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

CONTENTS

PDL CONSORTIUM MEMBERS

Amazon

Google

Hitachi, Ltd.

Honda

IBM Research

Intel Corporation

Meta

Microsoft

Oracle Corporation

Pure Storage

Salesforce

Samsung Semiconductor Inc.

Two Sigma

Western Digital

PDL PACKET

NEWSLETTER ON PDL ACTIVITIES AND EVENTS • FALL 2023

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

RECENT PUBLICATIONS

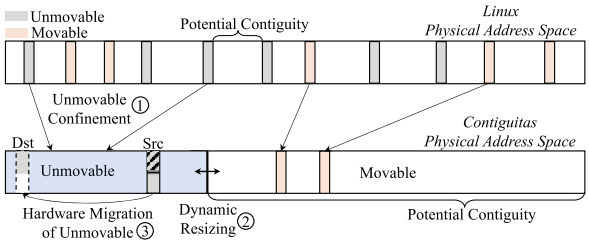
Contiguitas: The Pursuit of Physical Memory Contiguity in Datacenters

Kaiyang Zhao, Kaiwen Xue, Ziqi Wang, Dan Schatzberg, Leon Yang, Antonis Manousis, Johannes Weiner, Rik van Riel, Bikash Sharma, Chunqiang Tang, Dimitrios Skarlatos

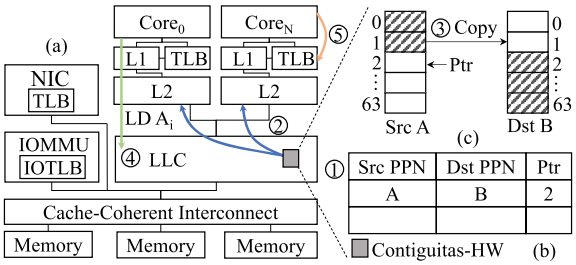
ISCA '23, June 17–21, 2023, Orlando, FL, USA.

BEST PAPER AWARD!

The unabating growth of the memory needs of emerging datacenter applications has exacerbated the scalability bottleneck of virtual memory. However, reducing the excessive overhead of address translation will remain onerous until the physical memory contiguity predicament gets resolved. To address this problem, this paper presents Contiguitas, a novel redesign of memory management in the operating system and hardware that provides ample physical memory contiguity. We identify that the primary cause of memory fragmentation in Meta's datacenters is unmovable allocations scattered across the address space that impede large contiguity from being formed. To provide ample physical memory contiguity by design, Contiguitas first separates regular movable allocations from unmovable ones by placing them into two different continuous regions in physical memory and dynamically adjusts the boundary of the two regions based on memory demand. Drastically reducing unmovable allocations is challenging because the majority of unmovable pages cannot be moved with software alone given that access to the page cannot be blocked for a migration to take place. Furthermore, page migration is expensive as it requires a long downtime to (a) perform TLB shootdowns that scale poorly



Contiguitas design overview.



Contiguitas hardware overview. (a) shows the hardware extension of Contiguitas-HW in the LLC. (b) shows the metadata table of Contiguitas-HW. (c) shows the page migration process followed by Contiguitas-HW.

continued on page 5

FROM THE DIRECTOR'S CHAIR

GREG GANGER

Yep, we are back! We enjoyed doing both a PDL Visit Day and a PDL Retreat in 2022, though corporate travel restrictions dampened attendance at each a bit... lingering pandemic restrictions for the May Visit Day and then economy-related ones for the Fall Retreat. Still, it was great to restart these important in-person interactions (and traditions :)), and both were successes. We are excitedly looking forward to full participation in this November's PDL Retreat!

It has been an great year for PDL, on the research and student accomplishment fronts, in addition to the increased interactions and a great PDL Talk Series over the summer. PDL researchers continue an amazing streak of research awards... though different conferences use different names for awards, they've brought home "Outstanding Paper", "Distinguished Paper", "Best Paper", etc. And the faculty are getting in on the prestigious awards action too, including Rashmi's Sloan Fellowship and Goldsmith Lecturer, Dave's Weiser Award for OS research, and several other awards and nominations. Behind the awards, of course, has been great progress on many fronts and new projects/collaborations initiated. A highlight has been strong continued interaction with PDL sponsors, including cool guest lectures to PDL's storage systems and cloud classes and co-authored papers (including some of the awards) with us in the context of research collaborations. I will not try to cover all of the PDL progress across storage systems, database systems, ML ↔ systems, and data processing infrastructure--specifics can be found throughout the newsletter—but I will briefly highlight a few things.

I'll start with storage and memory systems, where great research results, impact and awards have arisen from our collaborations with PDL sponsor companies. As one example, the Contiguitas memory management system (see paper abstract on front page of newsletter) achieves the physical memory contiguity needed to efficiently exploit large memory capacities and disaggregated memory... ISCA 2023 Best Paper, and the ideas are being incorporated for real deployment at a PDL sponsor company! We've also been excited to hear that at least one PDL sponsor is now integrating ideas from PDL's disk-adaptive redundancy research into production systems. Exciting new ideas and results have been achieved in effective high-throughput storage caches, ML-guided cache policies, redundancy and cache designs for emerging device interfaces (e.g., ZNS and FDP), flexible and "smart" storage systems, etc. And we look forward to sharing cool research in our new storage sustainability research thrust, this Fall and in the coming years. We thank our PDL sponsor companies who have enabled (and collaborated on) much of the research mentioned above by allowing us to experiment with real devices, workload traces, and failure logs!

