Vol.2011-HPC-129 No.3
2011/3/15

一次データを保存しない大規模科学計算の可能性

高 見 利 也[†1]　戸 田 幹 人[†2]　福 水 健 次[†3]

本研究報告では、大規模応用計算によって生成されるデータを最小化するための方法を検討する。超並列計算機の導入で従来では扱うことの出来なかった大規模問題も研究対象となりつつあるが、演算の高性能化に比較して相対的に低い入出力やストレージ性能のため、大規模科学計算による仮想実験では大量に生み出される数値データの処理が問題になる。ここでは、可逆な微分方程式に基づく時間発展計算によって新規の現象を発見するための物質系シミュレーションを対象とし、大規模計算による一次データ量の削減とトレードオフの関係にある計算の再現性について検証する。また、部分系のデータ保存という観点からのデータマイニング手法についても検討する。

Feasibility of Large-scale Scientific Computing without Saving Results

Toshiya Takami,[†1] Mikito Toda[†2]
and Kenji Fukumizu[†3]

In this report, methods of minimizing numerical data obtained by large-scale applications are investigated. Recent development of massively parallel machines enables us to challenge new problems in the field of computational sciences. However, it reveals also relatively low I/O and storage performance compared to the high-performance in floating-point calculations. The problem is how to manage such large amounts of data produced by scientific calculations as virtual experiments. For numerical simulations based on time-reversal differential equations, several data-reduction methods are analyzed, where the reduction conflicts with the possiblity of reproducing numerical results. It is also analyzed methods of numerical-data mining as a method of data reduction of partial systems.

†1 九州大学, Kyushu University
†2 奈良女子大学, Nara Women's University
†3 統計数理研究所, The Institute for Statistical Mathematics

1. はじめに

演算機コアの実装方法の進歩や並列数の増大で、単位時間あたりの演算性能は順調に伸びている。これには、Top500 リストなど演算性能を中心にした評価が世界的に注目を集めていることにより、次々と新しい超並列マシンが開発され続けていることにもよる。通常、これらのマシンには多額の予算が投入されるため、情報科学以外の様々な面での成果が求められ、最近では、特に計算科学の領域での学術的な成果も必要とされている。つまり、規模と性能において世界的に上位に位置づけられる計算機は、単に動くことだけでなく、科学や技術の研究などのために有用であることを実証しなくてはならなくなっている。しかし、大規模科学研究のために使う場合、現在作られている超並列計算機の性能は必ずしもバランスの良いものとなっていない。突出して高性能になっている演算機コアの性能に対して、データの保存と移動に関する性能が見劣りするようになってしまっているのである。

Top500 にリストされるような計算機は、天文台などの天体観測施設や素粒子や原子核の実験を行う加速器などの大規模実験施設と比較されることが多い。これらの実験施設は、いずれも観測や実験によるデータが大量に生み出されることが知られているが、「京」での仮想的な数値実験でも、一般的には計算結果として大量のデータ出力が予想される。数値実験にも様々な目的や手法があるが、ここでは、大規模な数値計算の中から天体観測や加速器実験と同様の未知の現象を探求するというタイプの計算を対象とする。具体的には、何らかの

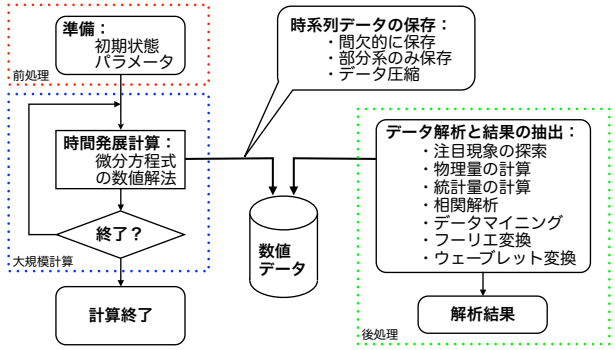

図 1　科学シミュレーションによる時系列データ解析の概念図。時系列を求めるための大規模計算は並列化されており効率的な実行が可能であるが、解析のための後処理まで含めた全体で見ると、時系列の数値データの保存と移動に時間がかかる。

初期条件や境界条件のもとで時間発展計算を行い、その中から興味ある新しい現象を見つけて詳しく解析することを目的とする。この場合、大規模計算で得られる時系列データを保存するが、通常はこの部分までが一次計算に位置づけられ、データの解析は後処理として改めて実施されることとなる。しかし、大規模観測・実験装置と同様に、データを解析して初めて新しい現象や科学的知見が得られるのであるから、データの解析過程までの一連の作業全体を効率よく実行することが重要である。

図 1 に示すように、解析まで含めた全体として眺めると、大規模計算になるほど増大する数値データの保存と移動が問題になる。通常のシミュレーションでは、前処理として初期条件などを確定する計算の後、高速計算機を用いた時間発展計算を行い、その出力を後処理として解析し必要な形式に加工する。数値演算の部分は、様々な手法による並列化によって効率的に実行されるようになってきているが、この部分が効率化されればされるほど単位時間当たりに生成されるデータ量が増えることとなり、データ処理部分が律速となることが予想される。これまで、入出力性能向上のための努力が払われてきているが、ここではデータ量を削減するという方向で、シミュレーションと結果の解析全体の効率化手法に関して検討することとする。

