グラフィクスとCAD 68－6
（１９９４．５．２０）

高並列コンピュータ AP1000 の分散ディスクビデオ ハードウエアとソフトウエア

大江 和一　稲野 聡　佐藤 弘幸
(株)富士通研究所 並列処理研究センター

概 要

高並列コンピュータＡＰ１０００の大容量２次記憶と画像出力機能を実現するために分散ディスクビデオハードウエア（Distributed Disk and Video: ＤＤＶ）を開発した。また、このハードウエアを利用して画像のファイル入出力、及びモニターへの表示等の機能をサポートするイメージファイルシステムの開発も併せて行なった。本論文では、このハードウエア、及びソフトウエアの詳細について述べ、性能について評価を行ない、その有効性を示す。

Distributed Disk and Video hardware and software for the AP1000 highly parallel computer

Kazuichi Ooe, Satoshi Inano, and Hiroyuki Sato
Fujitsu Laboratories Ltd.
e-mail: [ooe, inano, or hsat]@flab.fujitsu.co.jp

Abstract

We developed the distributed disk and video (DDV) hardware for the AP1000 highly parallel computer to realize high-performance I/O and video output facilities. We also developed Image-filesystem software that can access image files on DDV and display images on cell memory or DDV disks to a monitor connected with DDV. In this paper, we present the architecture of DDV hardware and the performance of Image-filesystem.

1 はじめに

並列計算機の普及に伴って、今までは不可能だった大規模シミュレーションが可能になってきている。この場合、CPU能力の増大に比例したＩ／Ｏ能力の強化が必須である。また、大規模シミュレーションで得られたデータを評価する場合、可視化処理が重要になってくる。特にシミュレーションで得られた結果をリアルタイムに可視化して評価したり、可視化した結果を連続した画像として蓄え、後でアニメーションにして評価する機能が重要視されている。

我々は、これら２つの要件を解決する目的で、AP1000用に分散ディスクビデオハードウエア（Distributed Disk and Video: ＤＤＶ）を開発した。このＤＤＶには、それぞれ分散フレームメモリとSCSIインターフェースが実装されており、上記２要件を解決するアーキテクチャを実現している。

本稿では、このＤＤＶのハードウエアとソフトウエアの実現方法、及びその性能評価について述べる。2節でDDVハードウエアの概説を行ない、3節でDDVソフトウエアの設計要件、実現方法、及び性能評価について述べ、4節を結論とし、5節にて今後の課題を述べる。

2 AP1000の分散ディスクビデオハードウエア

高並列コンピュータAP1000は、分散メモリ型の並列計算機であり、最大１０２４個のセルと呼ぶプロセッサから構成される（図1）。セル群は、ブロードキャストネットワーク(B-net)、トーラスネットワーク(T-net)、同期ネットワーク(T-net)の３種類のネットワークからなる。

このような多数のプロセッサを効率的に動作させるためには、高性能なプロセッサ間ネットワークに加えて、高性能な入出力性能が必要になる。AP1000のようにスケーラブルな（最小１６セルから最大1024セルまで拡張可能）並列計算機の入出力機構は、プロセッサの処理速度に見合ったスケーラブルな入出力性能が求められる。我々は、画像表示出力とディスクへの高速アクセスを可能にするAP1000の分散ディスクビデオハードウェア（ＤＤＶ）を開発した。

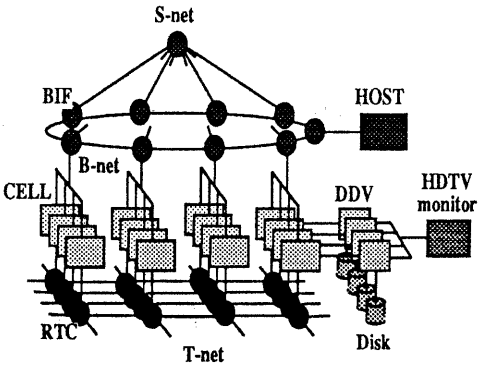


図 1: AP1000 アーキテクチャ

画像表示のために、各セルから直接アクセスが可能な、分散フレームメモリ(Partitioned Frame Memory: PFM)を実装した。また、この分散フレームメモリ上の画像データを高解像度のモニターに高速出力するため画像収集機構を開発した。また、ディスクは、全体としてスケーラブルな転送性能が得られるように、各ＤＤＶに個別に持たせ、分散させた。

