

並列計算機ノードのシステムオンチップ化とその性能

古川 文人[†] 渡辺 健太郎[†] 大津 金光[†]
吉 永 努[†] 馬場 敬信[†]

2005年の半導体技術では、1999年現在の並列計算機1ノードをそのままシステムオンチップ(SOC)化することが可能である。この場合、1)PEだけのSOC化、2)PEとルータのSOC化の2通りのSOC化方法に大別できる。本稿では、この2方法について比較した。その結果、並列計算機の規模が大きくなるほどルータまでSOC化した方が性能およびコストの面で有効であることがわかった。

The Case for System-On-Chip Nodes for Future Parallel Computers

FUMIHITO FURUKAWA,[†] KENTARO WATANABE,[†] KANEMITSU OOTSU,[†]
TSUTOMU YOSHINAGA[†] and TAKANOBU BABA[†]

In 2005, the parallel computer node of the year 1999 will be implemented as a system-on-chip processor. In this case, there are two choices. One is that only a PE is implemented on a chip. The other approach is to implement a PE and a router as a single chip. In this paper, we compared these two approaches. The result show that it is effective that a PE and a router are implemented on a chip, when total number of parallel computers nodes is large.

1. はじめに

半導体微細加工技術の進歩によって、汎用MPU、DRAM、特定用途向けハードウェアを1チップに集積するシステムオンチップ(以下、SOC)が実現可能になっている。SOC化の意義はシステムの、1)高性能化、2)低コスト化、3)低消費電力化、4)ダウンサイジング化の4点を達成することにある。システムの部分的なSOC化において、安易なSOC化では少なくとも上記1)を達成できない可能性がある¹⁾。このため、逐次プログラムあるいは、小規模並列プログラム*の高速実行を目的とした計算機システムのSOC化に関する研究が行われている²⁾³⁾。

また、米国SIAロードマップ⁴⁾(表1)では、2005

表1 MPUとDRAMのロードマップ(1998年版)

	1999年	2005年	2011年
最大集積Tr数(/チップ)	2千万	2億	14億
動作周波数(GHz)	1.25	3.5	10
ピン数(/パッケージ)	1400	2619	4898
最大搭載DRAM容量 (bytes/MPUチップ)	115M	1.43G	15.9G

[†] 宇都宮大学工学部

Faculty of Engineering, Utsunomiya University

* オンチップマルチプロセッサ(5台程度)での実行。

年時点のMPUチップ上の最大集積トランジスタ数は2億個、最大搭載オンチップDRAM容量は1.43GBと予測されている。これに従えば、少なくとも2005年には、1999年現在の並列計算機ノードをそのままSOC化することが可能である。

このとき、並列計算機を1システムとして各ノードをSOC化する場合、逐次プログラムを実行するノード単体の高性能化の追求だけでは、システムの高性能化が達成できるとは限らない。なぜなら、並列プログラムは、複数のノード間で通信をしながら処理を進めるため、データ配置の最適化を行なった上でもなお、他ノードとの十分な通信性能を要求するからである。

本稿では、逐次・小規模並列プログラムの高速実行のための計算機システムではなく、2005年における数十~数千台規模の並列計算機のノードのSOC化方法について考察する。以下、2章では、現在我々が想定しているSOC化ノードで構成する並列計算機とSOC化ノードの構成要素を示す。3章では、それらの構成要素によって実現可能なSOC化方法を全て列挙し、評価対象にするSOC化方法を選定する。4章では、選定したSOC化方法について、比較評価を行なう。5章では、その評価結果から、ルータをSOC化する有効性について考察して、6章でまとめる。

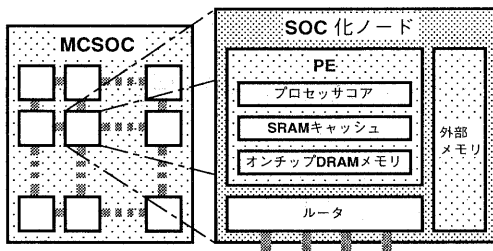


図1 MCSOCとSOC化ノード内構成要素

2. SOC化ノードで構成する並列計算機とSOC化ノード内構成要素

現在我々が想定しているSOC化ノードで構成する並列計算機MCSOC (MultiComputers with System On a Chip) とノード内の構成要素を図1に示す。

