社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2005－CG－120 (2)
2005／8／18

# イメージ合成装置を用いたシミュレーションと可視化の並列処理における通信コストの評価

緒方 正人† 菊川 孝明† 梶原 景範†

我々は，レンダリング及び計算をグラフィックボードと専用ハードウェアで高速化した，ボリュメトリックコンピューティンググラフィッククラスタ，VGCluster-Cluster システムを開発した．本システムの主な目的は，対話的にシミュレーションと可視化を行い，現象の直感的な把握を可能にすることである．並列処理に伴う空間分割は，シミュレーション及び可視化においてデータ通信が必須となり，性能の低下をもたらす．本システムの有効性を実験を行い評価した．実験結果に基づきパーフォーマンスモデルを検討し，実時間処理のための新しい性能基準を提案する．

# An Evaluation of Communications Cost for Simultaneous Processing with Simulation and Visualization using an Image-Composition Device

MASATO OGATA,† TAKAAKI KIKUKAWA† and KAGENORI KAJIHARA†

We have been developing a volumetric-computing graphics cluster system, a PC cluster with enhanced rendering and calculation by video boards and dedicated devices. The main purpose of the system is to interactively perform simulation and visualization in order to intuitively understand physical phenomena. The higher the computational power requested is, the larger the required number of PCs becomes, which inversely affects communication costs and creates a bottleneck for communications with both parallel simulation and parallel visualization. We study the effects of both types of communications by experiments. Based on the results, we discuss a performance model and propose a new performance metric for time-critical processing. We evaluate our VGCluster-cluster system using the proposed metric. The metric shows the effect of sustaining scalability by using a dedicated image-composition device.

## 1. はじめに

流体，ガス，気象など多くの自然現象のシミュレーションにおいては，摸擬対象はボリュメトリックである．これまで，複雑な自然現象を解明するために，ボリュメトリックなモデルを用いた大規模なシミュレーションが行われてきた．そのようなシミュレーションにおいて結果を直感的に把握するには，シミュレーションとその結果の可視化がインタラクティブに行えることが重要である．このことによって，思考を中断することなく実験をスムーズに繰り返すことができる．

近年の半導体や光学素子の急速な進歩によって，PC クラスタの性能が向上し，かつ安価に入手できるようになってきた．このような状況のもとで，伝統的な PC クラスタすなわち専用ハードウェアを持たないシステムで，インターラクティブなシミュレーションシステムすなわちシミュレーションと可視化が同時に行えるシステムが可能だろうか，という素朴な疑問を抱いたのが本研究の動機である．

本論文において，シミュレーションと可視化の並列処理における通信の影響について研究する．並列処理の実験を行い，結果を評価し，並列処理のパーフォーマンスモデルと評価メトリックについて考察する．

## 2. 従来研究

多くの大規模シミュレーションでは，数学モデルの計算にスーパコンピュータが使われ，その結果を SGI の Onyx のようなハイエンドのグラフィックワークステーションを用いてオフラインで表示している[6),10),14)]．この方式のシステムでは，地球シミュレータが有名である[14)]．伝統的なスーパコンピュータは，演算についてはきわめて強力であるが，シミュレーション結果の対話的な可視化には適していない．対話的なシミュレーションを行おうとすると可視化のための大量の通信を必要とし，通信がボトルネックとなる．

† 三菱プレシジョン株式会社
Mitsubisi Precision Co., Ltd.

近年，これまでスーパコンピュータが用いられた分野に，PC クラスタの利用が多くなってきた．これまで大規模なボリュームデータの対話的な可視化について報告されてきたが，最終的な目標は対話的なシミュレーションすなわち演算と可視化の同時処理である．対話的なシミュレーションと可視化を実現するために，画像合成用の専用装置を提案されている[2),13)]．筆者らは，文献[13)]において，空間分割による並列処理において，可視化のための通信による性能低下について評価し，文献[2),4),9),11)]に述べられているような対話的な可視化装置の実現には画像合成のための専用ハードウェアが必要性であることを提案した．

