Minsung Jang

mdpe36kr@gmail.com

https://minsungmj.github.io

AI Systems & Accelerated Computing · Full-Stack Systems for Scalable and Efficient AI

RESEARCH PROFILE

My research develops **full-stack systems software for large-scale AI infrastructure**, spanning accelerator-aware inference runtimes, heterogeneous GPU clusters, AI datacenter networking, cloud systems, and confidential execution. I study how AI workloads should be mapped, scheduled, and optimized across accelerators, memory, networks, and failure domains. My work combines systems research with large-scale deployment experience at Samsung SDS and AT&T Labs Research, with recent results in HPCA, EuroSys, IEEE BigData, IPCCC, and MCCSys.

RESEARCH INTERESTS

AI Infrastructure Systems · **Systems for Machine Learning** · **LLM Inference Serving** · **GPU Cluster Systems** · **AI Datacenter Networking** (RDMA/RoCE) · **Cloud Systems** · **Accelerator-aware Runtime Optimization** (PIM, quantization) · **Secure & Confidential AI Execution**

EDUCATION

Ph.D., Computer Science
Georgia Institute of Technology, Atlanta, GA, USA — August 2008 – August 2015
Dissertation: *Virtual Platforms: System Support to Enrich the Functionality of End Client Devices*
Advisor: Prof. Karsten Schwan

M.S., Mechanical Engineering
Yonsei University, Seoul, Korea — February 2000

B.S., Mechanical Design and Production Engineering
Yonsei University, Seoul, Korea — February 1998

SELECTED RESEARCH CONTRIBUTIONS

Selected for relevance to AI systems and accelerator research; publication status is stated explicitly.

- **PAISE** (IEEE HPCA 2025) — GPU–PIM scheduling that offloads memory-bound attention to HBM-PIM, reducing LLM inference time by up to 48.3% versus GPU-only execution.
- **Omni-MP** (MCCSys @ ACM ICS 2026, to appear) — generalized PIM-GEMV mapping for arbitrary LLM matrices and KV cache, accelerating attention and FFN computation by 40.2% on average versus GPU-only execution on real HBM-PIM.
- **FineServe** (preprint, under review; arXiv:2509.06261) — precision-aware KV-cache management and two-level scheduling for mixed-precision LLMs, achieving up to 2.2× higher SLO attainment and 1.8× higher throughput.
- **JABAS** (EuroSys 2025; collaboration with UNIST) — joint adaptation of global batch size and GPU allocation, reducing training time by 33.3% and cost by 54.2% on heterogeneous GPU clusters without accuracy loss.
- **CORN** (IEEE IPCCC 2023) — end-host RoCEv2 congestion control and adaptive multipathing for GPU interconnects, removing dependence on switch-level priority flow control.
- **DSDE** (IEEE BigData 2025, 11th Special Session on Intelligent Data Mining) — training-free, KLD-stability-driven adaptation of speculative-decoding length under low and variable acceptance rates.

PUBLICATIONS

Candidate name underlined. Contribution and leadership roles for recent collaborative work are described in the preceding and leadership sections.

