PARALLEL DATA LABORATORY

Carnegie Mellon University

AN INFORMAL PUBLICATION FROM ACADEMIA'S PREMIERE STORAGE SYSTEMS RESEARCH CENTER DEVOTED TO ADVANCING THE STATE OF THE ART IN STORAGE AND INFORMATION INFRASTRUCTURES.

PDL PACKET

NEWSLETTER ON PDL ACTIVITIES AND EVENTS • FALL 2025

http://www.pdl.cmu.edu/

CONTENTS

PDL CONSORTIUM MEMBERS

Bloomberg LP
Datadog
Google
Intel Corporation
Jane Street
LayerZero Research
Meta
Microsoft Research
Oracle Corporation
Oracle Cloud Infrastructure
Pure Storage
Salesforce
Samsung Semiconductor Inc.
Western Digital

RECENT PUBLICATIONS

LithOS: An Operating System for Efficient Machine Learning on GPUs

Patrick H. Coppock, Brian Zhang, Eliot H. Solomon, Vasilis Kypriotis, Leon Yang, Bikash Sharma, Dan Schatzberg, Todd C. Mowry, Dimitrios Skarlatos

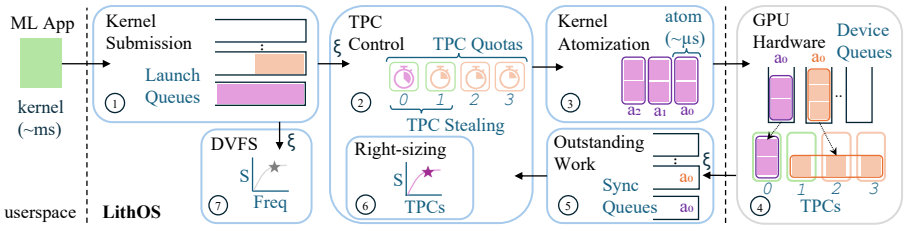
SOSP '25, October 13–16, 2025, Seoul, Republic of Korea.

The surging demand for GPUs in datacenters for machine learning (ML) workloads has made efficient GPU utilization crucial. However, meeting the diverse needs of individual ML models while optimizing resource usage is challenging. To enable transparent, fine-grained management of GPU resources that maximizes GPU utilization and energy efficiency while maintaining strong isolation, an operating systems (OS) approach is needed. Hence this paper introduces LithOS, a first step towards a GPU OS.

LithOS includes the following new abstractions and mechanisms for efficient GPU resource management: (i) a novel TPC Scheduler that supports spatial scheduling at the granularity of individual TPCs, unlocking efficient TPC stealing between workloads; (ii) transparent kernel atomization to reduce head-of-line blocking and allow dynamic resource reallocation mid-execution; (iii) a lightweight hardware rightsizing mechanism that dynamically determines the minimal TPC resources needed per atom; and (iv) a transparent power management mechanism that reduces power consumption based upon in-flight work characteristics.

We implement LithOS in Rust and evaluate its performance across a broad set of deep learning environments, comparing it to state-of-the-art solutions from NVIDIA and prior research. For inference stacking, LithOS reduces tail latencies by 13× compared to MPS; compared to the best-performing SotA, it reduces tail latencies by 3× while improving aggregate throughput by 1.6×. Furthermore, in hybrid inference-training stacking, LithOS reduces tail latencies by 4.7× compared to MPS; compared to the best-performing SotA, it reduces tail latencies by 1.18×

continued on page 5



LithOS operations overview.

GREG GANGER

Hello from fabulous Pittsburgh!

It has been another great year for PDL, on the research and student accomplishment fronts, including some big awards for PDL students and rising star faculty. And, although it has been a bumpy time for University research in general, PDL remains strong because of our great connections with and support from industry and national labs. Our many collaborations continue to produce cool and impactful results. Specifics can be found throughout the newsletter, but let me highlight a few things.

Naturally, the biggest shift in modern computing is a major focus of continuing and new PDL research activities, since data systems are central to AI/ML. Self-driving databases that combine extensibility and automation have been a long-running theme, and now is turning toward database system support FOR AI/ML. Our recent work on cluster scheduling for ML jobs has led to a novel "continual optimization" resource allocation approach, inspired by ML model fine-tuning, that allows optimization algorithms to be applied in large-scale settings. And our newest big project is a joint effort with LANL to build testbeds and help advance open-standard pNFS as a high-performance AI storage solution.

The efforts we started last year on enabling effective use of higher-capacity disks enabled by the arrival of HAMR and QLC are making great strides. These technologies will be crucial to satisfying exploding AI-driven capacity demands, but they come with less IO-per-TB capabilities and increased failure/wearout challenges. Our Declarative IO project promises to mitigate this issue by addressing the large amount (often over half) of IO dedicated to (critical) data maintenance in modern bulk storage systems, like compaction, capacity balancing, integrity checking, etc. Declarative IO interfaces would allow such processes to describe their IO needs so that a new IO planner component can exploit their order- and time-flexibility to eliminate redundant IO (e.g., two reads of the same data in the same day) that caching would not catch. Our focus on the role of QLC in exascale storage systems is creating new holistic TCO-driven tools for selective device type mixes and assigning workloads to them. And, to address one of the biggest impediments for high-density QLC, we are exploring storage system designs where QLC devices can be converted online and used as TLC devices, after no longer being able to function as QLC because of wearout... we call the approach Possum, because the devices are revived after appearing "dead".

Database systems and large-scale data processing remain central PDL focuses, even as they are crucial elements of AI/ML systems as mentioned above. In addition to the AI/ML <-> DB intersections, there is major focus on exploiting hardware accelerators of different kinds for DB functions,

THE PDL PACKET

THE PARALLEL DATA LABORATORY

School of Computer Science
Department of ECE
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213-3891
VOICE 412•268•6716

PUBLISHER

Greg Ganger

EDITOR

Joan Digney

The PDL Packet is published once per year to update members of the PDL Consortium. A pdf version resides in the Publications section of the PDL Web pages and may be freely distributed. Contributions are welcome.

THE PDL LOGO

Skibo Castle and the lands that comprise its estate are located in the Kyle of Sutherland in the northeastern part of Scotland. Both 'Skibo' and 'Sutherland' are names whose roots are from Old Norse, the language spoken by the Vikings who began washing ashore regularly in the late ninth century. The word 'Skibo' fascinates etymologists, who are unable to agree on its original meaning. All agree that 'bo' is the Old Norse for 'land' or 'place,' but they argue whether 'ski' means 'ships' or 'peace' or 'fairy hill.'

Although the earliest version of Skibo seems to be lost in the mists of time, it was most likely some kind of fortified building erected by the Norsemen. The present-day castle was built by a bishop of the Roman Catholic Church. Andrew Carnegie, after making his fortune, bought it in 1898 to serve as his summer home. In 1980, his daughter, Margaret, donated Skibo to a trust that later sold the estate. It is presently being run as a luxury hotel.

PARALLEL DATA LABORATORY

FACULTY

Greg Ganger (PDL Director)
ganger@ece.cmu.edu

George Amvrosiadis
David Andersen
Nathan Beckmann
Chuck Cranor
Lorrie Cranor
Christos Faloutsos
Phil Gibbons
Mor Harchol-Balter
Olivia Hsu
Zhihao Jia
Gauri Joshi
Todd Mowry
David O'Hallaron
Jignesh Patel
Andy Pavlo
Majd Sakr
M. Satyanarayanan
Dimitrios Skarlatos
Akshitha Sriraman
Rashmi Vinayak

STAFF MEMBERS

Bill Courtright, 412•268•5485
(PDL Executive Director) wcourtright@cmu.edu
Karen Lindenfelser, 412•268•6716
(PDL Administrative Manager) karen@ece.cmu.edu
Jason Boles
Joan Digney
Chad Dougherty

VISITING RESEARCHERS & POST DOCS

Ellango Jothimurugesan
Martin Prammer

GRADUATE STUDENTS

Nikhil Agarwal
Sam Arch
Daiyaan Arfeen
Sanjith Athlur
Abnash Bassi
Aditya Bhatnagar
Jennifer Brana
Frank Chen
Hilbert Chen
Siyuan Chen
Yiwei Chen
Xinhao Cheng
Zhuo Cheng
Val Choung
Patrick Coppock
Theo Gregersen
Ankush Jain
Neharika Jali
Siddharth Jayashankar
Jekyeom Jeon
Sheng Jiang
Hongyi Jin
Hyoungjoo Kim
Timothy Kim
Vasileios Kypriotis
Ruihang Lai
Christos Laspias
Jiaying Li
Edwin Lim
Sherman Lim
Wan Shen Lim
Yuchen Liu
Yixuan Mei
Deepanjali Mishra
Veronica Muriga
Nj Mukherjee
Gabriele Oliaro
Hojin Park
Ziyue Qiu
Minya Rancic
Drew Ripberger
Hugo Sadok
Sara M Shahri
Eliot Solomon
Minh Truong
Jaylen Wang
Daniel Wong
Mengdi Wu
Mingkuan Xu
Will Zhang
Kaiyang Zhao
Yiwei Zhao

UNDERGRADUATE STUDENTS

Kyle Booker
Sophia (Qingyang) Cao
William Courtright
Helen Wang
Lucy Wang

FROM THE DIRECTOR'S CHAIR

such as GPUs, processing-in-memory, computational storage, and other accelerators. The results show great promise and are providing effective programming models for exploiting such resources to gain much-needed efficiencies. At a much larger scale, we have created a new system (called Moirai) for automatically partitioning queries and tables for hybrid-cloud (i.e., on-prem plus public cloud) data processing that provides huge cost savings over state-of-the-art solutions, and it is already being put into practice at one of the PDL companies.

Indeed, real impact and winning awards have been two hallmarks of PDL research, and the past year is no exception. For example, (now Prof.) Juncheng Yang received the ACM SIGOPS Doctoral Dissertation Award (and the CMU SCS dissertation award), in part because his award-winning cache policies (e.g., SIEVE and S3-FIFO) cache policies have been adopted at multiple PDL companies. At least one PDL sponsor has integrated ideas from PDL's disk-adaptive redundancy research into production systems, and multiple are helping us to evaluate the Declarative IO concept's potential for real systems. We thank all of the PDL sponsor companies who have enabled (and collaborated on) much of the research mentioned above by working with us to enable evaluations with real devices, workload data, and failure logs!

Many other ongoing PDL projects are also producing cool results... too many for me to cover, especially as I strive to keep this note brief. But, this newsletter and the PDL website offer more details and additional research highlights.

We look forward to seeing many of you at this year's PDL Retreat, which will be in the Spring (for the first time), after many years of talking about the potential benefits of making the switch from Fall. We hope it indeed reduces conflicts for attendees and increases effectiveness of recruiting efforts!

I'm always overwhelmed by the accomplishments of the PDL students, staff, and faculty, and it's a pleasure to work with them. As always, their accomplishments point at great things to come.

Greg and his grad student crew, from L to R: Sanjith Athlur, Suhas J Subramanya, Sara McAllister, Daiyaan Arfeen, Greg Ganger, Hojin Park, Ziyue Qiu, Theo Gregersen,Timothy Kim, and PDL alum Saurabh Kadekodi (now with Google).

YEAR IN REVIEW

October 2025

- Juncheng Yang received the Dennis M. Ritchie Doctoral Dissertation Award!
- Valerie Choung gave her speaking skills talk on "Casma: Addressing Memory Bottlenecks with Compiler-Assisted Dynamic Memory Allocation."
- Three papers were presented at SOSP '25, Seoul, Republic of Korea: Eliot Solomon — "LithOS: An Operating System for Efficient Machine Learning on GPUs," Suhas J. Subramanya — COpter: Efficient Large-Scale Resource-Allocation via Continual Optimization," and Ziyue Qiu — "Moirai: Optimizing Placement of Data and Compute in Hybrid Clouds."

September 2025

- PDL Alum Jure Leskovec received a CMU 2025 Alumni Achievement Award!
- William Zhang presented his thesis proposal "On Holistic Database Optimization via Leveraging Similarity Across Actions, Workloads, Configurations, and Scenarios."
- Hyoungjoo Kim gave presented "No Cap, This Memory Slaps: Breaking Through the Memory Wall of Transactional Database Systems with Processing-in-Memory" at VLDB in London, UK.
- Sara McAllister is currently a Visiting Faculty Researcher at Systems Research@Google, before she begins her career as an assistant professor at the U. Wisconsin-Madison.
- Frank Chen presented his MS Thesis "Towards Utilizing Cached Context for Faster and Smarter Code Agents."
- Samuel Arch and his co-authors received the Runner Up Best Paper Award at VLDB 2025, London, UK.
- Also presented at VLDB in London: "The Key to Effective UDF Optimization: Before Inlining, First Perform Outlining" — Yuchen Liu, and "A Hot Take on the Intel Analytics Accelerator for Database Management Systems" — Christos Laspias.

