2015 年 11 月 16 日 情報システム基盤学合同輪講発表資料

# A Million Spiling-Neuron Integrated Circuit with a Scalable Communication Network and Interface

著者：
出典：　*2014 47th Annual IEEE/ACM International Symposium on Microarchitecture*
発表者：　1553004 大場百香

## 1　はじめに

機械学習の方法の 1 つに Convolutional neural network(CNN),Deep neural network(DNN) などの大規模ニューラルネットワークがある。ところが、大規模ニューラルネットワークの計算に必要なメモリ量と計算量は膨大であり、現在の CPU や GPU では十分な性能を達成できない。そこで先行研究では大規模ニューラルネットワークの計算を実行するアクセラレータ (DianNao[?]) を設計した。DianNao は、CPU,GPU に比べて実装面積が小さく、しかも大規模ニューラルネットワークの計算をより高速かつ低消費ネルルギで行うことができる。しかし DainNao はニューラルネットワークの計算ユニットである NFU とメインメモリ間のデータ転送がボトルネックになっており、NFU の性能を十分に引き出すことができていない。NFU を休みなく稼働させるためには 467.3GB/s のバンド幅が必要であるのに対して DianNao のバンド幅は最大で 120GB/s と不足している。このようにバンド幅が不足しているのは、図??に示すように、DianNao ではメモリと NFU が別々のチップに分かれているためである。

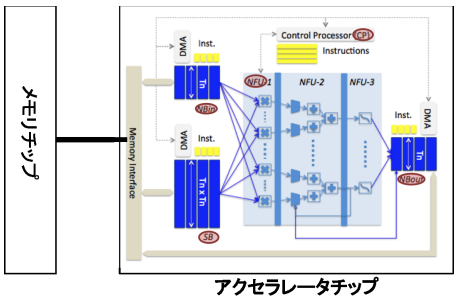


図 1: DianNao のアーキテクチャ

そこで本研究では、DainNao の問題点である NFU とメモリ間のバンド幅を拡大した新たなニューラルネットワークアクセラレータを開発する。大規模ニューラルネットワーク計算の更なる高速化と省エネルギ化を実現することを目的とし、アクセラレータの設計と評価を行う。

## 2　大規模ニューラルネットワーク

脳内の多数のニューロンが行っている、様々な情報処理の仕組みをコンピュータ内に実現したのが機械学習に使われているニューラルネットワークである (図??)。

$N_o = 1(\Sigma w_{ij} N_{in} - \theta \geq 0), 0(\Sigma w_{ij} N_{in} - \theta < 0)$

上の式のように、各ニューロンは複数のニューロンから入力 ($N_{in}$) 受け取り、入力とシナプス ($w_{ij}$) の積を求め、その総和が一定の閾値 $\theta$ を超えた場合に 1、超えない場合に 0 に出力する。

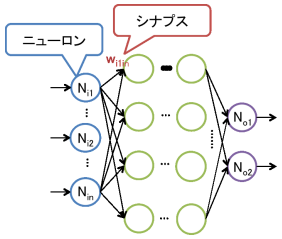


図 2: ニューラルネットワーク

本論文では、沢山の層で構成されている大規模ニューラルネットワークを対象とする。

## 3　提案アーキテクチャ

本研究では、メモリがアクセラレータチップ内にある on chip システムのアクセラレータ (DaDianNao) を設計する。大容量のメモリを使用可能にする eDRAM をチップ内に搭載し、チップ内のバンド幅を増やすためにタイル構成を採用する。また、複数チップが並列に動作することにより高い総バンド幅を実現する。

### 3.1　eDRAM の搭載

シナプスデータを格納するメモリとして eDRAM をアクセラレータチップ内に搭載する。on chip システムの場合 SRAM を使うことが多いが、eDRAM の SRAM に対するメリットとして低消費電力で動作すること、集積度が高いことが挙げられる。また、eDRAM と集積度がほぼ同じである DRAM は消費エネルギが大きいので今回は eDRAM を採用する。

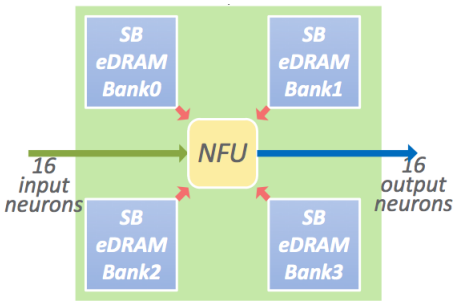


図 3: タイルのアーキテクチャ

eDRAM の欠点としてレイテンシが SRAM より大きいことが挙げられる。これを改善するため、提案アクセラレータでは eDRAM を 4 つに分けて NFU に配置した (図??)。1 つ当たりの面積を小さくすることによって配線を短くすることができ、シナプスを高速で NFU に送ることができる。また、eDRAM を NFU の中央に配置することで、eDRAM から NFU 間でのシナプス転送において低消費エネルギ、低レイテンシを実現することができる。