シミュレーションと可視化の同時処理を実現するための実験システムや提案が発表されている[1),3),8),12)]．GPU クラスタ[1)]は，数値計算を加速するために GPU が用いられている．480×400×80 格子の Boltzmann モデルのシミュレーションと可視化の同時処理を，30 台の PC を用いて，0.32 秒/ステップで達成している．

PC の台数を増やして行くことによってビデオレートでのシミュレーションと可視化の同時処理が可能かという疑問が生じるが，これまで高並列処理における通信のボトルネックの評価がなされてこなかった．本論文では，シミュレーションと可視化の同時処理における通信のボトルネックに関する実験を行い，システムパーフォーマンスについて評価する．

## 3. 並列処理における通信

大規模シミュレーションに PC クラスタの使用が一般的になって来た．その場合，データは各 PC に分割され処理されるので，ネットワークによる通信が不可欠となる．そのようなシミュレーションにおけるシステム性能を考える場合通信コストが重要な問題となる．

### 3.1 シミュレーションにおける並列処理

大規模なボリュメトリックなモデルのシミュレーションと可視化を同時処理する場合，空間分割による並列処理を行うのが自然である．空間分割は必然的にデータ通信を必要とする．分割されたボリュームをサブボリュームと呼ぶ．図 1 は，隣接するサブボリュームの境界層であるゴーストボクセルスライスを示す．ゴーストボクセル層は 1 シミュレーションサイクルごとに隣接するサブボリューム間すなわち PC 間で交換される必要がある．

### 3.2 可視化における並列処理

シミュレーションにおける並列処理に対して，可視化における並列処理には種々の方式がある[7)]．空間分割に対しては，sort-last 並列レンダリングが最も適している[5),7)]．図 2 は，sort-last 並列レンダリングの画像合成の手順を示している．各 PC はサブボリュームの 2 次元画像を生成し，最終画像を得るには 2 つの画像を合成しながら 2 進木を下から上に辿る．従って，最終画像を合成するには PC 間の大量のデータ通信を必要とする．通信すべきデータ量は PC の台数と画像の大きさにに比例する．

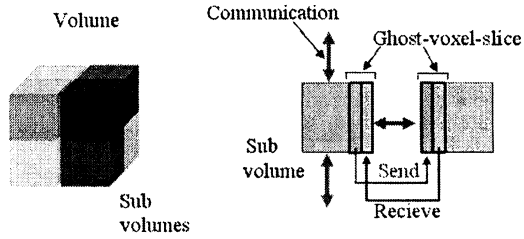



図 1 Space-partition for parallel simulation. The ghost-voxel slices are exchanged periodically between adjacent subvolumes

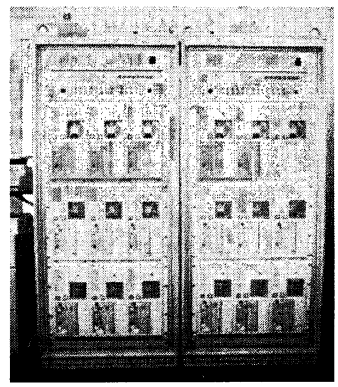


図 3 Overview of the VGCluster cluster.

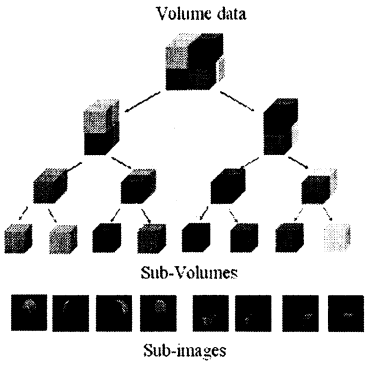
## 4. システム構成

### 4.1 VGCluster クラスタの構成

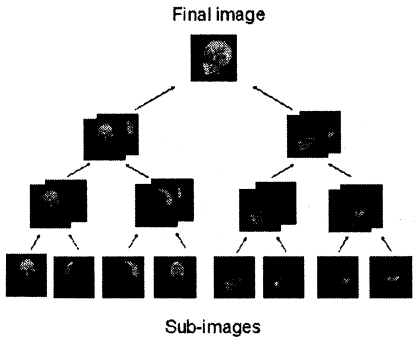
図 3 は，シミュレーションと可視化の並列処理における通信コストの評価に用いた VGCluster クラスタの概観を示す．VGCluster クラスタは，高速ネットワーク Myrinet を備えた 2 台の VGCluster と 1 台の専用画像合成装置から構成され，各 VGCluster はビデオボードを備えた 8 台のノード PC，1 台のホスト PC と専用画像合成装置で構成されている．その諸元と構成を表 1 及び図 4 に示す．

