分散共有メモリ型超並列計算機JUMP-1におけるスケーラブルI/Oサブシステムの構成

中條拓伯[†1] 中野智行[†1] 松本尚[†2]
小畑正貴[†3] 松田秀雄[†4]
平木敬[†2] 金田悠紀夫[†1]

本論文では，分散共有メモリ型超並列計算機JUMP-1の入出力サブシステムの構成について述べる．JUMP-1は，4つのCPUから構成される複数のクラスタをRDTと呼ばれる階層トーラス・ネットワークで接続した分散共有メモリ型のアーキテクチャをとる．クラスタと画像/ディスク入出力ユニットの間は，STAFF-Linkと呼ばれる高速シリアルリンクにより接続される．本稿では，JUMP-1の入出力サブシステムの構成と特徴およびSTAFF-Linkの概念と実現方法について述べ，ディスク入出力サブシステムを構成するディスク入出力ユニットのハードウェア構成について述べる．また，共有入出力バッファを用いたディスク入出力サブシステムへのアクセス方式について説明し，その基本的な性能評価を行った結果について報告する．

A Scalable I/O Subsystem of a Distributed Shared-Memory Massively Parallel Computer JUMP-1

HIRONORI NAKAJO,[†1] TOMOYUKI NAKANO,[†1]
TAKASHI MATSUMOTO,[†2] MASAKI KOHATA,[†3] HIDEO MATSUDA,[†4]
KEI HIRAKI[†2] and YUKIO KANEDA[†1]

We summarize a configuration of an I/O subsystem of a distributed shared-memory massively parallel computer JUMP-1. JUMP-1 consists of multiple clusters connected by a broad bandwidth hierarchical torus inter-connection network called RDT network , and supports an efficient distributed shared-memory. We introduce a scalable I/O subsystem configuration which consists of image and disk I/O units connected via fast serial links called Serial Transparent Asynchronous First-in First-out Link (STAFF-Link). In this paper, we describe features of the scalable I/O subsystem of JUMP-1, a concept and an implementation of STAFF-Link and the hardware configuration of a disk I/O unit. Finally, an I/O access method using shared I/O buffer and also preliminary evaluation of each I/O access are presented.

1. はじめに

現在，さまざまな大学および研究機関において，将来の超並列計算機のアーキテクチャ，オペレーティングシステム，I/Oシステムやアプリケーションについて研究が進められている．

文部省科学研究費補助金・重点領域研究においても，分散共有メモリ型の超並列計算機のプロトタイプマシンJUMP-1[1)~9)]の開発が行われている．

JUMP-1は，4つのCPUから構成されるクラスタ間をRDT（Recursive Diagonal Torus）ネットワーク[10)]と呼ばれる階層トーラス型ネットワークで接続し，共有メモリによる通信機構や同期機構を専用ハードウェアであるMemory-Based Processor (MBP)[11)]により強力にサポートすることによって，局所処理と非局所処理を分離分割した効率の良い分散共有メモリ型のアーキテクチャをとる．

多数の要素プロセッサから構成される超並列計算機

†1 神戸大学工学部情報知能工学科
Department of Computer and Systems Engineering, Faculty of Engineering, Kobe University
†2 東京大学大学院理学系研究科情報科学専攻
Department of Information Science, Faculty of Science, The University of Tokyo
†3 岡山理科大学工学部情報工学科
Department of Information and Computer Engineering, Faculty of Engineering, Okayama University of Science
†4 大阪大学基礎工学部情報工学科
Department of Information and Computer Sciences, Faculty of Engineering Science, Osaka University

システムにおいて，サイエンティフィック・ビジュアライゼーションのための画像I/Oシステムやディスク I/Oシステムを接続する場合に，十分な入出力バンド幅を確保しなければ，システム全体の性能を発揮することはできない．したがって，今後の高性能な超並列計算機のアーキテクチャについて研究するうえで，I/Oサブシステムの構成に対して十分考慮する必要があると考えられる．

今後の超並列計算機におけるI/Oサブシステムの研究目標として以下に示す項目をあげる．

- 拡張性（Scalability）
  拡張性はプロセッサの台数増加に対して，それに見合う性能向上として知られる指標である．そこで，I/Oサブシステムにおいても，ディスク等のI/O機器を容易に拡張でき，またそのことによりI/O性能も向上しなければならない．
- 柔軟性（Flexibility）
  今後の超並列計算機において入出力装置の拡張性を考慮した場合，種々の入出力装置の設置場所や接続形態に関しては柔軟性が要求される．
- アクセスの容易さ（Accessibility）
  超並列計算機において，以下に示す，
  - ファイル入出力を行うファイルシステム
  - ページングやスワッピングのためにディスクにアクセスするカーネル
  - 直接ディスクや画像I/Oアクセスを行うユーザプロセス

  などがディスク等のI/O機器にアクセスを行う．これらが多数のI/O機器に対して円滑にアクセスを行うためには，アクセスの容易さを確保する必要がある．そのためには，昨今の共有メモリアーキテクチャの商用ベースでの成功に見られるように，I/O機器に対するアクセスはメモリアクセスと同様に処理されるべきであると考える．
- 保護機能（Protection）
  多数のプロセッサで構成される超並列計算機を複数のユーザにおいて有効に利用するために，タイムシェアリングシステムを導入するのみならず，システム全体をいくつかのクラスタでパーティションに分割してユーザ/タスクに割り当てる．しかしながら，それぞれのパーティションにおけるI/Oアクセス時に，パーティション間で実行性能の影響が及ばないように，ハードウェア/ソフトウェアの両面から支援する必要がある．また，複数のユーザ間においてアクセス競合が生じた場合にも，それぞれの領域の破壊を防ぐための保護機能は必要不可欠である．

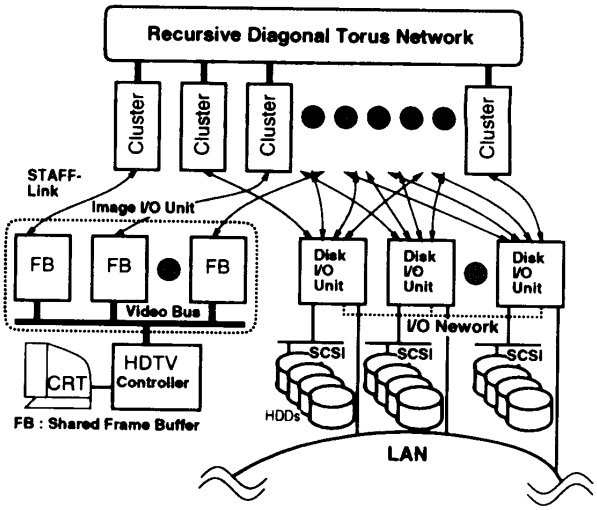

図1 超並列計算機JUMP-1のシステム構成
Fig. 1 System Configuration of JUMP-1.

