計算機アーキテクチャ　68－14
(1987・11・13)

ネットワーク仮想記憶方式による マルチプロセッサの試作について

An Experimental Distributed Multiprocessor Based on Networked Virtual Memory System

陣崎　明　　樋口　昌宏　　八星　禮剛
Akira Jinzaki　　Masahiro Higuchi　　Reigo Yatsuboshi

(株) 富士通研究所
Fujitsu Laboratories Ltd.

あらまし　分散型マルチプロセッサシステムの構成方式として提案しているネットワーク仮想記憶(NET-VMS) 方式に基づくマルチプロセッサシステムの試作について述べる．NET-VMS は分散した計算機の仮想記憶システムをブロードキャストネットワークで結合し全体で単一の仮想記憶システムを構成するもので，従来のバス結合システムの接続距離やプロセッサ数の限界を解消し，より大規模な密結合システムの構築を目指すものである．試作システムは最大16台の計算機を最大リング長1.6km，伝送路速度125Mbps の光リングネットワークで結合するもので，待行列モデルによる評価により最大11MB/ 秒の実効通信性能を得る見通しを得た．

Abstract　In this paper, we describe an experimental geographically distributed multiprocessor system based on Networked Virtual Memory System (NET-VMS) architecture. NET-VMS realizes single virtual memory system connecting distributed virtual memory systems with a high speed broadcasting network. The experimental system connects 16 computers using up to 1.6km long 125Mbps optical-fiber ring network. As the result of analysis, the experimental system achieves 11MB/s effective communication rate.

1. はじめに

1980年代にはいってから分散・並列処理システムの商用化が盛んになっている[1][2]．特に最近の動向として，画像処理，人工知能など従来から盛んに研究されている分野だけでなくオンライントランザクション処理など汎用処理分野における分散・並列処理システムの実用化に関心が高まっており，これに伴って処理の高速化だけでなく，システムの大規模化（多プロセッサ化，異機種間結合，地理的分散の実現），高信頼化（フォールトトレランシの実現）が重要なテーマとなっている．

従来いわゆるＬＡＮなどのネットワーク結合といえば疎結合システムを意味し，マルチプロセッサシステムなどの密結合システムは並列バス結合が主流であった．しかし上に述べたようなシステムの大規模化，さらに大規模システムにおける高信頼化に対応するためには接続できるプロセッサ数，接続距離，物理条件の制限が強いバス結合では限界であって，バス結合に置き換えうる性能をもつ新しいネットワーク結合方式が不可欠である．

我々は以上の観点から密結合型分散処理システムを狙いとしたプロセッサ結合方式としてネットワーク仮想記憶システム(NET-VMS: NETworked Virtual Memory System)方式及び本方式によるプロセス間通信方式として宣言的プロセス間通信方式[3]-[6] を提案した．現在本方式に基づく試作システムの開発を進めているが，ＮＥＴ－ＶＭＳ方式の実現性を確認すると共に１

6台の計算機を計算機間接続距離１００ｍ程度で分散配置し，最高８～１１ＭＢ／秒のプロセッサ間通信性能をもつマルチプロセッサシステムを構成できる見通しを得たので報告する．

2. ネットワーク仮想記憶方式

ネットワーク仮想記憶（ＮＥＴ－ＶＭＳ）方式の詳細については後に示す試作システムの説明及び文献(3)に譲り，ここでは論理的な構成と本方式の特徴について述べる．

ＮＥＴ－ＶＭＳは分散した計算機の仮想記憶システムをブロードキャストネットワークによって結合し，全体を単一階層の共有仮想記憶として構成することにより，ネットワークを介した通信を見掛け上共有メモリを介した通信として実現するシステムである．図１に構成を示す．

