jh140004-NA02

大規模データ系の VR 可視化解析を効率化する 多階層精度圧縮数値記録(JHPCN-DF)の実用化研究

萩田　克美（防衛大学校）

本研究では、VR 可視化等を目的として、大規模データのうち上位ビットの部分のみを分割転送することで、高圧縮でのデータ転送を実現することで、実用上の効率化を検討した。ビットを分割し下位ビットをゼロパディングすることと、Huffman 符号化圧縮とを組み合わせることで、可変長データの API の作成が不要となる。HDF5 のような汎用フォーマットを利用することで、可視化アプリなどにおいて何ら改造することなく、提案手法の恩恵を受けられることを示した。

1.　共同研究に関する情報

(1) 共同研究を実施した拠点名

名古屋大学情報基盤センター

北海道大学情報基盤センター

東北大学サイバーシナジーセンター

大阪大学サイバーメディアセンター

（東京工業大学学術国際情報センター）

(2) 共同研究分野

☑　超大規模数値計算系応用分野

□　超大規模データ処理系応用分野

□　超大容量ネットワーク技術分野

□　超大規模情報システム関連研究分野

(3) 参加研究者の役割分担

参加研究者の役割分担は次の通り。

★印は､JHPCN 拠点情報基盤センターの教員。

参加研究者	役割分担
防衛大　萩田克美	総括、有効性の検討
名大　荻野正雄★	総括補佐、FEA 解析での有効性の検討
核融合研　石黒静児	プラズマ粒子計算での有効性検証
核融合研　大谷寛明	
広島大　加藤恒彦（10 月～ 国立天文台）	天文・プラズマ粒子計算での有効性検証
北大　大宮学★	電磁界解析データでの有効性検証
東工大　青木尊之★	超多粒子系データでの可視化用データ圧縮に関する議論
阪大　下條真司★	ＶＲ可視化などでの可視化用データ圧縮の試験検討や議論
東北大　江川隆輔★	流体計算での有効性検証
RIST 東京　井上孝洋	気象系（NICAM）データでの有効性検証
諏訪東京理科大　河合浩志	FEA 解析での有効性の検討

2.　研究の目的と意義

本研究では、大規模計算で生み出される大容量な計算データについて、目的に応じて必要とされる数値精度に応じた階層的な圧縮記録を行うとともに、それらを結合することで精度を回復できるような記録方式を用いることで、データの詳細に対峙した VR 可視化解析の効率化が可能であることを、分野横断的に検討することを目的とする。特に、JHPCN が、ネットワーク型の拠点である利点を活かし、情報基盤センター教員との共同研究で、多様な対象分野での検討することが目的である。そして、スパコンから産業界まで、広く普及させることも、研究の目的である。

我々は、技法として「多階層的精度圧縮数値記録（JHPCN-DF; Jointed Hierarchical Precision Compression Number - Data Format)」を提案し、実用性を示す。

物理法則を基礎方程式とした数値シミュレーションでは、結果として得られる数値データは、次のような特徴があると考えられる。

・データの変化が時空間的に滑らかである

という特徴をもつ。この特徴から、仮数部の上位ビットは変化が少なく、下位ビットほど変化が大きいという特性を持つ。また、ある程度以下のビットは、ランダムに変化する特性を持つ。

- 可視化などの用途では、上位と下位のビットを分割し、上位ビットの部分を高い圧縮率でデータ圧縮し、ファイル転送などの作業効率化が期待できる。
- 用途の階層に応じて、上位、中位と下位と多階層的に分割して記録し管理することで、効率を向上させる可能性がある。

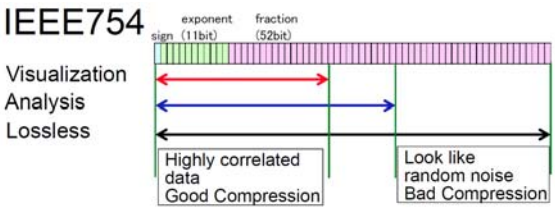


図 1　浮動小数点表現と、上位下位ビットの特性

実際の利用上の問題としては、可変ビット長のデータ記録をする場合、API が特殊となることが課題点と言える。それに対して、我々は、「下位ビットのゼロパディング（マスク）」と「Huffman 符号化圧縮」を組み合わせることで実現でき、HDF5 のような汎用フォーマットを用いれば、アプリを何ら改造することなく恩恵を受けられることに気がついた。