PDL and CMU CS alumni join for a photo at the PDL FAST reunion in February. From L to R, Dave Maltz, Ted Wong, Julio Lopez, Chenxi Wang and Chris Colohan.

August 2025

- Hojin Park presented his dissertation "Cost-Efficient Storage and Caching in Public Clouds."
- Nirav Atre defended his dissertation on "Refining Classical Abstractions of Network Subsystems."
- Yiwei Zhao presented "Optimal Batch-Dynamic *kd*-trees for Processing-in-Memory with Applications" at SPAA '25, Portland, OR.

July 2025

- Akshitha Sriraman won an inaugural 2025 Google ML and Systems Junior Faculty Award.
- Sara McAllister presented her PhD thesis "Toward Sustainable Datacenters through Efficient Data Retrieval."
- Sanjith Athlur presented "Okapi: Decoupling Data Striping and Redundancy Grouping in Cluster File Systems" at ODSI'25, Boston, MA.

June 2025

- Dimitrios Skarlatos and Todd Mowry received Amazon Research Awards.
- Dimitrios Skarlatos won the IEEE TCCA Young Architect Award at ISCA '25!
- Daiyaan Arfeen proposed his PhD research on "Designing Scalable DNN Training Systems to Overcome Algorithmic Constraints."
- Patrick Wang presented "Automated Database Tuning vs. Human-Based Tuning in a Simulated Stressful Work Environment: A Demonstration of the Database Gym" at the 2025 International Conference on Management of Data in Berlin, Germany.

May 2025

- Suhas Jayaram Subramanya defended his dissertation on "Efficient and Responsive Job-Resource Co-adaptivity for Deep Learning Workloads in Large Heterogeneous GPU Clusters."
- Ziyue Qiu had a summer internship with Uber.
- Yiwei Zhao was research scientist intern in Meta mentored by Ziyun Li and Chiao Liu. His research topic studied the design of efficient multimodal algorithms and systems.
- Eliot Solomon had a summer internship at Oracle in Redwood Shores, CA. He worked on user-space page fault handling techniques based on userfaultfd for GraalOS.
- Hyoungjoo Kim did a summer internship at Microsoft Research working with Yinan Li on GPU-based analytical databases.
- Daiyaan Arfeen presented "PipeFill: Using GPUs During Bubbles in Pipeline-parallel LLM Training" the 8th Annual Conference on Machine Learning and Systems, Santa Clara, CA.
- Mingkuan Xu proposed his PhD research "Optimization and Simulation of Quantum Circuits."

continued on page 22

continued from page 1

while improving aggregate throughput by 1.35×. Finally, for a modest performance hit under 4%, LithOS's hardware right-sizing provides a quarter of GPU capacity savings on average, while for a 7% hit, LithOS's transparent power management delivers a quarter of a GPU total energy savings on average. Overall, LithOS transparently increases GPU efficiency, establishing a foundation for future OS research on GPUs.

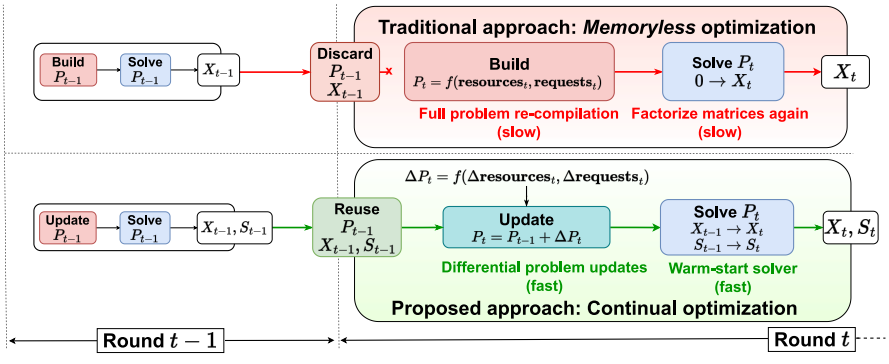
COpter: Efficient Large-Scale Resource-Allocation via Continual Optimization

Suhas Jayaram Subramanya, Don Kurian Dennis, Gregory R. Ganger, Virginia Smith

SOSP '25, October 13–16, 2025, Seoul, Republic of Korea.

Optimization-based resource allocation in large-scale systems often must trade-off responsiveness and allocation quality. Generally, allocations are reconsidered every few minutes (a round) by formulating and solving a new optimization problem. This paper introduces continual optimization, which reframes round-based resource allocation as a sequence of interconnected problems, leveraging the observation that these resource allocation problems often only change by small amounts across successive rounds to reduce solving times. COpter provides a method for continual optimization of Linear Programs (LP) and Mixed Integer Linear Programs (MILP) formulations of resource allocation problems by combining three innovations: (1) an efficient-to-update problem representation for incremental changes, (2) a proximal-point method implementation that can provably benefit from prior computational effort and allocations, and (3) lightweight heuristics for mixed-integer problems that recover feasible integer solutions with negligible quality loss. We evaluate COpter on problems in three domains: GPU cluster scheduling, shard load balancing, and WAN traffic engineering.

Overall, we find that COpter finds high-quality solutions while reducing solver runtimes by 57–83× compared to state-of-the-art commercial solvers. Compared to problem partitioning approaches (POP), COpter simultaneously improves allocation quality and reduces end-to-end allocator runtimes by 1.5–30×.



Existing approaches for RA problems cannot scale to large sizes because: (a) any changes to a problem (from adding/removing resources and/or requests) require problem recompilation; (b) recompiling a problem discards any solver work done for previous rounds; and (c) warm-starting is either not supported or often not beneficial in reducing solver runtimes. We propose continual optimization and implement a prototype (COpter) with techniques designed for (a) efficient problem manipulation, (b) solver state re-use, and (c) efficient warm-starting.

Moirai: Optimizing Placement of Data and Compute in Hybrid Clouds

Ziyue Qiu, Hojin Park, Jing Zhao, Yukai Wang, Arnav Balyan, Gurmeet Singh, Yangjun Zhang, Suqiang (Jack) Song, Gregory R. Ganger, George Amvrosiadis

SOSP '25, October 13–16, 2025, Seoul, Republic of Korea.

The deployment of large-scale data analytics between onpremise and cloud sites, i.e., hybrid clouds, requires careful partitioning of both data and computation to avoid massive networking costs. We present Moirai, a cost-optimization framework that analyzes job accesses and data dependencies and optimizes the placement of both in hybrid clouds. Moirai informs the job scheduler of data location and access predictions, so it can determine where jobs should be executed to minimize data transfer costs. Our optimizer achieves scalability and cost efficiency by exploiting recurring jobs to identify data dependencies and job access characteristics and reduces the search space by excluding data not accessed recently.

We validate Moirai using 4-month traces that span 66.7M queries accessing 13.3EB from Presto and Spark clusters deployed at Uber, a multinational transportation company leveraging large-scale data analytics for its operations. Moirai reduces hybrid cloud deployment costs by over 97% relative to the state-of-the-art partitioning approach from Alibaba and other public approaches. The savings come from 95–99.5% reduction in cloud egress, up to 99% reduction in replication, and 89–98% reduction in on-premises network infrastructure requirements. We also describe concrete steps being taken towards deploying Moirai in production.

continued on page 6

continued from page 5

A Hot Take on the Intel Analytics Accelerator for Database Management Systems

Christos Laspias, Andrew Pavlo, Jignesh M. Patel

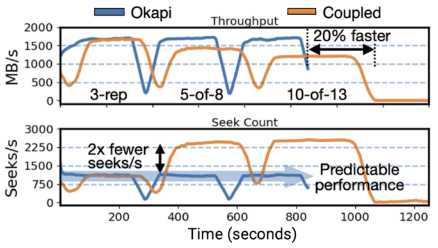
Sixteenth International Workshop on Accelerating Analytics and Data Management Systems Using Modern Processor and Storage Architectures in conjunction with VLDB 2025, London, U.K. September 1, 2025.

For as long as database management systems (DBMSs) have existed, there have been efforts to develop specialized hardware to accelerate their workloads. The goal is clear: to offload the DBMS's most common and repetitive tasks to hardware, thereby improving efficiency and performance. Recently, Intel has released CPUs with new accelerators located on the same die, such as the In-Memory Analytics Accelerator (IAA) that targets data processing tasks. In this work, we examine the Intel IAA's ability to optimize data compression and decompression operations for online analytical processing (OLAP) workloads. To evaluate the benefits of this accelerator, we added support for IAA compression into DuckDB. Our experiments comparing IAA with DuckDB's existing compression method (Snappy) show that it improves decompression speeds by up to 3.15× in microbenchmarks and the end-to-end TPC-H query latencies by up to 38%.

Okapi: Decoupling Data Striping and Redundancy Grouping in Cluster File Systems

Sanjith Athlur, Timothy Kim, Saurabh Kadekodi, Francisco Maturana, Xavier Ramos, Arif Merchant, K. V. Rashmi, Gregory R. Ganger

19th USENIX Symposium on Operating Systems Design and Implementation. OSDI '25. July 7–9, 2025, Boston, MA, USA.



Performance and resource efficiency of file accesses as they transition to wider EC schemes over their lifetimes. The figure shows seeks/s and throughput observed for 16 MB file reads, first when files are in a 3-way replicated format, second in 5-of-8 and then in 10-of-13 EC scheme.

The Okapi cluster file system decouples how data is spread across disks (data striping) for IO efficiency from how data is erasure coded together (redundancy grouping) for durability. Existing systems couple these two mechanisms' configurations, inducing significant inefficiencies. Decoupling allows grouping to be configured based on reliability and space efficiency goals, while simultaneously allowing striping to be configured based on performance goals. Decoupling also allows redundancy scheme changes from one EC scheme to another (e.g., to react to data temperature or disk failure rate changes) to occur without having to re-write data. Evaluation of an Okapi prototype shows that decoupling can be accomplished with <1% increase in metadata size and file manager memory, and minimal file creation and degraded read resource increase. Experiments demonstrate that decoupling can improve read throughput by 80% and reduce seeks per second by up to 70%, without yielding any data reliability, and reduce the overhead of redundancy transitions by up to 70%.

Anarchy in the Database: A Survey and Evaluation of Database Management System Extensibility

Abigale Kim, Marco Slot, David G. Andersen, Andrew Pavlo

Proc. VLDB Endow., Vol. 18, Iss. 6, pp. 1962-1976, August 2025.

Extensions allow applications to expand the capabilities of database management systems (DBMSs) with custom logic. However, the extensibility environment for some DBMSs is fraught with perils, causing developers to resort to unorthodox methods to achieve their goals. This paper studies and evaluates the design of DBMS extensibility. First, we provide a comprehensive taxonomy of the types of DBMS extensibility. We then examine the extensibility of six DBMSs: PostgreSQL, MySQL, MariaDB, SQLite, Redis, and DuckDB. We present an automated extension analysis toolkit that collects static and dynamic information on how an extension integrates into the DBMS. Our evaluation of over 400 PostgreSQL extensions shows that 16.8% of them are incompatible with at least one other extension and can cause system failures. These results also show the correlation between these failures and factors related to extension complexity and implementation.

Optimal Batch-Dynamic kd-trees for Processing-in-Memory with Applications

Yiwei Zhao, Hongbo Kang, Yan Gu, Guy E. Blelloch, Laxman Dhulipala, Charles McGuffey, Phillip B. Gibbons

SPAA '25, July 28–August 1, 2025, Portland, OR, USA.

The kd-tree is a widely used data structure for managing multidimensional data. However, most existing kd-tree designs suffer from the memory wall—bottlenecked by off-chip memory latency and bandwidth limitations. Processing-in-memory (PIM), an emerging architectural paradigm, offers a promising solution to this issue by integrating processors (PIM cores) inside memory modules and offloading computational tasks to these PIM cores. This approach enables low-latency on-chip memory access and

continued on page 7

continued from page 6

provides bandwidth that scales with the number of PIM modules, significantly reducing off-chip memory traffic.

