計算機アーキテクチャ 113−18
（1995. 8. 24）

超並列計算機 JUMP-1 における ディスク入出力サブシステムの実装

中野 智行[*1], 中條 拓伯[*1], 岡田 勉[*2], 松本 尚[*3], 小畑 正貴[*4], 松田 秀雄[*5], 平木 敬[*3], 金田 悠紀夫[*1]

[*1]神戸大学 工学部 情報知能工学科, [*2]古野電気株式会社 技術研究所,
[*3] 東京大学 理学部情報科学科 [*4]岡山理科大学 工学部 情報工学科
[*5]大阪大学 基礎工学部 情報工学科

e-mail:nakano@jedi.seg.kobe-u.ac.jp

JUMP-1 は要素プロセッサの他に MBP と呼ばれる通信，同期処理に特化したプロセッサを持つクラスタを RDT と呼ばれる相互結合網で結合させた分散共有メモリマシンである．クラスタと入出力ユニットは STAFF-Link と呼ばれる高速なシリアルリンクで接続され，各入出力ユニットに用意された入出力用のバッファが JUMP-1 のグローバルメモリ空間にマッピングされる．そのため，入出力操作がクラスタ上のメモリと同様に，入出力バッファへのメモリアクセスとして扱うことができる．本稿では，JUMP-1 のディスク入出力ユニットの実装とその性能評価について述べる．

A Implementation of Disk I/O Subsystem for Massively Parallel Computer JUMP-1

Tomoyuki Nakano[*1], Hironori Nakajo[*1], Tsutomu Okada[*2], Takashi Matsumoto[*3], Masaki Kohata[*4], Hideo Matsuda[*5], Kei Hiraki[*3] and Yukio Kaneda[*1]

[*1]Department of Computer and Systems Engineering, Faculty of Engineering, Kobe University,
[*2]Elemental Technology Development and Researching Laboratory,
FURUNO ELECTRIC CO., LTD.
[*3]Department of Information Science, Faculty of Science, The University of Tokyo,
[*4]Department of Information and Computer Engineering, Faculty of Engineering,
Okayama University of Science
[*5] Department of Information and Computer Sciences, Faculty of Engineering Science,
Osaka University

A massively parallel computer JUMP-1 is a distributed shared-memory machine which consists of multiple clusters which include processors for inter-processor communication and synchronization called MBP, via a inter-connection network called RDT. I/O units are connected to clusters via fast serial links called Serial Transparent Asynchronous First-in First-out Link (STAFF-Link), and each I/O buffer is mapped to global memory space. Thus, I/O access can deal with memory read/write access. In this paper, we describe the implementation of a disk I/O unit and evaluation of its performance.

1 はじめに

近年，さまざまな研究機関，及び大学において並列計算機に関する研究が盛んに行なわれている．なかでも，相互結合網を介して多数のプロセッサを結合させた超並列計算機は次世代のスーパコンピュータの候補として有力視され，いくつかの商用マシンも登場している．

文部省科学研究費補助金・重点領域研究においても，超並列計算機プロトタイプ JUMP-1 の開発が進められている [1]．JUMP-1 は，クラスタ間を Recursive Diagonal Torus(RDT) と呼ばれる階層トーラス状の相互結合網で結合させた分散共有メモリアーキテクチャのマシンである．各クラスタは，要素プロセッサ (PE) のとは別に Memory Based Processor(MBP) と呼ばれる通信，同期といった非局所処理に特化したプロセッサがあり，効率の良い分散共有メモリの実現が可能となっている．

しかしながら，これまで超並列計算機に関しての興味の対象は，そのプロセッサ能力やメモリ周辺技術に集中していた．このことは多くの計算機システムの性能評価尺度が入出力処理の部分についてほとんど考慮がなされていないことからもうかがえる．にもかかわらず，入出力性能がシステム全体の性能に大きく影響することは周知の事実である．とりわけ，超並列計算機のような多数のプロセッサを有するシステムでは，各プロセッサに効率の良いデータ配分が行なわれなければ，システムのプロセッサ能力を有効に利用することはできない．このため，超並列計算機における入出力サブシステムの形態については慎重に議論する必要がある．