上位、中位（複数）、下位のビットをうまく分割管理し、全体としては、Loss-less 圧縮を実現する仕組みや、ライブラリ・ミドルウェアなどのシステムソフトウェア化を図ることが、大規模データを伴うシミュレーションの効率化に有効であると考えている。

JHPCN-DF の活用により、手元の WS、近くのスパコン、大規模スパコンと、それぞれの階層で扱うデータを、分割し、階層的に保存することができる。本研究により、VR 可視化などの実用事例までを含めて、JHPCN-DF の有効性を示すことで、多分野でのデータハンドリングの問題に貢献することが可能であり、意義は大きい。

本提案手法は、可視化で扱うデータサイズの削減（ハンドリングの向上）のみならず、多数の研究分野でのシミュレーション結果のアーカイブや、産業界における作業効率の向上（産業界向けスパコン利用のワークフロー支援ソフトでの実装）、Amazon AWS(EC2)などでのデータ出力料金の節約（実質的には、金銭のみならず、電気通信の電気的資源のエコにも繋がる。）、上位ビットを割り符としたデータ機密管理法など、実用的なメリットが見出されており、大規模計算での検証事例を増やすことは意義が大きいと考えられる。

3. 当拠点公募型共同研究として実施した意義

JHPCN のネットワーク型拠点であることを最大限活かし、JHPCN 情報基盤センターの教員が専門とする分野でのシミュレーションデータに対して、提案手法の有効性確認を行う役割分担が実現できた。これは、JHPCN 拠点公募型ならではと言える。

また、提案手法の普及展開においても、JHPCN のシンポジウムでの成果公開や交流が大いに役立った点は意義が大きい。

さらに、センターの利用支援や利用技術の啓発活動などを通じて、提案手法が普及する可能性もあり、期待している。

加えて、本提案手法は、シミュレーション・データ・マネージメントや、ワークフロー（データフロー）効率化に関係してくることから、スパコンのシステムや利用方法と密接な形で検討できる素地が確立することは、今後の検討推進において望ましい形である。

4. 前年度までに得られた研究成果の概要

今年度において、新規に提案した課題である。

5. 今年度の研究成果の詳細

提案した手法 JHPCN-DF について実装コード例を作成し、複数のシミュレーション手法の大規模データに対して、可視化を目的とした場合に、データ圧縮が良好に行われることを確認した。

(1) JHPCN-DF の方法の概要

① 背景

大規模なシミュレーション結果のデータについては、データ転送が事実上困難であることが課題となっている。元データの内容確認のための圧縮形態の一つとして、スパコンでの「その場」可視化が行われている。この方針では、データ確認に必要な多視点の画像作成を行い、大幅にデータ量を縮減している。実際のデータ確認としては、画像を作成することでは不十分な場合もある。データ量を転送しやすい 1/100 程度にして、VR 可視化や解析（可視化表現の物理量計算や特徴点抽出・主成分分析など）を実現することが望まれている。また、データ圧縮は大規模可視化でのスループット改善としても期待される。

実際の圧縮率は、データ特性に依存することから、実際のデータを用いて、VR 可視化に耐えるレベル、サイエンスとしての解析に耐えるレベルなどの様々な利用レベルに必要な精度に応じた圧縮性能を評価する必要がある。また、圧縮したデータでの可視化や解析により、ディスク容量問題、IO 性能問題などの改善につながることの実証も、手法の提案として必要なことである。

② 手法の特徴

JHPCN-DF では、ある精度を基準とした「上位ビットと指数部を持つデータ」、階層的に精度を基準とした「中位ビットのデータ（複数）」、最後に Loss-less 圧縮を実現するための「下位ビットのデータ」をそれぞれ用意し、HDF5 の形式で、複数のファイルに分割して記録することを提案している。

