# NeuroCube: Low-Power Reconfigurable Neuromorphic Architecture Using Hybrid Memory Cube


## Abstract

*This document serves as a sample for submissions to MICRO 2014. We provide some guidelines that authors should follow when submitting papers to the conference.*


## 1. Introduction

THIS PART IS FOR INTRODUCTION

1. Machine Learning is important application in RMS (Recognition/Mining/Synthesis)

2. However, it's impossible to operate current ML techniques in embedded system due to massive computation

3. As demands for ML in embedded system such as IoT/Mobile platform increases, low-power and high power-efficiency ML is required

4. Therefore, Specific Architecture for ML is required which is better than General Purpose Micro-Arch such as CUDA

5. Deep Learning Network: compsed of multiple different type of NN is powerful tool in ML

6. To operated diff. NNs with single ASIC, Reconfigurbility is important in Neuromorphic Architecture

## 2. Previous Work

In this section, we will introduce recent machine learning techniques using different types of neural network and hardware implemenation based on ASIC or FPGA.

### 2.1. Neural network for machine learning

1. Why do we need different NNs other than Convolutional NNs?

2. How can we classify different NNs.

According to [1], neural network can be classified based on differernt characteristics.

3. What is the characteristic of diff. NNs (pro - cons) or appropriate application

Can we compute number of basic computations such as multiplications or additions for some applications?

Maybe table to compare diff NNs will be good.

### 2.2. Neuromorphic hardware implementation

text

## 3. Reconfigurable Neuromorphic Architecture

text

### 3.1. External memory

text

### 3.2. Cache memory

text

### 3.3. Processing elements

As we explained before, main basic operations for NNs are multiplication, addition, activation function, sampling ... So we will prepare all functions to cover different neural networks. Then controlling the data from memory to processing elements can implement different neural networks.

### 3.4. Network-on-chip

text

## 4. Traffic Analysis on Network-on-Chip

For different neural networks, our system described in previous sesction is simulated to operate given testcase for each neural network. For this system, standard DDR3-SDRAM (low bandwidth specification) is assumed as external DRAM.

## 5. Hybrid Memory Cube

hmc introduce

### 5.1. Architecture of HMC

hmc introduce

### 5.2. Intranetwork of HMC through logic die

hmc introduce

### 5.3. Internetwork of HMC through high speed links

hmc introduce

### 5.4. Neurocube using Intranetwork of HMC

Processor-in-Memory (PIM) design hmc introduce Area/power/thermal limitation

## 6. Traffic Analysis of NeuroCube

Introduce improvement of NeuroCube with HMC with high bandwidth

## 7. Conclusions

conclusions bla bla bla [1].

## References

[1] I. Basheer and M. Hajmeer, "Artificial neural networks: fundamentals, computing, design, and application," *Journal of microbiological methods*, vol. 43, no. 1, pp. 3–31, 2000.