3.2 ビッグデータの有効利用

3.2.1 計算科学基盤技術の創出と高度化

(1) 課題概要

HPC が貢献しうる社会的課題と対応するシミュレーション技術は多岐にわたる。実世界で起こる現象は現象自体が複雑であり、その因果関係を正確に把握することは難しい。そのため、多くの場合は現実世界のありのままを模擬することはできず、シミュレーションでは対象とする実現象（問題）を理想化して計算することになる。したがって、現実世界で起こる現象とシミュレーションする現象との間には、少なからずギャップが存在する。現実世界の問題の特徴をうまく再現できるような方程式群の形式を選定し、関連するパラメータを決めるプロセスをモデリングと呼ぶ。モデリングは観測や実験結果に対する洞察から得られた知見を基にしているが、シミュレーションの正しさや精度に影響する非常に重要なプロセスである。なかでもデータ同化は、実験や観測データとシミュレーションとを融合し、最適な初期条件や境界条件、その他のパラメータを推定してシミュレーションの不確実性を低減するものであり、実現象を扱う社会的課題では特に重要となる。モデリングに対して、計算結果として得られた巨大な数値データ群から、現象の理解や問題解決の指針となるような有用な情報を得るプロセスをシミュレーション結果の現実世界への投影という意味で、ここではプロジェクションと呼ぶ。このモデリングとプロジェクションは、現実世界とシミュレーションの世界とをつなぐプロセスであり、これらに関わる技術開発と人材育成は、HPC 成果の社会還元に必要不可欠である。

従来、これらの技術はアプリケーションの分野ごとに個別の技術的進化を遂げてきたが、共通部分も多い。例えば、現実世界の影響をモデリングに適用する技術であるデータ同化は、気象をはじめ、石油掘削、制御、ものづくり、創薬や分子シミュレーションなどの分野でも研究・活用されている。また、結果の解釈の一助となる可視化技術もさまざまな研究領域で利用され、個別の現象に適した表現方法が研究されている。今後、例として挙げたデータ同化と可視化、またその他の基盤技術の共通部分を横断的に統合し、情報共有を進め、技術を進化させることの重要性はますます増加する。加えて、今後の大規模化、多様化、リアルタイム化などに特徴付けられるビッグデータの扱いも共通的な課題となる。本節では、データ同化、可視化、ビッグデータ処理、知識処理を代表的な共通基盤技術として述べる。

大規模シミュレーションにおけるデータ同化は、数値天気予報のための初期条件を作成する際、さまざまな観測データと数値天気予報モデルとを統合的に扱う手法として気象分野を中心に発展してきた。近年、データ同化の役割は数値天気予報に限らず、幅広いシミュレーションに応用されてきている。データ同化で用いられている数学や統計数理は、最適化問題、制御工学、状態空間モデルなどいろいろな分野と類似したものであり、データ同化に関する計算科学はさまざまな分野に共通する基盤的技術としての重要な役割を果たしている。特に、地球環境観測データは種類も多様で、量も豊富ないわゆるビッグデータである。時空間的に変化が大きく、予測可能性に限界がある気象・気候という難しいシステムを対象としたデータ同化システ

ムは、実験データと数値モデルの統融合によって高付加価値データセットを作成するための実例としてさまざまな分野への応用が可能である。

可視化とデータ処理は計算結果の解釈を助ける共通的な技術と分野ごとの知見を利用する固有部分が混在する。共通部分の大きな課題は、急ピッチで規模が拡大するビッグデータに対応したシステム化である。シミュレーションの規模が大きくなると、必然的に分散ファイルとなるため、アプリケーションレベルでのファイルを管理し、シミュレータとポスト処理でスキームを共有することが必要となる。汎用的な枠組みとしてファイル交換や管理を一元化するフォーマットやライブラリが提案されているが、多機能で複雑なライブラリは、今後新しく出現するアーキテクチャの上で最適化することが難しいという課題を抱えている。また、高速化する演算装置と低速な記憶装置との間の処理速度の乖離は大きくなる一方であり、アプリケーションの性能を考慮すると低速なファイル入出力を抑制するほうがよい。この点で、計算と同時に可視化・データ処理を行うデータストリーミング手法の検討が必要となっている。