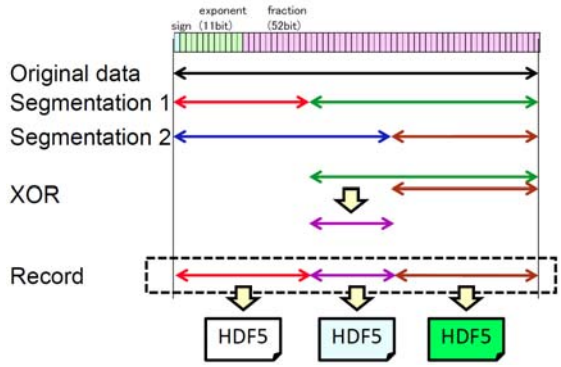


図２　JHPCN-DF でのファイル分割のイメージ

これらの複数の HDF5 ファイルは、コンバータにより、結合することで、可視化やデータ解析で用いるデータの精度を変更することが可能である。シミュレーションのワークフローとしては、Loss-less の HDF5 でデータ記録を行い、コンバータで、精度毎の圧縮ファイルとして記録する。このとき、圧縮率を高めるために、時間方向の複数のデータを１つのチャンクに格納する方式をとることが望ましい。さらに、下位ビット側のデータは保存せずに消去し、ステージアウトの制限やデータ容量問題に対応する使い方もある。

「上位ビットと指数部を持つデータ」は、高い圧縮率が期待できる。故に、シミュレーション直後のデータ転送での確認を可能とするとともに、手元の WS 等での可視化解析を可能とする。

「中位ビットのデータ（複数）」の圧縮率特性は、やや悪化し、データサイズは大きくなる。このデータは、サイエンスに基づいた精度（誤差精度）でのデータアクセスを可能とし、高度な解析用途にも利用できる。近くの情報基盤センターのスパコンへ、精度順にデータ転送することを想定する。

「中位ビットのデータ（複数）」の一部と、「下位ビットのデータ」は、データ転送の対象とせず、計算元のスパコンでの保存を想定している。これにより、Loss-less なデータ再現を担保する。

③ コーディング例

提案手法の理解の為に、本提案手法を実現するために必要な「規定した数値精度で浮動小数点デ

ータをビット単位で分割するコーディング例」について、C 言語版と Fortran 版を以下に示す。以下は、単精度の場合の例である。倍精度に対応するには、仮数部の 23bits を 52bits にする変更をすれば良い。

```
union fi32{
float f;
int i32;
};

// bit-segmentation for variable “fval0”
// resultant is “fval1”
      allowerr=0.001;
      logallo=log(allowerr)/log(2.0);
      fval=frexp(fval0,&ival);
      ival2=(int)(-logallo+ival-1);
      sval=(int)(23-ival2+1);
      if(sval>23) sval=23;
      do {
      sval--;
      fival.f=fval0;
      fival.i32=(fival.i32 >> sval);
      fival.i32=(fival.i32 << sval);
      fval1=fival.f;
      } while
((fval1-fval0)*(fval1-fval0)>allowerr*allowerr);

// XOR
fival.f=fval0;
fival1.f=fval1;
ixval=(fival1.i32   ^ fival.i32);
```

```
INTEGER(4) ::ddi32
REAL(4) ::ddf
EQUIVALENCE(ddi32,ddf)

!! bit-segmentation for variable “d”
!! resultant is “d1”
b = fraction(d)
a = exponent(d)
allo=0.001d0
v1=log(allo)/log(2.0)
v2=int(-v1+a-1)
s1=int(23-v2+1)
if (s1>23) s1=23
do
   s1=s1-1
   ddf=d
  ddi32=ISHFT(ddi32, -s1)
   ddi32=ISHFT(ddi32, s1)
   d1=ddf
   if((d1-d)*(d1-d).le.allo*allo) exit
enddo

!! XOR
ddf=d
dd1f=d1
ixval=ieor(ddi32,dd1i32)
```

(2)　有限要素解析のデータ

グリッドベースの手法の例として、有限要素解析(FEA)について調べた。大規模有限要素法シミュレータである ADVENTURE の出力部を HDF5 に対応させた。そして、JHPCN-DF での精度制限に対するデータ圧縮について評価を行った。

予備テストとして、24 万要素と 38 万接点の系を扱った。静的弾性解析の結果のファイルサイズは、各ノードに対して３次元の変位を倍精度で記録した HDF5 フォーマットで、8.6MB であった。可視化と評価に必要な２つの精度（10^{-3} と 10^{-6}）について、JHPCN-DF を適用した HDF5 ファイルの大きさを調べた。ここで、必要な精度（許容する誤差）は、線形ソルバーの収束判断基準から決定した。可視化と評価のそれぞれの精度に対して、1.3MB と 3.3MB であった。

