Vol.2009-ARC-186 No.1
Vol.2009-HPC-123 No.1
2009/11/30

機能メモリと COTS の PCI express 接続による ヘテロ環境向けボリュームレンダリングの 設計

芳野裕子[†]　田邊 昇[††]　小川裕佳[†]
高田雅美[†]　城 和貴[†]

本報告では，筆者らが提案している可視化システムにおける「アクセラレータが PCI express 越しに機能メモリをアクセスする」という概念を推進するために，ボリュームレンダリングに注目した提案システム向けアルゴリズムを検討し，PCI express の特性を評価する．さらに，その環境に適したボリュームレンダリングを C-ray をベースに設計する． C-ray は Cell/B.E.用に最適化されたボリュームレンダリングの実装例であり，我々の開発しているシステムは PPE が入っていないので，C-ray をそのまま用いることはできない．そのための移植手法の検討が本稿の目的である．

Design of a Volume Rendering Method for a Hetero Environment Constructed by PCI express with Functional Memory and COTS

Yuko Yoshino[†]　Noboru Tanabe[††]　Yuka Ogawa[†]
Masami Takata[†]　and　Kazuki Joe[†]

In this report, aiming at our concept of "Accelerators access function memory via PCI express ", we investigate a volume rendering algorithm for our system to evaluate the characteristics of PCI-express. In addition, we design a volume rendering algorithm based on the C-ray for the hetero environment. The C-ray is the optimized implementation of volume rendering for the Cell/B.E. Broadband Engine. Since our system does not have PPEs, we can not apply the C-ray to our system. The purpose of this report is to investigate the re-design of the C-ray for our system.

1. はじめに

ボリュームレンダリングによる３次元可視化手法は，大規模な数値シミュレーション結果や MRI・人工衛星・油田探索等各種センサーによる観測データを分析するために，幅広く利用されている．この手法を用いて大量の数値データを分析する際，拡大・縮小・自由視点変更などのインタラクティブな操作性は非常に重要である．数 GB から数十 GB に及ぶ大規模データのボリュームレンダリングにおいて，インタラクティブな操作を実現するためには，レンダリングの高速計算が必要不可欠である．しかしながら，ボリュームレンダリングにおけるメモリアクセスは，データの並びとアクセスの方向が異なるために不連続アクセスとなるため，CPU のキャッシュや GPU のデバイスメモリが有効に働かず，メモリアクセスが非効率的になる．そのためにリアルタイム処理を実現するのが難しい．

我々は Cell/B.E.プロセッサを改良し，大規模なボリュームデータをリアルタイムでレンダリングすると同時に，H.264 のライブストリーミングで配信し，遠隔のクライアントからインタラクティブに制御できるシステム[6][7][8]を提案している．提案システムは，高い演算能力を有するアクセラレータ（SpersEngine[9] または GPU）と，DIMMnet-3[10][11]のように Gather 機能を有する拡張大容量機能メモリが PCI express を介して接続された並列システムである．これまでの研究報告において用いられてきた Ray casting 法のプログラムには，Cell/B.E.用のボリュームレンダリングプログラムである C-ray などの他の Ray casting 法ベースのボリュームレンダリングプログラムで用いられてきた様々な最適化技法が実装されていない．ハードウェア量を少なくするためには，このプログラムを最適化する必要がある．本報告では C-ray において用いられた各種技法について，上記のような特殊な構成との相性や改良方針を検討し，それに合わせた実験を行う．実験環境として DIMMnet の入手が不可能であることから，実験は，PlayStation3 に搭載されている Cell/B.E 上で，エミュレートすることで行う．

第２章では提案システムの基本概念や設計事項などをまとめる．第３章では，Cell/B.E. 上で Ray casting 法を高速化するための手法の１つである C-ray や，その中で使われている高速化手法について紹介し，第４章はその予備評価を行う．第 6 章は具体的な移植方法について述べ，第７章にまとめを述べる．