MCSOCは複数のSOC化ノードが2D-メッシュのネットワークで結合されている。SOC化ノードは、1つのPEと0バイト以上の外部メモリ、および4本のネットワークへのリンクを持つルータから構成される。ただし、PEは1つのプロセッサコア、それぞれ0バイト以上のSRAM キャッシュとオンチップDRAMメモリを1チップ上に集積したDRAM混載SOCである。

また、ルータはルーティングアルゴリズムによらず高々百万トランジスタ程度で実現可能である⁵⁾。したがって、2005年以降のSOC化において、その消費トランジスタ数を考慮する必要はない。ルータのSOC化において、考慮しなければならないハードウェアリソースは、ネットワークへのリンクによって消費されるピン数である。

3. ノードのSOC化方法

SOC化方法の選択肢はPEとルータをSOC化する場合(図2)と、それぞれ別々のパッケージに実装する場合(図3)とに大別でき、さらにその中で細分化すると、以下の5通りに分けられる。

- PEとルータをSOC化する場合(図2)
 - 外部メモリを備える方法(図2(a))
 - 外部メモリを備えない方法(図2(b))
- PEとルータをそれぞれ別々のパッケージに実装する場合(図3)
 - 外部メモリを備え、PEパッケージのピンをシステムバスに全て使用する方法(図3(a))
 - 外部メモリを備え、PEパッケージのピンを外部メモリバスとルータ内のPE I/Fへのリンクで分けて使用する方法(図3(b))
 - 外部メモリを備えない方法(図3(c))

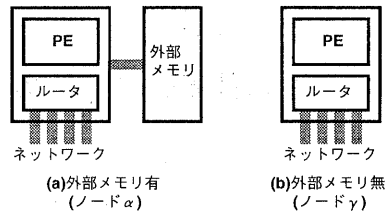


図2 PEとルータのSOC化

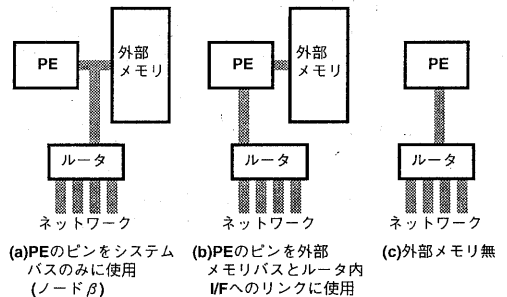


図3 PEとルータの非SOC化

ここで、図3(b)のSOC化方法は、図3(a)のSOC化方法と比べて、外部メモリ-ルータ間のデータ転送がPEを介した比較的低速な転送となり、有利な点がない。また、図3(c)のSOC化方法は、PE-ルータ間のデータ転送がパッケージ外のバスを使用した比較的低速な転送となり、図2(b)のSOC化方法に対して有利な点がない。PEとルータを別パッケージで実装する方法は、PEとルータをSOC化する方法に対して、比較的高コストであり、性能面で有利な点がないため選択する意義がない。

したがって、評価する意義があるのは、図2(a)、図3(a)、図2(b)の3方法である。

以降、これらの方法でSOC化するノードをそれぞれ、ノードα、ノードβ、ノードγと呼ぶ。

4. 評価

評価は、評価式を定義し、各パラメータを設定して計算することにより行なう。

4.1 評価式

プロセッサコア内でデータを要求する命令が発行されてから、そのデータがレジスタにセットされるまでの平均時間を平均データアクセス時間T(サイクル)とすると、Tは式(1)により得られる。この平均データアクセス時間はメモリ素子のアクセスタイム、命令発行からメモリ素子へアクセスを要求するまでの時間、メモリ素子からデータが返送されてからレジスタにセットされるまでの時間の合計である。

$$T = p_c T_c + (1 - p_c) p_{M_{on}} T_{M_{on}} + (1 - p_c) p_{M_{off}} T_{M_{off}} + (1 - p_c) \{1 - (p_{M_{on}} + p_{M_{off}})\} T_r \quad (1)$$

