Jongse Park

August 2025

Contact Information

School of Computing
KAIST
291 Daehak-ro, Yuseong-gu
Daejeon, South Korea, 34141

E-mail: jspark@casys.kaist.ac.kr
URL: https://jongse-park.github.io

Research Interests

Computer Architecture, Computer System, HW/SW Co-Design,
Generative AI Serving Systems, On-Device AI Systems, Processing-in-Memory

Employment

Visiting Associate Professor. **Stanford University** Jan. 2025–date
Associate Professor. **KAIST** Mar. 2024–date
Assistant Professor. **KAIST** Dec. 2019–Feb. 2024
System Architect. **Bigstream Solutions Inc.** Jun. 2018–Nov. 2019

Education

Ph.D. in Computer Science. Georgia Institute of Technology Aug. 2013–Aug. 2018

- Advisor: Prof. Hadi Esmaeilzadeh
- Dissertation: *Breaking the Abstractions for Productivity and Performance in the Era of Specialization*

M.S. in Computer Science. KAIST Feb. 2012

- Advisor: Prof. Seungryoul Maeng
- Thesis: *Dynamic Resource Reconfiguration on the Cloud for Improving Data Locality*

B.E. in Computer Science and Engineering. Sogang University Feb. 2010

- Graduated with Honors

Honors and Awards

Best Paper Award. IEEE Micro. 2025
"LPU: A Latency-optimized and Highly Scalable Processor for Large Language Model Inference"

Samsung Humantech Paper Award. 2025
Gold Prize (1st place in the Computer Science and Engineering track)

Best Paper Award & Distinguished Artifact Award. IISWC. 2024
"LLMServingSim: A HW/SW Co-Simulation Infrastructure for LLM Inference Serving at Scale"

Distinguished Artifact Award. ISCA. 2024
"DaCapo: Accelerating Continuous Learning in Autonomous Systems for Video Analytics"

ISCA 25-Year Retrospective 1996-2020 Inclusion 2023
"Bit Fusion: Bit-Level Dynamically Composable Architecture for Accelerating Deep Neural Networks"

ISCA 25-Year Retrospective 1996-2020 Inclusion 2023
"General-Purpose Code Acceleration with Limited-Precision Analog Computation"

Distinguished Paper Award. HPCA. 2016
"Tabla: A Unified Template-Based Framework for Accelerating Statistical Machine Learning"

Honorable Mention in IEEE Micro Top Picks from 2014 Computer Architecture Conferences. 2015
"General-Purpose Code Acceleration with Limited-Precision Analog Computation"

Refereed Conference Papers

1. C. Oh, S. Oh, J. Hwang, Y. Kim, H. Sharma, **J. Park**, "Neo: Real-Time On-Device 3D Gaussian Splatting with Reuse-and-Update Sorting Acceleration," in *ACM International Conference on Architectural Support for Programming Languages and Operating Systems (**ASPLOS**)*, March 2026. (To Appear)

2. W. Kim, Y. Lee, Y. Kim, J. Hwang, S. Oh, J. Jung, A. Huseynov, W. G. Park, C. H. Park, D. Mahajan, **J. Park**, "Pimba: A Processing-in-Memory Acceleration for Post-Transformer Large Language Model Serving," in *IEEE/ACM International Symposium on Microarchitecture (**MICRO**)*, October 2025.

3. W. Yang, Y. Shin, O. Woo, G. Park, H. Ham, J. Kang, **J. Park**, Gwangsun Kim, "PyTorchSim: A Comprehensive, Fast, and Accurate NPU Simulation Framework," in *IEEE/ACM International Symposium on Microarchitecture (**MICRO**)*, October 2025.

4. J. Hwang, D. Kim, S. Lee, Y. Kim, G. Heo, H. Kim, Y. Jeong, T. Meaza, E. Park, J. Ahn, **J. Park**, "Déjà Vu: Efficient Video-Language Query Engine with Learning-based Inter-Frame Computation Reuse," in *International Conference on Very Large Data Bases (**VLDB**)*, September 2025. (To Appear)

5. M. Kim, S. Hong, R. Ko, S. Choi, H. Lee, J. Kim, J-Y Kim, **J. Park**, "Oaken: Fast and Efficient LLM Serving with Online-Offline Hybrid KV Cache Quantization," in *ACM/IEEE International Symposium on Computer Architecture (**ISCA**)*, June 2025.