Database systems research continues strong, and I think that the arrival of Professor Jignesh Patel makes CMU/PDL the top database systems group in the world (see a short introduction to him on page 24). His continuing work on maximizing database system performance on next-generation systems combine with Andy's work on database system automation, as well as

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
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
Mitch Franzos
Baljit Singh

VISITING RESEARCHERS & POST DOCS

Ellango Jothimurugesan
Xupeng Miao

GRADUATE STUDENTS

Nikhil Agarwal
Sam Arch
Daiyaan Arfeen
Sanjith Athlur
Nirav Atre
Mohammad Bakhshalipour
Jennifer Brana
Matt Butrovich
Neville Chima
Yae Jee Cho
Val Choung
Patrick Coppock
Travis Hance
Ankush Jain
Neharika Jali
Jekyeom Jeon
Hongyi Jin
Akshath Karanam
Abigale Kim
Timothy Kim
Ruihang Lai
Wan Shen Lim
Hao Yang Lu
Francisco Maturana
Sara McAllister
Nj Mukherjee
Gabriele Oliaro
Hojin Park
Pratyush Patel
Ziyue Qiu
Minya Rancic
Suhas J Subramanya
Sara Mahdizadeh Shahri
Shalini Shukla
Minh Truong
Jaylen Wang
Daniel Wong
Mike Xu
Mingkuan Xu
Jason Yang
Tianyu Zhang
Will Zhang
Kaiyang Zhao
Yiwei Zhao
Giulio Zhou

UNDERGRADUATE STUDENTS

Jonathan Chiu
Yucong Wang

FROM THE DIRECTOR'S CHAIR

collaborations storage and ML systems activities, to really make this the top group going forward. And it's not just on the research front... Andy's online lectures on databases have become an invaluable international resource for anyone seeking to understand what they are and how they work; I hear about their impact anytime I visit PDL companies. It has been really cool to watch the re-emergence of database strength!

We also continue our extensive work in large-scale data processing systems, including systems for ML and schedulers for analytics clusters. For example, our latest new scheduler for shared GPU clusters tackles the issue of GPU heterogeneity, co-adaptively selecting GPU type, GPU number and job configuration parameters (e.g., batch size) for each submitted DNN-training job to achieve unprecedented efficiency... it will soon be presented at SOSP 2023. Our new automated storage selector for cloud infrastructures to cost-effectively handle large-scale data processing workloads was recently published, and ongoing work explores automated data caching strategies when cross-cloud data access is involved... like in other areas, interaction and collaboration with PDL sponsor companies facing this emerging challenge has been invaluable. And, excitingly, broad aggregated activities are emerging among the research into simplifying, automating, and improving efficiency in big-data ML systems, both in collaboration with other CMU groups and with the Army's Pittsburgh-homed AI Integration Center (AI2C), as it looks to adopt and adapt key concepts into the new mechanisms they are building for AI development and use in the Army.

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.

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

The PDL was pleased to welcome Daniel Stodolsky, SambaNova (formerly with Google and Akamai) to the 2022 Retreat and delighted to introduce him as the newest member of the PDL Distinguished Alumni roll.

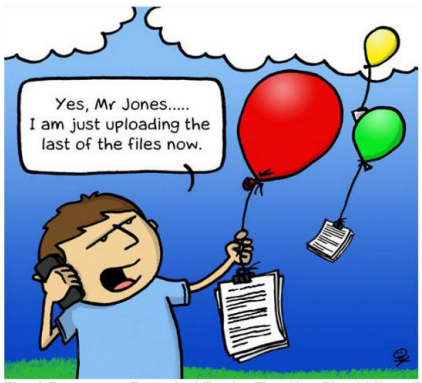
YEAR IN REVIEW

August 2023

- The paper "PIM-tree: A Skew-resistant Index for Processing-in-Memory" by Hongbo Kang, Yiwei Zhao, Guy E. Blelloch, Laxman Dhulipala, Yan Gu, Charles McGuffey, Phillip B. Gibbons appeared in the Proceedings of the VLDB Endowment and received Best Paper Runner Up!
- Juncheng Yang and Mohammad Bakshalipour Among the ML Commons Rising Stars!
- Travis Hance proposed his PhD research on "Verifying Concurrent Systems."
- Matthew Butrovich proposed his PhD research topic "On Embedding Database Management System Logic in Operating Systems via Restricted Programming Environments."
- Francisco José Maturana Sanguineti defended his PhD thesis on "Designing Storage Codes for Heterogeneity: Theory and Practice."

July 2023

- Mohammad Bakshalipour presented "Runahead A*: Speculative Parallelism for A* with Slow Expansions" at ICAPS 23 in Prague, Czech Republic.
- Rashmi Vinayak was named a Goldsmith Lecturer by IEEE.
- Brian Schwedock defended his PhD research on "Optimizing Data Movement Through Software Control of General-Purpose Hardware Caches."



Cloud Computing: Facts And Truths That You Should Know!!

June 2023

- Juncheng Yang presented "FIFO Can Be Better than LRU: The Power of Lazy Promotion and Quick Demotion" at HotOS '23 in Providence, RI.
- Kaiyang Zhao presented "Contiguitas: The Pursuit of Physical Memory Contiguity in Datacenters" at ISCA '23 in Orlando, FL. He and his co-authors received the Best Paper Award!
- Hojin Park presented "Mimir: Finding Cost-efficient Storage Configurations in the Public Cloud" at SYSTOR '23 in Haifa, Israel.
- Michael Kuchnik presented "Validating Large Language Models with ReLM" at the 6th MLSys Conference in Miami Beach, FL. He and his co-authors won the conference's Outstanding Paper Award!
- Sara Mahdizadeh Shahri was awarded a K & L Gates Presidential Fellowship.
- Juncheng Yang proposed his PhD research on "Designing Efficient and Scalable Cache Management Systems."
- Daniel Wong proposed "Machine Learning for Flash Caching in Bulk Storage Systems."

May 2023

- Ziyue Qiu presented "FrozenHot Cache: Rethinking Cache Management for Modern Hardware" at EuroSys 2023 in Rome, Italy.
- Akshitha Sriraman discussed Rethinking Datacenters in the CMU Engineering news.
- CyLab faculty and PDL alumni Michelle L. Mazurek, Lujo Bauer, Lorrie Faith Cranor, and Julio Lopez earned a "Test of Time" award at the IEEE Symposium on Security and Privacy for their 2012 paper "Guess Again (and Again and Again): Measuring Password Strength by Simulating Password-Cracking Algorithms."
- Michael Kuchnik defended his PhD research on "Beyond Model Efficiency: Data Optimizations for Machine Learning Systems."