式 (1) の右辺の各パラメータの意味を以下に示す。

- p_c
SRAM キャッシュにヒットする確率
- $p_{M_{on}}$
オンチップ DRAM メモリに要求データが存在する確率
- $p_{M_{off}}$
外部メモリに要求データが存在する確率。ノード γ の場合は 0。
- T_c
SRAM キャッシュヒット時のデータアクセス時間 (サイクル)
- $T_{M_{on}}$
オンチップ DRAM メモリアクセス時のデータアクセス時間 (サイクル)
- $T_{M_{off}}$
外部メモリアクセス時のデータアクセス時間 (サイクル)。ノード γ の場合は 0。
- T_r
他ノードへのリモートメモリアクセス時のデータアクセス時間* (サイクル)

以下では、リモートメモリアクセスの発生する確率を p_r と表記する。このとき、

$$p_r = 1 - (p_{M_{on}} + p_{M_{off}})$$

である。

また、ノード α およびノード β の各 T , $T_{M_{off}}$, T_r をそれぞれ、 T_α , $T_{\alpha M_{off}}$, $T_{\alpha r}$, T_β , $T_{\beta M_{off}}$, $T_{\beta r}$ 。ノード γ の T , T_r をそれぞれ、 T_γ , $T_{\gamma r}$ と表記する。

4.2 パラメータの設定

4.2.1 ハードウェアリソースと動作周波数

表 1 より、2005 年の半導体技術では、以下の SOC を実現可能と仮定する。

- 最大集積トランジスタ数：2 億個
- SOC 動作周波数：3.5GHz
- パッケージ信号線数：1310 本
- 最大搭載 DRAM 容量：1.43GB

上記のパッケージ信号線数は、表 2 より、Alpha21264 がパッケージピン数の約半数だけしか信号線に使用できないことをそのまま適用し、2005 年のパッケージピン数 2619 本の半数と仮定した。

4.2.2 SOC 化ノード構成要素

前述したように、2005 年の半導体技術ではルータの消費トランジスタは無視してもよい。したがって、

* 全要求データは他ノード PE 内メモリに存在すると仮定し、全ノードに対するアクセスの平均的な時間とする

SOC の面積を消費するのは、ほぼ PE 構成要素のみである。各 PE 構成要素を、表 3 に示すように仮定する。

表 2 Alpha21264 の消費トランジスタ数⁶⁾ とピン数⁷⁾

Tr 数	プロセッサコア	600 万
	キャッシュ(命令、データ各 64KB)	900 万
ピン数	信号線	367
	電源、グラウンド	300

表 3 各 PE 構成要素の仮定

プロセッサコア	Alpha21264 コア (5%)
命令・データキャッシュ	各 640KB (45%)
オンチップメモリ容量	700MB (50%)

(括弧内は消費チップ面積の割合)

4.2.3 評価式の各パラメータ

評価式の各パラメータ値と関係を以下のように仮定する。

- $p_c = 90\%$, $p_{M_{on}} = 70\%$
- $T_c = 1$ サイクル

このためには、SRAM キャッシュが SOC の動作周波数 3.5GHz 以上で動作しなければならない。つまり、演算器・レジスタから SRAM 素子までのデータパスのレイテンシと SRAM 素子自体のアクセスレイテンシの合計が 0.285ns 以下の必要がある。

- $T_{M_{on}} = 4$ サイクル

このためには、オンチップ DRAM メモリが 3.5GHz/4 = 875MHz 以上で動作しなければならない。つまり、演算器・レジスタからオンチップの DRAM 素子までのデータパスのレイテンシとその DRAM 素子自体のアクセスレイテンシの合計が 1.142ns 以下の必要がある。

- $T_{\beta M_{off}} = 20$ サイクル

このためには、外部メモリへのバスおよび外部メモリ自体が 3.5GHz/20 = 175MHz 以上で動作しなければならない。つまり、演算器・レジスタからオフチップのバスを介する外部メモリの DRAM 素子までのデータパスのレイテンシとその DRAM 素子自体のアクセスレイテンシの合計が 5.7ns 以下の必要がある。

