画像重畳装置とクラスタ装置を使った高速可視化システムの開発

林 幸子 (KGT), 宮地英生○ (KGT), 小野謙二 (東大)

High Performance Visualization on PC Cluster with Image Merge Unit

Sachiko Hayashi, Hideo Miyachi and Kenji ONO

ABSTRACT

A novel software "AVS/Express PCE", which runs on commodity-based PC cluster, was developed to visualize a large-scale data set efficiently. This system allows users to customize functions without any parallel knowledge and provides parallelized processes for almost all of processes such as I/O, filtering, mapping, and rendering. It is very effective to parallelize all processes but the image compositing becomes a bottleneck for a parallel rendering system. To avoid this problem, a special hardware that can synthesize an output image from input images without significant delay was combined to the software. According to preliminary test results, the present system performed about 4 frames per second to update iso-surfaces for a million grids field data.

Keywords: visualization, parallel, large data, PC cluster

1. はじめに

並列計算機の普及により、 10^9 を越える格子数の大規模数値計算が行われるようになってきた。このような大規模データの可視化に並列の可視化ライブラリ^{[1][2]}やシステム^[3]が提案されている。しかし、ライブラリやツールキットはプログラマしか利用できず、一方、並列対応の可視化アプリケーションは自由度が小さい。そこで著者らは、利用者のカスタマイズを容易にするビジュアルプログラミングを提供する可視化ツールである AVS/Express の並列版 AVS/Express PCE(Parallel Cluster Edition)^[4]を開発した。

並列可視化では、並列計算機の計算ノードと制御ノードの役割分担が重要である。一般に、三次元可視化では、計算時間を要する等値面の生成までを計算ノードで並列に行い、その結果を幾何データとして制御ノードに転送し、制御ノードのグラフィックスハードウェアを利用して表示を行う。しかし、この方法は制御ノードのグラフィックスハードウェアの性能（グラフィックスメモリ容量）に制限を受ける。そこで、画像生成まで計算ノードで並列処理し、奥行き情報を持った画像を制御ノードに転送し、制御ノードで複数の画像を合成する方法が有効となる。ところが、この方法は回転や拡大の幾何変換操作が行われるたびに計算ノードでの処理が必要となり、可視化にとって重要な対話性能が損なわれる。

そこで、本研究では VG クラスタ^[5]に利用されている画像重畳装置を画像合成に用いることで我々の並列可視

化ソフトの対話性能の改善を行った。本報では、ソフトウェアの実装方法と実験結果について報告する。

2. 実装方法

2.1 AVS/Express PCE

Fig. 1 に AVS/ExpressPCE の概念図を示す。

Fig. 1 AVS/Express PCE Concept Chart

ソフトウェアは制御ノードと計算ノード(クラスタ側)で動作するが、すべての対話操作は制御ノードから行われる。数値解析結果は、あらかじめ分割され計算ノードのローカルディスクに格納されているという前提で、計

算ノード側のソフトウェアは分割された各ブロックの可視化を行う。

計算ノードから制御ノードへ渡すデータは3種類である(Fig. 2)。等値面などのマッピングの前のFieldデータ、マッピング後の幾何データ、レンダリング後の奥行き情報(Z値)付き画像データである。この中で、画像データのサイズは、データの規模ではなく表示ウインドウのサイズにのみ依存するので、大規模データに対しては画像転送が最も有効な方法となる。

Fig. 2 Three Types of Data Transfer from Computation Nodes to Control Node

2.2 画像重畳装置

並列ボリュームレンダリングをリアルタイムに実行するVGクラスタは、透明度付きの画像をハード的に合成する画像重畳装置(Merge Unit)を持つ。これはボリュームレンダリング結果を合成するためのものであるが、サーフェスレンダリングに対しても同様に利用することができる。

Fig. 3 Image Merge Unit System Chart

画像重畳装置は優先度を持った画像をLVDSケーブルから入力し(最大8チャンネル)、画像を合成し出力する。

これをアプリケーションから使う場合、重畳装置専用のボード(IBout/IBin)をPCIスロットにさし、専用のドライバを通してアプリケーションから画像データを読み書きする。読み書きの同期を取るのはアプリケーション

の仕事で、クラスタの計算ノードのアプリケーションは1ms以内の精度でデータの読み書きをしなければならない。一方、画像重畳装置は、計算ノードでのIBoutへのデータの書き込みから、画像の合成、IBinへの画像の書き込みが1ms以内に終了するように設計されている。

2.3 AVS/Express PCEの画像重畳装置対応

奥行き付き画像を合成するプロセスについて詳細を説明する。

Fig. 4 Process Diagram of Standard AVS/Express PCE

Fig. 4は画像重畳装置を実装する前の、通常のAVS/Express PCE(以下、標準版)の処理を示している。まず、計算ノードのAVS/Expressが、OpenGL APIをコールして可視化結果のレンダリングを行う(Fig. 4 ①)。次に、グラフィックスボードのフレームバッファとZバッファから画像とZ値を獲得し、FIFOを経由して通信プロセスにそれらを渡す(Fig. 4 ②、③)。