April 2023

- Ellango Jothimurugesan presented "Federated Learning Under Distributed Concept Drift" at AISTATS '23 in Valencia, Spain.
- Lorrie Cranor was named a University Professor.
- Michael Rudow presented his PhD dissertation on "Efficient Loss Recovery for Videoconferencing via Streaming Codes and Machine Learning."

March 2023

- Huaicheng Li presented "Pond: CXL-Based Memory Pooling Systems for Cloud Platforms" at ASPLOS '23 in Vancouver, BC, Canada. The paper received the conference's Distinguished Paper Award!
- Thomas Kim presented "RAIZN: Redundant Array of Independent Zoned Namespaces" at ASPLOS '23 in Vancouver, BC, Canada.

February 2023

- Juncheng Yang presented "GL-Cache: Group-level Learning for Efficient and High-performance Caching" at FAST '23 in Santa Clara, CA.
- Rashmi Vinayak was awarded a 2023 Sloan Research Fellowship.
- Jack Kosiain defended his dissertation on "Practical Coding-Theoretic Tools for Machine Learning Systems and by Machine Learning Systems."
- Thomas Kim defended his research on "Design Principles for Replicated Storage Systems Built On Emerging Storage Technologies."
- Travis Hance gave his speaking skills talk on "Verifying Highly-Optimized Concurrent Systems with IronSync."

continued on page 23

continued from page 1

with the number of victim TLBs, and (b) copy the page. To this end, Contiguitas eliminates the primary source of unmovable allocations by introducing hardware extensions in the last-level cache to enable the transparent and efficient migration of unmovable pages even while the pages remain in use.

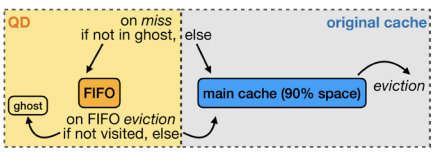
FIFO Can Be Better than LRU: The Power of Lazy Promotion and Quick Demotion

Juncheng Yang, Ziyue Qiu, Yazhuo Zhang, Yao Yue, K. V. Rashmi

HotOS '23, June 22–24, 2023, Providence, RI, USA.

LRU has been the basis of cache eviction algorithms for decades, with a plethora of innovations on improving LRU's miss ratio and throughput. While it is well-known that FIFO-based eviction algorithms provide significantly better throughput and scalability, they lag behind LRU on miss ratio, thus, cache efficiency.

We performed a large-scale simulation study using 5307 block and web cache workloads collected in the past two decades. We find that contrary to what common wisdom suggests, some FIFO-based algorithms, such as FIFO-Reinsertion (or CLOCK), are, in fact, more efficient (have a lower miss ratio) than LRU. Moreover, we find that quick demotion — evicting most new objects very quickly — is critical for cache efficiency. We show that when enhanced by quick demotion, not only can state-of-the-art algorithms be more efficient, a simple FIFO-based algorithm can outperform five complex state-of-the-art in terms of miss ratio.



An example of QD: add a probationary FIFO queue to an existing cache.

Validating Large Language Models with ReLM

Michael Kuchnik, Virginia Smith, George Amvrosiadis

6th MLSys Conference, Miami Beach, FL, USA, June 4-8, 2023.

OUTSTANDING PAPER AWARD AT MLSYS23!

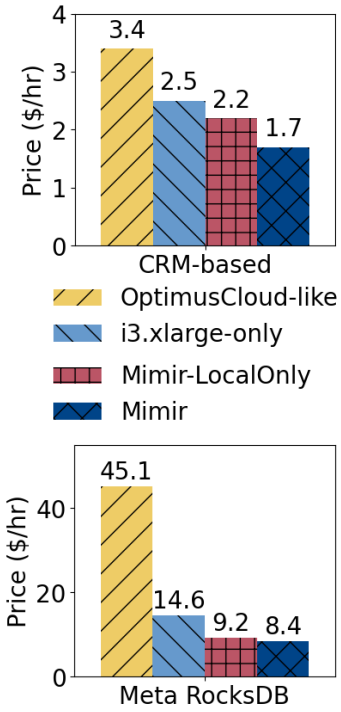
Although large language models (LLMs) have been touted for their ability to generate natural-sounding text, there are growing concerns around possible negative effects of LLMs such as data memorization, bias, and inappropriate language. Unfortunately, the complexity and generation capacities of LLMs make validating (and correcting) such concerns difficult. In this work, we introduce , a system for validating and querying LLMs using standard regular expressions. ReLM formalizes and enables a broad range of language model evaluations, reducing complex evaluation rules to simple regular expression queries. Our results exploring queries surrounding memorization, gender bias, toxicity, and language understanding show that ReLM achieves up to 15x higher system efficiency, 2:5x data efficiency, and increased statistical and prompt-tuning coverage compared to state-of-the-art ad-hoc queries. ReLM offers a competitive and general baseline for the increasingly important problem of LLM validation.

Mimir: Finding Cost-efficient Storage Configurations in the Public Cloud

Hojin Park, Gregory R. Ganger, George Amvrosiadis

SYSTOR '23: Proceedings of the 16th ACM International Conference on Systems and Storage, Haifa, Israel, June 5-7, 2023.

Public cloud providers offer a diverse collection of storage types and configurations with different costs and performance SLAs. As a consequence, it is difficult to select the most cost-efficient allocations for storage backends, while satisfying a given workload's performance requirements, when moving data-heavy applications to the cloud. We present Mimir, a tool for automatically finding a cost-efficient virtual storage cluster configuration for a customer's storage workload and performance requirements. Importantly, Mimir considers all block storage types and configurations, and even heterogeneous mixes of them. In our experiments, compared to state-of-the-art approaches that consider only one storage type, Mimir finds configurations that reduce cost by up to 81% for real-application-based key-value store workloads.



