社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

リングネットワークにおけるλコンピューティング環境に適した共有メモリアーキテクチャの設計と評価

久保　貴司† 　谷口　英二†† 　馬場　健一††† 　村田　正幸†

† 大阪大学大学院情報科学研究科 〒 565-0871 吹田市山田丘 1-5
†† （株）日本ビジネスデータープロセッシングセンター 兵庫県姫路市
††† 大阪大学サイバーメディアセンター 〒 567-0047 茨木市美穂ヶ丘 5-1
E-mail: †{t-kubo,murata}@ist.osaka-u.ac.jp, †††baba@cmc.osaka-u.ac.jp

あらまし　我々は，各ノード計算機間に光ファイバを直結して各ノード計算機上に存在する共有メモリを波長パスで結ぶことにより，高速計算を可能とするλコンピューティング環境を提案している．本稿では，λコンピューティング環境において波長数やハードウェアの制約を考慮した実現可能なリングネットワークにおける共有メモリアーキテクチャを提案し，その性能を評価する．具体的には，リングネットワークにおいて複数波長を用いてデータ転送やキャッシュ一貫性制御を行う方式の設計を行い，制御にかかる遅延時間を求め，セミ・マルコフ過程を用いて解析を行った．その結果，共有メモリアクセス頻度が大きい場合など性能向上が得られるパラメータ領域が存在することを明らかにした．

キーワード　λコンピューティング環境，共有メモリアーキテクチャ，キャッシュ一貫性制御，セミ・マルコフ過程

Design and Evaluation of a Shared Memory Architecture for λ Computing Environment in Ring Networks

Takashi KUBO†, Eiji TANIGUCHI††, Ken-ichi BABA†††, and Masayuki MURATA†

† Graduate School of Information Science and Technology, Osaka University
1-5 Yamadaoka, Suita, Osaka 565-0871, Japan
†† Nihon Business Data Processing Center Co.,Ltd
††† Cybermedia Center, Osaka University　5-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan
E-mail: †{t-kubo,murata}@ist.osaka-u.ac.jp, †††baba@cmc.osaka-u.ac.jp

Abstract　We proposed the λ computing environment, in which computing nodes with shared memories are connected by optical fibers, and their shared memories are also connected by wavelength paths directory, so that we can make a distributed computing. In this report, we model and analyze the shared memory architecture in the ring network considering the restriction of the number of wavelength and hardware in λ computing environment. Actually we design the control method for cache coherency by using multiple wavelength on the ring network, and calculate the control delay time. Then we analyze the model with semi-Markov process and evaluate the shared memory architecture. As a result, we showed that the performance improvement was achieved when the shared memory access frequency was large.

Key words　λ computing environment, Shared memory architecture, Cache coherency, Semi-Markov process

1. ま え が き

一台の計算機では実行できないような大規模計算を複数の計算機資源を利用して実行するグリッドコンピューティング技術に関する研究開発がさかんに進められている．グリッドコンピューティング環境では一般に現在のインターネットで用いられている TCP/IP を計算機間でのデータ交換に利用するが，TCP ではパケット処理などの通信に要するオーバヘッドが大きいため十分な性能を得るのは難しいと考えられる．そこで，ネットワーク上のノードや計算機群を光ファイバで接続し，エンドホスト間に光波長パスを設定することにより高速かつ高品質な通信パイプを提供することを考える．その上で，この波長パスを

専用の通信チャネルとして利用し，分散計算を行う新たなアーキテクチャとしてλコンピューティング環境を提案している．

関連研究[1]〜[3]では，λコンピューティング環境を実現するひとつの手法として，NTTフォトニクス研究所が開発した情報共有システム（AWG-STARシステム）[4]を利用している．AWG-STARシステムでは，各ノード計算機が共有メモリボードを有し，共有メモリボード間をAWGルータを介して波長パスで接続した上で共有すべきデータを転送し，すべてのノード計算機間で同一のデータを共有する．これらの研究では，AWG-STARシステムを用いたλコンピューティング環境において，MPIやOpenMPの設計と実装を行い，アプリケーションを用いてその性能を評価している．しかしながら，AWG-STARシステムによる共有メモリアーキテクチャにおいては，並列計算アプリケーションの実行時間に基づいて評価を行っているため，ネットワーク特性やキャッシュプロトコルの違いによる性能への影響などについては十分な評価ができていない．

