Taura Laboratory

Home Page: https://www.eidos.ic.i.u-tokyo.ac.jp

Keywords: ***Systems, High-Performance Computing (HPC), AI Accelerators, Trustworthy AI, Data Mining, Foundation Model***

Research Theme & Mottos

Main Research Theme of the Laboratory

The main interest of our group is system software and machine learning systems that enhance programming productivity, performance, scalability, and security. We have been actively engaged in developing runtime systems and programming systems for parallel computing environments, with a strong focus on GPUs and emerging accelerators such as Cerebras. Recently, our research has also expanded to foundation models for spatial-temporal data and human mobility, mechanistic interpretability of machine learning models, and privacy-preserving data analysis in general-purpose programming systems. Our ambition is to enable efficient, secure, and interpretable use of powerful computing resources for scientific computing, machine learning, and data-intensive applications.

Message to Prospective Students

Many of our research themes share a common objective of efficiently and securely leveraging powerful computing resources. We welcome students interested in system software, GPU programming systems, accelerator-based computing, foundation models, spatial-temporal data analysis, mobility and machine learning, mechanistic interpretability, and privacy-preserving data analysis. Students will have opportunities to work on both fundamental system techniques and practical machine learning applications that require high performance, scalability, security, and interpretability.

Topics

Systems / HPC / AI Accelerators

Programming Systems for GPUs

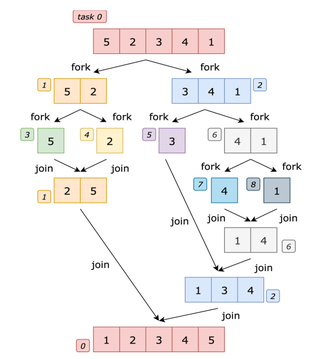
■ Task Parallelism on GPUs

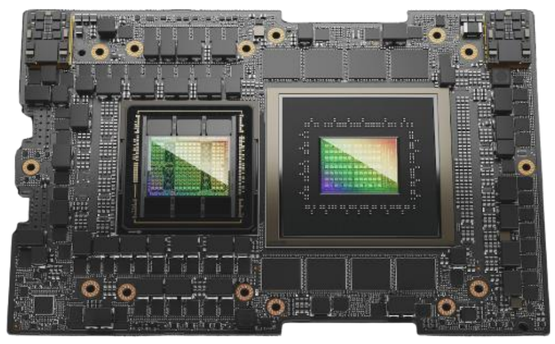
- Develops runtime systems that enable fork-join style task parallelism on GPUs, beyond conventional data-parallel loop execution.
- Aims to efficiently schedule irregular and recursive workloads while overcoming GPU-specific challenges such as warp divergence, kernel-launch overhead, and limited preemption.

■ Memory-Aware Programming on CPU-GPU Integrated Systems

- Explores programming techniques for integrated CPU-GPU architectures such as NVIDIA GH200, where GPUs can directly access CPU memory through a unified address space.
- Investigates how to improve memory efficiency and performance on LLM workloads by considering data placement, bandwidth differences, page faults, and memory pooling.

Collaborator: Naoya Maruyama @ NVIDIA





Wafer-scale Kernel Optimization and Performance Observability

■ Topic Impact

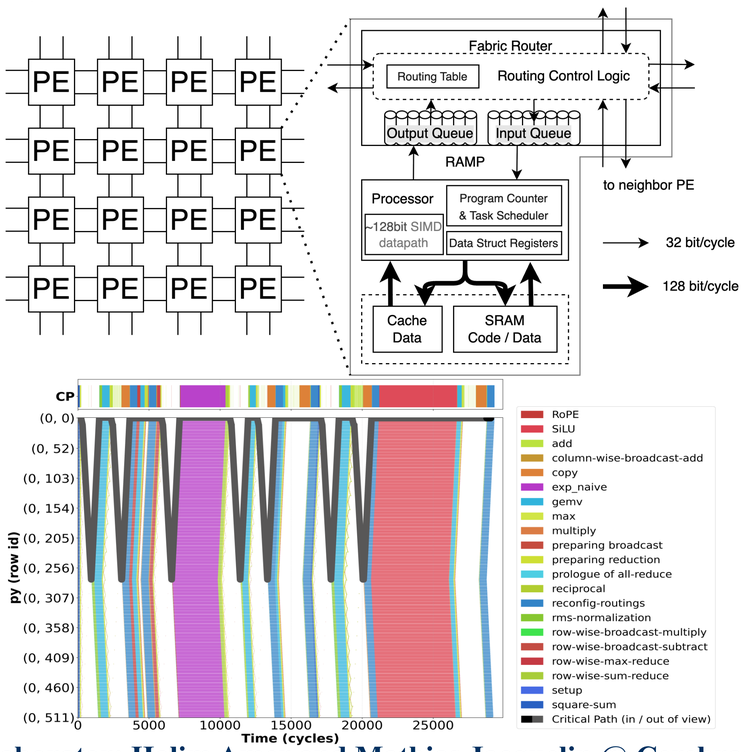
- Cerebras Wafer Scale Engine is a largest inference chip that achieves both high bandwidth and low latency. It consists of almost a million Processing Elements (PE; minimal cores) interconnected on a Network-on-Chip (NoC).