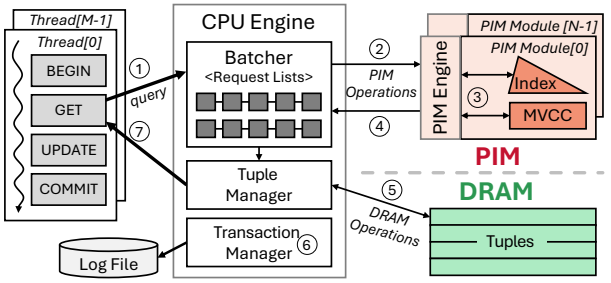
This paper introduces PIM-kd-tree, the first theoretically grounded kd-tree design specifically tailored for PIM systems. The PIM-kd-tree is built upon a novel log-star tree decomposition that leverages local intra-component caching. In conjunction with other innovative techniques, including approximate counters with low overhead for updates, delayed updates for load balancing, and other PIM-friendly aspects, the PIM-kd-tree supports highly efficient batch-parallel construction, point searches, dynamic updates, orthogonal range queries, and kNN searches. Notably, all these operations are work-efficient and load-balanced even under adversarial skew, and incur only O(log* P) communication overhead (off-chip memory traffic) per query. Furthermore, we prove that our data structure achieves *whp*, an optimal trade-off between communication, space, and batch size. Finally, we present efficient parallel algorithms for two prominent clustering problems, density peak clustering and DBSCAN, utilizing the PIM-kd-tree and its techniques.

EMT: An OS Framework for New Memory Translation Architectures

Siyuan Chai, Jiyuan Zhang, Jongyul Kim, Alan Wang, Fan Chung, Jovan Stojkovic, Weiwei Jia, Dimitrios Skarlatos, Josep Torrellas, Tianyin Xu

19th USENIX Symposium on Operating Systems Design and Implementation, July 7–9, 2025, Boston, MA.

With terabyte-scale memory capacity and memory-intensive workloads, memory translation has become a major performance bottleneck. Many novel hardware schemes are developed to speed up memory translation, but few are experimented with commodity OSes. A main reason is that memory management in major OSes, like Linux, does not have the extensibility to empower emerging hardware schemes. We develop EMT, a pragmatic framework atop Linux to empower different hardware schemes of memory translation such as radix tree and hash table. EMT provides an architecture-neutral interface that 1) supports diverse memory translation architectures, 2) enables hardware-specific optimizations, 3) accommodates modern hardware and OS complexity, and 4) has negligible overhead over hardwired implementations. We port Linux's memory management onto EMT and show that EMT enables extensibility without sacrificing performance. We use EMT to implement OS support for ECPT and FPT, two recent experimental translation schemes for fast translation; EMT enables us to understand the OS perspective of these architectures and further optimize their designs.



OLTPim Design. Query requests from M worker threads are served by N PIM modules. The batcher (1) combines the requests in the request list and (2) -(4) offloads index and MVCC operations to the PIM engines. Then, the CPU engine (5) fetches the tuples from DRAM, (6) records them in the transaction context, and (7) returns them to the thread.

No Cap, This Memory Slaps: Breaking Through the Memory Wall of Transactional Database Systems with Processing-in-Memory

Hyoungjoo Kim, Yiwei Zhao, Andrew Pavlo, Phillip B. Gibbons

PVLDB, 18(11): 4241-4254, July 2025.

Memory channel bandwidth imposes an upper bound on the performance of online transaction processing (OLTP) on in-memory database management systems (DBMS). Emerging processing-in-memory (PIM) hardware has the potential to overcome this barrier by using small cores in DRAM chips that can read and process data in situ, thereby avoiding moving these data across memory channels. However, naïvely offloading all database components to PIM does not solve the problem due to the characteristics of software components and the limitations of PIM hardware.

In this paper, we present OLTPim, the first end-to-end OLTP DBMS designed for PIM systems. We build a formalized model for the affinity of each database operation towards PIM and use it to decide the partitioning of components on different types of memory. We also design a lightweight batching algorithm to overcome the large PIM control latency while minimizing the batching overhead. We implement and evaluate OLTPim on the latest PIM system from UPMEM with 64 worker threads and 2048 PIM modules. Our results show that OLTPim achieves up to 1.71x throughput and up to 6.14x less per-transaction memory channel traffic over MosaicDB, a state-of-the-art in-memory system.

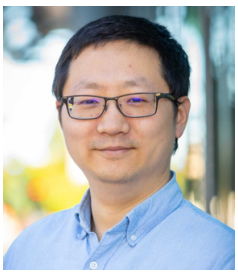
Mirage: A Multi-Level Superoptimizer for Tensor Programs

Mengdi Wu, Xinhao Cheng, Shengyu Liu, Chunan Shi, Jianan Ji, Man Kit Ao, Praveen Velliengiri, Xupeng Miao, Oded Padon, Zhihao Jia

arXiv:2405.05751v3 [cs.LG] 6 Jun 2025.

continued on page 15

October 2025
Juncheng Yang Receives DMR Thesis Award!



Congratulations to PDL Alum Juncheng Yang (advised by Rashmi Vinayak) on receiving the SIGOPS 2025 DMR Thesis Award for his dissertation on Designing Efficient and Scalable Key-value Cache Management Systems. The award, officially the ACM SIGOPS Dennis M. Ritchie Doctoral Dissertation Award, is an annual award that recognizes outstanding doctoral research in software systems. Created by the ACM Special Interest Group on Operating Systems (SIGOPS), it was established to honor the legacy of Dennis Ritchie, a key figure in computer science, and to encourage creativity in the field. Each year, the award is given to a recent Ph.D. graduate whose dissertation demonstrates exceptional innovation and impact on software systems.

Juncheng is now an assistant professor of the Harvard John A. Paulson School of Engineering and Applied Sciences in Cambridge, MA.

September 2025
PDL Alum Jure Leskovec Receives CMU 2025 Alumni Achievement Award!

Jure Leskovec's pioneering work in data science, machine learning and network science has shaped the way complex systems are studied and applied across academia, industry and society, proving what can be accomplished when talented, driven people put their hearts in the work. The awards are presented to alumni for exceptional accomplishment and leadership in their fields or vocations.

Jure is a professor of computer science at Stanford University, where he pioneered the field of graph machine learning. The technology has revolutionized the way in which complex data is analyzed and has become a cornerstone in modern artificial intelligence and machine learning. His innovative approach has drawn interest from computer scientists and researchers across the globe, who have made his Stanford course, Machine Learning with Graphs, one of the most popular machine learning courses on YouTube with more than 2 million views.

Jure is dedicated to open science and fostering collaboration between academia and industry. In September of 2021, he released pyg.org, an online library for machine learning on graphs that has more than 100,000 monthly downloads. His earlier projects include the Stanford Network Analytics Platform, which facilitates research and learning, and the Open Graph Benchmark, which offers diverse, large-scale datasets and tools.

Jure is also a successful entrepreneur, whose work has made an impact on both industry and global policy. He has founded several companies including Kumo, a startup that applies graph machine learning to enterprise data and following the COVID-19 pandemic, he published a study of reopening strategies that was used by the U.S., Poland and Japan to shape their country's policies.

Prior to Stanford, he worked as chief scientist at Pinterest, where he built several foundational AI platforms that lead to a significant improvement in key business metrics. His work has received more than 200,000 citations, has been published in top scientific journals like Nature and Science, and is featured regularly in major press outlets such as The New York Times, Wall Street Journal, Forbes and Business Insider. He is the recipient of numerous honors including the ACM SIGKDD Innovation Award, the Lagrange Prize, an Alfred P. Sloan Fellowship and an honorary doctorate from the University of Antwerp.

Jure earned a bachelor's degree in computer science from the University of Ljubljana in Slovenia and a master's degree in knowledge discovery and data mining and Ph.D. in machine learning from Carnegie Mellon's School of Computer Science. He completed postdoctoral training at Cornell University.

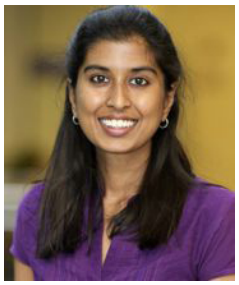
-- info from Engage With CMU, September 22, 2025

September 2025
Sam Arch and Co-authors Receive Best Research Paper Runner-up at VLDB'25

Congratulations to Sam Arch and co-authors Yuchen Liu, Todd Mowry, Jignesh Patel, and Andrew Pavlo on receiving the Best Research Paper Runner-Up award at the 51st International Conference on Very Large Data Bases, held in London, United Kingdom from September 1-5, 2025. Their paper "The Key to Effective UDF Optimization: Before Inlining, First Perform Outlining" discusses augmenting SQL's declarative approach with user-defined functions (UDFs). Problems present as the mismatch between programming paradigms creates a fundamental optimization challenge. UDF inlining automatically removes all UDF calls by replacing them with equivalent SQL subqueries.

July 2025
Akshitha Sriraman Wins an Inaugural 2025 Google ML and Systems Junior Faculty Award

Congratulations to Akshitha as she receives an inaugural 2025 Google ML and Systems Junior Faculty Award in recognition of the significance and



promise of her work in HW/SW Co-Design and Acceleration. These new awards from Google for ML and systems pioneers in academia are going to more than 50 assistant professors in 27 U.S. universities whose research is particularly noteworthy for Google. Akshitha will participate in a future symposium with her fellow awardees and will liaise with a Google technical point-of-contact to facilitate further research.

-- info from https://blog.google/products/google-cloud/ml-systems-junior-faculty-awards/

June 2025
Dimitrios Skarlatos and Todd Mowry Receive Amazon Research Awards!

Congratulations to Dimitrios and Todd on receiving 2025 Amazon Research Awards. Amazon Research Awards to support work in areas such as artificial intelligence, cryptography and automated reasoning. The awards recognize innovative academic work with the potential for broad societal and scientific impact, and provide recipients with unrestricted funding, Amazon Web Services (AWS) promotional credits, and access to Amazon's cloud computing tools and public datasets. Todd Mowry, a professor in the Computer Science Department (CSD), has been awarded for his project, "Efficient LLM Serving on Trainium via Kernel Generation." Dimitrios Skarlatos, an assistant professor in CSD, was awarded for his project, "Scale-Out FHE LLMs on GPUs." Recipients have access to more than 700 Amazon public datasets and can use AWS services and tools. They also consult with an Amazon research contact and can participate in Amazon events and training sessions.

-- info from CMU SCS News, June 13, 2025.

June 2025
Dimitrios Skarlatos Wins IEEE TCCA Young Architect Award at ISCA '25!

We are pleased to announce that Dimitrios has been recognized as the IEEE TCCA 2025 Young Architect! The award recognizes outstanding research contributions by an individual in the field of Computer Architecture, who received his/her PhD degree within the last 6 years. It is awarded annually at the International Symposium on Computer Architecture (ISCA), this year held in Tokyo, Japan. Dimitrios was named for his contributions to virtual memory management and computer security. Dimitrios is an Assistant Professor in the Computer Science Department at Carnegie Mellon University where he leads the CAOS group. His research bridges computer architecture and operating systems focusing on performance, security, and scalability, and current work follows two central themes: (a) uncovering security vulnerabilities and building defenses at the boundary between hardware and OS, and (b) redesigning abstractions and interfaces between the two layers to improve performance and scalability.

-- info from https://www.cs.cmu.edu/~dskarlat and https://ieeetcca.org/awards/young-computer-architect-award/

February 2025
Zhihao Jia is a 2025 Sloan Research Fellow

Congratulations to Zhihao Jia, who has been named a Sloan Research Fellows of 2025. The 126 scholars awarded this honor represent the most promising early-career scientists working today. Their achievements and potential place them among the next generation of scientific leaders in the U.S. and Canada. Winners receive $75,000, which may be spent over a two-year term on any expense supportive of their research.

Zhihao's research interests lie in the intersection of computer systems and machine learning (ML). In particular, his current research focuses on building efficient, scalable, and high-performance software systems for emerging ML applications, such as large language models and generative AI tasks.

-- info from https://sloan.org/fellowships/2025-fellows

January 2025
Gauri Joshi Named 2025 Goldsmith Lecturer

The PDL along with the IEEE Information Theory Society is pleased to announce that Gauri Joshi has been named the 2025 Goldsmith Lecturer. The Goldsmith Lecturer is a woman, no more than ten years beyond having her highest degree conferred, selected for the quality of her research contri-

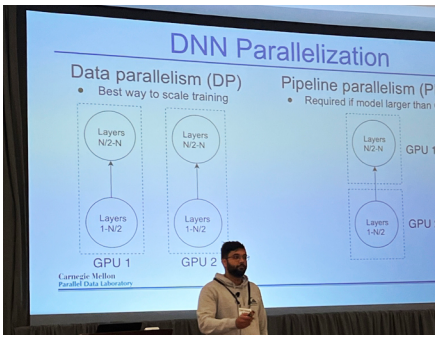
continued on page 24

THESIS PROPOSAL: On Holistic Database Optimization via Leveraging Similarity Across Actions, Workloads, Configurations, and Scenarios

William Zhang, CSD
Tuesday, September 23, 2025

Modern database management systems (DBMSs) have evolved to support increasingly sophisticated data-intensive applications, at the cost of substantial complexity to configure them for two reasons. First, DBMSs expose a vast configuration space with trillions of possibilities that encompass system knobs, physical design (e.g., indexes), and query options, amongst others. Second, these applications are constantly evolving with changes in data access patterns, query types, load intensities, hardware, and data distributions that necessitate continuous re-optimization.

