MICHAEL JAEMIN KIM

michael.kim.604@gmail.com ◇ michael604@scale.snu.ac.kr
website-link ◇ google-scholar-link

NATIONALITY

United States of America & Republic of Korea (Dual Nationality)

PROFESSIONAL APPOINTMENTS

Research Scientist, **Meta, AI and Systems Codesign** June 2025 - Present

INDUSTRY PROJECTS

Meta, AI and Systems Codesign — Research Scientist June 2025 - Present

- **AMD MI350X LLM/LDM Inference Enablement & Optimization.** Led enablement and optimization of LLM and LDM inference on AMD MI350X across multiple production models on vLLM and Gen2 stacks. Optimized attention kernels, Triton fused-MoE heuristics, kernel fusion, and auto-tuning configurations. Achieved 2× throughput over H100/MI300X at iso-latency for LLM (40% and 31% latency reduction), 2.14× on Flux diffusion (1.59× vs H100, deployed to production), 3.5–4.3× on AVG lite, 1.2× faster than B200 on Unigen-1.5-MoE.
- **AI-Assisted Iterative Kernel Optimization for MI350X.** Developed an automated kernel optimization loop combining AI-guided search with hardware-specific tuning for LLM inference. Wrote CK tile RMSNorm achieving 94% peak memory bandwidth (1.4–2.3× over baseline), FP8 fused MoE with 17% speedup via XCD tile remap, and offline-tuned AITER MoE kernels (15% faster at large batch).
- **Power Roofline Analysis** (submitted to MICRO 2026). Co-developed a power roofline methodology for GPU inference. Discovered energy-efficiency optimal operating point below max TDP for datacenter compute maximization under power cap. Explained kernel auto-tuning ranking flips from power excursion and characterized power-bound decode attention kernels.
- **Blackwell FP4 GEMM Root-Cause Analysis.** Investigated GB300 vs. GB200 FP4 GEMM regression (1.2× achieved vs. 1.5× expected). Root-caused to TMA load latency exposure: tensor core throughput scaled 1.5× but shared memory capacity and TMA latency remained unchanged (long_scoreboard 79–89%).
- **Delta Compression for GEMM Decomposition.** Explored delta compression as lossless sparsification to reduce GEMM power. Best-case yielded ∼22% exact zeros (from 0%), below the 75–90% threshold needed; temporal delta on diffusion cross-attention K/V reached 97% zeros, but spatial delta on weights decreased sparsity (0.73–0.84×).
- **GPU Benchmarking Infrastructure.** Built unified GPU benchmarking repository with automated sweep, power/clock telemetry, time-series dashboard, and auto-tuning integration.

EDUCATION

B.S.E. in Electrical and Computer Engineering, **Seoul National University** Aug 2019
Ph.D, **Seoul National University (Prof. Jung Ho Ahn)** Aug 2019 - 2025
Visiting Research, **University of Illinois Urbana-Champaign (Prof. Nam Sung Kim)** Mar 2023 - Mar 2024

RESEARCH INTERESTS

My research focuses on improving the efficiency of GPUs on Large Language Model (LLM) and Latent Diffusion Models (LDM) inference, with regards to power efficiency, higher performance kernels, KV cache offloading and prefix caching to CPU memory, disaggregated and/or CXL memory, or SSD for agentic and multi-turn inference. Also, I have interest in maximizing sparsity of the model weights and activations. I have also interest in improving the HBM architecture or improving the memory hierarchy of the GPU systems.

