報道各位

2017年4月25日
株式会社インプレスR&D
http://nextpublishing.jp/

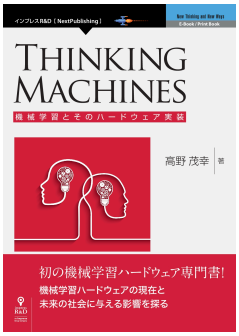
日本初の機械学習ハードウェア専門書！

「Thinking Machines 機械学習とそのハードウェア実装」発行

最先端開発動向と技術を解説

インプレスグループで電子出版事業を手がける株式会社インプレス R&D は、『Thinking Machines 機械学習とそのハードウェア実装』(著者:高野 茂幸)を発行いたしました。

『Thinking Machines 機械学習とそのハードウェア実装』
http://nextpublishing.jp/isbn/9784844397694



著者:高野 茂幸
小売希望価格:電子書籍版 1300円(税別)／印刷書籍版 1800円(税別)
電子書籍版フォーマット:EPUB3／Kindle Format8
印刷書籍版仕様:B5判／モノクロ／本文178ページ
ISBN:978-4-844397694
発行:インプレスR&D

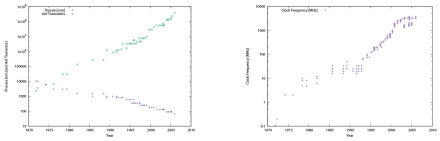
＜＜内容紹介＞＞

本書は、急速に研究が進んでいる機械学習専用ハードウェアについて、その開発方法や各研究機関の動向、各国の概要をまとめた日本初の書籍です。

機械学習の基礎的な情報だけでなく、ハードウェア実装する際のポイントなどを掲載。付録では機械学習の技術的な側面に加えて社会に与える影響についても考察しています。

機械学習の最新動向を把握できる開発者必読の一冊です。

(本書は、次世代出版メソッド「NextPublishing」を使用し、出版されています。)



(a) Process and Number of Transistors. (b) Clock Frequency.

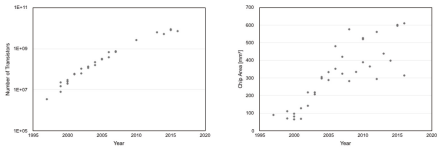
図 2.2 History of Intel Microprocessors.

2.1.2 GPU の計算機システムへの利用

GPU 以前のグラフィックス処理回路は、最初から固定設計されたグラフィックス処理専用の回路だった。90 年代当初は二次元画像処理用の回路であったが、マルチメディアアプリケーションの登場とともに三次元グラフィックス処理や精細なグラフィックス描画への要求が強くなった。グラフィックス専用回路へプログラミング性が導入されたことにより、グラフィックス・プログラミングが進展し、グラフィックス専用回路の利用者が自分のグラフィックス・アルゴリズムを実装できるようになり、固定されたグラフィックス・パイプライン回路からプログラム可能なグラフィックス処理プロセッサ（Graphics Processing Unit）へと発展したのである [190]。

ポリゴン系グラフィックス処理は、頂点座標データと最低 3 頂点で構成される面のマップデータで構成され、データが膨大である。しかし各ピクセル単位での処理は同じであり、三次元グラフィックスを二次元画面上にマッピングする過程まで同じ処理を繰り返しているので、データレベル並列性（DLP; Data-Level Parallelism）を得られやすい。この GPU の持つ DLP 特性を一般的な他の DLP 特性を持つアプリケーションへ利用しようと試みる動きが現れ [113]、今日の汎用処理用グラフィックス処理ユニット（GPGPU; General-Purpose GPU）や GPU Computing[190] と呼ばれる汎用処理への GPU 利用が進んだ経緯がある。この流れで、HPC（High-Performance Computing）用として利用が進んでおり、GPU への倍精度浮動小数点演算の実装により HPC 開発では GPU の採用が一般化している。

図 2.3 は GPU の実装傾向を示す。トランジスタ数は 2010 年頃まで対数スケールで上昇しているがその後の増加ペースは若干落ちている。そのトランジスタ数を実装するためのチップ面積は $600mm^2$ を超えている。チップ面積が $600mm^2$ は大体 24mm 角チップであり、一般的な 10mm 角チップサイズを大きく上回り、エンタープライズ用メニーコアプロセッサと同等のサイズまで大きくなっている。なお、GPU のアーキテクチャに関しては第 4 章で説明する。



(a) Number of Transistors. (b) Chip Area.