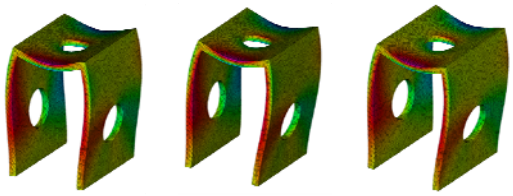


図３　FEM 計算の可視化結果
(左:フル精度、中央:許容誤差 10^{-6}、右:10^{-3})

現実的なテストとして、2.5 千万要素と 3.5 千万節点のパンテオンモデルを扱った。ここでは、可視化用の精度として 10^{-2} と 10^{-3} を、解析用の精度として 10^{-6} を許容誤差として設定した。表１と図４に結果を示す。可視化用と解析用のそれぞれで、オリジナルの HDF5 圧縮データと比べて、約 84%と約 58%までデータサイズを削減できた。なお、可視化用の 10^{-2} と 10^{-3} は、似た傾向を示していた。

表１　パンテオンモデルのデータサイズ

	Allowed Error	**Data size [MB] (space saving)**
HDF5	---	1,289 (---)
HDF5 with JHPCN-DF	10^{-2}	164 (87.2%)
	10^{-3}	200 (84.4%)
	10^{-6}	545 (57.7%)

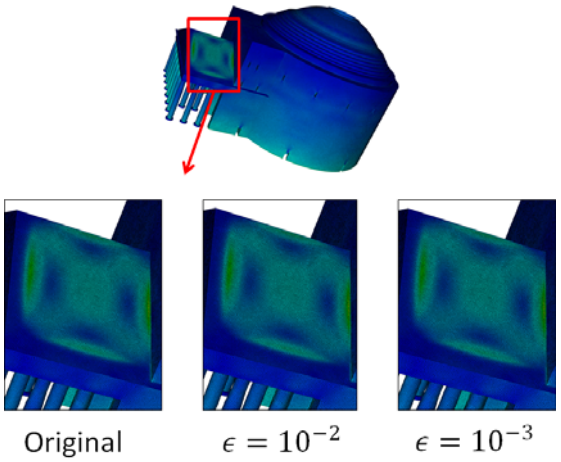


図４　パンテオンモデルの可視化結果

(3)　大規模粒子系データ（プラズマ PIC 計算）

大規模粒子系の例として、時系列を多項式で記述する圧縮法(TOKI 圧縮)の検討で利用したプラズマ PIC シミュレーションのデータを用いて、提案手法 JHPCN-DF の有効性の検証をした。ここでは、HDF5 で入力した Zindaiji3 で描画した図を元に必要精度を判断した。

ここでは、15 万のイオン粒子と、15 万の電子について、2000 時刻分の時系列データを扱った。100 時刻毎に記録した 20 個の HDF5 のサイズの合計は、6,336 MB であった。ここで、ID と index は、integer (4byte)で記録し、3 次元配置を float (4 bytes)で記録した。図５は、許容誤差を 0.01、 0.02、と 0.05 とした場合のスナップショットである。可視化の結果から、必要な精度は、概ね 0.02 と考えられる。JHPCN-DF の方法で上位ビットと下位ビットの２つに分割したファイルサイズについて調べた。結果を表 2 に示す。この結果から、可視化に必要な精度では 1/5 に圧縮できることが分かった。

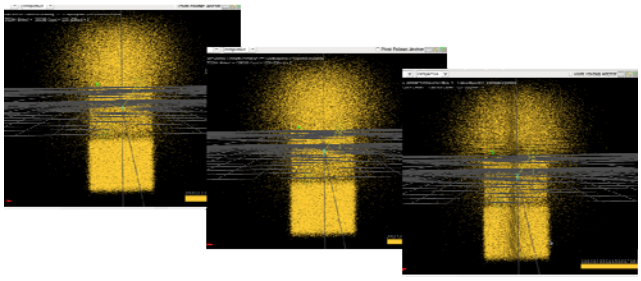


図５　プラズマ PIC シミュレーションの可視化
(左:許容誤差 0.01、中央: 0.02、右: 0.05)

