計算機アーキテクチャ 119－25
（1996. 8. 28）

光バスクラスタシステムの仕様と基本性能の評価

福井俊之[†,☆] 鈴木茂夫[†,☆] 中村秀一[†,☆] 下山朋彦[†,☆] 数藤義明[†]
濱口一正[†,☆] 柴山茂樹[†,☆]

本稿では，ワークステーションクラスタの分散共有メモリ機構を，光波長多重回線を用いてハードウェアでサポートした“光バスクラスタ”の第一次試作機“Euphoria”の仕様，及びEuphoriaを実際に稼動させて測定した基本性能の評価に関して述べる．Euphoriaはハードウェア・バスプロトコルを光回線により他ノードまで通信し，cache coherenceを保った分散共有メモリを実現している点でユニークである．現在のEuphoriaでは光バスアービタにおける回線設定及びキャッシュの一貫性保持動作をアービタ内でソフトウェアにより行っているが，このボトルネックを解消できた場合，ノード間距離100mのシステムでもレスポンスタイムが5μs以内で，自他ノードのメモリを区別なく自由に参照できるシステムの可能性を示すことができた．

Design and Preliminary Evaluation of an Optical Bus Computer Cluster

TOSHIYUKI FUKUI[†,☆] SHIGEO SUZUKI[†,☆] SHUUICHI NAKAMURA[†,☆] TOMOHIKO SHIMOYAMA[†,☆]
YOSHIAKI SUDOU[†] KAZUMASA HAMAGUCHI[†,☆] and SHIGEKI SHIBAYAMA[†,☆]

In this paper, we describe hardware design considerations and a preliminary performance evaluation of an Optical Bus Computer Cluster (OBCC). The OBCC is categorized in a class of workstation cluster having hardware-supported distributed shared memory. Furthermore, the OBCC employs optical wavelength-division multiplexing (WDM) technology to connect nodes (workstations) with high bandwidth. Euphoria consists of several nodes (workstations) and an arbiter which arbitrates optical bus requests and maintains cache coherency among nodes. Current implementation of the arbiter functions is done by software for experimentation purpose. By examining the basic analytic performance evaluation we conclude that a distributed shared memory system with one-hundred-meter nodes distance is possible with five microsecond range response times if we eliminate the software overhead associated with the current arbiter implementation.

1. はじめに

マイクロプロセッサの性能向上への努力は続けられているものの，現存する高性能チップをはるかに凌ぐ新世代のマイクロプロセッサを作ることはだんだん難しさを増してきている．そういう中で性能向上にかかるコストの増大を押さえる解法にいくつかのマルチプロセシング技術が存在する．単一のマシンの中でのマルチプロセッサ化は既に広く行なわれてきているが，一方，共有メモリを利用したワークステーションクラスタによる手法は，低コストであり，広帯域ネットワーク技術により最近可能になってきたものである．

ワークステーションクラスタではソフトウェアによるメモリ共有手法を採ることが多い．しかし，ソフトウェアが介在することにより，そのレイテンシは大きくなる．

本研究では，分散共有メモリプログラミングモデルを提供する上での通信のボトルネックを解消するために，ソフトウェアのオーバヘッドを極小化し，実行転送速度をハードウェアの速度に近付けることにより，WSクラスタ構成での共有分散型マルチプロセッサシステムの可能性を探ることを目的とした．

我々の提案する「光バスクラスタシステム」では，高速ネットワークの実現法として光ファイバを用い，更に光波長多重技術により複数ノード間での広帯域同時通信を可能にする．また，通信プロトコルも，ノードの内部バスハードウェアプロトコルを基本に設定し，通信オーバヘッドの低減を図った．これらにより，高速ネットワークによるノード間のメモリ共有を可能とし，より高性能なクラスタ型計算機を提供することを目的とした．

現在，光バスクラスタシステムの一次試作機“Euphoria[1),2),3)]”はノード間キャッシュ一貫性保持動作機構を実装したバージョンが稼動しており，テストプログラムを用いて評価中である．

本報告では，まず光バスクラスタシステムのコンセプト，及び一次試作機Euphoriaの仕様について述べた後，そのノード間におけるメモリアクセススピードなどの基礎データについて報告し，更にアプリケーションを動かした時の動作について述べる．

†キヤノン株式会社情報メディア研究所
Media Technology Laboratory, Canon Inc.
☆現在，キヤノン株式会社CyberMediaプロジェクト
CyberMedia Project, Canon Inc.