The cost-efficiency analysis of the optimization results of the two benchmarks, CRM and MR. Mimir finds the most cost-efficient VSC configuration compared to the other baselines.

continued on page 6

continued from page 5

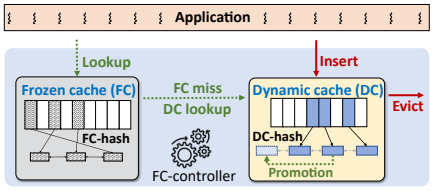
FrozenHot Cache: Rethinking Cache Management for Modern Hardware

Ziyue Qiu, Juncheng Yang, Juncheng Zhang, Cheng Li, Xiaosong Ma, Qi Chen, Mao Yang, Yinlong Xu

EuroSys 2023, Rome, Italy, May 8th-12th, 2023.

Caching is crucial for accelerating data access, employed as a ubiquitous design in modern systems at many parts of computer systems. With increasing core count, and shrinking latency gap between cache and modern storage devices, hit-path scalability becomes increasingly critical. However, existing production in-memory caches often use list-based management with promotion on each cache hit, which re- quires extensive locking and poses a significant overhead for scaling beyond a few cores. Moreover, existing techniques for improving scalability either (1) only focus on the indexing structure and do not improve cache management scalability, or (2) sacrifice efficiency or miss-path scalability.

Inspired by highly skewed data popularity and short-term hotspot stability in cache workloads, we propose Frozen- Hot, a generic approach to improve the scalability of list-based caches. FrozenHot partitions the cache space into two parts: a frozen cache and a dynamic cache. The frozen cache serves requests for hot objects with minimal latency by eliminating promotion and locking, while the latter leverages the existing cache design to achieve workload adaptivity. We built FrozenHot as a library that can be easily integrated into existing systems. We demonstrate its performance by enabling FrozenHot in two production systems: HHVM and RocksDB using under 100 lines of code. Evaluated using production traces from MSR and Twitter, FrozenHot improves the throughput of three baseline cache algorithms by up to 551%. Compared to stock RocksDB, FrozenHot-enhanced RocksDB shows a higher throughput on all YCSB workloads with up to 90% increase, as well as reduced tail latency.



Overview of FrozenHot. A lookup first goes to FC-hash (FC); if it is a cache miss, then look up in DC-hash (DC). Insertions and evictions occur only in DC.

Runahead A*: Speculative Parallelism for A* with Slow Expansions

Mohammad Bakhshalipour, Mohamad Qadri, Dominic Guri, Seyed Borna, Ehsani, Maxim Likhachev, Phillip B. Gibbons

ICAPS 2023, Prague, Czech Republic, July 8-13, 2023.

A* suffers from limited parallelism. The maximum level of traditional parallelism in A* is the same as the degree of the search graph nodes, which is too small in many applications. As such, A* cannot fully leverage the multithreading capabilities of modern processors. In this paper, we go beyond traditional parallelism and introduce speculative parallelism for A*. We observe that A*'s node expansions exhibit predictable patterns in applications like path planning. Based on this observation, we propose Runahead A* (RA*). When a node is being expanded, RA* predicts future likely-to-be-expanded nodes, performs their corresponding computation on separate threads, and memoizes the computation results. Later when a predicted node is selected for expansion, rather than performing its computation, the memoized results are used, saving significant time in slow-expansion applications. We study five applications of A*. We show that when its prediction accuracy is high, RA* offers significant speedup over vanilla A* for slow-expansion applications. With 16 threads, RA*'s speedup for such applications ranges from 3.1× to 14.1×. We also study and provide insight into when, why, and to what extent node expansions are predictable. We provide an implementation of RA* at: https://github.com/cmu-roboarch/runahead-astar/

GL-Cache: Group-level Learning for Efficient and High-performance Caching

Juncheng Yang, Ziming Mao, Yao Yue, K. V. Rashmi

21st USENIX Conference on File and Storage Technologies (FAST '23). Feb. 21–23, 2023, Santa Clara, CA.

Web applications rely heavily on software caches to achieve low-latency, high-throughput services. To adapt to changing workloads, three types of learned caches (learned evictions) have been designed in recent years: object-level learning, learning-from-distribution, and learning-from-simple-experts. However, we argue that the learning granularity in existing approaches is either too fine (object-level), incurring significant computation and storage overheads, or too coarse (workload or expert-level) to capture the differences between objects and leaves a considerable efficiency gap.

In this work, we propose a new approach for learning in caches ("group-level learning"), which clusters similar objects into groups and performs learning and eviction at the group level. Learning at the group level accumulates more signals for learning, leverages more features with adaptive weights, and amortizes overheads over objects, thereby achieving both high efficiency and high throughput.

We designed and implemented GL-Cache on an opensource production cache to demonstrate group-level learning. Evaluations on 118 production block I/O and CDN cache traces show that GL-Cache has a higher hit

continued on page 7

continued from page 6

ratio and higher throughput than state-of-the-art designs. Compared to LRB (object-level learning), GL-Cache improves throughput by 228× and hit ratio by 7% on average across cache sizes. For 10% of the traces (P90), GL-Cache provides a 25% hit ratio increase from LRB. Compared to the best of all learned caches, GL-Cache achieves a 64% higher throughput, a 3% higher hit ratio on average, and a 13% hit ratio increase at the P90.

Pond: CXL-Based Memory Pooling Systems for Cloud Platforms

Huaicheng Li, Daniel S. Berger, Lisa Hsu, Daniel Ernst, Pantea Zardoshti, Stanko Novakovic, Monish Shah, Samir Rajadnya, Scott Lee, Ishwar Agarwal, Mark D. Hill, Marcus Fontoura, Ricardo Bianchini

ASPLOS '23, March 25–29, 2023, Vancouver, BC, Canada.

DISTINGUISHED PAPER AWARD!