以上の研究目標を達成するために我々は，図1に示すようなJUMP-1における，I/Oサブシステムのテストベッドの設計・開発を進め評価を行っている．

JUMP-1では，ディスクI/O装置や画像I/O装置を本体のクラスタと独立させることにより拡張性を確保する．そして，それぞれのI/O装置と本体の要素クラスタ上のMBP間をSTAFF-Link（Serial Transparent Asynchronous First-in First-out）と呼ばれる高速シリアルリンクを複数本用いて接続し柔軟性を持たせている．また，アクセスの容易性については，本体の分散共有メモリアーキテクチャの特徴を引き出すために，メモリアクセスと同様の方式でI/Oアクセスを行う．保護機能についてはMBPのメモリ保護機能を有効に活用することにより実現される．

本論文では，まずSTAFF-Linkの概念と構成について説明し，続いてJUMP-1のI/Oサブシステムの特徴と具体的なディスクおよび画像I/Oユニットの構成について述べる．さらに，ディスクI/Oユニットに関してデバイスドライバの役割について説明し，基本性能を測定した結果を報告する．最後に現状の実装での問題点とその改善の方策について述べる．

2. STAFF-Link

2.1 入出力インタフェースと狭域・広域ネットワーク

並列計算機システムにおいて，要素プロセッサ同士はバックプレーン上の高速なバスで接続したり，物理的に離れたルータ間をフラットケーブルなどで結ぶのが一般的である．しかしながら，多数の信号線を有するケーブルで高速にデータ通信を行う場合，ケーブル

の持つ誘導特性などから生じるノイズの影響によりケーブル長には限界が存在する．また，ルータのポート数の増加によるコネクタの基板上の占有面積などの空間的な制約から逃れることはできない．

従来，LAN，WANなどにおけるネットワーク技術と，計算機システムにおける入出力インタフェース技術は，独立にそれぞれの分野を形成してきた．しかしながら，ネットワークの転送速度が上がり，入出力インタフェースの通信距離に対する要求が高くなるにつれ，これらの技術を統合した形のインタフェースに期待が寄せられている．最近ではFibre ChannelやSSA (Serial Storage Architecture)，IEEE1394などが実際にディスク装置の接続インタフェースに利用されつつある[12),13)]．しかしながら，これらの次世代周辺装置インタフェースのほとんどは，従来のSCSIプロトコルに基づいて高速シリアル転送を行うものであり，SCSI機器の接続には適しているものの，汎用的な通信路としての用途には難がある．

そこで我々は，物理的，空間的な制約と転送スピードとのトレードオフを考慮したうえで，通信路の両端間においてFIFOメモリのようにアクセスできるような通信ハードウェアを実現し，計算機内の入出力インタフェースや，入出力装置間の通信機構への応用を考えた．その通信リンクを *Serial Transparent Asynchronous First-in First-out Link* (STAFF-Link) と呼ぶ．

技術的な背景として，B-ISDNやATM技術などの広域ネットワークの分野で開発された高速シリアル通信用LSIが安価に提供されるようになったことがある．これらのLSIによるシリアル通信の信頼性が高まり，さらに基板への実装技術の進歩が加わり，シリアル通信が今後の並列計算機内における，計算機本体と入出力機器との通信形態のひとつとして有効であると考えた．

STAFF-Linkを用いたI/Oサブシステムにおける伝送速度と伝送距離の位置付けを示すために，現状の入出力インタフェース，LANおよびWANの伝送速度と伝送距離を**図2**に示す[14)]．STAFF-Linkによる通信路は，複数のリンクを束ねることにより，伝送速度においては数十Mbpsから，数Gbpsに及ぶ範囲をカバーし，通信距離においては数メートルから，数百メートルまたはそれ以上の距離に及ぶ入出力データの通信に利用できると考えられる．

2.2 STAFF-Linkの構成

従来のシリアル通信インタフェースにおいて，パラレルデータをシリアルに変換するために要する時間に

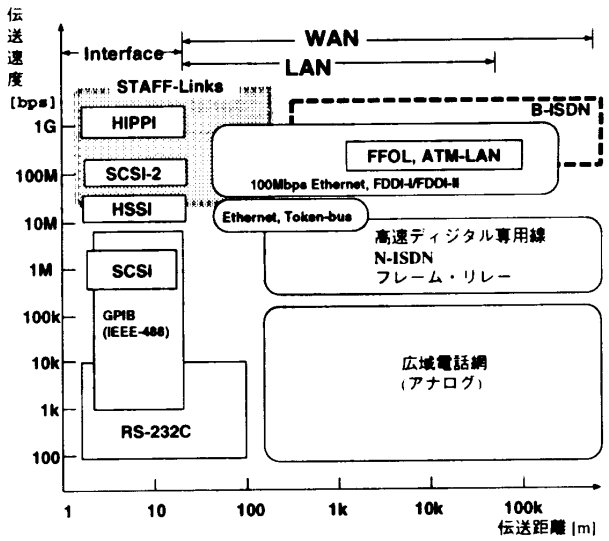

図2 入出力インタフェース/LAN/WANの伝送速度と伝送距離
Fig. 2 Transfer speed and distance in I/O interface/LAN/WAN.