まず 2 節では、ミクロ系の可逆なダイナミクスの例として分子動力学と量子力学の時間発展計算、代表的な時系列計算の例として数値流体計算に関して、数値計算によって生み出されるデータ量を見積もることにより、律速になるのがデータ処理であることを明らかにする。この状況を根本的に解消することは難しいため、第 3 節では、現在の状況でバランスよく大規模な科学計算を実施するために、データ量を削減するための手法を検討する。データの削減の手法として、保存間隔を積極的に延ばす方法を導入し、この有効性をデータの再現性という面から検討する。また別の方法として、最も注目する部分系のみをデータ記録の対象とし、それ以外の部分の時系列データは保存しないという方針を検討する。この手法を発展させるものとして、明確に対象系と部分系が分離できない場合の処理で重要となる数値データのマイニング手法についても検討する。最後に、全く異なる観点から、そもそも計算によって得られた一次データを全く保存しないインタラクティブシミュレーション手法の可能性についても検討する。

2. 演算性能と出力データ量

各時間ステップで M バイトのデータを記録することとし、1 ステップの時間発展計算に要する浮動小数点演算数を N 回とすると、実効性能 P FLOPS の計算機で出力されるデータ量は、毎秒 MP/N バイトということとなる。計算機のハードウェアとしては神戸に建設中の「京」(SPARK64VIIIfx 8 core, 128 GFlops) を想定する。以下では、分子動力学、量子ダイナミクス、数値流体力学の三種類に関して、出力されるデータ量を大まかに見積もることとする。

2.1 分子動力学

ミクロな多粒子系に対して広く利用される手法に分子動力学計算 (molecular dynamics, MD) があるが、領域分割することにより非常に大規模な計算が並列に実行できる。「京」での大規模計算では、1,000 万粒子の計算が実施できると言われているが、この場合、1 ノードあたり約 120 原子の計算が行われることとなる。

二体相互作用近似の時間発展計算では、1 階の $6n$ 次元連立常微分方程式

$$\dot{\mathbf{q}}_j = \frac{\mathbf{p}_j}{m_j}, \qquad \dot{\mathbf{p}}_j = -\nabla_j \sum_{k\neq j} V(|\mathbf{q}_j - \mathbf{q}_k|) \tag{1}$$

を数値的に解くこととなる。ここで、$\mathbf{q}_j$、$\mathbf{p}_j$、m_j は、それぞれ、j 番目の粒子の座標、運動量、質量を表し、$V(r)$ は二体相互作用近似のもとでのポテンシャル関数である。記録するべきデータは、各時刻の全原子の座標 $\{\mathbf{q}_j\}$ と運動量 (速度) $\{\mathbf{p}_j\}$ の値であり、そのサイズは原子数 n に対して $M = 6n \times 8$ バイトと計算される。短距離力だけに従う時、時間発展の 1 ステップで最も計算量の多いのは二体力の計算部分で、n 粒子に対して $O(n^2)$ となる。レナード・ジョーンズポテンシャルでは $V(r)$ は r の有理関数となり、30 回以下の浮動小数点演算で特定の二原子間の力を計算することが可能であるが、境界を越えた原子との相互作用などその他の演算も含めて仮に 100 回の浮動小数点演算が必要であるとすると、$n = 120$ に対する二次精度の時間発展 1 ステップについて、

$$N = \frac{n(n-1)}{2} \times 100 = 7.14 \times 10^5 \tag{2}$$

回の演算である。ピーク性能比 1 割程度の実効性能 $P = 12.8 \times 10^9$ FLOPS で計算すると、

$$\frac{MP}{N} = 6 \times 120 \times 8 \ \times \ 12.8 \times 10^9 \ / \ 7.14 \times 10^5 = 1.03 \times 10^8 \tag{3}$$

バイト、すなわち、毎秒約 100MB のデータ出力ということとなる。

この計算を「京」全体での実施を想定すると約 80TB/sec のデータ出力となり、全データの保存を目指そうとすると、分子動力学計算のように保存データ量が比較的少ない問題でも、既にデータ処理に問題があることがわかる。実は、短距離力だけの場合には、大幅に粒子数を増やすことが可能である。その場合、粒子数の二乗で計算量が増えることでデータ保

存の間隔は長くなるが、粒子数に比例してデータ量は増えることになる。単位時間当たりのデータ量は粒子数に反比例する形になるが、同じ時間ステップだけ計算を行う場合にはデータの総量は同じである。現実には、短距離力だけの分子動力学計算を行うことはまれで、電荷の相互作用 (クーロン力) などの長距離力を考慮することが多く、1 ステップ当たりの計算量は 2 桁から 3 桁増えることになる。それでも、データ量の削減を実施しない場合、数十分から数時間の長時間計算でデータ量の限界に到達する。

