WIDE Technical-Report in 2010

Handmade WG 2010年度活動報告

wide-tr-handmade-report-2010-00.pdf

WIDE Project : http://www.wide.ad.jp/

Title: Handmade WG 2010 年度活動報告
Author(s): 島 慶一 (keiichi@iijlab.net), 中野 博樹 (cas@trans-nt.com)
Date: 2010-1-3

Handmade WG 2010年度活動報告

島 慶一* 中野 博樹†

概 要

本書はHandmade WGの2010年度の活動をまとめたものである。

1 Handmade WGとは

ネットワークに関する研究を進めていく上で、ハードウェア技術とソフトウェア技術の連携は欠かせない。ネットワーク技術の研究が始まった当初は、計算資源の制限などにより、ハードウェア技術を進展させる方向での研究が主なものであったが、計算機の驚異的な発展により、ネットワーク技術の分野に限らず、多くの処理をソフトウェア側で実行する形態が一般的となっている。

しかしながら、近年はその傾向が強すぎて、ソフトウェアに依存しすぎた技術開発をしてしまう場合が散見される。ハードウェア技術との連携を考慮することで、より優れたデザイン、もしくは、よりよい性能を獲得できる余地があるにもかかわらず、単にソフトウェアでも可能だから、あるいはハードウェア技術に詳しくないから、といった理由で連携を考慮しないのは、研究活動の幅を狭め、また本来獲得できたであろう価値を低減させかねない。

また、ソフトウェアのみでは実現不可能な分野も当然ながら存在する。近年、組み込み用のプロセッサが安価に入手できるようになり、一般的な計算機との連携動作も容易になってきている。なかでも、センサーやUI技術の分野では、これらの組み込み機器と、データ処理を行うソフトウェア技術の連携による面白い研究活動が多く見受けられる。

残念ながら、WIDEプロジェクト内でのハードウェアに関する技術力は、その分野に詳しかった年代の参加率の減少とともに低くなってきており、また、技術、知識の継承も十分に行われてこなかった。

一方、ソフトウェア技術に関しても、デバイスドライバー、OSなど、低レイヤーのソフトウェアに関する活動は年々減少してきており、ウェブアプリケーションやスクリプティングへの偏重が見受けられる。元々アプリケーション分野が不得意といわれるWIDEプロジェクトにとって、アプリケーション層の技術開発がプロジェクトにとって欠かせない重要な項目であることは言うまでもないが、それもソフトウェア階層の知識がなければ十分に活用することはできない。

Handmade WGは、内容の如何に関わらず、ハードウェア技術を活用してネットワーク技術の研究活動に役立つ為の基礎をお互いに学習する場、またソフトウェアのみの技術では解決が困難な課題をハードウェア技術を用いて解決できるようなアイデアを考えていく場、また新しいソフトウェアアーキテクチャを自らの手で考えていく場として2010年10月に設立された。

2 今年度の活動

2.1 WG立ち上げのための活動

Handmade WGに関係する活動が最初に行われたのは2010年8月に開催されたWIDEボード合宿である。この時点では、明確にHandmade WGの立ち上げや、具体的な活動について内容が決まっていた訳ではないものの、ハードウェアをソフトウェアの連携による研究活動に関する紹介が行われ、出席したボードメンバーおよび有志参加者にハードウェア関連技術への取り組みを思い出させるきっかけとなっている。本節では、ここで議論された項目のいくつかをかいつまんで報告する。

- 汎用部品を用いたネットワークボードの開発の現状

*株式会社IIJイノベーションインスティテュート
†株式会社トランス・ニュー・テクノロジー

一橋大学の台坂博氏から、「汎用部品を用いたネットワークボードの開発の現状」と題して天文学シミュレーションへの応用を念頭においたネットワークアダプタが紹介された。並列計算の分野では、大量のデータを複数の計算機の間でやりとりする需要が存在する。既存の高速ネットワーク技術を利用することはもちろん可能であるが、それらのデバイスは一般的に高価であり、また標準化された技術では現状 10Gbps 程度までの転送能力しか提供されていない。台坂氏は、パーソナルコンピュータに使われている汎用バス規格のひとつである SATA に注目し、それを複数束ねることによって既存のネットワークデバイスを上回るネットワークアダプタを提案している。また、アイデアの検証のために FPGA を用いたプロトタイピングを行っている。