よる通信遅延が生じ，通信性能が上がらなかった．シリアル通信は，

(1) データの書き込み
(2) パラレル-シリアル変換
(3) データ転送
(4) シリアル-パラレル変換
(5) データの読み出し

の5つのフェーズに分けることができる．STAFF-Linkにおいて，(2)~(4)のフェーズは，シリアル通信用LSIにより高速に処理し，送信側と受信側にバッファを設けて，5つのフェーズをオーバラップさせることにより，通信スループットを向上させることができる．

図3に，STAFF-Linkの構成を示す．通信ブロック (Communication Block) は送受信用高速シリアル通信LSI (TAXIチップ[15)]) と送信用/受信用の2つのFIFO，さらにFIFOが溢れないようにハンドシェイク (Xフロー制御) を行いながら非同期通信制御を行なう通信コントローラから構成される．2つの通信ブロック間を，カテゴリ5のツイストペアケーブルで接続することによって，その両端には仮想的に双方向のFIFOが形成され，双方のノードから見た場合にノード間の物理的な通信距離は隠蔽されることとなり，透過な通信路を構成することができる．現在の実装では最高140 Mbpsの転送レートで通信を行うことができる．

このSTAFF-Linkにより複数のリンクを構成することによって，広いレンジの入出力通信バンド幅を持ち，超並列計算機の要素クラスタと入出力機器との伝送距離の制限を緩和した入出力システムの実現が可能

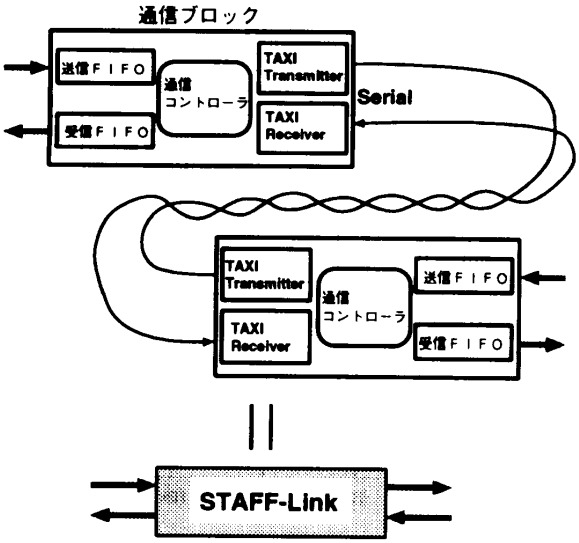

図3 STAFF-Link の構成
Fig. 3 Configuration of STAFF-Link.

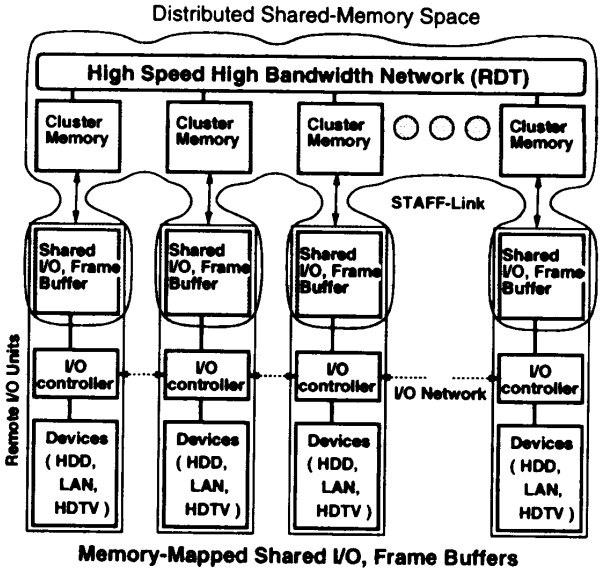

図4 JUMP-1 の共有入出力バッファ
Fig. 4 Shared I/O buffer of JUMP-1.

となる.

3. JUMP-1 の I/O アーキテクチャ

ここでは, JUMP-1 における I/O サブシステムの特徴について述べ, 現時点でのディスクおよび画像 I/O ユニットの構成について説明する.

3.1 拡張性 (Scalability) と柔軟性 (Flexibility)

I/O サブシステムの構成として, インテル社の Paragon[16] などの分散メモリ型の並列計算機においては, ディスク装置を要素クラスタや要素プロセッサに分散させた形態が実現されている. この場合, I/O システム全体で広い I/O バンド幅を得ることができるが, ディスク装置やそのインタフェースと要素プロセッサ間のケーブルを短くする, もしくは要素プロセッサボード上にディスクユニットを直接実装するなどの物理的な制限が大きい.

これに対して, ある特定のノードに専用の高速 I/O バスを設置し, そのバスに種々の I/O 機器を接続する形態が, CRAY などのスーパコンピュータでは一般的で, 高速 I/O バスの代表的なものに HiPPI[17] があげられる. しかしながら, 多数の要素プロセッサや要素クラスタから構成される超並列システムに対して専用バスを接続する場合, 接続されるノードやその近傍においてボトルネックが生じ, システム全体にわたる円滑なデータの入出力を行うことは困難となる.

そこで我々は, JUMP-1 本体から独立した複数のディスク I/O ユニットや画像 I/O ユニットを本体に密接した位置ではなく, 比較的離れた場所に設置する方針をとった. したがって, ディスク装置等の拡張については I/O ユニット内もしくはユニット単位で行うことが可能となり, 容易にディスク容量を増やしたり, 新たな I/O 装置を増設することができることになる. そして, ディスク I/O ユニットや画像 I/O ユニットと多数の要素クラスタとの接続には STAFF-Link を複数本設けて, 分散した共有入出力装置群を形成する (図 1). STAFF-Link で接続することにより, それぞれの I/O ユニットの設置場所や接続形態, さらには機器増設に対する柔軟性が確保できることになる.

3.2 I/O アクセスの容易さ (Accessibility)

JUMP-1 では分散共有型のコヒーレントメモリシステムは, 要素クラスタ間を接続する RDT ネットワーク上においてパケット転送を行うことによって実現される. そこで, I/O アクセスの容易さを実現するために, I/O ユニットと JUMP-1 本体とのデータ交換のためにバッファメモリを設け, このメモリや画像出力のためのフレームバッファ[5]を本体の共有メモリマップ上にマッピングする. すなわち, 図 4 に示すように入出力バッファを要素クラスタから見た拡張メモリと見なすことによって, 入出力機器を共有メモリとしてすべてのクラスタ間で共有することが可能となり, 種々の入出力機器の特性を吸収することができる. そのために, 画像データやディスクからのデータに対しても RDT パケットの形で転送することとなる.