Public cloud providers seek to meet stringent performance requirements and low hardware cost. A key driver of performance and cost is main memory. Memory pooling promises to improve DRAM utilization and thereby reduce costs. However, pooling is challenging under cloud performance requirements. This paper proposes Pond, the first memory pooling system that both meets cloud performance goals and significantly reduces DRAM cost. Pond builds on the Compute Express Link (CXL) standard for load/store access to pool memory and two key insights. First, our analysis of cloud production traces shows that pooling across 8-16 sockets is enough to achieve most of the benefits. This enables a small-pool design with low access latency. Second, it is possible to create machine learning models that can accurately predict how much local and pool memory to allocate to a virtual machine (VM) to resemble same-NUMA-node memory performance. Our evaluation with 158 workloads shows that Pond reduces DRAM costs by 7% with performance within 1-5% of same-NUMA-node VM allocations.

Federated Learning Under Distributed Concept Drift

Ellango Jothimurugesan, Kevin Hsieh, Jianyu Wang, Gauri Joshi, Phillip B. Gibbons

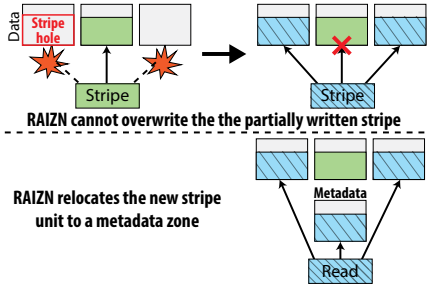
International Conference on Artificial Intelligence and Statistics (AISTATS), Apr 2023. In preprint arXiv:2206.00799v1.

Federated Learning (FL) under distributed concept drift is a largely unexplored area. Although concept drift is itself a well-studied phenomenon, it poses particular challenges for FL, because drifts arise staggered in time and space (across clients). Our work is the first to explicitly study data heterogeneity in both dimensions. We first demonstrate that prior solutions to drift adaptation, with their single global model, are ill-suited to staggered drifts, necessitating multi-model solutions. We identify the problem of drift adaptation as a time-varying clustering problem, and we propose two new clustering algorithms for reacting to drifts based on local drift detection and hierarchical clustering. Empirical evaluation shows that our solutions achieve significantly higher accuracy than existing baselines, and are comparable to an idealized algorithm with oracle knowledge of the ground-truth clustering of clients to concepts at each time step.

RAIZN: Redundant Array of Independent Zoned Namespaces

Thomas Kim, Jekyeom Jeon, Nikhil Arora, Huaicheng Li, Michael Kaminsky, David G. Andersen, Gregory R. Ganger, George Amvrosiadis, Matias Bjørling

ASPLOS '23, March 25–29, 2023, Vancouver, BC, Canada. Supoercedes Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-22-101, January 2022.



Only a subset of the stripe units are persisted before power is lost, resulting in a hole in the logical address space. The next stripe cannot be written at the correct address due to the persisted stripe unit.

Zoned Namespace (ZNS) SSDs are the most recent evolution of host-managed flash-based storage, enabling improved performance at a lower cost-per-byte compared to traditional block interface SSDs. To date, there is no support for arranging these new devices in redundant arrays (RAID), which may limit their deployment in environments where this is the favored mechanism for increasing reliability and throughput. This paper identifies key challenges in the design of a RAID-like mechanism for ZNS SSDs, such as the requirement to manage metadata updates and persist partial stripe writes in the absence of overwrite semantics in the device's interface. We present the design, implementation, and evaluation of RAIZN, a logical volume manager that exposes a ZNS interface and stripes data and parity across ZNS SSDs.

Experiments show that RAIZN provides full expected performance from the aggregate device set, successfully addressing the key challenges from the ZNS interface. RAIZN achieves throughput and latency comparable to the equivalent Linux software RAID implementation running on conventional SSDs that use the same hardware platform, and then RAIZN exceeds its performance once device-level garbage collection inhibits the conventional SSDs. Importantly, RAIZN retains

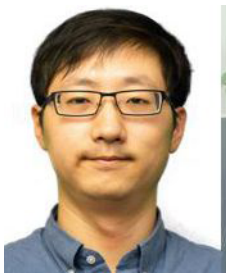
continued on page 18

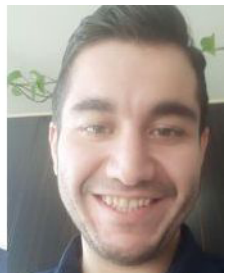
August 2023
PIM-Tree Awarded Best Research Paper Runner-up at VLDB!

Congratulations to Yiwei Zhao, Charles McGuffey, Phil Gibbons and their co-authors on receiving the best paper runner up award at VLDB! Their paper "PIM-tree: A Skew-resistant Index for Processing-in-Memory" discusses solutions to mitigate the bottleneck induced by the memory latency/bandwidth wall in memory indexing by enabling low-latency memory access whose aggregate memory bandwidth scales with the number of PIM nodes.

August 2023
Juncheng Yang and Mohammad Bakshalipour Among the ML Commons Rising Stars!

Congratulations to Juncheng and Mohammad, who have been listed among a stellar group of 35 current and recently graduated PhD students in the inaugural ML Commons Rising Stars cohort! These students, who work in the intersection of Machine Learning and Systems research have been chosen from among over 100 applicants globally. The students will be able to present their research to industry and academic experts and establish collaboration among their fellows, promoting diversity in the ML/systems research community. Juncheng's and Mohammad's first involvement with the group will be a workshop at Google this month where the students will showcase their work and meet their fellows. Juncheng is advised by Rashmi Vinayak and will be discussing his work on "GL-Cache: Group-level Learning for Efficient and High-performance Caching" at the workshop. Mohammad is advised by Phil Gibbons and is presenting his research on "Bridging Robotics and Architecture." Read more about the ML Commons Rising Star program at https://mlcommons.org/en/news/rising-stars-2023/.