2. 光バスクラスタシステムコンセプト

光バスクラスタシステムのコンセプトは以下の3点である．

・各ノード（単体のワークステーションに相当する）は，個々のユーザに従来通りのユーザインタフェースを与える．
・広帯域，低オーバヘッドの相互転送路を利用して，システム全体でCPUやメモリなどの計算資源を共有し，負荷の分散をうまく図ることによって，全体として高性能かつ可用性の高いシステムを提供する．
・クラスタ内各ノードの主メモリは全てのノード上のCPUから同一の方法でアクセスできることをハードウェアで保障し，その上で動作するソフトウェアには新たに特別な機構を導入すること無くノード間のメモリ共有を実現できるようにする．

3. Euphoria の概要

3.1 Euphoria設計時における方針

我々は光バスクラスタシステムの第一次試作機として“Euphoria”を製作した．その設計時における方針は，主に以下の三点である．

・Euphoriaは光バスの有効性を検証することを目的とした設計方針をとる．
・さまざまな光バス調停手法／キャッシュコヒーレンシ維持プロトコルを試せるように，光バスの調停部はソフトウェア化することとする．
・波長多重部分はエミュレーションで実現する．(設計時点では波長多重デバイスの入手が困難であったため)

3.2 Euphoriaの構成

図1にEuphoriaのシステム構成を示す．以下，

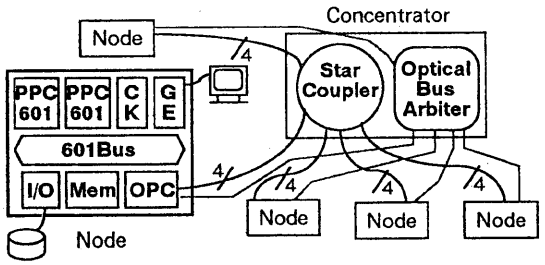


図1　Euphoriaシステム構成

Euphoria の各部の構成について簡単に説明する．

3.2.1 ノード構成

メインCPUにはIBM/ Motorolaの PowerPC601 (66MHz) を採用した．ノード内部の共有バスである601Bus は，PowerPC601のnative バスを拡張したものであり，split bus transaction をサポートし，33MHzで動作している．主記憶容量は128MByteである．グラフィック表示部はGE (Graphic Engine) が，入出力はIOC (I/O Channel) が担当する．OSにはCMUが開発したMach3.0を拡張したものを採用している[4]．

OPC (OPtical-Connection control)　OPC は光バスの制御を司る部分である．実際のデータの交換を行うデータ回線IF部とデータ回線の利用調停パケット及びノード間キャッシュの一貫性保持動作制御用パケットを光バスアービタに対して送るアービトレーション回線IF部の二つの部分からなる．光モジュールには1300nm帯のLED／PIN-PD モジュールを用いた．なお，光波長多重技術をエミュレートするために，データ回線には4波長分のモジュールを準備した．パラレル／シリアル信号変換器には，Fibre Channel用のIC (HOTLink) を利用した．符号化には8B/10B方式を採用し，光回線を流れるFrameの SD (Start Delimiter)，ED (End Delimiter) には8B/10B方式のコントロールコードを割り当てた．

CK (Coherency Keeper)　CKはノード間キャッシュの整合性を保障する．バスマスタから発行されたアクセスは，それが自ノード内アドレスであり，ノード間でキャッシュの一貫性保持動作等が必要とされる場合は，CKによる所定の動作の実施後許可される．また，CKは光バスアービタからの指示に従って，ノード内部にキャッシュ一貫性保持のためのバストランザクションを発行する．

3.2.2 光回線

光回線は，ノード間を相互に接続する波長多重データ回線，及び各ノード間のデータ回線利用の調停等を行うアービトレーション回線よりなる．データ回線では4波長を上り/下り回線の2波長ずつ2組に分け，全二重回線2本として利用する．データ回線のトポロジはスター型であり，Single Hop の Broadcast-and-Select方式を採用した．アービトレーション回線は光バスアービタとPoint-to-Point に接続される．光信号のデータ転送レートはいずれも200Mbps（搬送波250MHz）である．