そこで，JUMP-1 における入出力サブシステムでは，多数の入出力ユニットを Serial Transparent Asynchronous First-in First-out Link (STAFF-Link) と呼ばれる高速なシリアルリンクによってクラスタ側の MBP と接続する形態をとる．個々の入出力ユニットは，入出力用のバッファ(共有入出力バッファ) を持ち，それらのバッファは JUMP-1 のグローバルアドレス空間にマッピングされる．共有入出力バッファを用いることにより，全クラスタで入出力機器の共有をはかることができ，入出力機器へのアクセスをメモリアクセスとして等価的に扱うことができる．このような形態をとることにより，MBP, 及び RDT の能力を活かし，分散共有メモリアーキテクチャに適した入出力サブシステムの構築が実現できる．

本稿では，JUMP-1 の入出力サブシステムの一構成要素であるディスク入出力ユニットの実装とディスク入出力サブシステムのアクセスを支援するソフトウエアについて述べ，それらの性能評価を行なう．

2 JUMP-1 のディスク入出力ユニットの構成

2.1 JUMP-1 の入出力サブシステム

JUMP-1 では多数の入出力機器を複数のクラスタへ分散させ接続する形態をとる (図 1).

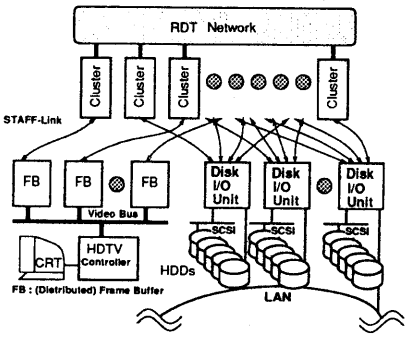

図 1: JUMP-1 の全体構成図

JUMP-1 における入出力サブシステムは以下に示す特徴を持つ．

STAFF-Link の利用

JUMP-1 では多数の入出力ユニットを接続させるため，入出力ユニットの設置場所に関する制限が少ない方が望ましい．また，多数のポートを実装する際，SCSI 等のパラレルリンクでは実装基板上におけるコネクタの占有面積が増大するという問題や，大きなコネクタや太いケーブルのために入出力機器の接続，変更，あるいはメンテナンスといった作業が容易ではないといった問題が生ずる．そのため，JUMP-1 での入出力サブシステムではクラスタと入出力ユニット間の接続には STAFF-Link と呼ばれる高速なシリアルリンクを用いる．シリアルリンクを用いることにより，ケーブル長による制限を少なくし，入出力ユニットを計

算機本体からある程度離れた位置に設置することができる．また，多数のポートを容易に確保することができ，一つの入出力機器に対して複数のリンクを (複数のクラスタに対して) 接続するといったことにも簡単に対応できる．

共有入出力バッファを用いたアクセス

各入出力機器には共有入出力バッファと呼ばれる入出力用のバッファメモリを持たせ，このメモリを JUMP-1 のグローバルメモリ空間にマッピングさせる．これにより，全てのクラスタから，全ての入出力機器へのアクセスをメモリアクセスとして透過的に行なうことができる．このとき，STAFF-Link は JUMP-1 クラスタの相互結合網である RDT の延長のように振舞う．入出力データがメモリアクセスとして扱われることから，入出力データのクラスタ間の移動においては，RDT 及び MBP の能力を有効に利用することができる．

2.2 STAFF-Link

STAFF-Link[2] とは，高速シリアル通信用 LSI(TAXI チップ [3]) と送受信用の FIFO，及びそれらの制御用の通信コントローラからなる通信ブロック間をカテゴリ 5 ツイストペアケーブルで接続させた最高 140Mbps の転送速度を持つリンクである (図 2)．