表２　分割したファイルの大きさ

Allowed Error	Data size	
	Higher bits	***Lower bits***
0.0	6,336 MB	----
0.001	3,369 MB	4,503 MB
0.01	1,940 MB	5,298 MB
0.02	1,414 MB	5,586 MB
0.05	883.1 MB	5,712 MB
0.1	568.5 MB	5,857 MB

(4)　高分子相分離系 OCTA/SUSHI のデータ

密度場データの例として、高分子相分離系の OCTA/SUSHI の結果について、JHPCN-DF の有効性について調べた。対象とする相分離構造は、図６に示すように、大きなドメイン構造と小さい球状のドロップレットを含んだ複雑な系である。工業的な応用として、512^3 等の大きなメッシュサイズで、OCTA/SUSHI の計算を実施している。ここでは、128^3 と 512^3 の場合について、AVS/Express で直接読み込める HDF5 形式のファイルサイズを評価した。等値面の可視化に必要な許容誤差は 0.05 程度であり、許容誤差を 0.05 として上位ビットのみ残し圧縮した後のデータサイズを評価した。その結果を表３に示す。比較のために、テキスト形式のオリジナルの出力ファイルサイズも示した。JHPCN-DF の方法では約 1/10 の圧縮が実現でき、有効性を確認した。

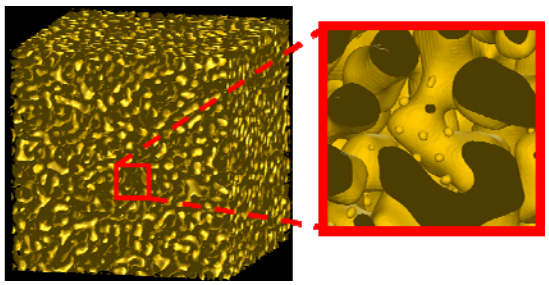


図６　高分子相分離構造の可視化例

表３　ASCII ファイルと圧縮後のファイルサイズ

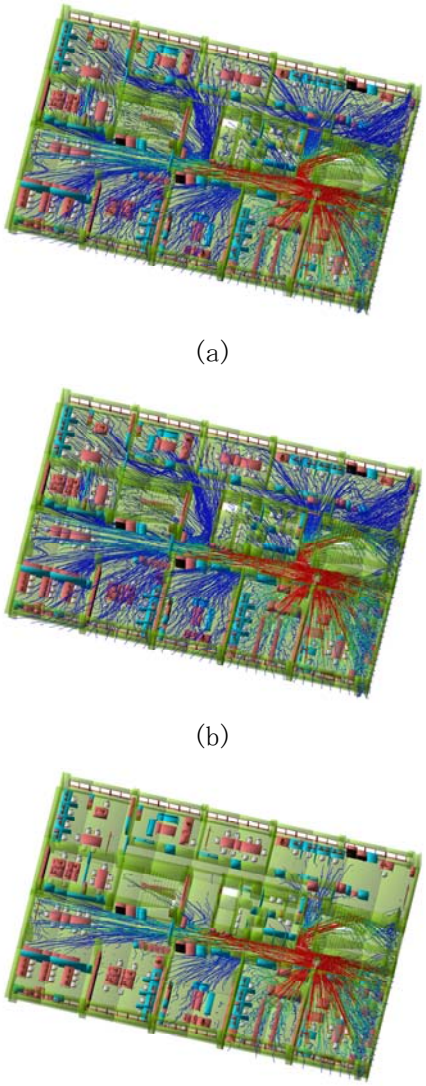
Meshsize	Data size		
	Ascii text	***HDF5***	***HDF5 with JHPCN-DF***
128^3	18,940 kB	6,770 kB	723 kB
512^3	1,182 MB	422 MB	41.8 MB

(5)　大規模電磁界解析のデータ

電磁界解析データについて、JHPCN-DF の有効性の検証を行った。数値シミュレーションには、入出力処理を HDF5 に対応させた時間領域差分(Finite Difference Time Domain、FDTD)法に基づくアプリケーションソフトウエア Jet-FDTD を利用する。