■ Problem Statement

- Because of its scale and exposure to low-level resource management, leveraging high performance from this architecture is ultimately challenging. Simulation on wafer-scale program has been impossible, and which prevents users from reasoning their optimization or performance insights.

■ Methodology

- We are building a platform that integrates high-level programming abstraction with critical-path analysis driven performance profiling and layout optimization, and the compiler feature in the future.



Collaborator: Halim Amer and Mathias Jacquelin @ Cerebras

Trustworthy AI/ Data Mining/ Foundation Model

Mechanistic Interpretability for AI Safety

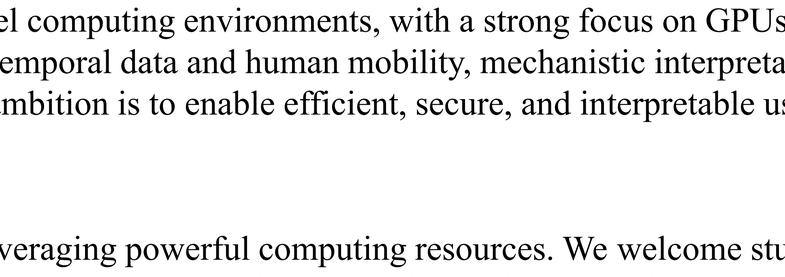
■ Background: Demystifying the Black Box

- Aims to reverse-engineer neural networks to uncover the mechanisms behind their outputs.
- Clarifies the "thinking process" of AI models to enhance operational transparency and safety.

■ Research: From Neurons to "Circuits"

- Shifts the focus from analyzing individual, isolated neurons to understanding functional "circuits" (subnetworks of interconnected neurons and attention heads).

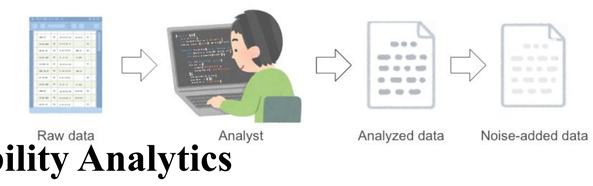
Collaborator: Gouki Minegishi @ Utokyo, Koshi Eguchi @ Sakana AI



Privacy-Preserving Data Analysis and Programming Systems

■Background: Sensitive Mobility Data

- Check-in and GPS trajectories can reveal individual movement patterns.
- Privacy-preserving programming systems are needed to analyze such data without exposing raw records.



■Research: Differentially Private Mobility Analytics

- We use *PrivJail* to release DP-noisy mobility statistics and train privacy-preserving prediction models.
- We explore both DP Markov-chain prediction from noisy transition counts and DP-SGD training for RNN/LSTM-based next-location prediction.

■Findings: Privacy, Utility, and Efficiency

- Larger privacy budgets achieve performance close to non-private baselines, while strict privacy reduces utility.
- Our goal is to build scalable support for privacy-preserving trajectory analysis.

Collaborator: Shumpei Shiina @ Toyota, Shohei Hanaoka and Renhe Jiang @ UTokyo

Foundational Model for Time Series

■Background: Time Series Forecasting

- Time series is important for intelligent society, such as analyzing human mobility and modeling climate change. Given historical values in a time series, the objective is forecasting its future values.

■Motivation: Universal Forecasting

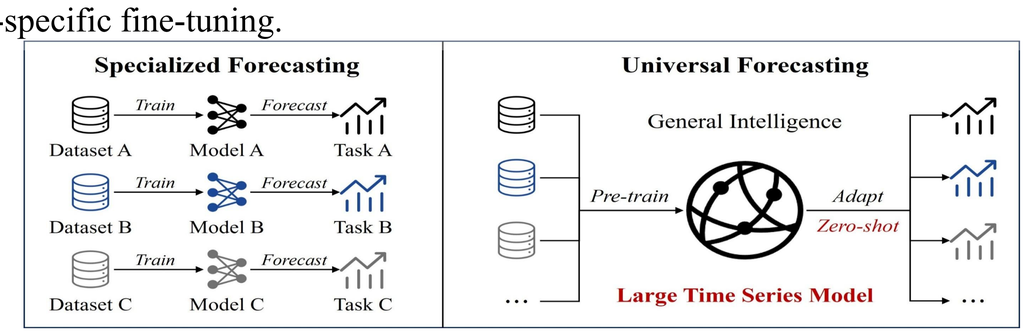
- Specialized forecasting (before): *one model for one dataset and one task*
- Universal forecasting (our goal): *one model for many datasets and many tasks*

■Research: Developing Foundational Models

- We aim to develop a single, large foundational model, pre-trained on large-scale and diverse datasets, that can be rapidly adapted to a wide range of forecasting tasks across different domains (like *"GPT for time series"*).
- We collected large time series pre-training corpus (1TB vs. 10MB before).
- We trained large time series models (200M vs. 200K parameters before).

■Experiment: Few/Zero-shot Forecasting

- The pre-trained foundational model can forecast on unseen datasets with little or even no task-specific fine-tuning.



Collaborator: Renhe Jiang @ UTokyo