RC-003

複数 FPGA によるアレイ型差分法専用計算機のための FPGA 間通信帯域評価

Evaluation of Inter-FPGA Communication Bandwidth for Array-Structured Custom Computers Dedicated to Difference Schemes

王 陸洲† 佐野 健太郎† 初田 義明† 山本 悟†
WANG Luzhou Kentaro SANO Yoshiaki HATSUDA Satoru YAMAMOTO

1. 緒言

近年，数値シミュレーションは様々な分野において必要不可欠な技術となっており，高性能計算の需要が高まっている．しかしながら，多数のマイクロプロセッサを用いた並列計算に頼る現在の大規模計算機では，プロセッサの増加に伴う並列処理効率の低下や消費電力の増大が問題となっている．これは，任意のプログラムを実行可能な汎用マイクロプロセッサが，対象とする計算に対して必ずしも効率的ではないことが原因である．マイクロプロセッサでは，チップ面積の大部分が，キャッシュメモリ，分岐予測機構，命令スケジューラといった計算以外の目的に使用されており，本来実装可能な演算器が犠牲となっている．また，複数のプロセッサを接続した並列計算機では，ネットワークや共有メモリを起因とするオーバヘッドのため，各プロセッサの稼働率はさらに低下する．このため，目標の計算性能を実現するため多数のプロセッサを用いることとなり，さらなる処理効率低下や消費電力の増大を招くことになる．

これに対して，計算問題に特化した構造により高い台数効果の下，効率良く高性能を実現する専用計算機が研究されている．特に，近年ではマイクロプロセッサを上回る浮動小数点演算性能を持ちつつある FPGA（Field-Programmable Gate Array）[1]を用いて数値計算のための専用計算機を構築する試みが為されている[2][3][4][5]．Smith らは，FPGA を用いて，流体計算の大半を高速化するアクセラレータを構築した[4]．このアクセラレータは，バッファを含む専用データパスによりデータ入出力の帯域低減を図りながら高スループットの計算を実現している．He らは，FPGA による電磁場計算専用の計算機を提案した[5]．He らの専用計算機でも，バッファを用いて入力データを活用し，入力データを再利用することにより，高次精度の差分計算を高速に実行するためにデータ入出力帯域を低減している．

本研究では，差分解法に基づく計算問題を指向した専用計算機を提案している．これまで，単一の FPGA を用いた試作実装では，96 MHz 動作にもかかわらず 18.3GFlops の単精度浮動小数点ピーク性能を実現し，差分法に基づく２次元の伝熱計算，流体計算および電磁波計算に対して 14.7 GFlops 以上の実効性能を確認できた[6]．提案する専用計算機はアレイの規模に比例した計算性能を有することから，複数 FPGA による大規模実装を次の目標としている．このためには，十分な FPGA 間帯域が必要となる．

本論文では，現在利用可能な FPGA が，複数 FPGA による実装に対して十分な I/O 帯域を持つことを検証する．まず，アレイの規模や動作周波数，計算問題における通信頻度のパラメータを用いて，FPGA 間要求帯域のモデルを与える．次に，専用計算機の試作結果に基づいて実際の要求帯域を求める．最後に，現在利用可能な FPGA について I/O 帯域の評価を行い，複数の FPGA を用いた実装において，提案するアレイ型専用計算機が FPGA 数に比例する性能を実現できることを示す．

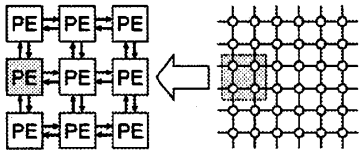

図 1 アレイ型専用計算機の構造と計算領域の関係

2. 差分法および専用計算機のアーキテクチャ

2.1 差分法

差分法に基づく数値計算では，微分方程式を近似した差分式に対して反復計算を行い，数値解を得ることが一般的である．例えば，2 次元ラプラス方程式(1)に対して中心差分を適用すると，差分式(2)が得られる．

$$\frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} = 0 \tag{1}$$

$$\phi_{i,j}^{n+1} = c_1\,\phi_{i-1,j}^{n} + c_2\,\phi_{i+1,j}^{n} + c_3\,\phi_{i,j-1}^{n} + c_4\,\phi_{i,j+1}^{n} \tag{2}$$