ＮＥＴ－ＶＭＳにおける通信は常に共有仮想記憶空間上の共有領域を介して行う（図２）．このためＮＥＴ－ＶＭＳ方式では共有仮想記憶に対するアクセス制御（排他アクセス制御，アクセス同期制御，メモリ複写／無効化制御）をネットワークシステムとメモリシステムが実現する．この結果オペレーティングシステムを含む全てのソフトウェアはセマフォなどを用いたアクセス制御のプログラムを必要としない．我々はこれを宣言的プロセス間通信と呼んでいる．

宣言的プロセス間通信とはおおよそ次のようなものである（図３）．まず送信側プロセスは共有仮想記憶空間上の通信領域に対して排他的書き込みアクセスをする由メモリシステムに通知（これを宣言という）してから実際のアクセスを行う．一方受信側のプロセスは排他的読みだしアクセスをする由メモリシステムに通知してから実際のアクセスを行う．通信領域は送信側と受信側のアクセスに従ってシステム内を移動するが，この移動制御や移動が完了したことをプロセスに通知する処理は先の宣言に基づいてＮＥＴ－ＶＭＳが行う．

ここで重要なのは宣言が実行に先立って行われる点である．通常の共有メモリシステムでは排他アクセスなどを共有メモリ上のセマフォによって解決するが，この方法にはセマフォへのアクセス競合（Ｈｏｔ　Ｓｐｏｔ[7]）の問題，セマフォアクセスなどアクセス制御処理をソフトウェアに埋めこむためソフトウェアが複雑化する問題，複数のプロセスが動的にセマフォを変更するためデバッグが困難となる問題などがある．これに対してＮＥＴ－ＶＭＳの宣言は各メモリシステムに対して行われるので宣言の競合がなく，ソフトウェアにアクセス制御処理を埋めこむ必要もない．

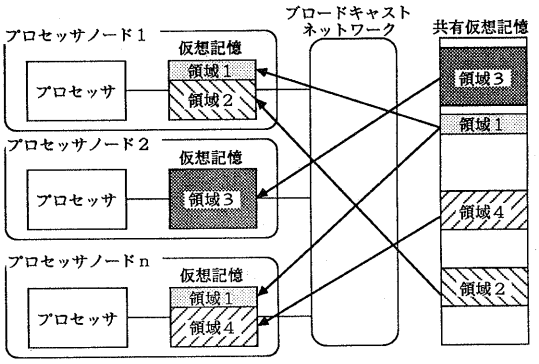

図１　ＮＥＴ－ＶＭＳの構成

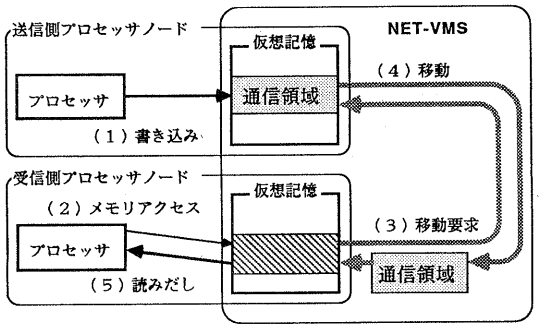

図２　ＮＥＴ－ＶＭＳの通信

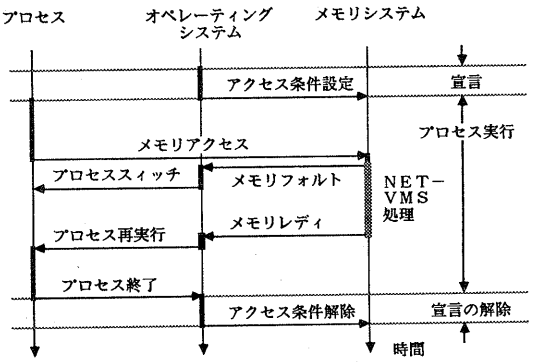

図３　宣言的プロセス間通信の動作

3. 試作システム開発の目的

試作システム開発の目的は次の点にある．

(1)　ＮＥＴ－ＶＭＳ方式の検証

第一の目的はＮＥＴ－ＶＭＳ方式及び宣言的プロセス間通信方式の実用化検討，性能評価，

有効性の検証である．このためにＮＥＴ－ＶＭＳ方式をほぼ完全に実現し，試作システム上で各種プロセス間通信方式（宣言的プロセス間通信，メッセージ通信，共有メモリ通信など），分散・並列処理アプリケーションの実装及び評価を行う．

