JUNYEOL RYU

jyeol.ryu@gmail.com junyeol.me junyeol gajagajago

RESEARCH INTERESTS

Computer architecture and systems, GPU and accelerators, heterogeneous and large-scale computing, shared memory architectures, multi-tiered memory systems, language and multi-modal models, graph neural networks

SHORT BIO

I am a researcher at Seoul National University. I began studying computer science in the later years of Bachelor's, fascinated by the beauty of general-purpose processor architecture. Parallel processors, such as GPUs, captivated me during Master's, and I built a strong background in heterogeneous system design through overcoming the challenges of making different architectures collaborate. I anticipate even greater challenges when specialized accelerators come into play, which shapes my research vision: **innovating practical architectures and systems for a massively heterogeneous paradigm**. I am particularly interested in enhancing the accessibility and efficiency of hardware and software for large-scale, performance-centric workloads such as deep learning and graph analytics. Meanwhile, I am a competitive programmer, symbiotic teammate, passionate mentor/learner, and always a friend.

EDUCATION

Seoul National University (SNU)

M.S. in Computer Science and Engineering — Sep. 2022 – Aug 2024

- Thesis: Optimization of Pipeline Parallelism using CPU Memory
- Advisor: Prof. Jaejin Lee

B.S. in Computer Science and Engineering and B.B.A. in Business Administration — Mar. 2016 – Aug 2022

- **Summa Cum Laude** with double major (B.B.A. 16 – 19, B.S. 20 – 22)
- Period includes two years of military service and six months of full-time work

RESEARCH EXPERIENCE

Training large models with pipeline parallelism on GPUs under memory constraints — Jan. 2024 – Present

- Advisor: Prof. Jaejin Lee (SNU Thunder Research Group)
- Designed mechanism for training large models that exceed GPU memory capacity by optimizing GPU pipeline parallel schedule and offloading memory requirements to host memory, with focus on maximizing overlap.
- Proposed *decoupled backward assignment* to reduce pipeline bubbles, where forward and backward pass of each pipeline stage are processed on different GPUs. Devised memory-efficient, low-latency techniques such as minimizing GPU memory usage with shared memory-based parameter offloading, decreasing optimizer latency by asynchronous parameter updates using CPU, and scaling to multiple machines by parameter prefetching using RDMA. Achieved 1.26-1.32× speedup against state-of-the-art for training up to 110B parameter model.

GPU communication library for mitigating congestion in PCIe systems — Mar. 2023 – Dec. 2023

- Advisor: Prof. Jaejin Lee (SNU Thunder Research Group)
- Discovered congestion patterns in host network, and found them unavoidable with heuristic-based pathfinding algorithms used in GPU communication libraries such as NCCL and MSCCL, which became key motivation for profiling-based pathfinding mechanism.
- Analyzed and implemented baselines [link], conducted multiple machine experiments having 2.07× speedup, and contributed to paper and artifact evaluation [link]. Documented key technologies [link], filed analysis of NCCL [link], and explored future work with in-network aggregation with Mellanox SHARP.

GPU kernel optimization and scheduling investigation — Mar. 2023 – Jun. 2023

- Advisor: Prof. Jaejin Lee (SNU Thunder Research Group)
- Open-sourced [link1,link2] and documented [link] step-by-step optimization of MatMul kernel, achieving 27.8 Tensor TFLOPS on NVIDIA Tesla V100 GPU, and implemented transformer inference in plain C++ with custom kernels [link] for performance experiments.
- Investigated thread block scheduling, implemented various kernels with different intensities on certain GPU hardware components [link], and filed report on their concurrent execution performance [link].

Shared cluster scheduling for mitigating network contention Jul. 2022 – Feb. 2023

- Advisor: Prof. Byung-Gon Chun (SNU Software Platform Lab)
- Designed Reinforcement Learning (RL) scheduler to mitigate network contention in shared GPU clusters.
- Implemented end-to-end framework that integrates RL training libraries, like OpenAI gym, with job schedulers such as Slurm, enabling online training of deployed scheduling policies [link].

Resource contention-aware GPU scheduling for ML models Mar. 2022 – Jun. 2022

- Advisor: Prof. Byung-Gon Chun (SNU Software Platform Lab)
- Proposed resource contention-aware scheduling strategies for colocation of multiple jobs on single GPU using CUDA Multi-Process Service (MPS) [link].
- Investigated the resource intensities of language, vision, speech, and recommendation models using profilers such as Nsight Compute and TensorBoard.

Selected Projects

LLM inference acceleration Sep. 2023 – Oct. 2023

- Competition: Samsung Computer Engineering Challenge 2023, **1st place** (Teammate: Heehoon Kim)
- Achieved fastest inference of HellaSwag with Llama 30B on single machine with four NVIDIA Tesla V100 GPUs.
- Designed batching with minimal padding using dynamic programming, developed hierarchical scheduling of inference phases for KV cache management, optimized communication with CPU-controlled mechanism to minimize contention, and wrote custom kernels.
- Open-sourced [link] and documented two round reports [link1,link2], gave a talk at Samsung Advanced Institute of Technology (SAIT) [link], and awarded at Samsung AI Forum (SAIF) 2023 [link].

