Tae Jun Ham

CONTACT INFORMATION

1600 Amphitheatre Parkway
Mountain View, CA 94043
United States of America

Voice: (+1)-609-455-0675
E-mail: ham.taejun@gmail.com
Website: https://taejunham.github.io

CURRENT POSITION

Google, Sunnyvale, California, USA **Sep 2021 - Present**
Senior Software Engineer

Hardware/Software Performance Optimization for Google Applications and Systems

EDUCATION

Princeton University, Princeton, NJ USA **Sep 2012 - Jun 2018**
School of Engineering and Applied Science

Ph.D and **M.A** in Electrical Engineering

Dissertation: Data Access Optimization in Accelerator-oriented Heterogeneous Architecture through Decoupling and Memory Hierarchy Specialization

Advisor: Professor Margaret Martonosi and Professor Juan Luis Aragon

Duke University, Durham, North Carolina USA **Aug 2009 - Dec 2011**
Pratt School of Engineering

B.S.E in Electrical and Computer Engineering (GPA : 3.95/4.00)
Summa Cum Laude, with Distinction in Electrical and Computer Engineering

PROFESSIONAL EXPERIENCE

Seoul National University, Seoul, Republic of Korea
Postdoctoral Researcher **Jul 2018 - Jul 2021**

Software-hardware co-design for emerging applications such as big data analytics and machine learning.
This position also fulfilled the **mandatory military service duty** required for all Korean men.

Microsoft Research, Cambridge, UK
Graduate Research Intern **May - Aug, 2016**

Research on an efficient secure memory design with near-data computation

Intel Labs — Parallel Computing Lab, Santa Clara, USA
Graduate Technical Intern **May - Nov, 2015**

Research on a custom hardware accelerator for graph analytics applications.

AMD Research, Austin, USA
Co-op Engineer **Jun - Aug, 2013**

Research on an efficient use of high-perf energy-efficient heterogeneous system consists of large, low memory bandwidth processors and small, high memory bandwidth processors.

Samsung Advanced Institute of Technology, Yongin, Republic of Korea
Research Intern **Jun - Aug, 2012**

Research on a GPU branch divergence problem.

Duke University — BCL Research Group, Durham, USA
Research Assistant **Jan - May, 2012**

Research on a heterogeneous memory system.

HONORS AND AWARDS

- **IEEE MICRO Top Picks (2021)**
- **IEEE MICRO Top Picks Honorable Mention (2021)**
- **ISPASS Best Paper Award Nominee (2020)**
- **MICRO-49 Best Paper Award (2016)**
- **IEEE MICRO Top Picks Honorable Mention(2016)**
- **Facebook Graduate Fellowship Finalist (2016-2017)**, Facebook, Inc.
- **Gordon Y.S. Wu Fellowship (2012-2017)**, Princeton University
- **Samsung Scholarship (2012-2017)**, Scholarship that supports up to $50,000 per year
- **Summa Cum Laude**, Duke University

Selected Publications

[ECCV '22] L3: Accelerator-Friendly Lossless Image Format for High-Resolution, High-Throughput DNN Training

Jonghyun Bae, Woohyeon Baek, **Tae Jun Ham**, Jae W. Lee

European Conference on Computer Vision (**ECCV**)
Acceptance Rate : 1650/5803 ≈ 28%

[IEEE TC'22] Architecting a Flash-based Storage System for Low-cost Inference of Extreme-scale DNNs

Yunho Jin*, Shine Kim*, **Tae Jun Ham**, Jae W. Lee

IEEE Transactions on Computers (**TC**)

[ACM TECS '22] MaPHeA: A Lightweight Memory Hierarchy-aware Profile-guided Heap Allocation

Deok-Jae Oh, Yaebin Moon, Do Kyu Ham, Yongjun Park **Tae Jun Ham**, Jae W. Lee, Jung Ho Ahn, Eojin Lee

ACM Transactions on Embedded Computing Systems (**TECS**)

[MLSys '22] ULPPACK: Fast Sub-8-bit Matrix Multiply on Commodity SIMD Hardware

Jaeyeon Won, Jeyeon Si, Sam Son, **Tae Jun Ham**, and Jae W. Lee

The 5th Conference on Machine Learning and Systems (**MLSys**)

[HPCA '22] ANNA: Specialized Architecture for Approximate Nearest Neighbor Search