2.2 量子ダイナミクス

もう一つの例として、量子系の時間発展計算を考察する。ここでは、化学反応過程を求めるための電子や陽子の波束動力学などの例として、有限基底上で時間発展をする量子ダイナミクスを考える。よく知られているように、有限自由度の孤立量子系では、ハミルトニアンを数値的に対角化すれば、時間発展は位相の回転だけとなり、時間発展計算が不要になる。一般には、外部自由度との結合や、古典パラメータ (例えば電場や磁場) が変動する場合など、非孤立系が興味の対象となることが多いため、ここでも、古典的な電場との相互作用を含んだ量子力学系を対象として時間発展計算を実施することとする。

具体的に話を進めるために、ここでは、短パルスレーザーで周期的に駆動される少数自由度系をモデル化した kicked rotor に、連続的に変化する外場 $\varepsilon(t)$ (例えば古典電場などに対応する) が作用している場合を考え、

$$H(q,p;\varepsilon(t)) = \frac{p^2}{2} + K\cos(q)\sum_{n=-\infty}^{\infty}\delta(t-n) + \varepsilon(t)\cos(q+\delta q) \tag{4}$$

というハミルトニアン[1] に従って時間発展させることとする。ここで、q は空間座標で 1 次元の周期的な空間 (例えば角度) を表し、p は運動量で q 表示 (シュレーディンガー表示) の量子力学では、q の微分演算子として与えられる。K はレーザーパルスの強さを表し、対応する古典力学系では $K \approx 1$ より強くなると、不安定な運動[⋆1] に転移するが、量子力学系では指数関数的に局在した状態が一般的で、明確な転移は存在しない。

この量子系の時間発展を計算する場合、相互作用を表す q の関数と自由粒子の運動エネルギー項 $p^2/2$ からそれぞれ時間発展演算子を構成し、これらを繰り返し波動関数に作用することとなる。波動関数を座標表示で表す時、q の関数による相互作用項の時間発展は対角演算子の演算 (計算量は $O(n)$) となるが、運動量による時間発展項は、$O(n^2)$ の演算が必要である。そこで、波動関数の表示を q から p へと変換して、対角的な運動量演算子を作用し、再びもとの q 表示に変換する方法を採用すると、FFT の利用により計算量は $O(n\log n)$ 程度になる。これをもとに時間発展計算のステップ当たりの計算量を見積もることとするが、係数まで含めて正確に見積もることは難しいが、次数として分子動力学に比べて計算量が少ないことは明らかである。各ステップでの保存データ量は $M = 2n \times 8$ であるから、仮に 1 ステップの計算量を $N = 10(n + n\log_2 n)$ とし、ノードあたりの実効計算性能をピーク性能比 1 割の $P = 12.8 \times 10^9$ FLOPS とするとき、単位時間あたりのデータ量は、

$$\frac{MP}{N} = \frac{16n \times 12.8 \times 10^9}{10(n + n\log n)} = \frac{20.48}{1 + \log_2 n} \times 10^9 \tag{5}$$

となる。例えば、$n = 1000$ に対して 2GB/sec 程度となる。分子動力学の場合には問題サイズ n が大きくなれば、逆に単位時間当たりのデータ量は減少したが、量子系の場合には計算量の増加があまりないため、データ量はあまり変わらないことになる。このように試算すると、高性能なノード上で量子ダイナミクスの計算を実施する場合には、各時間ステップに出力されるデータ量はハードウェア的な性能の限界に近い値になってしまう。

ただし、ここでの見積もりは、並列化を考慮したものではないことを注意しておく。既に見たように FFT による変換をステップごとに繰り返し行う部分が律速になるが、非常に大きな状態ベクトルを全ノードで領域分割して保持している場合には、超並列計算での実効性能の低さから、単位時間当たりのデータ量はさらに 1 桁以上少なくなる。これを考慮しても、8 万ノード全体では 16TB/sec という非常に大きな値となるため、何らかの方法でデータの削減を行うことが必要になる。

2.3 数値流体力学

可逆な時間発展方程式による問題ではないが、保存するデータ量の多い問題として流体計算についても検討しておく。この場合には、計算で得られた時系列のすべてのデータを保存するのは非現実的である。実際、流体の応用計算では、時間発展計算の時間ステップはクーラン数などに従って十分に小さく取るものの、それらすべての計算結果を保存する必要はない。では、どの程度の保存回数までなら耐えられるのかを試算しておこう。「京」の 1 ノードでユーザーが利用できるメモリーは 10GB あまりである。ここに領域分割された空間の時系列データを保存するとき、非圧縮流計算の場合、各メッシュ点毎に流速ベクトルと圧力の合計 4 成分の倍精度数値を保持できることが必要である。場の量を保存するだけなら、効率化を図ることで 500^3 よりは多くのメッシュ点を取れるが、1000^3 は不可能である。