λコンピューティング環境における共有メモリでは，ノード計算機が広域に配置されているため，マルチプロセッサシステムやクラスタにおける分散共有メモリなどに比べ，ネットワーク特性が共有メモリシステムを用いた計算環境の性能に大きな影響を与えるものと考えられる．このため，λコンピューティング環境に適した共有メモリアーキテクチャを設計するには，ネットワークが共有メモリの性能に与える影響を十分考慮しなければならない．また，現在の計算機アーキテクチャではCPUの計算性能にプロセッサのキャッシュメモリが大きな影響を与える．このキャッシュメモリの制御方式と共有メモリシステムの相互作用についても検討する必要がある．

そこで，本稿では，λコンピューティング環境における共有メモリアーキテクチャのモデル化を行い，ネットワークやキャッシュ一貫性制御のための処理がシステムの性能にどのような影響を与えるかを解明する．まず，ネットワークモデルとしてリングネットワークを対象とし，波長数やハードウェアの制約を考慮する．モデル化には状態の滞在時間が任意に設定できるセミ・マルコフ過程[5]を利用する．さらに定常状態確率から性能解析を行うことにより，どの共有メモリアーキテクチャの構成がλコンピューティング環境に適しているかについて検討していく．

2. λコンピューティング環境における共有メモリアーキテクチャ

2.1 λコンピューティング環境

λコンピューティング環境においては，WDM技術を利用して各ノード計算機，光スイッチを光ファイバで接続し，ノード計算機間に波長パスを設定する．このノード計算機間に設定した波長パスを利用して分散並列計算を行う．あらかじめ設定した波長パスを専用の通信チャンネルとして利用することにより，分散並列計算のデータ交換において，高速かつ高信頼な通信が実現できる．

2.2 共有メモリアーキテクチャの特性要因

λコンピューティング環境における共有メモリアーキテクチャはネットワークトポロジやメモリアクセスモデル，キャッシュプロトコルなどにより性能を大きく左右される可能性があり，どのような共有メモリアーキテクチャが適しているのかは，一概に決定することはできない．そこで，本節では共有メモリアーキテクチャの特徴を決定づける要因であるトポロジ，メモリアクセスモデル，キャッシュプロトコル，キャッシュと共有メモリ間のデータ一貫性プロトコルについて説明し，本稿において対象とする共有メモリアーキテクチャついて述べる．

a) トポロジ

トポロジとしてはリングトポロジとメッシュトポロジが考えられる．リングトポロジはブロードキャストが容易となるトポロジであるが，伝搬遅延として最低でもリング1周分を要することになる．一方，メッシュトポロジはリングトポロジに比べると伝搬遅延は短くなるが，ブロードキャストのために各ノードで送受信されるデータの複製が必要となる．

b) メモリアクセスモデル

メモリアクセスモデルとしてはUMA (Uniform Memory Access) モデル，NUMA (Non-Uniform Memory Access) モデル，NORMA (NO Remote Memory Access) モデルの3つが考えられる．UMAモデルはすべてのプロセッサがアドレス空間を共有し，同一時間でアクセス可能なモデルまたはそのようなメモリを持つ計算機である．NUMAモデルはすべてのプロセッサが，アドレス空間を共有するメモリを持つが，あるプロセッサから見た時のアクセス速度は，メモリの番地によって異なるモデルまたはそのようなメモリを持つ計算機である．NORMAモデルは各プロセッサは互いに独立したアドレス空間のメモリを持ち，メッセージのやりとりによって計算を進めていく，つまり共有メモリをもたないモデルまたは計算機である．

c) キャッシュプロトコル

キャッシュプロトコルとしてはスヌープ法とディレクトリ法が考えられる．スヌープ法はマルチプロセッサシステムで多用されており，各キャッシュが共有メディアを監視し，データや制御信号の発生などシステムの挙動を把握することにより，キャッシュ一貫性保持プロトコルを実現している．ディレクトリ法はNUMA型システムで使用されており，どのキャッシュがどのキャッシュラインを保持しているかという情報を各ノードのディレクトリに保持し，必要が生じた時にそれを検索することにより直接相手のキャッシュの無効化などの処理を行う．

d) キャッシュと共有メモリ間のデータ一貫性プロトコル

キャッシュと共有メモリ間のデータ一貫性プロトコルは，一致させるタイミング（ライトスルー，ライトバック）と方法（無効化，更新）の組み合わせによって，無効化型ライトスルー，無効化型ライトバック，更新型ライトスルー，更新型ライトバックの4種類が考えられる．