Yejin Lee, Hyunji Choi, Sunhong Min, Hyunseung Lee, Sangwon Baek, Dawoon Jeong, Jae W. Lee, and **Tae Jun Ham (Co-corresponding)**

The 28th *IEEE International Symposium on High-Performance Computer Architecture* (**HPCA**)
Acceptance Rate : 80/262 ≈ 30%

[HPCA '22] Mithril: Cooperative Row Hammer Protection on Commodity DRAM Leveraging Managed Refresh

Michael Jaemin Kim, Jaehyun Park, Yeonhong Park, Wanju Doh, Namhoon Kim, **Tae Jun Ham**, Jae W. Lee, and Jung Ho Ahn

The 28th *IEEE International Symposium on High-Performance Computer Architecture* (**HPCA**)
Acceptance Rate : 80/262 ≈ 30%

[ATC '21] ASAP: Fast Mobile Application Switch via Adaptive Prepaging

Sam Son, Seung Yul Lee, Yunho Jin, Jonghyun Bae, Jinkyu Jeong, **Tae Jun Ham**, Jae W. Lee, Hongil Yoon

USENIX Annual Technical Conference (**ATC**), 2021
Acceptance Rate : 64/341 ≈ 19%

[LCTES '21] MaPHeA: A Lightweight Memory Hierarchy-aware Profile-guided Heap Allocation

Deok-Jae Oh, Yaebin Moon, Eojin Lee, **Tae Jun Ham**, Jae W. Lee, Jung Ho Ahn

ACM SIGPLAN/SIGBED Conference on Languages, Compilers, Tools and Theory of Embedded Systems (**LCTES**), 2021
Acceptance Rate : 15/37 ≈ 40%

[IEEE Micro] 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, Lisa Wu

IEEE Micro, May/June 2021
Special Issue on Top Picks from the 2020 Computer Architecture Conferences

[ISCA '21] ELSA: Hardware-Software Co-design for Efficient, Lightweight Self-Attention Mechanism in Neural Networks

Tae Jun Ham*, Yejin Lee*, Seong Hoon Seo, Soosung Kim, Hyunji Choi, Sung Jun Jung, Jae W. Lee

The 47th *ACM/IEEE International Symposium on Computer Architecture* (**ISCA**), 2021
Acceptance Rate : 76/406 ≈ 19%
• *Two authors contributed equally.

[ISCA '21] BOSS: Bandwidth-Optimized Search Accelerator for Storage-Class Memory

Jun Heo, Seungyul Lee, Sunhong Min, Yeonhong Park, Sung Jun Jung, **Tae Jun Ham**, Jae W. Lee

The 47th *ACM/IEEE International Symposium on Computer Architecture* (**ISCA**), 2021

Acceptance Rate : 76/406 ≈ 19%

[ASPLOS '21] MERCI: Efficient Embedding Reduction on Commodity Hardware via Sub-Query Memoization
Yejin Lee, Seong Hoon Seo, Hyunji Choi, Hyoung Wook Sul, Soosung Kim, Jae W. Lee, **Tae Jun Ham (Corresponding)**
The 26th *ACM International Conference on Architectural Support for Programming Languages and Operating Systems* (**ASPLOS**), 2021
Acceptance Rate : 75/398 ≈ 19%
• Also presented in PeRSonAl Workshop @ MLSys 2021

[FAST '21] FlashNeuron: SSD Enables Large-Batch Training of Very Deep Neural Networks
Jonghyun Bae, Jongsung Lee, Yunho Jin, Sam Son, Shine Kim, Hakbeom Jang, **Tae Jun Ham**, Jae W. Lee
USENIX Conference on File and Storage Technologiess (**FAST**), 2021
Acceptance Rate : 28/130 ≈ 21%

[FAST '21] Behemoth: A Flash-centric Training Accelerator for Extreme-scale DNNs
Shine Kim, Yunho Jin, Gina Sohn, Jonghyun Bae, **Tae Jun Ham**, Jae W. Lee
USENIX Conference on File and Storage Technologiess (**FAST**), 2021
Acceptance Rate : 28/130 ≈ 21%

[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**, Jae W. Lee
The 27th *IEEE International Symposium on High-Performance Computer Architecture* (**HPCA**), 2021
Acceptance Rate : 63/258 ≈ 24%