To address these challenges, decades of autonomous DBMS optimization research have produced specialized tuning tools to assist human operators. Deploying these tools involves a complex multi-step workflow where an operator (1) observes the DBMS's behavior, (2) selects tools based on the objectives and their expertise, (3) configures them with an isolated environment, (4) orchestrates their execution to obtain recommendations, and (5) reviews those recommendations before deployment. This cumbersome process results in suboptimal configurations and slow adaptation to evolving applications' workloads due to isolated specialized tools, inefficient reuse of prior tuning knowledge, and the fallible human factor.



Daiyaan Arfeen presents his work on PipeFill: Using GPUs During Bubbles in Pipeline-parallel LLM Training at the 2024 PDL Retreat.

This proposal presents techniques for addressing those limitations with similarity to enable holistic database optimization. First, we present a holistic tuning tool that optimizes multiple DBMS aspects simultaneously by using action similarity to organize actions into neighborhoods conducive to exploration. We then present a framework that assists tuners in adapting to environment changes by leveraging workload and configuration similarity to re-mix historical knowledge.

We propose to extend our preliminary work by transforming the human-centric tuning workflow into an agentic process through scenario similarity. We will first investigate contextualizing deployments and creating semantic tool interfaces. We will then design an orchestrator that learns to select and deploy relevant tuning tools to obtain validated recommendations. With these efforts, the agentic process will enable holistic DBMS optimization throughout its lifetime.

MASTERS THESIS: Towards Utilizing Cached Context for Faster and Smarter Code Agents

Frank Chen
Thursday, September 4, 2025

Recent advances in Large Language Models (LLMs) have enabled the development of coding agents that can autonomously perform tasks such as code generation, debugging, patch validation, etc. Industry-proprietary systems (e.g., Colab AI, Claude Code, Cursor Agent) and open-source frameworks (e.g., Gemini Cli, SWE-Agent, AutoCodeRover, Agentless) have demonstrated both practicality and popularity in real-world software engineering workflows. Despite these successes, existing agentic frameworks face two common challenges: providing accurate and complete context information and ensuring low-latency patch generation under heavy workloads. Although prior work has proposed partial solutions addressing either context retrieval or latency, little attention has been paid to the joint optimization of both aspects. Ideally, joint optimization should enhance both performance and speed without incurring additional cost, which is particularly critical in modern fast-paced, iterative software engineering environments. This thesis investigates integrating existing state-of-the-art approaches to these challenges, specifically RepoGraph for smart context retrieval on the frontend and CacheBlend for fast inference on the backend. To verify feasibility of an integration, we then evaluate performance of the frontend across multiple LLMs under realistic code-agent scenarios, and measure the latency improvement of the backend against systems such as vLLM and SGLang on trace generated by frontend under the same benchmark. The results highlight the trade-offs between cost, efficiency, and performance and argue for the necessity of integrated solutions that achieve balance between the factors mentioned. Finally, we suggest a preliminary design for an end-to-end system that combines the benefits of RepoGraph and CacheBlend via a

continued on page 11

Jason Boles assisting Jaylen Wang with AV setup prior to Jason's retreat presentation.

continued from page 10

simple adapter module, along with optimizations in the original algorithms. Overall, the findings suggest promising directions toward building a robust and production-ready coding agent that is both fast and high-performing.

DISSERTATION ABSTRACT: Refining Classical Abstractions of Network Subsystems

Nirav Atre
Friday, August 29, 2025

We reason about computer systems via models of their behavior — whether implicit mental models, or explicit mathematical models. These models are the linchpins of our decision-making ability, e.g., in formulating service-level agreements (SLAs), or tendering performance claims. Unfortunately, a growing disconnect between how systems are modeled and how they are actually deployed has engendered a class of problems I call model incongruity: circumstances where a model's prediction deviates significantly from real-world behavior. Model incongruities are highly pervasive in modern systems, resulting in expensive performance anomalies, scalability bottlenecks, and security vulnerabilities.

In this thesis, we argue that many incongruities observed in practice today are not a fundamental limitation of our modeling capabilities, but rather artifacts of using the wrong models. We show that: (a) assumptions centrally underpinning contemporary models of network subsystems have drifted far from deployment realities; (b) these assumptions are frequently violated in the field, subverting the operator's expectations about key metrics in highly unexpected ways; and, (c) making modest model refinements not only yields designs with state-of-the-art performance, attack resilience, and scalability, but also enables us to make rigorous mathematical guarantees about the resulting system's behavior.

We exemplify this point using case studies of three ubiquitous network subsystems. First, I will describe "delayed hits", an incongruity arising in high-performance caching systems which breaks the textbook caching principle that maximizing cache hit-rate also minimizes latency, and causes every existing caching algorithm to make latency-suboptimal decisions; in this context, I will introduce Minimum-Aggregate-Delay (MAD), a turnkey augmentation to existing algorithms that makes them aware of delayed hits, yielding 5-35% lower request latencies. Second, I will describe "algorithmic complexity attacks" (ACAs), a highly potent class of Denial-of-Service attacks arising from transient workload incongruity; in this context, I will introduce SurgeProtector, an adversarial scheduling framework that provably protects network dataplanes against ACAs, resulting in 90-99% reduction in harm for the same volume of attack traffic. Finally, I will describe BBQ, a system borne out of addressing design incongruity in hardware packet schedulers which, for the first time, makes it feasible to deploy packet scheduling at line-rate on modern switches and SmartNICs.

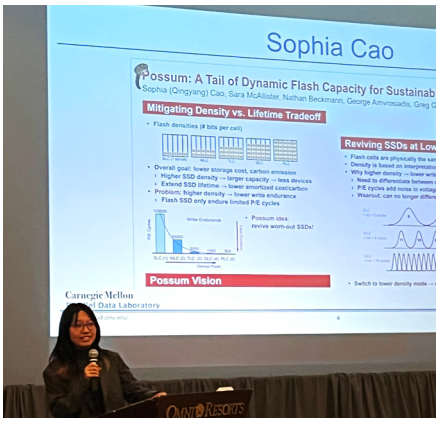
PDL's founder, Garth Gibson with Ippokratis Pandis (Databricks) and Greg Ganger at the FAST 2025 PDL Reunion.

DISSERTATION ABSTRACT: Cost-Efficient Storage and Caching in Public Clouds

Hojin Park
Tuesday, August 26, 2025

As modern data-intensive workloads increasingly migrate to the public cloud, managing the resulting costs has emerged as a pressing challenge despite the operational simplicity and elasticity that cloud environments offer. Although many efforts in cost optimization have focused on computation, storage-related costs have received comparatively less attention despite being a significant portion of total cloud spending. In particular, two categories dominate storage-related costs in public cloud: the cost of deploying and operating storage clusters in the cloud, and the cost of accessing data across geographically distributed regions or clouds. These challenges cannot be effectively addressed by existing optimization techniques developed for on-premise environments, since they often overlook the unique characteristics of public clouds, including elastic resource provisioning, diverse cost-performance trade-offs, and dynamic and unique access patterns found in cloud object storage workloads.

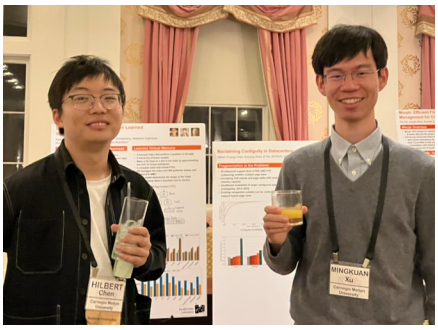
This dissertation addresses these challenges by proposing a cost-efficient approach to designing storage and caching systems that are cloud-aware, elastic, and adaptive to workload behavior. It introduces three systems that target key aspects of cloud storage cost optimization. First, Mimir reduces the cost of the deployment of storage clusters by automatically selecting cost-effective combinations of virtual ma-



Sophia Cao presents a preview of her retreat poster at the PDL Retreat Poster Minutes of Madness session.

continued on page 12

continued from page 11



Hilbert Chen and Mingkuan Xu ready to discuss their posters at the 2024 PDL Retreat.

chines and block storage types, based on profiling workload characteristics and benchmarking available resource options. Second, Macaron reduces cross-region and cross-cloud data access costs by auto-configuring a cache with a tiered storage architecture that leverages low-cost object storage and dynamically resizes the cache based on workload changes. Third, Macaron+ builds on Macaron by introducing a cost-aware prefetching technique that analyzes object-level access patterns to reduce latency in workloads with high cold miss ratios, while preserving cost-efficiency. Together, these systems demonstrate that by tailoring automated resource selection, adaptive configuration, and predictive techniques to the characteristics of the public cloud, it is possible to significantly reduce the cost of storing and accessing data.

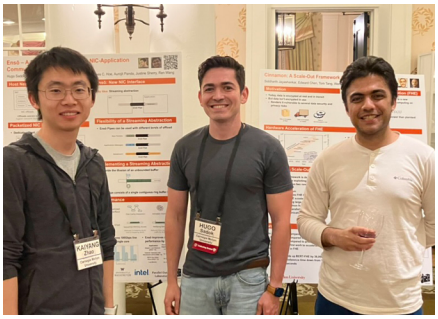
DISSERTATION ABSTRACT: Toward Sustainable Datacenters through Efficient Data Retrieval

Sara McAllister
Wednesday, July 2, 2025

Datacenters are projected to account for 33% of the global carbon emissions by 2050. As datacenters increasingly rely on renewable energy for power, the majority of datacenter emissions will be embodied — emissions from lifecycle stages including acquiring raw materials, manufacturing, transportation, and disposal. To reach the ambitious emission reduction goals set by both companies and governments, datacenters need to reduce emissions throughout their operations, including (and particularly relevant for this thesis) the storage system. Unfortunately, while data storage and retrieval systems are large contributors to embodied emissions, reducing their embodied emissions have largely been overlooked.

This dissertation addresses how to reduce emissions in data retrieval for large-scale storage systems. These storage systems can reduce their carbon footprint by enabling storage devices to have longer lifetimes and use denser media. However, storage hardware's IO limits combined with software's unnecessary additional IO often severely restrict emission reductions, or at worse cause increased emissions. Thus, this thesis focuses on reducing IO in several parts of the storage stack to enable efficient and sustainable data retrieval.

First, this dissertation addresses the sustainability of flash caching, a critical layer in datacenter storage systems that is limited by flash write endurance. This improvement results from two caching systems: Kangaroo and FairyWREN. Together, these caches dramatically reduce writes by over 28x, allowing flash devices to use denser flash for longer lifetimes, ultimately reducing emissions. Then, this thesis discusses enable more sustainable bulk storage, where bandwidth limitations prevent deployment of denser HDDs. Declarative IO, a new interface for distributed storage, empowers the storage system to eliminate duplicate IO accesses in maintenance tasks through exposing the time- and order-flexibility in maintenance tasks. This work enables deployment of larger HDDs, further reducing emissions from storage systems.



Kaiyang Zhao, Hugo Sadok, and Siddharth Jayashankar at a poster session during the 2024 PDL Retreat.

THESIS PROPOSAL: Designing Scalable DNN Training Systems to Overcome Algorithmic Constraints

Daiyaan Arfeen
Tuesday, June 17, 2025

LLM training requires massive amounts of compute due to large model and dataset sizes, so it is not unusual to train LLMs on tens or hundreds of thousands of GPUs to complete training in a reasonable amount of time (days or weeks). However, GPU failures (which are common at these scales) and data-dependencies (introduced by the training algorithms) can lead to severe GPU under-utilization.

In this talk, we present distributed LLM training systems which are efficient and fault-tolerant at these scales. We first present Nonuniform-tensor-parallelism (NTP), a technique which increases the fault-tolerance of tensor-parallel training, thereby reducing the blast-radius of GPU failures. NTP enables scale-up training with little-to-no loss in training efficiency from realistic rates of GPU failures. Next we present PipeFill, a system for recovering GPU utilization (lost due to scale-out training) by filling pipeline bubbles with third-party latency-insensitive jobs. We will discuss how PipeFill could be extended to support filling pipeline bubbles with online inference jobs, which are latency-sensitive.

continued on page 13

continued from page 12

DISSERTATION ABSTRACT: Efficient and Responsive Job-Resource Co-adaptivity for Deep Learning Workloads in Large Heterogeneous GPU Clusters

Suhas Jayaram Subramanya
Tuesday, May 27, 2025

Existing cluster schedulers face many limitations in scheduling adaptive deep learning training jobs on large heterogeneous GPU clusters – many are not heterogeneity-aware, few are adaptivity-aware, and none scale to large clusters without sacrificing allocation fidelity or cluster efficiency. Emerging clusters further complicate this problem with larger, more heterogeneous resources running more increasingly diverse jobs with more dimensions of adaptivity.