3.2.3 コンセントレータ

コンセントレータはデータ回線の波長利用調停等を行うための光バスアービタ，及び各ノードからの光信号を再分配するためのパッシブスターカプラから構成される．

光バスアービタ　本研究部で過去に開発されたワークステーションのドータボードとして構成される．光バス利用調停処理，及びノード間キャッシュの一貫性保持処理は種々の方式が実験できるように，ボード上のMPU (XC68040) によりソフトウェアで実現される．ノード間キャッシュのディレクトリ情報をキャッシュするICC (Internode Caching-information Cache) として64MByteのDRAMを実装する．システムバスクロックは33MHzである．

3.3 メモリアーキテクチャ

メモリアーキテクチャとしては，Euphoria は NUMA (Non-Uniform Memory Access) 型メモリアーキテクチャを採る．メモリ資源をアクセスするシ

ステム中の全てのプロセッサは，**図2**に示すアドレスマップに従ってシステム中の任意のメモリ資源に対するアクセスを区別なく行うことができる．

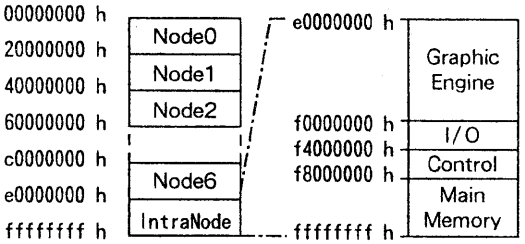


図2　Euphoria アドレスマップ

3.4 キャッシュ方式

PowerPC601のキャッシュはMESIプロトコルを採用している．キャッシュの一貫性保持はノード内部ではスヌープ方式により，ノード間ではディレクトリ方式により行う．

一貫性の保持単位はのキャッシュブロック単位である32Byteであり，各ノードには自ノード中のメモリデータブロック数に対応する4Mエントリのディレクトリがck内部に存在する．このディレクトリを管理するckとコンセントレータ部に存在する光バスアービタが一貫性保持に必要な動作を執り行う[2]．

4. ノード間メモリアクセス手順

ノード間メモリアクセスの手順の例を**図3**に示す．

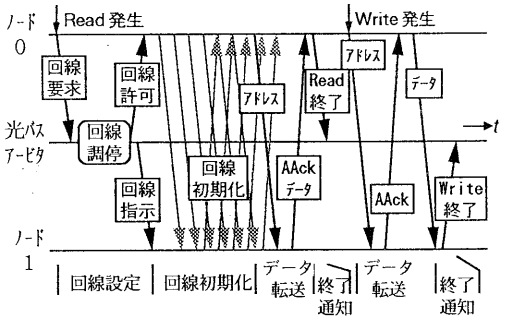
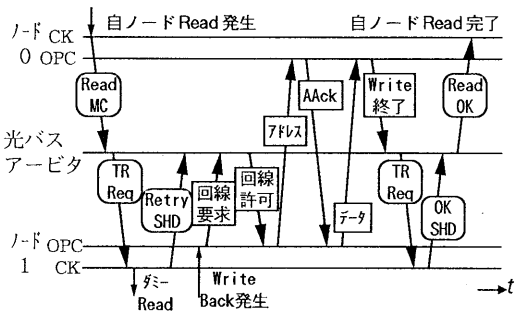


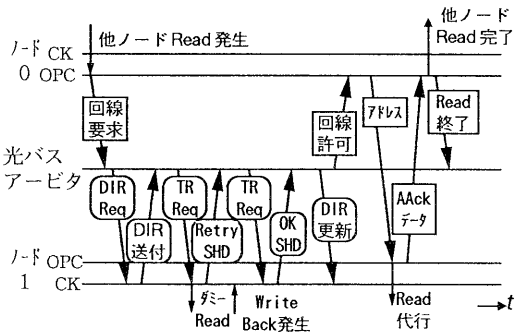
図3　ノード間メモリアクセス

図3は，ノード間のキャッシュ一貫性保持動作が必要ない場合で，かつデータ回線が確立していないときのパケットの流れを表わしている．図中四角で囲まれた文字がそれぞれパケットを表わしている．ノード0でノード1上のメモリへのReadが発生すると，まずノード0のOPCから光バスアービタに対して回線要求パケットが飛ぶ．光バスアービタは回線の調停をした後，対象ノードにデータ回線を設定するようにパケットを送る．それを受けて，まずノード0が回線初期化チェックパケットを繰り返し送出する．それを受信したノード1も同様に回線初期化チェックパケットを送出する．その後回線の確立を受けてノード0はアドレスパケットをノード1に送付する．ノード1では，OPCがメモリアクセスを代行した後，データパケットにAAck*をpiggybackしてノード0に送り返す．ノード0では処理が終了するとRead終了通知を光バスアービタに送付する．Writeの場合もほぼ同様である．