ディスク I/O のための共有メモリを共有入出力バッファ (Shared I/O Buffer), 画像表示のためのメモリを共有フレームバッファ (Shared Frame Buffer) と

呼ぶ．共有入出力バッファを用いて，要素クラスタ上のデバイスドライバからのアクセスをメモリアクセスとして提供することが可能となる．また，共有フレームバッファに対するアクセスもすべての要素クラスタからは，メモリのリード/ライトアクセスにより実現される．以上の機能をもとに我々は共有メモリアーキテクチャに適した I/O サブシステムの形態を提案する．

3.3 保護機能 (Protection)

I/O アクセス時に RDT パケットが中継される場合，パケットは途中経路の RDT ルータ間で中継される．そして，I/O ユニットに接続される要素クラスタにおいても MBP のハードウェアで実装されたパケット中継回路を経由するのみである．したがって，パーティション間で実行性能の影響はほとんどなく，パーティションの独立性は保証される．また，ユーザ間のディスク領域の保護については，共有メモリアーキテクチャにおけるメモリ保護機能を活用し，MBP により保証される[11]．したがって，ファイルシステム側から見た場合，保護機能を提供するためのシステムの負荷は軽減される．

3.4 ディスク I/O ユニットの構成

拡張性，外部ネットワークとの接続の親和性および評価実験システムの早期構築を考慮して，ディスク I/O ユニットとしてワークステーションを利用した．ディスク I/O ユニットは**図 5** に示すように以下の要素から構成される．

- **ディスク I/O コントローラ (Disk I/O Controller)**
  要素クラスタ上のデバイスドライバから転送されるディスクのブロックまたはトラック単位のリード/ライト要求に応じて，SCSI などのディスク I/O インタフェースを通じてディスク装置に対するアクセス制御を行う．また，ディスクに対して要求を行った要素クラスタへの割り込みパケットの生成などの役割を果たすことによって I/O アクセスの負荷分散をはかる．
- **大容量ディスク装置 (Disk Devices)**
  JUMP-1 のメインメモリに対応した大容量の 2 次記憶空間を確保するために，1 つのディスク I/O ユニットあたり，2〜8 GB の容量のディスク装置を接続する．
- **共有入出力バッファ (Shared I/O Buffer)**
  トラックバッファとして働き，要素クラスタからの要求にしたがってディスク入出力コントローラによりアクセスされる．JUMP-1 の共有メモリ空間にマッピングされ，要素クラスタからはメモリアクセスとして読み書きされる．
- **クラスタ接続用 STAFF-Link**
  通信処理の多重化による高速性と，シリアルリンクによる柔軟性を合わせ持つ．本システムでは JUMP-1 の要素クラスタ内の MBP とディスクおよび画像 I/O システムを接続するために使用される．1 ユニットからは複数のリンクにより，いくつかの要素クラスタと接続する．これより，データ転送と I/O アクセスとの間のバランスをとる．
- **DMA コントローラ (DMA Controller)**
  要素クラスタからのリード要求に対して，共有入出力バッファに格納されたデータを STAFF-Link に連続的に書き込みを行ったり，STAFF-Link を通じて送られてくるライトデータパケットを共有入出力バッファに連続的に格納する．
- **STAFF-Link I/O ネットワーク**
  STAFF-Link により I/O ユニット間の接続を行い，I/O ネットワークを構成する．I/O アクセスに関して，RDT ネットワークは，クラスタ本体から I/O アクセスのためのパケット転送に利用される．それに対して，信頼性の向上やディスクアクセスの負荷分散を目的としたファイルシステムの再構成のためのユニット間通信に I/O ネットワークを用いる．この I/O ネットワークにより，I/O サブシステムとしての独立性を保持し，インテリジェントな I/O 装置群を形成することが可能となる．

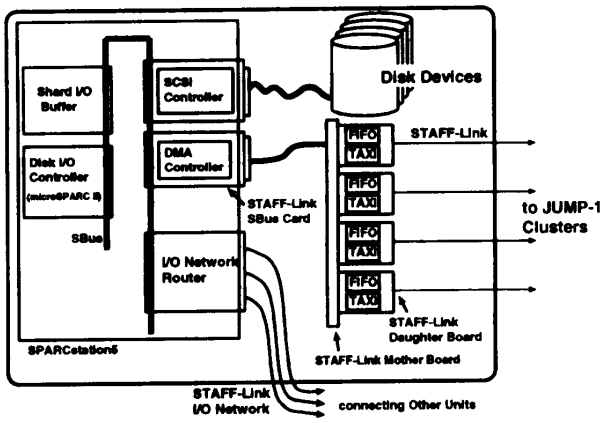

図 5 JUMP-1 におけるディスク I/O ユニット
Fig. 5 Disk I/O unit for JUMP-1.

3.5 画像 I/O サブシステム

ハイビジョン規格の画像データを処理する場合に，クラスタと画像 I/O インタフェース間とのリンクに要求される転送速度は，STAFF-Link の接続ポイントを持つクラスタの数を n とすると，

$$180/n \text{ MB/S} = 1440/n \text{ Mbps}$$

となる．したがって，90 Mbps 以上の転送速度を持つリンクを利用することにより，$n = 16$ として対応できるものと考えられる．

クラスタから出力される STAFF-Link の一端は，2 バンクの共有フレームバッファに接続される．STAFF-Link からの画像データが一方のフレームバッファ・バンクに転送されている間に，他方のバンクからハイビジョンモニタへ出力画像データを表示する．また，カメラなどの画像入力装置からのデータが共有フレームバッファへ入力画像を書き込んでいる間にもう一方のバンクから画像データを読み出し，STAFF-Link に流し込むことによって，クラスタに処理すべき画像データを転送する．このように画像データの入出力と STAFF-Link によるデータ転送をオーバラップさせることにより，リアルタイムの画像入出力を可能とする．

JUMP-1 のクラスタメモリの一部に共有フレームバッファのコピーを持たせることにし，以下のような方式でクラスタメモリの中に格納されている表示イメージと実際の表示が一致するようにする．

(1) STAFF-Link により共有フレームバッファに接続される各クラスタメモリは，分割された画面の部分イメージを持っている．クラスタ上の要素プロセッサは，共有空間にマッピングされた共有フレームバッファメモリ領域に対して直接ブロックデータ転送を行うことによって画像の入力・表示を行う．分割は静的で，ブロックやラインなどの単位で行う．

(2) フレームバッファは垂直同期信号に同期して，画像表示をになうクラスタに対して，次の表示を要求するために割り込みメッセージを転送する．

(3) リアルタイム表示のために，クラスタは 1/30 秒（ビデオ 1 フレーム）以内に自分の持っている分割部分のイメージデータをフレームバッファに送る．そして，この次の垂直同期信号のタイミングでバッファを切り替える．