- $T_{\alpha M_{off}} \geq T_{\beta M_{off}}$

ノード β はネットワークへのリンクに SOC のピン数を消費しないため、外部メモリへのバンド幅がノード α のものよりも高くなる。したがって、ノード β の外部メモリアクセス時間がノード α のものよりも短くなる可能性がある。

- $T_{\alpha r} \leq T_{\beta r}$

ノード β のルータは、自ノードのオンチップメモリ-ルータ間のデータ転送にオフチップのバスを

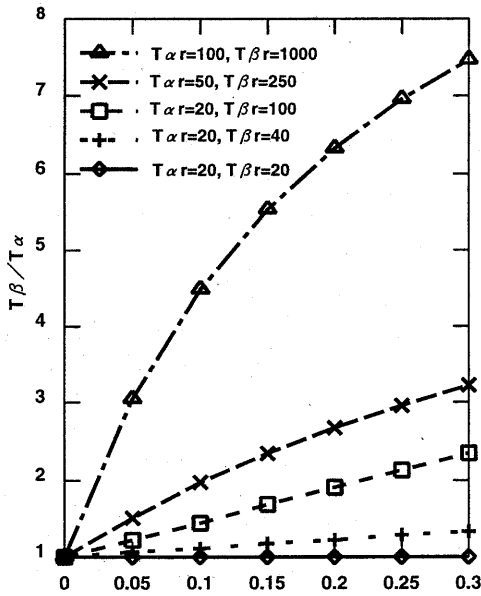


図4 $T_{\alpha M_{off}} = T_{\beta M_{off}}$ のときの T_{β}/T_{α}

介することにより、ノード β のリモートメモリアクセス時間がノード α のものよりも長くなる可能性がある。

- $T_{\alpha r} \geq T_{\gamma r}$
ノード α の SOC のピンは、ネットワークへのリンクだけではなく、外部メモリへのバスにも消費されるため、ネットワークバンド幅がノード γ のものよりも低い。したがって、ノード α のリモートメモリアクセス時間がノード γ のものよりも長くなる可能性がある。

4.3 評価結果

4.3.1 ノード α とノード β の比較

ノード α とノード β を以下の (1), (2), (3) の3つの場合に分けて比較する。

(1) $T_{\alpha M_{off}} = T_{\beta M_{off}}$ のとき

$$(T_{\alpha M_{off}} = 20, T_{\beta M_{off}} = 20)$$

ノード α とノード β の外部メモリへのデータアクセス時間に差がないときである。このときの T_{β}/T_{α} の計算結果を図4に示す。横軸は、キャッシュミス時のリモートメモリアクセスの発生確率 p_r である。縦軸は、ノード β の平均データアクセス時間 T_{β} のノード α の平均データアクセス時間 T_{α} に対する割合を示している。

図4から、 T_{α} は T_{β} よりも常に短く、ノード β の利点がないことがわかる。つまり、ルータを SOC 化しない選択肢には意義がない。

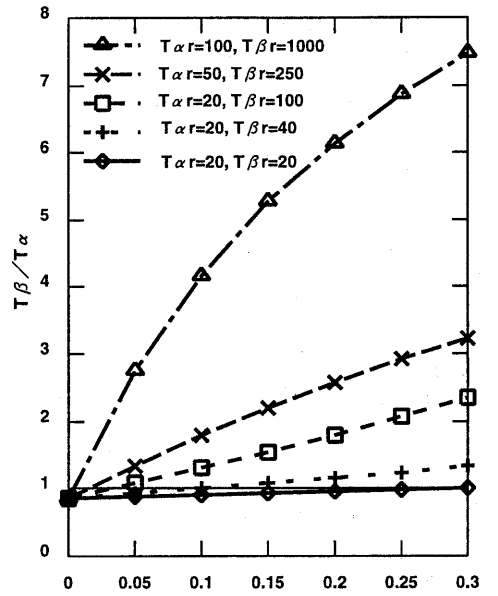


図5 $T_{\alpha M_{off}} = 30, T_{\beta M_{off}} = 20$ のときの T_{β}/T_{α}

T_{α} が T_{β} よりも短くなる場合は、 $T_{\alpha r}$ が $T_{\beta r}$ よりも短いときである。また、両ノードのリモートメモリアクセス時間 $T_{\alpha r}$ と $T_{\beta r}$ の割合が同じ場合でも、そのリモートメモリアクセス時間そのものが長くなれば、 T_{α} の T_{β} に対する短かさが顕著に現れる。例えば、同じ $T_{\alpha r}/T_{\beta r} = 1/5$ である、 $T_{\alpha r} = 20, T_{\beta r} = 100$ のときの T_{β}/T_{α} よりも、 $T_{\alpha r} = 50, T_{\beta r} = 250$ のときの T_{β}/T_{α} の方が大きくなっている。