[†]奈良女子大学
Nara women's university
[††]株式会社　東芝
Toshiba corporation

Vol.2009-ARC-186 No.1
Vol.2009-HPC-123 No.1
2009/11/30

2. 提案システム

本章では，我々が提案している可視化システムについて述べる．2.1 節では既存のアクセラレータの問題点について述べ，2.2 節は提案システムの基本概念や，処理の流れ，SpursEngine と DIMMnet-3 の組み合わせについて述べる．

2.1 既存のシステムの問題点

HPC 用途へのアクセラレータとして代表的なものは, Tsubame[12]システムに採用された ClearSpeed[13]や, RoadRunner[14]システムに採用された Cell/B.E.(PowerXcell)[15]などがあるが，これらのアクセラレータはメモリ容量の拡張性，アプリケーションによっては演算能力に見合った通信能力，不連続アクセスに対する実効バンド幅に問題を抱えている．

単一の GPU に装備されているメモリ容量を大幅に超えるボリュームデータのサイズでは，GPU から見てホストインターフェースの先のデータをアクセスする確率が高くなる．さらに，1 回の画面生成において原理的には 1 回しか各ボリュームデータの格子点はアクセスされないため，キャッシュフレンドリーではない．そのため，データの再利用性が極めて低く，時間的局所性も望めない．さらに，視線の方向とボリュームデータ配列の並びが違うために，キャッシュラインサイズより大きなストライドで隔てられた不連続アクセスとなる．従って，アクセス時の実行バンド幅を確保しなければ，本質的に高い性能を得られない．

2.2 アーキテクチャ

2.2.1 基本概念

図 1 は提案アーキテクチャの基本概念を示したものである．PCI express 等の高バンド幅な標準 I/O を介してアクセラレータと機能メモリを結合する．PCI express スイッチ等の共有アドレス空間上にデバイスをマップする機能を有する結合網を介してこれらを多数結合する．このような方式により，メモリ容量とメモリバンド幅と結合網バンド幅と演算能力のバランスを維持したスケーラビリティ向上，低消費電力化と低コスト化を実現する．

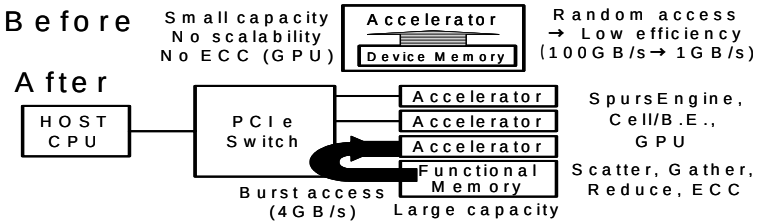

図 1　提案アーキテクチャの基本概念

機能メモリはアクセラレータの外付けデバイスとして，エラー訂正機能が付いた拡張メモリとして用いられる. さらに機能メモリはホストの主記憶と異なり, PCI express 等の標準 I/O を通過するデータ量を削減する機能や転送効率を向上させるための機能を有する．機能メモリが有する具体的な機能として代表的なものは，Scatter/Gather 機能である．その他にも機能メモリ上の複数のデータから中央値(Median)等を出力するような簡単な演算機能(Reduce 機能)も限られた I/O 転送バンド幅の節約に寄与する．

一部のスイッチ製品ではリーフノード間のルーティングをサポートしている．これらを階層的に接続すれば，低コストで PCI 空間上にマップされた多数のデバイス間での読み書きを多数並列に実行できる．

なお，上記の基本概念は国際会議 IWIA'09 における口頭発表および SWoPP'08 において論文発表された大容量データ向け可視化装置を実例として述べられている概念である．ただし，この概念は可視化装置に限定されるものではない．

Scatter/Gather 機能を有する機能メモリとして最も代表的なものは筆者らが研究開発してきた DIMMnet-2 およびその改良版である DIMMnet-3 である．米国の二つの国立研究所(ORNL および SNL)が 2008 年より Scatter/Gather 機能を有する機能メモリを有する超並列計算機の研究開発に着手しており, この種の機能メモリが HPC システムで実際に用いられていく兆しがある．膨大な国費を背景としてもベクトル型スーパーコンピュータが経済的問題から開発が困難な状況が現実化しており，この種の機能メモリが COTS にベクトル機的性質を付加する代替技術として有望である．よって，DIMMnet-3 に類似した機能メモリが今後市場に出てきて，本基本概念を用いたシステム構築に応用できる可能性が高まりつつある．