6. Y. Kim, I. Kim, K. Choi, J. Ahn, **J. Park**, J. Huh, "Interference-Aware DNN Serving on Heterogeneous Processors in Edge Systems," in *IEEE International Conference on Computer Design (**ICCD**)*, November 2024.

7. J. Cho, M. Kim, H. Choi, G. Heo, **J. Park**, "LLMServingSim: A HW/SW Co-Simulation Infrastructure for LLM Inference Serving at Scale," in *IEEE International Symposium on Workload Characterization (**IISWC**)*, September 2024.

   Best Paper Award & Distinguished Artifact Award

8. M. Kim, J. Hwang, G. Heo, S. Cho, D. Mahajan, **J. Park**, "Accelerating String-key Learned Index Structures via Memoization-based Incremental Training," in *International Conference on Very Large Data Bases (**VLDB**)*, August 2024.

9. Y. Kim, C. Oh, J. Hwang, W. Kim, S. Oh, Y. Lee, H. Sharma, A. Yazdanbakhsh, **J. Park**, "DaCapo: Accelerating Continuous Learning in Autonomous Systems for Video Analytics," in *ACM/IEEE International Symposium on Computer Architecture (**ISCA**)*, June 2024.

   Distinguished Artifact Award

10. G. Heo, S. Lee, J. Cho, H. Choi, S. Lee, H. Ham, G. Kim, D. Mahajan, **J. Park**, "NeuPIMs: NPU-PIM Heterogeneous Acceleration for Batched LLM Inferencing," in *ACM International Conference on Architectural Support for Programming Languages and Operating Systems (**ASPLOS**)*, April 2024.

11. S. Ghodrati, S. Kinzer, H. Xu, R. Mahapatra, Y. Kim, B. H. Ahn, D. K. Wang, L. Karthikeyan, A. Yazdanbakhsh, **J. Park**, N. S. Kim, H. Esmaeilzadeh, "Tandem Processor: Grappling with Emerging Operators in Neural Networks," in *ACM International Conference on Architectural Support for Programming Languages and Operating Systems (**ASPLOS**)*, April 2024.

12. Sunho Lee, Seonjin Na, Jungwoo Kim, Jongse Park, and Jaehyuk Huh, "Tunable Memory Protection for Secure Neural Processing Units," in *IEEE International Conference on Computer Design (**ICCD**)*, October 2022.

13. Bokyeong Kim, Soojin Hwang, Sanghoon Cha, Chang Hyun Park, Jongse Park, and Jaehyuk Huh, "Supporting Dynamic Translation Granularity for Hybrid Memory Systems," in *IEEE International Conference on Computer Design (**ICCD**)*, October 2022.

14. Joon Kyung Kim, Byung Hoon Ahn, Sean Kinzer, Soroush Ghodrati, Rohan Mahapatra, Brahmendra Yatham, Dohee Kim, Parisa Sarikhani, Babak Mahmoudi, Divya Mahajan, Jongse Park, Hadi Esmaeilzadeh, "Yin-Yang: Programming Abstraction for Cross-Domain Multi-Acceleration,", in ***IEEE Micro**, special issue on Compiling for Accelerators*, 2022.

15. Jinwoo Hwang, Minsu Kim, Daeun Kim, Seungho Nam, Yoonsung Kim, Dohee Kim, Hardik Sharma, Jongse Park, "CoVA: Exploiting Compressed-Domain Analysis to Accelerate Video Analytics,", in *USENIX Annual Technical Conference (**ATC**)*, July 2022.

16. Seungbeom Choi, Sunho Lee, Yeonjae Kim, Jongse Park, Youngjin Kwon, and Jaehyuk Huh, "Serving Heterogeneous Machine Learning Models on Multi-GPU Servers with Spatio-Temporal Sharing,", in *USENIX Annual Technical Conference (**ATC**)*, July 2022.

17. S. Lee, J. Kim, S. Na, **J. Park**, and J. Huh, "TNPU: Supporting Trusted Execution with Tree-less Integrity Protection for Neural Processing Unit," in *IEEE International Symposium on High-Performance Computer Architecture (**HPCA**)*, February 2022. [To appear]

18. S. Na, S. Lee, Y. Kim, **J. Park**, and J. Huh, "Common Counters: Compressed Encryption Counters for Secure GPU Memory," in *IEEE International Symposium on High-Performance Computer Architecture (**HPCA**)*, February 2021.