### 4.2 画像合成装置による並列レンダリング

並列度が増大するにつれ，PC 間の通信量が急激に増加しビデオレートでの可視化がソフトウェアでは困難になる[13)]．この問題を解決するために，筆者らは 8 入力の画像合成装置を開発した．図 5 は，VGCluster




(a) A volume is divided into subvolumes, then each PC generates 2D image.




(b) The composition of two images is repeatedly applied while traversing a binary tree from bottom to top.

図 2 Parallel rendering in the space-partition scheme.

表 1 Specification of the VGCluster cluster.

| No. | Item | VGCluster #1 and #2 |
|---|---|---|
| 1 | The number of PCs | 9 (8 nodes, 1 host ) |
| 2 | CPU | Xeron 2GHz x 2 |
| 3 | GPU | Geforce 4 |
| 4 | Memory | 2024MB |
| 5 | Network | Myrinet; 2Gbits/s |
| 6 | OS | Linux 7.2,Score5.0.1 |
| 7 | Image composition H/W | PCI32/33MHz |

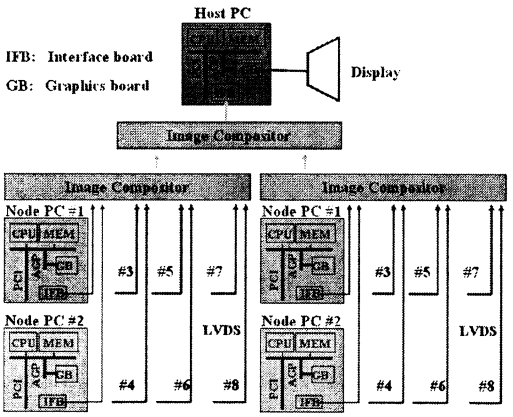



図 4 Schematic diagram of VGCluster cluster.

クラスタにおけるネットワークの通信量を低減する専用画像合成装置の概観を示す．本装置を階層的に接続することによって，512 入力まで拡張することができる．詳細は文献[9),13)] を参照されたい．

### 4.3 ソフトウェア構成

OS として Linux7.2，並列処理用ソフトとして SCore5.0.1 が用いられ，汎用性及び互換性の観点からレンダリングソフトとして OpenGL，通信ソフトとして MPI が用いられた．

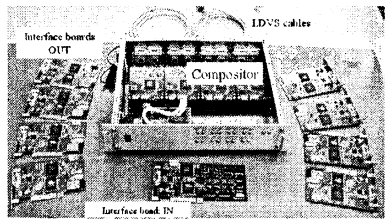


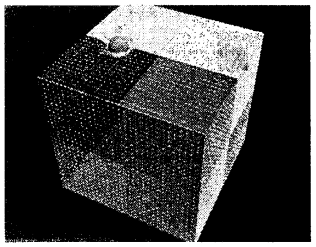

図 5 Overview of the Imgage composition device.

## 5. 実験及び性能評価

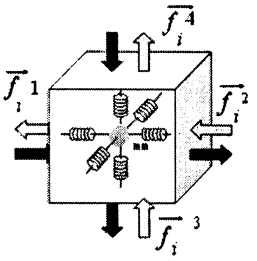
シミュレーションと可視化の同時処理の性能評価実験に，18PC の VGCluster クラスタが用いた．性能はシミュレーションと可視化の同時処理の更新レートで評価した．実験は，専用画像合成装置ありとなしの 2 つの条件下で，ノード PC を 2 台から 16 台まで変化させて行った．専用画像合成装置なし（今後，without-H/W という）の場合すなわち従来の方式では，シミュレーションと可視化における通信は Myrinet によって行われる．専用画像合成装置ありの場合（今後，with-H/W という）は，可視化に伴う通信負荷は専用画像合成装置が PCI bus 経由で担う．