(2)　プロセス間通信の性能測定

分散・並列処理におけるプロセス間通信はアプリケーションからネットワークハードウェアへ至る階層構造（図４）をなしているが，これらの階層間の通信量や処理量を個別に測定することは非常に重要である．特にアプリケーションやオペレーティングシステムが入り乱れて動作する状況において特定のソフトウェアに関するプロセス間通信の状況を実行している処理に影響を与えることなく把握することは従来のシステムでは行われていなかった．試作システム開発の第二の目的は本システムで実際に分散・並列処理ソフトウェアを実行し，プロセス間通信の状況を数量的に把握可能とすることである．

以上の目的を達成するために本試作では通信性能・処理性能評価機能のハードウェア組み込みを行うこととした．

4.　試作システム

4.1　システム構成

図５に試作システムの構成を示す．システムはプロセッサ，メモリシステム，ネットワークシステム，スーパバイザプロセッサからなるプロセッサノードを光ファイバリングで最大１６ノード接続したマルチプロセッサシステム，システム全体の制御を行うためのスーパ・スーパバイザプロセッサ，ソフトウェア開発及び実験結果の解析を行うためのホスト計算機からなる．

システム規模はネットワークの伝送路速度（１２５Ｍｂｐｓ）をベースとして決めた．まずプロセッサノード数はプロセッサのデータ処理能力を５００ＫＢ～７００ＫＢ／秒程度と見積もり，最大構成時にこの程度のメモリ間転送が可能となる（つまりプロセッサが処理するデータを全てネットワークから供給できる）よう１６台とした．

プロセッサノードに実装する物理メモリはまずプロセッサのメモリ空間（１６ＭＢ）を制御領域を除いて全て共有仮想記憶空間とすることとし，最大構成時（１６ノード）に共有仮想記憶空間の大きさと物理メモリ空間の大きさが同じになるよう１ＭＢとした．またＮＥＴ－ＶＭＳ方式ではネットワークのデータ転送を固定長のメモリページ単位で行うためメモリページサイズがシステムの性能に大きく影響することが考えられる．そこでページサイズを２５６Ｂ～４０９６Ｂの範囲で設定可能とした．

本試作の目的はＮＥＴ－ＶＭＳ方式の実現検討および性能の評価にあり，高速処理を行うことではない．またＮＥＴ－ＶＭＳはメモリシステムであるから基本的にプロセッサはなんでもよい．そこでプロセッサは，仮想記憶対応で，ソフトウェア開発環境が整い，従来ベンチマークが多く行われ他プロセッサとの比較が容易なものとしてＭＣ６８０１０を選んだ．

以下，ネットワークシステム，メモリシステム，性能測定部，システム制御についてさらに詳しく説明する．

4.2　ネットワークシステム

ネットワークシステムは伝送速度１２５Ｍｂｐｓ，リング長最長1.6Ｋｍの光ファイバリングネットワークで構成する．表１にネットワーク諸元を示す．本システムでは伝送符号として４Ｂ５Ｂ符号を用いるので，物理的な実効通信