ビッグデータへの対応は、大規模シミュレーションを実行するうえで、まさに共通的な課題である。この中には、新しい知識の原石が埋没しており、学術や防災をはじめ、設計、医療・創薬、サービス等の分野に有用な知識を発掘し提供することが重要である。ビッグデータの特徴は、大量、多様、リアルタイムという特徴を持つ。加えて、ノイズがあり、全体として異なるタイプのデータの集合であり、多様な構造を持つ場合もある。しかしながら、その多くのデータは構造化されておらず、計算量が指数関数的に増大し処理が困難となる問題点がある。観測や実験系、シミュレーションから生成される多様なデータに対して、安定的かつ効率の高い、汎用的なデータ処理の枠組みの研究とツール蓄積が必要となる。

プロジェクションプロセスの中でも、シミュレーション結果から得られる知見を非専門家に対してわかりやすく説明するアウトリーチは、今後ますます重要になってくる。例えば、天気予報はシミュレーションの結果を基に、一般層へわかりやすく説明する役割を気象予報士が担っているが、他の分野では研究者自身がアウトリーチの役割も兼ねていることが多い。アウトリーチの仕組みや、そのための人材の育成の点も含めて検討していく必要がある。シミュレーション結果の活用という観点からは、データの中に潜む有益な知見を引き出す技術の高度化が重要になる。これまでは、人の経験やある程度定式化されたプロセスにより情報が抽出され、知識へと昇華されてきた。ビッグデータの時代には、規模とスピードの点から、機械学習やデータマイニングによる自動的な知識形成の仕組みが必要であり、ボトムアップ的なエキスパートシステムや人工知能技術が役立つ。知識処理については、これまでもクラスタリングやパターンマイニング、グラフ分類、類似性検索、ベイジアンフィルタ、機械学習など、多くの技術が創出されてきた。ビッグデータの中から知識という抽象的なレベルの知見を得るには、データマイニングの前処理、本分析、後処理の各プロセスをうまく設計する必要がある。現時点では、汎用的なプロセスはなく、パターン抽出の観点は人間が指示しないといけないため、データから価値を発見するプロセスの高度化に取り組む必要がある。