(2) $T_{\alpha M_{off}} = 1.5 \times T_{\beta M_{off}}$ のとき

$$(T_{\alpha M_{off}} = 30, T_{\beta M_{off}} = 20)$$

$T_{\alpha M_{off}}$ が $T_{\beta M_{off}}$ に対して1.5倍と、比較的長くなるときである。このときの T_{β}/T_{α} の計算結果を図5に示す。横軸および縦軸の意味は、図4と同じである。

図5は、図4と同傾向は同じである。しかし、 $T_{\alpha r}$ が $T_{\beta r}$ よりも短いときでも、 T_{α} が T_{β} よりも長くなる場合がある。例えば、 $T_{\alpha r} = 20, T_{\beta r} = 40$ の場合では、 p_r が12%以下のとき、外部メモリアクセスコストの違いにより、 T_{β}/T_{α} が1未満になっている。

(3) $T_{\alpha M_{off}} = 100 \times T_{\beta M_{off}}$ のとき

$$(T_{\alpha M_{off}} = 2000, T_{\beta M_{off}} = 20)$$

$T_{\alpha M_{off}}$ が $T_{\beta M_{off}}$ に対して100倍と、圧倒的に長くなるときである。これは、ノード α の SOC がネットワークへのリンクにほとんどピンを消費してしまい、外部メモリへは極めてバンド幅の低い手段でアクセス

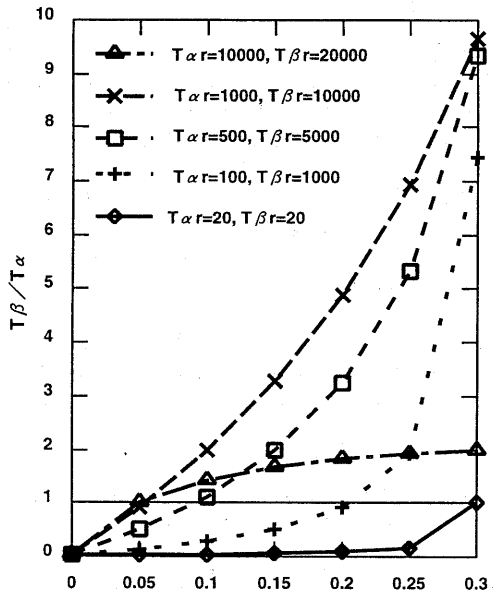


図6 $T_{\alpha M_{off}} = 2000$, $T_{\beta M_{off}} = 20$ のときの T_{β}/T_{α}

する場合を想定している。このときの T_{β}/T_{α} の計算結果を図6に示す。横軸および縦軸の意味は、図4と同じである。

図6から、 $T_{\alpha r}/T_{\beta r} = 1/10$ の場合でも、 p_r が高くないと、 T_{α} が T_{β} よりも短くならない。しかし、両ノードのリモートメモリアクセス時間自体が極端に長くなると、 T_{α} が T_{β} よりも短くなる。例えば、 $T_{\alpha r} = 10000$, $T_{\beta r} = 20000$ と、 $T_{\alpha r}/T_{\beta r} = 1/2$ の場合でも、 p_r に比較的依存することなく、 T_{α} が T_{β} よりも短くなっている。

4.3.2 ノード α とノード γ の比較

ノード α とノード γ 、すなわち、「ノード = SOC + 外部メモリ」と「ノード = SOC」の比較を以下に示す。

図7は、 $T_{\alpha M_{off}} = 20$ のときの T_{α}/T_{γ} の計算結果を示したものである。横軸および縦軸の意味は、図4と同様である。