STAFF-Link は以下に示すような特徴を持つ．

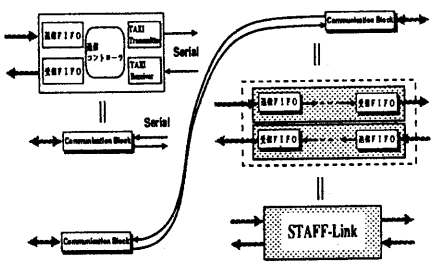

図 2: STAFF-Link の構成

簡便なインターフェース

細かな通信制御は通信ブロック上の通信コントローラがハードウェアで処理するため，通信路へのアクセスは FIFO メモリへのアクセスと全く同様に行なえる．

通信スループットの向上

シリアル通信は，データの書き込み，パラレル - シリアル変換，シリアル通信，シリアル - パラレル変換，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．このうち，パラレル- シリアル変換，シリアル通信，シリアル - パラレル変換のフェーズは，TAXI チップにより高速に処理し，送信側と受信側の FIFO メモリを利用したバッファを用いることで 5 つのフェーズをオーバラップさせ，通信スループットを向上させている．

現在，通信ブロック部，及び通信ブロックと SBus とのインターフェースボードの実装が終了し，STAFF-Link を通じたワークステーション間での通信が行なえることを確認している．

2.3 ディスク入出力ユニットの構成

JUMP-1 のディスク入出力ユニットは Sun のワークステーション SPARC Station 5(SS5) 上に構築される (図 12)．ディスク入出力ユニットにワークステーションを用いることには，次に示すような利点がある．

入出力操作の負荷分散

ディスクの入出力に関する低レベルの管理をディスク入出力ユニット側で処理することができ，入出力操作に関する負荷分散をはかることができる．

デバイス構成の変化への柔軟な対応

JUMP-1 クラスタ側ではデバイスの構成要素，形態について意識する必要がなく，単なるブロックデバイスとして統一的に扱える．必要であればディスク入出力ユニット側でさらに複数のディスクにデータを分散させるなどといったことも可能である．

ディスク入出力システムは以下の要素より構成される (図 3)．

- ディスク装置
  SCSI バスを介して SS5 に接続されているディスクである．
- 共有入出力バッファ
  JUMP-1 クラスタ側で管理されるディスク入出力用のバッファメモリである．JUMP-1 のグローバルアドレス空間にマッピングされ，全てのクラスタから直接アクセスすることが可能である．SS5 の内部メモリを用いて実装される．

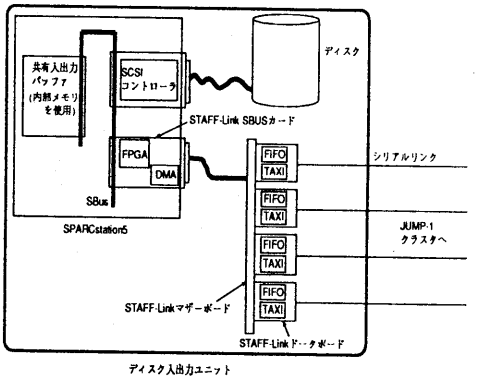

図 3: ディスク入出力ユニットの構成

- STAFF-Link インターフェース
  STAFF-Link の通信ブロックと SS5 の内部バスである SBus とのあいだのインターフェースボードである．DMA コントローラ (LSI Logic L64853A[4]) と制御用の LCA(Xilinx XC3120 [5]) が載った SBus カード (図 13) と STAFF-LINK の通信ブロックが載った STAFF-Link ドータボード，及びドータボードを 4 枚まで差すことのできる STAFF-Link マザーボード (図 14) からなる．

これらの構成要素間のデータのやりとりは全て SS5 の内部バスである SBus[6] を介して行なわれる．ディスク入出力ユニットの制御用 OS としては Solaris 2.3 を用いる．

3 ディスク入出力サブシステムのアクセス方式

JUMP-1 のディスク入出力サブシステムはクラスタ側の OS から論理的に単一のブロックデバイスとして扱われる．各ディスク入出力ユニットのディスクブロックにはシステムで一意のブロック番号が付けられ，空きブロック情報や共有入出力バッファの利用状況など，ディスク入出力に関する情報の管理は全て JUMP-1 クラスタ側の OS が行なう．