図 2.3 History of Graphics Processing Units.

2.1.3 FPGA の計算機システムへの利用

他方、LSI 製造前の動作検証目的で発明された FPGA（Field-Programmable Gate Array）デバイスは近年最終製品へ利用され始めている。FPGA は求められる論理回路を自由に構成できるデバイスで、一般にそのデバイスに記憶する構成プログラム（Configuration Data、構成データという）を変更すれば何度でも再構成が可能である。設計から市場投入までの期間（一般に Time-to-market という）を短縮し [19]、かつ多品種少量生産が主流になっている現在、特定用途向け LSI である ASIC の NRE（Non Recurrent Engineering）、実装、そして製造コストが見合わないため、FPGA の最終製品への利用が進んでいる。また、従来は計算機のソフトウェアで行われているバグに対処するためのパッチ（修正プログラム）当ても FPGA 上に構成された回路で可能になるので、些細な回路のバグであれば市場投入後に構成データを差し替えることで修正できるメリットもある。

現在の FPGA はメモリブロック、積算器と加算器（FPGA ではこの 2 つの演算器の組み合わせ回路を DSP と呼ぶ）を多数実装して、単一チップでのシステム（SoC; System-on-Chip という）実装に対応している。メモリブロックは外部メモリへのアクセス回数を減らす効果とアクセス遅延による待ち時間を削減する効果があり、このメモリブロックを大量にチップ上に実装することで、複数データへの同時並行アクセスを可能にし、低遅延なデータの並列処理による高性能化を実現している。また、DSP の実装により、デジタル信号処理（DSP; Digital Signal Processing）でよく利用される整数積算と整数加算をクロック周波数の低下を抑えた上で効率的に処理できるようにしている。現在の DSP は整数演算と固定少数点演算以外に単精度浮動小数点演算に対応できる回路構成になっており用途が広がっている [37]。

FPGA は汎用デバイスであるため、ASIC による実装と比較してトランジスタ数換算値（ゲート相当数という）、クロック周波数、そして消費電力それぞれの能力が低い（表 2.1 参照）。ベンチマーク回路ではゲート相当数、クロック周波数、消費電力がそれぞれ 3 ～ 4 世代

GPU と FPGA について、機械学習への利用を解説

おける電位を $V^{(j)}_{neuron}(t)$、先行接続しているニューロン i の時刻 t における電位を $V^{(i)}_{neuron}(t)$、特定ニューロン j と先行ニューロン i の間のシナプスに到着するそれぞれのニューロンからの電位を $V^{(j)}_{neuron}(t+\delta^{(j)}_i)$ と $V^{(i)}_{neuron}(t+\delta^{(i)}_j)$ とする。$\delta^{(j)}_i$ と $\delta^{(i)}_j$ はそれぞれのニューロンから、その特定シナプスまで到着するのに要する遅延時間である。特定ニューロン j の発火条件を定義する活性化関数を $f^{(j)}(*)$、そのシナプスの実効強度を時刻 t において $w^{(j)}_i(t)$ とすると、ニューロン j へ流入するイオン物質の総量 $I^{(j)}(t)$ は以下の様になる。

$$I^{(j)}(t) = \sum_i w^{(j)}_i(t-\delta^{(j)}_i)\, V^{(i)}_{neuron}(t-\delta^{(j)}_i-\delta^{(i)}_j) \tag{3.1}$$

