경희대학교
Kyung Hee University

Computer Architecture

Kim, Lokwon

Assistant Professor
Department of Computer Engineering,
Kyung Hee University, Korea

The Course in a Nutshell

- Title: Computer Architecture
- After finishing this course, you will be able to imagine how computers physically work with your programs.
- Grading will be mainly based on 2 exams, attendance some practices, and assignments.
- **Prerequisite: Digital Logic, Basic C language**

경희대학교 Kyung Hee University

Who am I?

- A New faculty in the department of Computer Science and Engineering.

SK hynix

KETI 전자부품연구원 Korea Electronics Technology Institute

BROADCOM.

IBM

CISCO

Professor

A Fully Pipelined FPGA Architecture of a Factored Restricted Boltzmann Machine Artificial Neural Network

LOK-WON KIM, Cisco Systems
SAMEH ASAAD and RALPH LINSKER, IBM T. J. Watson Research Center

Artificial neural networks (ANNs) are a natural target for hardware acceleration by FPGAs and GPGPUs because commercial-scale applications can require days to weeks to train using CPUs, and the algorithms are highly parallelizable. Previous work on FPGAs has shown how hardware parallelism can be used to accelerate a "Restricted Boltzmann Machine" (RBM) ANN algorithm, and how to distribute computation across multiple FPGAs.

Here we describe a fully pipelined parallel architecture that exploits "mini-batch" training (combining many input cases to compute each set of weight updates) to further accelerate ANN training. We implement on an FPGA, for the first time to our knowledge, a more powerful variant of the basic RBM, the "Factored RBM" (fRBM). The fRBM has proved valuable in learning transformations and in discovering features that are present across multiple types of input. We obtain (in simulation) a 100-fold acceleration (vs. CPU software) for an fRBM having N = 256 units in each of its four groups (two input, one output, one intermediate group of units) running on a Virtex-6 LX760 FPGA. Many of the architectural features we implement are applicable not only to fRBMs, but to basic RBMs and other ANN algorithms more broadly.

Categories and Subject Descriptors: B.5.1 [**Register-Transfer-Level Implementation**]: Design

General Terms: Design

Additional Key Words and Phrases: Restricted Boltzmann Machine, FPGA based system design, hardware acceleration of Neural Network, pipelined and parallel hardware architecture

ACM Transactions on
Reconfigurable Technology and Systems (TRETS)
a journal focused on research in, on, and with reconfigurable systems and the underlying technology that supports these systems for computing or other applications.

Association for Computing Machinery

- In 2010, Dr. Kim started with IBM T.J. Watson Research team for Deep Learning Accelerator.

Lok-Won Kim; Sameh Asaad; Ralph Linsker "A Highly-Scalable and High-Performance Hardware Architecture for Factored Restricted Boltzmann Machine Artificial Neural Networks", ACM Transactions on Reconfigurable Technology and Systems, Volume: 7, Issue: 1, Feb. 2014, Article No: 5.

Professor

IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 29, NO. 5, MAY 2018

DeepX: Deep Learning Accelerator for Restricted Boltzmann Machine Artificial Neural Networks

Lok-Won Kim

***Abstract*— Although there have been many decades of research and commercial presence on high performance general purpose processors, there are still many applications that require fully customized hardware architectures for further computational acceleration. Recently, deep learning has been successfully used to learn in a wide variety of applications, but their heavy computation demand has considerably limited their practical applications. This paper proposes a fully pipelined acceleration architecture to alleviate high computational demand of an artificial neural network (ANN) which is restricted Boltzmann machine (RBM) ANNs. The implemented RBM ANN accelerator (integrating 1024 × 1024 network size, using 128 input cases per batch, and running at a 303-MHz clock frequency) integrated in a state-of-the art field-programmable gate array (FPGA) (Xilinx Virtex 7 XC7V-2000T) provides a computational performance of 301-billion connection-updates-per-second and about 193 times higher performance than a software solution running on general purpose processors. Most importantly, the architecture enables over 4 times (12 times in batch learning) higher performance compared with a previous work when both are implemented in an FPGA device (XC2VP70).**