ここで，ϕ は位置(x, y)と時間(t)のスカラ関数，$\phi_{i,j}^{n}$ は格子点(i, j)に離散化されたϕであり，n は反復数である．c_1, ..., c_4 は差分格子に依存する定数である．反復法の一つであるヤコビ法では，式(2)の計算を全格子点に対して繰り返すことにより，$\phi_{i,j}$ の数値解を得る．このような反復法は，各格子点毎の計算の並列性および規則性に加えて，各格子点計算が近傍格子点にのみ依存する局所性を持つ．SOR 法や multi grid 法による他の反復解法や高次差分スキームを用いる場合でも，得られた差分式の計算が式(2)のような隣接格子点の値に対する累算の組み合わせに帰着できると考えられる．

2.2 差分法専用計算機のアーキテクチャ

本研究では，式(2)のような累算を高速に計算するためのアレイ型の専用計算機を提案している[3][6]．アレイ型専用計算機の構造と計算領域の関係を図 1 に示す．3 次元アレイでは大規模化に伴い配線が問題となるため，FIFO を介したメッシュネットワークによる 2 次元アレイを採用し

† 東北大学 大学院情報科学研究科
{limgys, kentah, hatsuda, yamamoto}@caero.mech.tohoku.ac.jp

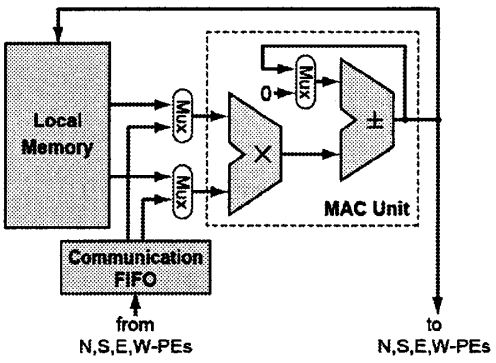

図2 計算要素のデータパス

```
1:    mul c1, W       ,           ,
2:    mul c1, phi[0,0],           ,
3:    mul c1, phi[1,0],           ,
4:    mac c2, phi[1,0],           ,
5:    mac c2, phi[2,0],           ,
6:    mac c2, phi[3,0],           ,
7:    mac c3, S       ,           ,
8:    mac c3, S       ,           ,
9:    mac c3, S       ,           ,
10:   mac c4, phi[0,1], phinew[0,0], SW
11:   mac c4, phi[0,1], phinew[0,0], S
12:   mac c4, phi[0,2], phinew[0,0], S
```

図3 累算を行うマイクロプログラム

ている．2 次元直交格子上の計算では，各計算要素(PE)は格子を分割して得られる部分格子を担当し，上下左右(NSWE)の PE と必要なデータを通信しながら任意の項数の累算を並列に実行可能である．3 次元格子の計算の場合には，例えば式(2)に z 方向の差分を加えた 6 項の累算を行ない，各 PE に柱状の部分領域を割り当てることにより，2 次元のアレイでも上下左右の隣接 PE との通信を行ないながら並列計算が可能である．

PE は図 2 のデータパスを持ち，マイクロプログラムにより指定された順番に複数の演算を実行可能である．浮動小数点積和演算器(MAC)は前の計算結果に加算を行なう累算モードを持ち，ローカルメモリへの中間結果の書き込み無しに，累算を連続して計算することができる．

