Tech Insights

MICROPROCESSOR report

Insightful Analysis of Processor Technology

AIのレイテンシを低減するEDGECORTIX IP

スタートアップ企業によるハイエンドなDNA IP、レベル3 ADASをターゲットに

著：Aakash Jani (April 26, 2021)

翻訳：EdgeCortix

EdgeCortixは、ハイエンドのエッジ市場をターゲットとしたカスタムIPで、デュアルアクセラレータアプローチによる業界最高レベルの効率性を実現しています。シミュレーションでは、同社のIPは、ResNet-50およびYolo v3において、レベル3のADAS(先進運転支援システム) アクセラレータで、バッチ =1 の性能でトップスコアを獲得しており、同様に優れた電力効率も実現しています。顧客にはまだ公表していませんが、IPとFPGAのビジネスで利益を上げています。

この性能を実証するため、同社はテストチップの開発を進めており、先日のLinley Spring Processor Conferenceで発表したように、8Wの消費電力で54TOPSを実行できると推定しています。800MHzのクロックと8つのコンボリューション・エンジンを使うことで、ピークスループットを達成するように設計されています。このチップはTSMCの12nmプロセスで製造され、EdgeCortixは、来年早々に最初のシリコンを出荷する予定です。

同社は、東京に本社を置き、サンフランシスコにデザインセンターを構えています。米国およびアジアの投資家を対象にした2回の資金調達ラウンドを行いましたが、わずか600万米ドルしか調達できていません。利益は出ていますが、より早く成長させるために、現在シリーズA資金を調達しています。EdgeCortixは、IBM ResearchでAI-optimizationチームを率いていたCEO兼創設者のサキャシンガ・ダスグプタを筆頭に、約20名の従業員を抱えています。

EdgeCortix は IP ライセンサーとしてスタートし、Xilinxの FPGAを組み合わせたDNAブランドのディープラーニングアクセラレータ (DLA) IPを、自動車、ドローン、ロボット市場における初期顧客向けに提供していました。IPは柔軟にコンフィギュレーションすることが可能ですが、表1に示すように、同社では設計済みのソリューションを顧客に提供しています。DNA-A050は、1つのコンボリューション・エンジンで1.8TOPSの性能を発揮する製品です。その上には、A100、A200、A400、A800があります。

IP事業と並行して、収益性の高いFPGA事業も展開しており、300MHzで動作するFシリーズで、組み込み（25W以下）およびエッジサーバ（50W以上）市場向けに設計をしています。このシリーズは、F200とF400の2つのラインアップを除き、同社の他のIPシリーズと同じものです。v1は64x64のシストリックアレイでコンボリューションエンジンの数が少なく、v2は32x32のシストリックアレイでコンボリューションエンジンの数が多くなっています。そのため、後者の方が小規模なネットワークに対してより高い並列度を実現することができます。他の多くのIPベンダやAIチップのスタートアップ企業とは異なり、Edge-Cortixは最近、Xilinx Alveo U50 FPGAカードを使用したDNA F200のMLPerf 1.0の結果を掲載し、そのソフトウェアスタックの堅牢性と達成した低レイテンシを公表しています。

	DNA-A800	DNA-A400	DNA-A200	DNA-A100	DNA-A050
Systolic Arrays	8 arrays	4 arrays	2 arrays	1 array	1 array
Internal Memory	30MB	16MB	10MB	9MB	3.5MB
MACs per Cycle	67,000	34,300	17,100	8,600	2,200
Peak INT8 Perf*	54.0 TOPS	26.0 TOPS	13.0 TOPS	6.5 TOPS	1.8 TOPS
ResNet-50*†	2,500 fps	2,000 fps	1,250 fps	571 fps	159 fps
Yolo v3*†	555 fps	333 fps	167 fps	83 fps	27 fps

表１：EdgeCortix DNA IP

DNA IPポートフォリオは、AIのトータル性能が大きく異なる製品を提供しています。対応範囲が広いことにより、組み込み型およびエッジサーバー型のAIアプリケーションをターゲットにすることができます。