通信プロセスの間はTCP/IPソケットを使い、LAN上を画像データが流れ、制御ノードの通信プロセスがFIFOを経由して制御ノードのAVS/Express PCEに画像を渡す(Fig. 4 ④、⑤)。最後にマスター側のAVS/ExpressがZ値を比較しながら画像を合成し、完成した1枚の画像をマスター側のフレームバッファに描くことでスクリーンに表示を行う(Fig. 4 ⑥、⑦)。

Fig. 5 Process Diagram of Merge Unit Version AVS/Express PCE

次に、画像重畳装置を使った場合(以下、重畳装置版)の処理をFig. 5に示す。こちらでは通信プロセスの代わをMPIプロセスが行う。通信プロセスも、命令を制御ノードから計算ノードへ通信するために動作しているが、画像合成には関与しないので省略している。計算ノードのグラフィックスボードでレンダリングした画像を獲得

するところまで、標準版と重畳装置版の処理は同じだが、重畳装置版では、Z 値を合成に利用しないのでボードからは取らない。代わりにカメラマトリックスから、ブロックの前後判定を行い、その情報は優先度として画像に添付されFIFOを通してMPIプロセスに渡される(Fig. 5 ①、②、③)。MPI プロセスは、画像を受け取ったあと、すべての計算ノードが画像を受け取るまで MPI のバリア同期で待ち状態になる。すべてのクラスタで画像の獲得が終わると、バリアが解消されクラスタ側の MPI プロセスは画像を IBout に書き込み、画像重畳装置で合成が行われ、制御ノードの MPI プロセスは画像を IBin から読み出す(Fig. 5 ④、⑤、⑥)。最後に MPI プロセスが自分の保持するウィンドウに、合成された画像を描画することで、画像はスクリーンに表示される(Fig. 5 ⑦)。

3. 性能測定

3.1 試験環境

試験は4台の計算ノードと1台の制御ノードのクラスタマシンで行った。詳細は Table. 1 に示す。

Table. 1 System Specifications for Test

Hardware	Linux Cluster System 5Node (1: control, 4: calculation), Memory 1GByte, GeForce4 4600Ti CPU: Control node: Intel Xeon 2GHz Calculation node: Intel P4 2.6GHz Image Merge Unit developed by Mitsubishi Precision Co., Ltd.
Network	1000BASE
Software	OS: Linux RedHat 9.0 MPI: CH 5.1.2 Open Motif 2.2 AVS/Express PCE 1.1 (Developer 6.2 base)

3.2 試験データ

ベンチマークのために直交等間隔格子（データとしては x,y,z 座標値を持っている）の同じサイズ、同じ分布のデータを、4 ノードそれぞれに与える。各ノードには立方体の中心からの距離をスカラー値として float で定義している (Fig. 6)。

Fig. 6 Visualization Result of Test Data

データサイズは次の3種類を準備した。

- 1) 1 ノードの格子数：64x64x64（ファイルサイ

ズ：4 Mbyte)、出力ポリゴン数：7 5 6 ライン、37,148 三角形（等値面 level=32）。

- 2) 1 ノードの格子数：128x128x128（ファイルサイズ：32Mbyte）、出力ポリゴン数：1,524 ライン、151,724 三角形（等値面 level=64）。
- 3) 1 ノードの格子数：256x256x256（ファイルサイズ：256Mbyte）、出力ポリゴン数：3,060 ライン、612,476 三角形（等値面 level=128）。

表示する画像のサイズは 506x506 ピクセルとした。

3.3 計測方法

等値面を生成して表示を行った後、回転操作を行ったときの再描画時間を測定し、標準版と重畳装置版との比較を行った。測定には 1 ms のオーダーで時刻を取るシステム関数を用い、次の3区間で測定した。

区間1は、計算ノードで再描画命令を受けた後、レンダリングを実施し通信プロセスまたは MPI プロセスへのデータ転送を開始するまでの時間(Fig. 4, 5 の①、②)。これを「レンダリング時間」と呼ぶことにする。

区間2は、そこから画像が制御ノードに渡され、Z 比較を行い1枚の最終画像が生成されるまで。標準版では、計算ノードの通信プロセスから制御ノードの通信プロセスを経由して奥行き付き画像が転送され、制御ノードで画像合成処理が終了するまでの時間、重畳装置版では計算ノードから制御ノードへ画像が転送される途中、ハードウェアにより合成されるので、制御ノードで最終画像の受け取りが終了するまでの時間となる (Fig. 4, 5 の③、④、⑤、⑥)。これを「通信合成時間」と呼ぶことにする。