19. S. Ghodrati, H. Sharma, S. Kinzer, A. Yazdanbakhsh, **J. Park**, N. Kim, D. Burger, and H. Esmaeilzadeh, "Mixed-Signal Charge-Domain Acceleration of Deep Neural Networks through Interleaved Bit-Partitioned Arithmetic," in *International Conference on Parallel Architectures and Compilation Techniques (**PACT**)*, October 2020.

20. Y. Li, **J. Park**, M. Alian, Y. Yuan, Q. Zheng, P. Pan, R. Wang, A. Schwing, H. Esmaeilzadeh, N. Kim, "A Network-Centric Hardware/Algorithm Co-Design to Accelerate Distributed Training of Deep Neural Networks," *IEEE/ACM International Symposium on Microarchitecture (**MICRO**)*, October 2018.

21. H. Sharma, **J. Park**, B. Samynathan, B. Robatmili, S. Mirkhani, H. Esmaeilzadeh, "From Tensors to FPGAs: Accelerating Deep Learning," *A Symposium on High Performance Chips (**Hot Chips**)*, August 2018.

22. H. Sharma, **J. Park**, N. Suda, L. Lai, B. Chau, J. Kim, V. Chandra, H. Esmaeilzadeh, "Bit Fusion: Bit-Level Dynamically Composable Architecture for Accelerating Deep Neural Networks," *ACM/IEEE International Symposium on Computer Architecture (**ISCA**)*, June 2018.

    ISCA 25-Year Retrospective 1996-2020 Inclusion

23. **J. Park**, H. Sharma, D. Mahajan, J. Kim, P. Olds, H. Esmaeilzadeh, "Scale-Out Acceleration for Machine Learning," in *IEEE/ACM International Symposium on Microarchitecture (**MICRO**)*, October 2017.

24. **J. Park**, E. Amaro, D. Mahajan, B. Thwaites, H. Esmaeilzadeh, "AxGames: Towards Crowdsourcing Quality Target Determination in Approximate Computing," in *ACM International Conference on Architectural Support for Programming Languages and Operating Systems (**ASPLOS**)*, April 2016.

25. H. Sharma, **J. Park**, D. Mahajan, E. Amaro, J. Kim, C. Shao, A. Mishra, H. Esmaeilzadeh "From High-Level Deep Neural Models to FPGAs," in *IEEE/ACM International Symposium on Microarchitecture (**MICRO**)*, October 2016.

26. D. Mahajan, **J. Park**, E. Amaro, H. Sharma, A. Yazdanbaksh, J. Kim, H. Esmaeilzadeh, "Tabla: A Unified Template-based Framework for Accelerating Statistical Machine Learning," in *IEEE International Symposium on High-Performance Computer Architecture (**HPCA**)*, March 2016.

    Distinguished Paper Award

27. D. Mahajan, A. Yazdanbaksh, **J. Park**, B. Thwaites, H. Esmaeilzadeh, "Towards Statistical Guarantees in Controlling Quality Tradeoffs in Approximate Acceleration," in *ACM/IEEE International Symposium on Computer Architecture (**ISCA**)*, June 2016.

28. A. Yazdanbakhsh, **J. Park**, H. Sharma, P. Lotfi-Kamran, H. Esmaeilzadeh, "Neural Acceleration for GPU Throughput Processors," in *IEEE/ACM International Symposium on Microarchitecture (**MICRO**)*, December 2015.

29. **J. Park**, H. Esmaeilzadeh, X. Zhang, M. Naik, W. Harris, "FlexJava: Language Support for Safe and Modular Approximate Programming," in *Joint Meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering (**ESEC/FSE**)*, September 2015.

30. A. Yazdanbakhsh, D. Mahajan, B. Thwaites, **J. Park**, A. Nagendrakumar, S. Sethuraman, K. Ramkrishnan, N. Ravindran, R. Jariwala, A. Rahimi, H. Esmailzadeh, K. Bazargan, "Axilog: Language Support for Approximate Hardware Design," in *Design Automation and Test in Europe (**DATE**)*, March 2015.

31. R. S. Amant, A. Yazdanbakhsh, **J. Park**, B. Thwaites, H. Esmaeilzadeh, A. Hassibi, L. Ceze, D. Burger, "General-Purpose Code Acceleration with Limited-Precision Analog Computation," in *ACM/IEEE International Symposium on Computer Architecture (**ISCA**)*, June 2014.