*800MHzで †バッチサイズ1での推論スループット

(出典: EdgeCortix)

コンボリューション・エンジンの省電力化

EdgeCortixのIPはDNAだとすれば、その塩基対となるものは互いに補完してアクセラレータを形成する二つの異なるエンジンです。

同社が標準的なコンボリューション・エンジンと呼ぶものは、処理要素（PE）のシストリックアレイを採用しており、3x3から64x64のアレイに拡張可能です。大きなモデルを使用するライセンシーは、コンボリューションを分割する無駄を省くために、より大きなアレイを選択することができます。しかし、64x64の構成を超えると、シストリックアレイは十分に活用されなくなり、多くのモデルに置いての性能向上は収穫逓減になる恐れがあります。EdgeCortixは、各PEに2つのウェイトレジスタを併せることで利用率を高めています。配列の各要素は列・行ごとに合計され、最後に累積されます。

同社がDepthwise畳み込みと呼ぶ別のエンジンでは、畳み込み演算の処理にベクトル法を用いています。これらのエンジンは、図1に示すように、通常の畳み込みの3Dカーネルではなく、2Dカーネル（ベクトル）を使用したDepthwise畳み込みを直接実装しています。Mobile-NetやShuffleNetのようなニューラルネットワークモデルでは、精度を落とさずにモデルパラメータを削減するためにこの技術を採用しています。その他のモデルでは、3Dカーネルを2Dスライスに変換し、コンボリューション・エンジンにフィードすることができます。

外部の重み値は畳み込みユニットに入り、重みバッファに格納され、外部の特徴値は活性化バッファに格納されます。特徴値と重み値は、その後、活性化バッファと重みバッファの異なるセットに入り、加算器ツリーにフィードされて、部分和を生成します。DNA IPは高度なコンフィギュレーションが可能です。ライセンシーは、パイプライン幅と並列パイプラインの数を増やすことができますが、前者ではかなりのパフォーマンスの低下が生じます。乗算ユニットが増えるごとに加算器ツリーも大きくなり、パイプラインのサイクル効率が低下します。また、コンボリューション・エンジンの幅が広くなると、より小さなフィルターやカーネルサイズの利用率が低下し、電力効率や面積効率が低下します。このため、コンボリューション・エンジンは通常、最も一般的なタイプである3x3カーネルに最適化されています。DNAの両エンジンとも、INT32 アキュムレータを使用した符号付きおよび符号なし INT8 乗算アキュムレート演算をサポートしています。

ベクターユニットには、プーリングエンジン、アクティベーションエンジン、スケーリングエンジンがそれぞれ複数搭載されており、その数は自由に設定可能です。プーリングエンジンは最大プーリングのみを処理し、アクティベーションエンジンはReLU、leaky ReLU、H-swish、ReLU6など、より多くのアクティベーション関数を高速化することが可能です。そして、スケーリングエンジンはバイリニアのアップサンプリング、およびダウンサンプリングを処理し、それぞれ線形補間により画像を拡大・縮小します。 アクティベーションエンジンとスケーリングエンジンのパラメータ値は、各サブユニットに存在し、シストリックアレイまたはコンボリューション・エンジンのいずれかから送られます。