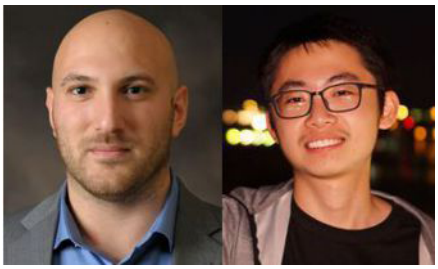


July 2023
Rashmi Vinayak Named Goldsmith Lecturer

Congratulations to Rashmi, a professor of computer science and faculty member of the PDL, who was named the 2023 Goldsmith Lecturer by the IEEE Information Theory Society for her professional and technical achievements in data systems. The Goldsmith Lecturer Program highlights the technical achievements of early-career researchers and helps build their professional career and recognition. The program contributes to the public visibility of the chosen lecturer and seeks to increase the diversity of IEEE. Vinayak, an assistant professor in the Computer Science Department, studies information and coding theory, computer and networked systems, and where these fields intersect. She currently focuses on robustness and resource efficiency in data systems, including storage and caching systems, systems for machine learning, and live-streaming communication. Vinayak will deliver a lecture at one of the IEEE Information Theory Society's Schools of Information Theory. These short workshops introduce students to new research frontiers in information theory. More information is available in the IEEE Information Theory Society's newsletter.

-- info from the Piper, by Aaron Aupperlee, July 25, 2023 https://www.cs.cmu.edu/news/2023/vinayak_goldsmith_lecturer

June 2023
Best Paper Award at ISCA '23!



Congratulations to Kaiyang Zhao, Dimitrios Skarlatos, and alum Ziqi Wang on winning the Best Paper Award at ISCA '23 this June in Orlando, Florida! Their paper "Contiguitas: The Pursuit of Physical Memory Contiguity in Datacenters" explores Contiguitas, a novel redesign of memory management in the operating system and hardware that provides ample physical memory contiguity. To boost physical memory contiguity Contiguitas first separates regular movable allocations from unmovable ones by placing them into two different continuous regions in physical memory and dynamically adjusts the boundary of the two regions based on memory demand.

June 2023
Sara Mahdizadeh Shahri Awarded K & L Gates Presidential Fellowship

Congratulations to Sara Mahdizadeh Shahri, a PDL and CMU doctoral student on being selected to receive a 2023-24 K & L Gates Presidential Fellowship. Sara is a Ph.D. student in electrical and computer engineer-

ing and her research, which bridges computer architecture and software systems, aims to introduce equity in the context of data center systems. The fellowship provides financial support to four doctoral students, enabling them to further their studies on ethical and policy issues surrounding artificial intelligence.

Sara notes: "Existing data center systems are fundamentally inequitable in several ways. So, it is critical to systematically analyze when such prioritization-based solutions can cause biases, compromising equity. I am excited about my research as I believe it can provide a basis of such analysis in the context of data centers, and consequently enable a data center design paradigm that prevents discrimination against users from under-resourced communities."

-- info from Carnegie Mellon University News, June 16, 2023 by Kelly Saavedra.

May 2023
Kuchnik, Smith, and Amvrosiadis Receive Outstanding Paper at MLSys 2023!

Congratulations to Michael, Virginia and George on receiving the award for Outstanding Paper at MLSys, held this year in Miami Beach, FL, for their paper "Validating Large Language Models with ReLM." The paper introduces ReLM, a system for validating and querying LLMs using standard regular expressions. ReLM formalizes and enables a broad range of language model evaluations, reducing complex evaluation rules to simple regular expression queries.

May 2023
Akshitha Sriraman is Rethinking Datacenters

Today's technology would not exist without datacenters. Six in 10 people use modern web services such as social media, web search, video streaming, online banking, and online healthcare that require datacenters that scale to hundreds of thousands of high-end computers or servers. Akshitha Sriraman, assistant professor of electrical and computer engineering, is rethinking datacenter computing across hardware and software systems to enable efficient, sustainable, and equitable large-scale web systems.

Current datacenters house thousands of servers that hold information and route signals for billions of users. With the surge of devices and users coming online daily and the growing amount of data being exchanged, the demand for faster, more efficient cloud services is drastically increasing. With this increase in demand, the logical answer is to keep building larger and more datacenters. However, this is not sustainable in the long-term. Not only are these colossal datacenters extremely expensive to build and maintain, but their carbon footprint is massive. Akshitha and her colleagues are "redesigning datacenters from first principles, thinking about what these servers should look like at the hardware level in a way that they can be cost- and energy-efficient. To enable sustainable datacenters, we must carbon-efficiently architect and manufacture hardware and make the most out of existing hardware. Datacenters must adopt the mindset of reducing, reusing, and recycling hardware."

-- from Carnegie Mellon Engineering News, May 16, 2023 - Krista Burns

May 2023
CyLab Faculty and PDL Alumni Earn 'Test of Time' Awards at IEEE Symposium on Security and Privacy

At its 44th Symposium on Security and Privacy, the IEEE recognized a paper by CMU/CyLab/PDL faculty and alumni with a 'Test of Time award' for the paper "Guess Again (and Again and Again): Measuring Password Strength by Simulating Password-Cracking Algorithms" (2012). The 'Test of Time' award recognizes published papers previously presented at the annual symposium that have had a broad and lasting impact on both research and practice in computer security and privacy.

In the paper, researchers Patrick Gage Kelley, Saranga Komanduri, Michelle L. Mazurek, Richard Shay, Timothy Vidas, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Julio Lopez analyzed 12,000 passwords collected under seven composition policies and developed an efficient distributed method for calculating how effectively several heuristic password-guessing algorithms guessed passwords. The study advanced the understanding of both password-composition policies and metrics for quantifying password security.

-- from CyLab News, May 23, 2023 by Ryan Noone

April 2023
Lorrie Cranor Named University Professor

School of Computer Science faculty member Lorrie Faith Cranor has

continued on page 10

continued from page 9

been elevated to the rank of University Professor, the highest distinction a faculty member can receive at Carnegie Mellon University. Distinguished by international recognition and for their contributions to education, artistic creativity and/or research, University Professors exemplify a high level of achievement and commitment to the university and the broader academic communities.

Cranor is the director and Bosch Distinguished Professor in Security and Privacy Technologies of CyLab and the FORE Systems Professor of Computer Science and of Engineering and Public Policy. She co-founded and co-directs the world's first privacy engineering master's program and was a founding co-director of the Collaboratory Against Hate: Research and Action Center. Her research focuses on usable privacy and security with contributions in a variety of areas, including antiphishing technologies, usable and secure password policies, privacy "nutrition" labels, and tools to make it easier for people to protect their privacy and security. She founded the Symposium on Usable Privacy and Security and co-founded the Conference on Privacy Engineering Practice and Respect.