### 5.1 シミュレーションモデル：ボリュームの変形

実験に，図 6 に示すばねモデルによるヒトの頭部の変形のシミュレーションを用いた．このシミュレーションは，ゴーストボクセルスライスの通信を必要とする．ボクセル間の力学的関係は式 (1) で表される．第 1 項，第 2 項及び第 3 項は，それぞれ粘性項，重力項及び弾性項である．グリッド数は $32^3$ グリッド，出力画像は $512^2$ ピクセルである．



(a) Example of simulation space partition scheme: The space is divided into 8 subspaces with 2 divisions along with each axis.




(b) Communications of a cellular automaton: At each boundary surface of subvolume, ghost-voxel-slices are exchanged between adjacent subvolume.

(c) Simulated image: Deformation of head

図 6 Spring model implemented using cellular automaton.

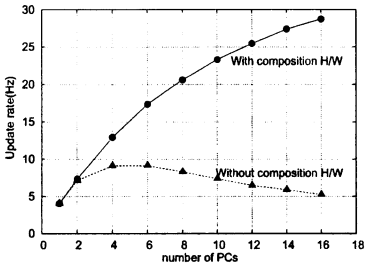



図 7 Comparing performance between the composition hardware and a tradition network.

$$\boldsymbol{f}_i = -\gamma_i \dot{\boldsymbol{x}}_i - m_i G - \sum_j k_{ij}(|\boldsymbol{x}_{ij}| - L_{ij}) \frac{\boldsymbol{x}_{ij}}{|\boldsymbol{x}_{ij}|} \quad (1)$$

where:

$\boldsymbol{f}_i$ is the external forece for grid $i$.
$\boldsymbol{x}_i$ is the position vector of grid $i$,
$m_i$ is the mass of grid $i$,
$\gamma_i$ is the viscosity at grid $i$,
$k_{ij}$ is Fuch's constant between grid $i$ and $j$,
$\boldsymbol{x}_{ij}$ is the vector from grid $i$ to grid $j$,
$L_{ij}$ is the intial length between grid $i$ and $j$.

### 5.2 実験結果

表 2 に実験結果を示す．図 7 はそのグラフで，横軸はノード PC の台数，縦軸は更新レートである．並列度が高くなるに従って with-H/W と without-H/W の差が大きく，専用画像合成装置の有効性を表している．

### 5.3 画像合成ハードウェアありとなしの性能比較

更新レートは，witout-H/W の場合 6 ノードでピーク 9.14Hz に達しその後減少するのに対し，with-H/W の場合は，大雑把に言ってノード数に比例し，16 ノード並列時に 28.75Hz である．without-H/W でのスケーラビリティの悪さはネットワークにおけるボトルネックが原因である．without-H/W でのネットワークの負荷はゴーストボクセルの情報の交換とサブイメージの転送であるのに対して，with-H/W ではゴーストボクセルの情報の交換のみである．

表 2 Result of experiment.

| Number of nodes | Average update rate: ( Hz ) | |
|---|---|---|
| | With H/W | Without H/W |
| 1 | 4.04 | 4.04 |
| 2 | 7.35 | 7.12 |
| 4 | 12.94 | 9.12 |
| 6 | 17.35 | 9.14 |
| 8 | 20.62 | 8.29 |
| 10 | 23.13 | 7.10 |
| 12 | 25.46 | 6.47 |
| 16 | 28.75 | 5.26 |

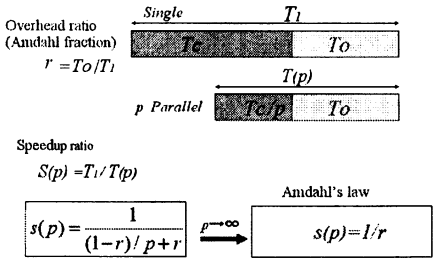



図 8 Theoretical limit by Amdharl's law.