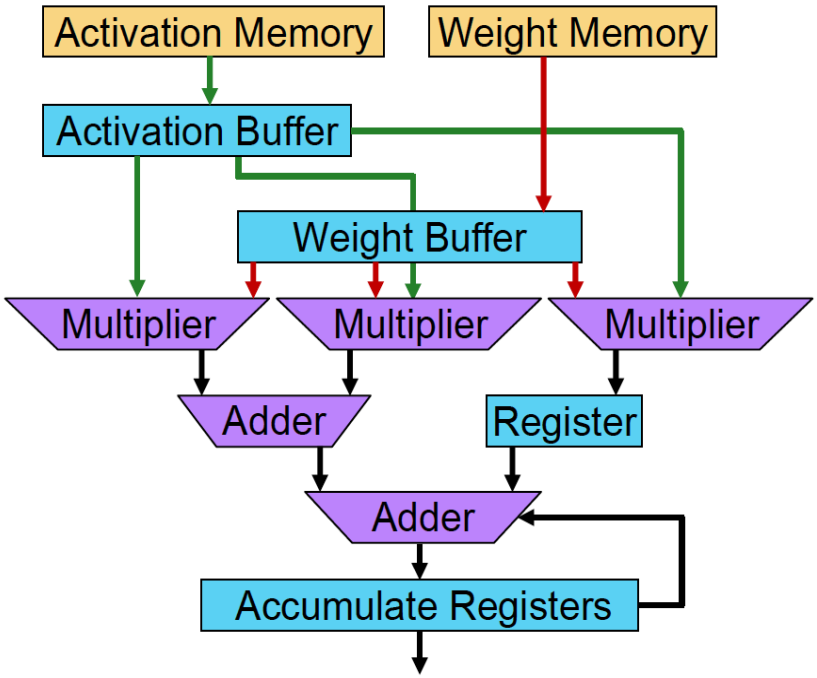

図１：コンボリューションエンジン

図では、1本のパイプラインで3ワイドのコンボリューションエンジンを表現しています。

ライセンシーがパイプラインの数やパイプライン幅を増やすことは可能ですが、後者は加算器ツリーを深くする必要があり、効率が落ちます。

柔軟なNoCでレイテンシーを半減

DLAがニューラルネットワークを処理する際に、チャネル数とマトリックスのサイズが異なる層に遭遇します。固定されたハードウェアの構造は、層ごとにコンピューティング使用率が異なります。EdgeCortixは、再構成可能なインターコネクトとメモリ構造によってポートフォリオを差別化するという機会に飛びつきました。

レイヤーごとに、DNAコアは回路切り替えを使ってデータフローを変更し、MACの利用率を最適化することができます。回路切り替えは、マクロスコーピック(巨視的)な対応と同様に、コンパイル時に各層に設定される「セレクト」信号に基づいて、さまざまなエンドポイントにデータを転送します。このアプローチにより、再構成可能なNoC（Network on Chip）は、データ依存のストールを減らし、タイル、カーネル、モデルレベルの並列性を向上させることができます。また、低ランクのカーネルを複数の畳み込みエンジンで並列化したり、3Dカーネルでは、シストリックアレイで処理を分割することで、並列化を実現します。例えば、DarkNet-53の中間層は、入力画像の16分の1の大きさです。この場合、DNAコアはNoCを再構成して、タイルレベルの並列性からチャネルレベルの並列性にシフトさせ、エンジンを動作させることができます。

EdgeCortix 社は、ケイデンス社の Xcellium Logic Simulation ツールを使用して、2つの一般的なニューラルネットワーク (ResNet-50 と Yolo v3) の消費電力、性能、領域に対する再構成可能なインターコネクトの効果を定量化しました。 A050、A100、A200、A400、A800 の 5 つのコンフィギュレーション済みの IP コアをシミュレーションしました。バッチ=1の推論では、A800は再構成可能なインターコネクトを使用しないバージョンと比較して、レイテンシが50%減少し、また、コンフィギュレーションが小さいほど改善度が低下しました。A800は最も多くのコンボリューションエンジンを搭載しており、コンパイラがリソースを再構成する際の粒度が大きくなっています。

この再構成機能により、ResNet-50の実行時にA800のスループットは2倍になります。また、このネットワークの層は、最初のイメージの後、指数関数的に縮小していきます。

シストリックアレイからコンボリューション・エンジンに切り替えることで、データ関連のストールを劇的に減らし、スループットを向上させることができます。EdgeCortixのIPの電力効率は、A400でピークに達し、A800では10%低下します。そのため、A800はResNet-50を扱うには必要以上の性能を備えており、MACの利用率が悪くなっているのではないかと推測しています。しかし、多くの顧客モデルはもっと複雑です。