MAC は 5 段のパイプラインとして実装されており，ローカルメモリから読み出されたデータや隣接 PE から送られてきたデータを入力できる．また，MAC の出力を上下左右の隣接 PE に送ることができる．MAC では，加算器の出力から加算器の入力への 3 段前へのフォワーディングにより累算を実現しているため，累算は 3 サイクル毎の入力に対して連続に行なわれる．このため，積和演算器の稼働率を高めるには，3 つの累算式の各項を交互に入力する必要がある．

$\phi_{0,0}^{n+1}$，$\phi_{1,0}^{n+1}$，$\phi_{2,0}^{n+1}$ のそれぞれに対し，式(2)を計算するためのマイクロプログラムを図 3 に示す．第 1，第 2 オペランドは入力を，第 3 オペランドは出力を書き込むローカルメモリのアドレスを，第 4 オペランドは MAC の出力を送信する隣接 PE を表す．この例では，1，4，7，10 サイクルの 4 命令が $\phi_{0,0}^{n+1}$ に対する 4 項の累算を行っている．10サイクルでは，計算結果をローカルメモリに書き込むと同時に S(下)と W(左)方向の 2 つの隣接 PE に転送する．

これまで，単一の ALTELA 社 StratixII FPGA を用い，12×8 の 96 個の PE からなるアレイ型専用計算機の試作を行った[3][6]．96 MHz で動作する本計算機のピーク性能は 18.3 GFlops である．伝熱問題，流体問題，電磁波問題の 2 次元計算に対して 8 割以上の MAC の稼働率と 14.7 GFlops の単精度浮動小数点実効性能が得られた．次の目標として，単一の FPGA を越える計算性能の実現に向けて，複数の FPGA を用いたアレイ型専用計算機の構築を目指しているところである．図 4 に示すような複数 FPGA の 2 次元配列上に大規模なアレイ型専用計算機を実装する場合、各 PE で滞りなく計算を行うためには FPGA 間の通信帯域が十分である必要がある。要求される帯域は、計算における隣接 PE へのデータ送信頻度により左右される。次章では、PE の性能と要求帯域のモデルを与え、PE の性能を低下させずに済むための FPGA 間通信帯域の条件を導く。

3. 複数の FPGA を用いた拡張

3.1 アレイ型専用計算機の性能モデル

アレイ型専用計算機の実効性能 P は，アレイの規模である PE 数 N，演算器の稼働率 U および動作周波数 F との積によって与えられる．

$$P = NUF \tag{3}$$

FPGA あたりの PE 数を N_{PE} とすると，FPGA 単体の実効性能は $P_{FPGA}=N_{PE}UF$ になる．FPGA を M 個用いる場合には，演算器の総数が M 倍になるものの，FPGA 間の通信帯域による制約等のために，実際に得られる性能は $P \leqq MP_{FPGA}$ となる．以下では，$P=MP_{FPGA}$ となる帯域条件について議論する．

3.2 通信帯域による制約

図 5 に隣接 FPGA 間の通信機構を示す．これをリンクと呼ぶ．各 PE の送信するデータサイズを s とすると，N_b 個の PE が同時に送信するリンクあたりのデータサイズは sN_b となる．各 PE は最大で毎サイクル 1 つの演算結果を隣接 PE に送信するため，動作周波数 F に対して，リンクあたりの最大通信帯域は sN_bF となる．

しかしながら，図 3 のマイクロプログラムのように，実際の計算ではデータ送信は全てのサイクルにおいて起こらない．ここで，プログラムにおけるある方向へのデータ送信命令の最小間隔を d サイクルとすると，リンクに対する必要帯域 w_b は次式で与えられる．