ここでは，本稿で用いる無効化型ライトバックの状態と状態遷移を図1を示す．無効化型ライトバックには3つの状態があり，それぞれキャッシュと共有メモリとの間でデータが一致しているClean状態（以下，*C*状態），キャッシュと共有メモリとの間でデータが一致していないDirty状態（以下，*D*状態），キャッシュが無効化されているInvalid状態（以下，*I*状態）で

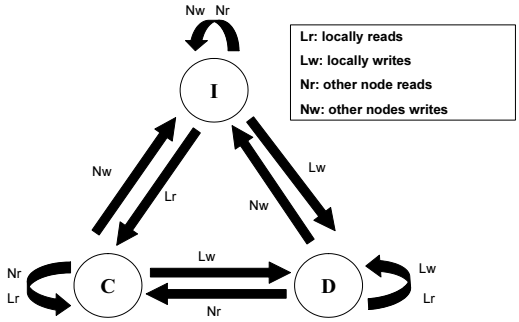


図 1　無効化型ライトバックの状態遷移図

表 1　共有メモリアーキテクチャの構成

トポロジ	メモリアクセスモデル	キャッシュプロトコル
リングトポロジ	UMA モデル	スヌープ法
リングトポロジ	NUMA モデル	スヌープ法

ある．

2.3　対象とする共有メモリアーキテクチャ

共有メモリアーキテクチャにおいて特に重要な機構は各プロセッサ間のキャッシュの一貫性と，キャッシュと共有メモリ間のデータの一致である．また，並列計算アプリケーションの観点ではプロセス間の同期も重要な点である．このようなデータの一貫性や同期の際にはブロードキャストが多用される．そこでまず，ブロードキャストが容易なリングトポロジを構成する共有メモリアーキテクチャについての解析を行う．この場合，リングトポロジを構成しているため，制御トークン用の波長を用意し，これを監視することにより従来のスヌープキャッシュプロトコルを自然に拡張することができる．また，キャッシュと共有メモリ間のデータ一貫性プロトコルは無効化型ライトバックを採用する．これは，プロトコル処理の際にネットワークに流れるデータ量が他に比べると少なく，プロトコルの処理もシンプルであることに加え，ノード計算機間の遅延時間が大きい場合に有効であると考えられる．リングトポロジを用いる場合，ネットワークにデータを送信することにより全てのノードにデータを伝えることができる．そのため，メモリアクセスモデルとしては UMA モデル（図 2）と NUMA モデルが考えられる．UMA モデルと NUMA モデルではネットワーク利用率が異なるため共有メモリアーキテクチャの性能が異なる可能性があり，両方のモデルについて解析を行い，性能を評価する（表 1）．

これらのアーキテクチャは，リングトポロジであるために，伝搬遅延として最低でもリング 1 周分の時間を要する．そのため遅延時間による影響を抑えるために複数波長を用い，ノードでの処理遅延時間を削減する方法についても検討し，性能評価を行う．具体的には波長数が 1 の単一波長リングトポロジ，波長数が 2 の複数波長のリングトポロジについて性能を評価する．

3.　複数波長を用いた光リングネットワークの構成

本稿では，対象とするリングネットワークを単一方向リングとする．前節で述べたように，リングトポロジを採っているために，全てのノードを経由することにより多くの処理遅延を要する．そこで，複数波長を用いてノード間に波長パスを設定し，ノードの処理遅延を減らすことにより処理遅延を抑える方法を提案する．複数波長を用いたネットワーク構成としては階層化したリング構造を有する Hierachical ring [6], [7] を採用する．このネットワークはノード数によらず少ない任意の波長数で実現することができ，また，ノードからネットワークへの接続に必要となる波長インターフェース数が少なく，ネットワークの構築コストを抑えることができる．

