと、、、
- 使えない！

さて、HPCの動向

- スカラー超並列
  - 地球シミュレータ1　640 -> K computer 80,000
  - エクサ機は？　10,000, 100,000,　1000,000
    - 壊れる！壊れる！壊れる！
- メニーコア化
  - AMD opteron, Intel sandy bridge
  - Knight, MICなど
- ヘテロジニアス化
  - GPGPUなどアクセラレータ装備（例：TSUBAME2）
- 主記憶メモリー性能の相対的低下
  - ES1: 4 -> ES2: 2.5 -> K: 0.5 -> エクサ：0.1???

うちらにとって好ましくないこと

- ヘテロジニアスアーキテクチャ：
  - アクセラレータどう使う？
  - うまく使えれば、とてもうれしい。
    - ASUCAなど
- 主記憶バンド幅低下：
  - 流体計算は全滅なのか？
- チェックポイントを強いられる。

BFあげることを要求するの？

- 電力問題

しかし、そもそもほんとのそんなにBFは必要か？

エクサコンピュータ10日間全系使用して、フルに性能出すために何が必要か？超理想論かも？

- エクサ10日間の総演算量：10^24＝10^6Exa
  - 1Exa = 10^(3(k)+3(m)+3(g)+3(t)+3(p)+3(e))=10^18
  - 1EFLOPSだと、10^6秒（11.57日）ということになる。
    - すなわち、1EFLOPSマシンを10日程度占有した時の最大可能な総演算量
- NICAMを例にとる。

NICAM基本データ

- 現行のNICAMgl10L40（dt=30秒）の一日積分
  - ステップ数は？　2880ステップ
  - 格子点数は？　419,430,400（4^11X10 X40)=420M
    - 力学 23PFLOP　：　RK3のため（2＋3＋6＝11）
      - 1ステップ1格子点：19014FLOP
      - DYNSTEP前後で1.20e+7 VI_SMALL_STEP前後で7.68e+6
        - これにより、small step 64% / largestep 36%
        - RK3 2step+3step+6step=11step(small) / 3step(large)
      - 1 small step、1格子点1small ステップあたりの演算量
        - 1100FLOP　X　11
      - 1large step、1格子点あたりの演算量：2280FLOP　X　3
    - 物理 21.64PFLOP
      - 放射　　15.43PFLOP（20ステップ毎）
        - ？？？1回あたり換算15.43x20=308PFLOP
      - 雲微物理 2.93PFLOP
        - ダブルモーメントで倍程度になる可能性あり
        - 清木スキームでどうか？
      - 乱流　　1.75PFLOP
      - 陸面・地表面過程　　0.56PFLOP
    - 合計 44.81PFLOP

おおざっぱに物理過程と力学過程の演算量を分けねばならない？

- 力学過程：1 large step 一格子点あたりの演算量：
  - Small stepは基本的力学過程なので不変：1100 X 11
  - Large step はほとんどがトレーサー移流。
    - 現在では、(u,v,wゴミ), qv, qc, qr, qi, qs, qg　6成分
      - 注：エネルギーは音波にも関係するのでsmall stepで計算
      - 1トレーサーあたり　1140　＝　2280X3 / 6
      - 今後double momentを想定し、乱流TKEも移流させると。nc, nr, ni, ns, ng, tke + 6成分
      - よって、1140 X (6+6)　= 13680
  - まとめよう。
    - Large step : 13,680FLOP
    - Small step : 1100X11 = 12,100FLOP
    - Total : 25,780FLOP
      - 25780/ 19014 = 1.35
      - つまり、トレーサーを加えることにより1.35倍増える
      - 137：121の比率
        - Small time step の占める演算量の比率は、121/257=0.47

- 物理過程
  - 放射過程： 一回当たり15.43PF/144（回）＝107 TF
    - 40層計算の時一回一格子あたりにかんさんすると、254761FLOP
    - 仮定： 1分で一回でよいとする。
      - 計算量は、鉛直格子の何乗で伸びる？
      - 線型に伸びるとする。（清木による実測で）
    - Gl17 dt=0.234375秒 とすると一分間隔で256largstepに一回となる。
  - 雲微物理：
    - 仮定： 1秒に一回でいいとする。
      - 4 large stepに一回。
      - Double moment で計算量は約2倍（清木より）
        - この場合、1回1格子あたり4844FLOP
  - 乱流：
    - 仮定： 1秒に一回でいいかな？
      - 4large stepに一回。
      - 1ステップ、1格子あたりの計算量：1446FLOP
        - u, v, w, T, qv, qc, qr, qi, qs, qg 10変数に対して、
        - 一変数あたりは145FLOP
        - 今後＋6なので 145X16＝2336FLOP
  - 陸面仮定：
    - 仮定： 乱流と同期なので1秒に一回。
    - 1ステップ1カラムあたり ：18518FLOP（ま、ゴミ）