$$w_b = \frac{sN_bF}{d} \tag{4}$$

例えば，図 3 の例では，S 方向への通信命令が 10，11，12 サイクルに連続して実行されるため，S 方向のリンクに対する d は 1 となる．

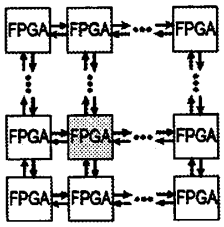

図4 2次元FPGAアレイ

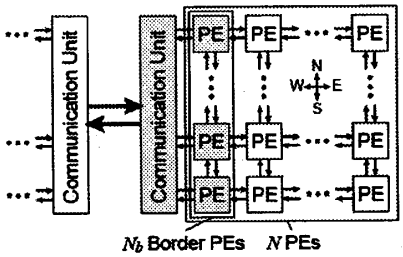

図5 FPGA間の通信機構

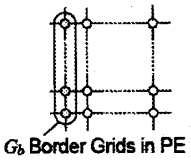

図6 PEが担当する部分格子

FPGA の I/O に必要な帯域 w_{FPGA} 各リンクおける必要帯域の和として与えられる．図4に示すように，FPGA を2次元アレイ状に配置する場合，w_{FPGA} は式(5)より与えられる．

$$w_{FPGA} = sF\left(\frac{N_{bN}}{d_N} + \frac{N_{bS}}{d_S} + \frac{N_{bE}}{d_E} + \frac{N_{bW}}{d_W}\right) \quad (5)$$

各 FPGA の I/O 帯域を W_{FPGA} とすると，PE をストールさせないためには $w_{FPGA} \leqq W_{FPGA}$ でなくてはならない．w_{FPGA} を小さくするために N_b，U または F を下げると FPGA あたりの計算性能 P も低下するため，好ましくない．一方，各方向の d を十分に大きくできれば，FPGA あたりの性能を低下させずに帯域条件を満たすことができる．

3.3 FPGA 間データ通信命令の最小間隔 d の見積もり

本節では，実際の計算問題に対し，d の見積もりを行う．不必要に nop 命令を挿入することにより d を大きくすることも可能であるが，同時に演算器の稼働率 U が低下してしまうため，ここでは U を変えない範囲で命令のスケジューリングを行う場合の d について考えることにする．

実際の計算問題ではデータを送信しない命令もあるため，データを送信する命令を等間隔に並べることができれば，d は1よりも大きい値となる．各格子点に対して差分式を計算するのに I 命令を必要とし，各 PE が G 個の格子点を計算する場合，各 PE が担当する部分格子に対して計算を行うには GI 命令必要である．次に，FPGA 間のデータ送信を生じる命令の数を考える．図6に示すように，隣接 PE に計算結果を送信する必要のある格子点は部分格子の端に並ぶ G_b 個であるとし，データ送信命令を均一に配置できるとすると，d は次式より求められる．

$$d = \frac{GI}{G_b} \quad (6)$$

式(6)において，I はアルゴリズムによる因子であり，格子点あたりの演算数が多いほど通信間隔を広くできることを意味する．対して G/G_b はハードウェア規模と計算規模による因子であり，2次元または3次元の計算格子では，G はそれぞれ $O(G_b^2)$，$O(G_b^3)$ となるため，総じて PE が担当する領域が大きいほど d も大きな値を持つことが予想される．

4. 2つの FPGA を用いた実装と性能評価

前節では，専用計算機のハードウェアパラメータおよびソフトウェアパラメータから FPGA 間に要求される通信帯域を求めた．本節では，これが現在利用可能な FPGA のチップ間 I/O 帯域よりも十分に小さいことを示す．

4.1 専用計算機の試作実装と各計算問題における d の値

本研究で過去に行った単一 FPGA による試作[3][6]の2倍の規模の PE アレイを，LVDS による高速差動通信を行うデータ通信ユニットにより接続された2つの FPGA 上に実装した．実装には，DiNI 社製の FPGA ボード DN7000k10PCI に搭載の2つの ALTERA Stratix II FPGA (EP2S180-F1508-C5) を用いた．各 FPGA に 12×8=96 個の PE を実装し，これらが 96 MHz で動作することを確認した．境界 PE 数は $N_{bE}=N_{bW}=8$ である．各 PE の送信するデータサイズ s は，単精度浮動小数点 32 bit に制御信号 1 bit を加えた 33 bit である．$d=1$ の場合における最悪値を考えた場合，E,W 方向の求帯域は式(4)より $w_{bE}=w_{bW}=25.3$ Gb/s となる．