現在、IEEE802.11n/ac 規格に準拠した高出力無線 LAN アクセスポイント(AP)が利用可能である。この AP は高速なデータ通信を実現し、フロア内および複数フロア間での通信も可能である。そこで、無線 LAN システムのオフィス環境での電波伝搬特性評価を目的に大規模電磁界シミュレーションを行い、ポインティングベクトルを可視化することで電波伝搬経路の推定を試みる。FDTD 法では、Yee セルと呼ばれる直方体で解析空間を離散することから、複雑な構造を容易に数値モデル化できる。さらに、マクスウェルの回転方程式を時間領域で直接解くことから、並列計算機向きの汎用かつ効率的な解析手法である。ただし、高精度な解析結果を得るためにはセル寸法を波長の 10 分の 1 以下にしなければならなく、潤沢な計算リソースを必要とする。解析での周波数を 5200MHz とし、一辺の長さが 5mm の立方体で数値モデルを作成した。この結果、オフィス全体を解析するために必要な総主記憶容量は約 3TB であった。なお、数値シミュレーションには HITACHI SR16000 モデル M1 の 40 演算装置を使用し、48000 タイムステップの計算に要した時間は約 20 時間であった。解析結果として最終タイムステップにおける 6 つの電磁界成分(単精度複素数)を HDF5 ファイルで保存したところ、その総容量は 126GB となった。

図 7 にポインティングベクトルの可視化結果を示す。可視化においては、保存した電磁界成分データを再サンプリングし、波長あたり 1 つのデータを使用した。この条件のもとで、電磁界成分からポインティングベクトルを評価し、単精度実数で表現される 3 つのベクトル成分を求めた。このときの HDF5 ファイルサイズは 258MB であった。図 7 (a)はオリジナルデータを利用して可視化した結果である。一方、下位 22 ビットをすべてゼロとしたとき、HDF5 のファイルサイズを 62MB まで減じることができた。そのときの可視化結果は図 6(b)である。これら図を比較することで、可視化結果における差異は明確ではないことが分かる。また、従来行っていた 2 バイト整数に変換した場合の可視化結果を図 7 (c)に示す。この場合、HDF5 ファイルサイズを 30MB とすることができるが、可視化結果は図 7 (a)あるいは(b)とは大きく異なり、多くの情報が失われていることが分かる。



(a)　(b)　(c)

図 7　ポインティングベクトルによる電波伝搬経路の可視化

図 8 は、5.(1)③に示すコーディング例を適用し、

単精度実数の仮数部の下位ビットから順にゼロパディング（マスク）処理を行ったときのビット数とファイルサイズの関係を示している。同図において、横軸はビット数、縦軸はファイルサイズと圧縮率である。ただし、圧縮率とはファイルサイズ比の逆数で、マスク処理なしのファイルサイズ 258MB を 1 とした。同図から、15 ビット以上のマスク処理を行うことで，線形にファイルサイズが減少することが確認できる。

以上の検討結果から、マスク処理によるファイルサイズの圧縮率は必ずしも大きくはないが、JHPCN-DF を利用することでファイルサイズの削減と実用上有効な可視化表示を同時に実現することが可能であることを示した。

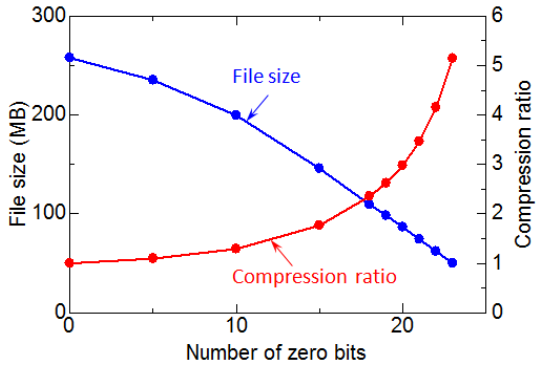


図８　ファイルサイズに対するマスク処理の効果

(6)　気象気候(NICAM)データ

気象気候のシミュレーションデータとして、NICAM（Nonhydrostatic ICosahedral Atmospheric Model）のシミュレーションの水平解像度 3.5km での、1 ヶ月分（2004 年 9 月）の Production run のデータについて調べた。2 次元量 54 個と 3 次元量 14 個について、仮に想定した 3 水準の許容誤差に対して、圧縮後のデータサイズを計測した。この準備的な評価をもとにして、今後、可視化での検証による変数毎に望ましい許容誤差の評価など、詳細の分析と更なる検討は現在進行中である。

