2J-06 柔軟な分散共有メモリボードの実現と広域分散予備方式への応用

向井 良　　田中 聡
NTT 未来ねっと研究所

1 はじめに

分散システムにおけるノード (計算機) 間の通信手段として，ハードウェアによる分散共有メモリ (DSM) を用いる方法がある．ハードウェア DSM は，通信処理に MPU が介在しないため，MPU にとって低オーバヘッドな通信が可能である．

筆者らはこれまでに，DSM を用いた軽いメッセージ通信機構 MESCAR(MEmory-coupled SCalable ARchitecture) [1] や，メモリ空間を全て DSM 上にマッピングして，広域ネットワーク上でメモリ情報を二重化することによってノードの高信頼化を図るネットワークワイド予備 (NWW 予備方式)[2] を提案してきた．

今回，DSM 上でトランザクション単位のメモリ書き替えを行うための機能拡張を行い，大規模なトランザクションシステムに適用することを目的として FPGA(Field Programmable Gate Array) と SRAM モジュールおよび交換可能な物理層ネットワークインタフェースから成る柔軟な分散共有メモリボード (Flexible DSM ボード，以下 FDSM) を作成した．本ボードはプロトコルの変更や機能拡張に対して柔軟に対応することができるため，ハードウェアの開発期間およびターンアラウンドタイムの短縮が可能となった．

以下に，ボードの構成と広域分散予備方式への適用例について述べる．

2 ハードウェア構成

FDSM ボードの構成を図 1 に示す．

(1) メインボード

メインボードは，FPGA とメモリモジュールを搭載し PCI インタフェース，メモリインタフェース，データリンク層ネットワークインタフェースを FPGA 内に実装した．必要な回路を可能な限り FPGA 内に実装することにより，システムの柔軟性を高めることができる．FPGA には，容量と I/O ピン数および実装パッケージの観点から，大容量で I/O ピン数が多くかつ実装が容易な Pin Grid Array 型のパッケージである ALTERA 社の APEX20K400GC655[3] を使用した．

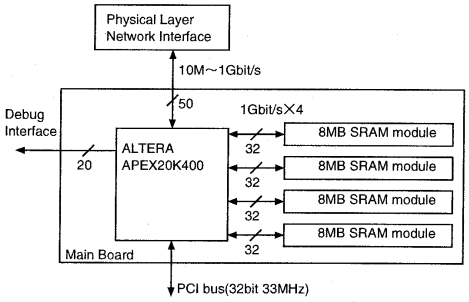

図 1: FDSM ボードの構成

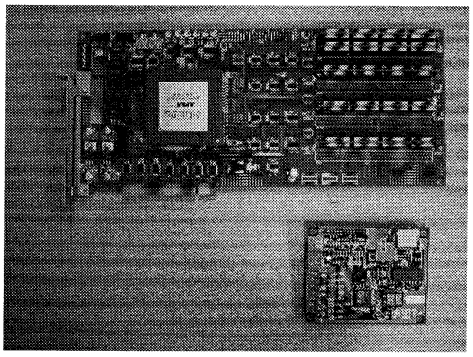

図 2: メインボード (上) と物理層インタフェースボード (下)

メモリはバンクあたり 8MB，データ幅 32bit の SRAM モジュールを 4 バンク構成で実装した．今回想定しているアプリケーションではバースト的な転送よりもランダムなアクセスが多いことと，今回開発したボードはプロトタイプ的な使用を想定していることから，容量よりも速度を優先して SRAM を使用した．メモリに対しては 1 バンクあたり 125MB/s 以上のランダムアクセスが可能となっている．アドレスバス，データバスは FPGA から各バンクに独立に設けているため，128bit 幅のメモリとしても使用することができる．

(2) 物理層インタフェースボード

現状では，物理層のネットワークインタフェースまでを FPGA 内に実装することは困難である，そこで FDSM

Implementation of Flexible Distributed Shared Memory Board and Application for Wide-area Back-up System,
Ryo Mukai and Satoshi Tanaka, NTT Network Innovation Laboratories, Musashino-shi ,180-8585 Tokyo, Japan