一方，隣接 FPGA 間に実装できた差動通信の帯域は $W_{bW}=36.1$ Gb/s であった．使用している FPGA ボードの配線上，同じ帯域を実現できる LVDS チャンネルが未使用状態にあるため，試作に用いた Stratis II FPGA では，$d=1$ の場合でも FPGA あたりの計算性能を落とさずに複数の FPGA を1次元配列にした拡張が可能であることが確認された．

しかしながら，FPGA を2次元配列として並べた場合，$N_{bE}=N_{bW}=8$，$N_{bS}=N_{bN}=12$を用いて要求帯域を計算すると，$w_{bE}=w_{bW}=25.3$ Gb/s，$w_{bS}=w_{bN}=38.0$ Gb/s，$w_{FPGA}=126.7$ Gb/s となる．対して，FPGA の I/O 帯域は，実装できた 36.1 Gb/s と未使用の 36.1 Gb/s を合わせて $W_{FPGA}=63.2$ Gb/s でしかないため，$d=1$ では $w_{FPGA} \leqq W_{FPGA}$ を満たさない．

次に，実際の計算問題における d を評価するため，ベンチマーク問題として，式(2)の Jacobi 法による2次元熱伝導計算(Jacobi)，フラクショナルステップ法による2次元正方キャビティ強制対流計算(Frac)[3]，FDTD 法による2次元電磁波伝播計算(FDTD)を作成し[6]，d が大きくなるよう手動により最適スケジューリングを行った．それぞれの計算条件および d を表1に示す．表1において，d の理論値は式(6)を用いて求めた値であり，実測値はスケジューリングを施したプログラムの持つ実際の d である．両方とも W 方向リンクにおける値である．Jacobi と Frac では，理論値になるまでスケジューリングできた一方，FDTD では理論値よりも小さな d が得られた．これは，FDTD では，

表 1 各ベンチマークの計算条件および結果

ベンチマーク		Jacobi	Frac	FDTD
差分式を計算する命令数 I		4	3	3
各 PE が担当する格子点数 G		8×12	4×6	6×9
部分格子の境界格子点数 G_b		12	6	9
データ送信命令の最小間隔 d	理論値	32	12	18
	実測値	32	12	6
FPGA 間の必要帯域 W_b [Gb/s]		0.79	2.11	4.22

境界条件のために格子点に対する演算順序の制約が厳しく，手動では式(4)の見積もりほど理想的なスケジューリングが困難であったためである．

しかしながら，これらのベンチマーク問題に対して得られた d は 6 ～ 32 と大きく，FPGA 間要求帯域は $d = 1$ の場合の 1/6 以下で済むことが明らかとなった．このため，Stratix II を用いた本試作に対しては，FPGA の 2 次元配列による拡張でも FPGA 間帯域が問題とならないことが確認された．

4.2 入手可能な FPGA の I/O 帯域の評価

現在入手可能な FPGA の LE (Logic Element)数，DSP ブロック数，チップ I/O 帯域を表 2 に示す[7]．また，36×36 bit 整数乗算を行う DSP ブロック数より，FPGA あたり実装可能な PE 数 N_{FPGA} を与えた．FPGA への要求帯域 w_{FPGA} は式(5)を用いて，$N_b=\sqrt{N}$，s=33 bit，F=96 MHz，d=1，2 次元 FPGA アレイとして求めた．対して，実装可能な FPGA 間の総通信帯域 W_{FPGA} は，各 FPGA において利用可能な LVDS チャンネル数と最大のビットレートとの積により求めた．表 2 より，ハイエンド向けの Stratix III と Stratix II においては，各 FPGA の持つ計算性能に対して W_{FPGA} は十分に大きいことが分かる．一方，低価額の Stratix II GX，Arria GX と Cyclone では，$w_{FPGA} \leqq W_{FPGA}$ を満さないため帯域が不足する．しかし，実際の計算問題では d が 1 より大きい値を持つことが予想されるため，これらの FPGA でも帯域条件が問題とならないことが期待される．