This thesis develops new scheduling approaches and algorithms that can (1) scale to emerging clusters with hundreds of thousands of GPUs and many GPU types, (2) quickly optimize high-fidelity allocations for adaptive DL training jobs with low scheduler overhead, and (3) efficiently adapt to changing cluster conditions to improve goodput on the limited GPU resources.

We first introduce Sia — a round-based scheduler that efficiently optimizes adaptive jobs in a heterogeneous cluster with many GPU types. Sia uses GPU resources judiciously to gather information on job-GPU fit-levels using a mix of online and offline profiling, and continuously co-optimizes the GPU resources allocated to jobs and their execution parameters at runtime to maximize cluster-wide training progress. Using job traces derived from real-world data centers, we find that Sia's allocations are fair and efficient, and are quickly computed using an efficient formulation, even for 1000-GPU clusters.

Second, we introduce continual optimization — a new paradigm that explicitly models the slow evolution of resource-allocation problems at scale to reduce solver runtime for quick responses to changes in jobs or resources. We then introduce COpter, our approach to continual optimization that (a) efficiently updates the optimization problems for job and resource changes using a differential interface, (b) implements a factorization-free warm-started LP solver to benefit from slowly-evolving nature of the allocations, and (c) implements lightweight heuristics to recover feasible integral solutions with negligible quality loss. In our evaluations, COpter speeds up Sia scheduler policy by a few orders of magnitude on clusters with tens of thousands of GPUs without sacrificing job completion times and makespan.

Third, COpter is easily applied to resource-allocation problems in other domains (e.g. shard load-balancing, WAN traffic engineering) and we see 57 - 83 × reductions in solver runtimes.

THESIS PROPOSAL: Optimization and Simulation of Quantum Circuits

Mingkuan Xu
Wednesday, May 7, 2025

Optimizing quantum circuits and simulating them at scale remain critical bottlenecks: manual design of quantum circuit optimizations is labor-intensive and device-specific, while simulators struggle with exponential resource costs. This thesis delivers tools to tackle these challenges.

Shuyi Pei (Samsung) and Yiwei Zhao discuss Samsung's research directions at the 2024 PDL Retreat Industry Poster Session.

Bob (Qinghan) Chen discusses his poster on "Lazy Promotion in Cache Eviction: What, How, and Why?" with Kim Keeton (Google) and Ron Minsky (Jane Street) at the 2024 PDL Retreat.

First, I introduce Quartz, a superoptimizer that automates the generation and verification of circuit transformations for arbitrary quantum gate sets. By systematically exploring small circuits and employing an automated theorem prover (Z3), Quartz discovers both expert-designed and novel optimizations, outperforming hand-tuned optimizers across various gate sets.

Next, I present Atlas, a distributed GPU-based simulator that hierarchically partitions circuits to exploit available data parallelism while minimizing communication costs, running over 2× faster than state-of-the-art GPU simulators. Atlas minimizes communication overhead via integer linear programming to allocate "nearby" gates to "nearby" GPUs and maximizes throughput through dynamic programming for kernel scheduling.

Finally, I propose an initial formal verification framework to certify each application of transformation-based optimizers like Quartz, paving the way for full correctness guarantees. Together, these contributions advance automated, scalable, and reliable quantum computing workflows for emerging devices.

continued on page 23

PDL ALUMNI NEWS

Niraj Tolia

PDL 2002 - 2007

Niraj's startup, Alcion was acquired and he took on the position of CTO at Veeam, a company in the data protection space and the largest in terms of market share.

Alexey Tumanov

PDL 2010 - 2016

Alexey has been promoted to Associate Professor with tenure in the College of Computing at the Georgia Institute of Technology.

Ippokratis Pandis

PDL Member: 2005 - 2011

Ippo recently moved from being VP/Distinguished Engineer at Amazon to the position of Distinguished Engineer at Databricks.

David Petrou

PDL Member: 1997 - 2005

David is the founder & CEO at Continua, a consumer-facing company that uses AI agents to enhance collaboration and interaction in group chats on SMS, iMessage, and Discord.

Juncheng Yang

PDL 2018 - 2024

Juncheng is now an assistant professor of computer science at Harvard University. More importantly, his son, Albert Yang was born June 27 2025!

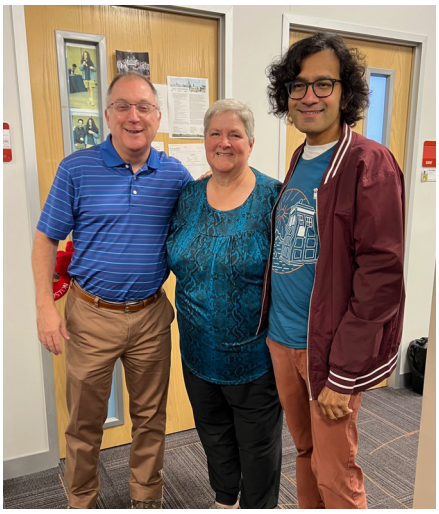
Eno Thereska

PDL 2002 - 2007

Eno is the Co-founder and CEO @ Trent AI, a new venture in the area of AI and security. We'll be in stealth for a while, so more news will follow in the future. Chaochen Qian (CMU alumni, MS 2013) has also joined the startup.

Raja Sambasivan

PDL 2006 - 2016



Raja is the Ankur & Mari Sahu Assistant Professor of Computer Science at Tufts University in Medford, MA. He dropped in to visit Karen and Bill a few weeks ago!

Steve Schlosser

PDL 1998 - 2004

Steve, along with the CMU Bagpipe band, attended the World Pipe Band Championships, held at Glasgow Greens in Glasgow, Scotland this summer. CMU placed 2nd in their grade (3B) for piping, 8th in drumming, and 5th overall. Steve is a software engineer at Google Pittsburgh.

Jure Leskovec

PDL 2004 - 2008

Jure, a professor at Stanford Computer Science and Co-Founder at Kumo.ai, has been awarded the 2025 CMU Alumni Achievement Award. Read more in the PDL News & Awards section.

continued from page 7

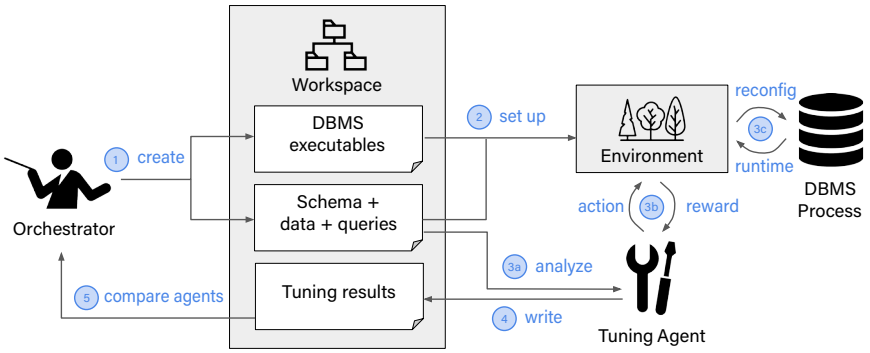
We introduce Mirage, the first multi-level superoptimizer for tensor programs. A key idea in Mirage is μGraphs, a uniform representation of tensor programs at the kernel, thread block, and thread levels of the GPU compute hierarchy. μGraphs enable Mirage to discover novel optimizations that combine algebraic transformations, schedule transformations, and generation of new custom kernels. To navigate the large search space, Mirage introduces a pruning technique based on abstraction that significantly reduces the search space and provides a certain optimality guarantee. To ensure that the optimized μGraph is equivalent to the input program, Mirage introduces a probabilistic equivalence verification procedure with strong theoretical guarantees. Our evaluation shows that Mirage significantly outperforms existing approaches even for DNNs that are widely used and heavily optimized. Mirage is publicly available at https://github.com/mirage-project/mirage.

Automated Database Tuning vs. Human-Based Tuning in a Simulated Stressful Work Environment: A Demonstration of the Database Gym

Patrick Wang, Wan Shen Lim, William Zhang, Samuel Arch, Andrew Pavlo

Companion of the 2025 International Conference on Management of Data, 2025, pp. 247-250. June 22 - 27, 2025, Berlin, Germany.

Machine learning (ML) has gained traction in academia and industry for database management system (DBMS) automation. Although studies demonstrate that ML-based tuning agents match or exceed human expert performance in optimizing DBMSs, researchers continue to build bespoke tuning pipelines from the ground up. The lack of a reusable infrastructure leads to redundant engineering effort and increased difficulty in comparing modeling methods. This paper demonstrates the database gym framework, a standardized training environment that provides a unified API of pluggable components. The database gym simplifies ML model training and evaluation to accelerate autonomous DBMS research. In this demonstration, we showcase the effectiveness of automated tuning and the gym's ease of use by allowing a human expert to compete against an ML-based tuning agent implemented in the gym.



DB-Gym Architecture – An overview of the three components of the DB-Gym: (1) Orchestrator, (2) Tuning Agent, and (3) DBMS Process. The DB-Gym also provides two interfaces (unfilled icons in gray boxes) that these components use to interact with each other: (1) Workspace and (2) Environment. The arrows with blue labels represent the steps in the workflow when using the database gym. First, the Orchestrator creates the DBMS executable and workload files (schema, table data, and queries). Second, the Environment is initialized using these files. Third, the Tuning Agent uses the Environment to tune the DBMS. Fourth, the Tuning Agent writes the results to the Workspace. Lastly, the Orchestrator compares the results across different Tuning Agents.

BPF-DB: A Kernel-Embedded Transactional Database Management System For eBPF Applications

Matthew Butrovich, Samuel Arch, Wan Shen Lim, William Zhang, Jignesh M. Patel, Andrew Pavlo

Proc. ACM Manag. Data, Vol. 3, No. 3 (SIGMOD), Article 135. Publication date: June 2025.

Developers rely on the eBPF framework to augment operating system (OS) behavior for the betterment of database management system (DBMS) without having to modify kernel code. But eBPF's verifier limits program complexity and data management functionality. As a result eBPF's storage options are limited to kernel-resident, non-durable data structures that lack transactional guarantees.

Inspired by embedded DBMSs for user-space applications, this paper present BPF-DB, an OS-embedded DBMS that offers transactional data management for eBPF applications. We explore the storage management and concurrency control challenges associated with DBMS design in eBPF's restrictive execution environment. We demonstrate BPF-DB's capabilities with two applications based on real-world systems. The first is a Redis-compatible in-memory DBMS that uses BPF-DB as its transactional storage engine. This system matches the performance of state-of-the-art implementations while offering stronger transactional guarantees. The second application implements a stored procedure-based DBMS that provides serializable multi-statement transactions. We compare this application against VoltDB, with BPF-DB achieving 43% higher throughput. BPF-DB's robust and high-performance transactional semantics enable emerging kernel-space applications.

continued on page 16

continued from page 15

PipeFill: Using GPUs During Bubbles in Pipeline-parallel LLM Training

Daiyaan Arfeen, Zhen Zhang, Xinwei Fu, Gregory R. Ganger, Yida Wang

Eighth Annual Conference on Machine Learning and Systems, Santa Clara, CA, May 12-15, 2025.

Training Deep Neural Networks (DNNs) with billions of parameters generally involves pipeline-parallel (PP) execution. Unfortunately, PP model training can use GPUs inefficiently, especially at large scale, due to idle GPU time caused by pipeline bubbles, which are often 15–30% and can exceed 60% of the training job's GPU allocation. To improve the GPU utilization of PP model training, this paper describes PIPEFILL, which fills pipeline bubbles with execution of other pending jobs. By leveraging bubble GPU time, PIPEFILL reduces the GPU utilization sacrifice associated with scaling-up of large-model training. To context-switch between fill jobs and the main training job with minimal overhead to the main job, and maximize fill job efficiency, PIPEFILL carefully fits fill job work to measured bubble durations and GPU memory availability, introduces explicit pipeline-bubble instructions, and orchestrates placement and execution of fill jobs in pipeline bubbles. Experiments show that PIPEFILL can increase overall utilization by up to 63% for GPUs used in large-scale LLM training, with <2% slowdown of the training job, and 5–15% even for low-scale LLM training. For large-scale LLM training on 8K GPUs, the 63% increase translates to up to 2.6K additional GPUs worth of work completed.