今回開発したオプションハードウェアは、AP1000に付加することで以下のような機能向上が図れる。

1. 高解像度画像のリアルタイム(30フレーム／秒)の画像表示
2. 大容量、高速なディスクへのアクセス
3. 分散ディスク上の画像データのアニメーション表示

オプションハードウェアは、オプションユニット、ディスク装置、表示フレームメモリユニット(Display Frame Memory: DFM)で構成されている（図2参照）。個々のオプションユニットは、それぞれセルにローカルバス(LBUS)経由で、ディスク装置にSCSIバスで、DFMへはビデオバス(VBUS)で接続される。分散ディスクビデオハードウェア（ＤＤＶ）の仕様を表１に示す。

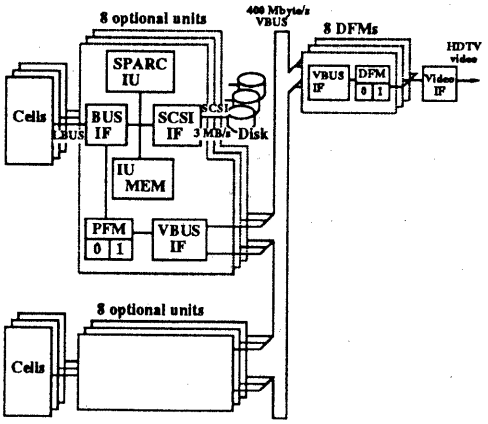


図 2: 分散ディスクビデオハードウエア

Parameter	Specifications	
Display resolution (pixels)	NTSC	720 * 486
	HDTV	1920 * 1035
Video interface	NTSC	D-1 component digital RGBA component digital RGBAS analog
	HDTV	Component digital RGBA component digital RGBAS analog
Number of outputs	One or two	
Pixel gradation	8 bits per r, g, b, and a attribute (total of 32 bits/pixel)	
Display mode	Single/double buffer	
Display refresh rate (Mbyte/s)	PFM to DFM: 400 (peak)	
	Disk to PFM: 3 per drive	
Disk drivers (Mbyte)	1,100 each	
System configuration	16, 32, 64, 128, 256, and 512 options (# options must be less than or equal to # cells)	

表 1: Video and disk specification

2.1 分散ディスク

各オプションユニットに各々SCSIインターフェースが実装されており、そこに1.1 GB 3.5インチのハードディスク装置が接続される。最大ディスク転送レートはディスク当たり、3 MB/sである。

データ転送は、ディスクとセルのメモリ、分散フレームフレームメモリ(PFM)、そしてIUのメモリ間で可能である。

2.2 ビデオ

ＤＤＶのPFMは、それぞれのセルによって書き込まれた画像データを保持している。複数のPFMに分散された画像データはVBUSを通じてDFMに書き込まれる。個々のPFMは、1MBづつの2バンクのメモリからなり、ダブルバッファとして使える。片方がセルからアクセスされている間に、他方から画像をDFMに転送する。片方のバンクのみを選択し、シングルバッファとして使うこともできる。

VBUSは、8チャンネル構成で、リング状に結ばれ、PFMとDFM間のデータ転送路として使う。各チャンネルは、50MB/sの最大転送速度をもち, 8チャンネル合わせてHDTVレベルの高解像度画像(1920 x 1035画素) のリアルタイム表示(30フレーム／秒) が可能な400 MB/sの最大転送速度をもつ。

DFMは、全体で16MBの容量があり、8グループに分割されている。個々のグループは、PFMと同様に1MBづつの2バンクのメモリからなりダブルバッファとして使われる。片方のバンクがVBUSからアクセスされ、もう一方から表示データが出力される。ビデオインターフェースは、HDTV, NTSC, PALフォーマットのそれぞれについて、アナログとデジタルの信号出力ポートを備えている。

2.3 画像表示

画像表示の機能を図 3により説明する。画像データは図に示すようにスクリーン上の縦ラインの単位でＤＤＶのPFMに分散されている。具体的には、全ＤＤＶ数をPとしたとき、N番目のＤＤＶはN = X modulo PであるX番目の縦ライン（複数）の画像データを担当する。例えば図 3では、Pは１６で、ＤＤＶ０は、0, 16, 32, ...のXアドレスの縦ラインを、ＤＤＶNは、N, N+16, N+32,...のXアドレスの縦ラ

