超並列計算機における並列2次記憶の基本アーキテクチャ

3B－5

大西一正　北村徹　大上靖弘　清水雅久
RWCP 超並列三洋研究室

1　はじめに

超並列計算機は、大規模なシミュレーションや音声・画像理解などのさまざまな分野で利用されると考えられる。それらのアプリケーションが扱うデータは膨大なものであり、この大量のデータを2次記憶に蓄え高速に効率良く運用管理するための手法を構築することが必要である。また、超並列計算機は膨大な演算資源を有しているため、多数のユーザが大量のプロセッサを分割使用することにより、システムを同時に利用することが多いと考えられる。このような環境においても、システムは柔軟に対応でき、ユーザに対して使い易く、かつ効率の良い2次記憶を提供しなければならない。

以上の観点から、筆者らは超並列計算機に適した並列2次記憶のアーキテクチャについて検討を行なってきた。本稿では、超並列計算機のための並列2次記憶に対する要件を述べたあと、それらの要件を満たすための並列2次記憶システムの基本アーキテクチャを提案し、その基本機能と特徴について述べる。

2　並列2次記憶に対する要件

超並列計算機における並列2次記憶について、前節で述べた「高速性」と「柔軟性」を重点として要件を抽出した。

超並列計算機は膨大な演算資源を有しており、複数のユーザが同時に使用するため、計算機上のプロセッサ空間がユーザによって分割利用される。そこで我々は、1000台規模のプロセッサがいくつかのプロセッサグループ（クラスタ）に分割され、複数のプロセスがクラスタを境界として実行される計算機を想定した[1][2]。

以上のような想定のもとに抽出した並列2次記憶に対する要件について述べる。

1. アクセスに関しプロセス間での相互干渉が少ないこと

   ファイルシステムは各プロセスの共有資源であるため、プロセス間相互でそのアクセスにおいて影響を及ぼし合うが、各プロセスのページングについては、相互干渉がないことを要件とする。

2. 十分なデータ転送能力を持つこと

   超並列計算機では、大量のプロセッサが同時にデータ処理を実行するため、2次記憶に対するアクセス要求は膨大である。並列2次記憶は、プロセスの進行を阻害することなく、プロセッサからの多数のアクセス要求に対して迅速に応答できるデータ転送能力を有することが必要である。

3. ファイルシステムがどのプロセッサからも一様に見えること

   超並列計算機として使い易いものであるためには、使用上の制限が少なく、プロセスも任意のプロセッサ（クラスタ）に割り当て可能な柔軟なシステムであることが望まれる。これを実現するには、プロセスがどのプロセッサに割り当てられても、同様な環境で実行できることが必要であり、ファイルシステムも各プロセスから一様に見えなければならない。

4. スケーラビリティを持つこと

   超並列計算機のスケーラビリティに応じて、並列2次記憶も柔軟に拡張できかつスケーラブルであることが必要である。

5. 十分な信頼性があること

   超並列計算機においては必要な2次記憶の容量は膨大であり、多数のディスクデバイスを用いることが考えられる。そのため、並列2次記憶システムの十分な信頼性を維持する必要がある。

3　並列2次記憶のアーキテクチャ

3.1　基本アーキテクチャ

前節に述べた要件を前提にして、図1に示すような並列2次記憶の新しい基本アーキテクチャを提案した。

提案した並列2次記憶は、ファイルシステムを蓄えるディスク（ファイルシステム用ディスク）をページングなどを行なうローカルなディスク（ページング用ディスク）から分離した形態のものであり、以下のような特徴を持つ。

- 各プロセスの共有資源であるファイルシステム用のディスクをページング用のディスクか

Basic Architecture of Parallel Secondary Storage for Massively Parallel Computers.