このイオン物質の総量 $I^{(j)}(t)$ により、ニューロンの発火条件を定義している活性化関数 $f^{(j)}(*)$ による時刻 t におけるニューロン j の電位 $V^{(j)}_{neuron}(t)$ は次の様に書ける。

$$V^{(j)}_{neuron}(t) = f^{(j)}(I^{(j)}(t)) \tag{3.2}$$

$V^{(j)}_{neuron}(t) \gg 0$ の時、ニューロン j は時刻 t で発火して活動電位を生成した事に相当する。シナプスにおける実効強度 $w^{(j)}_i(t)$ は学習係数 $1 > \epsilon > 0$ と実効強度の更新量である電導率の変化量 $S^{(j)}_i(t)$ を使用して以下の様になる。

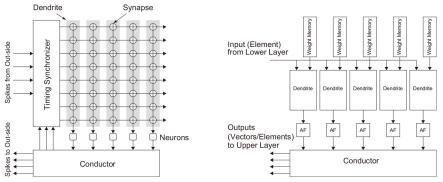
$$w^{(j)}_i(t+\delta_t) = w^{(j)}_i(t) - \epsilon S^{(j)}_i(t) \tag{3.3}$$

$\delta_t > 0$ は考慮可能な時間の最小単位とする。以上のことから脳の学習はニューロンのシナプスの実効強度、つまり電導率である w を更新することに相当する。

3.1.2 ニューロモルフィックコンピューティングハードウェア

ニューロモルフィックコンピューティングのハードウェア構成を図 3.2(a) に示す。一般にニューロモルフィックコンピューティング・アーキテクチャは活動電位を単位の大きさとした、ある時刻に発火したパルス波として STDP を模倣している。ニューロンは複数のシナプスで構成された Dendrite 部と発火を行う Neuron 部で構成され（この構成を一般に Soma という）、複数の Soma をクラスタ化し（図 3.2(a) では 5 つの列としてクラスタ実装している）、入力したスパイクを各 Soma で共有する構成を取って複雑なグラフを構築できるようにしている。

入力スパイクは Soma への入力タイミングが Timing Synchronizer で調整された後、複数のシナプスで構成されたアレイに送信される。各ニューロンは Dendrite からの出力値を入力して何らかの条件を満たした時にスパイクを発火させる。発火したスパイクは Conductor で収集され、出力できるデータフォーマットに整えた後に出力、あるいは自分のこのユニットにフィードバックする。このユニットをコア・モジュールとしてチップ上に複数個これを並べて実装して全体的な脳を構成する。



(a) Neuromorphic Computing Hardware Basic Structure. (b) Neural Network Hardware Basic Structure.

図 3.2 Basic Structure of Machine Learning Hardware.

このモデルの小分類としてその実装方法から二つに分類できる。入力スパイクと Soma 内のシナプスの電導率 w の積、その Soma における積算値の総和の仕方（つまり内積演算の仕方）が従来のデジタル回路実装に基づいたモデル（Digital Logic Circuit）と、アナログ回路実装に基づいたモデル（Analog Logic Circuit）である。

1. デジタル回路（Digital Logic Circuit）

   Dendrite 内の各シナプスはクロスバー・スイッチで実装され、図 3.2(a) の Dendrite アレイは全体としてクロスバー・スイッチアレイを構成する。任意の時刻に一つの入力スパイクを Timing Synchronizer から特定のクロスバーの行へ入力し、シナプスであるスイッチが “ON” の時にそのクロスポイントをスパイクが通過してニューロンへ入力される。

   複数の Dendrite はその入力スパイクを共有しているので、各 Dendrite においてそのスパイクは適宜クロスポイントを通り特定の一つ以上のニューロンへ入力される。各 Dendrite において一つの入力スパイクのみクロスバーに入力されるので、特定のニューロンが先行接続している N 個のニューロンと接続している場合、N 個のスパイクをクロスバーに入力させてニューロンは Dendrite からの入力値を累算（Accumulation）する必要があるので、ニューロンへ流入するイオン物質の総量 I を得るまで N ステップを要する。