ディスク入出力サブシステムへのアクセスは，デバイスドライバのレベルにおいては全て共有入出力バッファへのメモリアクセスとして，JUMP-1 のクラスタメモリへのアクセスと全く等価に扱われる [7]．

ディスク入出力サブシステムに対するデバイスドライバは，JUMP-1 クラスタ側の OS 上のデバイスドライバ (JDD) とディスク入出力ユニット制御用 OS 上のデバイスドライバ (IDD) からなり，両者の協調動作によってディスクアクセスが実現される．IDD と JDD の役割を以下に示す．

JDD

JUMP-1 クラスタ上の OS のファイルシステムやメモリオブジェクトマネージャ等からのディスク入出力要求を RDT のメモリアクセスパケットに変換し，変換したパケットをディスク入出力ユニットに対して転送する．

IDD

JDD から送られてきたパケットを解釈し，実際に共有入出力バッファ，及びディスク装置へのアクセスを行ない，結果を要求元へ返す．

3.1 アクセスの手順

JUMP-1 のディスク入出力サブシステムでは，実際のディスク装置への読み書きは共有入出力バッファに対して行なわれる．例えば，あるディスクブロックを読み出したいとき，そのディスクブロックが共有入出力バッファにまだ存在していなければ，ディスクリード要求を発行して，読みたいディスクブロックを指定した共有入出力バッファのアドレスへ転送させ，転送が終了した後，改めて目的のデータが格納された共有入出力バッファへの読み出しを行なう．同様に，ディスクブロックの書き込みの場合も，書き込むべきデータと共に共有入出力バッファへのアドレスを指定し，データを共有入出力バッファへ転送してから指定されたディスクブロックへ書き込みを行なう．具体的なアクセス手順のフローを図 4，5 に示す．

また，このようなディスクリード，及びディスクライト要求は共有入出力バッファのディスク入出力要求用アドレスへのメモリアクセスとして，RDT メモリアクセスパケットと同様の形式で扱われる．

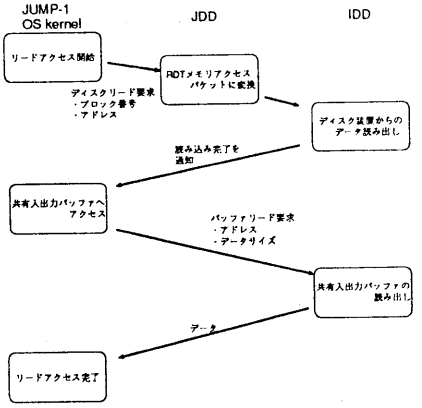

図 4: リードアクセス手順

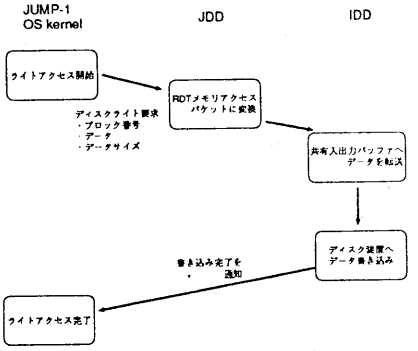

図 5: ライトアクセス手順

3.2 IDD の実装

IDD は JDD から送られてきた要求パケットにしたがって各種の処理を行なう．送られてくる要求パケットの種類は，

(1) 共有入出力バッファのリード要求
共有入出力バッファのアドレス，データサイズなどがパラメータとして渡される．基本的に RDT におけるメモリアクセスパケットと同じである．

(2) ディスクリード要求
ブロック番号，共有入出力バッファのアドレスなどがパラメータとして渡される．ディスクリードが完了したとき，ディスク入出力ユニットはクラスタに対して割り込みパケットを発行する．