よって，本基本概念は，可視化装置のみならず，COTS の CPU と GPU 等の演算アクセラレータとベクトル機的なメモリアクセラレータ（機能メモリ）を組み合わせた将来の COTS ベースのベクトル機代替システムにも敷衍できる可能性を秘めている．

2.2.2 処理の流れ

図 2 に基本概念に基づく機能メモリアクセス処理の流れを示す.

Vol.2009-ARC-186 No.1
Vol.2009-HPC-123 No.1
2009/11/30

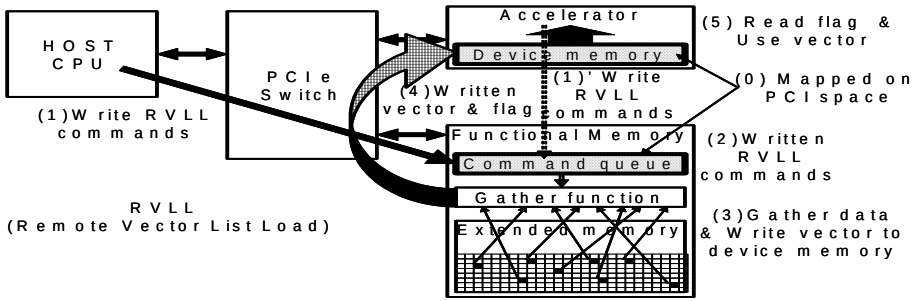

図 2　基本概念に基づく機能メモリアクセス処理の流れ

(1)機能メモリ(例えば PCI express 子基板を実装した DIMMnet-3)へのコマンドキューは PCI 空間上にマップされる．よって，ホストまたはアクセラレータはアクセスしたい機能メモリに割り当てられた上記のコマンドキューに対応するアドレスに所定フォーマットでコマンドを書き込む.

(2)上記に応じて，PCI express スイッチによって構成された木状結合網によって上記書き込みトランザクションは実行され，アクセスする機能メモリのコマンドキューにコマンドが書き込まれる.

(3)機能メモリはコマンドキューからコマンドを取り出して，記載された内容の機能(例えば遠隔リストベクトルロード)を実行し，指定があれば応答データや完了フラグをコマンドに記述されたアドレスに書き込む.

(4)上記に応じて，PCI express スイッチによって構成された木状結合網によって上記書き込みトランザクションは実行され，コマンドは終了する.

(5)アクセラレータは十分に余裕のあるタイミングでコマンドを起動できない場合は必要に応じて上記完了フラグをポーリングし，完了していれば後続の処理(例えば連続化かつアライメント調整されてプリフェッチ済みのベクトルデータに対する SIMD 演算ループ)を実行する.

2.2.3 SpursEngine と DIMMnet-3 の組合せ

SWoPP'08 において発表した可視化装置のハードウェア構想では，前述の基本概念を適用し，アクセラレータとして SpursEngine を用いるシステム構成を提案した．SpursEngine を用いる場合は H.264 ハードウェアエンコーダが内蔵されているため，遠隔可視化の際に有利であるとともに，情報家電機器に利用されることを想定された設計であるため GPU と比べて消費電力とコストの将来性の面で有利である.

SpursEngine と DIMMnet-3 を接続する下位階層は PCI express スイッチで接続されるが，その構想段階では上位階層を Infiniband スイッチによって 4 台の PC が接続されることとしていた.その後，Infiniband の部分も PCI express スイッチに置き換える案を検討中である.この変更によりスイッチ部分の低コスト化と，PCI express と Infiniband 間のルーティング機能といった新規開発部分を省略し，開発リソースや部品コストの削減を図ることができる.なお，図 3 に示すように最上位階層のスイッチのトポロジーがクロスバ接続(Infiniband)からトリー接続(PCI express)に変更になるため，下位階層は PCI express Gen1，上位階層は PCI express Gen2 での接続とするなどで，上位階層ほど高いバンド幅の通信路で接続することが望ましい.