ネットワークにおける画像転送の負荷は，without-H/W の場合，画像の大きさと並列度に比例するのに対し，with-H/W の場合は画像データは PCI bus 経由で画像合成 H/W に送られるのでネットワークの負荷は 0 である．ゴーストボクセルに関する情報の交換については次節で詳しく検討するが，並列度が大きくなると各ノード PC のゴーストボクセルの量は減少するので並列度に比例して増大することはない．

## 6. 結果に対する考察

### 6.1 パーフォーマンスモデル

図 8 は，並列プロセッサの数と speedup-ratio の関係を表す，有名な Amdahl の法則を示している．この

表 3 Estimation of parameters in performance model.

| No. | Parameter | Estimated Value $T_1 = 0.2473s$ : $v^2 = 32^2$; $m^2 = 512^2$ | |
|---|---|---|---|
| | | with H/W | without H/W |
| 1 | r | 0.078 | 0.075 |
| 2 | $cv^2$ | 0.00082 | 0.00082 |
| 3 | $bm^2$ | 0.0 | 0.0106 |

法則は，並列処理による speedup ratio は並列化し得ない処理量 $To$ の全体の処理量に対する割合 $r = To/T1$ によって制限されること表している．

しかし，我々の実験では専用ハードウェアなしの場合は更新レートは 0 に収斂し (図 7)，Amdahl の法則と矛盾する．これは，Amdahl の法則が通信コストを加味していないことによる．この状況を表現できるパーフォーマンスモデルを検討する．我々のパーフォーマンスモデルを式 (2) で表す．

$$T(p) = T_1 r + T_1(1-r)/p + cv^2(p^{\frac{1}{3}} - 1) + bm^2 p \quad (2)$$

where:

| | |
|---|---|
| $T(p)$ | parallel processing time with p PCs, |
| $p$ | number of PCs; $p \geq 2$, |
| $T_1$ | processing time on unit processor, |
| $r$ | overhead ratio $T_o/T_1$ or Amdahl fraction, |
| $v$ | number of voxles for each axis, |
| $cv^2$ | communications cost for boundary voxels in simulation; A volume consists of $v^3$ voxels, |
| $bm^2$ | communications cost for subimage in rendering, A screen consists of $m^2$ pixels. |

式 (2) の第 1 項及び第 2 項は Amdahl のモデルと同じである．第 3 項及び第 4 項は，シミュレーションと画像合成のための通信負荷に対応する項である．詳細は apenndix A を参照されたい．speedup ratio $S(p)$ は式 (3) で定義され，並列処理の効率を表す Efficiency $E(p)$ は式 (4) で定義される．

$$S(p) = T_1/T(p) \quad (3)$$

$$E(p) = S(p)/p \quad (4)$$

表 3 は，表 2 の実験結果から式 (2) のパラメータを最小 2 乗法に基づいて推定した値である．図 9 は，$S(p)$ と $E(p)$ を，推定したパラメータを用いて with-H/W と without-H/W に関してプロットしたものである．

### 6.2 実時間処理用 Metric の提案

並列処理システムの Metric として一般に speedup ratio$S(p)$ と効率 $E(p)$ が用いられる．しかし，実時間シミュレーションのような実時間性の重要なシステムには，この Metric はあまり適していない．なぜなら，扱う問題を大きくして行くと $S(p)$ や $E(p)$ は改善されて行くが，システムの実時間性は損なわれて行く．

そこで，更新レートが 30Hz とか 60Hz に厳しく要求されるシステムに向く Metric，time-ristricted efficiency $E_c(p)$ を提案する．$E_c(p)$ を式 (5) で定義する．

$$E_c(p) = (T_{target}/T(p))E(p) = \mu S(p)E(p) \quad (5)$$

ここで，$T_{target}$ はアプリケーションに要求される更新時間であり，例えば 33.3ms である．図 10 は，$T_{target} = 33.3ms$ として提案の Metric と従来の Metric をプロットしたものである．with-H/W において，ほぼ 30Hz の更新レートを達成する 16 ノード構成で，従来の Metric が非常に低い値を示しているのに対して，提案の Metric が最大値に近い値を示している．

