Junsu Kim

145, Anam-ro, Seongbuk-gu, Seoul, Republic of Korea (Korea University)
☎ (+82) 10-8684-3631 ✉ j0807s@korea.ac.kr github.com/j0807s 🏠 j0807s.github.io

Research Interests

Computer Architecture, Memory Systems, Systems for ML & ML for Systems

Education

Korea University, Seoul, Korea Sep. 2023 - Current
M.S. in Electrical Engineering (Advisor: Prof. Yunho Oh)
Cumulative GPA: 4.0/4.0

Hanyang University, Seoul, Korea Mar. 2014 - Feb. 2021
B.S. in Electronic Engineering (Advisor: Prof. Ki-Seok Chung)
Cumulative GPA: 3.81/4.0 (Graduating with Honors - Summa Cum Laude)

Publications

Conference Papers

[C4] **Junsu Kim**, and Suhyun Kim, "Salient Frequency-aware Exemplar Compression for Resource-constrained Online Continual Learning", The 39th Annual AAAI Conference on Artificial Intelligence (AAAI) 2025.

[C3] Minseong Gil, Jaebeom Jeon, **Junsu Kim**, Sangun Choi, Gunjae Koo, Myung Kuk Yoon, and Yunho Oh, "TLP Balancer: Predictive Thread Allocation for Multi-Tenant Inference in Embedded GPUs", The IEEE Embedded Systems Letters.

[C2] Jaebeom Jeon, Minsung Gil, **Junsu Kim**, Jaeyoung Park, Gunjae Koo, Myung-Kuk Yoon, and Yunho Oh. "VitBit: Enhancing Embedded GPU Performance for AI Workloads through Register Operand Packing". The 53rd International Conference on Parallel Processing (ICPP), 2024

[C1] Kwangrae Kim, Jeonghyun Woo, **Junsu Kim**, and Ki-Seok Chung. "HammerFilter: Robust Protection and Low Hardware Overhead Method for RowHammer". The 39th IEEE International Conference on Computer Design (ICCD), 2021

Preprints (Project Names Only)

[P1] First author, "Memory Oversubscription-Aware Tensor Migration Scheduling for GPU Unified Storage Architecture" *Under Review in CAL*

[P2] Co-author, "Mitigating Software Overhead in Tiered Memory-based Accelerator for Training" *Under Review in ISCA'25*

[P3] Co-author, "Hardware Supports for Enabling Arbitrary Numeric Format on GPUs" *Under Review in ISCA'25*

[P4] Co-author, "A Behavioral Analysis of CXL Memory Systems" *Under Review in SIGMETRICS'25*

[P5] Co-author, "Accelerating Yinyang K-Means on Heterogeneous Platform" *Under Review in IPDPS'25*
Got 3 reviews of borderlines, stating "This paper raise concerns that can be addressed during revision"

[P6] Co-author, "Improving Performance of Data-Free Continual Learning" *Ready for Submission*

Work Experience

Korea University, Seoul, Korea Sep. 2023 - Current
Research Assistant at Computer Architecture and System Software Lab (ComSys) Advisor: Prof. Yunho Oh

Korea Institute of Science and Technology, Seoul, Korea May. 2022 - Aug. 2023
Research Assistant at Korea Data Science Team (KDST) Supervisor: Dr. Suhyun Kim

Hanyang University, Seoul, Korea Dec. 2019 - Mar. 2020, Aug. 2020 - Nov. 2020
Research Assistant at Embedded System on Chip Laboratory (ESOC Lab) Advisor: Prof. Ki-Seok Chung
Research Assistant at Computer Architecture and System SW Lab (CASS Lab) Advisor: Prof. Yongjun Park

Teaching Experience

Korea University, Seoul, Korea Spring 2024, Fall 2024
Teaching Assistant for Computer Architecture

School for the Blind, Chuncheon, Korea Mar. 2017 - Feb. 2019
Assistant Teacher (Alternative Military Service)

Selected Research Projects

Mitigating Software Overhead in Tiered Memory-based Accelerator for Training [Under Review]
Advisor: Prof. Yunho Oh, Korea University Mar. 2024 - Nov. 2024

- ◇ Analyzed FTL overhead due to frequent promotion and demotion bottlenecked AI training as model sizes grow
- ◇ Proposed a unified address translation with dedicated IOMMU for each accelerator to reduce address translation overhead
- ◇ Proposed a migration scheduler that prefetches tensors at runtime, leveraging the predictability of AI workloads
- ◇ Contributions: co-author, motivation study, idea, implementation, paper write-up

Hardware Supports for Enabling Arbitrary Numeric Format on GPUs [Under Review]
Advisor: Prof. Yunho Oh, Korea University May. 2024 - Nov. 2024