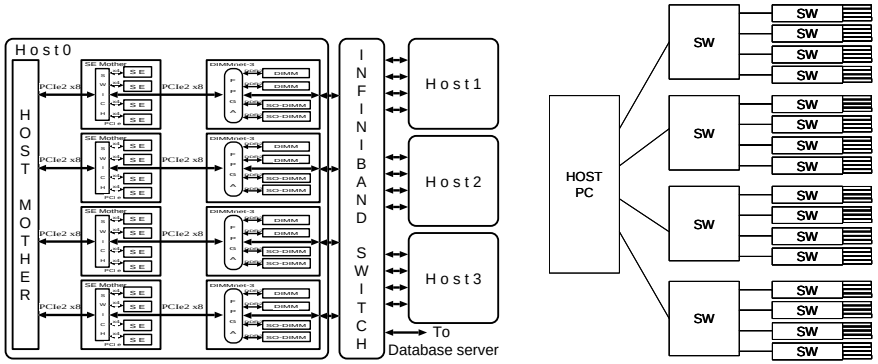

図 3　市販 PCI express スイッチ BOX を用いた結合方法の変更案

3. C-ray

C-ray は IBM 社主催のコンテストである Cell/B.E. Challenge’07 において America region の第 3 位に輝いた Cell/B.E.向けのボリュームレンダリングプログラムである.このプログラムは early termination や，empty skipping など，様々な最適化手法を用い，高速化を行っているとても完成度の高いプログラムである．3.1 節は今回の可視化の方法であるボリュームレンダリングについて述べ，3.2 節は C-ray に使用されている具体的な高速化手法について述べる.

3.1 ボリュームレンダリング

本研究の検討対象はランダムアクセス型アプリケーションの典型例としてボリュームレンダリングとする.

ボリュームレンダリングは半透明の微小物体が浮遊する空間を画像化する方法で，科学技術計算結果の可視化や医療撮影結果などを 1 画面で表現できるので，非常に可視化の効果が高いレンダリング手法である．空間を微小なサイズの等間隔な空間に切

Vol.2009-ARC-186 No.1
Vol.2009-HPC-123 No.1
2009/11/30

り分け，その１つ１つをボクセルと呼ぶ．そのボクセルの並びを３次元配列を用い格納し，任意の視点によりスナップショット画面を計算する方法である．

ボリュームレンダリング手法にはいくつかの基本アルゴリズムがあるが，その中で今回は Ray Casting 法を用いる．Ray Casting 法はレイトレーシング法における物体表面との交点の計算をしないかわりに，ボリュームデータ内でレイをまっすぐに通過させ，その過程に設定した複数のサンプリング点における輝度計算結果を重ね書きする方法である．計算量が多い上に，大容量メモリに対してのアクセスがランダムになるため．高速化にあたっては工夫が必要である．

3.2 C-ray

次に，C-ray に用いられている高速化手法について説明する．

3.2.1 負荷分散戦略

ターゲットシステムでは各アクセラレータまたはホスト PC から Gather 機能付メモリ(DIMMnet-3)に対して遠隔リストベクトルロードコマンドによってデバイスメモリへのプリフェッチを起動することができるものとする．

DIMMnet-3 上の大容量メモリ(最大 28GB)に入ってしまう範囲では 1 個の DIMMnet-3 上にボリュームデータ全体を格納できるので，PCI express スイッチによる木状結合網のバンド幅制約をあまり気にすることなく，個々のアクセラレータから最もアクセスしやすい位置にある DIMMnet-3 に対して遠隔リストベクトルロードコマンドによってプリフェッチを起動する．