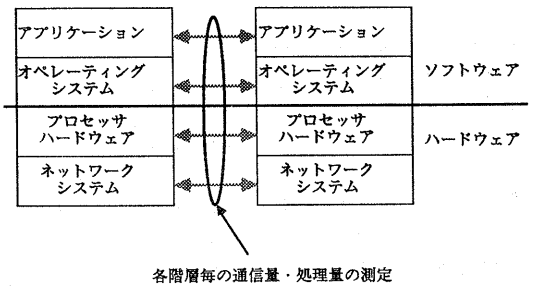

図４　プロセス間通信の階層

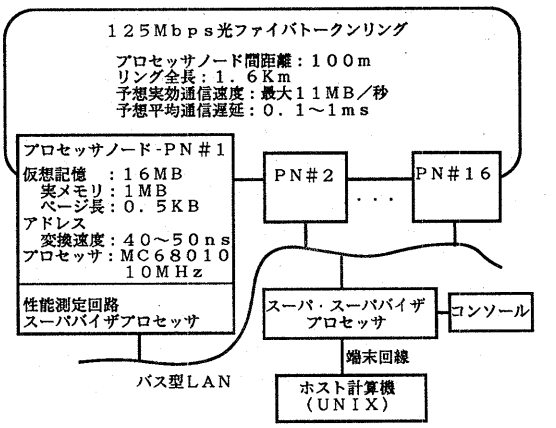

図５　試作システムの構成

表1　ネットワーク部仕様

項目	仕様
伝送速度	１２５Ｍｂｐｓ
伝送符号	４Ｂ５Ｂ符号　ＦＤＤＩとは全く異なる符号を用いる
制御シンボル	伝送路制御シンボル　３種 トークンシンボル　１種 フレームシンボル　１種 データシンボル　１６種 ＦＤＤＩシンボル　９種＋１６種
ネットワーク制御方式	アクセス方式はシングルトークン方式， ネットワーク接続管理，トークン管理は全てハードウェア制御， 単リング動作，二重リングに拡張可
フレーム長	最大８２０４シンボル　－　ページサイズ４０９６バイト時 最小　５２４シンボル　－　ページサイズ　２５６バイト時 固定フレーム長
ノード内遅延	最大１６０ｎｓ／ノード
リング長	1.6Ｋｍ，ノード間距離１００ｍ

速度は１００Ｍｂｐｓである．ネットワークアクセス方式としては独自のシングルトークンリング方式を採用している．

〔通信フレーム〕

図６に通信フレームを示す．ＮＥＴ－ＶＭＳ方式ではフレームは要求部分と応答部分からなる．図７に通信の様子を示す．要求を行うプロセッサノードはフレームの要求部分を他の全てのプロセッサノードに対してブロードキャストする．他のノードは要求部分に含まれるコマンドとメモリページアドレスをメモリシステムでアドレス変換することによって解析し，要求部分に引き続いて即座に応答する．このため要求部分と応答部分の間にはアドレス変換及び応答処理に見合うだけのギャップが設けられている．

〔ネットワークコマンド〕

コマンドとしてＣｏｐｙ，Ｕｎｉｆｙ，ＰａｇｅＯｕｔ，Ｍｏｎｉｔｏｒを設けた．Ｃｏｐｙは他プロセッサノードからメモリページを複写するコマンド，Ｕｎｉｆｙは他プロセッサノードからメモリページの複写をすると同時にそれらのノードのメモリページを無効化（Invalidate）するコマンド，ＰａｇｅＯｕｔは自メモリページを他プロセッサノードに複写させるコマンド，Ｍｏｎｉｔｏｒは他プロセッサノードのメモリページを排他アクセス中であるなしにかかわらず複写するコマンドである．

〔ネットワーク管理，エラー処理〕

ネットワーク管理制御はリング接続管理（リングが形成されたか切断されたかの検出），送信処理，受信処理をハードウェアで，トークン管理（トークン消失時の再発生）をスーパバイザプロセッサの制御で行う．