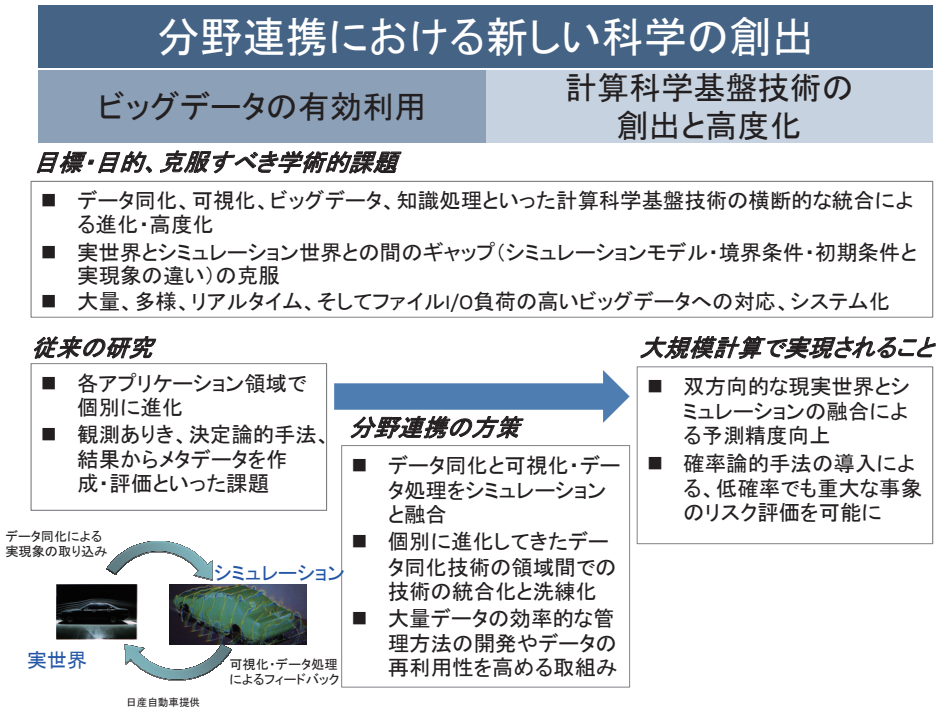

図 3.2.1-1　計算科学基盤技術の創出と高度化

(2)　サイエンスの質的な変化と長期的目標

計算機と計算科学技術が発展した現在でも、シミュレーションの問題設定と初期条件・境界条件などには実現象との乖離があり、結果として得られるデータの解釈について専門的な知識や経験を必要とする。このため、シミュレーションの結果を直接そのまま社会へ反映することは難しく、実社会とシミュレーションの結びつきには改善されるべき点が多い。HPC 技術の社会的意義の点からは、双方のフィードバックを実現し、相互の結びつきの深化が期待される。そのためには、以下に述べるサイエンスの質的変化を意識して、共通基盤技術を進展させることが重要である。同じ計算技術であっても個別の分野で独立に開発されてきた技術は多い。今後は、分野の垣根を越えて統合化できる基盤部分を強化し、熟成することにより、HPC 利用技術全体としての底上げが促進される。

モデリングプロセスにおける共通基盤技術としてデータ同化が重要だが、そのサイエンスの質的な変化として着目すべき点として、以下の 2 点が挙げられる。

① 従来は観測をシミュレーションに生かすことが主体であったが、将来はシミュレーションに基づいて効果的な観測システムを設計し効率的にデータを収集するという逆向きの応用が行われるようになり、現実世界とシミュレーションとの融合が双方向的なものとなる。

② 従来は決定論的であったシミュレーション技術は、将来の計算機資源の増大により確率論的な表現が実現され、より詳細な不確実性の評価が可能となる。これにより、低い確率であるが重大な影響をもたらす事象の予測、つまり「想定外」を減らすこととなり、シミュレーション結果の信頼性および可用性が高まる。一方、データ同化技術

にとってはシミュレーションの確率表現が高度化することで、統計数理手法のブレークスルーが期待される。

プロジェクションプロセスでは、シミュレーションで生成されるビッグデータが、計算プロセス自体を変えていく。従来は計算機資源の制限や計算手法の模索のため、固定的な計算プロセスを主体とするアプローチであった。今後、生成されるデータが次の処理プロセスを誘導し、一連の計算プロセスが実行されるデータドリブンへとパラダイムが変化していく。このプロセスの変化は、シミュレーションの結果を解釈し、有用な知識として社会還元する要求にも合致する。すなわち、知識処理技術を用いてシミュレーション結果を自動的に知識化するプロセスは、データドリブン的な処理に移行しつつある。このようにプロジェクションプロセスは、HPCの結果を社会に還元する技術と捉えることもできる。

更に、上述したモデリングとプロジェクションをシミュレーションと融合し、シミュレーション利用技術の高度化を加速することが、サイエンスの方法論の革新とそこから生み出される科学的な発見につながる。

このように、計算科学技術の進化により、得られるデータとそこから抽出される情報の質的な向上が見込まれるが、一方で、情報技術やデバイスの進化が HPC 技術や使い方の変化をもたらすだろう。例えば将来的には、携帯端末のサービスの一環として HPC 利用が想定され、「いつでも、どこでも、だれでも」というユビキタスなキーワードが表すようなオンデマンド的な利用が、HPC 技術の社会活用を大きく変えるであろう。より具体的な例を挙げると、緊急時に有害物質の分布予測が知りたければ、稼働中のジョブをすべてキャンセルし、最優先で物質拡散シミュレーションを実行する。それに先立ち、実際の初期条件と境界条件を収集し、データ同化を行ってシミュレーションを実行する、その全体の流れを適切にスケジューリングし、可能な候補を示して、計算指示者の実行判断を支援する。計算した結果から、特徴的で重要な点を自動抽出し、専門家向けの情報提示から、行政の意思決定者向け、更には一般市民といったエンドユーザ向けの情報提示までつなげる。このようなシナリオは、HPC 技術、ネットワークインフラ、ユーザインターフェイス、スケジューリング、データマイニングなどさまざまな情報処理技術をベースとして積み上げて初めて成立する。

