分散共有メモリ型超並列計算機の入出力システムとその評価

4F－5

中野 智行　吉山 晃　中條 拓伯　金田 悠紀夫
神戸大学 工学部 情報知能工学科

1. はじめに

多数のプロセッサを相互結合網で結合させた超並列計算機による並列処理は次世代のハイパフォーマンスコンピューティングの最有力候補として注目されている．しかしながら，近年のプロセッサ技術の飛躍的な向上に対して，ディスク装置などの入出力装置の性能はあまり向上していない．このため，大量のデータを扱う超並列計算機の入出力システムは，計算処理性能と入出力性能とのバランスを十分考慮して設計されなければならない．

我々は分散共有メモリ型超並列計算機 JUMP-1 のディスク入出力システムの実装を行っている．JUMP-1 のディスク入出力システムは，入出力ネットワークで結合された複数の入出力ノードからなり，クラスタと入出力システムは高速シリアルリンクで結合される．また，入出力ノードのバッファメモリを JUMP-1 の共有メモリ空間にマッピングすることにより，メモリアクセスを基本とした入出力アクセスが，すべてのクラスタから透過的に実現できる．

本稿では，ディスク入出力システムのプロトタイプの実装の現状について報告する．

2. JUMP-1 の概要

JUMP-1 は文部省科学技術研究補助金・重点領域研究で実装が進められた分散共有メモリ型の超並列計算機である．JUMP-1 [1] は，要素プロセッサ (PE) と MBP と呼ばれる通信，同期といった処理に特化したプロセッサを含むクラスタを，RDT と呼ばれる相互結合網で結合させることにより構成される．各種の入出力機器は複数のクラスタへ分散させて接続させ，アクセスの並列化をはかる (図 1).

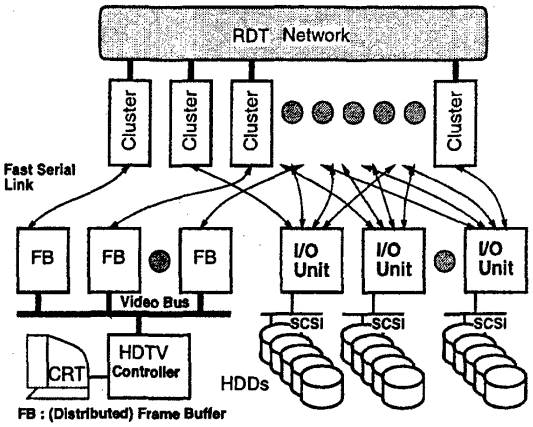

図 1: JUMP-1 全体構成図

クラスタと入出力機器との接続には，我々が独自に開発した STAFF-Link(Serial Transparent Asynchronous First-in First-out Link)[2] と呼ばれる高速なシリアルリンクが用いられる．ケーブルの取り回しが容易なシリアルリンクを用いることにより，多数の入出力機器の設置場所や接続機器の変更，メンテナンスといった作業に柔軟に対応できる．

3. ディスク入出力システムの構成

ディスク入出力システムは多数のディスク入出力ユニットから構成される [3]．各ユニットは STAFF-Link で構成された入出力ネットワークを介して結合され，また，クラスタとは 4 本の STAFF-Link で直接接続される (図 2).

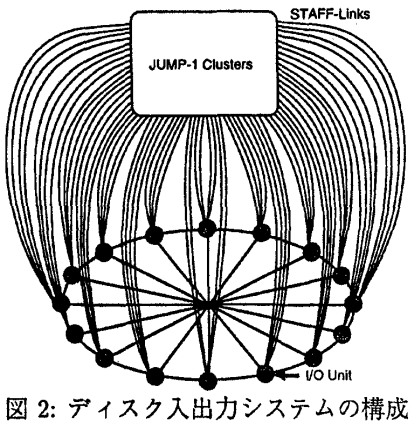

図 2: ディスク入出力システムの構成

I/O System for Distributed Shared Memory Massively Parallel Computer, and its Evaluation

Tomoyuki Nakano, Akira Yoshiyama, Hiroyuki Nakajo and Yukio Kaneda
Department of Computer and Systems Engineering, Faculty of Engineering, Kobe University
1-1 Rokkoudai,Nada,Kobe 657,Japan