Refereed Conference & Workshop Publications

- H. Lee, D. Baek, J. Son, J. Choi, K. Moon, M. Jang. *PAISE: PIM-Accelerated Inference Scheduling Engine for Transformer-based LLM.* In *Proceedings of the 31st IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 2025.
- G. Yun, J. Kang, H. Jeong, S. Eom, M. Jang, Y.-r. Choi. *JABAS: Joint Adaptive Batching and Automatic Scaling for DNN Training on Heterogeneous GPUs.* In *Proceedings of the Twentieth European Conference on Computer Systems (EuroSys)*, 2025.
- M. Yang, J.-Y. Choi, K. Moon, M. Jang, E. Jeon. *DSDE: Dynamic Speculative Decoding with KLD Stability for Real-World Serving.* In *Proceedings of the 2025 IEEE International Conference on Big Data (IEEE BigData)*, 11th Special Session on Intelligent Data Mining, pp. 3115–3124, 2025.
- D. Baek, J. Son, J. Choi, K. Bin, S. Choi, K. Moon, M. Jang, H. Lee. *Omni-MP: Practical PIM Matrix Mapping for Accelerating LLM Inference on Real PIM Hardware* (to appear). In *Proceedings of the 7th Workshop on Memory-Centric Computing Systems (MCCSys), in conjunction with the ACM International Conference on Supercomputing (ICS)*, 2026.
- H. Lee, L. Vu, M. Jang. *Economics of Spot Instance Service: A Two-Stage Dynamic Game Approach.* In *Proceedings of the 2023 IEEE International Conference on Cloud Computing (IEEE CLOUD)*, 2023.
- J.-H. Cha, S. Kang, Y. Kang, H. Seo, J. Lee, J. Kim, M. Jang. *CORN: Cloud-optimized RDMA Networking.* In *Proceedings of the 2023 IEEE International Performance, Computing, and Communications Conference (IPCCC)*, 2023.
- R. M. Krishnan, W.-H. Kim, X. Fu, S. K. Monga, H. W. Lee, M. Jang, A. Mathew, C. Min. *TIPS: Making Volatile Index Structures Persistent with DRAM-NVMM Tiering.* In *Proceedings of the 2021 USENIX Annual Technical Conference (USENIX ATC)*, 2021.
- E. F. Boza, C. L. Abad, S. P. Narayanan, B. Balasubramanian, M. Jang. *A Case for Performance-Aware Deployment of Containers.* In *Proceedings of the 5th International Workshop on Container Technologies and Container Clouds (WoC)*, 2019.
- H. Gupta, A. Sharma, A. Zelezniak, M. Jang, U. Ramachandran. *A Black-Box Approach for Estimating Utilization of Polled IO Network Functions.* In *Proceedings of the 11th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud)*, 2019.
- M. Jang, H. Lee, K. Bhardwaj, K. Schwan. *SOUL: An Edge-Cloud System for Mobile Applications in a Sensor-Rich World.* In *Proceedings of the 2016 IEEE/ACM Symposium on Edge Computing (SEC)*, 2016.
- M. Jang, H. Lee, K. Bhardwaj, K. Schwan. *vSensor: Toward Sensor-rich Mobile Applications.* In *Proceedings of the Sensors to Cloud Architectures Workshop (SCAW), in conjunction with HPCA-21*, 2015.
- M. Jang, K. Schwan, K. Bhardwaj, A. Gavrilovska, A. Avasthi. *Personal Clouds: Sharing and Integrating Networked Resources to Enhance End User Experiences.* In *Proceedings of the 2014 IEEE International Conference on Computer Communications (IEEE INFOCOM)*, 2014. (Acceptance rate: 19.4%)
- M. Jang, K. Schwan. *Clouds4Users: From Isolated Devices to Rich Device Platforms* (poster). In *Proceedings of the 3rd Workshop on SoCs, Heterogeneous Architectures and Workloads (SHAW-3), in conjunction with HPCA-18*, 2012.
- M. Jang, K. Schwan. *STRATUS: Assembling Virtual Platforms from Device Clouds.* In *Proceedings of the 2011 IEEE International Conference on Cloud Computing (IEEE CLOUD)*, 2011. (Acceptance rate: 18%)
- S. Heo, M. Jang, S. Suh. *Software Mobility for Personalized Computing Environments on Removable Storage.* In *Proceedings of the 2008 IEEE Consumer Communications and Networking Conference (IEEE CCNC)*, 2008.
- J. Hwang, J. Bae, A. Kirnasov, M. Jang, H. Kim. *A Reliable and Portable Multimedia File System.* In *Proceedings of the 2006 Linux Symposium*, 2006.

Preprints / Under Review

- K. Bin, S. Choi, J. Son, J. Choi, D. Bae, D. Baek, K. Moon, M. Jang, H. Lee. *FineServe: Precision-Aware KV Slab and Two-Level Scheduling for Heterogeneous Precision LLM Serving.* Preprint (under review), arXiv:2509.06261, 2025.
- D. Baek, J. Choi, J. Son, K. Bin, S. Choi, K. Moon, M. Jang, H. Lee. *FireQ: Fast INT4-FP8 Kernel and RoPE-aware Quantization for LLM Inference Acceleration.* Preprint (under review), arXiv:2505.20839, 2025.
- S. Lee, B. Woo, H. Kang, H. Choe, H. Lee, K. Moon, M. Jang, B. Kang. *Lightweight and Efficient Nested Enclaves for AMD SEV-SNP Confidential VMs.* Under review
- S. Choi, J. Goo, E. Jeon, M. Yang, M. Jang. *ELIS: Efficient LLM Iterative Scheduling System with Response Length Predictor.* Preprint, arXiv:2505.09142, 2025.
- S. Kim, Y. Kim, K. Moon, M. Jang. *LaDiMo: Layer-wise Distillation Inspired MoEfier.* Preprint, arXiv:2408.04278, 2024.