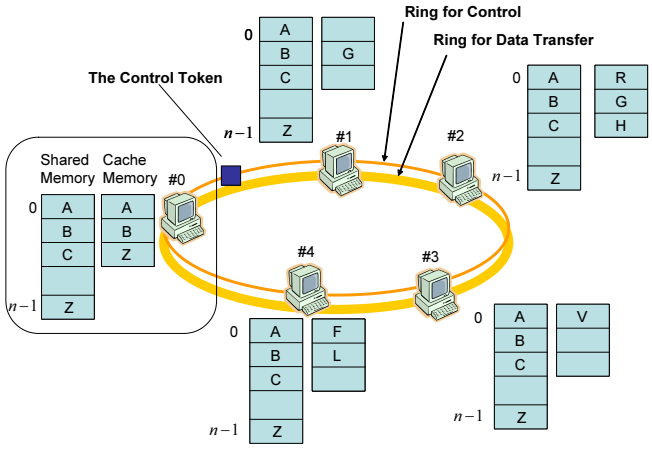


図 2　Ring-UMA モデル

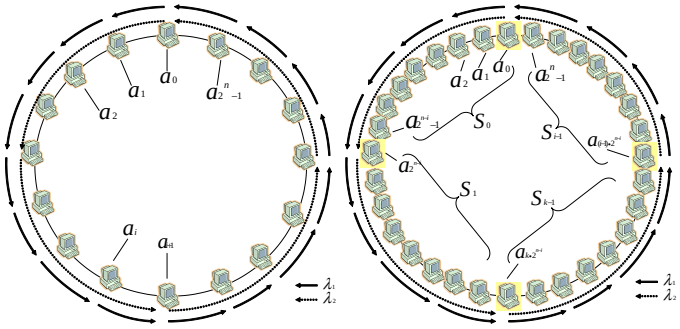
3.1　光リングネットワーク構成と記号の定義

ここでは，簡単のためにリングトポロジを形成するノード数 N を 2^n とする．これらのノードをリング状に配置し，各ノードに対して，ノード番号を 0 から順に反時計回りにつけ，$a_0, a_1, \cdots, a_{2^n-1}$ と表す．この状態のリングトポロジに，波長 λ_2 を分割数 $I = 2^i$，$I = 2^{i+1}$ で全ノードを均等に分割できるように設定する（図 3(a)）．

複数の波長を利用できるノードを複数波長ノードと呼び，複数波長ノードから半時計回りに次の複数波長ノード手前までのノードを 1 セットとして扱い，$\mathcal{S}_k$ $(k = 0, 1, \cdots I-1)$ と表す．ここで，複数波長ノードを a_{w_k} と定義する（図 3(b)）．すなわち，ノード数 $N = 2^n$，分割数 $I = 2^i$ とすると，a_{w_k} となるノードはノード番号が $a_0, a_{2^{n-i}}, a_{2\cdot 2^{n-i}}, a_{3\cdot 2^{n-i}}, \cdots, a_{(i-1)\cdot 2^{n-i}}$ のノードとなる．つまり $w_k = k \cdot 2^{n-i}, 0 \leqq k \leqq i-1$ と表すことができる．次に，ノード a_{w_k} から反時計回りに隣のノード $a_{w_{k+1}}$ の手前までを 1 セットとして区切る．このとき各セットに属するノード数は 2^{n-i} となる（図 3(b)）．

ノード間の転送遅延時間を求めるため，送信ノードを a_s とし，目的ノードを a_d $(0 \leqq s \leqq 2^n-1, s \neq d)$ とする．各セットはそれぞれ対称であるので送信ノードが属するセットを $\mathcal{S}_0$ に固定する．そのため，セット内のノード数は 2^{n-i} となり，すなわち $0 \leqq s \leqq 2^{n-i}-1$ である．

遅延時間の要素には，データが光リングを伝わるのに要する伝播遅延時間，ノードにおける処理遅延時間，波長変換に要する遅延時間がある．そこで，(a_s, a_d) 間における通信の遅延時間は，各遅延時間に対する定数 T_{TD}, T_{PD}, T_{CD} と係数



(a) 2 波長のノード構成図　　(b) セットへの分割

図 3　ノードの構成図

$F_{TD}(a_s, a_d)$, $F_{PD}(a_s, a_d)$, $F_{CD}(a_s, a_d)$ とした場合，式 (1) と表すことができる．

$$T_{Delay}(a_s, a_d) = F_{TD}(a_s, a_d) \cdot T_{TD} + F_{PD}(a_s, a_d) \cdot T_{PD} + F_{CD}(a_s, a_d) \cdot T_{CD} \quad (1)$$

3.2　Point-to-Point による平均遅延時間

キャッシュの状態の回答や，キャッシュの書き戻しは Point-to-Point の通信を利用する．Point-to-Point 通信における平均遅延時間は，式 (1) を用いて，式 (2) のように表すことができる．

$$T_{AveDelayPP} = \frac{1}{2^{n-i}} \sum_{s=0} \frac{\sum_{d=0} T_{Delay}(a_s, a_d)}{2^n} \quad (2)$$