-- from School of Computer Science News, April 28, 2023 - Christa Cardone and Kristen Bayley

March 2023
Distinguished Paper Award at ASPLOS '23!

Congratulations to PDL Alums Huaicheng Li and Daniel Berger on being awarded the Distinguished Paper Award for their research on "Pond: CXL-Based Memory Pooling Systems for Cloud Platforms" at ASPLOS '23, held in March in Vancouver, BC, Canada. The paper discusses memory pooling to improve DRAM utilization and thereby reduce costs. Pond is the first memory pooling system that both meets cloud performance goals and significantly reduces DRAM cost.

February 2023
Rashmi Vinayak Earns 2023 Sloan Research Fellowship

Congratulations to Rashmi Vinayak, who has earned a 2023 Sloan Research Fellowship in recognition of her research accomplishments. She is among 125 early career researchers from 54 institutions to receive the award from the Alfred P. Sloan Foundation.

"Sloan Research Fellows are shining examples of innovative and impactful research," said Adam F. Falk, president of the Alfred P. Sloan Foundation. "We are thrilled to support their groundbreaking work and we look forward to following their continued success."

Rashmi Vinayak is an assistant professor in the Computer Science Department (CSD). Her current work focuses on increasing reliability and efficiency in robust data systems by combining information and coding theory. Phillip Gibbons, a professor in CSD and the Department of Electrical and Computer Engineering, nominated Vinayak for the fellowship. Vinayak will receive a two-year, $75,000 fellowship that can be used flexibly to advance her research. To read more about the 2023 Sloan Research Fellows, visit the foundation's website.

-- from SCS News, Feb. 17, 2023. Aaron Aupperlee

November 2022
CMU Professors Awarded NSF Future of Work Grant

An NSF Future of Work grant is designed to increase opportunities for U.S. workers and generate positive societal and economic impacts both locally and nationally, will fund a

project that will investigate how AI-augmented learning can help accelerate student progress in community college information technology (IT) courses. This will in turn provide pathways to family-sustaining careers by making more comprehensive technical education accessible to students. The project team of "A New Bridge to the Digital Economy: Integrated AI-Augmented Learning and Collaboration" includes SCS faculty members Carolyn Rose, Majd Sakr, Lauren Herckis and Bruce McLaren. The SCS team, working in conjunction with The Community College of Allegheny County, will use not only decades of research into computer- and technology-assisted learning science but also the latest developments in the field to enhance lessons.

-- from School of Computer Science News Nov 21, 2022 - Aaron Aupperlee and Heinz College

November 2022
David Andersen wins ACM SIGOPS Mark Weiser Award

Congratulations to Dave on receiving the 2022 Mark Weiser Award for his ongoing insightful and innovative work in systems design! The Mark Weiser Award was created in 2001 by ACM SIGOPS to be given to an individual who has demonstrated creativity and innovation in operating systems research. The award is named in honor of Mark Weiser, a computing visionary recognized for his research accomplishments during his career at Xerox PARC.

DISSERTATION ABSTRACT: Designing Storage Codes for Heterogeneity: Theory and Practice

Francisco José Maturana Sanguineti
Carnegie Mellon University, SCS

PhD Defense — August 18, 2023

Distributed storage systems support many essential applications, and thus need to be highly reliable. To achieve this goal at a low cost, most systems use erasure codes. The parameters of the erasure code (which affect the cost and level of protection) are set based on the expected operating conditions. However, conditions vary significantly across time and across the system. For example, failure rates, workloads, and density of devices can change with time and in different locations. Many existing systems fail to accommodate these variations, or do so in inefficient ways. My thesis focuses on making distributed storage systems more robust and efficient by enabling them to automatically adapt to these variations. To make progress towards this goal, I develop and use tools from both Coding Theory and Computer Systems research.

The first part focuses on variations in the system across time. Our main contribution here is the "convertible codes" framework, designed to study and construct erasure codes that can efficiently change their parameters over time. We propose the framework, derive the fundamental limits of this problem and design optimal codes. Additionally, we propose two distributed storage system designs, which automatically decide when and how to convert between codes.

PDL faculty member Zhihao Jia discusses Machine Learning at the 2022 PDL Retreat.

The second part focuses on heterogeneity across the system. Specifically, we consider a geo-distributed storage system, where the density of nodes and latencies between nodes vary significantly, and the cost of sending data across the wide-area network (WAN) is crucial. Our main contribution is a new class of codes that optimizes both the storage overhead and WAN bandwidth given the parameters of the system. We additionally propose a new strongly-consistent geo-distributed storage system that jointly optimizes its consensus protocol and erasure code.

THESIS PROPOSAL: On Embedding Database Management System Logic in Operating Systems via Restricted Programming Environments

Matthew Butrovich, CSD
August 23, 2023

The rise in computer storage and network performance means that disk I/O and network communication are often no longer bottlenecks in database management systems (DBMSs). Instead, the overheads associated with operating system (OS) services (e.g., system calls, thread scheduling, and data movement from kernel- space) limit query processing responsiveness. To avoid these overheads, user-space applications prioritizing performance over simplicity can elide these software layers with a kernel-bypass design. However, extracting benefits from kernel-bypass frameworks is challenging, and the libraries are incompatible with standard deployment and debugging tools. For these reasons, few DBMSs employ a kernel-bypass approach.

This proposal presents user-bypass — an approach to designing DBMS software that complements OS extensibility. With user-bypass, developers write safe, event-driven programs to push DBMS logic into the kernel's stack and avoid user-space overheads. We demonstrate user-bypass in the context of two different applications for DBMSs. First, we present TScout, a framework for training data collection in self-driving DBMSs. user-bypass accelerates TScout's metrics collection by not requiring multiple round trips to kernel-space to retrieve performance counters and other resource counters. Then, we present Tigger, a PostgreSQL- compatible DBMS proxy similar to RDS Proxy, PgBouncer, and ProxySQL. Through user-bypass, Tigger supports features like connection pooling, transaction multiplexing, and workload mirroring without user-space interaction.