プラン1(ブレークスルー的案)

- グランドチャレンジ的(次世代(ポストペタ))問題サイズ:
  - gl17L500(水平50m格子、鉛直50m格子) 10日積分
    - 意義:全球LES
    - dt=0.234375秒
    - 総ステップ数:3686400= 3.7M
    - 格子点数:86T
    - 演算数: 86TX3.7MX25780=8,203,196EFLOP
  - 力学過程は、計算時間の8割ぐらいと考えたい。(根拠次のページ)
    - 0.8 X 10^6 EFLOP程度力学過程に割り当てられる。
    - 800,000EFLOP

これは、どだい無理である!却下。

gl17L500における力・物比率

- 以下の計算は、カラムで規格化（陸面過程も同時評価のため）
- 力学過程　（256ステップ）
  - 25780X500X256=3300MFLOP
- 物理過程　（256ステップ）
  - 放射一回：　254761X500 = 127MFLOP　(256)
  - 雲微物理:　4844 x 500 X 64 = 155MFLOP (4)
  - 乱流：　2336X500X64＝149,500=75MFLOP (4)
  - 陸面：　18518X64＝1.2MFLOP　(4)
  - 計　360MFLOP
- この計算では、圧倒的に、力学過程が支配的になる。
- 比率は、9.1：1である。
- 物理過程大目に見て、8：2と見積もる。
- つまり、計算リソースの8割は力学にあてがってよい。

プラン2（ちょっと中途半端か？）

- グランドチャレンジ的（次世代（ポストペタ））問題サイズ：
  - gl16L250(水平100m格子、鉛直100m格子)　10日積分
    - 意義：積分期間、格子サイズともにちょっと中途半端か？全球LESの布石にはなるが、、、
    - dt=0.46875秒
    - 総ステップ数：1,843,200= 1.85M
    - 格子点数：10.7T
    - 演算数：　10.7TX1.85MX25780=510,315EFLOP
  - 力学過程は、計算時間の6割ぐらいと考えたい。（次ページ）
    - 0.6 X 10^6 EFLOP程度力学過程に割り当てられる。
    - 600,000EFLOP

演算的量には、まあまあ、イケてる！

gl16L250における力・物比率

- 力学過程　(128ステップ)
  - 25780X250X128=824MFLOP
- 物理過程　(128ステップ) :1分
  - 放射一回:　254761X250 = 64MFLOP　(128に一回)
  - 雲微物理:　4844 x 250X 64 = 310MFLOP (2回)
  - 乱流:　2336X250X64＝37MFLOP (2ステップ一回)
  - 陸面:　18518X64＝1.2MFLOP　(2)
  - 計　412MFLOP

- 比率は、800:400＝2:1である。
- つまり、計算リソースの3分の2は力学にあてがってよい。

ここから、見積もり

- システム： 100TFLOPS X 10000　コア数：1000
- 1ノードのメッシュ数：水平4295K X 鉛直250 = 1.0Gメッシュ

- 0.66 X 10^6秒で、タイムステップ1.85M step進めるためには、1stepあたり、0.36秒で行う必要がある。
  - S: 0.36X 0.47 = 0.17秒（small step）
    - 1 small step 0.17/11=0.0154秒で行う必要あり。（15ms）
  - L: 0.36X0.53 = 0.19秒
    - Large stepはトレーサー移流

- 必要な主記憶： とりあえずペンディング

- 1small stepで、メモリーロードしながら計算すると、
  - スイープしながら計算するとき、
    - 基本変数、rho, T, u, v, w, qv, ql, qi 8変数
      - 6隣接x8＝48変数
    - メトリクス：各方向のdx　5変数
    - 53個　＝　424byte
    - 1100FLOPなので、BF=424/1100=0.4
      - エクサでこれはきつい。
  - 結論：どう考えても、オンチップ計算でないと無理！（ロードしながらは諦める）
    - 基本変数は、ずっとキャッシュに乗せっぱなしとしないと無理。
- オンチップ計算するとなると
  - 1Gmesh X 500=500Gbyte
    - 1コアあたり500G/1000=500MBいることになる。

ここですでにこの規模は無理！！！！**ここで終わり**

却下！！！

プラン3：規模小さく積分長く

- 問題サイズ：
  - gl14L100(水平400m鉛直250m格子) 6か月X10ケース
    - 意義（京コンピュータでのデモランの統計的評価）
      - 水平500mを切る計算においては、convectionの表現が格段によくなる。（真のglobal-cloud-system resolving）。北半球夏の熱帯計算を10ケース行い、季節内振動の再現性評価を行う。
      - 来るべき全球LESへの布石と位置付け。
  - dt=1.875秒
  - 総ステップ数：1,843,200= 83M
  - 格子点数：268G
  - 演算数： 286GX83MX25780=611,965EFLOP
  - 力学過程は、計算時間の7割ぐらいと考えたい。（次ページに根拠）
    - 0.7 X 10^6 EFLOP程度力学過程に割り当てられる。
    - 700,000EFLOP

gl14L100の時の物理／力学計算量比

- 力学過程　(32ステップ1分)
  - 25780X100X32=82.5MFLOP
- 物理過程　(32ステップ1分):
  - 放射一回:　254761X100X0.5 = 12MFLOP　(64に一回、2分に一回)
  - 雲微物理:　4844 x 100x 32 = 15.5MFLOP (毎回)
  - 乱流:　2336X100X32＝7.5MFLOP (毎回)
  - 陸面:　18518X32＝0.6MFLOP　(毎回)
  - 計　35.6MFLOP