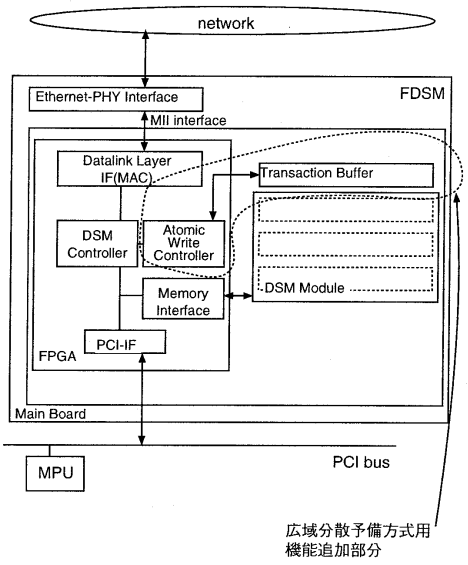

図 3: 広域分散予備方式用ノードの構成

は，専用 LSI を使用した物理層インタフェースを用意し，それを着脱可能にすることで複数の物理レイヤ (Ethernet, ATM 等) に対応できるようにしている．

今回の実装ではメインボードからのインタフェースに MII(Media Independent Interface)[4] を使用し，100Base-TX の Ethernet 用の物理層インタフェースボードに接続している．インタフェースの性能的には Gigabit Ethernet にも対応可能である．

3 アプリケーション

本ボードをメモリ同期型広域分散予備方式 [5] のためのノードに適用した例について説明する．

3.1 メモリ同期型広域分散予備方式

広域分散予備方式は，DSM を用いて運用ノードの状態を遠隔地にある予備ノードに逐次コピーし，運用ノードの障害時に予備ノードに切り替わることによってシステム全体のアベイラビリティの向上を図ることを目的としている．従来はログをファイルレベルでコピーして同期させていたものを，メモリレベルでの同期にまで細かくすることにより，切り替えの高速化や処理の継続性を高めるというねらいがある．

広域分散予備方式用ノードの構成を図 3 に示す．

ユーザプログラムは DSM の置かれている I/O 空間を，自分の論理空間にマップしてアクセスする．DSM へのライトアクセスはネットワークを通して他ノードの DSM 上にも書き込まれる．通常の DSM では，あらかじめ決められた単位のメモリアクセス毎に逐次他系に送信するため，複数ワードのアクセスから成るトランザクションを，他系に送信する際に，書き込みのアトミック性が失われるおそれがある．そのため，可変長の複数のメモリアクセスを 1 つのパケットに乗せて送信する機能 (アトミックライト機能) を追加する．これは通常の DSM ボードの機能にアトミックな書き込みをするための蓄積バッファを追加することで実現する．

今回想定しているアプリケーションでは，アトミックな書き替えの量は 16 ワード以内としているため，FPGA 内蔵のメモリをバッファに使用しているが，もっと大きなバッファが必要な際には，DSM 用メモリとして使用しているメモリバンクのうちの一つを割り当てるといった変更も FPGA のプログラム書き替えのみによって可能である．

このように，FDSM ボードを用いることにより，アプリケーションに特化した機能追加を容易に行うことができる．

4 まとめと今後の課題

FDSM ボードの構成と広域分散予備方式への適用例について説明した．今回の実装では，回路を可能な限り FPGA 内に実装し，専用部分も着脱可能にすることにより，DSM 機構への機能追加や変更を柔軟に行うことができた．今後は，物理インタフェース部分の高速化，他のアルゴリズムによる分散予備方式システムへの適用を考えている．

最後に，本研究を進めるにあたり，ご討議いただいた (株)NTT データ 横山和俊氏に感謝いたします．

参考文献

[1] Yamada, S., Tanaka, S. and Maruyama, K.: A Message-Coupled Architecture (MESCAR) for Distributed Shared-Memory Multiprocessors, in *ICPP*, pp. I–19–23 (1995).

[2] 向井良, 山田茂樹, 田中聡, 田中晶：メモリ間コピー機構を活用したネットワークワイド予備方式のハードウェア構成, 信学会春季総合大会 D-10-14 (1999).

[3] Altera Corp., : APEX20K Data Sheet (1999), http://www.altera.com/.

[4] Local and Metropolitan Area Netowrks (1998), IEEE Std 802.3.

[5] 横山和俊, 箱守聰：ハードウェア分散共有メモリを用いた広域分散予備方式の検討, 情報処理学会第 60 回全国大会 2J-05 (2000).