ノード間のキャッシュ一貫性保持動作が行われる際のパケットの流れの例（回線確立後）を**図4**に示す．



(a) 自ノードメモリアクセス（他ノードdirty保持）



(b) 他ノードメモリアクセス（他ノードdirty保持）

図4　一貫性保持動作の動作例

角が丸い四角は一貫性保持制御関連パケットを，角のある四角はメモリアクセス関連パケットを表す．

自ノード内メモリアクセスの場合，一貫性保持動作が必要であるとckが判断すると，そのバスアクセスはアドレス転送時（address tenure）でリトライされる．ckは次にOPCを通して光バスアービタに一貫性保持動作マルチキャスト要求パケットを送付する．光バスアービタはそのパケットに基き，該当キャッシュブロックを保持するノードに対して一貫性保持トランザクション発行要求を送付する．それを受けたノードのckはダミーバスアクセスを発行し，それを受けて必要に応じてキャッシュブロックのwritebackが実施される．一貫性保持動作の終了を確認した光バスアービタはその結果を要求したノードに送付する．そのパケットを受けとったckは，自ノード内のディレクトリのメンテナンスを行なった後，再びアクセスしてきたバスマスタのアクセスを許可し，その結果メモリが応答する．他ノードメモリアクセスの場合は，回線要

求パケットが一貫性保持要求パケットの代わりをする．なお，光バスアービタ上のICCにディレクトリがヒットしない場合は，ディレクトリの存在するノードからその中身を送付させた後，光バスアービタが一貫性保持動作の必要性を判断し一連の動作が行われる．

5. システム性能の測定と評価

5.1. 基本性能の測定

まずノード間のキャッシュの一貫性保持動作を必要としない場合（CK非動作時）の性能について述べる．以後本稿では特に断りのない限り全ての値は5回以上の実測に基く測定値の平均値であり，式はそれらの値より導出した実験式及びそれと理論値との組み合わせである．また転送タイプのsingleは1～8Byteのデータ転送を，burstは32Byteのバースト転送を意味する．なお，ノードと光バスアービタ間の距離はノード間距離の1／2とする．

5.1.1 ノード内部メモリアクセススピード

PowerPC601による自ノード内部の主記憶へのメモリアクセス実行時のレスポンスタイム，スループットを**表1**に示す．() 内はバスサイクル数を示す．

表1 自ノードメモリへのアクセス性能

	転送ﾀｲﾌﾟ	read	write
ﾚｽﾎﾟﾝｽﾀｲﾑ	single	394 ns(13)	333 ns(11)
	burst	606 ns(20)	424 ns(14)
ｽﾙｰﾌﾟｯﾄ	burst	48MByte/s	48MByte/s

writeが速いのはwriteバッファが利くからである．また，スループットがバースト転送時にread/write共に48MByte/sであるのは，EuphoriaでPowerPC601が主記憶に繰り返しアクセスする場合の平均アクセス間隔（アドレス転送が許可されるサイクル）が22サイクル（リトライ2回を含む）であるためである．

5.1.2 光回線の基本的遅延要素

光回線にHOTLinkをパラレル／シリアル変換器に用いた関係上，発生する遅延の主な遅延構成要素とそこでの遅延値（単位ns）を**図5**に示す．これはアービトレーション回線での例であるが，データ回線でもスターカプラによる遅延が加わる以外は変わりはない．

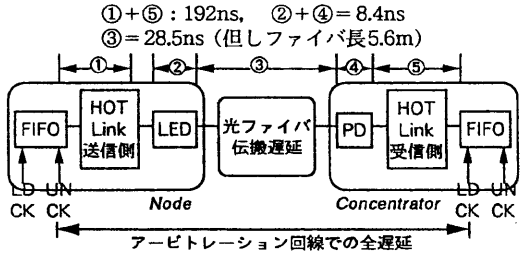


図5 アービトレーション回線上の主要遅延構成要素

図5で示した主な遅延構成要素に基盤上の配線遅延等を加えた実際の遅延の実験式は以下のようになる．

アービトレーション回線上でのパケット伝搬遅延

Darb[ns]= 207.5 + 5.085d (d[m]:Fiber長)