### 3.2 バンド幅の増加

1 つの chip は 16 個のタイルで構成されているので、1chip のバンド幅は、

(タイル数) ×（1 サイクルのデータ量）×（周波数）

$= 16$ 個$\times\ 1024bit \times 4$ 個$\times\ 608MHz = 4.8TB/s$

となり、先行研究で設計したアクセラレータのバンド幅より大きくすることができた。

## 4 評価

提案アクセラレータの性能と消費エネルギを評価する。提案アクセラレータを Verilog-HDL を用いて実装し、SYNOPSYS 社の CAD ツールを用いて消費電力の見積もりを行った。また、SYNOPSYS 社の VCS(functional verification solution) を用いて RTL(Register Transfer Level) シミュレーションを行うことにより、アクセラレータの性能を求めた。また、ニューラルネットワークを構成している Convolutional(CONV) 層,Pooling(POOL) 層,Local Response Normalization(LRN) 層,Classifier(CLASS) 層のベンチマーク（表??）を用いて、GPU(NVIDIA K20M GPU) と DaDianNao を比較することで性能を評価する。

### 4.1 マルチチップシステムによる性能

図??より、1chip でも GPU より提案アクセラレータの処理スピードが速いことがわかる。CONV3,CONV4 は計算量が多いため、64chip でしか実行できなかった。また、ほとんどのニューラルネットワークにおいてチップ数に比例して性能が向上していることがわかる。これはマルチチップ構成の有効性を表している。

| Layer | $N_x$ | $N_y$ | $K_x$ | $K_y$ | $N_i$ or $N_{if}$ | $N_o$ or $N_{of}$ | Synapses | Description |
|---|---|---|---|---|---|---|---|---|
| CLASS1 | - | - | - | - | 2560 | 2560 | 12.5MB | Object recognition and speech recognition tasks (DNN) [11]. |
| CLASS2 | - | - | - | - | 4096 | 4096 | 32MB | Multi-Object recognition in natural images (DNN), winner 2012 ImageNet competition [32]. |
| CONV1 | 256 | 256 | 11 | 11 | 256 | 384 | 22.69MB | |
| POOL2 | 256 | 256 | 2 | 2 | 256 | 256 | - | |
| LRN1 | 55 | 55 | - | - | 96 | 96 | - | |
| LRN2 | 27 | 27 | - | - | 256 | 256 | - | |
| CONV2 | 500 | 375 | 9 | 9 | 32 | 48 | 0.24MB | Street scene parsing (CNN) (e.g., identifying building, vehicle, etc) [18]. |
| POOL1 | 492 | 367 | 2 | 2 | 12 | 12 | - | |
| CONV3* | 200 | 200 | 18 | 18 | 8 | 8 | 1.29GB | Face Detection in YouTube videos (DNN), (Google) [34]. |
| CONV4* | 200 | 200 | 20 | 20 | 3 | 18 | 1.32GB | YouTube video object recognition, largest NN to date [8]. |

表 1: ベンチマーク

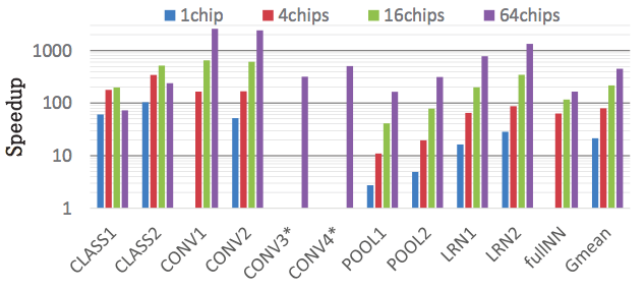


図 4: GPU と比較したときの DaDianNao の処理スピード

### 4.2 マルチチップシステムによるエネルギ

全てのチップ構成において消費エネルギを減らすことができた。CONV 層,POOL 層,LRN 層はチップの数が増えても消費エネルギの減少量は変わらないが、CLASS 層はチップ数が増えるとエネルギ減少量が減っている。これは CLASS 層で実行されている計算が他の層に比べて通信時間が長いことが原因だと考えられる。

## 5 まとめ

先行研究で設計したアクセラレータ（DianNao）の問題点を改善するために、マルチチップシステムを搭載したアクセラレータ (DaDianNao) の設計した。大規模ニューラルネットワークの計算を高性能かつ低消費エネルギで行うことができるアクセラレータを実現することができた。

## 参考文献

[1] T. Chen, Z. Du, N. Sun, J. Wang, C. Wu, Y. Chen, and O. Temam. DianNao: A Small-Footprint High-Throughput Accelerator for Ubiquitous Machine-Learning. In Interna- tional Conference on Architectural Support for Programming Languages and Operating Systems, 2014.