***Index Terms*— Deep belief networks (DBNs), hardware-based computational acceleration, pipeline and parallel architecture, reconfigurable computing, restricted Boltzmann machine (RBM).**

computation time (e.g., days to weeks or more) due to its poor computing power. Integrating the software implementations into large scale clusters of processors (e.g., supercomputers) for parallel processing is not very efficient as well. The all-to-all communication nature of processing RBM ANNs requires excessive communication among the processors. Therefore, exploiting fine-grain parallelism and enormous on-chip interconnect resources enabled by custom hardware architectures can enable to train such all-to-all connected neural networks in a reasonable amount of processing time so that researchers can develop new applications and solutions within practically reachable time.

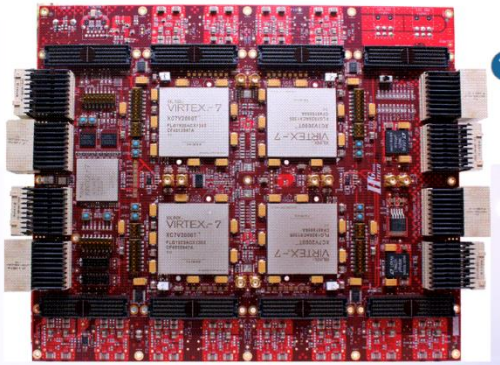
Rich publications have been focused on hardware-based acceleration of processing ANNs [6]–[8]. Reference [9] proposed special-purpose analog circuits used in a highly parallel architecture to enhance ANN learning performance. Field-programmable gate array (FPGA) ANN designs described in [7] and [10] have used fine-grain parallelization; the highly parallel processing of graphics processing units (GPUs) has been used in [11], and various commercial hardware solutions, including dedicated Application Specific

Browse Journals & Magazines > IEEE Transactions on Neural Ne...

IEEE Transactions on Neural Networks and Learning Systems

Home | Popular | Early Access | Cu

11.683
Impact Factor



- Published a research paper for Deep Learning Accelerator in the top journal for AI (IEEE Transactions on Neural Networks and Learning Systems).

Lok-Won Kim "DeepX: Deep Learning Accelerator for Restricted Boltzmann Machine Artificial Neural Networks", IEEE Transactions on Neural Networks and Learning Systems, Volume: 29, Issue: 5, May 2018, Page(s): 1441-1453.

Professor

- Played a key role in the development of the world first AI processor integrated Application Processor (Apple A11 Bionic) used in iPhone X

Professor

Core ML

Integrate machine learning models into your app.

Overview

With Core ML, you can integrate trained machine learning models into your app.

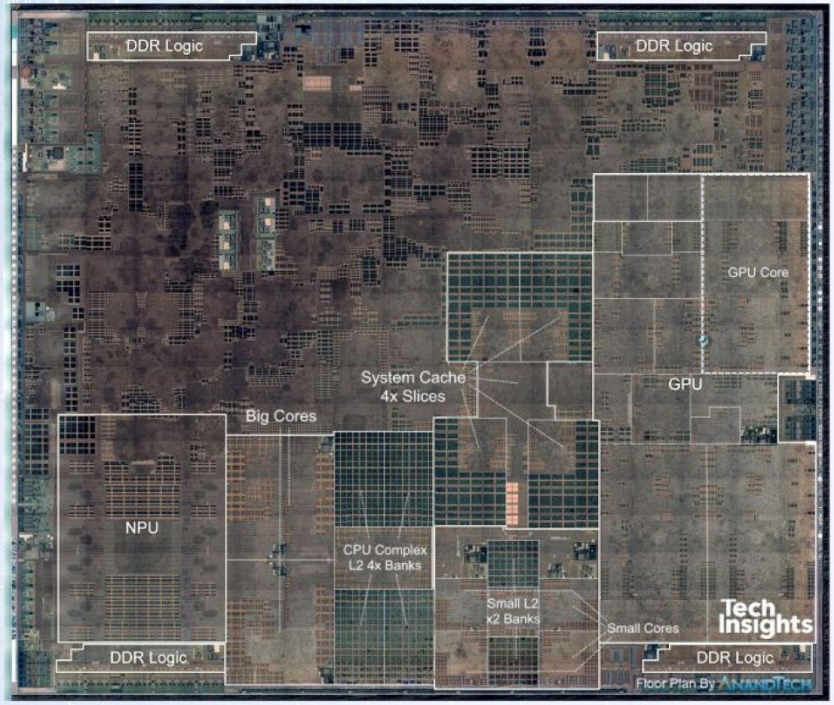
A *trained model* is the result of applying a machine learning algorithm to a set of training data. The model makes predictions based on new input data. For example, a model that's been trained on a region's historical house prices may be able to predict a house's price when given the number of bedrooms and bathrooms.

