Taura Laboratory

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

Key Words: ***Highly Productivity, High Performance, Parallel Programming, Big Data Processing, Machine Learning, 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.

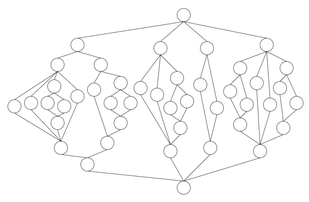
Message to Prospective Students

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

Itoyori: A Scalable Task-Parallel Runtime System

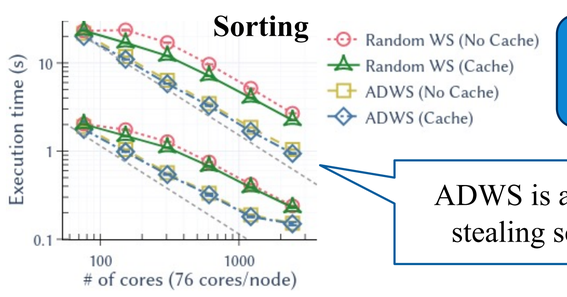
- **Task Parallelism:** a parallel model based on (possibly many) tasks and their dependencies
- **Itoyori** enables *efficient dynamic load balancing for task parallelism on large-scale systems* (from multi-core CPUs to supercomputers)



Task Graph

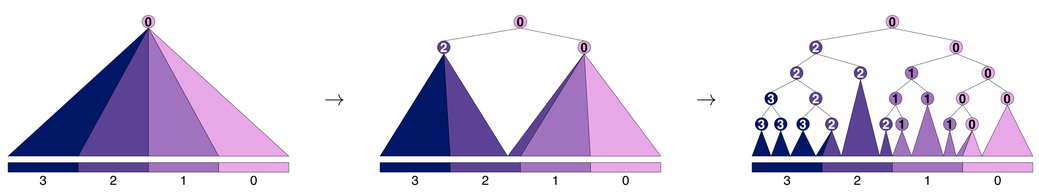
Specifically, we investigated...

1. Efficient *dynamic thread migration* across machine boundaries
2. *Global address space* with software caching for remote memory access
3. Locality-aware scheduling: Almost Deterministic Work Stealing (**ADWS**)



Achieve a good balance between productivity and performance

ADWS is about 2x faster than the standard work stealing scheduler and scales to > 2000 cores!



A Programming System for Privacy-Preserving Data Analysis

■Big-Data and Privacy

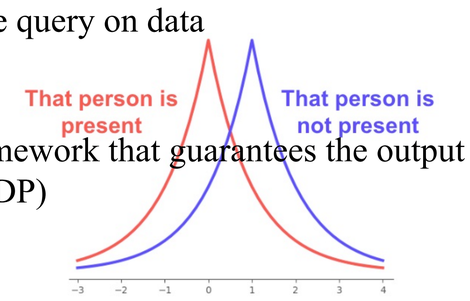
- Data useful to make a good, evidence-based decision are often *personal* (location, trajectories, health records, etc.)
- Many useful insights could be drawn from aggregated statistics without violating privacy, yet fears remain about *potential* bleach of privacy

■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

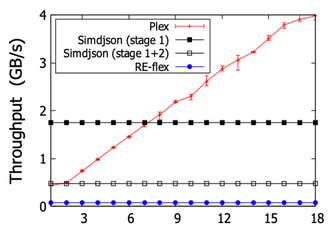
■ Approach

- *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

Lightning-Fast Lexer/Parser for Massive Data

■Plex: Lightning-Fast Lexer

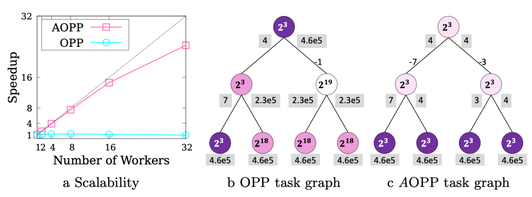
- Scaling Parallel Lexing with Backtrack-Free Prescanning
- An automated tool for generating parallel lexers from user-defined grammars.
- *9.8-11.5X speedups using 18 threads.*



(a) JSON

■AOPP: Lightning-Fast Parser

- **A**ssociative **O**perator **P**recedence **P**arsing
- Improved parallelism by allowing ambiguous grammars for *associative* operators



■VRegex

- A regular expression searching tool that uses SIMD instructions to process multiple characters of input data per instruction
- Improved data extraction tasks in JSON by *10x* when compared to existing *format-specific* parsers, like RapidJSON.

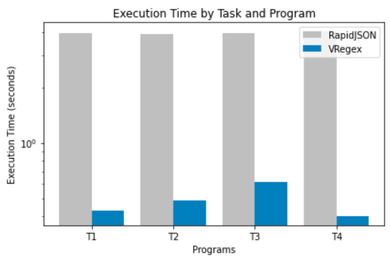
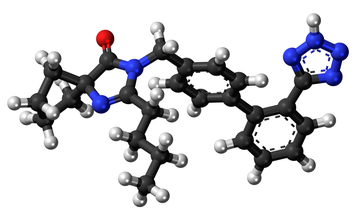


Figure 9: Data extraction on the Yelp business dataset.

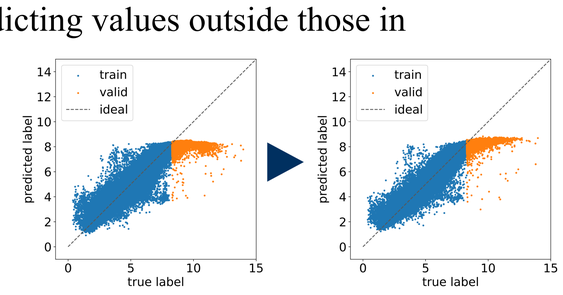
Machine Learning for Chemical Structure Discovery



Fast calculation of physical properties using predictive models

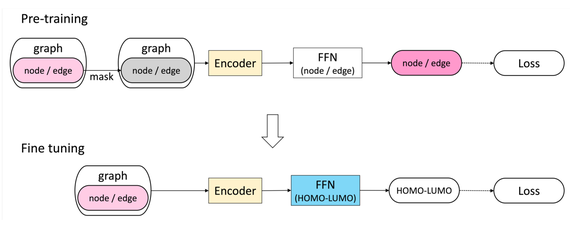
■Extrapolation problem

- 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*



Direct generation of structure using generative models

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

(collaborating with Masatoshi Hanai and Suzumura Lab.)