- ◇ Observed GPU supports a limited set of numeric formats, wasting register files when processing arbitrary numeric formats
- ◇ Employed bitslice representation, which transposes the data elements, packing arbitrary numeric formats without register wastage.
- ◇ Proposed Bitslice Vector multiplier and adder, constructing a tree structure to replace a multiplication-adder tree in a Tensor core
- ◇ Contributions: co-author, motivation study, idea, implementation, paper write-up

Memory Oversubscription-Aware Tensor Migration Scheduling for GPU Unified Storage Architecture [Under Review]
Advisor: Prof. Yunho Oh, Korea University Feb. 2024 - Sep. 2024

- ◇ Analyzed page faults due to memory oversubscription stalled AI workloads when expanding GPU memory with SSD using UVM
- ◇ Proposed a tensor migration scheduling algorithm considering GPU memory oversubscription for GPU unified storage architecture
- ◇ Achieved the averaged speedup by 12.9% compared to G10, which was presented at MICRO 2023
- ◇ Contributions: 1st author, motivation study, idea, implementation, experiment, paper write-up

A Behavioral Analysis of CXL Memory Systems [Under Review]
Advisor: Prof. Yunho Oh, Korea University Sep. 2023 - Sep. 2024
Collaborator: SK hynix

- ◇ Observed the behavior of a real CXL-based system on datacenter and AI workloads in the CXL-based platform
- ◇ Analyzed how the different promotion and demotion methods for CXL devices affected the performance of the workloads
- ◇ Presented performance modeling for datacenter workloads using different system factors (e.g., memory bandwidth, memory latency)
- ◇ Contributions: co-author, experiment, analysis, paper write-up

Accelerating Yinyang K-Means on Heterogenous Platform [Under Review]
Advisor: Prof. Yunho Oh, Korea University Jun. 2024 - Sep. 2024

- ◇ Analyzed that CPU was underutilized while executing Yinyang K-Means clustering on embedded GPUs
- ◇ Observed warp divergence caused by checking boundary conditions for skip clustering degraded performance
- ◇ Developed a software technique to cooperate with CPUs to enhance performance on resource-constrained embedded GPUs
- ◇ Proposed an adaptive reordering for the condition check to balance the warps
- ◇ Contributions: co-author, idea, implementation, paper write-up

TLP Balancer: Predictive Thread Allocation for Multi-Tenant Inference in Embedded GPUs [ESL'24]
Advisor: Prof. Yunho Oh, Korea University Mar. 2024 - Aug. 2024

- ◇ Observed fused kernels for multi-tenant AI workloads relied on the sub-optimal thread configuration
- ◇ Presented modeling to find the optimal thread configuration for fused AI kernels that balanced the warp-level computation
- ◇ Proposed a runtime system that dynamically fused AI kernels with the modeling
- ◇ Contributions: co-author, idea, paper write-up

VitBit: Enhancing Embedded GPU Performance for AI Workloads through Register Operand Packing [ICPP'24]
Advisor: Prof. Yunho Oh, Korea University Sep. 2023 - May. 2024

- ◇ Observed under-utilization of floating CUDA cores or Tensor cores when processing integer-quantized AI workloads
- ◇ Proposed a software technique for simultaneous computation on all heterogeneous cores on GPU to support arbitrary integer formats
- ◇ Proposed a software-based packing policy to support simultaneous processing of packed integers
- ◇ Contributions: co-author, motivation study, idea, implementation, paper write-up

Salient Frequency-aware Exemplar Compression for Online Continual Learning [AAAI'25]
Supervisor: Dr. Suhyun Kim, Korea Institute of Science and Technology Jan. 2023 - Nov. 2023

- ◇ Observed exemplar compression methods occupied limited GPU resources during online continual learning
- ◇ Proposed a computationally efficient compression algorithm using salient frequency
- ◇ Proposed a buffer management scheme to alleviate harmful effects from the compression artifacts remaining in the buffer
- ◇ Contributions: 1st author, motivation study, idea, implementation, paper write-up

HammerFilter: Robust Protection and Low Hardware Overhead Method for RowHammer [ICCD'21]
Advisor: Prof. Ki-Seok Chung, Hanyang University Aug. 2020 - Nov. 2020

- ◇ Motivated by the fact that newer DRAM chips are more vulnerable to Rowhammer (i.e., Rowhammer threshold has decreased from 139K to 10K)
- ◇ Proposed a robust and low overhead RowHammer protection scheme by modifying counting bloom filter
- ◇ Contributions: co-author, motivation study, experiment, paper write-up

Skills

C/C++, Python, Tensorflow, Pytorch, Git, Verilog, Shell script