Linux operating system Mar. 2022 – Jun. 2022

- Course: Operating Systems, Spring 2022 (Instructor: Prof. Byung-Gon Chun)
- Hacked Tizen Linux kernel embedded on Raspberry Pi 3.
- Implemented weighted round-robin scheduler, file system with access control mechanism based on GPS, synchronization primitive using reader-writer lock that changes its behavior based on 3D orientation, and system call for process tree.

RISC-V pipelined processor Mar. 2021 – Jun. 2021

- Course: Computer Architecture, Spring 2021 (Instructor: Prof. Jihong Kim)
- Implemented five-stage pipelined processor with data hazard minimization techniques, including scoreboard and forwarding, and experimented using SPEC benchmarks.
- Volunteered for post-semester homework enhancement project and implemented branch predictor based on Bray et al. [link].

Work Experience

SNU Center for optimizing Hyperscale AI Models and Platforms (CHAMP), Researcher Sep. 2024 – Present

- Developing deep learning systems for training large language models.

FriendliAI, Software engineer intern Jan. 2022 – Feb. 2022

- PeriFlow: Participated in developing chat interface and web graph analytics for inference service [link] powered by Orca [link].

Waffle Studio, Frontend team lead Mar. 2021 – Dec. 2021

- Guam: Led frontend development [link] from scratch to launch on marketplaces [link1,link2] of teammate recruiting app for programmers, managers, and designers. Organized weekly sprints, facilitated code reviews, and mentored junior frontend developer during whole period.

Vanilla Bridge, Full-time software engineer intern Jul. 2020 – Dec. 2020

- VB: Introduced DevOps for analyzing UX data collected from human matchmaker-based dating app [link]. Enhanced matching algorithm, developed swipe-based UI, implemented gifticon purchases, social login, etc.

Consulate of the Republic of Korea in Melbourne, Intern Jul. 2016 – Aug. 2016

- Assisted with consular/visa services at the consulate and attended civil servant business trip to rural areas like Wollongong, Australia.

Teaching Experience

Principles and Practices of SW Development (Instructor: Prof. Byung-Gon Chun) Fall 2022

- Provided lab lectures (3hr/week) and prepared exercises [link] on full stack development of web service.
- Led seven teams through bi-weekly sprint meetings (0.5hr/team) for final term project of developing creative web service and delivering offline presentation to invited audiences.

Publications

- **Junyeol Ryu**, Yujin Jeong, Daeyoung Park, Jinpyo Kim, Heehoon Kim, Jaejin Lee. SysX: System for Training under Memory Constraints (Anonymized title for review). *Under submission to peer-reviewed conference*.
- Heehoon Kim, **Junyeol Ryu**, Jaejin Lee. TCCL: Discovering Better Communication Paths for PCIe GPU Clusters. *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3 (ASPLOS '24)*.
- **Junyeol Ryu**, Jeongyoon Eo. Network Contention-Aware Cluster Scheduling with Reinforcement Learning. *Proceedings of the IEEE 29th International Conference on Parallel and Distributed Systems (ICPADS '23)*.
- **Junyeol Ryu**, Jinpyo Kim, Jaejin Lee. A Fast and Scalable Generative Model Inference on Distributed Multi-GPU Environment. *Proceedings of the 2023 Korean Computing Congress (KCC '23)*.
- **Junyeol Ryu**, Byung-Gon Chun. Investigating Contention Sensitivity of DL Training Workloads in Shared GPU Cluster. *Proceedings of the 2022 Korean Software Congress (KSC '22)*. **Best paper award**.

Honors and Awards

Samsung Computer Engineering Challenge, **1st place** (USD 7,300) Nov. 2023

SNU BK21 Scholarship (USD 4,400) Dec. 2023

SNU Merit-based Scholarship (USD 600) Feb. 2023

SNU Teaching Assistance Scholarship (USD 3,800) Sep. 2022

IMM Hope Foundation Outstanding Student Scholarship (USD 7,200 over two years) Apr. 2021

Changgang Foundation Outstanding Student Scholarship (USD 1,800) Mar. 2020

SNU Merit-based Scholarship (USD 900) Aug. 2019

Skills

- Programming: Proficient in C/C++ and Python for building parallel software with PyTorch, CUDA, C parallel libraries, and MPI/NCCL, with 5+ relevant project experiences. Expertise in core mechanics of deep learning frameworks such as Megatron-LM and DeepSpeed. Knowledgeable in Verilog with 1 project experience and ongoing self-induced learning for preparation purposes.
- Languages: Korean (native), English (fluent), Chinese (survivable)

Related coursework

Seoul National University (SNU)

- Computer architecture (**A+**), Adv. computer architecture (**A+**), Computer interconnection networks (**A+**), Adv. compilers (**A+**), Studies in computer systems (**A+**), Logic design (A0), OS (A0), Scalable HPC (A-)

Outreach

Dasom, SNU volunteer club member Mar. 2016 – Feb. 2017

- Served as teacher for providing education to elementary school children from low-income families residing in Sillim-dong, Seoul [link]. Helped keep up with school curriculum and focused on fostering positive minds.

Shelter, DFLHS volunteer club member/president Mar. 2013 – Feb. 2016

- Served as club member and president (one year, 2014) for supporting refugees in Korea in collaboration with pNan [link]. Organized cultural exchange events and student flash mobs for improving refugee awareness, helped job placement at local stores, and supported application process for legal refugee recognition.