式 (1) における各係数は使用する各波長のホップ数を計算することにより求めることができる．また，ある送信ノード a_s からあるセット $\mathcal{S}_k$ にある全てのノードへ通信する遅延時間を $T_{Delay}(a_w, \mathcal{S}_0)$ とすると，全てのノード間の通信における遅延時間の総和は式 (3) となる．

$$T_{Delay} = \frac{1}{2^n}\Big(T_{Delay}(a_w, \mathcal{S}_0) + \sum_{k=1}^{2^i-1} T_{Delay}(a_w, \mathcal{S}_k) + \sum_{s=1}^{2^{n-i}-1} (T_{Delay}(a_s, \mathcal{S}_0) + \sum_{k=1}^{2^i-1} T_{Delay}(a_s, \mathcal{S}_k))\Big) \quad (3)$$

したがって，平均遅延時間 $T_{AveDelayPP}$ は式 (4) となる．

$$T_{AveDelayPP} = \frac{T_{Delay}}{2^{n-i}} \quad (4)$$

3.3　ブロードキャストによる平均遅延時間

キャッシュ状態の問い合わせやキャッシュ一貫性制御はブロードキャストで行う．ブロードキャストを行う場合，送信ノードは各 $\mathcal{S}_k$ 宛てにメッセージを複製し，送信する必要がある．また，全てのノードに送信したメッセージが届いたことを確認する必要がある．そのため，送信した全てのメッセージが再び自ノードに帰ってくるのを待たなければならない．よって，ブロードキャストに要する時間は最後の送信メッセージが戻ってくるまでの時間となる．送信ノードが複数波長ノードとそれ以外のノードで遅延時間が異なるためそれぞれについて求める．送信ノードが a_{w_0} の場合には式 (5)，送信ノードが a_{w_0} 以外の場合には式 (6) となる．

$$T(\mathcal{S}_0) = (2^{n-i}) \cdot T_{\lambda_1} + (2^i - 1) \cdot T_{\lambda_2} + 2 \cdot T_{CD} \quad (5)$$

$$T(\mathcal{S}_0) = (2 \cdot 2^{n-i}) \cdot T_{\lambda_1} + (2^i - 2) \cdot T_{\lambda_2} + 4 \cdot T_{CD} \quad (6)$$

表 2　モデルで用いるパラメータ

キャッシュヒット率	h
メインメモリアクセスの読み込み割合	r
メインメモリアクセスの書き込み割合	w
メインメモリアクセスの共有メモリへのアクセス割合	s

表 3　モデルで用いる変数

D 状態のキャッシュラインを持つ確率	P_D
あるノードだけが D 状態を持つ確率	P_d
C 状態のキャッシュラインを持つ確率	P_C
あるノードが C 状態を持つ確率	P_c
キャッシュメモリの空きが全くない確率	P_x
キャッシュラインが無効化される確率	P_{inv}

4.　共有メモリアーキテクチャのモデル化と評価

4.1　セミ・マルコフ過程

本稿では，セミ・マルコフ過程 [5] によって共有メモリアーキテクチャをモデル化する．セミ・マルコフ過程では，それぞれの状態に任意の滞在時間を設定することができる．したがって，セミ・マルコフ過程を用いることにより，キャッシュ一貫性制御などの複雑な要求が起こる共有メモリアーキテクチャのモデル化が容易になる．

確率過程を $\{X(t), t \geqq 0\}$ とし，有限状態を持っているとすると，定常状態確率を求める具体的なアルゴリズムは以下の通りである．

(1) 離散時間型マルコフ連鎖のための定常状態確率を π とし，状態遷移行列 $p = (p_{i,j})$ を用いてこの定常状態確率を計算する

(2) セミ・マルコフ過程の全ての状態 $\{i\}$ について，平均滞在時間 η_i を計算する．

(3) CPU の各滞在時間を求めることにより，セミ・マルコフ過程の定常状態確率を計算する．

$$P_i = \frac{\pi_i \eta_i}{\sum_j \pi_j \eta_j} \quad (7)$$

また，脱出確率 $\psi_i = P_i/\eta_i$ となる．

4.2　モデルで用いる変数の定義

モデルに用いるパラメータ，変数をそれぞれ表 2，表 3 に示す．変数の値はそれぞれ $P_D = hws$, $P_d = (N-1)hws(1-hws)^{N-1}$, $P_C = hrs$, $P_c = 1-(1-hrs)^{N-1}$, $P_x = (1-P_{inv})^{N-1}$, $P_{inv} = hP_cw + (1-h)(1-P_d)w$ となる．