   また、活動電位の生成はイオン物質の総量 I に基づいて発火するように論理演算回路で実装されている。シナプスがクロススイッチで構成されているので、シナプスの伝導率はスパイク同様に単位値であり実効強度を表していないので、それに合わせて Neuron の発火条件を調整する仕組みが必要である。さらにクロックを利用する（同期回路）

ニューロモルフィックコンピューティングでのハードウェア構成

62 第 4 章 機械学習ハードウェア

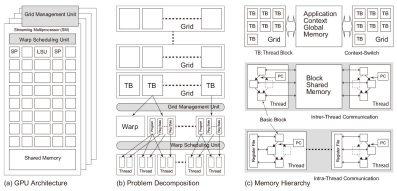


図 4.2 GPU Architectures.

(Streaming Processor) と呼ぶスーパースカラプロセッサで構成されており (ベクトル演算型プロセッサではない)、従来のコンパイラを利用可能かつ相対的に簡素なスケジューリングで十分なスカラ演算を行う。

シェーディングやバーテックス処理等、同じタイプのスレッド[*2]は、thread block と呼ぶ単位にまとめて管理される。SM コントローラは thread block 中 warp とよぶ最大 32 スレッドの集合で各スレッドを各 SP に割り当てる。SM 中の各スレッドは個々の命令アドレスとレジスタ状態を持ち独立して実行される。つまり、thread block 内のスレッドは同じプログラムであるが、その制御フローは個別である。また、バリア命令をサポートしており、スレッド間の同期がとれるようにしている。

NVIDIA 社では同じ命令列で構成された複数のスレッドがそれぞれ個別に実行するこの仕組みを SIMT (Single Instruction, Multiple Threads) と呼んでいる。SIMT 方式により、プログラマは最大 warp 数といった SIMT を基にしたハードウェア構成を意識する事なくプログラミングが可能になっている。GPU プログラムは Grid と呼ぶ複数の thread block 単位で管理し、一つの thread block を一つの SM へ割り当て、SM は thread block のスレッドと個々のスレッドのレジスタステートを各 SP へ割り当て、各 SP は同じ命令列を個別の制御フローで実行する。

さらに最近ではキャッシュメモリを増強してメモリアクセス周りを強化したり、順不同 (Out-of-Order) での thread block 実行のサポート、warp スケジューラの多重化、スレッド切り替え (context-switching) を高速化している [24][25]。2016 年、積層型メモリである HBM2 (High Bandwidth Memory) が採用され、深層学習でよく使用する半精度浮動小数点ユニットなどを実装した Pascal アーキテクチャが発表された [52]。HBM2 は帯域/消費電力比の改

[*2] プログラムの実行単位であるプロセスを細分化した、プロセスの状態情報を共有するプログラムの最小単位を一般にスレッドと呼ぶ。

4.1 実装プラットフォーム 63

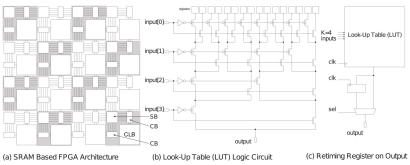


図 4.3 SRAM Based FPGA Architectures.

善と実装面積の最小化に寄与している。

4.1.4 FPGAs; Field-Programmable Gate Arrays

図 4.3 (a) は SRAM を回路の構成データを保持する記憶素子とした Field-Programmable Gate Array (FPGA) の基本アーキテクチャを示す [78]。論理回路の組み合わせ回路を構成するための回路 (CLB; Configurable Logic Block)、CLB 間を接続する相互接続網である CB (Connection Block) と SB (Switch Block) で構成される。CLB と複数の CB、SB をまとめて一つのタイルを作り、このタイルを並べて全体を構成する。ここでは利用者が設計した回路をユーザー回路と呼ぶ事にして FPGA 自身の回路と区別する。

