異なるアーキテクチャを持つ複合計算機システムの検討

7M−10

阿部 薫、岡本 弘、西井 龍五

三菱電機(株) 情報電子研究所

1　はじめに

近年、より高いパフォーマンスを目指して、CPU単独の性能向上以外に複数の計算機を複合するいわゆる複合計算機システムが多数提案されている.これらのうち並列型とよばれるものを除けば、主にタスクの負荷分散を行なうことでパフォーマンスの向上を目指すものであることは周知のとおりである.

こういった複合計算機システムのうち密結合型のものは一般にマルチプロセッサと呼ばれ、同じタイプのアーキテクチャを持つCPUをバス結合し、シェアドメモリ構造をとり唯一のOSのもとで動作するものである.

これは、個々のタスクをそれぞれのプロセッサに割り当てるというリソース管理により全体としての処理時間の短縮を図るもので、プログラム処理をタスク処理の集合という面でとらえたときには良い方法である.しかしながら、割り込み応答性やグラフィクス処理といったシステム機能やプロセッサ構成に密接に関連するパフォーマンスは、いわばアーキテクチャによる得手不得手が影響するためにマルチプロセッサ構成としても期待通りの性能向上が実現できない場合が多々存在する.

そこで、筆者らは異なるアーキテクチャのコンピュータ群をバス結合した複合計算機システム（ヘテロジニアスマルチコンピュータシステム）を構築し、それぞれのタスクの処理形態により適合したコンピュータに処理分散する方式について検討を行なった.本報告では、検討に用いた計算機モデル及びこの方式実現のために必要な計算機間でのデータ共有、割り込み、排他競合制御の具体的方策について述べる.

2　計算機モデル

異なるアーキテクチャのコンピュータ群を結合して分散環境を構築することについては、従来からLANなどのネットワークを用いた方法が広く行なわれてきた.

ネットワーク結合は柔軟性が高い反面、データ転送の絶対性能が低いと言うことだけでなく、転送動作のオーバヘッドが非常に大きいという欠点を有する.これを解決するために、2つの独立したコンピュータをハードウェアによるリンケージ機構を用いてバス結合し、シェアドメモリ構造とするモデルを想定した.

2つの独立したヘテロジニアスなコンピュータの主記憶をシェアドメモリ構造とするうえにおいて問題となるのはメモリ管理である.主記憶はOSからは一般に論理（仮想）アドレスで管理されているので双方の主記憶のアクセスをOSの主記憶管理機構を用いて対称に行なうためには双方のOSにそれぞれ両方の主記憶管理機構を組み入れなければならず、また構成としても冗長である.そこで今回の検討では、片側のコンピュータのメモリ（主記憶）の一部を交信バッファとして物理モードで共有し、もう一方のコンピュータからはIO空間として透過的にアクセスすることとした.このように非対称な構成とすることで、以下に挙げる利点を得ることができる.

- OSを含むそれぞれのコンピュータ上のソフトウェアは自己の主記憶管理およびIOアクセスとして交信が可能となる.
- 物理的なメモリ共有のためデータ転送を高速化できる.
- それぞれのコンピュータはシステムの中でマスタスレーブとして動作するため、システム構築が容易となる.

3　構成と各機能

図1に検討に用いたモデルの構成を示す.

ここで、コンピュータAの主記憶上に交信バッファを共有しコンピュータBは図中リンケージ機構と示されたハードウェアを用いてアクセスする.

このリンケージ機構においては、機能検討はLANを用いたシミュレーション、またデータ転送能力などの性能検討はハードウェアモデルによるシミュレーションを行なった.

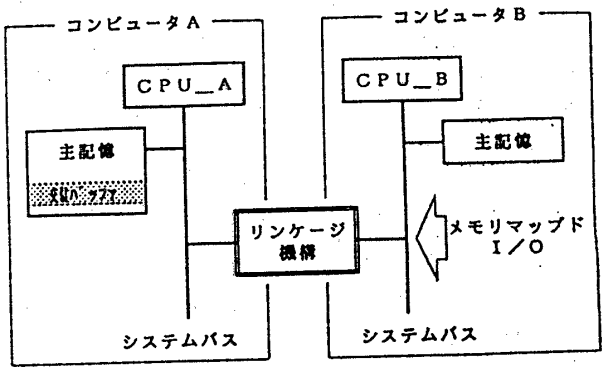

図1：モデルの構成

Multi-Computer system with Heterogeneous Architechture
Kaoru ABE, Hiromu OKAMOTO, Ryugo NISHII
Mitsubishi Electric Corp.