カメラインターフェースの欠落

EdgeCortixは、従来のIP/FPGA設計会社から移行して、テストチップの動力源としてDNA-A800を選択しました。図2に示すように、この設計では、ResNet-50の全モデルをINT8で保持するのに十分な30MBのオンダイメモリを搭載しています。しかし、ほとんどの顧客モデルはこのメモリには収まらず、レイヤーの入れ替えが必要になります。そこで、このチップには、ピーク帯域幅 51GB/秒を駆動する 2 つのLPDDR4X DRAM コントローラが組み込まれています。

EdgeCortixでは、オンダイメモリを3つの仮想ブロックに分割し、重み、中間蓄積、活性化値に連続したアレイを割り当てています。これらのパーティションのサイズは柔軟で、ランタイムで構成することができます。ウェイトブロックとデータブロックは、LSU（ロードストアユニット）を介して LPDDR4X コントローラに直接接続され、外部からの読み取りと書き込みのレイテンシを低減します。同社のこのテストチップは、8つの64x64シストリックアレイと4つのコンボリューション・エンジンを採用し、MAC演算を高速化します。そして、アクティベーションやプーリングのような補助的な機能を処理するために4つのベクトルエンジンが統合を統合しています。

コプロセッサとして、チップはホストCPUまたはSoCに接続する必要があります。ホストとの接続には、最大16レーンのPCIe Gen3インターフェースを使用しますが、小型のM.2モジュールでは4レーンのみが実装されてます。カメラやマイク用のインターフェースはなく、GPIOのみ搭載がされています。その代わり、これらの周辺機器の駆動はホストSoCに依存しています。EdgeCortixは、DNA IPとMIPI-CSIカメラインターフェース、その他の周辺機器を備えた、車載用Arm Cortex-A CPU群を組み合わせたSoCを開発しています。

同社は、オープンソースのディープラーニング用コンパイラ「Apache TVM」のフレームワークを拡張して、「Mera」コンパイラを開発しました。PytorchやTensorflow Liteのトレーニング後のINT8量子化モデルに対応しています。また、ONNXモデルも受け入れることができます。オープンソースのTVMコンパイラはターゲットに依存しないグラフの最適化を処理し、DNA IPがサポートする演算子は量子化を意識した中間表現である量子化ニューラルネットワーク（QNN）に変換され、機能シミュレーションのための内蔵インタプリタ上で実行できるようになっています。次に、Meraは定数折りたたみ、演算子マージ、レイヤー融合などのDNA固有の最適化を行い、外部メモリアクセスを削減します。その後、演算がスケジューリングされ、ハードウェアに割り当てられます。Meraが認識できない演算は、専用ハードウェアよりも効率の低いCPU用のLLVMを使ってコンパイルされます。