HPC 技術が単に科学者のツールとして利用されるだけでなく、これまで利用促進が進んでいなかった領域への浸透し、得られた知見が広く一般社会へと還元されることが重要であり、また価値を高めていく手段でもある。HPC は科学者や技術者のツールであるという殻を破り、広く一般層への門戸を開く技術的な道筋をつける努力が必要である。

(3) コミュニティからの意見

大規模データ同化は、これまで気象・気候分野での技術的発展が顕著であった。しかし、設計における制御、石油掘削、分子シミュレーションなど他の分野においても適用が図られている。各アプリケーション分野のコミュニティに閉じていたデータ同化技術は、現在、コミュニティを越えて共通化する動きが起きている。技術の統合化とその過程で得られる機能や性能の向上、ロバスト性、可用性などが期待される。

計算機科学分野でもビッグデータ処理、マイニング、大規模データ処理系の技術的な融合が始まっている。また、可視化分野では、これまでの要素技術だけではなく、大規模データの扱いが必須であり、アプリケーションレベルのインタフェースが重要との認識が高まっている。将来的には、ファイルを介さない計算と同時の可視化処理や可視化と分析の融合技術の実現が知識創出を大きく促進する。

(4) 必要な計算機資源

データ同化に関しては、数値解法としていくつかの手法が提案されているが、いずれの手法においても、必要となる計算機性能は数値モデルの繰り返し計算を何回行うか、もしくは予報誤差情報を得るためにいくつのアンサンブル計算を行うかによって決まっており、おおよそシミュレーションを一つ実行する計算量の数十倍から数百倍の計算量であると見積もることができる。更に、データ同化特有の計算要求として、観測データとの比較を行い、最適化問題を解くために、シミュレーション結果を詳細に保存する必要があることが挙げられる。シミュレーションの結果がメモリ上に保存できる量の場合にはメモリ性能が、メモリ量が足りない場合にはディスクに保存する必要があるためにディスク I/O（入出力）の性能がボトルネックとなる。メモリ、ディスク I/O の量を高分解能の大気海洋結合モデルを用いたデータ同化システムにおいて見積もると、数百 PFLOP 程度の計算を 1 時間で行う場合（100TFLOPS 程度の実行性能）に、1TB 以上のメモリ量と数十～数百 TB のディスク I/O が行われる。すなわち、計算の実行性能を向上させるためにはメモリ性能、ディスク性能の向上が不可欠である。

可視化については、表示デバイスの要求から、これまでより大解像度（8096×8096 画素 =256MB/フレーム）の画像が生成される。このため、30fps 程度の場合、実効ネットワーク帯域として双方向 15.6GB/s 程度が必要になる。

また、運用技術に関しては、ポスト処理の段階ではインタラクティブ性が重要であるため、スケジューラのインタラクティブ予約や、ある計算プロセスに対して別プロセスからのメモリアクセスなどの機能実現が要求として挙げられる。

課題	要求性能 (PFLOPS)	要求メモリバンド幅 (PB/s)	メモリ量/ケース (PB)	ストレージ量/ケース (PB)	計算時間/ケース (hour)	ケース数	総演算量 (EFLOP)	概要と計算手法	問題規模	備考
並列レンダリング	200	61	0.8	10	0.5	1	360	ボリュームレンダリング（レイキャスト、ファイルベース）		対象によって問題規模等は異なるため、典型的な例で概算
並列レンダリング	200	61	2	1	0.5	1	360	ボリュームレンダリング（In situ）		対象によって問題規模等は異なるため、典型的な例で概算
データ圧縮	500	25	8	10	0.5	1	900	POD圧縮（ファイルベース）		対象によって問題規模等は異なるため、典型的な例で概算