Building an Elastic Block Storage over EBOFs Using Shadow Views

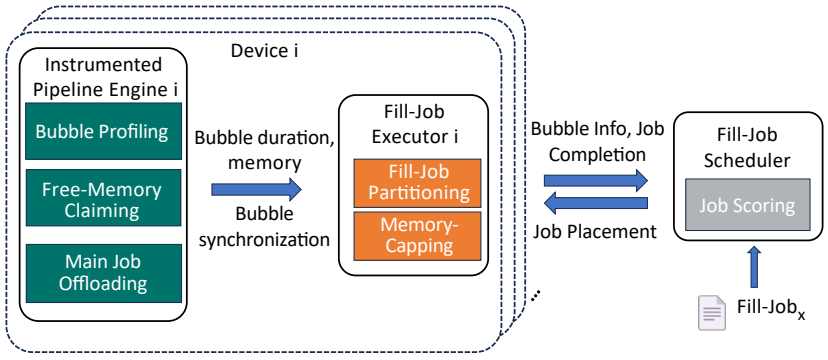
Sheng Jiang, Ming Liu

Proceedings of the 22nd USENIX Symposium on Networked Systems Design and Implementation. April 28–30, 2025 • Philadelphia, PA.

The EBOF (Ethernet-Bunch-Of-Flash) has emerged as an enticing and promising disaggregated storage platform due to its streamlined I/O processing, high scalability, and substantial energy/cost-efficiency improvement. An EBOF applies a smart-sender dumb-receiver design philosophy and provides backward-compatible storage volumes to expedite system deployment. Yet, the static and opaque internal I/O processing pipeline lacks resource allocation, I/O scheduling, and traffic orchestration capabilities, entailing bandwidth waste, workload non-adaptiveness, and performance interference.

This paper presents the design and implementation of a distributed telemetry system (called shadow view) to tackle the above challenges and facilitate the effective use of an EBOF. We model an EBOF as a two-layer multi-switch architecture and develop a view development protocol to construct the EBOF running snapshot and expose internal execution statistics at runtime. Our design is motivated by the observation that fast data center networks make the overheads of interserver communication and synchronization negligible.

We demonstrate the effectiveness of shadow view by building a block storage (dubbed Flint1) atop EBOFs. The enhanced I/O data plane allows us to develop new three techniques–an elastic volume manager, an eIO scheduler, and a view-enabled bandwidth auction mechanism. Our evaluations using the Fungible FS1600 EBOF show that a Flint volume achieves 9.3/9.2 GB/s read/write bandwidth with no latency degradation, significantly outperforming the defacto EBOF volume. It achieves up to 2.9× throughput improvements when running an object store. Flint is tenant-aware and remote target-aware, delivering efficient multi-tenancy and workload adaptiveness.



PipeFill system overview.

Nonuniform-Tensor-Parallelism: Mitigating GPU Failure Impact for Scaled-up LLM Training

Daiyaan Arfeen, Dheevatsa Mudigere, Ankit More, Bhargava Gopireddy, Ahmet Inci, Gregory R. Ganger

arXiv:2504.06095v1 [cs.DC] 8 Apr 2025.

LLM training is scaled up to 10Ks of GPUs by a mix of data- (DP) and model-parallel (MP) execution. Critical to achieving efficiency is tensor-parallel (TP; a form of MP) execution within tightly-coupled subsets of GPUs, referred to as a scaleup domain, and the larger the scale-up domain the bet-

cont. on page 17

continued from page 16

ter the performance. New datacenter architectures are emerging with more GPUs able to be tightly-coupled in a scale-up domain, such as moving from 8 GPUs to 72 GPUs connected via NV-Link. Unfortunately, larger scale-up domains increase the blast-radius of failures, with a failure of single GPU potentially impacting TP execution on the full scale-up domain, which can degrade overall LLM training throughput dramatically. With as few as 0.1% of GPUs being in a failed state, a high TP-degree job can experience nearly 10% reduction in LLM training throughput.

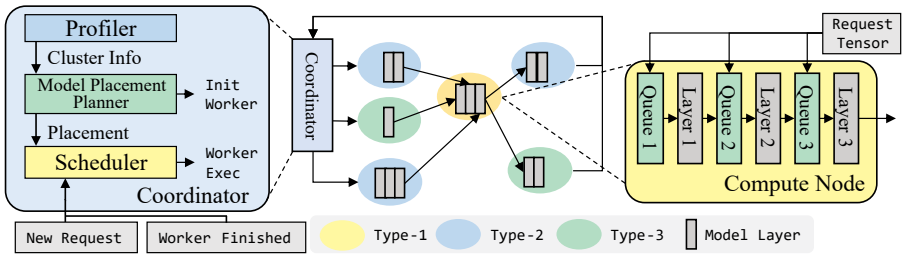
We propose nonuniform-tensor-parallelism (NTP) to mitigate this amplified impact of GPU failures. In NTP, a DP replica that experiences GPU failures operates at a reduced TP degree, contributing throughput equal to the percentage of still-functional GPUs. We also propose a rack-design with improved electrical and thermal capabilities in order to sustain power-boosting of scale-up domains that have experienced failures; combined with NTP, this can allow the DP replica with the reduced TP degree (i.e., with failed GPUs) to keep up with the others, thereby achieving near-zero throughput loss for large-scale LLM training.

Helix: Serving Large Language Models over Heterogeneous GPUs and Network via Max-Flow

Yixuan Mei, Yonghao Zhuang, Xupeng Miao, Juncheng Yang, Zhihao Jia, and Rashmi Vinayak

ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), March 30–April 3, 2025, Rotterdam, Netherlands.

This paper introduces Helix, a distributed system for high-throughput, low-latency large language model (LLM) serving in heterogeneous GPU clusters. The key idea behind Helix is to formulate inference computation of LLMs over heterogeneous GPUs and network connections as a max-flow problem on directed, weighted graphs, whose nodes represent GPU instances and edges capture both GPU and network heterogeneity through their capacities. Helix then uses a mixed integer linear programming (MILP) algorithm to discover highly optimized strategies to serve LLMs on heterogeneous GPUs. This approach allows Helix to jointly optimize model placement and request scheduling, two highly entangled tasks in heterogeneous LLM serving. Our evaluation on several heterogeneous clusters ranging from 24 to 42 GPU nodes shows that Helix improves serving throughput by up to 3.3× and reduces prompting and decoding latency by up to 66% and 24%, respectively, compared to existing approaches. Helix is available at https://github.com/Thesys-lab/Helix-ASPLOS25.



Helix overview. In Helix, the coordinator plans model placement as described in Sec. 4.4. We only need to run model placement once for each cluster. When a new request arrives, the coordinator node runs Helix scheduler to assign it a per-request pipeline and sends it to the first node in the pipeline. Each compute node in the pipeline performs inference on the request on the layers it is responsible for and sends the (output for the) request to the next node in the pipeline. When the last node in the pipeline finishes performing inference on its layers, it will send the output token for the request to the coordinator (Worker Finished). The coordinator schedules generation of the next token for the request using the same pipeline.

GraphPipe: Improving Performance and Scalability of DNN Training with Graph Pipeline Parallelism

Byungsoo Jeon, Mengdi Wu, Shiyi Cao, Sunghyun Kim, Sunghyun Park, Neeraj Aggarwal, Colin Unger, Daiyaan Arfeen, Peiyuan Liao, Xupeng Miao, Mohammad Alizadeh, Gregory R. Ganger, Tianqi Chen, Zhihao Jia

ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), March 30–April 3, 2025, Rotterdam, Netherlands.

Deep neural networks (DNNs) continue to grow rapidly in size, making them infeasible to train on a single device. Pipeline parallelism is commonly used in existing DNN systems to support large-scale DNN training by partitioning a DNN into multiple stages, which concurrently perform DNN training for different micro-batches in a pipeline fashion. However, existing pipeline-parallel approaches only consider sequential pipeline stages and thus ignore the topology of a DNN, resulting in missed model-parallel opportunities.

This paper presents graph pipeline parallelism (GPP), a new pipeline-parallel scheme that partitions a DNN into pipeline stages whose dependencies are identified by a directed acyclic graph. GPP generalizes existing sequential pipeline parallelism and preserves the inherent topology of a DNN to enable concurrent execution of computationally-independent operators, resulting in reduced memory requirement and improved GPU performance. In addition, we develop

continued on page 18

continued from page 17

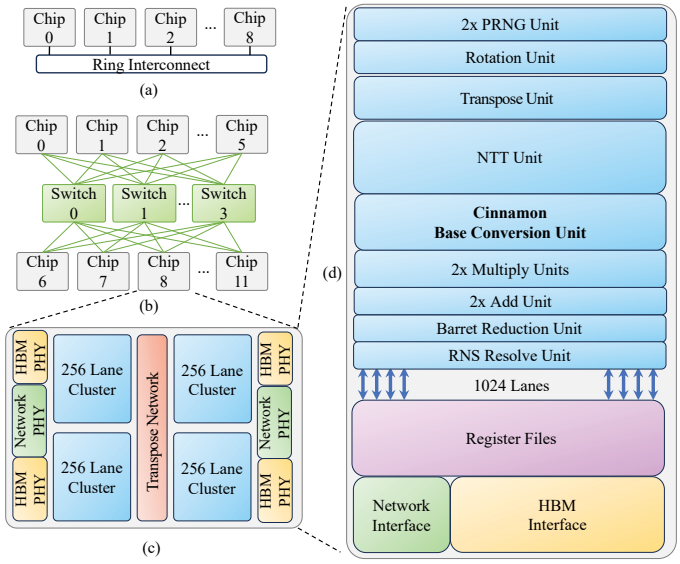
GraphPipe, a distributed system that exploits GPP strategies to enable performant and scalable DNN training. GraphPipe partitions a DNN into a graph of stages, optimizes micro-batch schedules for these stages, and parallelizes DNN training using the discovered GPP strategies. Evaluation on a variety of DNNs shows that GraphPipe outperforms existing pipeline-parallel systems such as PipeDream and Piper by up to 1.6×. GraphPipe also reduces the search time by 9-21× compared to PipeDream and Piper.

Cinnamon: A Framework for Scale-out Encrypted AI

Siddharth Jayashankar, Edward Chen, Tom Tang, Wenting Zheng, Dimitrios Skarlatos

Proceedings of the 30th Intl. Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), Rotterdam, The Netherlands, March 2025.

Fully homomorphic encryption (FHE) is a promising cryptographic solution that enables computation on encrypted data, but its adoption remains a challenge due to steep performance overheads. Although recent FHE architectures have made valiant efforts to narrow the performance gap, they not only have massive monolithic chip designs but also only target small ML workloads. We present Cinnamon, a framework for accelerating state-of-the-art ML workloads that are encrypted using FHE. Cinnamon accelerates encrypted computing by exploiting parallelism at all levels of a program, using novel algorithms, compilers, and hardware techniques to create a scale-out design for FHE as opposed to a monolithic chip design. Our evaluation of the Cinnamon framework on small programs shows a 2.3× improvement in performance compared to prior state-of-the-art designs. Further, we use Cinnamon to show for the first time the scalability of large ML models such as the BERT language model in FHE. Cinnamon achieves a speedup of 36,600× compared to a CPU bringing down the inference time from 17 hours to 1.67 seconds thereby enabling new opportunities for privacy-preserving machine learning. Finally, Cinnamon's parallelization strategies and architectural extensions reduce the required resources per-chip leading to a 5× and 2.68× improvement in performance-per-dollar compared to state-of-the-art monolithic and chiplet architectures respectively.



Cinnamon scale-out architecture shows: (a) Cinnamon with eight chips connected over a ring interconnect, (b) Cinnamon scaling to twelve chips with a switch interconnect, (c) Cinnamon's organization composed of four 256 lane clusters, (d) the logical organization of a Cinnamon chip.

FairyWREN: A Sustainable Cache for Emerging Write-Read-Erase Flash Interfaces

S. McAllister, Y. Wang, B. Berg, D. Berger, N. Beckmann, G. Amvrosiadis, G. Ganger

ACM Trans. on Storage, March 2025.

Datacenters need to reduce embodied carbon emissions, particularly for flash, which accounts for 40% of embodied carbon in servers. However, decreasing flash's embodied emissions is challenging due to flash's limited write endurance, which more than halves with each generation of denser flash. Reducing embodied emissions requires extending flash lifetime, stressing its limited write endurance even further. The legacy Logical Block-Addressable Device (LBAD) interface exacerbates the problem by forcing devices to perform garbage collection, leading to even more writes. Flash-based caches in particular write frequently, limiting the lifetimes and densities of the devices they use. These flash caches illustrate the need to break away from LBAD and switch to the new Write-Read-Erase iNterfaces (WREN) now coming to market. WREN affords applications control over data placement and garbage collection. We present FairyWren†, a flash cache designed for WREN. FairyWren reduces writes by co-designing caching policies and flash garbage collection. FairyWren provides a 12.5× write reduction over state-of-the-art LBAD caches. This decrease in writes allows flash devices to last longer, decreasing flash cost by 35% and flash carbon emissions by 33%.

continued on page 19

continued from page 18

†Fairywrens are vibrant birds native to Australia. Common varieties include Superb Fairywrens, Splendid Fairywrens, and Lovely Fairywrens.