(3) ディスクライト要求
ブロック番号，共有入出力バッファのアドレス，及びデータなどがパラメータとして渡される．

となる．これらの要求を円滑に処理するため，IDD は Solaris のマルチスレッド機能を用いて実装される．具体的には，4 つある STAFF-Link のチャンネルそれぞれにパケットの到着の監視とパケットの解釈を行なうパケットマネージャ・スレッド (PM スレッド) を生成し，さらに，ディスクの入出力の管理を行なうディスクマネージャ・スレッド (DM スレッド) を生成する (図 6)．

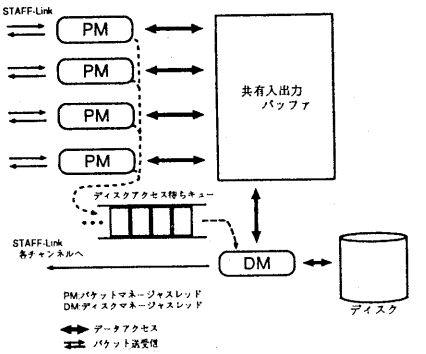

図 6: IDD の動作

PM スレッドは要求パケットが共有入出力バッファへのアクセスであった場合にはそのパケットで要求された処理を行なった後，結果を返して次のパケットの処理に移る．要求パケットがディスクへのアクセスであった場合には，その要求をディスクアクセス待ちキューに入れ，次のパケットの処理に移る．一方，DM スレッドはディスクアクセス待ちキューから要求を 1 つ取り出し，そこで要求されたディスクへのアクセスを行なった後に割り込みパケットを要求元へ送り，次の要求の処理へ移る．このようなスレッドを並行に動作させることによりディスク入出力ユニットのスループットの向上をはかる．

4 ディスク入出力ユニット実験システムの構成

ディスク入出力ユニットの性能を評価するために，実験システムの構築を行なう．図 7 に実験システムの構成を示す．

実験システムは 3 台のワークステーション (SS5) を用いて構成する．そのうち 1 台はディスク入出力ユニットとして，残りの 2 台を JUMP-1 クラスタ及び MBP と見立ててディスク入出力

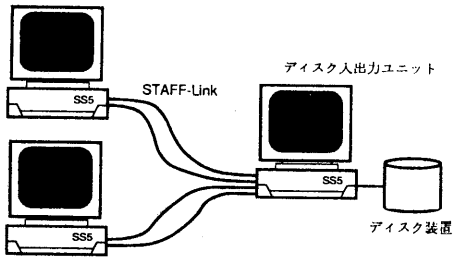

図 7: 実験システムの構成

ユニット単体でのアクセス性能の評価を行なう．今回の実験で用いたワークステーションの仕様を表 1 に，使用したディスク装置 (Quantum 社製 Empire シリーズ) の仕様を表 2 に示す．

表 1: 使用したワークステーション (SS5) の仕様

プロセッサ	microSPARC-II(70MHz)
メインメモリ	16Mbytes
SCSI インターフェース	SCSI-2 (同期 10Mbytes/秒)
OS	日本語 Solaris2.3

表 2: 使用したディスク装置の仕様

総容量	1.08 Gbytes
平均シーク時間	9.5 ミリ秒
SCSI タイプ	FAST SCSI-2

5 評価結果

5.1 転送速度の評価

図 8 に，2 台のワークステーション間での STAFF-Link を通じたデータ転送に要した時間を示す．ただし，現段階では，1 バイトずつのアクセスをデータ転送量分だけ繰り返すのに要した時間を計測した．

これによると，実効データ転送速度はせいぜい 500K バイト/秒程度であり，現状では STAFF-Link の性能が全く発揮されていないことがわかる．ただし，ここでの転送速度はディスク入出力ユニットの OS 上の余分なプロセス，あるいはドライバを取り除く (あるいはワークステーション自体の性能をあげる) ことで多少は改善できると思われるが，いずれにせよ STAFF-Link の性能を有効に利用するためには DMA 転送がぜひとも必要となるであろう．