以上のシーケンスを定期的に行うことにより一定時間内に JUMP-1 クラスタ上のメモリ内の画面イメージとフレームバッファのデータを一致させる．これにより，計算結果や途中過程がリアルタイムで表示できる．図 1 における画像入出力ユニット（Image I/O Unit）の実装については，VME ダブルハイトの基板を 16 枚用い，標準 VME ラックに収めるように設計・制作を行い，複数台のワークステーションに接続して動作を確認した．

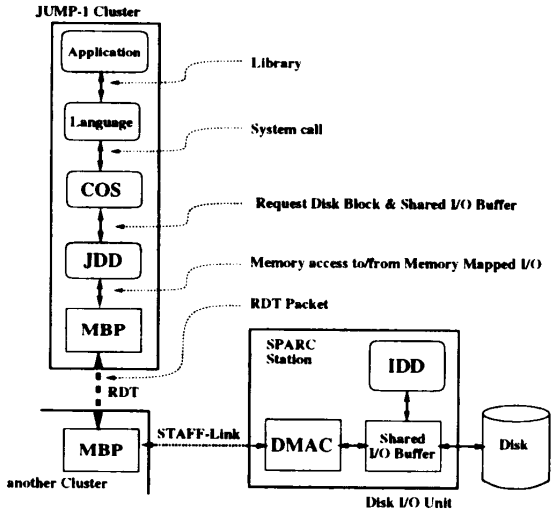

図 6 ユーザから見たディスクアクセスインタフェース
Fig. 6 Disk access interface view from users.

4. ディスク I/O サブシステムにおけるアクセス方式

4.1 共有入出力バッファを用いたディスク I/O アクセス

JUMP-1 において，そのハードウェア上における並列処理機能と性能の確保，アクセス保護およびユーザプログラムへのハードウェアの開放を共存させるための機構は COS (Collaborative Operating System)[18] により提供される．COS 上において，さまざまな並列処理言語により作成された種々の並列アプリケーションプログラムが実行される．ユーザプログラムにおいてファイルアクセスを行う場合における，ユーザレベルからハードウェアまでの処理の流れを図 6 に示す．

ディスク I/O システムにおいてはディスクブロック単位で番号付けを行い，その番号はシステム全体で一意とする．そして，JUMP-1 のクラスタ側のファイルシステムは，ディスクブロックや共有入出力バッファの使用状況に関する情報を，分散共有メモリ上においてすべてのクラスタ間で共有し管理する．ファイルシステムから I/O サブシステムへのアクセス要求は，JUMP-1 クラスタ上のデバイスドライバ（JDD）に渡される．ディスク入出力ユニット側において，ディスク装置へのアクセスなどの低レベルの処理はユニット側のデバイスドライバ（IDD）が行う．

以下では，ディスク入出力システムに対するリードアクセスとライトアクセスの手順を示す．

4.2 リードアクセス

(1) ファイル管理情報（UNIX における i-node 等）

をもとに，クラスタ上のファイルシステムからリード要求として，ブロック番号と共有入出力バッファのアドレスが JDD に渡される．

(2) そのブロックがすでに共有入出力バッファ上に存在していれば，以下の (a)～(e) の手順は省略される．

(a) JDD は，トラックリード要求としてブロック番号および格納すべき共有入出力バッファのアドレスを I/O コマンド用のアドレスに書き込み，応答待ち状態となる．このトラックリード要求は，MBP によりパケットとして転送される．

(b) 上記パケットはブロック番号および格納すべき共有入出力バッファのアドレスをトラックリード要求として RDT ネットワークを経由し，要求先のディスク I/O ユニットに転送され，ユニット上の IDD により解釈される．

(c) IDD は，要求されているブロック番号のあるトラックから 1 トラック分のディスク領域のデータを，指示された共有入出力バッファのアドレスが示す領域に転送を行う．

(d) 転送終了後，IDD は割り込みパケットを要求元クラスタの JDD に発行する．

(e) 割り込みパケットを受け取った JDD が応答待ち状態から復帰し，共有入出力バッファ上にデータが整ったことをファイルシステムに通知する．

(3) ファイルシステムは，要求されたブロックが格納されている共有入出力バッファのアドレスにリードアクセスを行う．このリード要求は，バイト，ワードまたはページ単位で行われ，実際には MBP によりパケットとして転送される．

(4) ディスク I/O ユニット上の IDD がリードパケットを解釈し，要求されたデータを共有入出力バッファから読み出し，RDT パケットの形式として STAFF-Link に転送する．

(5) パケットは RDT ネットワークを経由し，要求したクラスタのメモリに転送され，ファイルシステムが管理するメインメモリ上のバッファに格納される．

4.3 ライトアクセス

(1) ファイルシステムから，ディスクおよび共有入出力バッファの空き領域管理情報より，ブロック番号および，書き込むべきデータが JDD に渡される．

(2) JDD は，ライト要求としてブロック番号およびデータを，入出力コマンド用のアドレスおよび共有入出力バッファ領域に書き込み，応答待ち状態になる．このライト要求は，同様に MBP によりパケットとして転送される．

(3) ディスク I/O ユニットにパケットが到着すると，IDD がパケットを解釈し，共有入出力バッファ領域に格納を行った後，通知されたブロック番号を認識する．

(4) IDD は，共有入出力バッファ領域に格納したデータを，指定されたディスクブロック領域に書き込みを行う．

(5) ディスクへの書き込みが終了した後，IDD は書き込み終了の割り込みパケットを要求元のクラスタの JDD に発行する．

(6) 割り込みパケットを受け取った JDD が応答待ち状態から復帰して，ファイルシステムに書き込み終了を通知する．

5. ディスク I/O ユニットの基本性能評価

5.1 基本性能評価の目的

前章で示したディスク I/O ユニットの基本性能を評価するために，**図 7** に示す実験システムの構築を行った．この実験システムを用いて，STAFF-Link の現状での転送効率やユニット上で起動されるデバイスドライバ (IDD) のオーバヘッドを計測する．その結果から現状での到達点と問題点を示し今後の指針を提供することを目的とする．

実験システムは 12 台のワークステーション（サン・マイクロシステムズ：SPARCstation5: microSPARC-II (110 MHz), 32 MB）により構成される．そのうち

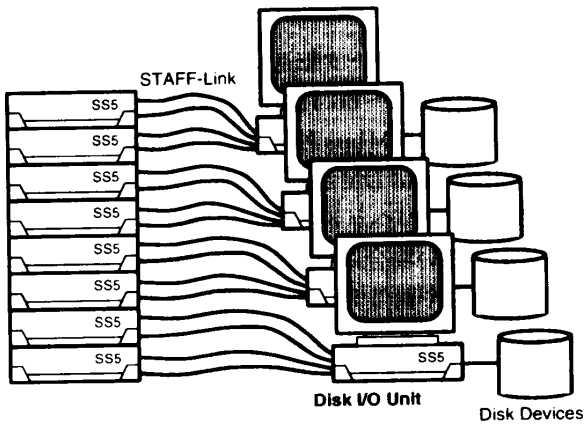

図 7 実験システムの構成
Fig. 7 Experiment system configuration.