このようなシステム上では単純なレイ単位の負荷分散が有効と考えられる．その理由は以下の 3 点である．

(1)Ray casting 法に基づくボリュームレンダリングでは，一般にレイ間の演算に依存関係はないため並列化に伴う計算途中の通信が不要である．

(2)上記の機能メモリを用いた遠隔リストベクトルロードを用いたプリフェッチにはレイ方向に分割されないようにすることで PCI express 上での転送のバースト長を大きく取ることができ転送効率が向上する．

(3)1 本のレイに伴う計算は 1 個のアクセラレータで閉じるので後述する Early ray termination (ERT)法が有効に機能する．

3.2.2 Empty space skipping

Empty space skipping (ESS)法はボリュームデータ内で，何も表示するものが無い空領域の部分が大きい時に効果を発揮する手法である．ボリュームデータが空の時，空であるという情報を格納することで，計算を省略する方法である．この手法で見込める加速率は，例えば空領域が 90%程度と多いボリュームでは 10 倍程度となる．

この方法は C-ray をはじめとする実用的なボリュームレンダリングプログラムにはほぼ必ず組み込まれている，歴史ある方法である．

ESS は予備評価に用いた結果には組み込まれておらず, 効果が高い方法なので今後，実装する．

3.2.3 Early ray termination

Early ray termination (ERT)法は，視点から飛ばしたレイの輝度値の計算を，ある値に達した後は飛ばす作業を中止にすることで計算量を削減する高速化手法である．輝度値が所定の閾値に達すると，一つのレイ上のある点より奥のサンプリング点に関する計算が最終的な見え方に殆ど影響がないことに基づく一種の近似である．本手法は ESS 法とは逆の性質があり，ボリュームに空領域が少ない時に役立つ．

本手法も歴史が古く，C-ray をはじめとする実用的なボリュームレンダリングプログラムにはほぼ必ず組み込まれている方法である．

3.2.4 Streaming

C-ray では Cell/B.E. の PPE 上に，Empty space skipping を適用しつつ空でない領域の Ray が交差するサブボリューム(レイセグメント)を決定して SPE に仕事(レイの視点からのオフセットと長さ)を割り当てるステージを割当て, ERT を適用しながら不透明度の積算をするステージを複数の SPE 上に設けてそれらをパイプライン的処理する．この場合，PPE がネックになりやすいので，次節のような PPE の負担を軽減する最適化が必要になる．

3.2.6 近似による空判定計算の削減

PPE の負担を軽減する最適化として C-ray は近似による空判定計算削減と，それに伴う SPE 側の処理増加の削減を，ハッシュテーブルを用いた Refining という方法で行っている．その効果は高いので，空領域判定部分が相対的に重くなるターゲットシステム上では有効性が高いと考えられる．

3.2.7 PPE 有無の違いへの対応

C-ray は Cell/B.E.上での最適化を行ったプログラムであり, PPE の処理を最適化することをメインにしている．ところが，ターゲットシステムで用いられる SpursEngine には分岐処理に強いとされる PPE は存在しない．そのため，C-ray で PPE に実行させていた処理を分岐処理に強くない SPE や GPU に担当させることが良いかどうか検討する必要がある．

3.2.8 プリフェッチスレッドの動作場所

ターゲットシステムのアーキテクチャ上は機能メモリへのプリフェッチ(遠隔リストベクトルロードコマンド)を発行するプリフェッチスレッドはホスト PC 上や，アクセラレータ上の CPU コア(例えば SpursEngine の SPU)で実行させることもできる．この項目については双方を実装し，評価によって優れた方式を選択するものとする．

Vol.2009-ARC-186 No.1
Vol.2009-HPC-123 No.1
2009/11/30

4. 空領域判定部の実行場所選定のための予備評価

Empty space skipping は処理すべきレイの本数を大幅に削減し，性能向上に有効な方法である．しかし，この中核をなす空領域判定部は C-ray においては，PPE 上で実行する．しかし，PPE は提案するターゲットシステムには存在しない．そこで本章では，この計算部分をホスト CPU に移動したらどのような性能が得られるかについて評価を行う．