PUBLICATION

- [S&P 2026] "SoK: Systematizing a Decade of Architectural RowHammer Defenses Through the Lens of Streaming Algorithms",
  M. J. Kim, Seungmin Baek, Jumin Kim, Hwayong Nam, Nam Sung Kim, Jung Ho Ahn, in IEEE Symposium on Security and Privacy, 2026
- [HPCA 2026] "RoMe: Row Granularity Access Memory System for Large Language Models",
  Hwayong Nam, Seungmin Baek, Jumin Kim, **M. J. Kim**, Jung Ho Ahn, in Proceeding of 32nd IEEE International Symposium on High Performance Computer Architecture, 2026
- [HPCA 2026] "ReScue: Reliable and Secure CXL Memory",
  Chihun Song, Austin Antony Cruz, **M. J. Kim**, Minbok Wi, Gaohan Ye, Kyungsan Kim, Sangyeol Lee, Jung Ho Ahn, Nam Sung Kim, in Proceeding of 32nd IEEE International Symposium on High Performance Computer Architecture, 2026
- [CAL 2025] "Per-row activation counting on real hardware: Demystifying performance overheads",
  Jumin Kim, Seungmin Baek, Minbok Wi, Hwayong Nam, **M. J. Kim**, Sukhan Lee, Kyomin Sohn, Jung Ho Ahn, in IEEE Computer Architecture Letters, 2025.
- [ASPLOS 2025] "Marionette: A RowHammer Attack via Row Coupling",
  Seungmin Baek, Minbok Wi, Seonyong Park, Hwayong Nam, **M. J. Kim**, Nam Sung Kim, Jung Ho Ahn, in ACM International Conference on Architectural Support for Programming Languages and Operating Systems, 2025
- [ASPLOS 2025] "M5: Mastering page migration and memory management for CXL-based tiered memory systems",
  Yan Sun, Jongyul Kim, Douglas Yu, Jiyuan Zhang, Siyuan Chai, **M. J. Kim**, Hwayong Nam, Jaehyun Park, Eojin Na, Yifan Yuan, Ren Wang, Jung Ho Ahn, Tianyin Xu, Nam Sung Kim, in ACM International Conference on Architectural Support for Programming Languages and Operating Systems, 2025
- [CAL 2025] "X-PPR: Post Package Repair for CXL Memory",
  Chihun Song, **M. J. Kim**, Yan Sun, Houxiang Ji, Kyungsan Kim, TaeKyeong Ko, Jung Ho Ahn, Nam Sung Kim, in IEEE Computer Architecture Letters, 2025
- [IEEE ESL 2024] "Hechi: A Hybrid Approach for Efficient Memory Reclamation Techniques in Mobile Systems",
  Wanju Doh, Seoyoung Ko, **M. J. Kim**, Jung Ho Ahn in IEEE Embedded System Letters, 2024
- [ISCA 2024] "DRAMScope: Uncovering DRAM Microarchitecture and Characteristics by Issuing Memory Commands",
  Hwayong Nam, Seungmin Baek, Minbok Wi, **M. J. Kim**, Jaehyun Park, Chihun Song, N. S. Kim, J. Ahn, in Proceedings of the 51st annual international symposium on computer architecture, 2024
- [ASPLOS 2024] "TAROT: A CXL SmartNIC-Based Defense Against Multi-bit Errors by Row-Hammer Attacks",
  C. Park, **M. J. Kim**, Tianchen Wang, Houxiang Ji, Jinghan Huang, Ipoom Jeong, Jaehyun Park, Hwayong Nam, Minbok Wi, J. Ahn, N. S. Kim, in ACM International Conference on Architectural Support for Programming Languages and Operating Systems, 2024
- [ASPLOS 2024] "AttAcc! Unleashing the Power of PIM for Batched Transformer-based Generative Model Inference",
  J. Park, J. Choi, K. Kyung, **M. J. Kim**, Y. Kwon, N. S. Kim, J. Ahn, in ACM International Conference on Architectural Support for Programming Languages and Operating Systems, 2024
- [MICRO 2023] "How to Kill the Second Bird with One ECC: The Pursuit of Rowhammer Resilient DRAM",
  M. J. Kim, J. Park, M. Wi, S. Ko, J. Park, H. Nam, N. S. Kim, E. Lee, J. Ahn, in IEEE/ACM International Symposium on Microarchitecture, 2023, DOI:10.1145/3613424.3623777
- [CAL 2023] "X-ray: Discovering DRAM internal structure and error characteristics by issuing memory commands.",

H. Nam, S. Baek, M. Wi, **M. J. Kim**, J. Park, C. Song, N. S. Kim, J. Ahn, in IEEE Computer Architecture Letters, 2023. DOI:10.1109/LCA.2023.3296153

- [HPCA 2023] "SHADOW: Preventing Row Hammer in DRAM with Intra-Subarray Row Shuffling",
  M. Wi, J. Park, **M. J. Kim**, S. Ko, N. S. Kim, E. Lee, J. Ahn, in Proceeding of 29th IEEE International Symposium on High Performance Computer Architecture, 2023. DOI:10.1109/HPCA56546.2023.10070966
- (arXiv) "AESPA: Accuracy Preserving Low-degree Polynomial Activation for Fast Private Inference",
  J. Park, **M. J. Kim**, W. Jung, J. Ahn, 2022. arXiv
- [IEEE TC 2022] "Future Scaling of Memory Hierarchy for Tensor Cores and Eliminating Redundant Shared Memory Traffic Using Inter-Warp Multicasting",
  S. Lee, S. Hwang, **M. J. Kim**, J. Choi, and J. Ahn, IEEE Transactions on Computers, Early Access, 2022. DOI:10.1109/TC.2022.3207134
- [ISCA 2022] "BTS: An Accelerator for Bootstrappable Fully Homomorphic Encryption",
  S. Kim, J. Kim, **M. J. Kim**, W. Jung, M. Rhu, J. Kim, J. Ahn, in Proceedings of the 49th annual international symposium on computer architecture, 2022. DOI:10.1145/3470496.3527415
- [HPCA 2022] "Mithril: Cooperative Row Hammer Protection on Commodity DRAM Leveraging Managed Refresh",
  M. J. Kim, J. Park, Y. Park, W. Doh, N. Kim, T. Ham, J. Lee, and J. Ahn, in Proceeding of 28th IEEE International Symposium on High Performance Computer Architecture, 2022. DOI:10.1109/HPCA53966.2022.00088

SERVICE

- MICRO PC member 2026
- IEEE Transactions on Computer reviewer 2024
- IEEE Computer Architecture Letters reviewer 2025-2026

EXPERIENCE

- SNU MMS Lab Internship (Professor Dongsuk Jeon) Dec 2018 - Feb 2019
  - Verilog design project for spike neural network accelerator.
- SK Hynix Internship Dec 2017
  - Verilog design project for toy memory controller.
- Mandatory military service for ROK army, at Korea Military Academy Nov 2014 - Aug 2016
  - Key Resolve ROK-US Combined Exercise at Ministry of National Defense. Mar 2016

SKILLS

Languages C/C++, Verilog, Python
Tools Architectural simulation (*e.g.*, Ramulator, McSimA+), Intel Pin, SPEC2006/2017, SPLASH-2/3, GAPBS, GoogleTest, Pandas, Docker