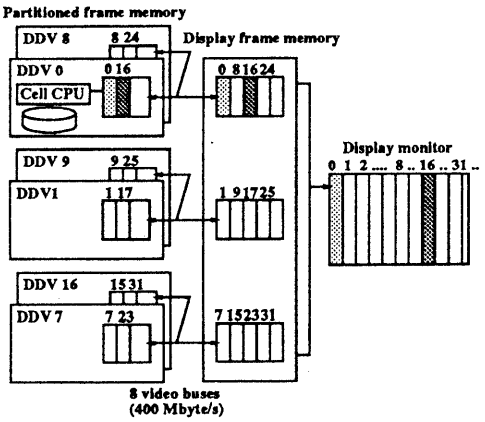


図 3: 画像出力系ハードウエア

インを担当する。

DFMは、VBUSに対応してやはり8チャンネルからなる。画像データは8チャンネルに8画素おきの縦ライン分割で分散される。

画像表示機能は以下のシーケンスで実行される。

1）PFMへの画像の書き込み

個々のセルは、縦ライン分割された画像の一部を生成し、ＤＤＶのPFMに書き込む。画像生成プログラムは、縦ライン分割と異なる分割で画像を生成した場合には、縦ラインへ分割変更処理を事前に行なっておく。PFMは、ダブルバッファになっているので、書き込み側に設定されているメモリバンクが使用される。

2）PFMからDFMへのデータ転送

表示リクエストがすべてのプロセッサから発行された時、画像データは、PFMからDFMに8チャンネルのVBUSを通じて並列に転送される。それぞれのチャンネルでは、リングの接続順に出力する。例えば、図3ではＤＤＶ０および８がグループ化されてチャンネル０のリングに接続されている。最初にＤＤＶ０がデータ転送を行ない、次にＤＤＶ８がデータ転送を行なう。このようなデータ転送が８チャンネル並列に同時実行される。

3）画像表示出力

新しい画像データのDFMへの書き込みが終わったら、DFMのフォアグラウンド（書き込み側）とバックグラウンド（読みだし側）が交換されて、新しい画像フレームが表示される。DFMはそれぞれFIFO(first-in-first-out)バッファを備えており、画像データは、FIFOバッファに書き込まれる。バッファからの出力は、表示のシーケンスに従って適当な表示クロックのタイミングで読まれ、８画素毎にマルチプレックスされ、ビデオ信号として出力される。

4）分散ディスクからの画像表示

個々のディスクは、縦ライン分割された画像の一部をあらかじめ保持しておく。表示コマンドが全セルから発行されると、オプションユニットのIUが画像データをディスクから読み、DMAでPFMに転送する。この後、上記2）、3）の処理を行なって画像が表示される。以上の処理を連続する画像フレームデータに対して行ない動画表示を行なう。

3 AP1000の分散ディスクビデオソフトウエア

上記ハードウエアを用いて目的とする高速・大容量Ｉ／Ｏ機能と可視化機能を実現する手段としてローカルファイルシステムとイメージファイルシステムを開発した。

ローカルファイルシステムは、UNIXと同じユーザインターフェースを採用しており、汎用ファイルを扱うことを目的とする。

イメージファイルシステムは、独自のユーザインターフェースを採用しており、イメージファイルのディスクへの格納・モニターへの表示等を可能にすることを目的とする。

本節では、この内イメージファイルシステムについて 設計要件、実現方法、及び性能評価について述べる。

3.1 設計要件

イメージファイルシステムは、ＤＤＶハードウエアの画像出力機構、及びディスクアクセス機構を利用して以下の機能を提供することを目

的とする。

- イメージファイルのディスクへの格納、及びディスクからの読み出し。
- セルでの生成画像をＤＤＶに接続されたモニターへリアルタイム表示。
- ディスクに蓄えられた連続画像をモニターへアニメーション表示。

これらの機能をユーザが高速、且つ容易に利用出来るようにするため以下のような要件を満たすべく設計した。

1. ユーザにＤＤＶのシステム構成を意識させず、且つ全セルで分割して画像生成を可能にする。（要件 1）
2. システムに接続された全ディスクを用いて、最大Ｉ／Ｏバンド幅を得る。（要件 2）
3. ディスクの回転待ち、シーク等のオーバーヘッドを最少にする。（要件３）
4. ユーザ～システム間のデータ転送に伴う通信オーバーヘッドを最少にする。（要件 4）
5. 複数タスクからの同時アクセス要求を可能にする。（要件 5）