評価環境は PPE と SPE については Playstation ®3，ホスト CPU は Intel® Xeon(TM) 3.00GHz，主記憶 1GB，OS は Linux，コンパイラは全て gcc である．

PPE とホスト PC 上で空領域判定部を実行させた場合の実効時間の結果を表 2 と図 4 に示す．ボリュームサイズの一辺が 128 を超えて大きくなるにつれ PPE とホストの性能差は拡大する．PPE もホスト PC もキャッシュベースのほぼ同じ周波数の CPU であるが，キャッシュ容量は PPE が 512KB に対し，ホスト PC は 2MB のキャッシュを内蔵しており大きな差がある．ボリュームデータそのものをアクセスする処理が SPE 側に移動すると，空領域判定のためにたどる木状構造体の全てまたは大半がホスト PC ではキャッシュに納まり，PPE では溢れて大きな差が生じていると考えられる．キャッシュの容量が性能に敏感であることから，アクセス範囲をより小容量で済ませなければ性能が出にくい SPE にこの処理を行わせるのは適切ではないと考えられる．ホスト CPU は PPE よりも大幅に高性能という結果は，空領域判定部をホストに移動すべきことを示唆している．しかし，1 台のホスト PC を図 3 の右図のように木状の PCI express の頂点とする単純なシステム構成では問題があるので，今後，ホスト PC の構成や台数，DIMMnet-3 へのコマンドの与え方について改善すべきであると考える．

表 2　空領域判定部の実行時間比率

Volume size	PPE [ms]	HostPC [ms]	Ratio
16x16x16	0.00011476	0.0000450227	2.549
32x32x32	0.0001182896	0.0000464604	2.546
64x64x64	0.0001778318	0.0000624361	2.848
128x128x128	0.0005335296	0.0000808137	6.602
256x256x256	0.001187948	0.0000954884	12.441

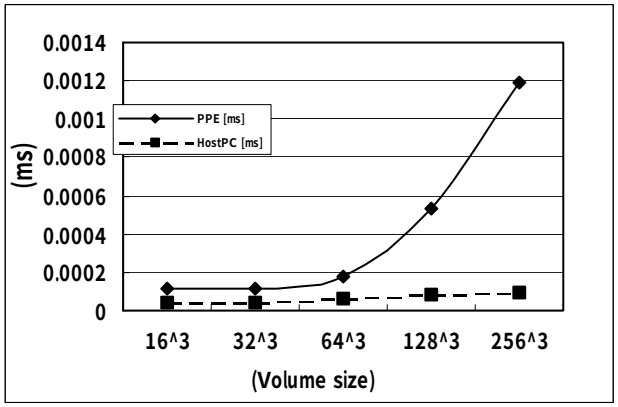

図 4 空領域判定部の実行時間

5. 移植方針

SpuersEngine は PPE を持たないなど Cell/B.E.と違う点が多く，移植する際 PPE で計算していた前処理を他のプロセッサで担当しなければならない．

C-ray は基本的に SPE で可視化の画面やデータの重ね合わせを行い，その前処理を PPE で行っている．4 章の結果より，有効なレイは，PPE よりホスト CPU にやらせる方がよい結果が得られるということがわかった．ポインタ処理や条件判定の多い今回の処理では，SPE による処理が PPE より劣ることは間違いなく，さらに SpursEngine ではクロックも通常の Cell/B.E.の半分なので，一層遅くなる．一方，ホスト PC のコア数は年々増加傾向にあり，多くの SPE を複数のコアで担当できる．以上の根拠から，今回の処理を SPE で動かすのは論外であり，空領域の判定の処理はホスト CPU にて行わせるべきである．

SPE が実行しやすいように DIMMnet が α ブレンディングで使うデータをリストアクセスして連続化するので，その前の処理であるレイの選択判定やレイとボリュームの交点の計算などはホスト CPU で行われるべきである．

