高速シリアルリンクを用いた並列分散入出力システム

1B－7

高橋 淳　佐伯 靖　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

マイクロプロセッサの高性能化・低価格化とネットワーク技術の発達に伴って，ネットワークに接続されたワークステーション群 (ワークステーションクラスタ) を利用する分散処理環境が普及してきた．ワークステーションクラスタを用いた並列処理では，ファイル入出力が頻繁に起こる場合，一つのディスクへのアクセスが集中すると実行効率が低下するため，アクセスの負荷を分散させることが重要である．また，イーサネット (10Mbps) 等のバス型ネットワークを前提とした従来のワークステーションクラスタでは，音声や動画像・静止画像などマルチメディアデータを処理する場合，通信するデータがが大きくなるにつれてネットワークがボトルネックとなり，競合が起こるので接続するノード数がある程度制限される．

我々はそれらの問題を解決するために，高速シリアルリンクを用いたネットワークでの並列分散入出力システムの実装を進めている．本稿では，そのシステムの構成と，プロトタイプの実装について述べる．

2　システムの概要

高速シリアルリンクを用いて隣接ノード間通信を行なう形態の並列分散入出力システムを構築する．シリアルリンクを用いることにより，一つのノードに物理的に多数のポートを実装することが可能となり，高次元のネットワークを構成することができる．ハイパーキューブ等の高次元ネットワークはリング等の低次元ネットワークに比べて転送時のホップ数が少なく，ネットワーク全体のバンド幅が広い．このようなネットワークを用いてデータ転送を行なうことにより，効率のよい入出力をサポートする．

2.1　システムの構成

システムの構成を図1に示す．I/O ノードはそれぞれローカルなディスクを持ち，分割されたファイルを格納している．各I/O ノードには分割されたファイルを管理するデータサーバが存在し，計算ノードからの入出力要求に対して互いに通信・協調して一つのファイルイメージを提供する．

ノード間は高速シリアルリンクで接続されており，本システムでは STAFF-Link[1][2] という高速シリアルリンクを用いる．ノード間の接続トポロジーは柔軟に変更することができ，対象とする問題に適したネットワークを構成できる．

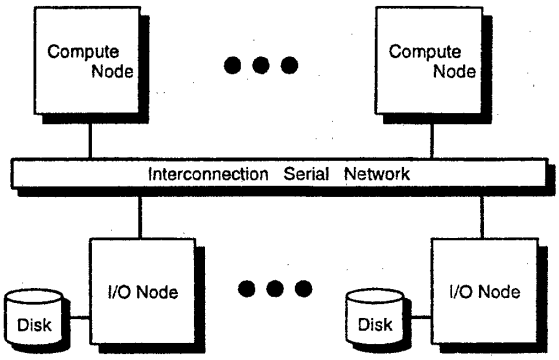

図 1: システムの構成

2.2　高速シリアルリンク STAFF-Link

STAFF-Link は自身が持つ FIFO メモリへ読み書きを行なうことによって通信することができるようになっており，最高 140Mbps でデータ転送を行なうことができる．また，コントロールプロセッサを搭載することによって，種々のルーティングアルゴリズムに対応することが可能となる．

Parallel and Distributed I/O System using High-speed Serial Links

Atsushi Takahashi, Yasushi Saeki, Hironori Nakajo and Yukio Kaneda

Department of Computer and Systems Engineering, Faculty of Engineering, Kobe University

1-1 Rokkoudai,Nada,Kobe 657,Japan

3 プロトタイプの実装

プロトタイプはSunのSPARCstation 5 (SS5)上に実装される．STAFF-LinkはSS5の拡張I/OバスであるSBusに接続され，4系統のリンクを持つ(図2)．SBusとのインタフェースカードには主記憶とSTAFF-LinkのFIFOメモリの間でDMA(Direct Memory Access)転送を行なうためのDMAコントローラが搭載されている．

