計算機アーキテクチャ 119－35
（1996．8．29）

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と評価

吉山 晃† 中野智行†
中條拓伯† 金田悠紀夫†

内容梗概

超並列計算機 JUMP-1 のディスク I/O サブシステムは多数の I/O ユニットから構成されており，各 I/O ユニットは STAFF-Link と呼ばれる高速シリアルリンクで相互に接続されている．この I/O サブシステム内の相互接続網（I/O ネットワーク）を用いて，ディスクアクセス要求の転送やディスクブロックの再配置，I/O サブシステム内の同期制御などに利用する．また，I/O サブシステムが JUMP-1 からのディスクアクセス要求を I/O ネットワークを使用して内部処理することで，JUMP-1 はディスク I/O サブシステム全体を 1 つのディスクシステムとしてアクセスすることができる．本稿では，JUMP-1 のディスク I/O サブシステムの I/O ネットワークの実装とその性能評価について述べる．

A Implimentation and Evaluation of Disk I/O Subsystem for Massively Parallel Computer JUMP-1

Akira Yoshiyama,† Tomoyuki Nakano,† Hironori Nakajo† and Yukio Kaneda†

Abstract

The disk I/O subsystem of a massively parallel computer JUMP-1 consists of many units for I/O. Each I/O unit is connected to others via fast serial links called Serial Transparent Asynchronous First-in First-out Link (STAFF-Link), and these inter-connection network called I/O Network. It makes the disk I/O subsystem able to forward disk-access requests from a JUMP-1 cluster, in order to replace disk blocks for disk access optimization, and to transport or broadcast synchronous-control signals. The inter-processing of I/O subsystem for disk-access request from a JUMP-1 cluster via I/O network enable JUMP-1 clusters to be recognized as the single disk system. In this paper, we describe the implementation of a I/O network for I/O subsystem and evaluation of its performance.

† 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

1. はじめに

計算機の利用形態がさらに高度化かつ大規模化した現在，計算機はただ演算が速いだけではなく，高いデータスループットも要求されている．

しかしながら，半導体技術が既に物理的な限界近くまで高度化しているように，数々の技術革新に支えられてきた従来のアーキテクチャの延長上にある計算機では，これまでのような計算機の発展が期待できないのが実情である．

そこで，従来の大型計算機に代わる次世代の計算機システムとして，超並列計算機が各研究機関や企業で開発され，商用機として実用化もされている．これらは主に演算処理を並列化する事で処理速度の向上を計った物であるが，その他の機能，特に入出力システムの処理速度の向上は，演算速度の向上に対して充分とは言えない．大規模データベースやリアルタイム動画処理など，取り扱うデータサイズが計算機システムに搭載されるメモリ容量よりはるかに大きい利用形態も増えている．その中で，ハードディスクなど補助記憶装置や，その他のI/O装置の速度がシステム全体に与える影響は，今後ますます大きくなると思われる．これを受け，入出力システムに関する研究が様々な研究機関で進められている．

文部省科学研究費補助金・重点領域研究において，これまで超並列計算機プロトタイプ JUMP-1 の開発が進められてきた．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マシンとして構成されている．各クラスタには，要素プロセッサ (PE) 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を可能にしている．

JUMP-1 では，I/Oサブシステムを構成する多数のI/Oユニットを，Serial Transparent Asynchronous First-in First-out Link(STAFF-Link) と呼ばれる高速シリアルリンクを用いてクラスタ側の MBP と接続している．各I/Oユニットには共有I/Oバッファがあり，各種のI/O装置は JUMP-1 のグローバルアドレス空間にマッピングされている．これにより JUMP-1 では各クラスタがこれらのI/O装置を共有し，I/O装置へのアクセスをメモリアクセスとして行なえるようになっている．このような形態をとることにより，MBP, RDT の能力を活かす，共有分散メモリアーキテクチャに適したI/Oサブシステムを構築する．

またI/Oサブシステムは，I/Oノード間をSTAFF-Linkで接続したI/Oネットワークを持つ．I/Oネットワークによって，各I/Oノードは JUMP-1 クラスタ側の相互結合網を介さずに高速に通信を行ない，I/Oサブシステムとしての協調動作を行なう事ができる．