4.3　共有メモリのモデル化

λ コンピューティング環境における共有メモリアーキテクチャをセミ・マルコフ過程を用いてモデル化する．各アーキテクチャの振る舞いにしたがって，各ノードの CPU の観点から状態遷

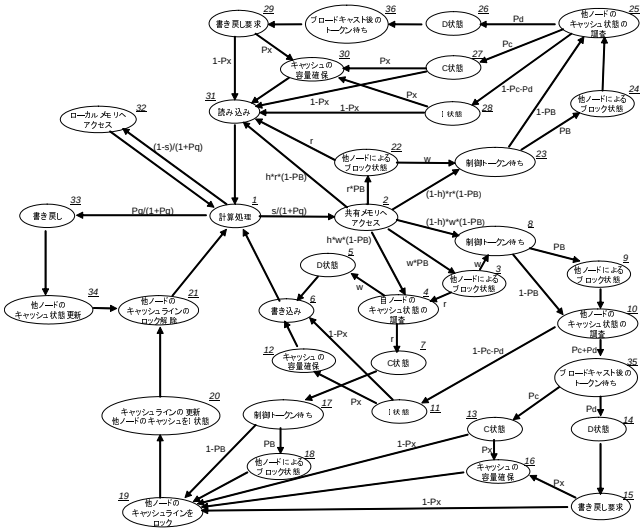


図 4　2 波長を用いるリング UMA アーキテクチャの状態遷移図

移図を作成する．2 波長を用いるリング UMA アーキテクチャにおける状態遷移図を図 4 に示す．リング NUMA アーキテクチャについても同様に表現できるが，紙面の都合上割愛する．

状態 1 を CPU における計算状態とし，メモリアクセスを必要としない計算処理を行うものとする．次に CPU がローカルメモリにアクセスする場合を考え，読み込みアクセス状態 (4〜18, 35) と書き込みアクセス状態 (22〜31, 36) に分ける．

また，自ノード以外のノードが，あるキャッシュラインにアクセスしている間，自ノードから同じ共有メモリに対するアクセスはブロックされる．この確率を P_B とする．CPU が共有キャッシュにアクセスしている状態の集合を S_a とする．キャッシュラインへのアクセスがブロックされるのは，状態集合 S_a において確率 s で起こる．各キャッシュラインへアクセスされる確率を α_1 とすると，$\alpha_1 = \frac{l}{C}\sum_{s\in S_a} P_s$，$P_B = 1-(1-\alpha_1)^{N-1}$ となる．

自ノード以外のノードから書き戻し要求メッセージが送られた場合，CPU は共有メモリに該当キャッシュを書き戻さなければならない．書き戻し要求の起こる確率を P_q とする．書き戻し要求メッセージが送られるのは CPU が読み込み，または書き込みアクセスしているキャッシュの状態が P_d であった場合である．

4.4　セミ・マルコフ過程による解析

4.4.1　解 析 方 法

本節では，セミ・マルコフ過程による定常状態確率の求め方を述べ，数値例により解析結果を示す．定常状態確率を次の手法で求め，解析する．

(1) 状態確率 $P = (P_{i,j})$ を初期化する

(2) 離散時間型マルコフ連鎖で方程式 $\pi = \pi P$ を解くことによって，定常分布ベクトル $\pi = \{\pi_i\}$ を得る

(3) 式 (7) により定常状態確率を求める

(4) $\{P_i\}$ を用いて状態確率を更新する

(5) 直前の $\{P_i\}$ の値と現在の $\{P_i\}$ の値の差が，設定した十分小さな閾値より大きければ (2) に戻る

表 4　対象モデルにおけるパラメータの値

CPU クロック	2	[GHz]
CPU–L2 キャッシュのアクセス時間 t_1	0.01	[μs]
L2 キャッシュ–メインメモリのアクセス時間 t_2	1	[μs]
ネットワークインターフェイス処理時間 t_3	3	[μs]
各ノードの共有メモリ容量 M	1024	[MB]
L2 キャッシュ容量 C	1024	[KB]
キャッシュラインのサイズ l	4	[KB]
ネットワーク容量 B	10	[Gbps]
キャッシュヒット率 h	0.95	
波長変換に要する時間 T_{CD}	0	[μs]