3.1 データ共有

複合計算機システムによる分散処理では、処理の依頼と処理データの転送を高速に行なう必要がある.
一般に、処理データについては例えばグラフィクス処理などでは非常に大きなデータ量となるのに対し、処理の依頼はフラグやメッセージデータ等データ量は僅少である.大量のデータを効率良く得るためにはいわゆるDMA転送が有利であるが、フラグ操作や小量のデータ転送ではワードアクセス（CPUによるリードライト）が有利である.したがってリンケージ機構の機能としてはDMA転送とワードアクセスの両方を可能とし、これら2つの転送方法を統一的にIOドライバでサポートし転送量に応じて使い分けることとした.交信バッファは、フラグやコマンドに用いる領域と処理データの格納に用いる領域を分けることにする.こうすることで、交信バッファのアドレス管理や容量を上回るような大量データのチェイニングを容易にすることが出来る.

3.2 割り込み

複合計算機システムでは、アプリケーションタスクを複数のコンピュータ間で分散して行なうためにプログラムの開始や終了などの通知と電源断や暴走などそれぞれのコンピュータの状態を相互監視する必要がある.
これらの機能は、コンピュータ間の同期をとりながら行なう必要があるので双方向の割り込み通信で行なう.

この割り込みは、外部割り込みとして認識される.今回の検討では、相互監視とアプリケーションタスク用のそれぞれに1本ずつ割り当てた.

相互監視については、以下の方法による.

- ステータス情報などと共に割り込みを相手側に発生する.
- 割り込みを受けた側は一定時間以内にその応答としてステータス情報などと共に割り込みを送り返す.

このプロセス（割り込みタスク）を相互に繰り返すこととし、一定時間以上割り込みが返ってこなければ相手側が異常であると判断する.またこの監視のための割り込みタスクは、アプリケーションタスクの起動時にタスクプライオリティをより低いものに変更する.こうすることで、起動されたアプリケーションタスクの暴走は相手側における割り込みのタイムアウトとして検知することが出来る.

またアプリケーションタスクの同期は、以下に示すようにコマンド及び終了ステータスとともに割り込みを相互に通信することで行なう.

- コマンドと共に割り込みを相手側に送信して処理依頼をする.
- 処理を受けたコンピュータは処理の終了後、終了ステータスと共に割り込みを処理依頼元に送信する.

3.3 排他制御

交信バッファ領域などリソースの競合制御を行なうために、セマフォによる排他競合制御を行なう.このセマフォは、予めソフトウェアの規約で交信バッファ上の特定番地に設けられたフラグを比較／置換することでセマフォロックを行なう.このフラグの比較／置換はアトミシティを保証しなければならないので、ハードウェア・セマフォを併用する.このハードウェア・セマフォの実装箇所によってさまざまな方法があるが、バス上のセマフォ線及びメモリ装置にセマフォ機構を設けるには、既存のバス仕様やメモリ装置に変更を加えなければならない.今回の検討モデルでは、結合のための機能をリンケージ機構に集中させる形態としたため、リンケージ機構上にハードウェア・セマフォフラグを設ける方法とした.

4 性能評価

例として、熱解析プログラムの処理をシミュレートしその開始と終了に要する処理時間（すなわち実計算時間以外の処理時間）をLAN(IEEE 802.3)結合の場合と比較した.
このプログラムは、10Kバイトの温度データを処理し約30Kバイトの計算結果を得るもので、コンピュータAがFEPとして動作することにより得た温度データを、コンピュータBに処理依頼をして、再びコンピュータAに送り返すというプロセスのシミュレーションである.尚、これらの処理時間は複合計算機システムのオーバヘッドに相当し、処理プログラムのアルゴリズムには依存しない.結果を下に示す.

処理	LAN結合モデルに対する時間比	
	本検討モデル	LAN結合モデル
プログラム開始指示	0.05	1
バッファからデータリード	0.15	1
バッファへデータライト	0.25	1
TOTAL	0.2	1

上記結果から明らかなように、開始および終了に要する処理時間の合計はLAN結合の場合のほぼ1／5である.これらのうち、特にプログラム開始指示はLAN結合の場合の1／20と非常に高速であり複合計算機システムの結合手段として本検討方式が有効である結果を得た.

5 おわりに

複合計算機システムの構成とその結合手段であるリンケージの機能について検討した.
検討の結果、LAN結合によるものと比較してパフォーマンスの著しい向上が見込めることが判った.今後は、2台の対向のモデルから発展したクラスタの構成、及びデータ転送系の性能向上を中心に耐環境性などを考慮した更に高性能で堅牢な複合計算機システムとその実現化方式について検討を続けて行きたい.