流体計算の大部分はいわゆるステンシル計算であるため、陽的な時間発展計算を行う場

⋆1 運動量 p の方向への拡散。つまり、非常に激しい運動をする状態に加速される粒子が生ずる

Vol.2011-HPC-129 No.3
2011/3/15

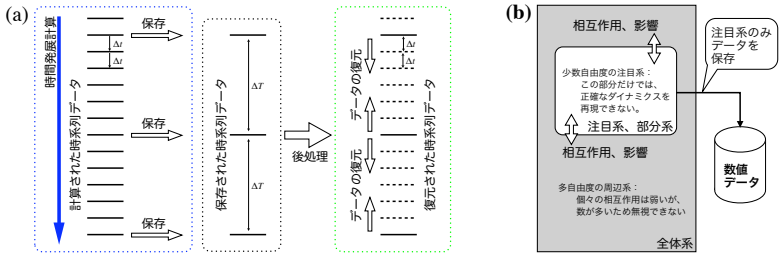

図 2　保存するデータを削減する方法。(a) 間欠的にデータを保存する。後処理で、データ復元のための計算が必要になる。(b) 部分系だけのデータを保存する。全体系の時系列データの復元は不可能。

合には、各メッシュ点当たりの計算量を見積もることも可能である。しかし、計算手法によって大きく異なることや、陰的解法の場合には収束条件などに依存して計算時間の見積もりが困難なため、ここでは、データ量からの見積もりとする。各時間ステップでノードあたり 10GB のデータの保存が必要になるとき、全体ではそのノード数倍のデータ量になるため、「京」の場合は毎回 800TB という非現実的な値となる。これはメモリー上のデータの全保存を行おうとしているようなものであるため、非現実的になるのは当然である。では、仮に各ノードで 100^3 メッシュの計算を行うこととすると、保存する場の量はノード毎に 32MB/回である。これはそれほど大きな値でないように思えるが、これでも 8 万ノード全体では 2.5 TB/回程度となる。

以上のように見てくると大抵の時間発展計算においては、計算量や計算速度ではなくデータ量とデータのアクセス速度が全体の計算規模を決定することになる。このような場合にも、部分系だけに注目して保存する方法や、空間方向に関して大きなスケールのデータだけを抽出するなど、いくつかの手法が考えられる。次節からは、このような場合も含めていくつかのデータ削減手法を検討する。

3. 保存データの削減と再現性の確保

前節で試算したように時系列計算では大量のデータが生成されるが、そのデータをそのまますべて保存するというのは現実的でない。実際の応用プログラムでも、ある程度の間隔を空けて保存する形になっているが、ここでは、データを削減するためにこの手法を積極的に利用することを考える。

以下では、科学的な目的のために数値実験を実施する時に最も重要な再現性について述べ

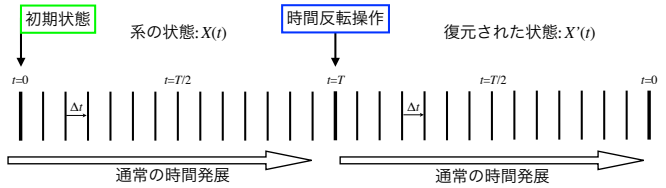

図 3　時間反転によるデータの復元可能性の検証実験。時間発展方程式が可逆な場合は、原理的には $X'(t) = X(T)$ となるはずだが、数値表現が有限桁であるために、数値計算では厳密には成立しない。

た後、再現性の観点から間隔を空けて保存する手法の適性に関して、分子動力学と量子力学、流体力学の各計算手法を考察する。次に、別の方法として、部分系のみのデータを保存する方法、データの解析を数値計算と同時に実施する方法を考える。最後に、全く別の方向として、数値計算による一次データを全く保存しないインタラクティブシミュレーションと適応的計算手法を検討する。

3.1 再現性の要求

科学実験が第三者の検証に耐える必要があるのは、実験結果にある種の一般性が求められているためである。通常の実験の場合、科学理論に照らして理解可能な場合には、個別の実験について再現性や検証可能性が問題になることは少なく、逆に、現象の原因や仕組みに関して報告者本人から納得できる解説がない場合には科学論文として認められないこともある。その意味では、報告される現象が全く新規であっても実際に再現実験による検証が行われることはまれであるが、どのような場合でも検証可能性を保証することは基本的な要求である。仮想実験として実施される数値計算においても同様に、報告した内容の検証として、科学理論に照らして理解でき、かつ、再現実験が可能なことが必要である。

ここでは、新規の現象を発見するためのシミュレーションにおいて、再現性という観点から何が必要かを考察しておく。現在行われる数値計算では様々な近似や高速化の手法が組み合わせて構成されており、全く同じ計算を再現することは難しい。特に、独立に開発されたプログラムによって実施する場合、様々なパラメータの設定を同一にしなくては厳密に同じ計算は出来ない。例えば、時間発展時に外力としてランダム力を与える計算や、初期状態として乱数列から設定される配置を使う場合に、厳密に同じ計算を再現するためには同じアルゴリズムと seed から生成される乱数を利用することが必要になる場合もある。

新規の現象の報告に際して、論文中にこれらすべてのパラメータを記述するわけではないため、公開された情報からだけでは現実的には再現が難しい場合は多いが、一般に、何らか