本稿では，JUMP-1 のI/Oサブシステムにおける I/O ネットワークの果たす役割を述べ，I/O ネットワークを構成する通信ボードの実装とその性能評価を行なった結果について報告する．

2. JUMP-1 のI/Oサブシステム

JUMP-1 のI/Oサブシステムは以下の特徴を持つ．

- **STAFF-Link による接続**
  JUMP-1 のI/Oサブシステムは多数のI/Oユニットから構成されているため，I/Oユニットの設置スペースを考慮する必要がある．また，多数のI/OユニットをMBPと接続する際，SCSI等のパラレルリンクを使用した場合，双方の実装基板上に占めるコネクタ部の占有面積の増大，接続に使用するケーブル長や太さ，コスト等も問題になる．JUMP-1 では，I/Oサブシステムと MBP の間を STAFF-Link と呼ばれる高速シリアルリンクで接続する．これにより，ケーブル長による制限を緩和し，I/Oユニットの設置場所を JUMP-1 本体から分離する事ができるので，I/Oサブシステムの保守性が向上する．また，I/Oユニット，MBP 双方に複数のポートを確保できるので，多対多の接続を容易に行なう事ができる (図1)．

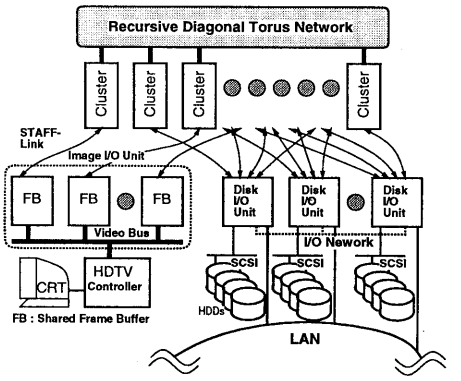

図1　JUMP-1 の全体構成

- **共有I/Oバッファを用いたアクセス**
  JUMP-1 のI/Oサブシステムは I/O 用の共有バッファを持っており，このバッファは JUMP-1 のグローバルアドレス空間にマッピングされている．JUMP-1 のクラスタは，I/O装置へのアクセスをこれらのマッピングされた領域へのメモリアクセスとして行なう事ができる．つまり，分散共有メモリアーキテクチャマシンである JUMP-1 に，I/O の為の特別なアクセス方式を用意する必要はない．

- I/O ネットワークによるディスク I/O サブシステム全体の仮想化
  各クラスタ間を接続する相互結合網とは独立して，JUMP-1 のディスク I/O サブシステムは入出力ノード間を接続する I/O ネットワークを持つ．JUMP-1 のディスクデバイスドライバは各ディスクブロックに一意な論理ブロック番号を付けて管理している．JUMP-1 クラスタからの論理ブロック番号によるディスクブロック要求は，ディスク I/O サブシステム側のデバイスドライバによって，実際のブロックが存在する I/O ノード番号とその物理ブロック番号に変換され，I/O ネットワークを通じてデータを持つ I/O ユニットに転送される．こうして，I/O ネットワークを利用して，JUMP-1 クラスタからのディスクブロック要求をサブシステム内で処理する事により，クラスタ側から見たディスク I/O サブシステムを，1 つのディスクシステムとして仮想化する事ができる (図 2)．

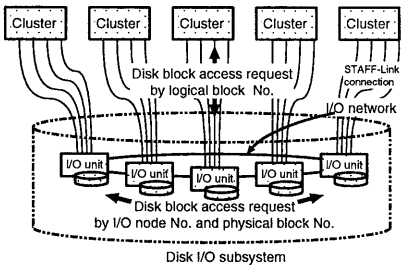

図 2　disk I/O subsystem

2.1 ディスク I/O サブシステム

本稿では，JUMP-1 クラスタ数を 256，I/O ノード数を 16 とし，1 つの I/O ノードは 4 つのクラスタに接続されているものとする．I/O ネットワークについては，各 I/O ノードは 1 つの I/O ネットワークルータボードを搭載し，2 つの STAFF-Link ポートでループを，残り 1 つの STAFF-Link ポートでループの最遠ノードと接続する (図 3)．