    ISCA 25-Year Retrospective 1996-2020 Inclusion

    Nominated for CACM Research Highlights; Honorable Mention in IEEE Micro Top Picks

32. B. Thwaites, G. Pekhimenko, A. Yazdanbakhsh, **J. Park**, G. Mururu, H. Esmaeilzadeh, O. Mutlu, T. Mowry, "Rollback-Free Value Prediction with Approximate Loads," in *International Conference on Parallel Architectures and Compilation Techniques (**PACT**)*, August 2014.

33. J. Choi, **J. Park**, J. Seol, and S. Maeng, "Isolated Mini-domain for Trusted Cloud Computing," in *IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing (**CCGrid**)*, May 2013.

34. **J. Park**, D. Lee, B. Kim, J. Huh, S. Maeng, "Locality-aware Dynamic VM Reconfiguration on MapReduce Clouds," in *IEEE InternationalSymposium on High-Performance Parallel and Distributed Computing (**HPDC**)*, June 2012.

Refereed Journal Articles

1. J. Cho, H. Choi, **J. Park**, "LLMServingSim2.0: A Unified Simulator for Heterogeneous Hardware and Serving Techniques in LLM Infrastructure,"in *IEEE Computer Architecture Letters (**CAL**)*, November 2025. (To Appear)

2. D. Kim, J. Hwang, C. Oh, **J. Park**, " MixDiT: Accelerating Image Diffusion Transformer Inference with Mixed-Precision MX Quantization," in *IEEE Computer Architecture Letters (**CAL**)*, June 2025.

3. H. Ham*, W. Yang*, Y. Shin, O. Woo, G. Heo, S. Lee, **J. Park**, G. Kim, "ONNXim: A Fast, Cycle-level Multi-core NPU Simulator," in *IEEE Computer Architecture Letters (**CAL**)*, December 2024.

4. S. Moon, S. Hong, M. Kim, D. Seo, J. Kim, R. Ko, S. Choi, J. Cha, J. Kim, S. Lim, H. Lee, H. Park, G. Choi, J. Kim, J. Lee, **J. Park**, J. Kim "LPU: A Latency-optimized and Highly Scalable Processor for Large Language Model Inference" in ***IEEE Micro**, special issue on Contemporary Industry Products*, 2024.

    Best Paper Award

5. S. Hwang, D. Baek, **J. Park**, J. Huh, "Cerberus: Triple Mode Acceleration of Sparse Matrix and Vector Multiplication," in *IEEE Transactions on Architecture and Code Optimization (**TACO**)*, 2024.

6. J. Park, S. Kang, S. Lee, T. Kim, **J. Park**, Y. Kwon, and J. Huh, "Hardware Hardened Sandbox Enclaves for Trusted Serverless Computing" in *IEEE Transactions on Architecture and Code Optimization (**TACO**)*, 2023.

7. S. Noh, J. Koo, S. Lee, **J. Park**, and J. Kung, "FlexBlock: A Flexible DNN Training Accelerator with Multi-Mode Block Floating Point Support" in *IEEE Transactions on Computers (**TC**)*, 2023.

8. S. Lee, R. Hwang, **J. Park**, and M. Rhu, "HAMMER: Hardware-friendly Approximate Computing for Self-attention with Mean-redistribution and Linearization" in *IEEE Computer Architecture Letters (**CAL**)*, 2023.

9. W. Seo, S. Cha, Y. Kim, J. Huh, and **J. Park**, " SLO-aware Inference Scheduler for Heterogeneous Processors in Edge Platforms" in *Transactions on Architecture and Code Optimization (**TACO**)*, 2021.

10. D. Mahajan, K. Ramkrishnan, R. Jariwala, A. Yazdanbakhsh, **J. Park**, B. Thwaites, A. Nagendrakumar, A. Rahimi, H. Esmaeilzadeh, K. Bazargan, "AXILOG: Abstractions for Approximate Hardware Design and Reuse," in ***IEEE Micro**, special issue on Alternative Computing Designs and Technologies*, October 2015.