[ICCAD '20] Unlocking Wordline-level Parallelism for Fast Inference on RRAM-based DNN Accelerator
Yeonhong Park, Seung Yul Lee, Hoon Shin, Jun Heo, **Tae Jun Ham**, Jae W. Lee
The 39th *IEEE/ACM International Conference on Computer-Aided Design* (**ICCAD**), 2020
Acceptance Rate : 127/470 ≈ 27%

[MICRO '20] Graphene: Strong yet Lightweight Row Hammer Protection
Yeonhong Park, Woosuk Kwon, Eojin Lee, **Tae Jun Ham**, Jung Ho Ahn, Jae W. Lee
The 53rd *IEEE/ACM International Symposium on Microarchitecture* (**MICRO**), 2020
Acceptance Rate : 82/424 ≈ 19%
• IEEE Micro Top Picks Honorable Mention

[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, Lisa Wu
The 47th *ACM/IEEE International Symposium on Computer Architecture* (**ISCA**), 2020
Acceptance Rate : 77/428 ≈ 18%
• IEEE Micro Top Picks

[ISCA '20] A Case for Hardware-based Demand Paging
Gyusun Lee*, Wenjing Jin*, Wonsuk Song, Jeonghun Gong, Jonghyun Bae, **Tae Jun Ham**, Jae W. Lee, Jinkyu Jeong
The 47th *ACM/IEEE International Symposium on Computer Architecture* (**ISCA**), 2020
Acceptance Rate : 77/428 ≈ 18%
∗ Two authors contributed equally.

[ISCA '20] A Specialized Architecture for Object Serialization with Applications to Big Data Analytics
Jaeyoung Jang, Sung Jun Jung, Sunmin Jeong, Jun Heo, Hoon Shin, **Tae Jun Ham**, Jae W. Lee
The 47th *ACM/IEEE International Symposium on Computer Architecture* (**ISCA**), 2020
Acceptance Rate : 77/428 ≈ 18%

[ISPASS '20] MosaicSim: A Lightweight, Modular Simulator for Heterogeneous Systems
Opeoluwa Matthews, Aninda Manocha, Davide Giri, Marcelo Orenes-Vera, Esin Tureci,

Tyler Sorensen, **Tae Jun Ham**, Juan Luis Aragon, Luca P. Carloni, Margaret Martonosi

IEEE International Symposium on Performance Analysis of Systems and Software (**ISPASS**), 2020
Acceptance Rate : 25/73 ≈ 34%
• **Nominated for the Best Paper Award**

[ASPLOS '20] IIU: Specialized Architecture for Inverted Index Search

Jun Heo, Jaeyeon Won, Yejin Lee, Shivam Bharuka, Jaeyoung Jang, **Tae Jun Ham**, Jae W. Lee

The 25th *ACM International Conference on Architectural Support for Programming Languages and Operating Systems* (**ASPLOS**), 2020
Acceptance Rate : 86/476 ≈ 18%

[HPCA '20] A^3: Accelerating Neural Network Attention Mechanism with Approximation

Tae Jun Ham, Sung Jun Jung, Seonghak Kim, Young H. Oh, Yeonhong Park, Yoonho Song, Jung-Hun Park, Sanghee Lee, Kyoung Park, Jae W. Lee, Deog-Kyoon Jeong

The 26th *IEEE International Symposium on High Performance Computer Architecture* (**HPCA**), 2020
Acceptance Rate : 48/235 ≈ 20%

[MICRO '19] Charon: Specialized Near-Memory Processing Architecture for Clearing Dead Objects in Memory

Jaeyoung Jang, Jun Heo, Yejin Lee, Jaeyeon Won, Seonghak Kim, Sung Jun Jung, Hakbeom Jang, **Tae Jun Ham**, Jae W. Lee

The 52nd *IEEE/ACM International Symposium on Microarchitecture* (**MICRO**), 2019
Acceptance Rate : 79/343 ≈ 23%

[IEEE Micro] SSDStreamer: Specializing I/O Stack for Large-Scale Machine Learning

Jonghyun Bae, Hakbeom Jang, Jeonghun Gong, Wenjing Jin, Shine Kim, Jaeyoung Jang, **Tae Jun Ham**, Jinkyu Jeong, Jae W. Lee

IEEE Micro, September 2019

[ATC '19] Asynchronous I/O Stack: A Low-latency Kernel I/O Stack for Ultra-Low Latency SSDs

Gyusun Lee, Seokha Shin, Wonsuk Song, **Tae Jun Ham**, Jae W. Lee, Jinkyu Jeong

USENIX Annual Technical Conference (**ATC**), 2019
Acceptance Rate : 71/356 ≈ 20%