2.2 I/O ネットワーク

I/O ネットワークの主な機能を以下に示す．

- JUMP-1 クラスタからのディスクブロック要求の転送
  JUMP-1 クラスタからのディスクブロック要求を受け取った I/O ユニットに要求されたブロックがなかった場合，I/O ユニットは I/O ネットワークを通じて，要求されたディスクブロックを持つ I/O ノードに要求を転送する．これにより，

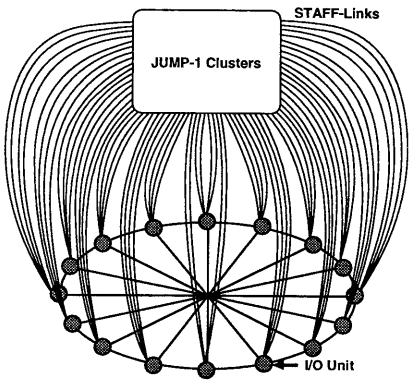

図 3　JUMP-1 Cluseters and Disk I/O Subsystem

JUMP-1 クラスタは任意の I/O ユニットにディスクブロックを要求する事ができる．すなわち，I/O サブシステムを論理的に 1 つのディスクとしてアクセスする事ができる．

- ディスクブロックの再配置
  JUMP-1 クラスタからのディスクアクセス履歴を参照し，ディスクブロックと各クラスタの依存度を調べ，特定クラスタの要求する頻度の高いブロックをそのクラスタに接続されている I/O ノードに転送する．これにより，JUMP-1 クラスタのディスクブロック要求に対する応答時間を短縮すると共に，ディスクブロックの転送に必要なトラフィックを最小限にする．
- ガベージコレクション
  各 I/O ノードのディスクブロックの利用状況を監視し，ブロックの再配置を行なってディスクブロックの最適化を行なうと共に，使用されていないブロックを集め，新たにディスクブロックを確保する際に効率の良い利用を行なえるようにする．

3. I/O ネットワークアクセス方式

JUMP-1 のディスク I/O サブシステムにおけるアクセス方式を以下に示す．

3.1 リードアクセス

JUMP-1 クラスタがディスクリード要求を I/O サブシステムに要求してから，I/O サブシステムが要求されたディスクブロックを JUMP-1 クラスタに転送するまでの過程を以下に示す (図 4)．リードアクセスでは，JUMP-1 クラスタ側の相互結合網 (RDT ネットワーク) の方が I/O ネットワークよりもバンド幅が広いので，I/O ネットワークはアクセス要求のみ転送し．要求されたディスクブロックを持つ I/O ノードが STAFF-Link で接続されている JUMP-1 クラスタにディスクブロックを転送する．

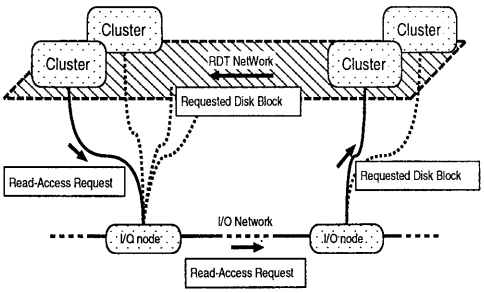

図4 Read Access

3.2 ライトアクセス

JUMP-1 クラスタがディスクライト要求を I/O サブシステムに要求してから，I/O サブシステムがディスクブロックを適切な I/O ノードに転送するまでの過程を以下に示す (図 5)．リードアクセスと異なり，ライトアクセスではディスクブロックデータが I/O ネットワークを介して目的ノードまで転送される．

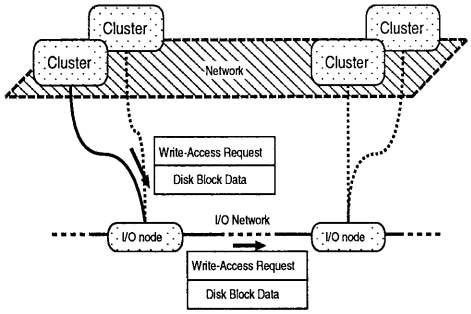

図5 Write Access

4. I/O ネットワークルータボード

I/O ネットワーク内の各 I/O ノードは STAFF-Link によって他のノードと接続される．STAFF-Link は point-to-point 接続の通信リンクであるので，1 つの I/O ノードが STAFF-Link によって直接通信できるノード数の上限は，1 ノードが持つ STAFF-Link のポート数に等しい．そのため，多数の I/O ノードを接続する為には，各ノードがそのノード宛でないパケットのルーティングを行なわなければならない．