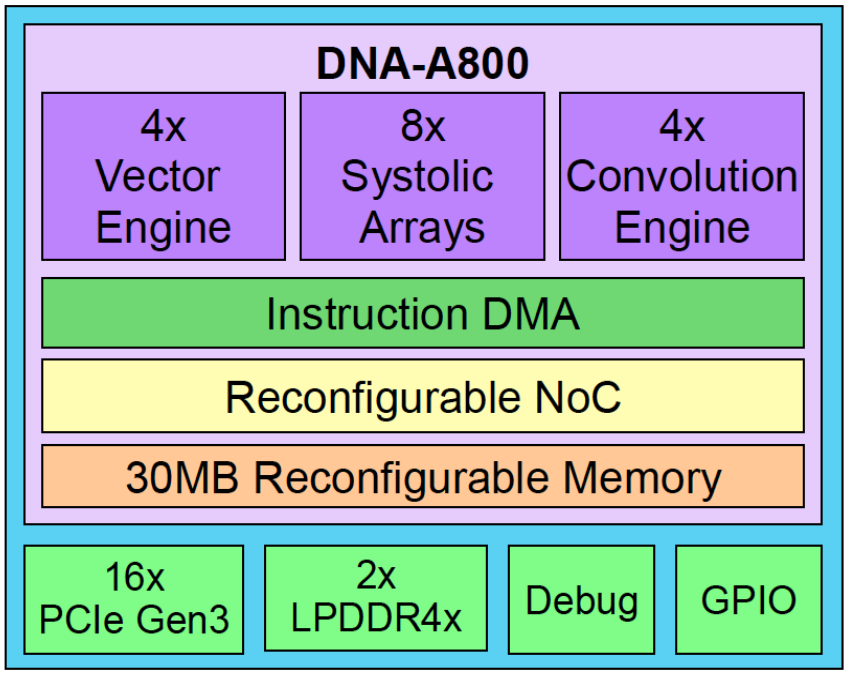

図 2：DNA-A800のブロック図

このコプロセッサは、シストリックアレイとコンボリューションエンジンの両方を使用して、ニューラルネットワークを処理します。プーリングとスケーリングの処理には、ベクトルエンジンの集積を利用。A800テストチップは、PCIe　Gen3インタフェースでホストに接続します。

A800のレイテンシは5倍のリード

このテストチップのマシンビジョンシステムに最も近い競合製品は、Nvidia社のXavierとHailo社のHailo-8です。（MPR 6/24/19, "Hailo Illuminates Low-Power AI Chip "を参照）これら3つのチップはレベル3のADASを駆動することができ、そのDLAを使用して前方の道路を監視することができます。Hailo-8は、ヘテロジニアス・リソース・マップを採用おり、ニューラルネットワークの計算ニーズを細かく追跡することで、26TOPSの性能を実現しています。A800はこの数字を2倍にし、さらに電力効率（TOPS/w）でもリードしていると推定しています。両チップとも、実世界のモデルでは畳み込みユニットの使用率を慎重に管理しているため、ResNet-50の効率ではこの差は縮まると予想されます。

A800の試算では、トータルパフォーマンスとResNet-50のレイテンシーでHailo-8とXavierの両社を上回っています。リアルタイム映像を解析する自動車用アプリケーションで一般的な1バッチサイズでは、ResNet-50をNvidiaの製品に比べて5倍速く推論し、シミュレーション値に基づく消費電力は73%も少ないことが分かります。 しかし、Xavierには高性能なホストCPU複合機が搭載されています。（MPR 2/19/18, "Nvidia Xavier Drives to Carmel "を参照）A800では、ホストプロセッサを別途用意する必要があるため、システム消費電力が増加します。しかし、コンパニオンSoCを搭載しても、現在の格差を考えれば、電力効率でリードしていることに変わりはないでしょう。

A800はオンダイメモリではHailo-8とほぼ同じですが、LPDDR4Xをサポートすることで差別化を図っています。DRAMの推論がないため、Hailo-8では小型のモデルしか扱えません。また、A800はテストチップのため、マシンビジョンアクセラレータでは標準的なカメラインターフェースを搭載していません。XavierとHailo-8はMIPIインターフェースを搭載していますが、A800はコンパニオンチップに依存する必要があり、システムレベルの画像遅延とコストが発生します。

	EdgeCortix DNA-A800	Nvidia Xavier	Hailo Hailo-8
Main-CPU Type	None	8x Carmel VLIW	1x Cortex-M4
CPU Speed	0.8GHz	2.5GHz	Undisclosed
DLA Type	Convolution and systolic engines	NVDLA + GPU	Custom
Peak AI Perf (INT8)	54 TOPS	30 TOPS	26 TOPS
ResNet-50 Latency	0.4ms	1.5ms	1.5ms
On-Die Memory	30MB	Undisclosed	32MB
DRAM Interface	2x 64-bit LPDDR4X-3200	1x 256-bit LPDDR4-2166	None
Camera Interface	None	16x MIPI CSI	2x MIPI
Power (TDP)	8W	30W	7W*
Efficiency (INT8)	6.8 TOPS/W	1.0 TOPS/W	3.7 TOPS/W
IC Process	12nm	12nm	16nm
Production	2H22 (est)	3Q18	1H20