図7から、 $T_{\alpha r} = 20$, $T_{\gamma r} = 20$ のとき、つまり、リモートメモリアクセス時間が自ノード外部メモリアクセス時間に等しい場合、 T_{α} と T_{γ} の間に差はなく、外部メモリをノードに持たせる意義はない。しかし、 $T_{\alpha r} = 40$, $T_{\gamma r} = 40$ のとき、つまり、ノード α とノード γ のリモートメモリアクセス時間は等しいが、その時間が外部メモリアクセス時間より長いときには、 T_{γ} が T_{α} のものよりも短くなることはない。

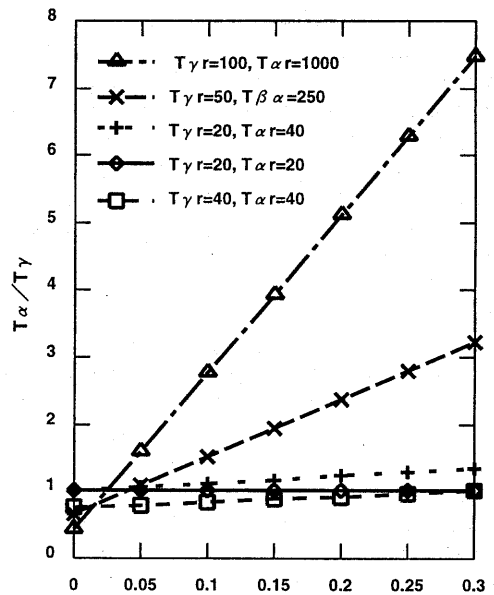


図7 $T_{\alpha M_{off}} = 20$ のときの T_{α}/T_{γ}

T_{γ} が T_{α} よりも短くなるときは、 $T_{\gamma r}$ が $T_{\alpha M_{off}}$ と $T_{\alpha r}$ のいずれよりも短くなることである。これは、両ノードのリモートメモリアクセス時間自体が、 $T_{\alpha M_{off}}$ に対して比較的長くなるほど顕著に現れる。

5. ルータのSOC化の有効性

以下では、ルータをSOC化する有効性について考察する。まず、コスト面について考えると、ルータをSOC化の方が優れている。なぜなら、ルータをSOC化すると、ノードを構成するパッケージおよびそれらを相互に結ぶバスが減るため、パッケージ・配線のコスト削減とノード自体の低消費電力化に結びつくからである。

次に、性能面について考える。前章より、 T_{α} が T_{β} よりも短くなる、すなわちルータをSOC化したノードが比較的高性能になる要因として、以下の3点があげられる。以下では、これら3点の可能性について考察する。

- $T_{\alpha M_{off}}$ と $T_{\beta M_{off}}$ の差が小さいこと
- $T_{\alpha r}$ が $T_{\beta r}$ よりも短くなること
- リモートメモリアクセス時間自体が外部メモリアクセス時間に対して比較的長くなること

はじめに、(a)について考察する。2005年のパッケージ信号線数は1310であるから、外部メモリへのデー

タブに 512 本、アドレスバスに 64 本を使用したとすると、残り 734 本使用可能となる。ルータの物理チャネル幅を 90 ビットとすると、ルータの消費ピン数は $90 \times 2(\text{全二重}) \times 4(\text{リンク数}) = 720$ 本となる SOC が実現可能である。いま、キャッシュラインサイズを 64 バイト (8 ワード = 512 ビット) とすると、これは外部メモリへのデータバス幅と一致する。よって、外部メモリアクセスへのバンド幅は十分に提供されていることになり、ノード β の外部メモリアクセス時間の方が短くなることはない。したがって、キャッシュラインサイズが非常に大きくならないと、 $T_{\alpha M_{off}}$ が $T_{\beta M_{off}}$ よりも長くない。

また、ネットワークバンド幅が要求され、外部メモリバスの消費ピン数を極端に減らす必要があるときは、高速シリアルリンクを複数本用いることにより $T_{\alpha M_{off}}$ が $T_{\beta M_{off}}$ より長くなることを抑制できる⁸⁾。以上より、 $T_{\alpha M_{off}}$ と $T_{\beta M_{off}}$ の差は大きくなりにくいといえる。