3.2 実現方法

3.1 節で示した設計要件の全てを満たすべく設計を行なった結果 図 5のようなシステム構成を採用した。以下、各要件の実現に関してその詳細述べる。

3.2.1 要件 1

この要件を満たすため、イメージファイルシステムでは以下のようなユーザインターフェースを採用した。

- SPMD (Single Program Multiple Data) モデルのタスクを対象にしており、全セルで同時に関数を呼び出す。
- 全画像を縦ラインで分割し、それをセル数でのround-robin orderで分散したデータを各セルが生成。（図 4参照。）

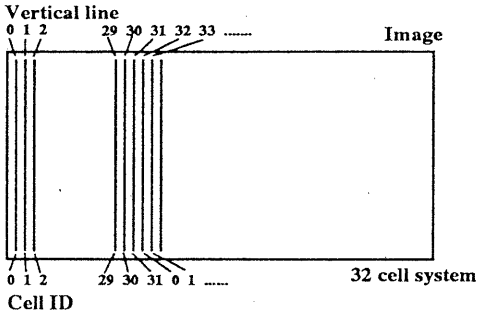


図 4: ユーザーインターフェースでのデータ分散

3.2.2 要件 2

システム内でのデータ分散に関してシステム内の全ＤＤＶディスクを利用し、イメージデータを縦ラインで分割しＤＤＶ数でのround-robin orderで分散して格納する方式を採用した。（図 6）この方式の採用により、ファイルアクセス時システムに接続されている全ディスクをアクセスすることになり、最大Ｉ／Ｏバンド幅を取り出すことが可能になった。さらに、アニメーション表示時、この縦ライン分割で分散を行なったため、ディスクから取り出したイメージデータを直接フレームメモリに書き込むことが出来、高速な画像出力が可能となった。

セル数とＤＤＶ数が一致しないシステムでは、図 7に示すようにＤＤＶ付きのファイルサーバがデータの合成／分散を行なうことによってユーザへのシステム構成の透過性と最大Ｉ／Ｏバンド幅を取り出すことの両方を実現した。

3.2.3 要件 3

この要件を満たすため、ファイル生成時にディスクの連続領域を予約するファイルシステム構造を採用した。この構造によりファイルアクセス時のシーク回数は、最大１回で済む。

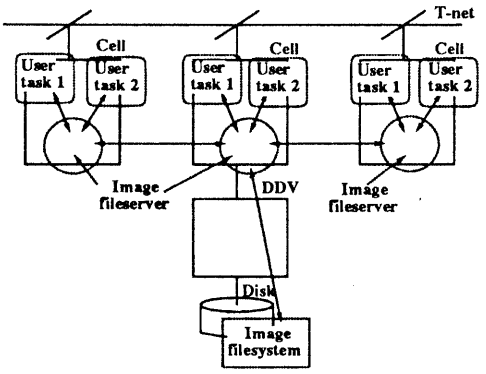


図 5: システム構成

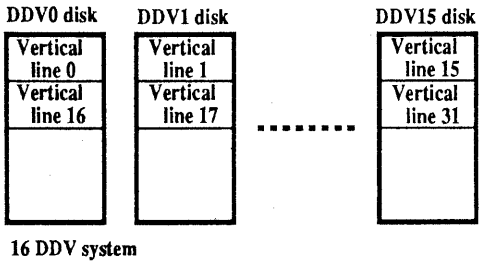


図 6: 各分散ディスクでのデータ分散

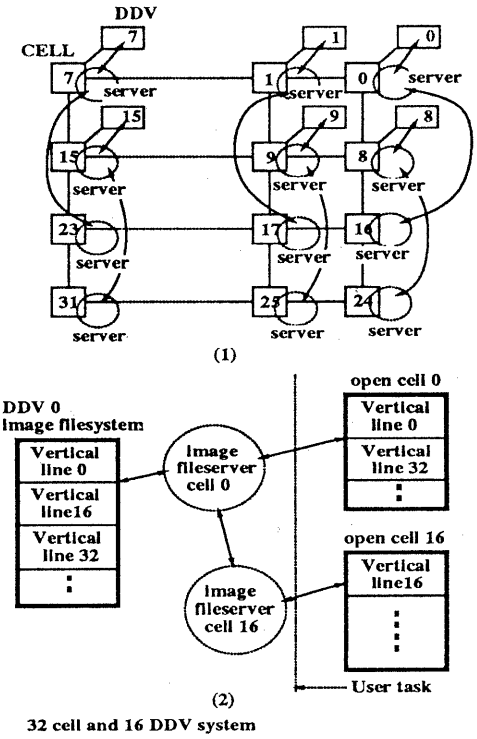


図 7: システム内でのデータの流れ

3.2.4 要件 4

要件4を満たすため、ファイルサーバからユーザタスクのデータ領域に直接データを転送する通信方式の採用を行なった。この構造の採用により、全セルにＤＤＶが接続されたシステムではディスクからユーザのデータ領域に直接アクセスが可能となった。