区間3は、最終画像を制御ノードのディスプレイに表示する時間である (Fig. 4, 5 の⑦)。これを「表示時間」と呼ぶことにする。

3区間の時間の合計は、ほぼ再描画の合計時間になる。制御ノードから計算ノードへ命令が伝達される時間が不足するが、これは 2/100 程度の時間である。

3.4 測定結果

測定結果をグラフで表したものを Fig. 7 に示す。

Fig. 7 Measurement Result

レンダリング時間は共通処理で、両者に差は無い。データサイズが大きくなるにつれて等値面で生成されるポリゴン数が増えるので時間がかかるようになる。

通信時間はデータサイズに依存せず一定。標準版で 0.15-0.16 秒、重畳装置版で 0.04-0.05 秒要している。

表示時間は、標準版では 0.25 秒、重畳装置版では 0.01 秒と、重畳装置版の方が速い。

3.5 実用データへの適用

Fig. 8 Application Example (Air Conditioning Simulation in a Car Cabin)

重畳装置版で車の居室内の熱流体解析結果^[6]を可視化した例を Fig. 8 に示す。計算格子は 215x282x134 で、表示されているポリゴン数は約 15 万ポリゴンである。Fig. 8 では幾何形状とともに、断面の温度分布も表示している。図中の白い枠線は分割した小領域を示している。この可視化結果の再レンダリングの時間は 0.363 秒で、約 3 フレーム/秒の対話性能が得られた。

4. 考察

画像合成に画像重畳装置を利用することで、通信合成時間を約 3 分の 1 に短縮することができた。1 ノード当たり 64x64x64 (全体で約 100 万格子点) のサイズのデータに対しては、再レンダリングの時間が 0.225 秒で、4 フレーム/秒で回転や拡大処理ができることが解った。

転送合成時間の詳細を見ると、重畳装置による画像合成時間と LVDS ケーブルを通してのデータ転送時間は一定で約 0.02 秒程度である。残りの 0.13-0.14 秒は、レンダリングされた画像の IBout への書き込みと、すべての計算ノードの画像が揃うまでの待ち時間である。計算ノードの数を増やしても画像合成と転送の時間は、増えないと思われるが、待ち時間に関しては増える可能性はある。この時間が一定で抑えることができるならば、計算ノード数を増やすことで大きなデータに対しても、秒 4 フレーム程度の対話処理が可能となる。

また、合成された後の画像を表示する時間に大きな差がある。これは実装方法の違いに起因する。標準版は AVS/Express のオブジェクトマネージャを経由して画

像を表示し、重畳装置版は MPI プロセスが直接 OpenGL API で画像を表示する。対話性能を求める場合、0.25 秒は非常に大きなウェイトを占めるので、標準版も直接描画することが望ましい。しかし、システム全体として操作の統一感を維持するには AVS/Express を経由する方が良い。表示が速いだけでは可視化システムと言うことはできないので、高速でかつユーザインターフェイスの一貫性を維持する工夫が必要となる。

5. まとめ

並列可視化ソフトウェア AVS/Express PCE に画像重畳装置を組み合わせることで高速な可視化システムを開発した。

ベンチマークの結果、100 万格子程度の数値計算結果に対して約 4 フレーム/秒で再レンダリングできることがわかった。また、実用的な 800 万格子のデータにも適用し 3 フレーム/秒の性能を得ることができた。

今後は、表示速度の改善と計算ノード数を増やしたときの性能特性について検討したい。

参考文献

- 1) J.Ahrens, C.Law, K.Martin, M.Papka, W.Schroeder: A Parallel Approach for Efficiently Visualizing Extremely Large, Time-Varying Data Sets, Los Alamos National Laboratory, Technical Report #LAUR-00-1620
- 2) 中島研吾、奥田洋司：戦略基盤ソフトウェアの開発：HPC ミドルウェア、計算工学講演会論文集, Vol. 8, No. 2, (2003), pp. 707-710
- 3) CEI home page
<http://www.ceintl.com/products/ensightgold.html>
- 4) 宮地英生、林幸子、中井陽一、伊藤嘉晃、白崎実、姫野龍太郎：クラスタシステムで動作する並列可視化システムの開発、計算工学会講演論文集, Vol. 8, No2, (2003), pp. 893-894, 2003
- 5) Shigeru Muraki, Masato Ogata, Kwan-Liu Ma, Kenji Koshizuka, Kagenori Kajihara, Xuezheng Liu, Yasutada Nagano, Kazuro Shimokawa, Next-Generation Visual Supercomputing using PC Clusters with Volume Graphics Hardware Devices, Proc. IEEE SC2001, November 2001
- 6) Kenji Ono, Hideki Matsumoto and Ryutaro Himeno: Visualization of Thermal Flows in an Automotive Cabin with Volume Rendering Method, Data Visualization 2001, D. Ebert, et al. (eds), Springer(2001)pp.301-308.