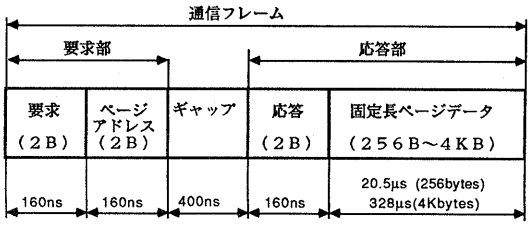

図６　通信フレームの構造

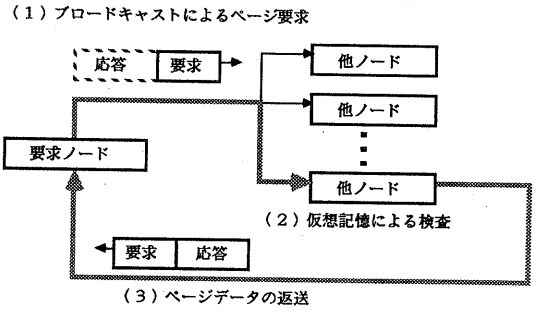

図７　通信の様子

ＮＥＴ－ＶＭＳ方式はフレームにＣＲＣコードなどを付加することにより，伝送エラーの検出及びエラー発生時の再送制御を完全にハードウェア制御で実現可能であるが，本試作ではハードウェアを単純化するため一切のエラー制御を行わないこととした．光伝送路のエラーレートは非常に小さいので試作システムでは実用上問題とならない．

4.3 メモリシステム

メモリシステムはアドレス変換用メモリ，仮想記憶制御回路，物理メモリ（各プロセッサノード当たり１ＭＢ）からなる（図８）．仮想記憶方式はデマンドページング方式を採用し，ページサイズは２５６Ｂ～４０９６Ｂの範囲で設定可能である．プロセッサの記憶空間１６ＭＢのうちアドレス変換メモリ，タグ用メモリ及び制御レジスタ領域に約５００ＫＢ使用するので共有仮想記憶空間は１５.５ＭＢである．

〔アドレス変換メモリ〕

ネットワークの説明で述べたようにＮＥＴ－ＶＭＳではフレームの要求部分をあらかじめ定められた時間（フレームのギャップの時間）以内にアドレス変換し，データの受信，送信，無効化などの処理を行う必要がある．従ってアドレス変換に要する時間はそのまま通信性能に影響を及ぼすので高速化が必要であるが，通常の仮想記憶システムで行われる階層型のページエントリテーブルをソフトウェアで検索する方法は

・ワーキングセットが共有仮想記憶空間全域に渡り，
・ネットワークからのアドレス変換要求をプロセッサが行わねばならない

ため高速化できない．

アドレス変換をハードウェアで行う方法としてはＣＡＭ（Content Addressable Memory）を用いる方法があるが，ＣＡＭはエントリ数が制限される（すなわち物理メモリの大きさが制限される）点が問題である．試作では６４Ｋエントリ（１６ＭＢ／２５６Ｂ）のアドレス変換テーブルを高速ＳＲＡＭで構成し，ページアドレスでこのテーブルをアクセスすることによってアドレス変換する方法をとった．アドレス変換テーブルは６４ＫビットＲＡＭ１８個（物理ページアドレス１２ビット，タグ６ビット）で構成されサイクルタイム５０ｎｓで動作する．

〔タグ〕

タグはメモリページ単位に設ける制御情報であって，ＮＥＴ－ＶＭＳ方式は基本的にＣＯＰＹ，ＶＡＬＩＤ，ＬＯＣＫ，ＳＹＮＣの４種類のタグを用いるが，本試作ではさらにＷＡＩＴＥＲ，ＴＥＳＴを追加した．それぞれの役割を表２に示す．

ここでＷＡＩＴＥＲ[8]は排他アクセスしているページに対して他のプロセッサノードがＵｎｉｆｙ要求を行った場合などに用いるもので，この場合ＮＥＴ－ＶＭＳはＵｎｉｆｙ要求を拒否する一方ＷＡＩＴＥＲタグをアサートすることによって他ノードから要求があったことをプロセッサに通知することができる．プロセッサは排他アクセスを終了する際ＷＡＩＴＥＲを検査して待ちプロセッサがいることを知り，ＰａｇｅＯｕｔ等でアクセス権の譲渡を行うことができる．ＴＥＳＴタグは性能測定用である．