- 汎用品を組み合わせたストレージ・ネットワークチューニング
  国立天文台の大江将史氏から、「汎用品を組み合わせたストレージ・ネットワークチューニング」と題して、高速な NAS を構築するためのチャレンジが紹介された。数種類のハードディスク、SSD、RAID カード、ネットワークアダプタ、マザーボードなど、利用する部品はすべて汎用品を用いる事で、全体のコストを下げつつ、それらが持つ性能を最大限引き出す組み合わせを模索した結果が報告されている。知見としては、単にデバイスを組み合わせただけでは、各々が持っている性能の合計性能を出せる訳ではなく、マザーボードの選定、RAID の構成方法などによって限界性能が変わってくる事があげられる。本チャレンジは継続中であり、今後も WIDE 研究会その他で報告が継続される予定となっている。

- メニーコアアクセラレータプログラミング
  会津大学の中里直人氏より、「メニーコアアクセラレータプログラミング」と題して、ハードウェアアクセラレーションを用いた並列計算プログラムを行う場合のテクニックが紹介された。近年、GPU などを用いた並列計算環境が整いつつあり、必ずしもハードウェアの知識がなくても莫大な計算資源を用いる事が容易になりつつある。しかしながら、ただ漫然とプログラムを書いているだけではアクセラレータが持つ本来の性能を引き出す事はほとんどできない。アクセラレータの構成を把握し、計算処理が均等にすべてのアクセラレータノードに分散するようなプログラムを記述することが重要である。プレゼンテーションでは AMD 社の GPU を例にとって効率的なプログラムを記述する方法が紹介された。

- IP 接続による密結合型分散コンピューティング
  慶應義塾大学の松谷健史氏からは、「IP 接続による密結合型分散コンピューティング」と題して、IP を計算機のバスとして用いる IP バス型計算機アーキテクチャの紹介と、その実証実験基盤となるプロトタイプが紹介された。本研究は慶應義塾大学村井研究室にて IP バスコンピュータとして研究が続いてきている内容の延長であり、その最先端の研究成果である。目標とするものは、アプリケーションに制約のない分散計算環境を提供し、自由にあるいは自動的にスケールアウトが可能な計算機アーキテクチャを定義することである。このような課題は、これまでの並列計算機の研究開発で長年議論された内容であり、ここではアーキテクチャの基礎通信技術として IP を用いた際の実現性の検証が行われている。

- 10GbE on FPGA
  東京大学の小泉賢一氏より、FPGA を用いた 10G スイッチのデザインとそのプロトタイプ実装の成果が披露された。高速なネットワークでは、長い時間単位では帯域以内のトラフィックしか流れていないように見えても、観測時間を短くすると一時的には帯域を越えるトラフィックが流れている事があることが分かる。このようなバーストトラフィックはパケットロスを引き起こす原因となり、結果的に全体のスループットを低下させてしまう。小泉氏の提案するスイッチでは、スイッチの出力ポートと入力ポートの間でフロー制御を行い、入力ポートから出力ポートへのデータの流れをシェーピングすることで、前述のバーストトラフィックを抑え、ワイヤーレートでの通信を実現している。また、そのアイデアを実証する FPGA ボードの試作機も紹介された。

ボード合宿では、既存のネットワーク機器にこだわらない、設計の本質を見据えたシステムデザインを志

す事で、研究品質が向上することが再確認された。また、WIDE プロジェクト内でのハードウェアを活用した研究の取り組みが近年著しく減少、また活用技術の練度も大きく後退している事が心配事項として挙げられた。これを受けて、WIDE 合宿などでの活動、および Handmade WG への設立へとつながっていくこととなった。

