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

[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

[Poster] Kwangrae Kim, **Junsu Kim**, Jeonghyun Woo, and Ki-Seok Chung. "HammerFilter: Robust Protection and Low Hardware Overhead Method for Row-Hammering". The 58th IEEE Design Automation Conference (DAC) Work-in-Progress, 2021

Preprints

[P1] **Junsu Kim**, Jaebeom Jeon, Jaeyoung Park, Seokin Hong, Gunjae Koo, Myung-Kuk Yoon, and Yunho Oh. "Memory Oversubscription-Aware Tensor Migration Scheduling for GPU Unified Storage Architecture" *Under Review*

[P2] **Junsu Kim**, and Suhyun Kim. "Salient Frequency-aware Exemplar Compression for Resource-constrained Online Continual Learning" *Under Review*

[P3] Jongmin Kim, Munsung Gil, Sangun Choi, **Junsu Kim**, Seondeok Kim, and Yunho Oh. "Exploring Datacenter Workloads: A Comprehensive Behavioral Analysis of CXL Memory Systems" *Under Review*

[P4] Minsung 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" *Under Review*

[P5] Yujin Kim, Minsoo Kang, **Junsu Kim**, and Suhyun Kim. "Integrative Solution for Catastrophic Forgetting in Model-Only Class Incremental Learning" *Under Review*

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 Jul. 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

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

Research Projects

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

- Analyzed the page faults due to GPU memory oversubscription stalled AI workloads despite prior migration scheduling methods
- Proposed a tensor migration scheduling algorithm considering GPU memory oversubscription for GPU unified storage architecture
- Contributions: 1st author, motivation study, idea, implementation, experiment, paper write-up

Exploring Datacenter Workloads: A Behavioral Analysis of CXL Memory Systems
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 Embedded GPU via Warp Balancing
Advisor: Prof. Yunho Oh, Korea University Jun. 2024 - Sep. 2024

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

TLPBalancer: Predictive Dynamic Thread Allocation for Fused Kernels in Embedded GPUs
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 inter-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 Resource-constrained Online Continual Learning
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

Integrative Solution for Catastrophic Forgetting in Data-Free Class Incremental Learning
Supervisor: Dr. Suhyun Kim, Korea Institute of Science and Technology May. 2022 - May. 2023

- Observed synthetic data for incremental learning caused bias in classification
- Developed a weight-balancing method to correct the bias in the classification head and a hybrid knowledge distillation approach
- Contributions: co-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, Pennylane, Qiskit, Git, Shell script, ARM assembly, Verilog