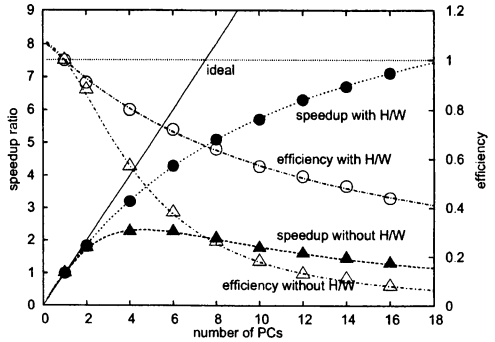



図 9 Comparing performance between the composition hardware and a tradition network.

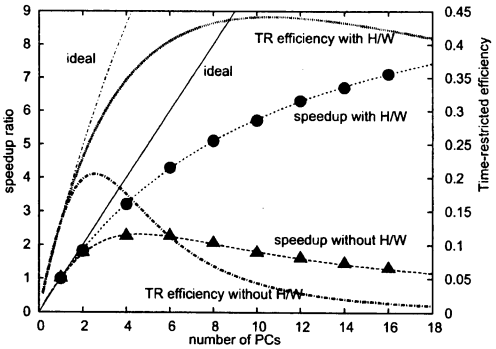



図 10 Parallel speedup ratio and time-restricted efficiency factor

## 7. おわりに

筆者らは，ビデオボードと専用画像合成ハードウェアを備えた volumetric-computing graphics クラスタを開発した．このシステムの主目的は，摸擬された自然現象を直感的に把握できるように対話的にシミュレーションと可視化を行うことである．クラスタシステムにおいては，要求される処理能力が高いほど PC の数が多くなり，PC 相互の通信がボトルネックとなる．

筆者らは，シミュレーションと可視化のための通信コストを実験で調べ，専用画像合成装置がシミュレー

ションと可視化の同時処理においても有効であることを明らかにした．また，実験に基づきパーフォーマンスモデルを考案し，提案モデルがシステムの性能をよく表すことを確認し，実時間並列処理用の新しいMetricを提案した．

本論文では，可視化のための専用ハードウェアを開発したが，今後計算能力を増強する専用ハードウェア例えばFEMを高速に計算するハードウェアを開発し，手術シミュレータなどに応用したいと考えている．

