SwFormer: Enabling Faster Foundation Models on new Sunway Supercomputer via Holistic Kernel Tiling and Scheduling

Ruo-han Wu(吴若晗), Xian-Yu Zhu(朱先语), Jun-Shi Chen(陈俊仕), Hong An(安虹)

DOI: 10.1007/s11390-025-4761-0

Research Objectives

- Address the performance bottleneck of AI operators on SW26010pro in all-shared mode
- Overcome the challenges of scaling single-CG operator performance to 6 CGs
- Enhance the computational efficiency of AI operators in all-shared mode

Research Method

- An intra-op tiling method that breaks operators into fine-grained tiled kernels
- Offline profiling-based approach to determine the optimal tiling strategy
- Combine intra-op tiling and inter-op scheduling for holistic optimization

Research Results

- Greatly improve the performance of all-shared mode AI operators
- Accelerate end-to-end foundation model training such as GPT-3 13B and 6.7B by up to 1.27x
- Further boost performance through inter-op scheduling policies

Research Conclusions

- We design and implement SwFormer, a framework for accelerating foundation models on the new Sunway Supercomputer
- Our work simplifies the design of 6-CG operators and provides valuable insights for the AI computing ecosystem of future SW26010pro processors