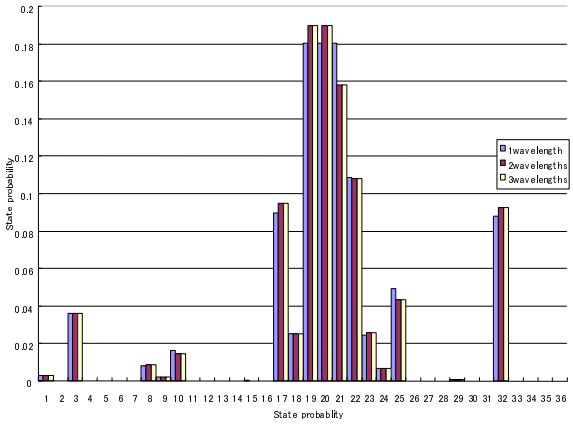


図 5　リング UMA アーキテクチャの定常状態確率 ($L = 100$km, $N = 32, s = 10^{-3}$)

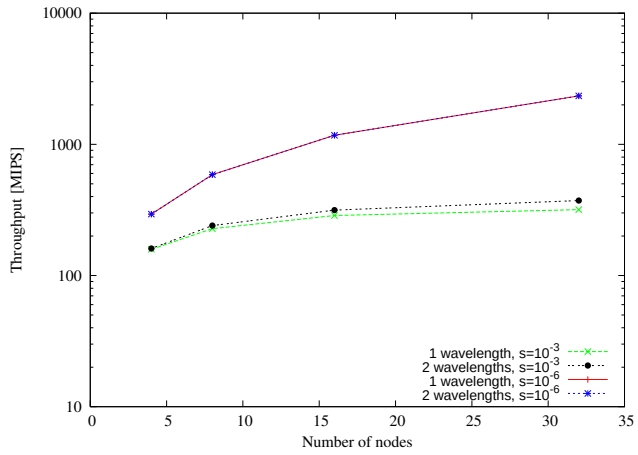
4.4.2　数　値　例

共有メモリアーキテクチャの解析のために，表 4 のようにパラメータを設定する．CPU の使用割合として，読み込みが 15%，書き込みが 5%，その他の演算処理が 80%とする．つまり $r = 0.75$，$w = 0.25$ とする．また，ノード数 N，ネットワークのリング長 L，共有メモリへのアクセス率 s を設定し，これらは独立したパラメータとする．

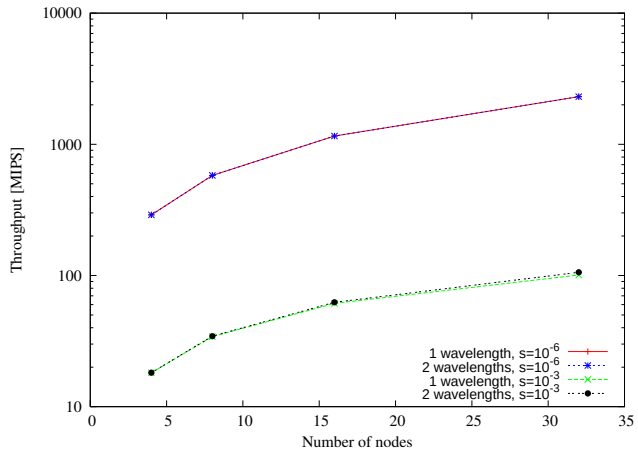
ここで，定常状態確率を求める．リング UMA アーキテクチャの定常状態確率を図 5 に示す．状態 1 は CPU の計算状態であり，状態 32 は CPU のローカルメモリへのアクセス状態である．また，状態 19, 20, 21 は一貫性制御を行っている状態であり，共有メモリへのアクセス頻度やデータ共有に要する遅延時間に影響される．遅延時間が増加する場合，共有メモリアクセスの頻度が増加した場合，この状態確率は大きくなる．すなわち，遅延時間の多くはノードの処理遅延やデータの伝播遅延である．これは，リング長やノード数が増大すると大幅に増加する．逆に，リング長が短くノード数が少ない場合は，データ共有における遅延時間が小さく，共有メモリへのアクセス頻度によらず一貫性制御の状態確率が小さく，計算処理の状態確率が大きくなる．