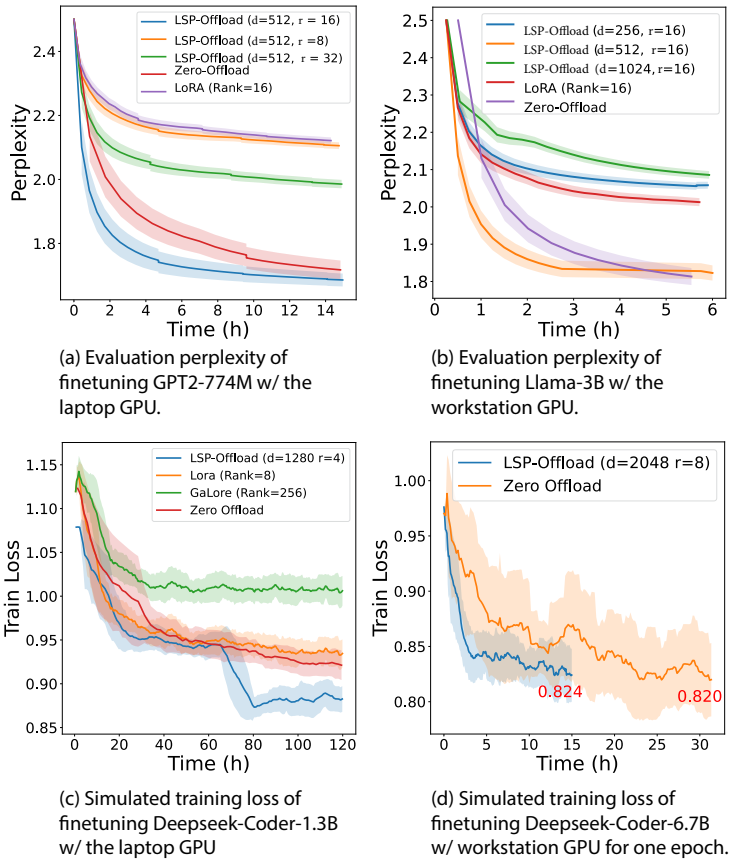
Practical Offloading for Fine-Tuning LLM on Commodity GPU via Learned Sparse Projectors

Siyuan Chen, Zhuofeng Wang, Zelong Guan, Yudong Liu, Phillip B. Gibbons

39th Annual AAAI Conference on Artificial Intelligence, February 25 – March 4, 2025. Philadelphia, PA.

Fine-tuning large language models (LLMs) requires significant memory, often exceeding the capacity of a single GPU. A common solution to this memory challenge is offloading compute and data from the GPU to the CPU. However, this approach is hampered by the limited bandwidth of commodity hardware, which constrains communication between the CPU and GPU, and by slower matrix multiplications on the CPU.

In this paper, we present an offloading framework, LSPOffload, that enables near-native speed LLM fine-tuning on commodity hardware through learned sparse projectors. Our data-driven approach involves learning efficient sparse compressors that minimize communication with minimal precision loss. Additionally, we introduce a novel layer-wise communication schedule to maximize parallelism between communication and computation. As a result, our framework can fine-tune a 1.3 billion parameter model on a 4GB laptop GPU and a 6.7 billion parameter model on an NVIDIA RTX 4090 GPU with 24GB memory. Compared to state-of-the-art offloading frameworks, our approach reduces end-to-end fine-tuning time by 33.1%-62.5% when converging to the same accuracy.



(a) Evaluation perplexity of finetuning GPT2-774M w/ the laptop GPU.

(b) Evaluation perplexity of finetuning Llama-3B w/ the workstation GPU.

(c) Simulated training loss of finetuning Deepseek-Coder-1.3B w/ the laptop GPU

(d) Simulated training loss of finetuning Deepseek-Coder-6.7B w/ workstation GPU for one epoch.

End-to-end evaluation of LSP-Offload. Rolling average is applied for drawing each curve. The shaded area around the line shows the standard deviation.

RTBAS: Defending LLM Agents Against Prompt Injection and Privacy Leakage

Peter Yong Zhong, Siyuan Chen, Ruiqi Wang, McKenna McCall, Ben L. Titzer, Heather Miller, Phillip B. Gibbons

arXiv:2502.08966v2 [cs.CR], 14 Feb 2025.

Tool-Based Agent Systems (TBAS) allow Language Models (LMs) to use external tools for tasks beyond their standalone capabilities, such as searching websites, booking flights, or making financial transactions. However, these tools greatly increase the risks of prompt injection attacks, where malicious content hijacks the LM agent to leak confidential data or trigger harmful actions.

Existing defenses (OpenAI GPTs) require user confirmation before every tool call, placing onerous burdens on users. We introduce Robust TBAS (RTBAS), which automatically detects and executes tool calls that preserve integrity and confidentiality, requiring user confirmation only when these safeguards cannot be ensured. RTBAS adapts Information Flow Control to the unique challenges presented by TBAS. We present two novel dependency screeners–using LM-as-ajudge and attention-based saliency–to overcome these challenges. Experimental results on the AgentDojo Prompt Injection benchmark show RTBAS prevents all targeted attacks with only a 2% loss of task utility when under attack, and further tests confirm its ability to obtain near-oracle performance on detecting both subtle and direct privacy leaks.

continued on page 20

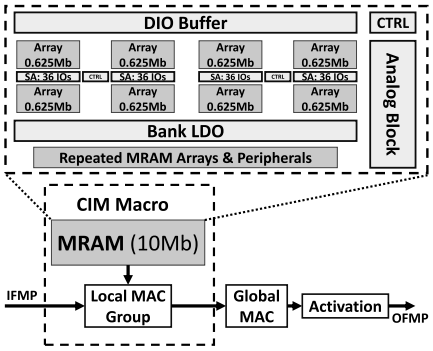
continued from page 19

H4H: Hybrid Convolution-Transformer Architecture Search for NPU-CIM Heterogeneous Systems for AR/VR Applications

Yiwei Zhao, Jinhui Chen, Sai Qian Zhang, Syed Shakib Sarwar, Kleber Hugo Stangherlin, Jorge Tomas Gomez, Jae-Sun Seo, Barbara De Salvo, Chiao Liu, Phillip B. Gibbons, Ziyun Li

30th Asia and South Pacific Design Automation Conference, January 20–23, 2025, Tokyo, Japan.

Low-latency and low-power edge AI is crucial for Augmented/Virtual Reality applications. Recent advances demonstrate that hybrid models, combining convolution layers (CNN) and transformers (ViT), often achieve a superior accuracy/performance tradeoff on various computer vision and machine learning (ML) tasks. However, hybrid ML models can present system challenges for latency and energy efficiency due to their diverse nature in dataflow and memory access patterns. In this work, we leverage architecture heterogeneity from Neural Processing Units (NPU) and Compute-In-Memory (CIM) and explore diverse execution schemas for efficient hybrid model executions. We introduce H4H-NAS, a two-stage Neural Architecture Search (NAS) framework to automate the design of hybrid CNN/ViT models for heterogeneous edge systems featuring both NPU and CIM. We propose a two-phase incremental supernet training in our NAS to resolve gradient conflicts between sampled subnets caused by different block types in a hybrid model search space. Our H4H-NAS approach is also powered by a performance estimator built with NPU performance results measured on real silicon, and CIM performance based on industry IPs. H4H-NAS searches hybrid CNN-ViT models with fine granularity and achieves significant (up to 1.34%) top-1 accuracy improvement on ImageNet-1k. Moreover, results from our algorithm/hardware co-design reveal up to 56.08% overall latency and 41.72% energy improvements by introducing heterogeneous computing over baseline solutions. Overall, our framework guides the design of hybrid network architectures and system architectures for NPU+CIM heterogeneous systems.



Architecture layout of our MRAM CIM macro. IFMP/OFMP stand for input/output feature maps.

Towards Functional Decomposition of Storage Formats

Martin Prammer, Xinyu Zeng, Ruijun Meng, Wes McKinney, Huanchen Zhang, Andrew Pavlo, Jignesh M. Patel

15th Annual Conference on Innovative Data Systems Research (CIDR '25). January 19-22, Amsterdam, The Netherlands.

The rise of data lakes containing mostly semi-structured and unstructured data has changed how traditional data platforms interact with collections of stand-alone files. Horizontally partitioned arrays are a fundamental construction in these columnar-like file formats, such as those partitioned into a column-chunk, row-group hierarchy (e.g., Parquet, ORC). Compressing each horizontal partition results in storage savings. Simultaneously, row-skipping metadata is a popular, lightweight indexing technique for accelerating columnar scans. Thus, existing storage-layer partitions are also used for general-purpose search acceleration. However, no single horizontal partition size optimizes both compressibility and row-skipping-driven scan performance.

Instead of settling for a suboptimal configuration, we return to the age-old wisdom of physical data independence: data should be kept separate from indexing structures. We propose splitting the current status quo into a "storage layer" and a "search acceleration layer" (SAL). By splitting these layers, row-skipping metadata is no longer stuck using the same partition sizes as compression blocks, allowing for fine-grained tuning of the SAL. In this paper, we explore the impact of such a split; not only do we find that search acceleration metadata is regularly optimal at small partition sizes (10-100), but also that optimal sizing depends on the metadata type, underlying data, and applied query. By separating the storage layer and SAL, we enable each to evolve independently, allowing for greater flexibility as datasets and application needs evolve.

Can Increasing the Hit Ratio Hurt Cache Throughput?

Ziyue Qiu, Juncheng Yang, Mor Harchol-Balter

EAI International Conference on Performance Evaluation Methodologies and Tools, December 12-13, 2024 Milan, Italy. BEST PAPER AWARD AT VALUETOOLS '24!

Software caches are an intrinsic component of almost every computer system. Consequently, caching algorithms, particularly eviction policies, are the topic of many papers. Almost all these prior papers evaluate the caching algorithm based on its hit ratio, namely the fraction of requests that are found in the cache, as opposed to disk. The "hit ratio" is viewed as a proxy for traditional performance metrics like system throughput or

continued on page 21

continued from page 20

request latency. Intuitively it makes sense that higher hit ratio should lead to higher throughput (and lower request latency), since more requests are found in the cache (low access time) as opposed to the disk (high access time).

This paper challenges this intuition. We show that increasing the hit ratio can actually hurt the throughput (and request latency) for many caching algorithms. Our investigation follows a three-pronged approach involving (i) queueing modeling and analysis, (ii) simulation to validate the accuracy of the queueing model, and (iii) implementation and mea- surement. We also show that the phenomenon of decreasing throughput at higher hit ratios is likely to be more pronounced in future systems, where the trend is towards faster disks and more cores per CPU.

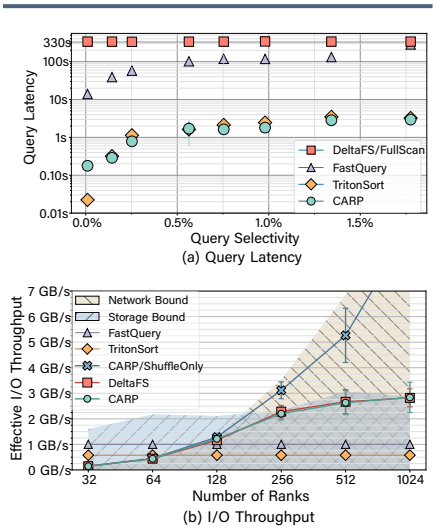
CARP: Range Query-Optimized Indexing for Streaming Data

Ankush Jain, Charles D. Cranor, Qing Zheng, Bradley W. Settlemeyer, George Amvrosiadis, Gary Grider

SC'24, November 17-22, 2024, Atlanta, Georgia, USA.

Ingestion of data generated by high-performance scientific applications continues to stress available storage resources. Efficient range-based analyses on this data can be enabled by reordering it on attributes of interest, but require expensive post-processing sorts to realize the query benefits of reordering. In-situ indexing techniques, while write-efficient, are orders of magnitude slower at range queries than sorted indices. Range queries are necessary for analyzing continuous physical attributes and tracking phenomena such as energy bands and wave fronts.