〔仮想記憶制御回路〕

仮想記憶制御回路はタグの検査，メモリシステムのマルチポート化を実現する．高速のＰＬＡ（Programmable Logic Array）で構成されている．

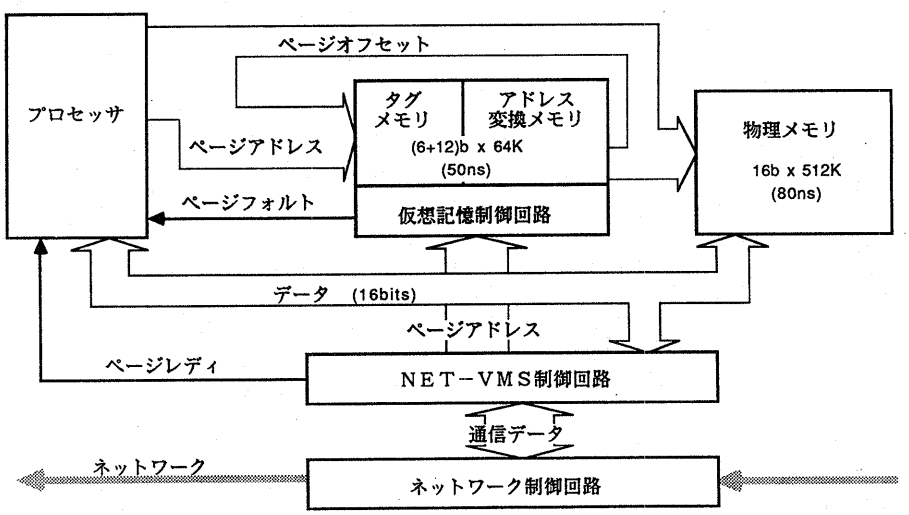

図８　プロセッサノードとメモリシステム

表2　メモリタグ

タグ	意味	設定	解除
COPY	メモリページが複数のプロセッサノードに存在することを示す	NET－VMS	NET－VMS
VALID	自ノードにメモリページが存在することを示す	NET－VMS ソフトウェア	NET－VMS ソフトウェア
LOCK	自ノードがメモリページに対して排他アクセス中であることを示す	ソフトウェア	ソフトウェア
SYNC	自ノードのメモリページを他ノードがアクセスするまで自プロセッサはアクセスできないことを示す	ソフトウェア	NET－VMS
WAITER	自ノードが排他アクセス中に他ノードがアクセス要求したことを示す	NET－VMS	ソフトウェア
TEST	メモリページが性能測定の対象となっていることを示す	ソフトウェア	ソフトウェア

〔物理メモリ〕

ＮＥＴ－ＶＭＳ方式では高速のネットワークを使用するので，物理メモリはネットワークとプロセッサの両方に対して充分なスループットを提供する必要がある．具体的にはネットワークの伝送中にオーバランあるいはアンダーランが起こるのを避けるために伝送路の転送速度を上回る程度のメモリスループットが最低限必要である．ネットワークは他のプロセッサノードと同期して動作しているので，もしネットワークとプロセッサのメモリアクセスが競合した場合はネットワークが優先し，プロセッサはウエイトしなければならない．試作の場合，物理メモリアクセススループットはネットワークが１２.５ＭＢ／秒，プロセッサは最大５ＭＢ／秒である．従って物理メモリのスループットは１２.５～１７.５ＭＢ／秒は必要ということになる．

本試作では実際のソフトウェアを動作させた時のプロセッサ処理時間やメモリアクセス性能を実時間測定することを目的としているが，ネットワークとプロセッサの競合がおこると測定結果に不確定な影響を与える可能性が生じる．この影響をさけるため，仮想記憶を完全にマルチポート化し，プロセッサ，ネットワークシステム，スーパバイザシステムそれぞれが見掛け上アクセス競合なく同時アクセス可能とした．これを実現するためにメモリのサイクルタイムは物理メモリで８０ｎｓ（２５ＭＢ／秒）とし，高速ＳＲＡＭを用いて実現している．