RESEARCH & TECHNICAL LEADERSHIP

Samsung SDS August 2021 – Present

Executive Advisor (VP-level), Seoul, Korea January 2026 – Present

Vice President & Head, Cloud Research Team, Seoul, Korea August 2021 – December 2025

- **Research organization and agenda:** directed a multi-site research organization spanning computing systems, networks, and high-performance computing; established an AI-systems agenda across GPU clusters, LLM serving, AI datacenter networking, and accelerator-aware runtimes.
- **Accelerator-aware AI inference:** led research on real HBM-PIM hardware, producing GPU–PIM scheduling (PAISE) and practical PIM matrix mapping (Omni-MP), together with low-precision kernels (FireQ) and precision-aware serving (FineServe).
- **GPU cluster and serving systems:** led the architecture and development of Samsung's Kubernetes-based GPU-as-a-Service and x.Cloud distributed training/inference platform, covering GPU-aware scheduling, multi-tenant isolation, inference endpoints, observability, and resource-efficiency controls.
- **AI datacenter networking and cloud systems:** directed CORN, an end-host RoCEv2 congestion-control and adaptive-multipath design later implemented as a hardware-NIC prototype with AMD and demonstrated at SC25; led architecture for cloud and storage control-plane technologies on Samsung Cloud Platform.
- **Research translation and collaboration:** built an industry systems-research program with results at HPCA, IEEE BigData, IPCCC, EuroSys, and MCCSys; led public academic collaborations with UC Berkeley, UNIST, KAIST, and Seoul National University.

Peraton Labs (formerly Perspecta Labs) September 2020 – July 2021

Senior Research Scientist, Basking Ridge, NJ, USA

- Applied research on networking and distributed systems under performance, resilience, and operational constraints.

AT&T Labs Research August 2015 – July 2020

Principal-Inventive Scientist, Bedminster, NJ, USA

- **GPU-accelerated networking systems:** architected a GPU-based NFV data plane to accelerate stateful, compute-intensive network functions to NVIDIA GPUs (presented at NVIDIA GTC 2018, "Practical GPU-Based Network Packet Processing"); foundation for granted patents.
- **Accelerator abstraction for high-performance software-defined network:** led R&D on cloud-native network services (5G and Internet) on the commodity server with hardware acceleration (e.g., smartNICs, FPGAs, and ASICs); contributed to the **O-RAN Alliance Working Group 6** on accelerator abstraction and O-Cloud architecture.
- **Persistent-memory systems:** co-designed and evaluated a persistent-memory key-value store for telco workloads (Intel Optane), contributing to systems research published at USENIX ATC.
- **Cloud-native infrastructure:** contributed to the design and implementation of the AT&T Network Cloud for carrier-grade 5G and Internet services across the United States.

Samsung Electronics & Samsung Advanced Institute of Technology August 2003 – August 2008
Research Staff Member, Kiheung, Korea

- Built **XenARM**, hypervisor technology to virtualize ARM-based platforms for embedded devices; systems-software and virtualization research for next-generation mobile/consumer platforms.

GRADUATE RESEARCH INTERNSHIPS

- **Intel Labs**, Integrated Platform Research, Hillsboro, OR Summer 2012
- **Nokia Research Center**, North America Research, Palo Alto, CA Summer 2011
- **IBM Almaden Research Center**, Storage Systems Group, San Jose, CA Summer 2010

SPONSORED & COLLABORATIVE RESEARCH

- **UC Berkeley Sky Computing Lab** (Prof. Ion Stoica), May 2022 – December 2025 — led Samsung SDS's participation as a founding sponsor and research collaborator on AI infrastructure systems, intercloud orchestration, LLM serving, and AI datacenter networking.
- **UNIST** (Prof. Young-ri Choi), January – December 2022 — joint research on adaptive batching and automatic scaling for heterogeneous-GPU training; research outcome: JABAS, EuroSys 2025.
- **KAIST Cyber Security Systems Research Lab** (Prof. Brent Byunghoon Kang), November 2024 – November 2025 — joint research on confidential computing and lightweight nested enclaves for secure AI execution; manuscript under review.
- **Seoul National University** (Prof. Taekyoung Kwon), January – December 2025 — joint research on replay-attack-resistant token validation; research outcome by the SNU team: "A Replay-Attack Resistant and Lightweight Token Validation Framework with Hash Chain," IEEE ICDCS 2026.