3.2.5 要件 5

複数タスクからの同時アクセスを可能にするため、サーバタイプで実現を行なった。（図5参照。）

3.3 性能評価

性能評価は、６４セル６４ＤＤＶシステムと６４セル３２ＤＤＶシステムを用いて、ユーザタスクからのファイルアクセス性能について行なった。以下、それぞれのシステムでの性能評価について報告する。

3.3.1 ６４セル６４ＤＤＶシステム

図 8にリード時の性能を示す。縦軸がシステム全体でのＩ／Ｏバンド幅を表し、横軸がシステム全体で読み込んだデータサイズを表す。Idealは、ドライバーを直接アクセスした時の性能を表す。Activeは、このイメージファイルシステムを用いて読み込んだ場合の性能を表す。このActiveには、ディスクアクセス処理(Ideal)以外にファイルシステム処理、そしてタスク切替え処理が含まれる。Image sizeが２MBの方が４MBの時より性能が良いのは、ディスクキャッシュヒットのためである。８MBの画像（１９２０×１０３５）を読み込んだ場合で、Idealの90%の性能がユーザに提供出来る。

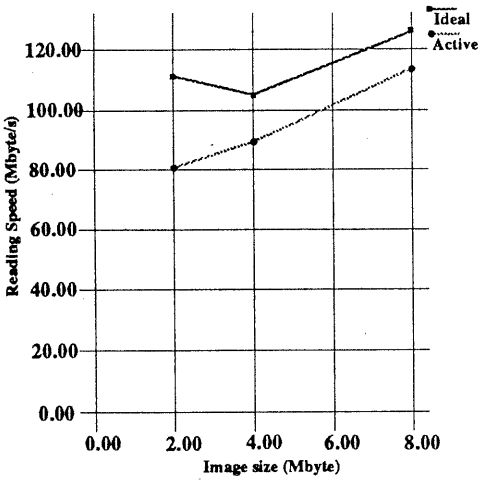


図 8: リード性能 (64cell+64DDV)

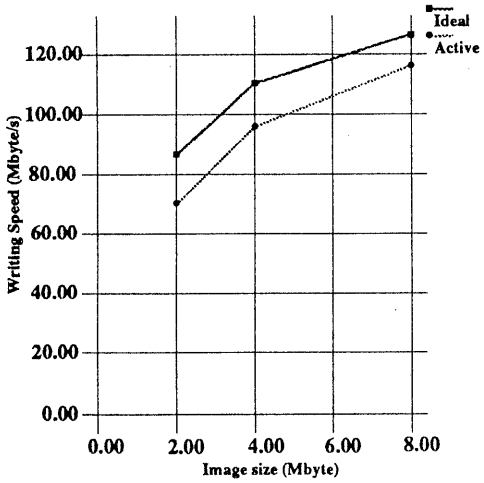


図 9: ライト性能 (64cell+64DDV)

図 9にライト時の性能を示す。ライトの場合でも、8MBの画像を読み込んだ場合で、Idealの92%の性能をユーザに提供出来る。

3.3.2 64セル32DDVシステム

図 10にリード時の性能、図 11にライト時の性能を示す。8MBの画像をアクセスしたケースで各々Idealの54%と65%の性能がユーザに提供出来る。

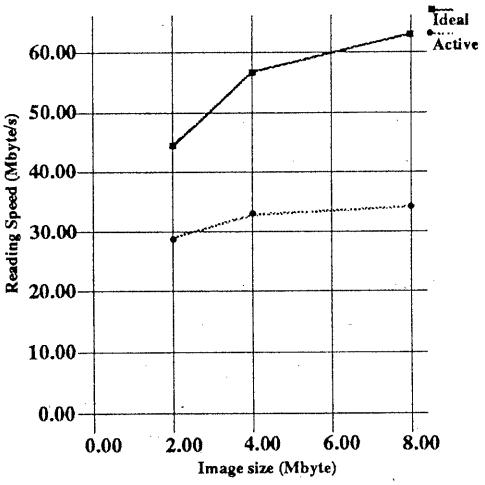


図 10: リード性能 (64cell+32DDV)