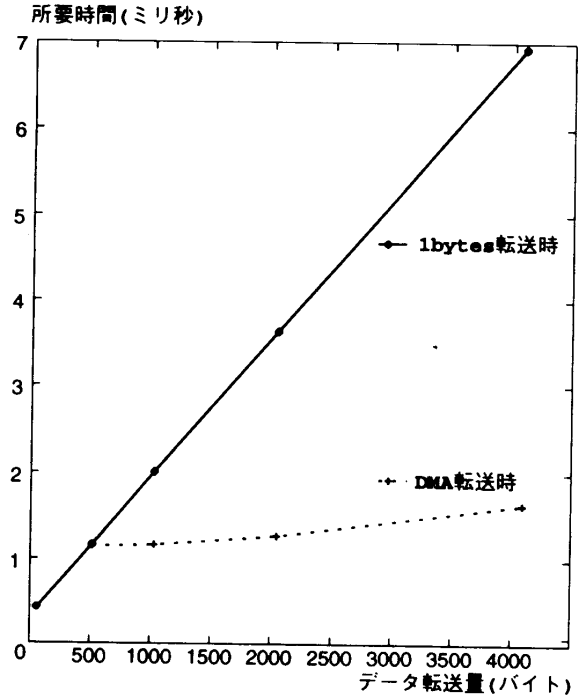

図8 データ転送時間
Fig. 8 Data transferring time.

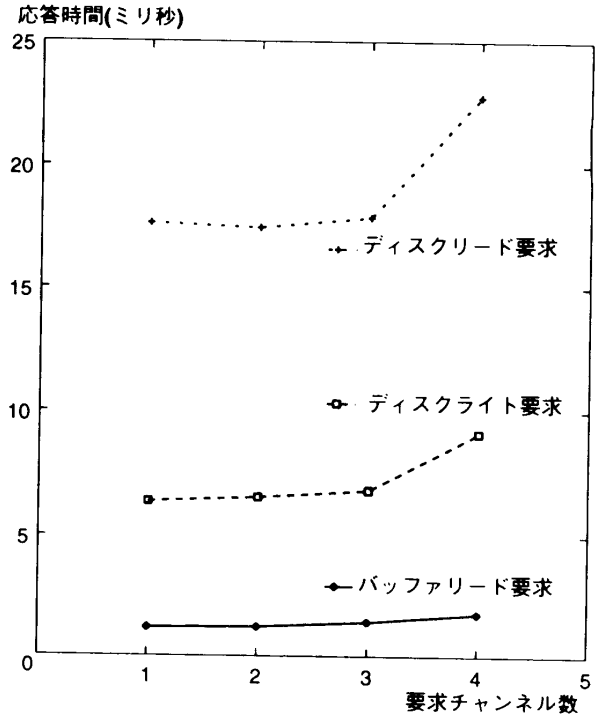

図9 応答時間
Fig. 9 Response time.

4台はディスクI/Oユニットとして，残りの8台をJUMP-1クラスタと見なすことができる．今回はディスクI/Oユニット単体でのアクセス性能の評価を行うため．図7において，1台のディスクI/Oユニットに関して性能評価を行った．

5.2 実験結果

5.2.1 転送速度の評価

図8 に，2台のワークステーション間において，STAFF-Linkを通じたデータ転送に要した時間を示す．ここでは，バイト単位で転送を行った場合と，DMAコントローラを用いて連続転送を行ったときに要した時間を計測した．この転送時間は，図6においてMBPとDMAC間のSTAFF-Linkにおける転送時間を表すことになる．

この結果より，データ転送レートは図8から，DMA転送の場合約44.1～66.4 Mbpsとなり，またバイト転送においては4.9 Mbps秒程度となっている．

現在の実装ではDMA転送を行うためのセットアップに要するオーバヘッド（約1ミリ秒）により，データサイズが小さい場合はバイト転送が有効であり，ディスクブロックや画像データなど粒度の大きいデータの転送にはDMA転送が有効である．

5.2.2 ディスクI/Oユニットの応答時間

次に，ディスクI/OユニットにおけるIDDの処理性能を評価するために，以下に示す要求パケットについての応答時間を計測した．

- ディスクリード要求：ディスクブロックからのリード
- ディスクライト要求：ディスクブロックへのライト
- バッファリード要求：共有入出力バッファからのリード

共有入出力バッファにのみ書き込みを行うというアクセスは本システムでは採用していないため，バッファライト要求の計測は行っていない．I/Oアクセスを行うワークステーションからディスクI/Oユニットに対して各要求パケットをランダムに送り，その応答時間を計測し，まとめた結果を**図9** に示す．この応答時間は，図6においてMBPとディスクI/Oユニット間における通信時間，IDDの処理時間およびディスク装置へのアクセス時間を含んだものとなる．

ここでは，データサイズは1024バイトとしている．図の横軸はディスクI/Oユニットに対してI/Oアクセスを行うために使用するSTAFF-Linkのチャンネル数を示す．図7において，2台のワークステーションから各2本のリンクで計4本のSTAFF-Linkがディスク I/Oユニットに接続されている．しかしながら，要求パケットサイズは小さいため，実質的には4台のワークステーションから同時にアクセスしたものと変わらない．

図9より，共有入出力バッファのリード要求には1.4～1.8ミリ秒前後，ディスクリードに17.5～22.7ミリ

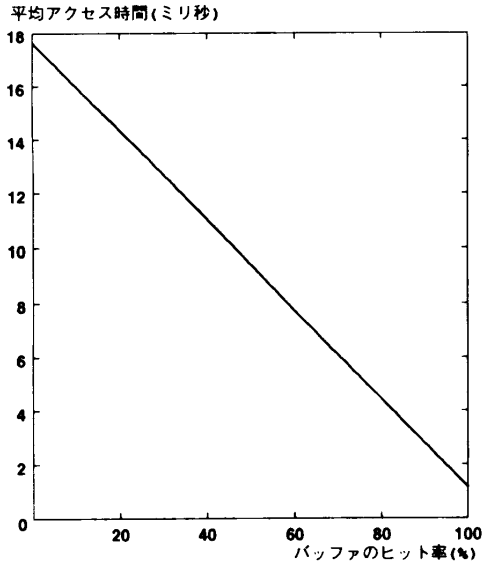

図10 平均リードアクセス時間
Fig. 10 Average read access time.