データ回線上の遅延は以下のとおりである．

Ddat[ns]= 235.1 + 5.085d (d[m]:Fiber長)

5.1.3 ノード間メモリアクセススピード

実際の処理の流れの上で，必要となる時間を測定しそこからEuphoriaにおける遅延構成要素の計算式及びソフトウェア処理時間等の平均値を導出した．それらの各フェイズにおける遅延要因別構成を**図6**に，処理時間の流れの例として回線確立後のノード間バーストwrite実行時の時間経過図を**図7**に示す．

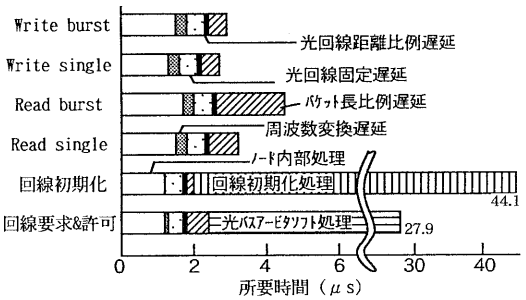


図6 各フェイズにおける遅延要因の構成

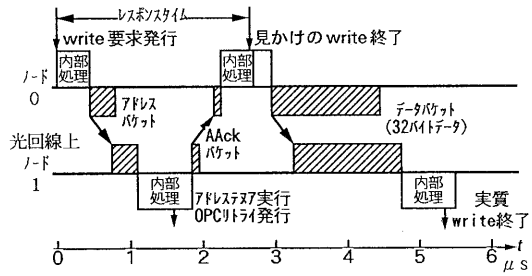


図7 バーストwrite実行時の時間経過図

なお、ここでの各パラメータは以下のとおりである．

ノード間距離：11.2m,

光バスアービタ上での波長利用調停に要する時間 = 25,506ns,

回線セットアップ時間（PLLのロックなどにかかる時間）= 44,143ns

レスポンスタイム及びスループットを**表2**に示す．

表2 他ノードメモリへのアクセス性能

	転送ﾀｲﾌﾟ	read	write
ﾚｽﾎﾟﾝｽﾀｲﾑ	single	3,133 ns	2,583 ns
	burst	4,516 ns	2,764 ns
ｽﾙｰﾌﾟｯﾄ	burst	10.7MByte/s	6.97MByte/s

表2より，Euphoriaでのバースト時の転送スループ

ットは，Readを利用して転送するほうがwriteを利用する場合より，約1.5倍の性能が出ることになる．

5.2 CK動作時のメモリアクセススピード

5.2.1 ノード内完結アクセス時の性能

一貫性保持動作実行時で，その保持動作がノード内部で完結するときの主記憶へアクセス実行時のメモリのレスポンスタイムとスループットを**表3**に示す．

表3 CK動作時の自ノードメモリアクセス性能

	転送タイプ	read	write
レスポンスタイム	single	758ns (25)	727ns (24)
	burst	848ns (28)	455ns (15)
スループット	burst	55.6MByte/s	42.2MByte/s

メモリのレスポンスタイムがCKの非動作時より大幅に伸びているが，これはアドレス転送が許可されるまでにかかる時間がディレクトリのチェックのため，13サイクル余分にかかっているからである．バースト時のwriteのレスポンスタイムが短いのは，キャッシュのwritebackが起こりうるときはExclusive状態であり，CKが動作しないからである．

readのスループットが表1のCK非動作時に比べて上がるのは，CK動作時にはアドレス転送のサイクル周期とCKの動作タイミングがほぼ同期しており，無駄なウェイトが入らないからである．

5.2.2 CKパケットの交換に伴うオーバヘッド

CKパケットはOPC及びアービトレーション回線を通じて光バスアービタに送られる．その様子は既に図4で一例を示している．現状では光バスアービタ上でのパケットの解釈やノード間のキャッシュメモリの一貫性保持処理のためのパケット生成等の全てをソフトウェアで行なっているために，パケットの行き来に応じてソフトウェア処理の時間がかかることになる．