5. 結言

本論文では，差分法のためのアレイ型専用計算機を複数の FPGA を用いて拡張する場合の FPGA 間帯域について，ハードウェアおよびソフトウェアのパラメータによりモデル化を行った上で，試作による評価を行った．その結果，現在利用可能な FPGA では，FPGA の 2 次元配列による拡張において通信帯域に余裕があり，FPGA 数に比例した計算性能を実現可能であるとの見通しが得られた．これは，実際の計算問題では，最適スケジューリングにより通信命令の間隔を広げることにより，要求される通信帯域を低く抑えられるためである．

現在，ボード上にある 2 つの FPGA を用いて，単体の FPGA による実装と比べて 2 倍の規模のアレイによる計算に成功したところである．今後は，複数の FPGA ボードを高速差動通信により接続し，提案するアレイ型専用計算機の高い台数効果を実証する予定である．

参考文献

[1] K. Underwood, "FPGA vs. CPUs: Trends in Peak floating-point Performance", Proceeding of the International Symposium on Field-Programable Gate Arrays, 171-180, 2004.

[2] M. deLorimier and A. DeHen, "Floating-point Sparse Matrix-vector Multiply for FPGAs", Proceedings of the International Symposium on Field-Programmable Gate Array, 75-85, 2005.

[3] K. Sano, T. Iizuka and S. Yamamoto, "Systolic Architecture for Computational Fluid Dynamics on FPGAs", Proceedings of the 15th Annual IEEE Symposium on Fild-Programmable Custom Computing Machines (FCCM2007), 107-116, 2007.

[4] W. D. Smith and A. R. Schnore, "Towards an RCC-based Accelerator for Computational Fluid Dynamics Appilication", The Journal of Supercomputing, 30, 239-261, 2004.

[5] C. He and W. Zhao, M. Lu, "Time Domain Numerical Simulation for Transient Waves on Reconfigurable Coprocessor Platform", Proceedings of the 13th Annual IEEE Symposium on Field-Programmable Custom Computing Machines(FCCM'05), 2005.

[6] 佐野 健太郎, 王 陸洲, 初田 義明, 山本 悟, "アレイ構造に基づく差分法専用計算機の FPGA 実装", 第 13 回計算工学講演会, D-8-1, 2008.

[7] Altera Corporation, "Altera Product Catalog", http://www.altera.com/literature/, 2007.

表 2 現在利用可能な FPGA のハードウェア資源と予測される PE 数および要求帯域

シリーズ	Stratix III		Stratix II	Stratix II GX	Arria GX	Cyclone III	Cyclone II
FPGA	EP3SL340	EP3SE260	EP2S180	EP2SGX90F	EP1AGX90	EP3C120	EP2C70
LE 数	338 000	254 400	179 400	90 960	90 220	119 088	68 416
ALM 数	135 200	101 760	71 760	36 384	36 088	—	—
ALUT 数	—	—	143 520	72 768	72 176	—	—
メモリの総容量 [Kb]	20 491	17 876	9 163	4 415	4 477 824	3 888	1 125
DSP ブロック数	72	96	96	48	44	288	150
PE 数 N_{FPGA}	144	192	96	48	44	72	37
2 次元配列での要求帯域 w_{FPGA} [Gb/s]	152	176	124	88	84	108	155
LVDS チャンネル数	132	240	152	59	45	110	128
チャンネルあたりの通信帯域[Mb/s]	1 250	1 250	1 000	1 000	840	400	640
FPGAのI/O帯域 W_{FPGA} [Gb/s]	165	300	152	59	38	44	82