CLB は K 個の 1-bit 入力の真理値表に相当し、図 4.3 (b) に示すように一般に LUT (LookUp Table) と呼ぶマルチプレクサで構成されている [90][89]。従って LUT は K-bit の入力から 2^K 個の 1bit レジスタのデータを選択する。LUT の出力と前クロックサイクルでの出力値を保持するレジスタ (retiming register と言う) から選択できるようになっている (図 4.3 参照)。Retiming register の値をユーザー回路にフィードバックすることで有限状態機械 (FSM; Finite State Machine) 等のユーザー回路を実装できるようにしている。CLB の入出力は相互接続網の配線に CB を経由して接続されており、横方向の配線は縦方向に、縦方向の配線は横方向に SB を経由して方向転換する。複数の配線はトラックを構成し、複数種類の配線長で構成されたトラックがあり、信号の送信先までの長さに合った配線を利用する事で配線の使用効率を上げ、かつ無駄に配線遅延が発生しないようにしている (この方法を Channel Segmentation Distribution と言う [78])。Virtex アーキテクチャ以降ではこのタイル型アーキテクチャを改良し、タイルが構成する行列について CLB、DSP、RAM ブロックといった複数の種類の内一つの要素で構成した列を用意して適宜並べるレイアウトを行っている。

図 4.4(a) は Xilinx のシングルチップ FPGA 製品の持つ LUT 数を、一番左を XC2000、一番右を Virtex7 としてプロットしている (時間軸のプロットではない事に注意が必要である。ま

FPGA での機械学習ハードウェアの基本設計と性能指標

＜＜目次＞＞

＜＜著者紹介＞＞

高野 茂幸(たかの しげゆき)

2008 年三洋半導体株式会社に入社。その後株式会社ドワンゴを経て、現在は電気・電子機器メーカー研究所勤務。業務としてデジタル信号処理プロセッサーの設計開発からビデオトランスコーダーの開発などの傍ら、次世代プロセッサーの一つである自立型再構成可能プロセッサーの個人研究に携わる。最近は機械学習プロセッサーを個人研究中。

＜＜販売ストア＞＞

電子書籍:

Amazon Kindle ストア、楽天 kobo イーブックストア、Apple iBookstore、紀伊國屋書店 Kinoppy、Google Play Store、honto 電子書籍ストア、Sony Reader Store、BookLive!、BOOK☆WALKER

印刷書籍:

Amazon.co.jp、三省堂書店オンデマンド、honto ネットストア、楽天ブックス

※ 各ストアでの販売は準備が整いしだい開始されます。

※ 全国の一般書店からもご注文いただけます。

【株式会社インプレス R&D】 http://nextpublishing.jp/

株式会インプレス R&D（本社：東京都千代田区、代表取締役社長：井芹昌信）は、デジタルファーストの次世代型電子出版プラットフォーム「NextPublishing」を運営する企業です。また自らも、NextPublishing を使った「インターネット白書」の出版など IT 関連メディア事業を展開しています。

※NextPublishing は、インプレス R&D が開発した電子出版プラットフォーム(またはメソッド)の名称です。電子書籍と

印刷書籍の同時制作、プリント・オンデマンド（POD）による品切れ解消などの伝統的出版の課題を解決しています。これにより、伝統的出版では経済的に困難な多品種少部数の出版を可能にし、優秀な個人や組織が持つ多様な知の流通を目指しています。

【インプレスグループ】 http://www.impressholdings.com/

株式会社インプレスホールディングス(本社：東京都千代田区、代表取締役：唐島夏生、証券コード：東証１部 9479)を持株会社とするメディアグループ。「IT」「音楽」「デザイン」「山岳・自然」「モバイルサービス」を主要テーマに専門性の高いコンテンツ＋サービスを提供するメディア事業を展開しています。

【お問い合わせ先】

株式会社インプレス R&D NextPublishing センター
〒101-0051　東京都千代田区神田神保町 1-105
TEL 03-6837-4820
電子メール: np-info@impress.co.jp