Taura Laboratory

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

Key Words: ***Highly Productivity, High Performance, Parallel Programming, Machine Learning, ML System, Privacy Protection***

Research Theme & Mottos

Main Research Theme of the Laboratory

The main interest of our group is system software that aims to enhance programming productivity, performance, and security. We have been actively engaged in developing runtime systems for parallel programming languages, with a strong focus on achieving both programmability and performance. Our ambition is to revolutionize the way people program parallel computers and future machines. Additionally, we are interested in applications that require substantial computing resources, particularly in the fields of scientific computing and machine learning. Furthermore, we have recently delved into privacy-preserving techniques in general-purpose programming languages.

Message to Prospective Students

Many of our research themes share a common objective of efficiently and securely leveraging powerful computing resources. We invite enthusiastic students who align with this goal and/or possess a broad interest in system software (such as programming languages, operating systems, etc.). We also welcome students who are keen on exploring the applications of big data processing, high-performance computing, and machine learning, as well as addressing security and privacy challenges associated with these applications.

Topics

InterconnectLens: Enhancing Observability in GPU Clusters

- **Recent GPU Clusters:** With the advancement of large language models (LLMs), the demand for faster computation has increased, leading to more complex communication paths within GPU clusters.
- **InterconnectLens (ICLens)** visualizes GPU clusters in a component-based manner (e.g., RNICs, root complexes, GPUs, CPUs), enhancing observability of the increasingly complex communication paths.

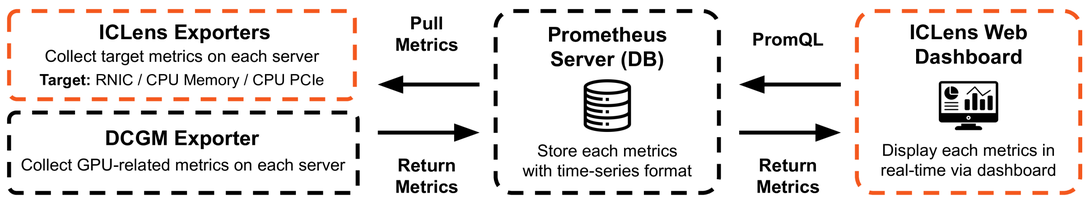
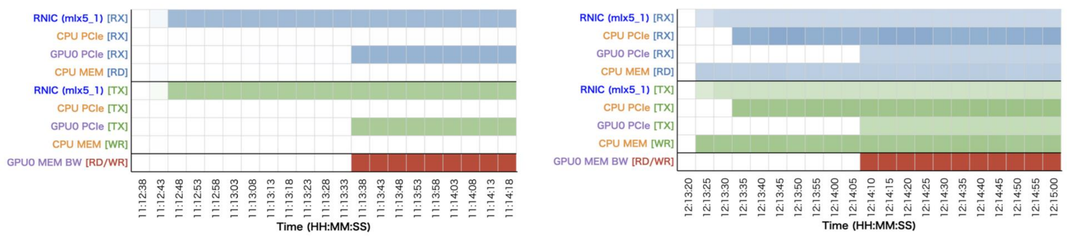


Figure) System Overview



(a) AllReduce with GPUDirect RDMA

(b) AllReduce without GPUDirect RDMA

Figure) How ICLens visualizes GPU-to-GPU data transfers

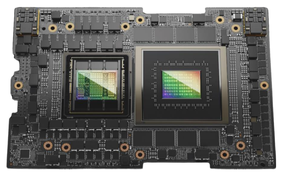
Heterogeneous Computing for Machine Learning

■VRAM poor problem

- The memory footprint needed for training machine learning models is exponentially increasing
- There is a strong need to efficiently utilize hardware for training machine learning models.