の決定論に従ってシミュレーションが実施されている以上、原理的に不可能であるということではない。現実的には、厳密な再現を可能にするために必要な情報の量はどの程度かという点で、結局データ量の問題となるが、実験の検証を目的とした再現においては、背後にある科学理論に照らして問題のない部分は、必ずしも厳密な再現が求められるわけではないため、この点でも問題となることはない。

しかし、以下で検討するデータの削減手法を利用して保存されたデータからの復元の場合、科学的検証のために必要とされる実験の再現性よりも復元に対する要求は高い。積極的にデータを削減する場合、解析対象のデータまで省く可能性があるため、解析時に必要な精度で復元できなくてはならない。科学的な結果だけが再現されれば良い場合に比べて、数値を元通り復元するのはより精密な計算が必要である。逆に言えば、削減されたデータからの復元を保証する数値計算では、科学実験として要求される再現性を満足するものとなることが保証される。以下では、削減されたデータから再計算して数値を復元する際の精度を、時間反転操作による再現性のテストを通して検証する。

3.2 時系列間隔の拡大による削減と復元

通常、時系列データを保存する場合、何ステップ毎のデータを保存するかを入力ファイル等で指定し、プログラム側ではそれに応じて間欠的に保存する形になっている。これは主に、計算精度を確保するために短い時間ステップが採用されている場合などにおいて、興味のある現象を記録するための時間ステップがあらかじめ設定できるためである。

ここではデータ削減という目的のために、解析に必要な間隔を超えて保存間隔をのばすことを考える。例えば、保存間隔を 1,000 回に一度だけとすれば、すべての時間ステップのデータを保存する場合に比べて、データ量は 1,000 分の一になり、非常に効果の大きい方法である。この時、保存されたデータからすべてのデータを復元できることが必要となるため、それぞれの計算について復元可能性を検証する。間欠的に保存したデータからもとの時系列を復元する時に、時間反転対称な微分方程式に従う問題では、図 2(a) に示すように正方向の時間発展によるだけでなく逆方向にも復元することが可能であるが、ここでは、時間反転時の復元精度を調べることにより、削減されたデータからの復元可能性を検証する。

3.2.1 分子動力学データの復元

まず分子動力学計算について考える。古典多粒子系の計算は、一般に初期値や誤差に敏感な計算である。そのために、後日同じデータを再計算により求めることが保証されないと、途中のデータを間引くことが難しい。ここでは、どの程度までなら間引くことが可能なのかを見積もるために、分子動力学の時間反転実験の結果を見てみることとする。

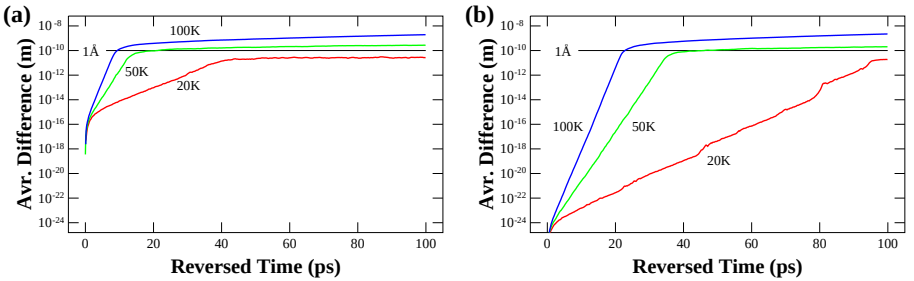

図 4 分子動力学計算の時間反転時の誤差を、一粒子平均の軌道の乖離距離として表示する。(a) は単精度、(b) は倍精度表現による結果。

外部と相互作用のない孤立系では、全エネルギーが保存し、時間反転に対して可逆な常微分方程式に従うが、現実には有限精度での数値表現に起因する誤差が蓄積するため、ある時点で反転させても、厳密に同じ軌道を逆にたどることはできない。図 4 は、実際に分子動力学計算で、時間反転した場合の軌道の誤差を全粒子の平均で表示したものである。ここでの計算は、Ar 510 原子のクラスタを、レナード・ジョーンズポテンシャルによる相互作用で時間発展させたものである。1 ステップの時間間隔を $\Delta t = 1$ fs としているため、全体の 100 ps の時間発展では 10^5 ステップの計算となっている。ベルレ法 (二次精度の時間発展) を使用。乱数で生成された状態から速度スケールで温度制御 (10ps) した後の状態を初期値とし、100ps 時間発展した時点で速度を反転した。ここで表示している誤差は、軌道の位置座標の差を一自由度当たりで平均した

$$R(t) = \sqrt{\frac{1}{n}\sum_k \left|q_k(t) - q'_k(t)\right|^2} \tag{6}$$

であるが、数値表現によらず、反転する時間に対して指数関数的に誤差が増えている様子が読み取れる。このうち、20K は固体の状態で、各原子の可動距離がポテンシャルの底の狭い範囲に限られているため、誤差が蓄積しても 1Å まで広がらないが、液体状態 (50K) や気体に蒸発しつつある状態 (100K) では、全く別の場所へ移動することが可能なため、軌道の誤差もクラスタサイズ程度まで広がることとなる。これは本来、常微分方程式の可逆なダイナミクスに従って、全く同じ軌道を戻ってくるはず ($R(t)$ は常にゼロ) のものである。