2.2 WIDE 合宿での集中議論

ボード合宿での議論を受け、ハードウェアとソフトウェアの融合についてより多くの WIDE メンバーに考えてもらうため、プレナリー枠にて議論を行った。本節では、WIDE 合宿で取り上げられたトピックをかいつまんで紹介する。

- 高速ネットワーク環境でのログ採取技術に関する研究経過報告
  東京大学の小泉賢一氏から、研究活動の途中経過報告として超高速ネットワーク環境でのログ採取技術に関する発表が行われた。

- NetFPGA の紹介
  慶應義塾大学の堀場勝広氏から、FPGA を用いてプログラムできるギガビットイーサネットカードである NetFPGA[1] が紹介された。NetFPGA には 4 ポートのギガポートと Xilinx 社 [2] の Virtex-II Pro 50 が搭載されており、ネットワークアダプタ、スイッチングハブ、ルータなどの内部構造を直接プログラムできる環境となっている。ハードウェアレベルでパケット処理できるため、ソフトウェアのみでは実現できないワイヤーレートでの転送が実現可能となっており、安価 (アカデミック価格で$499、一般向けで$1,199) でありながら十分な性能を引き出すことができる。

- DIY (handmade) HPC
  会津大学の中里直人氏より、「DIY (handmade) HPC」と題して発表していただいた。初期の計算機はほとんど自作のものであり、主となる研究目的を達成するために副次的にデザインされたものであったが、近年、計算機の汎用化が進み、計算機自体をデザインするということが少なくなっている。ここでは、中里氏の活動のひとつとして、4 倍精度演算機の設計と実装が取り上げられている。これは中里氏が関わっている天文学シミュレーションに必要な演算を実行する為に必要だった機能であり、この機能により本来の研究活動が加速されている。近年のコンピュータシステム設計における分業化傾向は、仕事の細かい分割と再割り当てを可能とし、業務効率は向上しているかもしれないが、最終的に得られる効果は必ずしも向上したかどうか十分考慮する必要がある、との問題提起で発表を締めくくっている。

WIDE 合宿では、ボード合宿での議論を受け、ボード合宿に参加できなかったメンバー間でその内容を共有し、さらに新しい事例を挙げながら、自分の研究の質を向上させるための手法のひとつとしてハードウェア活用の道を紹介した。この合宿での活動を受けて、Handmade WG が設立されている。

2.3 FPGA 勉強会の開催

WIDE プロジェクトには優秀なソフトウェア技術者が数多く在籍しているものの、ハードウェア記述言語に精通した技術者はあまり多くない。FPGA でのプログラムが Handmade のすべてではないものの、手法のひとつとして活用できる技能を身につけておくのは各自の今後の研究にとって有効と考え、FPGA 勉強会を開催中である。これまでに 2 回の勉強会を開催しており、今後数回に渡って継続して勉強会を開催する予定となっている。開催済みの会の概要を以下に示す。

- 第 1 回 ディジタル回路入門
  開催日: 2010 年 10 月 2 日
  講師: 慶應義塾大学 松谷健史
  概要: WIDE の研究者が主にソフトウェア研究者であることを踏まえ、FPGA プログラムに必要となる基礎的な論理回路の知識を復習した。WIDE メンバーのみの閲覧となるが、資料は `http://member.wide.ad.jp/~shima/tmp/handmade-study/handmade_01.pdf` から入手できる。

- 第 2 回 HDL 輪講とシミュレータ環境の構築
  開催日: 2010 年 11 月 6 日
  講師: 慶應義塾大学 松谷健史

概要: ハードウェア記述言語のひとつである Verilog と、Xilinx 社の評価ボード Spartan 3AN Starter Kit[1] を用い、実際の FPGA プログラム開発を実習形式で学習した。Handmade WG では、国立天文台の大江氏のご好意により、数枚の Spartan 3AN Starter Kit が貸し出し可能となっているので、興味のある WIDE メンバーの方は大江氏と連絡を取り、勉強会に参加してもらえたらと考えている。