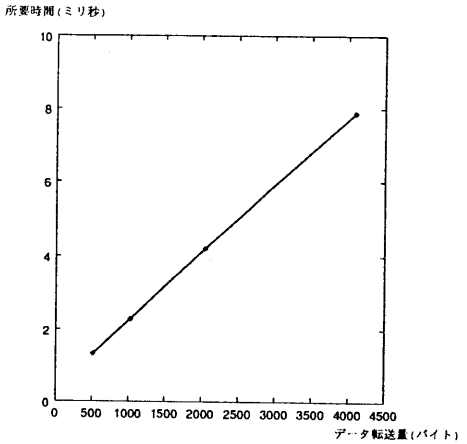

図 8: データ転送時間

5.2 応答時間の評価

ディスク入出力ユニット及び IDD の処理性能を評価するために，各要求パケットについての応答時間を評価する．JUMP-1 クラスタに見立てたワークステーションから STAFF-Link を通じてディスク入出力ユニットへランダムに要求パケットを送り，応答が帰ってくるまでの時間を各要求パケット毎に計測した結果が図 9, 10, 11 である．ただし，全てのパケットで要求するデータのサイズは 1024 バイトとしている．図の横軸はディスク入出力ユニットに対して同時に要求を送った STAFF-Link のチャンネル数である．

図より，共有入出力バッファのリード要求には 140 ミリ秒前後，ディスクリード / ライト要求には 350 ミリ秒前後かかっていることがわかる，同じ SS5 とディスクを用いたイーサネットを介したリモートディスクへのアクセス応答時間が約 120 ミリ秒となることから，現状では満足できる結果が得られたとは言い難く，今後，IDD のパフォーマンスチューニング，あるいは設計の変更等が必要となってくるだろう．

6 現状と今後について

現在，ディスク入出力ユニットのハードウェアの実装はほぼ終了しているが，細かいところ

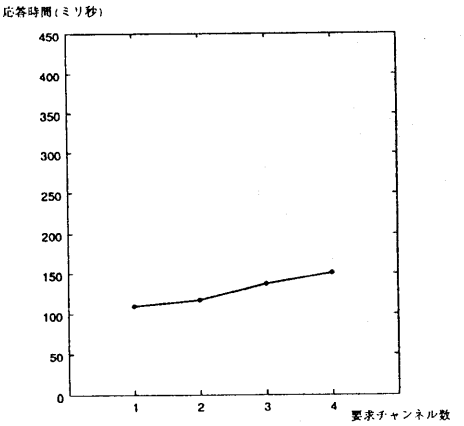

図 9: 応答時間 (バッファリード)

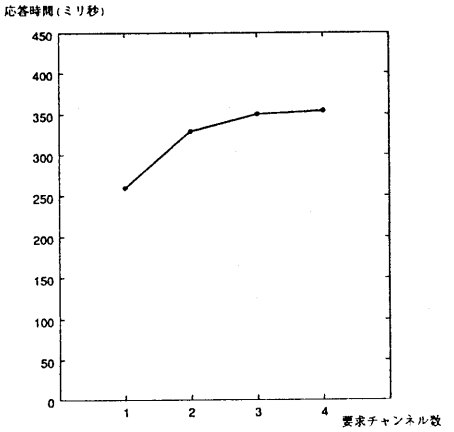

図 10: 応答時間 (ディスクリード)

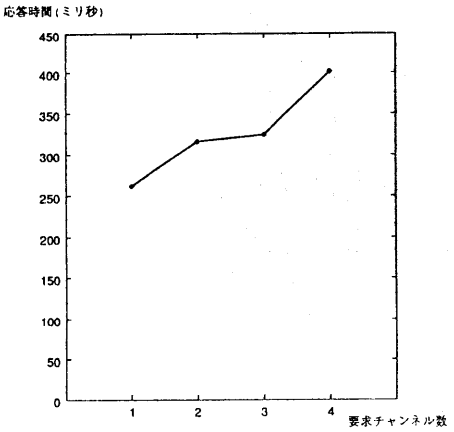

図 11: 応答時間 (ディスクライト)