CK関連動作の遅延要素別のグラフを**図8**に，メモリをバーストreadしたときのレスポンスタイムを構成する各動作フェイズの時間を**図9**に示す．

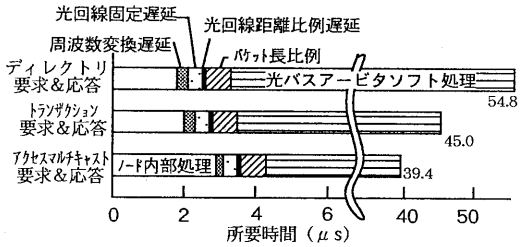


図8 CK動作時の遅延要因

図8から現在の遅延は光バスアービタ上のソフトウェアによる一貫性保持動作処理に起因すること，及びそれらを除くとノード内部のアクセス処理時間が大きく，本システムの距離ではその距離比例部の遅延要因は全体に比べて無視し得ることがわかる．

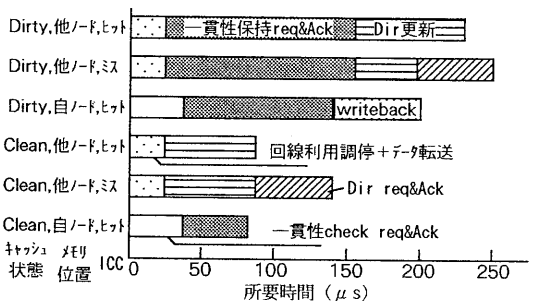


図9 各フェイズにおける遅延要因の構成

図9からはキャッシュがCleanでICCがヒットした場合（最良値）と，キャッシュがDirtyでICCがヒットしなかった場合（最悪値）とでは約3倍のレスポンスタイムの開きがあることがわかる．また，現在のシステムでは自ノードのメモリをアクセスした場合でそのブロックを他のノードのCPUがDirtyで保持していた場合のwritebackのオーバヘッドがかなり大きいことがわかる．

5.3 実際のソフトウェア動作時の性能

実際にEuphoriaの上でソフトウェアを実行させたときの性能を示す．一貫性保持動作実行時に，4KByteのメモリコピーを連続して1000回行ったときに要した時間を**表4**に示す．

表4 一貫性保持動作実施時の4KByteメモリコピーを1000回実施したときの所要時間

コピー元＼コピー先	local	remote
local	0.280s	28.2s
remote	20.1s	62.5s

local→remoteよりremote→localの方が性能が良いのは，PowerPC601がライトアロケイト方式を採用しているため，writeを実行するためにはその実行前に対象ブロックをCPUのキャッシュに読み込まねばならず，そのオーバヘッドがノード間アクセスとして大きいからである．remote→remoteの時間が他の倍以上かかっているのは，主にCPU内部のバスアクセスの優先度変更によるノード外アクセスの順序変更に伴うオーバヘッド処理のためである．

現在Euphoriaには並列クィックソートなどのプログラムも実装して評価している最中である．2ノードでそれぞれCPUを動かした場合，現在のシステムでは1CPUでのローカル実行時より性能が低下する現象などが観測されている．本システムを活かすためのスケジューリング手法等に関しては現在検討中である．

6. 考察

6.1 現在のシステムのボトルネック

図6，図8から，光バスアービタ上での回線の調停及び一貫性保持動作にかかるソフトウェア処理の時間，

及び回線繋ぎ替えに伴う回線初期化の時間がノード間のメモリアクセス遅延の大部分を占めていることがわかる．また，それらを除いた部分の，回線確立後の定常状態におけるデータ転送に要する時間では，システム内部での処理にかかわるオーバヘッドが多くの場合に律速の原因になっている．そこで，まず，これらのオーバヘッドの削減がどこまで可能かを考え，その上で回線速度，及びノード間距離等のパラメータを振った場合の影響を考える．

6.2 処理プログラムの高速化

現在XC68040を用いて行なっているソフトウェア処理をハードウェア化すれば，各処理が数100ns程度で完了するくらいにまでは高速化できると考えられる．

6.3 システム内部のオーバヘッドの除去

システム内部でのオーバヘッドとしてはシステムで使われているクロックが2系統（25MHzと33MHz）存在することの影響がある．まずそれを除去し，次いで現在のシステムアーキテクチャを維持したままで最適化することを考える．システムのバスサイクルがC(ns)で統一されたとして，光バスクラスタシステムにおける最適化された基本性能実験式を求めると，以下のようになる．（バースト転送時一貫性保持動作非実行時のものを示す。）