第 3 回は後述する東京エレクトロン様のプライベートセミナーとして開催することとなっており、より細かい FPGA プログラミングの学習の場を提供予定である。

以降のスケジュールについては、Handmade WG の WIKI ページ `https://member.wide.ad.jp/wg/handmade/wide-confidential/wiki/` に随時最新情報を掲載する予定としているので、ご興味のある方はそちらを参照して欲しい。

2.4 Handmade Grand Challenge

2010 年 12 月の WIDE 研究会では、Handmade Grand Challenge として、今後ハードウェアアクセラレーションを利用して進めていきたい研究項目を自由に挙げてもらう時間を設けた。ここで挙がったアイデアは以下の通りである。

- Network Earthquake Generator
  FPGA を用いて、ワイヤーレートで任意のパケットを送出するネットワークトラフィックジェネレータのアイデアである。

- IMS IPsec packet snooper
  IMS システムの相互接続テストにおいて、IMS クライアントとサーバー間のシグナルメッセージの内容を確認することが重要であるが、通常この区間は IPsec を用いて暗号化されている。そのため、この間に割り込み、通信遅延を発生させない高速な IPsec 復号/符号機があると、テストの品質が飛躍的に向上すると考えられる。

- 6lowpan テストスイート
  WIDE プロジェクトでは、IPv6 の普及政策のひとつとして、TAHI プロジェクト [3] による IPv6 機器の仕様適合性テストスイートの提供を実施してきた。この活動は現在 IPv6 Forum[4] での活動と統合され、IPv6 機器の品質向上に貢献している。近年、センサーネットワークなどの分野で省電力無線機器を用いたノードの利用が増えてきている。これらの機器で IPv6 をサポートするための規格として IETF 6lowpan (IPv6 over Low power WPAN) WG がプロトコルを定めている。6lowpan が無線ネットワーク上での動作を規定する規格であるため、テストスイートも無線ネットワーク上でテスト対象クライアントと接続する必要があるが、現在のところ無線上で動作する検証システムが存在しない。6lowpan 機器のテストスイート作成協力者が広く募集された。

2.5 FPGA 活用への展望

FPGA を使ってネットワークを流れるパケットの処理を行うことは、随分前から行われてきたが、ネットワークにおける FPGA の活用について基本に立ち返って議論し、研究を進める時が今こそ来ていると考える。ここでは、FPGA に関連する現在の状況を簡単にまとめた後、この領域において FPGA の利用を進めていく上での課題を提起し、これを解決する 1 つの提案として、ソフトコアネットワークプロセッサを活用して、論理回路と汎用プロセッサとの組み合わせと使い分けを提案する。

全世界におけるインターネットの利用者の増加に加え、動画、特に HD や 3D といった映像技術の進化に伴ない、インターネットが扱うデータ量が加速度的に増加している。100Gbit Ethernet の実用化やハードディスクの記録密度向上と大容量化の順調な進展などにより、通信路やストーレジの進化は進んでいるが、ルータを始めとするネットワーク内のデータプロセッシング機器の高速化では、単体のプロセッサの速度向上が頭打ちになっていることから新たなアプローチが求められている。例えば、マルチコアやメニーコアといった複数のプロセッサを並列に並べる試みが行われている。また、専用回路を組むことによる非ノイマン型のアーキテクチャも多数試みられている。