秒，ディスクライト要求には6.4～9.1ミリ秒程度の時間を要していることが分かる．共有入出力バッファからのリードはディスクへのアクセスをともなわないため高速なアクセスが可能となる．ディスクからのリードはトラック単位で行われるため，最初のアクセスには時間を要するものの，続くアクセスは共有入出力バッファに対するアクセスになるため，平均アクセス時間は向上できるものと考えられる．

次に，1つのI/Oユニットに対して，I/Oアクセスを行うクラスタ数の増加に対するユニット上のIDDの負荷状況について考察を行う．使用するリンク数が1～4に増えた場合，4リンク使用時には応答時間に若干の上昇傾向が見られる．しかしながら，増加時間にそれほど差がないことから，IDDの処理がボトルネックとはなってはいないと考えられる．これは，ディスクリード要求などの実際にディスク装置にアクセスする要求を処理している間に，バッファリードなどディスクにアクセスを行わない処理を並行して行え，安定したスループットが得られているからである．図9において，同時に使用するチャンネル数の増加に対する応答時間の上昇傾向から，ディスクI/Oユニットに接続するSTAFF-Linkのリンク数は4本が上限であると考えられる．

5.3 ディスクI/Oサブシステムの性能見積り

以上の基本性能評価の結果から，ディスクI/Oユニットに対するアクセス時間の見積り値を求める．要求ブロックが共有入出力バッファに存在する場合をヒットと呼び，そのヒット率をパラメータとして平均リードアクセス時間を示したものを**図10**に示す．

図11 STAFF-Linkドータボード
Fig. 11 STAFF-Link daughter board.

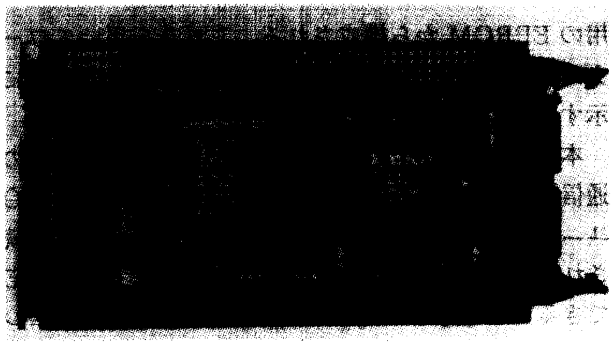

図12 STAFF-Link S-Busインタフェース
Fig. 12 STAFF-Link S-Bus interface.

これより，バッファへのヒット率が90%において平均アクセス時間は2.8ミリ秒となり，高速アクセスが可能となる．ヒット率を高めるためには，ファイルアクセスのプリフェッチやディスクアクセスのスケジューリングが重要な今後の課題となる．

6. まとめ

分散共有メモリ型超並列計算機JUMP-1のI/Oアーキテクチャについて述べ，その特長であるメモリアクセスによるI/Oサブシステムへのアクセス方式について述べた．また，ディスク装置を接続した数台のワークステーションを，STAFF-Linkにより相互に結合した実験システムを構築し，基本的な性能について評価を行った．

実装に関して，**図11**にSTAFF-Linkドータボードの PCB基板を示す．送受信シリアル通信LSIおよび送受信用FIFOと通信コントローラとなるFPGAから構成され，4層基板で実装した．**図12**にワークステーションに接続するためのインタフェース基板を示す．DMAコントローラLSIおよび周辺デコードのためのFPGA，さらにSBusコンフィギュレーション

図13 STAFF-Link インタフェースのワークステーションへの実装
Fig. 13 Connection of STAFF-Link interface to a workstation.

用のEPROMから構成される．これらを用いてワークステーションに接続した場合の実装状態を**図13**に示す．

本システムを用いて，リンク内では140 Mbpsでの通信性能を確認した．また，クラスタからディスクI/Oユニットへのアクセス時間を計測し，ユニットに接続されるSTAFF-Linkのリンク数は4本までは十分対応できることを示した．ディスクに対する応答時間についてはバランスのとれたアクセス時間を提供しており，共有入出力バッファを有効利用することにより，ある程度の平均アクセス時間を確保できることを確認した．今後はJUMP-1本体の実装と併行してDMAコントローラの転送性能を高めるとともに，STAFF-Linkの柔軟性を考慮して種々の形態の実験システムにより，超並列計算機システムのディスクおよび画像I/Oの性能を評価していく予定である．

謝辞 本研究を進めるうえで，プロジェクトを牽引くださった東京大学工学部電気工学科田中英彦教授に感謝いたします．また，有益なご助言をいただいた京都大学工学部情報工学科富田眞治教授に感謝いたします．STAFF-Linkインタフェースの実装に貢献いただいた古野電気株式会社岡田勉氏に感謝いたします．実験環境の整備についてご指導いただきました神戸大学工学部情報知能工学科瀧和男教授に感謝いたします．また，DMACボードに関しましてサポートいただきました中央システム技研株式会社製品企画部金井淳一氏に感謝いたします．なお，本研究の一部は文部省科学研究費（重点領域研究（1）課題番号04235130「超並列ハードウェア・アーキテクチャの研究」および，試験研究（A）（1）課題番号06508001「超並列計算機プロトタイプの開発と試作」）による．