誤差の大きさは系の状態によっても変わりうるが、この場合の計算では、倍精度表現でせいぜい 10 ps (1 fs × 10,000 ステップ) 程度の時間しか、事実上可逆であるとは見なせない

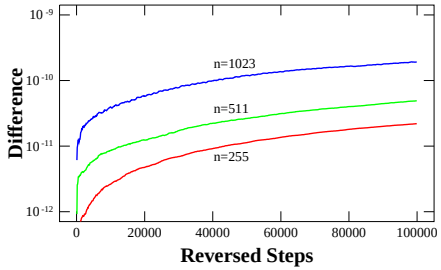

図 5　量子系の可逆性の数値実験。横軸は逆向きの時間発展のステップ数。連続外場を印可された kicked rotor ($K = 5$) を対象とし、kick 間隔は 100 ステップに一回、外場 $\varepsilon(t)$ はすべての時間に作用させている。

ということになる。時系列データの保存間隔をこれ以上離すと、一度は計算した時系列が容易にはつながらないものになるということになる。つまり、間隔を広げることでデータの削減を目指しても、後日、詳細な時系列が必要になった時に、同じデータの復元が難しくなるということを意味している。それでもこの場合には、10,000 ステップ程度は間隔をあけても大丈夫だということが保証されるため、データ量を 10,000 分の 1 程度まで削減することが可能である。

3.2.2　量子ダイナミクス時系列の復元

量子力学に従うダイナミクスも、電磁相互作用のない場合には時間反転に対して対称なシュレーディンガー方程式に従うため、原理的には可逆である。ここでは、ハミルトニアン (4) に従う時間発展の、数値的な可逆性を調べることにする。具体的には、分子動力学計算の場合と同じで、あらかじめ正方向に時間発展した結果を記録しておき、最終時刻から逆に時間発展させた時の波動関数の誤差を求める。ここでは、外部系との相互作用を取り入れるため、あらかじめ与えられた外場 $\varepsilon(t)$ に対して二次のシンプレクティック積分法による時間発展を計算する。図 5 に示す誤差は、波動関数の差から計算した、

$$R(t) = \sqrt{||\psi(t)\rangle - |\psi'(t)\rangle|^2} \qquad (7)$$

である。ここでの計算対象の量子 kicked rotor では、パラメータ K の値や $\hbar$ の値によって系の挙動が変わる可能性があるが、誤差が線形に増えていく様子は基本的に同じである。つまり、分子動力学のように指数関数的な誤差の増大がないため、かなり大きく間隔を取って保存しても、データの再現が可能である。今回の検証のステップ数を超えるが、10^6 ステップ程度の広い間隔をとってもデータの再現性には問題が生じないと考えられる。

最後に、時間発展計算の精度についてコメントをしておく。分子動力学と量子力学の時間発展に関しては、高次のシンプレクティック積分法を使った計算により、有限の時間刻みに対して高精度の計算を実現することが可能である。通常の 2 次精度の計算量を $k = 1$ とすると、4 次、6 次、8 次精度ではそれぞれ $k = 3,\ 7,\ 15$ だけの計算量が必要となる。しかし、この計算量の増大のために復元精度の面からは逆効果になる。高精度計算を利用して時間ステップを通常より大きく取るということをしない限り、逆に復元が難しくなってしまうのであるが、外部との相互作用がある場合など、時間ステップの大きさがあらかじめ指定されている場合があるため、いつでも高精度計算を利用して時間ステップを大きく取るという方法がとれるわけではないことにも注意しておく。

3.2.3　数値流体力学データの復元

流体力学の場合、ナビエ・ストークス方程式は時間反転に対して対称でないため、ある時点で流速をすべて逆向きにしても元通りの状態を回復することはない。しかし、方程式を逆向きに解くことで、時間をさかのぼることは原理的に可能である。では、数値的な安定性の面ではどうだろうか。通常の時間発展では、移流項の非線形性が不安定性の原因となることが多いが、時間を反転するときには、拡散項が問題になる。拡散方程式を逆向きに時間発展させればわかるように、非物理的な先鋭化の現象を安定に計算することは難しい。数値流体力学計算においては、時間間隔を積極的に広く取って保存した場合、逆向きに時間をさかのぼることで精度を高めるという方法が使えないのである。正方向だけの時間発展で再現する場合の時間間隔を評価することは可能だが、本稿での検証方法から外れるため、これ以上は議論しない。レイノルズ数が低い流れ現象は正方向の時間発展で安定に再現が可能であるが、高レイノルズ数の乱流となると再現性が期待できないため、解析に必要なデータを間引いて保存するのは危険な場合がある。