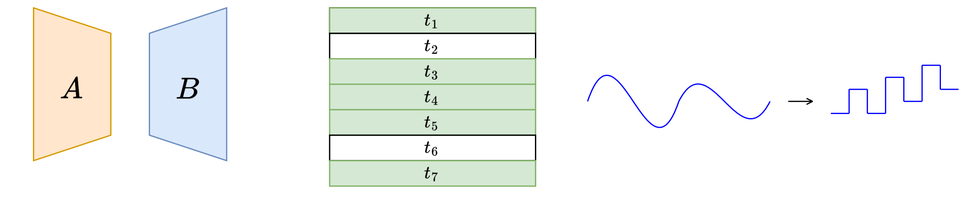
■NVIDIA GH200

- NVIDIA GH200 is a heterogeneous NUMA processor that can run CUDA kernels even on CPU data.
- We analyzed the bottlenecks of training when using the GH200.
- We modified jemalloc in the NUMA-aware way.
- Modified jemalloc enabled the training of the 3b models 10x faster than the naive mmap allocator did.



KV Cache Compression for faster LLM inference

- **KV Cache (Key-Value Cache)** stores the key and value tensors generated for previously processed tokens during inference, enabling the model to reuse past computations instead of recalculating.
- However, its memory budget has become a major memory and speed bottleneck. Its size becomes **1.5TB** given OTP-175B with a batch size 512.
- KV Cache Compression improves LLM inference speed.



Low Rank Decomposition | Token Eviction | Quantization

Programming System for Privacy-Preserving Data Analysis

■Big-Data and Privacy

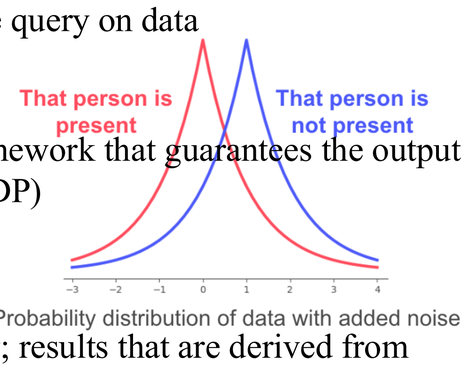
- Data useful to make a good, evidence-based decision are often *personal* (location, trajectories, health records, etc.)
- Many useful insights could be drawn from aggregated statistics without violating privacy, yet fears remain about *potential* bleach of privacy

■Differential Privacy (DP)

- A mechanism which provides mathematically provable privacy assurance while maintaining statistical usefulness by adding appropriate noise to the output of the query on data

■Goal

- A general-purpose programming framework that guarantees the output is privacy-protected (in the sense of DP)

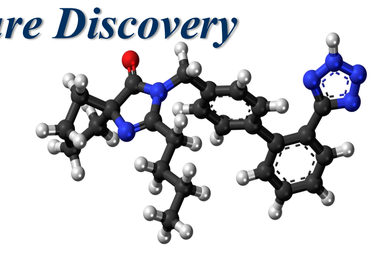


Probability distribution of data with added noise

■ Approach

- *Control what the program can output*; results that are derived from personal data can't go out; those who have gone through a DP mechanism can

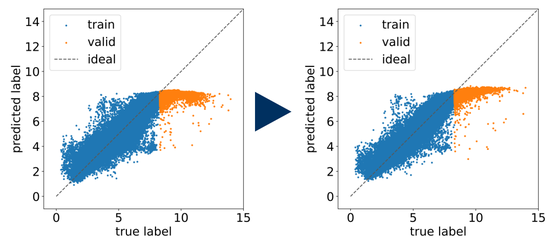
Machine Learning for Chemical Structure Discovery



<u>Fast calculation of physical properties using predictive models</u>

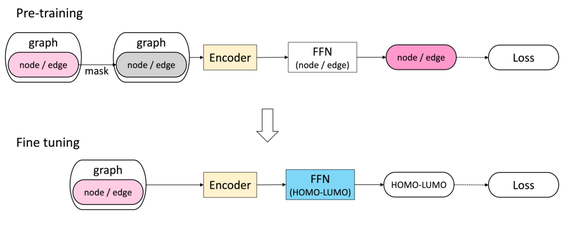
■Extrapolation problem

- Machine learning face difficulty in predicting values outside those in training data
- Improved it by *self-supervised pretraining*



■Mixed-Data and Data Imbalance

- Less training data containing information on atom's three-dimensional coordinates (3D data) than data that does not (2D data)
- Proposed a learning method that works under such Imbalance: *Pretraining and 2D-3D two-stage learning*



<u>Direct generation of structure using generative models</u>

- Structure generation using Diffusion Models, etc., is being pioneered

(collaborating with Masatoshi Hanai and Suzumura Lab.)