[ATC '19] Practical Erase Suspension for Modern Low-latency SSDs

Shine Kim, Jonghyun Bae, Hakbeom Jang, Wenjing Jin, Jeonghun Gong, Seungyeon Lee, **Tae Jun Ham**, Jae W. Lee

USENIX Annual Technical Conference (**ATC**), 2019
Acceptance Rate : 71/356 ≈ 20%

[ACM TACO] Efficient Data Supply for Parallel Heterogeneous Architectures

Tae Jun Ham, Juan L Aragon, Margaret Martonosi

ACM Transactions on Architecture and Code Optimization (**TACO**), June 2019
• Presented on **HiPEAC** 2020 Conference

[ACM TACO] Decoupling Data Supply from Computation for Latency-Tolerant Communication in Heterogeneous Architectures

Tae Jun Ham, Juan L Aragon, Margaret Martonosi

ACM Transactions on Architecture and Code Optimization (**TACO**), June 2017

[MICRO '16] Graphicionado: A High-Performance and Energy-Efficient Accelerator for Graph Analytics

Tae Jun Ham, Lisa Wu, Narayanan Sundaram, Nadathur Satish, Margaret Martonosi

The 49th *IEEE/ACM International Symposium on Microarchitecture* (**MICRO**), 2016
Acceptance Rate : 61/283 ≈ 22%
• **MICRO-49 Best Paper Award**

[MICRO '15] DeSC: Decoupled Supply-Compute Communication Management for Heterogeneous Architectures

Tae Jun Ham, Juan L Aragon, Margaret Martonosi

The 48th *IEEE/ACM International Symposium on Microarchitecture* (**MICRO**), 2015
Acceptance Rate : 61/283 ≈ 22%
• **IEEE Micro's Top Picks from the Computer Architecture — Honorable Mention** (Top 23 Computer Architecture Papers of 2015)
• **Motivated $5.8million DARPA-funded DECADES project** (https://decades.cs.princeton.edu/)

[HPCA '13] Disintegrated Control for Energy-Efficient and Heterogeneous Memory Systems
Tae Jun Ham, Bharath K. Chelepalli, Neng Xue, Benjamin C. Lee
The 19th *IEEE International Symposium on High Performance Computer Architecture* (**HPCA**), 2013
Acceptance Rate : 51/249 ≈ 20%

Skills

- **Languages** : C/C++, CUDA, Python, Chisel, Verilog
- **Applications/Frameworks** : PyTorch, Numpy, SciPy, Pandas, Intel Pin, LLVM, Cadence C-to-Silicon, LaTeX

Patents

Instruction, Circuits, and Logic for Graph Analytics Acceleration (US20170286122A1; WO/2017/172173)
with Lisa Wu, Nadathur Satish and Narayanan Sundaram

Method For Accelerating Candidate Selection based on Similarity and Accelerator for Performing Candidate Selection (Pending - US20200311182A1)
with Jae W. Lee, Deog-Kyoon Jeong, Seonghak Kim, Sung Jun Jung, and Minsoo Lim

Hardware Accelerator Performing Search using Inverted Index Structure and Search System Including the Hardware Accelerator (Pending - US20210374131A1)
with Jae W. Lee, Jun Heo, Jaeyeon Won, and Yejin Lee

Method for processing page fault by processor (Pending - US20210374063A1)
with Jinkyu Jeong, Jae W. Lee, Gyusun Lee, and Wenjing Jin

Scheduler, Method for Operating the Same and Neural Network Accelerator System Including the Same (Pending - US20210373944A1)
with Jae W. Lee, Young H. Oh, and Seonghak Kim

Electronic Device and Method with Scheduling (Pending - US20220114015A1)
with Seung Wook Lee, Jae W. Lee, Young Hwan Oh, Sam Son, and Yunho Jin

Input-Aware Current Compensation for Reliable NVM Crossbar based In-Memory Computing (Pending)
with Jae W. Lee, Yeonhong Park, Seungyul Lee, Hoon Shin, and Jun Heo

Method for Operating the same, and Electronic Device including the same
(for efficient quantized matrix multiplications) **(Pending)**
with Jae W. Lee, Jaeyeon Won, and Seungwook Lee

Research Mentoring

I closely worked with these students and supervised their work (along with their primary supervisor Jae W. Lee) through frequent (often more than once a week) meetings with each group of students. I provide advices and help on research, technical implementation, and writing.