3.3　部分系の抽出とデータマイニング

注目する系が少数自由度系であっても、相互作用している周りの大自由度系も含めた全体の系でシミュレーションを実施することがある。逆に、大自由度系のシミュレーションの中から、特定の少数自由度の部分系を抽出し、詳しく解析したいという場合もある。いずれの場合にも、図 2(b) のように、周りの大自由度系の情報を保存せず、注目系だけの情報を保存することで時系列データの大幅な削減が可能であるが、注目系以外の周辺系の再現が難しくなるため、注目系に関しても、厳密な意味では数値実験の再現は不可能となる。

このような選択的に特定の部分のデータだけを保存する方法は、一種の不可逆な圧縮をしている状態とも言える。復元可能な圧縮法を利用してデータを削減する方法は、一般の応

用プログラムでも広く利用されており、時系列を、時分割で個別のファイルに保存している場合は、個々のファイルを保存後に圧縮することでも効率化が可能である。画像の圧縮で一般に使われている不可逆圧縮の方法は、数値計算データに対してはまだ一般的ではないが、映像のエンコーディングのように時間方向に圧縮をかける方法として、変化を抽出して差分を保存するなど有効に応用可能である。一方、部分系の抽出による削減手法は、あらかじめ部分系が明確に分離している場合には問題ないが、雑音中に埋もれた運動を抽出するなど未知の部分系を探索する場合には、データマイニングの手法と組み合わせる必要がある。

現時点では、数値データのマイニングに関して汎用の手法や一般的な手順の形で確立しているものはなく、効率の面でも理論的な面でも手探りの状態である。一般に生体分子の機能に関連する運動は、速い運動の中に埋もれたゆっくりとした運動であると考えられているため、引用文献[2] に示すように、時間方向のウェーブレット変換の後に相関を見ることで、タンパク質の残基間の正負の相関が明確に現れる場合がある。このように特徴的な運動のみを自動的に抽出することが出来れば、時系列データ量の削減が出来るだけでなく、データ解析の手間も省けることとなる。ここで問題となるのは、このような抽出が完了するまでに、一旦大規模なデータ処理が必要になることである。

大自由度系の時系列データは、空間自由度方向と時間自由度方向の二次元データの形で表現することが可能であるが、時間発展計算では、空間自由度方向をメモリー上に持ち、時間方向に順に保存していく形を取るのが通常である。しかし、ウェーブレット変換などの時系列解析を行うためには、時間順に生成されるデータに対して一種の転置操作を行い、時間方向のデータを同時に処理できる形にする必要がある。二次元の全データをすべて保存することが出来ないため、現状では比較的小規模な時系列に対して、試験的に検証が行われている状態であるが、大規模なデータに対してこのような解析を行うためには、大規模データを扱うための情報科学的な手法が必要となる。このようなデータの爆発を計算量でカバーしようというアイデアが、カーネル法[3] を応用したデータマイニングであるが、こちらも有効性の検証段階にとどまっている。

注目する部分系が明確に与えられている場合も、データマイニングによって部分系を抽出する場合も、保存されるデータは全体の一部となり、データの削減をするという意味では有効であるが、厳密な意味での再現性を満足することは難しい。しかし、厳密に数値を再現するような再計算は不可能でも、科学的な意味での検証可能性は満足されていることが多い。この場合、通常の科学的な意味での実験の再現が可能かどうかは、与えられた初期条件からの時間発展で、目的とする現象が安定に発見できるかという点に関係があるので、結局は、

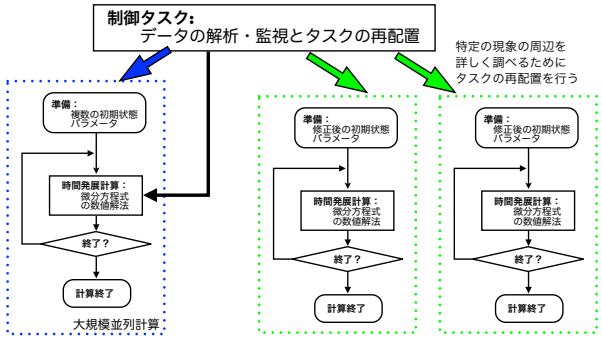

図 6 適応的計算の概念図。データの解析によって特定の状態を発見し、その周辺部を詳しく調べるために、新たにシミュレーションを計画し、自動的に再配置して実行する。インタラクティブシミュレーションの場合は、制御タスクの部分を人間が操作することになる。

目的の現象の一般性に関連がある。最初の発見が何らかの偶然による幸運の結果であった場合を除いて、通常の意味での再現性は満足されると考えられる。

3.4 インタラクティブシミュレーション

そもそも数値計算で得られる一次データとしての時系列を全く保存しないシミュレーション方法は考えられないのだろうか？本稿で考察しているシミュレーションは、天体観測施設や加速器などの大規模観測・実験装置と同じ位置づけで、大規模数値計算の中から未知の現象を発見して解析することが目的である。これまでは、まず一次データを保存した後に、ポスト処理として解析を実施する形を考察してきたが、ここでは、計算中に観測に相当する可視化や解析計算を実施し、新規現象の発見と同時にインタラクティブに操作することで、データ保存を行わない手法を考察する。