レスポンスタイム（ns）：

read: $Trb=28C+2(235.1+5.085d)+2F+V(P_{addr}+P_{datab})$

write: $Twb=21C+2(235.1+5.085d)+2F+VP_{addr}$

スループット（MByte／s）：

read: $Thr=32/(27C+2(235.1+5.085d)+2F+VP_{addr})$

write: $Thw=32/(28C+2(235.1+5.085d)+2F+V(P_{addr}+P_{datab}))$

但し，P：Packet長（Byte）$P_{addr}=9$, $P_{datab}=38$,

V：光回線で1Byte分のデータパケットを伝送するのに要する遅延（$=10^4/f$（f：搬送波周波数））

F：ノード間のクロックの位相差（≒C／2）

6.4 回線の伝搬速度が速くなった場合の性能

図10に最適化されたシステムの下で，回線速度を変化させた場合（ノード間距離10m）のスループットとレスポンスタイムを示す．

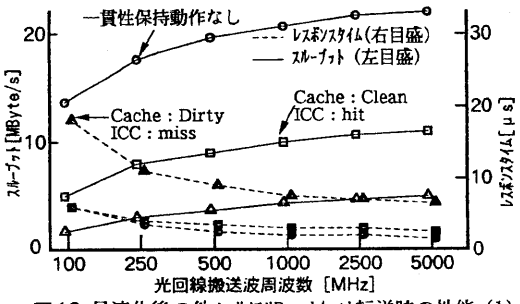


図10 最適化後の他ノードメモリReadバースト転送時の性能（1）

キャッシュがClean状態でのレスポンスタイムは，回線速度が現状（250MHz）でもICCヒット時には5μs以下，1GHzになればICCをミスしても5μs以下に押えられることがわかる．

6.5 回線の距離が長くなった時の伝搬遅延の影響

図11に最適化されたシステムの下でノード間距離を変化させた場合（搬送波周波数250MHz）のスループットとレスポンスタイムを示す．

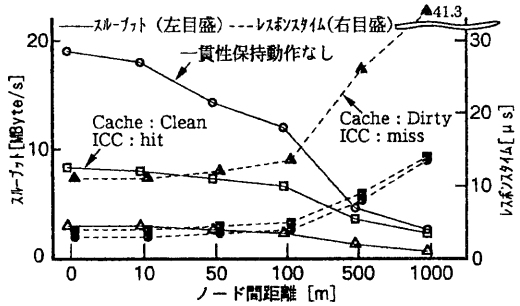


図11 最適化後の他ノードメモリReadバースト転送時の性能（2）

現在と同じ回線速度であってもノード間距離100m以内であればキャッシュClean状態でのレスポンスタイムがICCヒット時で5μs以下のシステムを提供できることがわかる．

7. おわりに

本稿では，ワークステーションクラスタの分散共有メモリ機構を，光波長多重回線を用いてハードウェアでサポートした“光バスクラスタシステム”の第一次試作機”Euphoria”の仕様，及びその基本性能の評価に関して述べた．

現在のEuphoriaでは特に光バスアービタにおける回線設定及びキャッシュの一貫性保持動作ソフトウェアの処理等がボトルネックとなる．しかし，それらのボトルネックを解消できた場合，ノード間距離100mのシステムでもレスポンスタイムが5μs以内で，自他ノードのメモリを区別なく自由に参照できるシステムの可能性を示すことができた．

参考文献

1) 福井俊之，濱口一正，下山朋彦，小杉直人，柴山茂樹：光バスクラスタ計算機Euphoriaの開発（1）概要，第49回情報処理学会全国大会5K-05(1994).

2) 下山朋彦，濱口一正，福井俊之，柴山茂樹：光バスクラスタ計算機Euphoriaの開発(2)メモリアクセス機構，第49回情報処理学会全国大会5K-06(1994).

3) S. Shibayama, K. Hamaguchi, T. Fukui, Y. Sudo, T. Shimoyama and S. Nakamura, "An Optical Bus Computer Cluster with a Deferred Cache Coherence Protocol." In Proceedings of the 1996 ICPADS, pp. 175-182, Tokyo, June (1996).

4) 鈴木茂夫，福井俊之，数藤義明，柴山茂樹：光バスクラスタシステム対応の分散タスク／スレッド機構の実現，情処研報ARC-119（1996).