また，ディスク入出力ユニットには共有入出力バッファ(Shared I/O Buffer) と呼ばれる入出力用専用のバッファメモリを設け，JUMP-1 のグローバルアドレス空間に直接マッピングさせる．これにより，各ディスク入出力ユニットに対してのアクセスが，すべてのクラスタから，メモリアクセスとして透過的に行うことができる．

ディスク入出力ユニットは SUN の SPARC Station 5 (SS5) の内部バスである SBus 上に STAFF-Link のインタフェースボードと入出力ネットワークのインターフェースボードを実装することで構築される．STAFF-Link のインタフェースボードには DMA コントローラが搭載され，SS5 のメモリと STAFF-Link 間のデータのやり取りを行う．入出力ネットワークインタフェースボードには DSP が搭載され，パケットのルーティングを行う．現状でのディスク入出力システムの実装の様子を図 3 に示す．

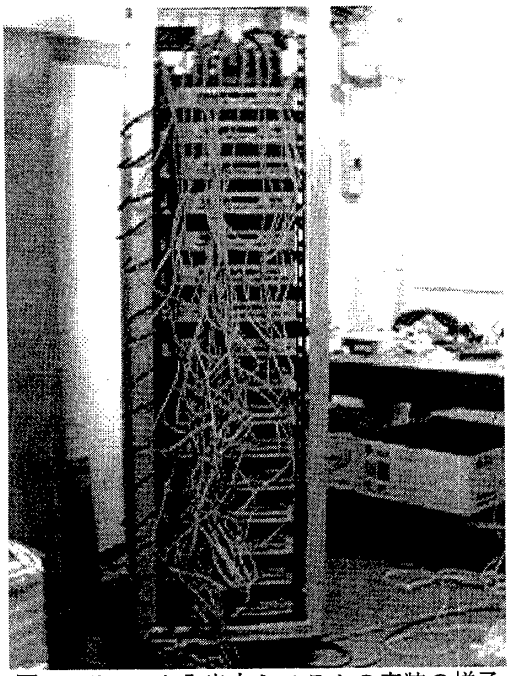

図 3: ディスク入出力システムの実装の様子

4. JUMP-1 クラスタ上の OS との関係

JUMP-1 クラスタとディスク入出力サブシステム間では，ディスクブロックを基本単位としたデータ転送が行われる．任意のディスク装置の任意のディスクブロックはブロックは，ディスク入出力システムで一意に定まるブロック番号で指定できるようにする．これにより，JUMP-1 クラスタ上の OS(以下 JUMP-1 OS) はディスク入出力システム全体を 1 つの超並列ディスクアレイとして仮想化することができ，メモリ資源と同様に，すべてのクラスタから大域的な 2 次記憶資源を利用することができる (図 4)．

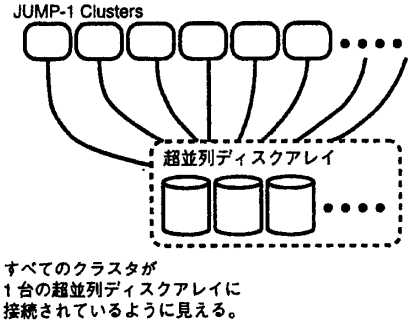

図 4: 超並列ディスクアレイ

すべての入出力装置へのアクセスは，共有メモリへのアクセスとして扱われるため，JUMP-1 OS はユーザプロセスからのディスク入出力要求をメモリアクセスの形へ変換し，ディスク入出力ノードへ転送する．実際にはすべてのクラスタがディスク入出力ノードへの直接のリンクを持ってはいないので，ディスク入出力要求を適切なクラスタへ転送しなければならないが，この処理は MBP により高速に処理されるのでクラスタ側の PE に負荷をかけることはない．

5. おわりに

現在，SS5 から STAFF-Link へのインターフェース部は実装が完了し，入出力ネットワークのルータボードも実装が進められている．JUMP-1 は現時点で開発中であるので，JUMP-1 の代わりに 4 CPU の SPARC Station 20 を用い，16 台のディスク入出力ユニットを用いた実験システムを構築し，評価を進めている．今後は，評価結果をもとに JUMP-1 OS のファイルシステムについて検討していきたい．

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 7 回シンポジウム予稿集 (CD-ROM), Mar 1996.

[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] 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, 37 巻 7 号, 1996.