6.　今年度の進捗状況と今後の展望

今年度の検討で、多くのシミュレーション対象についての検証を行うことができた。また、SC14 の査読付きポスターや ACSI2015 での発表等で広く研究紹介することもできた。さらに、名古屋大学情報基盤センターの主催で実施した「第 1 回名古屋大学情報基盤センターネットワーク型共同研究シンポジウム」において、本研究課題の成果発表と議論を実施していただき、情報交換や議論検討を大いに進めることができた。

上記のことから、進捗は概ね順調であると考えられる。なお、当初の予定からの変更や遅れがあった。それらへの対応は次のように予定している。

- 本手法でデータ転送において有効であることは明らかだが、大規模可視化装置などでの大規模可視化データにおけるスループット改善に関する評価については実施できていない。これは、評価方法の設定が難しい問題である。試行的に、データ読み込み処理の時間改善などについて検討を進める考えである。
- 比較的密な超多粒子系データの検討として、当初、約 1 億粒子の粉体データを扱う予定としていたが、MPS 法（粒子法）による流体計算のデータを扱うこととした。この方針変更は、JHPCN 中間報告会の発表での情報交換を通じて、東京大学大学院工学研究科 室谷先生に協力を頂けることとなったためである。室谷先生には、H27 年度の JHPCN 課題では共同研究者として協力いただく予定である。
- NICAM データの検討については、可視化による圧縮後データの評価確認環境の準備等が遅れた。H27 年度の JHPCN 課題において、引き続き検討する予定である。また、データ圧縮特性の季節による変化などは興味深いテーマであるが、多くの資源を要するため、将来の課題と考えている。
- 流体系のデータやその他のデータ系での応用展開については、引き続き、H27 年度の JHPCN 課題で実施する予定である。

H27 年度の JHPCN 課題研究では、本検討課題を

継続するとともに、大規模 VR 可視化を検討してきた H26 年度 JHPCN 課題（14-NA28）の検討を吸収統合し、検討を進める予定である。

さらなる展望としては、本提案手法を、トップスパコンのユーザーから、HPC クラウドを利用する産業界ユーザーまで、広く普及させたいと考えている。特に後者のボリュームゾーンへの展開については、民間企業との連携により進めていきたいと考えている。その一歩として、理研 AICS 小野謙二博士との共同研究で、JHPCN-DF のデータ処理を高速化するライブラリを開発し配布することを予定している。

7. 研究成果リスト

(1) 学術論文

・K. Hagita, H. Ohtani, T. Kato, and S. Ishiguro, "TOKI compression for plasma particle simulation", Journal of Plasma and Fusion Research, Vol.9 (2014) 3401083.

(2) 国際会議プロシーディングス

なし

(3) 国際会議発表

・K. Hagita, M. Omiya, T. Honda, M. Ogino, Efficient Data Compression by Efficient Use of HDF5 Format, SC14 Refereed Poster (2014)

(4) 国内会議発表

・萩田 克美,「可視化用途向けの Lossy 圧縮手法 JHPCN-DF の提案」, 核融合科学研究所 先進的描画装置を用いた可視化表現法の研究会(2015)

・萩田 克美,「JHPCN-DF の概要と、多分野での利活用に向けて」, 第１回名古屋大学情報基盤センターネットワーク型共同研究シンポジウム(2015)

・大宮 学,「FDTD 法による電磁界解析での活用」, 第１回名古屋大学情報基盤センターネットワーク型共同研究シンポジウム(2015)

・荻野 正雄,「FEM による構造解析での活用」, 第１回名古屋大学情報基盤センターネットワーク型共同研究シンポジウム(2015)

・井上 孝洋,「雲解像大気大循環モデル NICAM によるシミュレーションと、大規模出力データアーカイブでの JHPCN-DF 活用の展望」, 第１回名古屋大学情報基盤センターネットワーク型共同研究シンポジウム(2015)

・K. Hagita, Study of Efficient Data Compression by JHPCN-DF, Annual Meeting on Advanced Computing System and Infrastructure 2015 (2015)

(5) その他（特許，プレス発表，著書等）

なし