- 比率は、82.5:35.6＝2:1である。
- つまり、計算リソース7割は力学にあてがってよい。

ここから、見積もり。

- システム： 100TFLOPS X 10000　コア数：1000
- 1ノードのメッシュ数：水平268K X 鉛直100 = 26.8Mメッシュ
  - 1コアあたりは？　26.8Kメッシュ
- 0.7 X 10^6秒で、タイムステップ83M step進めるためには、1stepあたり、0.0084秒（8.4msec）で行う必要がある。
  - S: 8.4X 0.47 = 3.95msec（small step）
    - 1 small step 0.17/11=0.35msecで行う必要あり。
  - L: 8.4X0.53 = 4.45msec
    - Large stepはトレーサー移流
- 必要な主記憶： とりあえずペンディング

- オンチップ計算する。
  - 基本変数5変数： rho, T, u, v, w, 5変数トレーサー12変数＝17変数
  - ＋メトリクス（(6方向＋2方向)X(3オペレータ)）24変数
    - 24＋17＝41 変数
  - 作業変数として10変数あれば十分か？
    - 計50変数＝400byte
  - 26.8M mesh X 400=10.7Gbyte
    - 1コアあたり10.7G/1000=10.7MBいることになる。

- 通信
  - 考え方： 内部を計算している間にハローの通信もロードストアも済ませる。
  - smallstep毎に発生
  - 基本変数分　5変数
  - 1node: 517X517X100のメッシュ。
  - ハロー領域1として、
    - 517X4X100X5変数X8byte =8.3MB
    - 0.35msecの間に行うとすると（オーバーラップ）、
    - 24GB/sは必要。大目に見て倍の50GB/s
    - レイテンシー：0.35usecぐらい？（1small stepの10^-3程度）
- 主記憶バンド幅：
  - 考えたかた：ハローだけロードストアする。
    - Small step毎
      - 1ノードあたり、8.3MB/0.00035=24GB/sですむ。
    - Large time step
      - 1ノードあたり、517X4X100 X 12X8byte=19.8MB
      - 4.45msecの間に行うとすると、4.4GB/sec必要。
    - 未来永劫のせるとすれば、100GB/sあれば十分。（BF0.001）
    - 1stepでキャッシュを載せ替えるとすれば、8.4msecが計算時間なので、その10分の1でキャッシュに乗せるとして、11Gbyte/0.0084*10=13TB/sec(BF:0.13)
- 主記憶容量：
  - 20GBあれば実は十分。（前主記憶200TBに相当）

100TFLOPS X 10000の場合に要求仕様は、こんなかんじ

- 仮定：
  - 内部計算と袖計算に分ける。
  - 袖計算部分のみ、ロードストア／通信して、内部計算をやる間に、すべて済ませる。
- ネットワーク速度　：50GB/s以上
- ネットワークレイテンシー：　0.35usec以下
- 主記憶バンド幅：
  - 基本変数は、未来永劫キャッシュに乗せておく場合
    - 100GB/s以上（BF=0.001で済む）
  - 1step あたりキャッシュパージするときは、13TB/s以上
    - BF＝0.13ぐらいは必要。
- 主記憶サイズ：　20GB以上
- オンチップメモリサイズ：
  - コアあたり12MB以上（ノードあたり12GB以上）

原理的には、、、

- 思ったほどBFいらなさそう。
- でも、そうつくるには？？
  - 単盤でだらだら書く。
  - 通信オーバーラップなど、面倒くさいことといっぱい。

- そもそもライブラリー化に反する。
- どうしたらいいか？？？
  - コンパイラーにがんばってもらう。
  - プレコンパイラーを作って、プリプロセスしたのちにコンパイルする。
  - ステンシル計算言語を作る。
    - 東工大丸山さん

まとめ

- 地球科学標準ライブラリーの構築
  - 力学、物理どちらも。カップラー。
  - オールジャパンで
  - インターフェースと仕様の統一
    - 誰でもコミットできるように
  - 政治的に大変？！
- 今後のHPC
  - どんどん使いにくい（性能でない）方向に？！
  - なんでか？
    - アプリ側もちゃんと考えるべき。