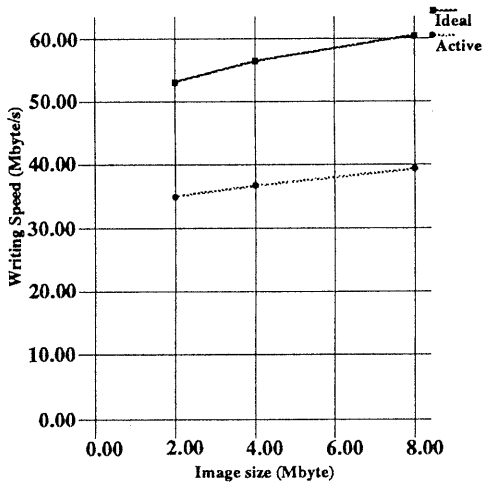
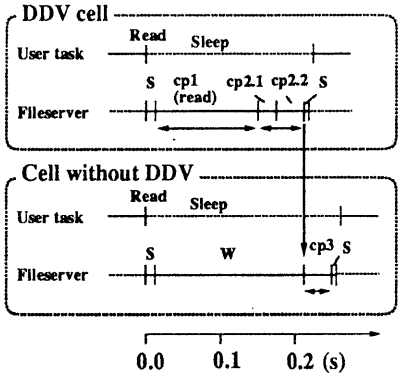


図 11: ライト性能 (64cell+32DDV)

64DDVシステムに比べて性能が低い原因は、データ読み込み時に生じるデータ分散・通信等による一時的なコピーである。図 12にリード処理の内部解析結果を示す。これより、ディスクアクセス以外のオーバーヘッドの大部分が一時的なコピーによるものであることがわかる。



S: Task switching and filesystem management
cp1: Copy from disk to temp.buffer 1
cp2.1: Stride read and copy from temp.buffer 1 to DDV cell's user buffer
cp2.2: Stride read and copy from temp.buffer 2, and to a cell without DDV
w: Wait
cp3: Copy from message buffer to user buffer

図 12: リード処理の詳細

4 結論

高並列コンピュータAP1000の大容量２次記憶、及び可視化機能を実現するハードウエアとして、分散ディスクビデオハードウエア（ＤＤＶ）を開発した。この分散ディスクビデオハードウエアによって、AP1000にて大容量Ｉ／Ｏ処理、リアルタイムな画像出力、及びアニメーションが可能になった。

この分散ディスクビデオハードウエアを駆動するソフトウエアとして、イメージファイルシステムの開発を行なった。性能評価を行なった結果、全セルにＤＤＶが接続されたシステムにおいて、ハードウエアが提供するＩ／Ｏバンド幅の90％（６４セル６４ＤＤＶシステムで、118 Mbyte/s）をユーザに提供出来ることがわかり、イメージファイルシステムの高速性が証明された。

全セルにＤＤＶが接続されていないシステムでは、ハードが提供する性能の６割程度しかユーザに提供出来ないことがわかった。この原因は、内部でのデータの合成・分散や通信に伴う一時的なコピーであることもわかった。

5 今後の課題

イメージファイルシステムにおいて全セルにＤＤＶが接続されていないシステムでは、データの合成・分散や通信に伴う一時的なコピーのためハード性能の６割程度しかユーザに提供出来ない。この問題を解決するためには、

- ディスクから読み込んだデータを複数のメモリ領域に一定間隔で分割して格納する機能。さらに、分割するデータの一部もしくは全部が他セルで用いるデータの場合、そのままメッセージパケットとして目的セルに送出する機能。

  また、複数メモリ領域のデータ、もしくはネットワーク経由で受信した他セルからのパケットデータを一定の割合で合成して、ディスクに書き込む機能。(ストライドディスクアクセス方式)
- PUT/GETを利用したデータ通信。

が有効と考えられる。今後、これらの機構を実装したシステムを用いて再度評価し、報告を行ないたい。

謝辞

日頃より御指導，御助言を頂いている，並列処理研究センター石井センター長，白石担当部長，池坂主任研究員，斉藤主任研究員，ならびに研究室の同僚諸氏に感謝いたします．

参考文献

[1] H. Ishihata, T. Horie, S. Inano, T. Shimizu, and S. Kato, "An Architecture of Highly Parallel Computer AP1000," *IEEE Pacific Rim Conf. on Communications, Computers, and Signal Processing* , May 1991, pp.13-16.

[2] H. Sato, S. Inano, and H. Yoshijima, "Parallel Visual Computing on the AP1000: Hardware and Software," *FUJITSU Scientific and Technical Journal Vol.29, No.1* (Kawasaki, Japan), March. 1993.