Refereed Workshop Papers

1. J. Cho, M. Kim, H. Choi, **J. Park**, "LLMServingSim: A Simulation Infrastructure for LLM Inference Serving Systems", in *ISCA Workshop on ML for Computer Architecture and Systems (**MLArchSys**)*, June 2024.
2. Y. Lee, **J. Park**, "LVS: A Learned Video Storage for Fast and Efficient Video Understanding" in *Efficient Deep Learning for Computer Vision (**ECV**) in conjunction with CVPR*, June 2024 (To Appear).
3. H. Sharma, **J. Park**, E. Amaro, B. Thwaites, P. Kotha, A. Gupta, J. Kim, A. Mishra, H. Esmaeilzadeh, "DnnWeaver: From High-Level Deep Network Models to FPGA Acceleration," in *The Second Workshop on Cognitive Architectures (**CogArch**) in conjunction with ASPLOS*, April 2016.
4. D. Mahajan, A. Yazdanbakhsh, **J. Park**, B. Thwaites, H. Esmaeilzadeh, "Prediction-Based Quality Control for Approximate Accelerators," in *The Second Workshop on Approximate Computing Across the System Stack (**WACAS**) in conjunction with ASPLOS*, March 2015.
5. **J. Park**, K. Ni, X. Zhang, H. Esmaeilzadeh, M. Naik, "Expectation-Oriented Framework for Automating Approximate Programming,", in *The First Workshop on Approximate Computing Across the System Stack (**WACAS**) in conjunction with ASPLOS*, March 2014.
6. A. Yazdanbakhsh, B. Thwaites, **J. Park**, H. Esmaeilzadeh, "Methodical Approximate Hardware Design and Reuse," in *The First Workshop on Approximate Computing Across the System Stack (**WACAS**) in conjunction with ASPLOS*, March 2014.
7. A. Yazdanbakhsh, R. Amant, B. Thwaites, **J. Park**, H. Esmaeilzadeh, A. Hassibi, L. Ceze, D. Burger, "Toward General-Purpose Code Acceleration with Analog Computation," in *The First Workshop on Approximate Computing Across the System Stack (**WACAS**) in conjunction with ASPLOS*, March 2014.
8. B. Thwaites, A. Yazdanbakhsh, **J. Park**, H. Esmaeilzadeh, "Bio-Accelerators: Bridging Biology and Silicon for General-Purpose Computing," in *Wild and Crazy Ideas (**WACI**) in conjunction with ASPLOS*, March 2014.

Research Experience

Research Assistant. Alternative Computing Technology (ACT) Lab — Aug. 2013–Aug. 2018
- Georgia Institute of Technology
- Advisor: Prof. Hadi Esmaeilzadeh

Visiting Researcher. Alternative Computing Technology (ACT) Lab — Jan. 2018–Aug. 2018
- University of California, San Diego
- Advisor: Prof. Hadi Esmaeilzadeh

Research Intern. Architecture Research Group (ARG) — May 2017–Aug. 2017
- NVIDIA Research
- Mentors: Dr. Arslan Zulfiqar and Dr. Eiman Ebrahimi
- Manager: Dr. Stephen Keckler

Research Intern. Catapult team — Jan. 2016–May 2016
- Microsoft Research
- Mentor: Dr. Eric Chung
- Manager: Dr. Doug Burger

Research Assistant. Computer Architecture (CA) Lab — Feb. 2010–Jul. 2013
- Korea Advanced Institute of Science and Technology (KAIST)
- Advisor: Prof. Seungryoul Maeng

Teaching Experience

Instructor.

• CS230:	System Programming	Fall 2024
• CS610:	Parallel Processing	Spring 2024
• CS311:	Computer Organization	Spring 2024
• CS411:	System for Artificial Intelligence	Fall 2023
• CS510:	Computer Architecture	Spring 2023
• CS230:	System Programming	Fall 2022
• CS311:	Computer Organization	Spring 2022
• CS230:	System Programming	Fall 2021
• CS492:	Special Topic in Computer Science: System for Artificial Intelligence	Spring 2021
• CS230:	System Programming	Fall 2020
• CS492:	Special Topic in Computer Science: System for Machine Learning	Spring 2020

Teaching Assistant.

• CS3220:	Processor Design	Georgia Institute of Technology	Fall 2016
• CS3220:	Processor Design	Georgia Institute of Technology	Fall 2014
• CS8803:	Alternative Computing Technology	Georgia Institute of Technology	Spring 2014
• CS211:	Digital System and Lab	KAIST	Spring 2011
• CS311:	Embedded Computer Systems	KAIST	Fall 2010

References Available to Contact

Hadi Esmaeilzadeh. Professor, UCSD — hadi@eng.ucsd.edu

Nam Sung Kim Professor, UIUC — nskim@illinois.edu

Doug Burger. Technical Fellow and Corporate VP, Microsoft Research — dburger@microsoft.com

Eric Chung. VP of AI Computing, NVIDIA — eschung@nvidia.com