## 謝　辞

本研究の一部は独立行政法人 情報通信研究機構 (NICT) の助成の下に行われた．

## 参 考 文 献


1) Fan, Z., Qiu, F., Kaufman, A. and Yoakum-Stover, S.: GPU Cluster for High Performance Computing, *SuperComputing 2004*, ACM SIGARCH and IEEE Computer Societ, ACM Press (2004).
2) Heirich, A. and Moll, L.: Scalable distibuted visualization using off-the-shelf components, *IEEE Symposium on Parallel Visualization and Graphics*, IEEE Computer Society, IEEE CS Press, pp. 55–118 (1999).
3) Kruger, J. and Westerman, R.: Linear Algebra Operators for GPU Implementation of Mumerical Algorithms, *SIGGRAPH2003*, ACM SIGGRAPH, ACM Press (2003).
4) Lombeyda, S., Moll, L., Shand, M., Breen, D. and Heirich, A.: Scalable interactive volume rendering using off-the-shelf components, *IEEE Symposium on Parallel and Large-Data Visualization and Graphics*, IEEE Computer Society, IEEE CS Press, pp. 115–1158 (2001).
5) Ma, K.-L., Schussman, G., Wilson, B., Ko, K., Qiang, J. and Ryne, R.: Advanced visualization technology for terascale particle accelerator simulations, *SuperComputing 2002*, ACM SIGARCH and IEEE Computer Society, ACM Press (2002).
6) Matsuo, Y. and Tsuchiya, M.: Early Experience with Aerospace CFD at JAXA on the Fujitsu PRIMEPOWER HPC2500, *SuperComputing 2004*, ACM SIGARCH and IEEE Computer Society, ACM Press (2004).
7) Molnar, S., Cox, M., Ellsworth, D. and Fuchs, H.: A Sorting Classification of Parallel Rendering, *IEEE CG & Application*, Vol. 14, No. 4, pp. 23–32 (1994).
8) Muraki, S., B.Lum, E., Ma, K.-L., Ogata, M. and Liu, X.: A PC Cluster System for Simultaneous Interactive Volumetirc Modeling and Visualization, *IEEE Symposium on Parallel and Large-Data Visualization and Graphics*, IEEE Computer Society, IEEE CS Press, pp. 95–102 (2003).
9) Muraki, S., Ogata, M., Ma, K.-L., Koshizuka, K., Kajihara, K., Liu, X., Nagao, Y. and Shimokawa, K.: Net-Generation Visual Supercompusting using PC Clusters with Volume Graphics Harware Devices, *SuperComputing 2001*, ACM SIGARCH and IEEE Computer Society, ACM Press (2001).
10) Nakano, A., Kalia, R. K. and Vashishta, P.: Scalable Atomaistic Simulation Algorithms for Materials Resarch.
11) Nonaka, J., Kukimoto, N., Sakamoto, N., Hazama, H., Watashiba, Y., Ogata, M., Kanazawa, M. and Koyamaza, K.: Hybrid Hardware-Acceleratated Image Composition for ort-Last Parallel Rendering on Graphics Clusters with Commodity Image Compositor, *Volume Graphics 2004*, IEEE Computer Society, IEEE CS Press, pp. 17 – 24 (2004).
12) Ogata, M., Kajihara, K., Kurita, T. and Fujishiro, I.: Volumetric Computing Graphics Cluster, *VSMM2004*, International Society on Virtual Systems and MultiMedia, IOS Press,Inc., pp. 220–224 (2004).
13) Ogata, M., Muraki, S., Ma, K.-L. and Liu, X.: The Design and Evaluation of a Pipelined Image Composition Device for Massively Parallel Volume Rendering, *Volume Graphics 2003*, Eurographics Organization, Eurographics Organization, pp. 61–68 (2003).
14) Oliker, L., Carter, A. C. J. and Shalf, J.: Scientific Comuputations on Modern Parallel Vector Systems, *SuperComputing 2004*, ACM SIGARCH and IEEE Computer Society, ACM Press (2004).


## 付　録

### A.1 パーフォマンスモデル

空間分割による並列処理のパーフォーマンスモデル式 (2) を導出する．処理時間は次式で表される．第1項及び第2項はAmdahlの公式と同じである．第3項はシミュレーションのための境界ボクセル情報の通信コスト，第4項はサブ画像の合成のための通信コストである．

$$T(p) = rT_1 + (1-r)T_1/p + c^* N_p + bp \qquad (6)$$

where:

| | |
|---|---|
| $T(p)$ | parallel processing time with p PCs, |
| $p$ | number of PCs; $p \geq 2$, |
| $T_1$ | processing time on unit processor, |
| $r$ | overhead ratio $T_o/T_1$ or Amdahl fraction, |
| $c^*$ | coefficient of communications cost for simulation, |
| $N_p$ | number of separating planes, |
| $b$ | coefficient of communications cost for visualization. |

ボリュームは平面で分割されると仮定したので平面の数とノードPCの数との間に次式が成り立つ．

$$p = (N_x + 1)(N_y + 1)(N_z + 1) \qquad (7)$$

$$N_p = N_x + N_y + N_z \qquad (8)$$

ここで，$N_x$，$N_y$ 及び $N_z$ は，それぞれx軸，y軸及びz軸に垂直な分割面の数である．次の不等式が成り立つ．

$$(N_x+1)+(N_y+1)+(N_z+1)/3 \geq ((N_x+1)(N_y+1)(N_z+1))^{(1/3)} \qquad (9)$$

従って，次式が成り立つ．

$$N_p \geq 3(p^{1/3} - 1) \qquad (10)$$

上式と式 (6) から，$T(p)$ の下限は，$3c^*$ を $c$ と記して，

$$T(p) = rT_1 + (1-r)T_1/p + c(p^{\frac{1}{3}} - 1) + bp \qquad (11)$$