そこで，I/O ネットワーク用にルーティング機能を持った STAFF-Link インターフェースを開発した．このボードは I/O ユニット (SUN: SPARCstation 5) の拡張バス (SBus) に搭載されるので，SBus ルータボードと呼ばれている (図 6)．

以下に I/O ユニットに搭載するルータボードについて述べる．

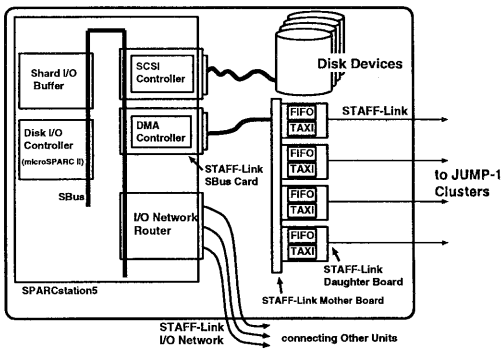

図6 I/O node host

4.1 ルータボードの仕様

ルータボードの仕様を以下に示す．

ボード	SBus ダブルハイト
ルーティング制御	DSP (TMS320C40)
インターフェース	STAFF-Link × 3, RDT 拡張ボード用インターフェース, SBus インターフェース
Local Program RAM (LPR)	512 kwords (2MB)
Local Data RAM (LDR)	512 kwords (2MB)
ホストコンピュータ	SPARCstation 5

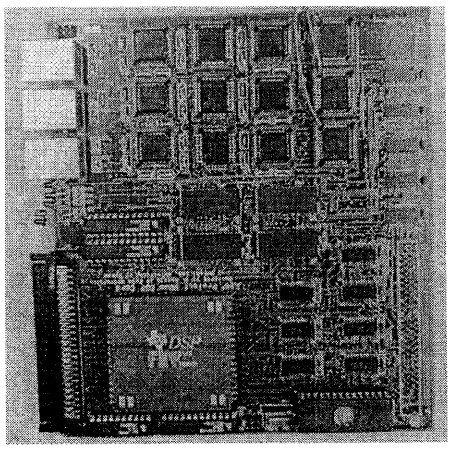

図7 SBus router board

4.2 ルータボードのアクセス方式

ルータボード上の各種インターフェースとステータスレジスタ，ホストマシンである SPARCstation 5 の仮想メモリ空間は DSP のアドレス空間にマッピングされている．DSP はこれらを実メモリと同様にアクセスすることでこれらの入出力を行なう．

4.2.1 SBus ルータボードのプログラム，データ転送

SBus ルータボードの DSP のルーティングプログラムのローディング方法と，SPARCstation 5 - DSP 間のデータ通信の方法について述べる．

Program Loading
SS5 から SBus 経由でロードされる

SPARCstation → DSP
ルータボード上の DSP の LDR は SBus のアドレス空間にマッピングされており，mmap システムコールによってアクセス可能になる．

DSP → SPARCstation
SS5 の仮想メモリは DSP のアドレス空間にマッピングされており，DSP は SBus マスタとして SS5 内の仮想メモリ空間にアクセス可能である．

5. ディスク I/O サブシステム性能評価

I/O ノードホストが，SBus ルータボードに出力するパケットの仕様を以下に示す．

- Read Access Packet
  制御情報のみ (10 bytes)
- Write Access Packet
  制御情報 (10 bytes) + ディスクブロックデータ (1024 bytes)

クラスタから要求を受信した I/O ノードによって，最初にパケットがルータボードに出力されてから，目的 I/O ノードに届くまでの所要時間を計算した．

Hop Counts	Read Access	Write Access
(Packet size)	10 bytes	1034 bytes
4(Maximum)	$6.2\mu s$	$88\mu s$
2.44(Average)	$4.1\mu s$	$86\mu s$