PROFESSIONAL & TECHNICAL SERVICE

- **Program Committee:** SOSP 2025 Poster Session.
- **External Reviewer:** IEEE Transactions on Circuits and Systems I: Regular Papers (May 2026); IEEE Transactions on Computers (June 2026); IEEE Transactions on Parallel and Distributed Systems (June 2026); IEEE Computer (March 2017); Future Generation Computer Systems (July 2016); IEEE Internet Computing (March 2016); IEEE Transactions on Industrial Informatics (December 2015); IEEE Micro (August 2014).
- **Ultra Ethernet Consortium (UEC):** Samsung representative for next-generation AI-datacenter networking November 2023 – Present
- **O-RAN Alliance Working Group 6:** AT&T representative and contributor on O-Cloud and accelerator abstraction 2019 – 2020
- **K-Perf NPU Performance-Evaluation Initiative:** represented Samsung SDS as the demand-side cloud-service-provider participant in the Ministry of Science and ICT / NIPA-led public–private initiative; contributed evaluation conditions and metrics for real-world LLM serving.

INVITED TALKS (selected)

- **SC25 (Supercomputing 2025)**, Technical Contributor and Co-presenter with AMD — CORN hardware-NIC prototype, November 2025.
- **Intel AI Summit Seoul 2025**, Invited Speaker — "Performance Analysis of Intel Gaudi 3 for LLM Inference," Seoul, July 2025.
- **Seoul National University**, Invited Lecture — "Trends in AI Infrastructure Technologies," Seoul, September 2024.
- **Korea Computer Congress (KCC) 2024**, Invited Speaker — "Cloud Technologies and Vision Behind Generative AI," Jeju, June 2024.
- **UC Berkeley Sky Computing Retreats 2024**, Collaboration Partner Speaker — "SkyAirFlow: Sky-powered Intercloud Workload Orchestration" (May 2024) and "Samsung Cloud Platform's Integration with Sky" (Jan. 2024).

- **High Speed Network (HSN) 2024**, Invited Speaker — "Emerging Technologies in Cloud Computing 2024 and Beyond," Jeju, January 2024.
- **KIISE Computer System Society Winter Workshop 2023**, Invited Speaker — "x.Cloud: Samsung SDS GPU-based AI Development Platform," Pyeongchang, February 2023.
- **NVIDIA GTC 2021**, Co-speaker — "High-performance/High-efficiency AI Model Training Cluster Based on Kubernetes (x.Cloud)," November 2021.
- **NVIDIA GTC 2018**, Technical Contributor — "Practical GPU-Based Network Packet Processing: Building the Next Generation Internet Router," San Jose, CA.
- Research presentations at **IEEE/ACM SEC** (2016), **IEEE INFOCOM** (2014), and **IEEE CLOUD** (2011); invited research talks at IBM Almaden (2014) and AT&T Labs Research (2015).

SELECTED PATENTS

- **Persistent kernel for GPU direct-memory-access network packet processing** — co-inventor; persistent GPU kernels + DMA for high-throughput NFV packet processing. (US 10,795,840 and related family)
- **Direct memory access for GPU packet processing** — co-inventor; efficient NIC–GPU DMA to reduce data movement for GPU-based network functions. (US 10,332,235 and related family)
- **Scaling network capability using baseband unit pooling in 5G and beyond** — co-inventor; BBU pooling and dynamic resource allocation for cloud-native vRAN. (US 11,582,642 and related family)
- **Operating a mobile virtual environment upon connection to a host computer** — co-inventor; running personal virtual environments from removable storage via virtualization. (US 10,007,541 and related family)
- **Managing file-system data using a database management system** — co-inventor; DBMS-backed file-system metadata for reliability and scalability. (US 9,384,201 and related family)

HONORS & AWARDS

- Intel Ph.D. Fellowship nominee, Georgia Institute of Technology, 2013–2014
- Intel Kudos Award, Intel CountryFair, Intel Corporation, 2012
- Excellent Research Project Award, Software Laboratories, Samsung Electronics, 2003

TECHNICAL EXPERTISE

AI/ML Systems & Accelerators: LLM serving, GPU-aware scheduling, speculative decoding, KV-cache management, quantization, HBM-PIM, CUDA, GPUDirect RDMA, INT4/FP8 kernels, performance profiling **Networking & Data Plane:** RDMA/RoCEv2, DPDK, XDP/eBPF, SDN, 5G vRAN, AI-datacenter networking **Cloud, OS & Storage:** Kubernetes, OpenStack, virtualization, distributed systems, storage control planes, persistent-memory systems