[1] http://www.xilinx.com/products/devkits/HW-SPAR3AN-SK-UNI-G.htm

データプロセッシングのために専用に論理回路を設計する場合、ASIC の開発と FPGA の利用の 2 種類の方法が考えられる。ASIC 開発の場合、費した費用に応じた性能を得ることができるが、コスト面の負担が大きい。すなわち、プロセス技術の進化に伴い、マスク代や設計・検証・試験のコストが急増しており、高性能な ASIC の製造には高額の投資が必要になる。一方、近年の FPGA の利用技術の発展には注目すべきものがある。FPGA は標準品として設計製造されるため十分な開発費を費すことができ、均質な構造を持つ特性を利用して最先端のプロセス技術が適用されている。この結果として、価格対性能比において FPGA が ASIC を上回るケースが見受けられるようになった。それに加えて、FPGA には、容易に書き換えられるという性質が備わっているため、FPGA を採用する優位性は多くの場合において現実のものになりつつある。

ところで、ASIC や FPGA によるパケット処理は特に目新しいものではない。今までも多くの研究が行われており、高速ルータなどにおいてはその成果を専用 ASIC として搭載しているものも多い。特に、ワイヤスピードの確保など、確定的に動作時間を設計し性能を保証する必要がある場合においては、汎用プロセッサでは対応できず、専用回路を設計する場合が多い。この場合における設計コストの低減はネットワーク機器の設計において重要である。しかしながら、専用回路によるネットワーク機器の設計に広く普及している手法はなく、専ら各社の努力に依存している。

データプロセッシングの高速化という要求に対してはネットワークプロセッサも多く利用されている。ネットワークプロセッサは様々なアーキテクチャが考案され研究されており、まとめて論じるのは難しいが、その多数に共通する特徴がいくつかある。一つには基本的な動作がイベント駆動処理であることである。ネットワークアプリケーションはパケットの到着に同期して処理を行うため、汎用プロセッサを用いると割り込み処理やポーリング処理が必要になり効率が悪い。ネットワークプロセッサでは、実行ユニットを複数用意するなど、イベント駆動処理に適した工夫がされていることが普通である。また、他の特徴として、比較的低い周波数のクロックでも十分な性能が得られることがある。パケット処理の特性として、単位時間当たりのデータ転送量が多いことが挙げられるが、この点を専用のバッファ管理回路によって克服し、パケット毎に必要な処理の並列化と絞り込みによって、400〜500MHz 程度でも 10Gbps のデータを処理することができる。以上に挙げた点は、FPGA の特性とよく適合する。昨今の微細化により有り余るゲート数を備えた FPGA にとって実行ユニットを多数搭載することは簡単である。また、動作速度としても、FPGA にとって実現可能なクロック周波数となっている。これらのネットワークプロセッサの特徴は、それを FPGA 上に実装することを後押しするものであると考える。

2010 年 12 月の WIDE 研究会では、ネットワークを流れるパケットを FPGA 上で処理するアーキテクチャとして、プログラマブルな三層の構造が株式会社トランス・ニュー・テクノロジーから提案された。第一の層は、Verilog や VHDL といったハードウェア記述言語による論理回路である。記述された論理回路は FPGA 上にプログラムする。第二の層は、ネットワークプロセッサによるプログラムである。ネットワークプロセッサは FPGA 上に構成し、その上でネットワークプロセッサに特化したソフトウェアを実行する。最後の層は、汎用プロセッサによるプロトコル処理である。プロトコルの最も複雑な部分の処理のためにこの層を利用する。これら各層は、得意とする処理の内容が異なり、処理内容に対して設計段階において可能な保証も異なる。また、プログラム書き換えにかかる時間やコストも異なるため、これらを総合して、ネットワーク機器としての機能を適切に分割し、実装する必要がある。

また、研究会では、トランス・ニュー・テクノロジーの技術者から、前述のアーキテクチャの解説と実装サンプル (図 1) が紹介された。紹介された実装では、論理回路の層として、ARP テーブル・経路表の検索やカウンタを用いた統計情報処理を用意し、その上のネットワークプロセッサ層において IPv4、IPv6 のフォワーディングを行い、汎用プロセッサ層では NetBSD[5] と経路制御プログラムを動作させることによって IP ルータが FPGA 上に構築されている。