Graduate Students

- **Jaeyoung Jang**, Ph.D from Sungkyunkwan University — **Jul 2018 - Jan 2020**
  Now at Samsung Electronics
- **Jeonghun Gong**, M.S from Seoul National University — **Jan 2019 - Jan 2021**
  Now at Samsung Electronics
- **Young H. Oh**, Ph.D from Sungkyunkwan University — **Jul 2018 - Jul 2021**
  Now at Samsung Electronics

- **Jun Heo**, Ph.D from Seoul National University **Jul 2018 - Jul 2021**
  Now at Samsung Electronics
- **Jonghyun Bae**, Ph.D from Seoul National University **Sep 2018 - Jul 2021**
  Now a Postdoc at Seoul National University
- **Shine Kim**, Ph.D from Seoul National University **Sep 2018 - Jul 2021**
  Now at Samsung Electronics
- **Wenjing Jin**, Ph.D Student at Seoul National University **Sep 2018 - Jul 2021**
- **Sung Jun Jung**, M.S/Ph.D Student at Seoul National University **Sep 2018 - Jul 2021**
- **Yejin Lee**, M.S/Ph.D Student at Seoul National University **Jan 2019 - Jul 2021**
- **Yeonhong Park**, M.S/Ph.D Student at Seoul National University **Jun 2019 - Jul 2021**
- **Yunho Jin**, M.S from Seoul National University **Jun 2019 - Jul 2021**
  Now a Ph.D student at Harvard University
- **Sam Son**, M.S from Seoul National University **Jun 2019 - Jul 2021**
  Now a Ph.D student at UC Berkeley
- **Seung Yul Lee**, M.S/Ph.D Student at Seoul National University **Jun 2019 - Jul 2021**
- **Seong Hoon Seo**, M.S/Ph.D Student at Seoul National University **Jun 2019 - Jul 2021**
- **Soosung Kim**, M.S/Ph.D Student at Seoul National University **Jan 2020 - Jul 2021**
- **Hyunji Choi**, M.S from Seoul National University **Jan 2020 - Jul 2021**
  Now at Meta

Undergraduate Students

- **Jaeyeon Won**, B.S.E from Seoul National University **Jan 2019 - Aug 2019**
  Now a Ph.D student at MIT. **May 2020 - Aug 2020**
- **Wookyung Song**, Undergraduate at Seoul National University **Jun 2019 - Aug 2019**
- **Hyung Wook Sul**, Undergraduate at Seoul National University **Jun 2020 - Aug 2020**

Professional Activities

Paper Reviews

- IEEE Transactions on Very Large Scale Integration Systems (TVLSI) - 2015
- IEEE Transactions on Mobile Computing (TMC) - 2016
- IEEE Transactions on Computer (TC) - 2017, 2018, 2020
- IEEE Transactions on Computer-Aided Design (TCAD) - 2021
- IEEE Computer Architecture Letters (CAL) - 2018
- IEEE Micro - 2019
- ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS) - 2017
- ACM/IEEE International Symposium on Computer Architecture (ISCA) - 2020
- ACM Transactions on Architecture and Code Optimization (TACO) - 2014
- ACM Transactions on Parallel Computing (TOPC) - 2019
- Elsevier Future Generation Computer Systems (FGCS) - 2019

Invited Talks

- KAIST, POSTECH (Sep 2017)
  DeSC: Decoupling Data Supply from Computation for Latency-Tolerant Communication in Heterogeneous Architectures
- DARPA HIVE PI Meeting (Oct 2017)
  Graphicionado: A High-Performance and Energy-Efficient Accelerator for Graph Analytics
- HiPEAC 2020 (Jan 2020)
  Efficient Data Supply for Parallel Heterogeneous Architectures
- Seoul National University AI Summer School (Aug 2020)
  Accelerating Neural Network Attention Mechanism with HW/SW Codesign
- POSTECH Summer AI Seminar (Aug 2020)
  Hardware/Software Co-design for Modern AI and Data Analytics Applications

Services

- Program Committee Member, IEEE MICRO Top Picks (2022)
- Program Committee Member, ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS '23 - Spring, Summer, Fall)
- External Review Committee Member, ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS '19)
- Web Chair, ACM International Symposium on Code Generation and Optimization (CGO '21, CGO '22)