We present CARP, a scalable data partitioner for range queries that reorders data in-situ as it is streamed to storage during application I/O. Motivated by our findings that real application distributions tend to be highly skewed and dynamic, CARP dynamically discovers



(a) CARP matches the query latency of TritonSort's fully-sorted data layout and outperforms both FastQuery by 100×. (b) CARP incurs no overhead over unpartitioned I/O and is 3-5× faster than post-processing.

and adapts its data partitions to track these characteristics. As a result, CARP can approximate the query performance of a sort without any ingestion overhead, making it 5× faster than prior work.

The Key to Effective UDF Optimization: Before Inlining, First Perform Outlining

Samuel Arch, Yuchen Liu, Todd Mowry, Jignesh Patel, Andrew Pavlo

Proceedings of the VLDB Endowment, Vol. 18, No. 1., December 2024.

Although user-defined functions (UDFs) are a popular way to augment SQL's declarative approach with procedural code, the mismatch between programming paradigms creates a fundamental optimization challenge. UDF inlining automatically removes all UDF calls by replacing them with equivalent SQL subqueries. Although inlining leaves queries entirely in SQL (resulting in large performance gains), we observe that inlining the entire UDF often leads to sub-optimal performance. A better approach is to analyze the UDF, deconstruct it into smaller pieces, and inline only the pieces that help query optimization. To achieve this, we propose UDF outlining, a technique to intentionally hide pieces of a UDF from the optimizer, resulting in simpler UDFs and significantly faster query plans. Our implementation (PRISM) demonstrates that UDF outlining improves performance over conventional inlining (on average 1.29× speedup for DuckDB and 298.73× for SQL Server) through a combination of more effective unnesting, improved data skipping, and by avoiding unnecessary joins.

Atlas: Hierarchical Partitioning for Quantum Circuit Simulation on GPUs

Mingkuan Xu, Shiyi Cao, Xupeng Miao, Umut A. Acar, Zhihao Jia

SC'24, November 17-22, 2024.

This paper presents techniques for theoretically and practically efficient and scalable Schrödinger-style quantum circuit simulation. Our approach partitions a quantum circuit into a hierarchy of subcircuits and simulates the subcircuits on multi-node GPUs, exploiting available data parallelism while minimizing communication costs. To minimize communication costs, we formulate an Integer Linear Program that rewards simulation of "nearby" gates on "nearby" GPUs. To maximize throughput, we use a dynamic programming algorithm to compute the subcircuit simulated by each kernel at a GPU. We realize these techniques in Atlas, a distributed, multi-GPU quantum circuit simulator. Our evaluation on a variety of quantum circuits shows that Atlas outperforms state-of-the-art GPU-based simulators by more than 2x on average and is able to run larger circuits via offloading to DRAM, outperforming other large-circuit simulators by two orders of magnitude.

continued from page 4

- Sanjith Athlur gave his speaking skills talk "Okapi: Decoupling Data Striping and Redundancy Grouping in Cluster File Systems."

April 2025

- Sheng Jiang gave a talk at NSDI '25 in Philadelphia, PA: "Building an Elastic Block Storage over EBOFs Using Shadow Views."
- Mengdi Wu presented "GraphPipe: Improving Performance and Scalability of DNN Training with Graph Pipeline Parallelism" ASPLOS'25 in Rotterdam, Netherlands.
- Wan Shen Lim proposed his PhD Thesis research on "Database Gyms: Towards Autonomous Database Tuning."
- Daiyaan Arfeen gave his speaking skills talk on "Nonuniform-Tensor-Parallelism: Mitigating GPU failure impact for Scaled-up LLM Training."
- Kaiyang Zhao gave his speaking skills talk on "Contiguitas: The Pursuit of Physical Memory Contiguity in Datacenters."
- The Student/Advisor team of Siyuan Chen & Phil Gibbons were CMU's Random Distance Run winners in that category.

March 2025

- Ziyue Qiu gave her speaking skills talk on "Moirai: Optimizing Placement of Data and Compute in Hybrid Clouds."
- Siddharth Jayashankar presented "Cinnamon: A Framework for Scale-out Encrypted AI" at the 30th Intl. Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), Rotterdam, The Netherlands.
- Yixuan Mei presented "Helix: Serving Large Language Models over Heterogeneous GPUs and Network via Max-Flow" at ASPLOS.

PDL students William Zhang and Wan Shen Lim, taking a break from retreat proceedings with a game of chess.

February 2025

- Zhihao Jia was named a 2025 Sloan Research Fellow.
- Siyuan Chen presented "Practical Offloading for Fine-Tuning LLM on Commodity GPU via Learned Sparse Projectors" at the 39th Annual AAAI Conference on Artificial Intelligence in Philadelphia, PA.
- Sam Arch gave his speaking skills talk on "The Key to Effective UDF Optimization: Before Inlining, First Perform Outlining."

January 2025

- Gauri Joshi was named a 2025 Goldsmith Lecturer.
- Hojin Park proposed his PhD research "Cost-Efficient Storage and Caching in Public Clouds."
- Yiwei Zhao presented "H4H: Hybrid Convolution-Transformer Architecture Search for NPU-CIM Heterogeneous Systems for AR/VR Applications" at the 30th Asia and South Pacific Design Automation Conference (ASPDAC '25), in Tokyo, Japan.
- Martin Prammer gave a talk on "Towards Functional Decomposition of Storage Formats" at the 15th Annual Conference on Innovative Data Systems Research (CIDR '25). in Amsterdam, The Netherlands.

December 2024

- Ziyue Qiu presented "Can Increasing the Hit Ratio Hurt Cache Throughput?" to the EAI International Conference on Performance Evaluation Methodologies and Tools in Milan, Italy, winning Best Paper Award!

Sara McAllister discusses her research with Jai Menon of Microsoft at the 2024 PDL Rtreat.

- William Zhang gave his speaking skills talk on "The Holon Approach to Holistic Database Optimization."

November 2024

- Sophia (Qingyang) Cao won the ACM Student Research Competition at SOSP 2024.
- Mingkuan Xu spoke on "Atlas: Hierarchical Partitioning for Quantum Circuit Simulation on GPUs" at the Int'l Conference for High Performance Computing, Networking, Storage and Analysis, SC'24, held in Atlanta, GA.
- Ankush Jain presented "CARP: Range Query-Optimized Indexing for Streaming Data" at SC'24.
- Nikhil Agarwal gave a talk on "The TYR Dataflow Architecture: Improving Locality by Taming Parallelism" at the 57th IEEE/ACM International Symposium on Microarchitecture (MICRO) held in Austin, TX.
- Timothy Kim presented "Morph: Efficient File-Lifetime Redundancy Management for Cluster File Systems" at the 30th Symposium on Operating Systems Principles (SOSP '24) in Austin, TX.
- Also presented at SOSP was "Reducing Cross-Cloud/Region Costs with the Auto-Configuring MACARON Cache" by Hojin Park.

October 2024

- 30th Parallel Data Lab Retreat and Workshop, Bedford Springs, PA.
- Wan Shen Lim gave his speaking skills talk on "Accelerating Machine Learning for Database Systems."

Olivia Hsu

Assistant Professor, ECE & CS

We are very pleased to welcome Olivia Hsu as a new faculty member of the PDL! Olivia will be joining Carnegie Mellon University as an Assistant Professor in Electrical and Computer Engineering (ECE) and, by courtesy, the Computer Science Department (CSD) starting Summer/Fall 2026.

Olivia is a PhD candidate in Computer Science at Stanford University, advised by Professor Kunle Olukotun and Professor Fredrik Kjolstad. Her work focuses on mapping and compiling sparse applications to domain-specific hardware, architectures, and accelerators. Her research interests also broadly include computer architecture, computer and programming systems, compilers, programming models and languages, and digital circuits/VLSI.

Prior to this, she graduated from the University of California, Berkeley in 2019 with a B.S.in Electrical Engineering and Computer Science (EECS). At Berkeley, Olivia was advised by Professor Vladimir Stojanovic and worked with Panagiotis Zarkos on novel applications of silicon-photonics.

Olivia and her co-authors recently received the Best Paper Award at the Deep Learning for Code (DL4C) Workshop at International Conference on Learning Representations (ICLR), in May 2025 for their work on "Adaptive Self-improvement LLM Agentic System for ML Library Development. The paper discusses the challenges of using LLMs to develop ML libraries using ASPLs and presents an adaptive self-improvement agentic system that enables LLMs to perform such complex reasoning under limited data by iteratively improving their capabilities through self-generated experience.

Olivia is a 2019 NSF Graduate Research Fellow, and her research won a distinguished paper award at PLDI 2023.

DEFENSES & PROPOSALS

continued from page 13

THESIS PROPOSAL: Database Gyms: Towards Autonomous Database Tuning

Wan Shen Lim
Tuesday, April 8, 2025

Database management systems (DBMSs) are the foundation of modern data-intensive applications. But as more features are developed to support new workloads, they become increasingly complex and difficult to configure. Decades of research on autonomous DBMS configuration have largely produced advisory tools that still rely on human expertise for their deployment into database tuning pipelines. Using these tools involves a multi-step process where a human operator (1) determines an optimization objective, (2) selects a suitable tool to improve the objective, (3) sets up and configures the DBMS to run a particular workload, (4) runs the workload to collect telemetry, (5) uses the collected telemetry to calibrate the tool, and (6) operates the tool to obtain recommendations, which the operator must then review and apply. Because of the ad-hoc nature of these pipelines, they require significant human effort to set up, extend, and deploy. Moreover, these tools are difficult to compose and swap.

This proposal presents the database gym, an integrated framework that systematizes and automates the DBMS configuration pipeline. The gym eliminates repetition in the setup and operation of such pipelines by providing a set of reusable, interoperable, and interchangeable components that simplify the development and integration of ML-driven DBMS configuration tools. It leverages its complete control over the tuning process to enable optimizations that require end-to-end knowledge. First, it eliminates step-level repetition by skipping over redundant computation during telemetry collection to reduce the latency of the tuning pipeline. Next, it eliminates pipeline-level repetition by reusing past experience to improve tool performance. For example, it adapts models of DBMS behavior to account for how operator semantics differ across DBMS versions.

We propose to extend our preliminary work by developing a tool for DBMS upgrades that uses version-aware models to predict performance improvements and regressions, addressing another database administration task with significant human involvement. Lastly, we will leverage recent advances in agentic artificial intelligence to orchestrate tools on behalf of a human operator. These efforts will transform the database gym from a platform for developing and deploying DBMS configuration tools into an autonomous database administrator for production environments.

Students preparing to present their research at the 2024 PDL Retreat.

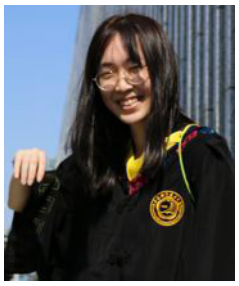
continued from page 9

butions in information theory and related areas and her ability to deliver an excellent lecture at one of the Society's Schools of Information Theory. Gauri will deliver the Goldsmith Lecture at one of the 2025 Information Theory Schools.

Gauri's research focuses on performance analysis and optimization of computing systems using a broad range of tools from probability, coding theory, and machine learning. She is an associate professor in the Electrical and Computer Engineering (ECE) department at CMU, with courtesy/affiliate appointments in the Machine Learning department (MLD) and the Robotics Institute (RI). Before joining CMU in Fall 2017, she was a Research Staff Member at the IBM T. J. Watson Research Center. She completed her Ph.D. from MIT EECS in 2016 and received my B.Tech and M. Tech in Electrical Engineering from IIT Bombay in 2010.

-- info from IEEE Information Theory Society and CMU ECE News

January 2025
Ziyue Qiu Wins Best Paper at ValueTools 2024!



Congratulations to Ziyue Qiu, and her co-authors Juncheng Yang and Mor Harchol-Balter who brought home the best paper award at the 17th EAI International Conference on Performance Evaluation Methodologies and Tools (ValueTools) held in Milan, Italy in December 2024. Their work asks "Can Increasing the Hit Ratio Hurt Cache Throughput?" and ultimately shows that increasing the hit ratio can actually hurt the throughput (and request latency) for many caching algorithms.

November 2024
Sophia (Qingyang) Cao Wins ACM Student Research Competition at SOSP 2024!

Congratulations to Sophia on winning the ACM Student Research contest at SOSP this year. To participate, for the first round, she had to submit an abstract of her research. In the second round, semi-finalists had to present a poster detailing their research. In the final round she gave a presentation. Evaluations are based on the presenter's knowledge of his/her research area, the contribution of the research, and the quality of the oral and visual presentation. Sophia's research on "Possum: A Tail of Dynamic Flash Capacity for Sustainability" investigates managing flash storage density for improved performance and device endurance.

Attendees of the 30th PDL Retreat held at the Bedford Springs Resort, in Bedford, PA, October 14-16, 2024.