Young H. Oh

Building #944, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, South Korea
(+82)-10-8309-5123 younghwan@skku.edu younghwanoh.github.io
youngh-oh younghwanoh

Education

Sungkyunkwan University (SKKU), *Suwon, Korea*

Combined M.S. and Ph.D. in *Electrical and Computer Engineering* **Mar. 2013 - Present**
Visiting student at Architecture and Code Optimization Lab (ARC).
Advisor: Jae W. Lee

B.S.E. in *Electronic and Electrical Engineering* **Mar. 2009 - Feb. 2013**

Experience

Seoul National University, Seoul, Korea **Sep. 2016 - Present**
Research Assistant, ARC Lab
(2021~Present) Hardware & Software Co-design for Sparse DNN Training.
(2018~2021) System-level Optimization of Cloud DNN Accelerator Stacks (PCIe-based).
(2016~2018) Building Open-source DNN Acceleration Stacks with Automatic Quantization.
Keywords: Simulation (C++, Python), Performance Modeling, High-level Synthesis and FPGA

Sungkyunkwan University, Suwon, Korea **Feb. 2013 - Aug. 2016**
Research Assistant, PAPL Lab
(2016~2016) Specialized Hardware Support for Script Languages (Python & JavaScript part)
(2015~2015) QoS-aware Dynamic Power Optimization for Data-Parallel JavaScript Applications
(2013~2014) Parallelization of JavaScript Applications
Keywords: JavaScript, WebCL, OpenCL, WebKit, and Parallelization

Samsung Software Membership (Intern) **Jan. 2012 - Jan. 2013**
Hardware Engineer
User-Interactive Gaming Interface by Real-Time Fingertip Tracking
Keywords: OpenCV, Hardware

Papers

[IEEE MICRO '21] "Accelerating Genomic Data Analytics with Composable Hardware Acceleration Framework", Tae Jun Ham, David Bruns-Smith, Brendan Sweeney, Yejin Lee, Seong Hoon Seo, U Gyeong Song, **Young H. Oh**, Krste Asanovic, Jae W. Lee and Lisa Wu, *IEEE MICRO: Special Issue on Top Picks from the 2020 Computer Architecture Conferences.*, 2021.

[HPCA '21] "Layerweaver: Maximizing Resource Utilization of Neural Processing Units via Layer-Wise Scheduling", **Young H. Oh**, Seonghak Kim, Yunho Jin, Sam Son, Jonghyun Bae, Jongsung Lee, Yeonhong Park, Dong Uk Kim, Tae Jun Ham, and Jae W. Lee, *The 27th IEEE International Symposium on High Performance Computer Architecture*, 2021.

[IEEE Access '21] "An 8-bit Ring-Amplifier based Mixed-Signal MAC Circuit with Full Digital Interface and Variable Accumulation Length", Jongho Kim, Beomkyu Seo, **Young H. Oh**, Jung-Hoon Chun, Jae W. Lee, and Jintae Kim, *IEEE Access 2021, Vol. 9*, 2020.

[ISCA '20] "Genesis: A Hardware Acceleration Framework for Genomic Data Analysis", Tae Jun Ham, David Bruns-Smith, Brendan Sweeney, Yejin Lee, Seong Hoon Seo, U Gyeong Song, **Young H. Oh**, Krste Asanovic, Jae W. Lee and Lisa Wu, *The 47th ACM/IEEE International Symposium on Computer Architecture*, 2020. **— Selected as IEEE MICRO Top Picks from Computer Architecture Conferences in 2020.**

[HPCA '20] "A3: Accelerating Neural Network Attention Mechanism with Approximation", Tae Jun Ham, Sung Jun Jung, Seonghak Kim, **Young H. Oh**, Yoon Ho Song, Junghoon Park, Sanghee Lee, Kyoung Park, Jae W. Lee, Deog-Kyoon Jeong, *The 26th IEEE International Symposium on High Performance Computer Architecture*, 2020.