4.5　共有メモリアーキテクチャの性能評価

本節では，4.4 節で行った数値解析の結果を用いて，計算ス



(a) $L = 1$km



(b) $L = 100$km

図 6　リング UMA アーキテクチャの計算スループット

ループットを求め，これを指標として共有メモリアーキテクチャの性能評価を行う．

計算スループットを以下のように定義する．また，計算スループットの単位は MIPS (Million Instructions Per Second) とする．

$$\text{Throughput} = \frac{N}{0.8\eta_1 + 0.2st_{share} + 0.2(1-s)t_{private}} \quad (8)$$

ここで $t_{private}$ はローカルメモリにアクセスする平均時間であり，式 (9) で表す．

$$t_{private} = h \times t_1 + (1-h) \times (t_1 + t_2) \quad (9)$$

図 6 にリング UMA アーキテクチャにおいて 1 波長および 2 波長を用いた場合の計算スループットを示す．図に示していないが，リング UMA，リング NUMA アーキテクチャの計算スループットはほぼ同様の結果となっている．

リング長が短く，共有メモリへのアクセス頻度が低い場合には 2000MIPS 以上の計算スループットを計測することができている．しかし，リング長が長い場合，共有メモリへのアクセス頻度が高い場合には著しい性能低下がおきている．これはデータ共有における遅延時間が原因である．また，1 波長と 2 波長のアーキテクチャの性能差を比較すると，リング長が短い場合，複数波長を用いるとノード数が増えるにしたがって計算スループットは向上しており，リング長 1km，ノード数 32 の場合には約 15%向上している．しかし，計算スループットの数値としては 1000MIPS にも達しておらず，共有メモリへのアクセス頻度が高い並列計算では十分な性能は得られない．また，リング長が長い場合や共有メモリアクセス頻度が低い場合には波長数による性能差は生じていない．これはノードの処理遅延に比べてデータの伝播遅延の割合が大きく，リングトポロジでは波長数を増やしても長いリング長で性能のよい計算ループットを得ることが難しいことを示している．

5. ま と め

本稿では，光リングネットワークを用いた λ コンピューティング環境に適したリングトポロジを用いた共有メモリアーキテクチャの設計と評価を行った．設計したリングトポロジを用いる共有メモリアーキテクチャをモデル化し，セミ・マルコフ過程を用いて解析を行った結果，リング長が短く，共有メモリへのアクセス頻度が低い場合には 2000MIPS 以上の計算スループットを計測することができた．また，2 波長を用いた場合，リング長が短く，共有メモリへのアクセス頻度が高い場合に性能が向上することが明らかになった．しかしながら，共有メモリへのアクセス頻度が高い並列計算では，リングトポロジを用いる共有メモリアーキテクチャではデータの伝播遅延がボトルネックとなり，複数波長を用いてさえ十分な計算スループットをを得ることが難しい．今後は伝播遅延が小さく抑えられるメッシュトポロジを用いた共有メモリアーキテクチャの性能の評価を行い，報告する予定である．

文　献

[1] M. Imoto, E. Taniguchi, K. Baba and M. Murata: “Implementation and evaluation of MPI library with globus toolkit for establishing λ computing environment”, Proceedings of 6th Asia-Pacific Symposium on Information and Telecommunication Technologies, pp. 421–426 (2005).

[2] 井本 舞, 合田 圭吾, 馬場 健一, 村田 正幸：“λ コンピューティング環境における OpenMP ライブラリのためのデータ共有機構の設計”, 電子情報通信学会技術報告 (PN2006-28), **106**, 281, pp. 19–24 (2006).

[3] 合田 圭吾, 井本 舞, 藤本 典幸, 馬場 健一, 村田正幸：“λ コンピューティング環境における OpenMP ライブラリの設計と実装”, 電子情報通信学会技術報告 (PN2006-40), **106**, 419, pp. 5–8 (2006).

[4] A. Okada, H. Tanobe and M. Matsuoka: “Dynamically reconfigurable real-time information-sharing network system based on a cyclic-frequency AWG and tunable-wavelength lasers”, Proceedings of ECOC 2003, Vol. 4, pp. 978–979 (2003).

[5] O. R. Haverkort: “Performance of Computer Communication Systems”, WILEY (1998).

[6] A. K. Somani: “Survivability and Traffic Grooming in WDM Optical Networks”, CAMBRIDGE (2005).

[7] O. Gerstel, P. Lin and G. Sasaki: “Wavelength assignment in a WDM ring to minimize cost of embedded SONET rings”, Proceedings of IEEE INFOCOM ’98, pp. 94–101 (1998).