例えば、GPGPU を使うことで高速化と状態の可視化を実現し、インタラクティブな操作と組み合わせてシミュレーションを実施するという試み[4] が存在する。残念ながら、通常の大規模超並列マシンではバッチ形式での実行が多く、さらに入力・出力等に関してもステージングを採用するなどインタラクティブな操作を可能にする仕組みからは遠くなっている。演算性能を高めるためには外部の擾乱から隔離した形で実行することが必要になるのは理解できるが、データの保存と移動が律速になっている状態では、科学的な成果を得ることを第一目的とするのであれば、性能面での追求をあきらめるという選択肢も存在してよい。

とはいえ、現実を見ると、「京」の計算ノードでの計算内容をインタラクティブに操作す

るというのは、実現性はほとんど期待できない。そこで、インタラクティブな操作ではなく、実行するプログラム中に何らかの適応的動作を組み込むことで実現するという方策を取ることを考える。大規模な計算の中から特定の現象を計算中に探索し、発見と同時にその周辺を詳細に計算するという手法、いわば、計算対象のダイナミックな修正まで含めた適応的計算を実施するのである (図 6)。計算機の効率を高めるための動的な負荷分散の手法に関しては、様々な仕組みを利用して試みられている[5] が、このような適応的手法を数値実験の解析にも適用しようという試みである。

インタラクティブなシミュレーションの場合にデータの保存が削減できるのは、時系列データそのものではなく、計算に対して与えた操作を記録することで再現性を確保するからである。しかし、この場合には、安定な再現性の確保と同時に、部分的に再現することの困難さという問題が存在する。最初の環境と同じ規模の計算機環境がないと、現象の検証が困難になるということで、大規模超並列計算機ではなく、GPGPU などの身近な高性能演算器の利用に適した方法であるといえる。適応的計算はインタラクティブ操作が不可能な計算機向けではあるが、現時点では、このような方法は単なるアイデア段階であり、あまり一般的に利用されているプログラミング手法とはいえず、これを実現するためのプログラミング手法や構成の研究から始める必要がある。しかし、データの保存という方向に今後、革新的な発展がないならば、計算科学的な立場からこの手法を実現するフレームワークの構築は、十分に意味のあることであると考える。

4. 結　　論

前半部分では、分子動力学、量子力学、流体力学のそれぞれの時間発展計算に対して、保存するデータ量の見積もりを行った。ここでは比較的単純な計算を対象としたために、保存データ量が課題に算出されている感は否めないが、今後、演算性能がエクサフロップスに向けてさらに高まる時には、無視できないものとなることは確実である。いずれにしても、現時点の科学の枠組みの中で意味のある計算を行う時に、演算性能自体が律速になっていることはあまりないように思われる。計算機科学の研究対象として演算性能を極限まで高めるための苦労や様々な面での試行錯誤を否定するつもりは毛頭ないし、その成果の華々しさが大規模研究開発の牽引力となっていることも否定できないが、科学研究にも使える大規模計算機を構築するためにも、計算科学の領域からの具体的な提案を増やしていきたい。

後半部分では、データの再現性に関して詳しく検討した。これは、現時点で最も容易に利用でき効果の大きいデータ削減手法が、時系列の間欠保存であるためだが、可逆時間発展方程式に従うものでも、古典動力学と量子動力学では明確な差が現れることがわかった。数値表現はここでは倍精度実数を基準としたが、4 倍精度実数表現を利用する場合、時系列長に線形に誤差が増えるにすぎない量子系の場合には有効だが、指数関数的に誤差が増える古典力学系に対しては、それほど大きな効果は期待できないことになる。今後、データの再現性に関して、非可逆時間発展を行う流体力学系などで、定量的な評価を行うこととしたい。また、その他のデータ削減手法に関しても、有効性の評価をさらに詳しくしていきたい。

謝辞　本研究は、科学研究費補助金 (挑戦的萌芽研究「揺らぐ環境にある生体分子における機能発現機構の『頑健性』の解明」課題番号 22654047) の支援を受けています。

参　考　文　献

1) T. Takami, H. Fujisaki, and T. Miyadera, “Coarse-Grained Picture for Controlling Quantum Chaos,” Adv. Chem. Phys. **130A**, 435–458 (2005).
2) M. Kamada, M. Toda, M. Sekijima, M. Takata, and K. Joe, “Analysis of motion features for molecular dynamics simulation of proteins,” Chem. Phys. Lett. **502**, 241–247 (2011).
3) 福水健次「カーネル法入門 — 正定値カーネルによるデータ解析 —」朝倉書店 (2010).
4) 楠 昌紘, 高見 利也, 小林 泰三, 稲富 雄一, 西田 晃, 青柳 睦「GPGPU を用いたリアルタイムシミュレーションの有効性と課題」, 火の国情報シンポジウム 2010 講演予稿集 (CDROM), A-2-4 (2010).
5) Y. Inadomi, T. Takami, J. Maki, T. Kobayashi, and M. Aoyagi, “RPC/MPI Hybrid Implementation of OpenFMO — All Electron Calculations of a Ribosome,” Advances in Parallel Computing **19**, 220–227 (2010).