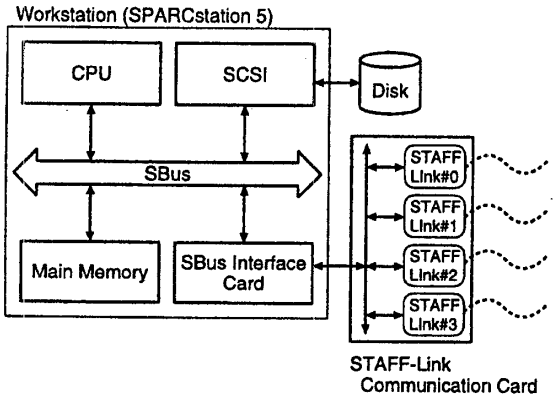

図 2: I/O ノードの構成

STAFF-Linkを用いたメッセージパッシング機能を提供するソフトウェアインタフェースとしてPVM[3]を実装し，その上にPVMをベースにした並列分散ファイルシステムであるPIOUS(Parallel Input/OUtput System)[4]を構築する．

PVMはワークステーションクラスタ上で並列処理を行なうためのシステムであり，通信の管理・実行を司るPVMデーモンと並列処理ライブラリ群とで構成されている．I/Oノード上のデータサーバや計算ノード上のタスクはPVMのライブラリを使ってデータ転送行なう．ライブラリが呼び出されるとPVMデーモンはそのデータを，STAFF-Linkを通じて適切なPVMデーモンへ転送し，そのデーモンが目的のタスクにデータを渡すことによって通信が完了する．

PVMデーモンがSTAFF-Linkを使ってデータを転送するときに2種類の転送モードをサポートする．1つ目は1バイトずつのデータの読み書きを行なうバイト転送である．これは転送時のオーバヘッドが少ないため，確認メッセージなどの少量のデータ転送に向いている．2つ目はDMAコントローラによるDMA転送である．これはある程度大きいサイズのデータ転送に向いている．現時点でのSTAFF-Linkを使ったPVMのデータ通信のグラフを図3に示す．

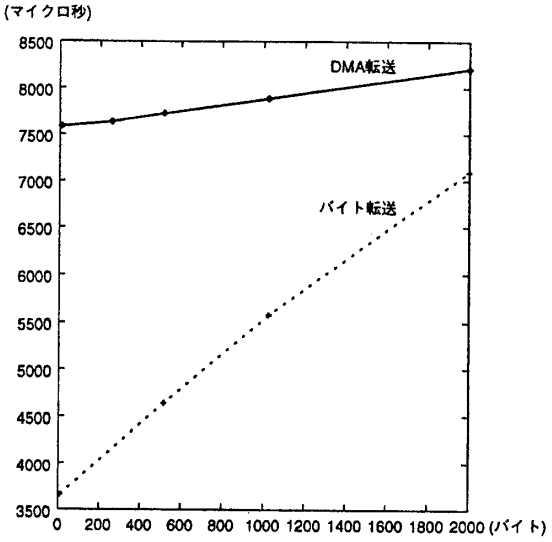

図 3: STAFF-Linkを使ったPVMの通信

4 おわりに

現在，STAFF-Linkを用いたネットワークでルーティングを行なうための，コントロールプロセッサを搭載したボードを実装中で，ワークステーション16台を接続したシステムの構築，評価を進めている．現状ではDMA転送のオーバヘッドが大きいためSTAFF-Linkの性能が十分に発揮できておらず，その原因を調査し改善を行なっている．

参考文献

[1] 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”, 情処研報 ARC107-24, pp.185-192, 1994.

[2] 高橋 淳, 中條 拓伯, 小畑 正貴, 金田 悠紀夫, “STAFF-Linkを用いたワークステーションクラスタ上へのPVMの実装とその評価”, 情処研報 HPC57-1, pp.1-6, 1995.

[3] A. Geist, A. Beguelin, J. Dongarra, W. Jiang, R. Manchek, V. Sunderam, “PVM : Parallel Virtual Machine - A Users' Guide and Tutorial for Networked Computing”, MIT Press, 1994.

[4] S. A. Moyer, V. S. Sunderam, “Characterizing Concurrency Control Performance for the PIOUS Parallel File System”, Dept. of Math and Computer Science, Emory University, CSTR-950601, 1995.