4.4　性能測定

本システムではプロセス間通信をすべて共有仮想記憶上で行うことを利用して非常にきめこまかな性能測定が可能である．具体的にはＴＥＳＴタグを用いて特定のメモリページに関して選択的な性能測定を行う．ソフトウェア実行前にあらかじめＴＥＳＴタグを設定しておき，実行中にメモリアクセスやメモリページの転送時にこのＴＥＳＴタグ設定されている場合のみカウントや時間の測定を行うことによって多重動作している複数のプログラムの中から特定の処理を行っている部分や特定のプログラムモジュールの動作をとらえることができる．この結果前に述べたプロセス間通信の階層毎の通信量や処理量を容易に測定可能となる．

以上のような測定は例えばバス結合やＬＡＮ結合システムではネットワーク上を流れるデータをすべてモニタし，データのアドレスや通信を実行しているプロセス名などの内容を解析しなければ実現できないし，十数台のプロセッサが並列動作している状況をリアルタイムでモニタするのは非常に困難である．

性能測定は８組の１６ビットカウンタ及び２組のインターバルタイマを用い，メモリアクセス回数，ページヒット／ミス回数，ネットワーク転送遅延を測定できる．またインターバルタイマはバッファメモリを持ち，１０００回分の測定データを保持することが可能である．性能測定及び結果の収集は後に述べるスーパバイザプロセッサの制御によって自動的に行うことができる．

4.5　システム制御

システム制御はスーパバイザプロセッサ（ＳＶＰ）によって行う．ＳＶＰは８ビットマイクロプロセッサをもつシングルボードコンピュー

タであって，次の機能を持つ．
① メインプロセッサの停止，起動，割り込みの発生．
② メインプロセッサを停止した状態で，物理メモリ及びアドレス変換・タグ用メモリの設定，メインプロセッサ用プログラムのダウンロード．
③ ネットワークシステムの状態監視及びフリートークンの生成．
④ 性能測定部の初期化，測定起動，停止，測定結果の収集，ホストへの通知．

ＳＶＰはプロセッサノード１台毎にある他，全体をまとめるためのスーパ・スーパバイザプロセッサ（ＳＳＶＰ）があり，バス型ＬＡＮによって相互接続されている．通常のシステム制御はＳＳＶＰのコンソールからＳＶＰにコマンドを送ることによって行う．またＳＳＶＰはホストと回線によって接続され，プロセッサノードプログラムのダウンロード，測定データのアップロードを行う．

ホスト計算機はＵＮＩＸワークステーションであって，実験ソフトウェアの開発，収集した測定データの解析を行う．

5. 予想性能

本試作システムの最大構成時の予想性能を待ち行列モデルによる解析を用いて評価した結果を図９に示す．１６プロセッサノードが全て同じ量のメモリページ転送を行うと仮定し，メモリページサイズを５１２Ｂとすると全体で１１ＭＢ／秒（物理伝送路速度の約７０％）程度のデータ通信速度が得られ，この時のメモリページ単位の転送遅延は約７００μsとなる．６８０１０を１０ＭＨｚで動作させた時のデータ処理性能を概ね５００～７００ＫＢ／秒程度とすると，本システムではプロセッサが処理するデータを全てネットワーク経由で供給できることになる．１６台のプロセッサがそれぞれ５００ＫＢ／秒程度のデータを転送した場合の転送遅延は１００μs程度となるが，これは１００命令程度の処理時間に相当する．オペレーティングシステムのプロセススイッチのオーバヘッドが数百命令を要することを考えると，ほぼシングルプロセッサ内のプロセス間通信遅延に匹敵する速度でプロセッサ間通信可能であるといえよう．この予測から本システムはマルチプロセッサシステムとして動作可能と考えられる．

また図９には本試作システムの構成で伝送路速度のみを４００Ｍｂｐｓとした場合の通信性能も示している．ＮＥＴ－ＶＭＳ方式は通信制御を単純化し，ハードウェア化しているため伝送路速度にほぼ比例した実効通信性能を得ることができる．従ってより高速な伝送路を用いることによってさらに高速なプロセッサによるマルチプロセッサシステムを実現可能である．

