CURRICULUM VITAE

BASIC INFORMATION

Name: Zhiyao Li
Gender: Male
E-Mail: ziozhiyao.lee@gmail.com
Age: 28

EDUCATION BACKGROUND

Ph.D. —— 2019 - present:

Institude for Interdisciplinary Information Sciences, Tsinghua University
Research fields: Computer system and architecture
Advisor: Prof. Mingyu Gao

Bachelor —— 2015 - 2019:

School of Computer Sciences, Chongqing University
GPA: 3.86 Rank: 1/208 CET-6: 572 TOEFL: 97
Advisor: Prof. Jiang Zhong

RESEARCH INTERESTS

Dataflow accelerator design and scheduling for AI;
Dynamic/Sparse workload optimization and hardware acceleration;
High-performance network for LLM serving/training;

RESEARCH PAPERS

SPADA: Accelerating Sparse Matrix Multiplication with Adaptive Dataflow

- Appeared in ASPLOS' 23, First Author
- https://dl.acm.org/doi/10.1145/3575693.3575706
- Current sparse matrix multiplication acceleration hardware tends to perform well only for specific sparsities. We propose a hardware-software design that adaptively adjusts the dataflow and hardware configuration according to the sparsity patterns of matrices, and achieves an average speedup of 1.43x under different sparse workloads.

Adyna: Accelerating Dynamic Neural Networks with Adaptive Scheduling

- Appeared in HPCA' 25, First Author
- https://people.iiis.tsinghua.edu.cn/~gaomy/pubs/adyna.hpca25.pdf
- Dynamic neural networks are the mainstream method to reduce the computation workload per sample and increase network parameters. Based on the existing

dataflow architecture, we optimize from both the software scheduling and on-chip architecture design to solve the load imbalance and redundancy computation caused by the dynamism, which achieves an average 1.87x acceleration.

KAPLA: Pragmatic Representation and Fast Solving of Scalable NN Accelerator Dataflow

- Appeared in ASPDAC' 25, First Author
- https://dl.acm.org/doi/10.1145/3658617.3697549
- Dataflow scheduling for neural network accelerators is critical to computational efficiency. Existing methods rely on exhaustive search or complex machine learning models, which are slow to solve and difficult to apply in practice due to the large design space. Our proposed efficient scheduling optimization method decouples multiple levels of design space and achieves up to 4x speedup over the best existing methods.

FastSwitch: Optimizing Context Switching Efficiency in Fairness-aware Large Language Model Serving

- Submitted to ATC' 25, Co-first Author
- https://arxiv.org/abs/2411.18424
- Serving numerous users and requests concurrently requires good fairness in Large Language Models (LLMs) serving system. To achieve better fairness, the preemption-based scheduling policy dynamically adjusts the priority of each request to maintain balance during runtime. However, existing systems tend to overly prioritize throughput, overlooking the overhead caused by preemption-induced context switching, which is crucial for maintaining fairness through priority adjustments. In this work, we introduce FastSwitch, a fairness-aware serving system that not only aligns with existing KV cache memory allocation policy but also mitigates context switching overhead. Our evaluation shows that FastSwitch outperforms the state-of-the-art LLM serving system vLLM with speedups of 1.4-11.2x across different tail TTFT and TBT.

INTERN & VISITING EXPERIENCE

Stanford University

- Pervasive Parallel Lab, led by Prof. Kunle Olukotun
- 2023.07 – 2023.12

Ali DAMO Academy

- Computing Technology Lab, led by Prof. Yuan Xie
- 2022.08 – 2023.07

Huawei

- 2018.07 – 2018.11

HONORS AND AWARDS

National Scholarship
National College Student Mathematics Competition, Second Prize

T. A . EXPERIENCE

Undergrad course: Computer system and Architecture, 2022 Fall
Graduate course: Computer system and Architecture, 2020 Fall