[PACT '18] "A Portable Automatic Data Quantizer for Deep Neural Networks", **Young H. Oh**, Quan Quan, Daeyeon Kim, Seonghak Kim, Jun Heo, Jaeyoung Jang, Sung Jun Jung, and Jae W. Lee, *The 27th IEEE International Conference on Parallel Architectures and Compilation Techniques*, 2018.

[ASPLOS '17] "Typed Architectures: Architectural Support for Lightweight Scripting", Channoh Kim, Jaehyeok Kim, Sungmin Kim, Dooyoung Kim, Namho Kim, Gitae Na, **Young H. Oh**, Hyeon Gyu Cho, and Jae W. Lee, *The 22nd ACM Architectural Support for Programming Languages and Operating Systems*, 2017. **— Selected as ASPLOS Highlight Session.**

[ISCA '16] "Short-Circuit Dispatch: Accelerating Virtual Machine Interpreters on Embedded Processors", Channoh Kim, Sungmin Kim, Hyeon Gyu Cho, Dooyoung Kim, Jaehyeok Kim, **Young H. Oh**, Hakbeom Jang, and Jae W. Lee, *The 43rd IEEE/ACM International Symposium on Computer Architecture*, 2021.

[IEEE D&T '16] "An eDRAM-Based Approximate Register File for GPUs", Donghwan Jeong, **Young H. Oh**, Yongjun Park, Jae W. Lee, *IEEE Design & Test: Special Issues on Approximate Computing*, 2016.

[PPoPP-Poster '15] "JAWS: A JavaScript Framework for Adaptive CPU-GPU Work Sharing", Xianglan Piao, Channoh Kim, **Younghwan Oh**, Huiying Li, Jin Cheon Kim, Hanjun Kim, and Jae W. Lee, *The 20th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, 2015.

[ISLPED '14] "eDRAM-based Tiered-Reliability Memory with Applications to Low-Power Frame Buffers", Kyungsang Cho, Yongjun Lee, **Young H. Oh**, Gyoo-cheol Hwang, and Jae W. Lee, *IEEE/ACM International Symposium on Low Power Electronics and Design*, 2014.

[PRISM '14] "Automatic Runtime Selection of Best Hardware for Data-Parallel JavaScript Kernels via Lifelong Profiling", **Younghwan Oh**, Xianglan Piao, Channoh Kim, and Jae W Lee, *The 2nd International Workshop on Parallelism in Mobile Platforms*, 2014.

[WWW-Poster '14] "Efficient CPU-GPU Work Sharing for Data-Parallel JavaScript Workloads", Xianglan Piao, Channoh Kim, **Younghwan Oh**, Hanjun Kim, and Jae W. Lee, *The 23rd International World Wide Web Conference*, 2014.

Skills

- **Programming Languages:** C/C++, Python, CUDA, OpenCL, JavaScript, Verilog, High-level Synthesis
- **Applications & Frameworks:** TensorFlow, PyTorch, Caffe, Numpy, Pandas, Nvprof, SDAccel, Matplotlib, Seaborn, LaTeX

Awards

- IEEE MICRO Top Picks 2021
- ASPLOS Best Paper Nominee (2017)
- SimSan Scholarship (2014-2016)

Open-Source Projects

OpenDNN: An Open-source, cuDNN-like Deep Learning Primitive Library, 2018

OpenDNN is an open-source, cuDNN-like deep learning primitive library to support various framework and hardware architectures such as CPU, GPU and FPGA. OpenDNN is implemented using CUDA and OpenCL and ported on popular deep learning frameworks (Caffe, Tensorflow, and DarkNet).

Implementation of Iterative Pruning on TensorFlow (Personal Project), 2016

This is an open-source implementation of *Song et al., "Learning both Weights and Connections for Efficient Neural Network.", NIPS '15*. I used *Embedding Lookup Sparse* operators in TensorFlow 1.0 and evaluated its performance impacts. Although it has some limitations, I observed some tangible performance benefits on GPU.

Invited Talks

FuriosaAI, South Korea, April 2021

Layerweaver: Maximizing Resource Utilization of Neural Processing Units via Layer-Wise Scheduling