図１０にメモリページサイズを変化させた場合のデータ転送性能を示す．フレームヘッダ部やギャップ部の長さを固定としている．小さいページサイズではフレーム長が短いためフレーム転送に要する時間（通信遅延）は短かくなるが，同時にヘッダ部などに対するデータの割合が小さくなるため実効的な転送速度が悪化する．反対に大きなページサイズでは通信遅延は長くなるが，実効的な転送速度が向上する．評価の結果では１６台のプロセッサがそれぞれ５００ＫＢ／秒のデータを転送した場合，ページサイズが５１２Ｂの時平均転送遅延が最小となっている．

以上については試作システム完成後，実際に

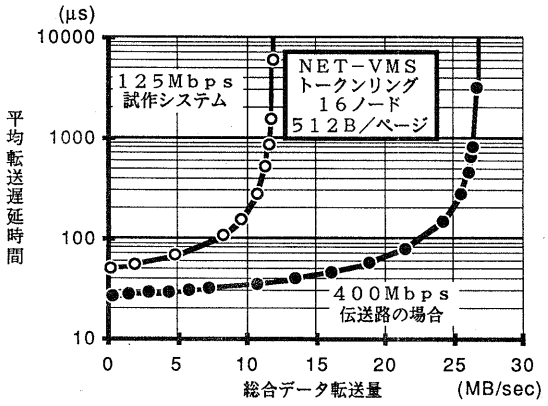

図９　予想通信性能－１

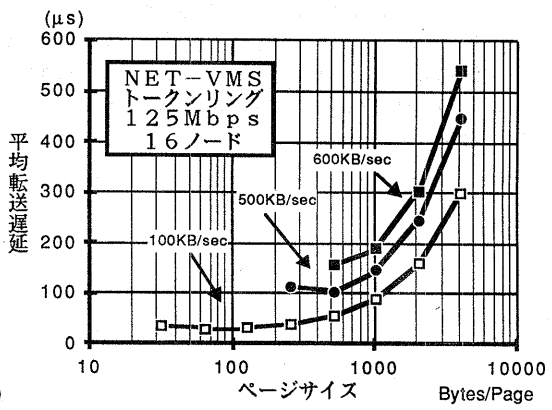

図１０　予想通信性能－２

アプリケーションを動作させて検証する予定である.

6. おわりに

ネットワーク結合による密結合型分散処理システムアーキテクチャとして開発を進めているネットワーク仮想記憶システム方式の実現性と性能の有効性を検証し，分散・並列処理ソフトウェアにおけるプロセス間通信の特性を測定することを目的とした試作システムについて説明した．完成後は宣言的プロセス間通信方式を始めとするプロセス間通信方式，分散・並列処理アプリケーションの実装及び評価を順次行っていく予定である．

参考文献

(1) C.G.Bell; Multis: A New Class of Multiprocessor Computers, Science, Vol.228, Apr. 1984

(2) 新たなコンピュータ世界を切り開くマルチマイクロ，相次ぎ登場，日経コンピュータ，1986.8.4号，Aug. 1986

(3) 陣崎，八星；ブロードキャストネットワークによる分散型単一階層仮想記憶システム 信学会研究会，CPSY 86-20, Jul. 1986

(4) 八星，陣崎；ネットワーク仮想記憶システム：NET-VMS (1)システムアーキテクチャ 情処33回全国大会，3T-8, Oct. 1986

(5) 陣崎，八星，樋口；ネットワーク仮想記憶システム：NET-VMS (2)プロセス間通信方式 情処33回全国大会，3T-9, Oct. 1986

(6) 陣崎，樋口，八星；ＬＡＮにおける通信制御高速化の検討，信学総全大，S25-5, Mar. 1987

(7) M.Kumer, G.F.Pfister; The Onset of Hot Spot Contention, ICPP'86, Aug. 1986

(8) P.Bitar, A.M.Despain; Multiprocessor Cache Synchronization Issues, Innovations, Evolution, ISCA'86, Jun. 1986