よってヘテロ環境での大まかな分担は，以下のようにすべきである．

Vol.2009-ARC-186 No.1
Vol.2009-HPC-123 No.1
2009/11/30

ホスト CPU：前処理＆DIMMnet へのコマンド起動
DIMMnet：不連続アクセス＆SpursEngine のデバイスメモリへの書き込み
SPE：デバイスメモリ上のデータを使ったαブレンディング計算

ホスト PC では，αブレンディング以外の処理，データの読み込みなどや視点の指定などを行う．そこで読み取ったデータを SPE で実行しやすいように DIMMnet3 でリストアクセスし，アクセスを連続化させる．ここのリストというものは，視線方向にデータを並び替えたものである．そのリストを元に SPE で，αブレンディング処理を行う．

具体的には，SPE の処理を，ホストから送られる画像データのリストと関連するその他データを読み込み，レイの計算を行うことだけにする．

6. おわりに

本報告では，これまで我々が提案し，評価してきた可視化システムの基本概念を説明するともに，ボリュームレンダリングを高速に提案システムで行うために， C-ray を移植することを述べた．C-ray は Cell/B.E.用の最適化プログラムであり，我々の開発しているシステムは PPE が入っていないので，C-ray をそのまま用いることはできない．そのため移植方針を探ることは重要であった．

この移植の際考慮しなければならない Empty spaces kipping について予備実験を行い, SPE で実行するよりも CPU で実行したほうが高速になることを確認した. ただし，アクセラレータを多数用いる場合はホスト CPU をマルチコア化するなどの強化が必要である．

今後は，C-ray の最適化技法を実際に本提案システム用に移植し，その評価を行う予定である．

謝辞 本研究の一部(DIMMnet-3 の開発)は総務省戦略的情報通信研究開発推進制度(SCOPE)の一環として行われたものである. C-ray のソースコードをご提供いただいた Cell/B.E. challenge’07 米国第 3 位入賞者 Jusub Kim 氏に感謝いたします.

参考文献

1) M. Levoy : “Display of surface from volume data”, IEEE Computer Graphics and Applications, Vol.8, No.3, pp.29-37 (May 1988)
2) M. Levoy : “Efficient Ray tracing of volume data”, ACM Trans. Graphics, Vol.9, No.3, pp.245-261 (1990)
3) W.M. Hsu : “Segmented ray casting for data parallel volume rendering”, Proceedings of 1993 Parallel Rendering Symposium (PRS’93), pp.7-14 (Oct.1993)
4) T. Porter and T. Duff. : “Compositing Digital Images” ACM Computer Graphics (SIG-GRAPH’84), Vol.18, No.3, pp.253-259 (1984)
5) Jusub Kim, Joseph JaJa : “Streaming model based volume ray casting implementation for Cell Broadband Engine”, Scientific Programming, Vol.17, No.1-2, pp.173-184 (2009)
6) N. Tanabe, M. Nakatake, H. Hakozaki, Y. Dohi, H. Nakajo, H. Amano : "A New Memory Module for COTS-Based Personal Supercomputing", Innovative Architecture for Future Generation High-Performance Processors and Systems (IWIA'04), pp.40-48 (2004)
7) N. Tanabe, H. Nakajo : “High Performance Computing and Database Processing with COTS and Extended Memory Modules”, HPC Asia’09 (Mar. 2009).
8) N. Tanabe, H. Nakajo : " An Enhancer of Memory and Network for Cluster and Its Applications", PDCAT’08 (Dec. 2008)
9) N. Tanabe, M. Sasaki, H. Nakajo, M. Takata, K. Joe : “The Architecture of Visualization System using Memory with Memory-side Gathering and CPUs with DMA-type Memory Accessing”, PDPTA’09 (Jul. 2009).
10) Toshiba Corp. : “Toshiba starts sample shipping of SpursEngine(TM) SE1000 high-performance stream processor”, Press release 08 April, 2008, http://www.toshiba.co.jp/about/press/2008_04/pr0801.htm
11) PLX Technology : "PCI Express 2.0 Switches - PCIe ExpressLane I/O Interconnect", http://www.plxtech.com/products/expresslane/gen2.asp