でまだ動作が不安定な部分があり，今後，性能の改善と合わせてバグフィックスを行なっていく．特に 現状では DMA 転送が正常に動作していない．DMA 転送はディスク入出力ユニットとしての性能に大きく関わってくるところであるので，早急に対処しなければならない．IDD に関しても，完全な動作を行なうだけの完成度にはまだ達していない．今後，JUMP-1 クラスタの MBP と RDT パケットの必要なやりとりができるように実装をすすめていく．

謝辞

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」による．

参考文献

[1] 文部省重点領域研究「超並列原理に基づく情報処理基本体系」第 6 回シンポジウム予稿集, pp 4-42–4-49, Mar 1995.

[2] 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”, 情報処理学会計算機アーキテクチャ研究会報告 ARC107-24, Jul 1994.

[3] Advanced Micro Devices, Inc, “TAXIchip^{TM} integrated Circuits Technical Manual”, AM7968/AM7969 TAXIchip^{TM} Handbook, 1992.

[4] Xilinx, Inc. “The Programmable Logic Data Book”, 1994.

[5] LSI Logic Corp. “L64853A Enhanced SBus DMA Controller Technical Manual”, LSI Logic Corporation, Milpitas, CA., 1993.

[6] Susan A. Mason, “SBus Handbook”, Sun Microsystems, Inc., 1994.

[7] 岡田 勉, 中條 拓伯, 松本 尚, 小畑 正貴, 松田 秀雄, 平木敬, 金田 悠紀夫, “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 ARC-107-23, Jul 1994.

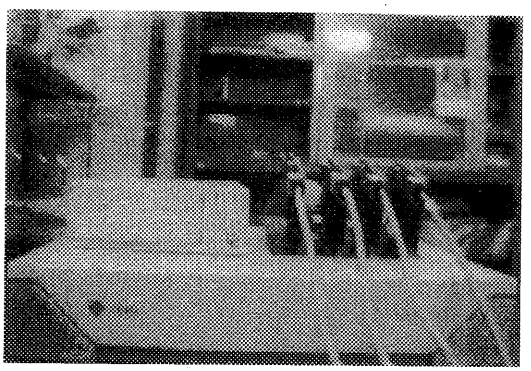

図 12: ディスク入出力ユニット

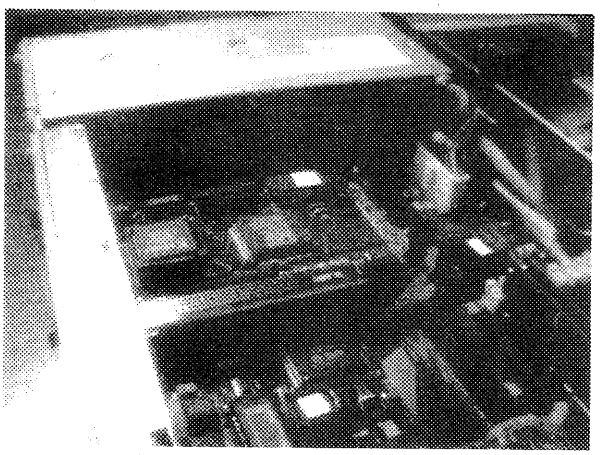

図 13: STAFF-Link SBus カード

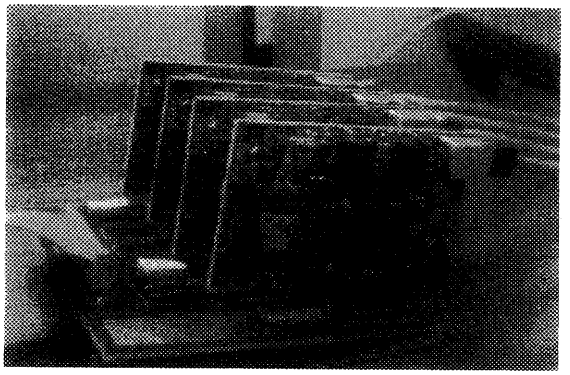

図 14: STAFF-Link マザーボードとドータボード