5.1 応答時間の評価

ここでは，I/O ネットワークの転送能力がディスク I/O サブシステムの処理能力に対して充分であるかを考察する．

ディスク I/O ユニットの応答時間は以下の通りである[2]．

mode	access time
read access from shared I/O buffer	$1.4\sim1.8ms$
disk read access	$17.5\sim22.7ms$
disk write access	$6.4\sim9.1ms$

ここで示した応答時間とは，JUMP-1 クラスタ側のディスクデバイスドライバ (JDD) がディスクアクセス要求を I/O ユニットに出力してから，I/O ユニット側のディスクデバイスドライバ (IDD) がディスクアクセスまたは共有入出力バッファへのアクセスを行ない，処理が終ってから JDD に結果を返すまでの時間である．したがって，JUMP-1 クラスタの要求したディスクブロックが要求を受信した I/O ユニットに存在する場合を前提としている．

ここで，JUMP-1 クラスタが要求したディスクブロックが，要求を受信した I/O ユニットに存在せず，要求を他の I/O ノードに転送する場合を考察する．

各アクセス方式の応答時間に対して，I/O ネットワークによるアクセス要求の転送時間が占める割合は，共有入出力バッファからのリード要求では平均 0.26%，ディスクリード要求では平均 0.02%，ディスクライト要求では平均 1.1%となり，いずれも無視できる範囲である．

6. 現状と今後について

本稿では，超並列計算機 JUMP-1 の I/O サブシステムにおける I/O ネットワークの概要について述べ，I/O ネットワークを構成するハードウェアである SBus ルータボードについての詳細を述べた．

考察の結果，DSP のルーティングプログラムは現在の処理速度で STAFF-Link の能力に見合った速度で転送を行なえると予想されるが，あくまで途中の転送に対して障害のない状態での予想であり，現実のシステムで評価を行なう必要がある．

今後の課題を以下に示す．

- ルータボードの実装と性能評価
  これまで進めてきた作業により，ルータボードは実装と基本機能のデバッグをほぼ終了した．今後台数が揃い次第，実験システムを構築してルータボードの転送能力が当初の予想通りの性能を発揮できるかを検証する．
- ルータボードとホスト間の通信方式の検討
  現段階では SPARCstation 5 とルータボードの通信にルータボード上の LDR のみを使用しているため，SS5 による LDR の write に $365 + 0.276 \times words(\mu sec.)$，read に $5115 + 0.524 \times words(\mu sec)$．かかる．今後はルータボードが SBus マスタとして，SS5 の内部メモリに直接アクセスする事によって，SS5・ルータボード間の

アクセス速度を削減しなければならない.

- I/O サブシステムの実装
  実際に SS5 16 台で I/O ネットワークを構成し, テストプログラムを I/O ユニットホスト上で実行して, I/O サブシステムとして稼働した状態と同様の通信要求を行ない, I/O ネットワークルータとしてのルータボードの評価を行なう (図 8).
- JUMP-1 への実装
  I/O サブシステムとしての検証が終れば, 本研究の最終目標である JUMP-1 MBP との接続を行ない, 全体として超並列計算機としての性能評価を行なう中で, I/O サブシステムの評価を行なう.

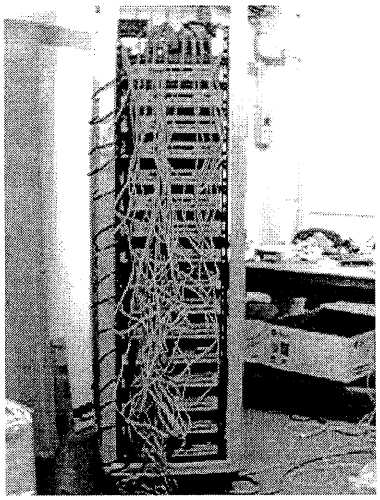

図 8 Disk I/O subsystem

参 考 文 献

1) 中條拓伯, 中野智行, 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 におけるスケーラブル I/O サブシステムの構成" 情報処理学会論文誌 vol.37, pp.1429-1439, 1996.

2) 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, pp.67-74

3) 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.4-11–4-16, 1995.

4) 中野智行, 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, " 超並列計算機 JUMP-1 における入出力サブシステムの実装" 情報処理学会研究会報告 ARC113-18, pp.137–144, 1995.

5) 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 ARC104-15, pp.113–120, 1994.

6) 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.