Core ML is the foundation for domain-specific frameworks and functionality. Core ML supports Vision for image analysis, Natural Language for natural language processing, and GameplayKit for evaluating learned decision trees. Core ML itself builds on top of low-level primitives like Accelerate and BNNS, as well as Metal Performance Shaders.

Source: APPLE

- Participated in the development of the Apple Neural Engine used in iPhone Xs.

Professor



A12 Die Photo

Die Block Comparison (mm²)		
SoC Process Node	Apple A12 TSMC N7	Apple A11 TSMC 10FF
Total Die	83.27	87.66
Big Core	2.07	2.68
Small Core	0.43	0.53
CPU Complex (incl. cores)	11.90	14.48
GPU Total	14.88	15.28
GPU Core	3.23	4.43
NPU	5.79	1.83

- The NPU has become one of main modules in the A12 AP.

Graduate Students

- Research: **Intelligent Computer Systems lab**

- **SW: Deep Learning Algorithms, SW programing, Python, C++**
  - Deep Learning Algorithm Optimization (model compression)

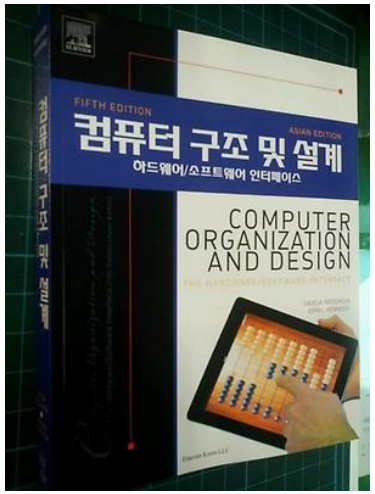
- **HW: Computer Architecture, VLSI, FPGA, Verilog, Deep Learning Algorithms, etc**
  - Deep learning HW processor (NPU)
  - 지능형 반도체 (시스템 반도체)

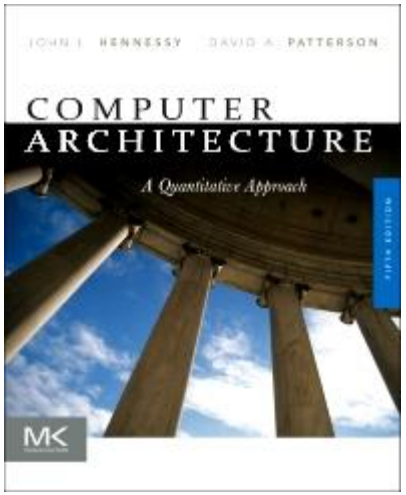
Professor

- Email: lwk@khu.ac.kr
- Office: Room 325-4
- Tel: 031-201-2547

Texts

- Text
  - David Patterson and John L. Hennessy, “Computer Organization and Design: The Hardware/Software Interface,” 5th Edition, Morgan Kaufmann.





경희대학교
Kyung Hee University

Language (부분 영어 강의 Partial English Class)

- Class will be tought in Korean.
- But all materials including class lecture notes, practice class notes, homework problems, and exam problems will be written in English
- Strongly recommend to buy English textbook

Grading

- Mid-term exam : 35%
- Final exam (final) : 40%
- Attendance : 5%
- **And some adjustment scores.**
- **Any kinds of dishonest behaviors (e.g., copying, cheating) will result in the F grade**
- **No negotiation on Grade!!!**