Altera 社 [6] の Stratix III[2] 上に、論理回路によるいくつかの機能モジュールと 1 つのネットワークプロセッサ IP、1 つの Nios II 汎用プロセッサ[3] を実装し、3.05M パケット /秒の性能を得た。現在入手可能な最大サイズ

[2] http://www.altera.com/products/devices/stratix-fpgas/stratix-iii/st3-index.jsp

[3] http://www.altera.com/products/ip/processors/nios2/ni2-index.html

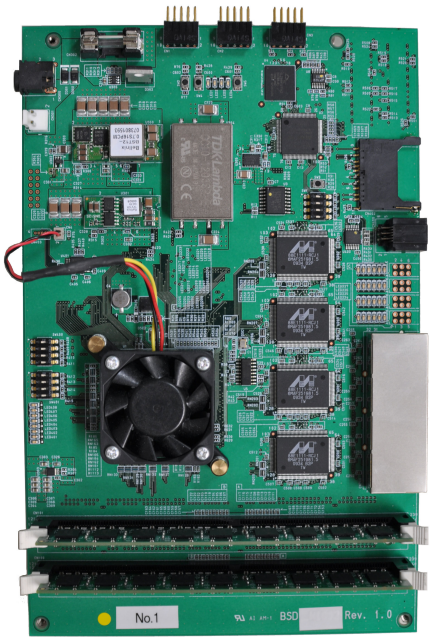

図 1: プログラマブル三層構造 FPGA ボードの試作機 (トランスニュー・テクノロジー社提供)

の FPGA は、今回サンプル実装で採用された FPGA の 10 倍程度の規模を持っているので、機能モジュールおよびネットワークプロセッサの数を増やして並列に処理できるパケット数を増やすことにより、この数倍の性能の実現が可能であると考えられる。

また、この実装の過程において、ルータに必要な各機能を実現する層を変更することによって容易に性能やコストの調整ができることも確認できた。今後、このアーキテクチャに従った設計手法を体系的に整理して、広く利用できるようにしていきたいと考えている。

2.6 Xilinx プライベートセミナーの開催

今後の FPGA でのプロトタイプ実装に備え、現在利用している Xilinx のボードをより深く学習する場として、東京エレクトロン様の協力のもと 2010 年 12 月 20 日にプライベートセミナーを開催した。本セミナーに関しては、WIDE プロジェクトの研究者を対象に、今後不定期に開催する予定となっている。

3 まとめ

7 月のボード合宿での議論に端を発する、自家製の (handmade) 研究機材による高品質な研究活動の実現に向け、Handmade WG が立ち上げられた。今年度は、自家製機材を活用して研究成果をあげてきた先人の経験やノウハウを共有し、今後の活動に向けての準備をする段階としての活動を行った。具体的には、2.2 節にて報告した WIDE 合宿での議論、また 2.3、2.6 節で報告した FPGA 勉強会などを実現し、12 月の研究会においては、実際に FPGA 製品の開発を行っている技術者から、FPGA を活用した研究の今後の方向性についての専門家としての観点から意見をいただくことができた (2.5 節)。

実際にハードウェアアクセラレーション機能を活用した研究活動を始めた研究者の支援と議論の場、WIDE 内から沸き上がったハードウェアアクセラレーションのアイデアを実現していく場、また基本的な知識を身につける学習の場としての活動を続けていく予定である。また、自家製という意味では、ハードウェアのみにとらわれる事なく、研究を促進するための自家製ソフトウェアなどの側面にも目を向けつつ、個々人がそれぞれのネットワーク研究活動の質を向上させるために役立つ WG を目指す。

参考文献

[1] NetFPGA. NetFPGA, December 2010. http://netfpga.org/.

[2] Xilinx, Inc., December 2010. http://www.xilinx.com/.

[3] WIDE project. TAHI project, December 2010. http://www.tahi.org/.

[4] IPv6 Forum, December 2010. http://www.ipv6forum.org/.

[5] The NetBSD Project, December 2010. http://www.netbsd.org/.

[6] Altera Corporation, December 2010. http://www.altera.com/.