次に、(b) について考察する。MPU を搭載するような SOC の VLSI 実装法としては、フルカスタム方式が用いられることが多く、表 1 に示したような動作周波数は、この実装法を想定して予測されている。一方、ルータのような特定用途向けのハードウェアの VLSI 実装法としては、ゲートアレイやスタンダードセルへ実装するセミカスタム方式が用いられることがほとんどである。この方式で実装した VLSI は、フルカスタム方式よりも低い周波数で動作する。

よって、ノード α のルータはフルカスタム方式で SOC 内に実装され、ノード β のルータはセミカスタム方式で 1 パッケージとして実装される可能性が高い。したがって、ノード α のルータはノード β のルータに対して比較的高い周波数で動作する。ルータの高周波数動作は、自ノードのリモートメモリアクセス時間だけでなく、他ノードのものも短くする可能性がある。なぜなら、他ノードからのアクセス先ではないノードのルータでも、アクセスの中継地点になることがあるためである。これは、並列計算機の規模が大きくなるほど顕著に現れる。また、ルータを SOC 化する場合は、PE-ルータ間のバスはオンチップに実装するため、比較的高バンド幅かつ低レイテンシなバスとなる。以上より、並列計算機の規模が大きくなるほど、 $T_{\alpha r}$ が $T_{\beta r}$ よりも短くなる可能性がある。

次に、(c) について考察する。並列計算機の規模が大きくなるほど、リモートアクセス先までの距離は長くなる可能性が高い。したがって、並列計算機の規模が大きいほどリモートメモリアクセス時間自体が外部メモリアクセス時間に対して比較的長くなる可能性がある。

以上をまとめると、大規模な並列計算機ほどルータの SOC 化の有効性が高いといえる。

6. おわりに

本稿では、2005 年における大規模並列計算機のノードの SOC 化方法について考察した。その結果、PE とルータの SOC 化は、並列計算機の規模が大きくなるほど、性能・コストの両面で有効性が高いことがわかった。

今後は、本稿で述べた SOC 化ノードで構成する並列計算機 MCSOC を実アプリケーションを使ってシミュレーションし、いくつかの SOC 化方法を検討する予定である。

謝辞 本研究は、一部文部省科学研究費 基盤研究 (B) 課題番号 10558039、奨励研究 (A) 課題番号 11780190 の援助によるものである。

参考文献

- 1) 新井智久、矢野陽一：“マイクロプロセッサ”，電子情報通信学会誌 Vol.81 No.11, pp.1107-1112, 1998 年 11 月
- 2) 宮嶋浩志、岩下茂信、村上和彰：“高性能システム・オン・チップ構成法に関する性能評価”，情報処理学会研究報告 96-HPC-62-6, pp.33-38, 1996 年 8 月
- 3) 坂井修一：“オンチップマルチプロセッシングに関する初期的検討”，情報処理学会研究報告 97-ARC-122, pp.33-38, 1997 年 2 月
- 4) Semiconductor Industry Association：“International Technology Roadmap for Semiconductors 1998 Update”，[http://notes.sematech.org/ntrs/PubLNTRS.nsf/Lookup/98Update/\\$file/98Update.pdf](http://notes.sematech.org/ntrs/PubLNTRS.nsf/Lookup/98Update/$file/98Update.pdf), 1998
- 5) 林匡哉、堀田真貴、吉永努、大津金光、馬場敬信：“適応ルータの効率的な並列デッドロックリカバリ方式の提案”，並列処理シンポジウム JSPP'99, pp.55-62, 1999 年 6 月
- 6) Robert Yung：“Evaluation of a Commercial Microprocessor”，<http://www.sunlabs.com/technical-reports/1998/sml-tr-98-65.pdf>, pp.38, 1998 年 6 月
- 7) COMPAQ：“Alpha 21264 Microprocessor Data Sheet”，<http://ftp.digital.com/pub/Digital/info/semiconductor/literature/21264ds.pdf>, 1999 年 2 月
- 8) 松本尚、平木敬：“Memory String Architecture —メモリウォールを越えて—”，情報処理学会研究報告 96-ARC-120-1, pp.1-6, 1996 年