Kazumasa OHNISHI, Toru KITAMURA, Yasuhiro OUE, Masahisa SHIMIZU.
Massively Parallel Systems Sanyo Laboratory, RWCP
研究実施場所：三洋電機(株) 東京情報通信研究所

ら分離し、ページング用のディスクを各クラスタ内でのみ利用することにより、ディスクアクセスが他のクラスタのページングを阻害するのを防ぐ。このような構成をとることによりプロセス間での相互干渉が少ない並列2次記憶システムを実現できる。

- ファイルシステム用のディスクを各クラスタに直接接続するのではなく、相互結合網を介して接続するため、各クラスタから見たファイルシステムの一様性を実現することが可能となる。
- 同様の理由でファイルシステム用のディスクの接続台数などに関する自由度が高く、スケーラビリティを実現するために適している。
- 本方式では永続記憶がページング用のディスクから独立しているため、ページング用のディスクはファイルシステム用のディスクと同等の信頼性を有する必要はない。例えば、故障前のプロセスを復元できるような構成のＯＳを用いることにより、システムとしては十分な信頼性を保つことが可能となる。

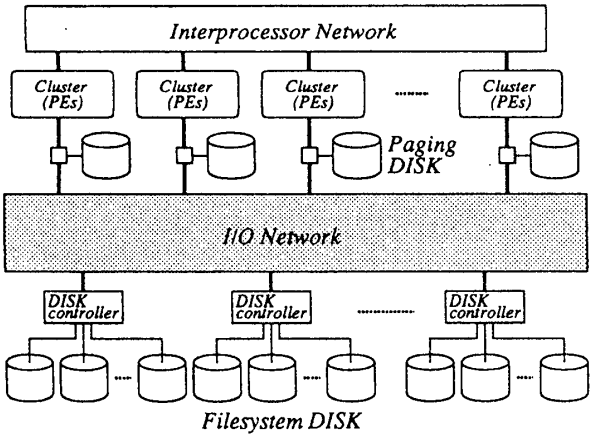

図 1: 並列2次記憶の基本アーキテクチャ

3.2 相互結合網

図1に示すように、相互結合網には複数のディスクシステムが接続される。これらのディスクシステムは、相互結合網を介して、超並列計算機上の多数のプロセッサからのデータ要求等に応答しなければならないため、相互結合網には高いデータ転送能力が要求される。また同時に、相互結合網には、ディスクシステムのスケーラビリティを実現するための機能や耐故障性が必要となる。

このような要求を満たすものとして、クロスバー網、fat-tree、階層型のリング網などが考えられる。クロスバー網は非閉塞網であるがノード数が多くなった場合にハードウェア量がＮの2乗オーダで増加し、超並列計算機に応用するには実装面での課題がある。fat-tree も非閉塞網であるが完全に実現しようとすればハードウェア量の増大が課題となる。また、階層型のリング網は構造が簡単で実現が容易であり、結合網にバッファを持たなくて良いという利点があるが、閉塞網でありブロッキングが発生してしまう。

このようにいずれの結合網にも課題が存在する。従って、シミュレーションによる比較評価結果等を用いて、結合網を決定していく予定である。

3.3 ディスクシステム

図1に示すように、ファイルシステム用のディスクについては、複数のディスクをコントローラで制御する構成をとる。コントローラで制御する各ディスクについては、高い信頼性が要求されるため、例えばデータに冗長性を持たせたディスク構成 (RAID など) にすることを検討している。

相互結合網には、以上に述べたようなディスクシステムが複数接続され、超並列計算機上のプロセッサからのアクセス要求は、このディスクコントローラに対して行なわれる。従って、多数のプロセッサからの要求を効率良く高速に行なうために、コマンドキューイングやバッファキャッシュなどを用い、ディスクシステムの制御を行なう方法が考えられる。

4 まとめ

超並列計算機のための並列2次記憶システムに対する要件を抽出し、並列2次記憶についてそれらの要件を実現するための新しい基本アーキテクチャを提案した。また、本アーキテクチャの特徴と相互結合網等について述べた。

今後、ソフトシミュレーションによる予備評価の結果等に基づき、相互結合網のトポロジーやディスクシステムの詳細な制御方法などを決定するとともに、並列2次記憶のプロトタイプを開発し、提案した手法の総合的な検証を行なう予定である。

参考文献

[1] 坂井他. 超並列計算機 *RWC-1* の基本構想, 並列処理シンポジウム JSPP'93, pp. 87-94,(1993).

[2] 廣野他. 超並列計算機 *RWC-1* における入出力機構, 情報処理学会研究報告 93-ARC-101, pp. 33-40,(1993).

[3] 大上他. 超並列計算機のための並列ファイルシステムの基本構成, 情報処理学会第 48 回全国大会講演論文集, 3B-6,(1994).