参考文献

1) 平木 敬ほか：超並列プロトタイプ計算機JUMP-1の構想，情処処理学会研究会報告，94-ARC-102, pp.73–84 (1993).
2) 松本 尚：Memory-Based Processorを使用した汎用超並列計算機の基本アーキテクチャ，並列処理シンポジウムJSPP'94論文集，pp.409–418 (1994).
3) 中條拓伯，松本 尚，小畑正貴，松田秀雄，平木敬，金田悠紀夫：分散共有メモリ型超並列計算機JUMP-1の入出力サブシステム，情報処理学会研究会報告，94-ARC-104, pp.113–120 (1994).
4) 岡田 勉，中條拓伯ほか：超並列計算機JUMP-1における入出力サブシステムのアクセス方式，情報処理学会研究会報告，94-ARC-107, pp.177 184 (1994).
5) 小畑正貴，中條拓伯：超並列計算機JUMP-1におけるハイビジョン画像表示システム，情報処理学会研究会報告，94-ARC-108, pp.17–23 (1994).
6) 中條拓伯，小畑正貴，金田悠紀夫：高速シリアル・リンクを用いた分散画像生成実験システム，信学報，CPSY93-33, pp.39–46 (1993).
7) 中條拓伯，松田秀雄，金田悠紀夫：超並列計算機におけるワークステーションクラスタ・ファイルシステム，情報処理学会研究会報告，94-ARC107-24, pp.185–192 (1994).
8) 中條拓伯，岡田 勉，松本 尚，小畑正貴，松田秀雄，平木 敬，金田悠紀夫：分散共有メモリ型超並列計算機JUMP-1のディスク入出力サブシステム，並列処理シンポジウム，JSPP'95論文集，pp.67–74 (1995).
9) Nakajo, H., Matsumoto, T., Kohata, M., Matsuda, H., Hiraki, K., and Kaneda, Y.: High Performance I/O System of the Distributed Shared-Memory Massively Parallel Computer JUMP-1, *Proc. 7th IASTED-ISMM Int. Conf. on Parallel and Distributed Computing and Systems*, pp.470–473 (1995).
10) Yang, Y., Amano, H., Shibayama, H., and Sueyoshi, T.: Recursive Diagonal Torus: An Interconnection Network for Massively Parallel Computers, *Proc. 1993 IEEE Symposium on Parallel and Distributed Processing* (1993).
11) 松本 尚，平木 敬：Memory-Based Processorによる分散共有メモリ，並列処理シンポジウム，JSPP'93論文集，pp.245–252 (1993).
12) 次世代の周辺装置インタフェース，ポストSCSIに向けて動き出す，日経エレクトロニクス，94年5月9日号（No.607），pp.83–103 (1994).
13) シリアルSCSIがいよいよ市場へ，次世代周辺装置インタフェースの評価定まる，日経エレクトロニクス，95年7月3日号（No.639），pp.75–105 (1995).

14) 上原政二（編）：標準 LAN 教科書，アスキー出版局 (1993).
15) Advanced Micro Devices, Inc.: Am7968/Am7969-175 TAXI-175 Transmitter/Receiver Data Sheet and Technical Manual (1992).
16) Intel Corporation: Paragon XP/S, Product Overview (1991).
17) ANSI Document: High Performance Parallel Interface, Document #X3T9.3
18) 猪原茂和，松岡 聡，松本 尚：分散共有記憶型超並列オペレーティングシステム COS マイクロカーネルの保護機構，並列処理シンポジウム，JSPP'94 論文集，pp.349-356 (1994).

(平成 7 年 9 月 4 日受付)
(平成 8 年 4 月 12 日採録)

中條 拓伯（正会員）

昭和 36 年生．昭和 60 年神戸大学工学部電気工学科卒業．昭和 62 年同大学院工学研究科電子工学修士課程修了．平成元年神戸大学工学部情報知能工学科助手．計算機アーキテクチャ，並列処理，並列入出力システムの研究に従事．電子情報通信学会，人工知能学会，システム制御情報学会各会員．

中野 智行（学生会員）

昭和 47 年生．平成 7 年神戸大学工学部システム工学科卒業．現在同大学院自然科学研究科博士前期課程（情報知能工学専攻）に在学中．並列入出力の研究に従事．

松本 尚（正会員）

1962 年生．1985 年東京大学工学部計数工学科卒業．1987 年大阪市立大学大学院理学研究科物理学専攻修士課程修了．日本アイ・ビー・エム（株）東京基礎研究所研究員を経て，1991 年 11 月より東京大学大学院理学系研究科情報科学専攻助手．並列計算機アーキテクチャ，OS，最適化コンパイラに関する研究に従事．他に制約解消系，グラフィックス，ニューラルネットワーク等に興味を持つ．電子情報通信学会，日本ソフトウェア科学会，ACM 各会員．

小畑 正貴（正会員）

1957 年生．1980 年神戸大学工学部電子工学科卒業．1985 年同大学院自然科学研究科博士課程修了．学術博士．1984 年岡山理科大学理学部助手，1996 年同大学工学部教授，現在に至る．計算機アーキテクチャ，並列処理の研究に従事．電子情報通信学会，ACM 各会員．

松田 秀雄（正会員）

昭和 34 年生．昭和 57 年神戸大学理学部物理学科卒業．昭和 62 年同大学院自然科学研究科（博士課程）修了．同年同大学工学部助手となり，同大学講師，助教授を経て，平成 6 年 10 月より大阪大学基礎工学部情報工学科助教授，現在に至る．この間，平成 3 年 4 月より 10 カ月間米国アルゴンヌ国立研究所客員研究員．学術博士．論理型言語による並列処理，遺伝子情報処理の研究に従事．電子情報通信学会，IEEE CS，ACM 各会員．

平木 敬（正会員）

昭和 51 年東京大学理学部物理学科卒業．昭和 57 年同大学院理学研究科博士課程修了．同年電子技術総合研究所入所．理学博士．計算機アーキテクチャ全般，特にデータフローマシン，分散共有メモリマシンの研究に従事．元岡賞，市村賞各授賞．平成 3 年から東京大学理学部情報科学科助教授を経て，平成 7 年から同大学院理学系研究科教授，現在に至る．昭和 63 年から平成 2 年まで IBM ワトソン研究センター招聘研究員．

金田悠紀夫（正会員）

昭和 15 年生．昭和 39 年神戸大学工学部電気工学科卒業．昭和 41 年同大学院電気工学専攻修士課程修了．昭和 41 年電気試験所（現電総研）入所．電子計算機研究に従事．昭和 51 年神戸大学工学部システム工学科，現情報知能工学科教授．工学博士．コンピュータシステムのハードウェア，ソフトウェアの研究に従事．高級言語マシン，並列マシン，AI に興味を持っている．