表 2：マシンビジョンDLAの比較

A800はResNet-50のレイテンシで5倍の差をつけ、競合製品を圧倒しています。しかし、ホストCPUとカメライン ターフェースを持たないため、システムコストと消費電力が増加しています。（出典：ベンダ、※The Linley Group推定を除く）

ハード化されたシリコンを待つ

新規参入企業として、EdgeCortix IP は、マシンビジョン分野で大きな反響を呼んでいます。同社は、初期のFPGA展開の経験を生かし、再構成可能なインターコネクトアーキテクチャと2つのコンボリューション・エンジン機能を持つIPのライセンスを取得し、さまざまなAI処理に効果的に取り組んでいます。同社のIPライセンスビジネスは、ソフトウェアスタックを強化するのに役立っています。EdgeCortixは、オープンソースのコンパイラフレームワークを自社のアーキテクチャに移植し、顧客のシリコンでテストしました。テープアウトが近いというプレッシャーを感じることなく、積極的にソフトウェアをテストする機会を得ることができる新興企業はほとんどないでしょう。

このIPの性能を証明するために、同社では、XavierやHailo-8を上回るTOPS性能やストリーミング推論レイテンシ、そして驚異的な省電力を実現するテストチップを開発中です。コプロセッサであるため、外部のホストSoCが必要ですが、システムレベルのアドオン用に十分な電力ヘッドルームを残しています。Edge-Cortixは、当初、このテストチップの量産化を計画していませんでしたが、その素晴らしいスペックがチップの顧客の関心を集めています。スタートアップ企業が本格的にテープアウトを行うためには、追加の資金が必要です。また、テストチップのテープアウトは今年末の予定のため、生産は2022年末以降になります。その頃には、競争環境は完全に一変していることでしょう。

一方、EdgeCortix は、他のチップ設計用の IP のライセンス供与を継続する予定です。IP自体はADASアプリケーションに最も適していますが、ドローンやロボットなど、様々なカメラベースのシステムに使用することができます。しかし、IP企業として同社はArm、Cadence、CevaやSynopsysなど、より広い分野で事業を展開する競合他社と直面しています。これらのIPベンダは、それぞれ長年にわたって信頼できるソリューションを提供してきた実績があり、新興企業にとって高いハードルとなっています。

EdgeCortixのDNA IPは、有力なポートフォリオと言えます。シミュレーション結果に基づき、この独自のアーキテクチャが業界最高のレイテンシを実現すると期待しています。このIPは構成可能で拡張性があるため、自動車や消費者向けのさまざまなアプリケーションに対応することができます。潜在的な顧客は、同社のFPGAシステムを使ってDNA設計を実験し、来年に生産されるテストチップを使って検証することができます。 この組み合わせは、このスタートアップ企業がさらなるライセンシーを見つけるのに、一役買うことになるでしょう。

本記事は、当初1Q21に予定していたEdgeCortixのテストチップが1Q22まで延期されたことを反映したものです。♦

価格と入手方法

EdgeCortixは、DNA IPポートフォリオの価格とロイヤリティを非公開としていました。現在、FPGAへの展開やシミュレータによるモデルのベンチマーク用に、DNA IPとMeraソフトウェアへのアクセス権を提供しています。2022年中頃には、A800チップをサンプル出荷する予定です。

詳細については、www.edgecortix.com をご覧ください。Linley Spring Processor Conferenceのプレゼンテーションは、イベント終了後、www.linleygroup.com/でご覧いただけます。※SPC21の終了後すぐにご覧いただけますが、登録が必要です。