We propose to extend our preliminary work by building a DBMS that executes queries with user-bypass. We will investigate the opportunities and limitations of placing core DBMS components in kernel-space. First, we will design and evaluate a storage manager that stores database entries in kernel-resident data structures, eliding the need to go to user-space to execute queries. Second, we will investigate concurrency control methods to enforce ACID properties within the constraints of kernel execution (e.g., no waiting). Lastly, we will create a framework for logging and checkpointing the database contents stored in kernel-space. This effort will fulfill the goal of crash recovery and inform the design of further uses for logging, like replication.

THESIS PROPOSAL: Verifying Concurrent Systems

Travis Hance, CSD
August 11, 2023

Concurrent software is notoriously difficult to write correctly, so to increase confidence in it, it is often desirable to apply formal verification techniques. One technique that is especially promising for verifying concurrent software

continued on page 12

continued from page 11

is concurrent separation logic (CSL), which uses reasoning principles based on resource ownership. However, even with CSL, verifying complex systems at scale (e.g., those with 1000s of lines of code) remains challenging. The reasons it remains challenging include,

1. The manual proof effort required by many existing CSL frameworks.

2. The inherent complexity of the target systems. Sophisticated systems may have custom, low-level synchronization logic, which may be deeply intertwined with domain logic, in the interest of performance.

We posit that a promising way to overcome (1) is, rather than using CSL directly, to use an ownership type system such as Rust's, taking advantage of its sophisticated but efficient type-checking algorithms. To demonstrate this, we develop a full methodology (from theory to implementation) based around this core idea. We show that it has numerous advantages, and in particular, it is rich enough to support the verification of inherently complex systems as in (2).

DISSERTATION ABSTRACT: Optimizing Data Movement Through Software Control of General-Purpose Hardware Caches

Brian Schwedock
Carnegie Mellon University, SCS

PhD Defense — July 3, 2023

Computer systems are increasingly burdened by the rising cost of data movement. Moving data across chip in a modern processor consumes orders-of-magnitude more energy than performing a floating-point operation on the data. On-chip caches also constitute more than half of a chip's area. The severity of these problems will continue to grow alongside rising core counts and data-processing requirements.

The underlying issue is that chip multiprocessors (CMPs) provide a compute-centric programming interface where software views the entire memory hierarchy as a black box. Software issues loads and stores, and it's entirely up to hardware to manage all data movement between the core and main memory. Although this interface simplifies software, hardware is forced to resort to overly general application-agnostic optimizations.

To overcome the limitations of compute-centric CMPs, prior work has proposed specialized hierarchies which add custom logic to the hierarchy to enable novel data-movement-reducing features. Specialized hierarchies reduce data movement by either moving data closer to compute (data placement) or moving compute closer to data (near-data computing). These data-centric systems often provide significant benefits by customizing data movement within the memory hierarchy to specific applications. Unfortunately, adding custom logic to CMPs for every possible application is not a scalable solution.

The goal of this thesis is making specialized hierarchies practical by letting software customize data movement, eliminating the need for application-specific custom hardware. Proposed systems address both data placement and near-data computing (NDC). First, Jumanji shows how software-controlled data placement for distributed last-level caches enables optimizing for a variety of performance objectives on a single processor. Specifically, Jumanji targets a datacenter environment where co-running applications either care about tail latency or throughput, and all applications care about security. Second, täko generalizes the broad category of data-triggered NDC to let software observe and manipulate data as it traverses the cache hierarchy. Finally, Leviathan unifies multiple types of NDC paradigms under a single architecture and programming interface to provide a truly practical NDC system. Together, these contributions demonstrate the feasibility of programmable data movement in general-purpose processors.

THESIS PROPOSAL: Machine Learning for Flash Caching in Bulk Storage Systems

Daniel Lin-Kit Wong, CSD
June 30, 2023

Flash caches are used to reduce peak backend load for throughput-constrained data center services, reducing the total number of backend servers required. Bulk storage systems are a large-scale example, backed by high-capacity but low-throughput hard disks, and using flash caches to provide a more cost-effective storage layer underlying everything from blobstores to data warehouses.

However, flash caches must address the limited write endurance of flash by limiting the long-term average flash write rate to avoid premature wearout. To do so, most flash caches must use admission policies to filter cache insertions and attempt to maximize the workload-reduction value of each flash write.

We present the Baleen flash cache, which uses coordinated ML admission and prefetching to reduce peak backend load. After learning painful lessons with early ML policy attempts, we exploit a new cache residency model (which we call episodes) to guide models used and model training, and focus on optimizing for an end-to-end system metric (disk-head time) balancing IOPS and bandwidth rather than hit rate. Evaluation using Meta traces from seven storage clusters shows that Baleen reduces Peak Disk-head Time (and backend capacity required) by 11.8% over state-of-the-art policies.

In proposed work, we apply ML to item placement to improve eviction and optimize the use of DRAM in hybrid caches. To improve eviction, we will reduce cache dead time by classifying items by their eviction age and placing them into different eviction queues. We will use ML to select a few items for placement in DRAM that are most helpful for reducing flash writes, instead of letting every item pass through DRAM.

continued on page 13

continued from page 12

Workloads change over time, requiring the cache to adapt to maintain performance. We propose strategies to actively target peak load reduction and to mitigate workload drift. We plan to augment admission to prioritize items based on their benefit during peak load and to adapt to load levels.

THESIS PROPOSAL: Designing Efficient and Scalable Cache Management Systems

Juncheng Yang, CSD
June 8, 2023

Software caches, e.g., key-value caches and object caches, are widely deployed in today's system stacks to support the ever-growing digital society. DRAM scaling has slowed down, with the price per Gigabyte staying stable over the past few years. In the meantime, computing devices, e.g., CPUs, continue to scale horizontally, with server CPUs reaching over a hundred cores per socket. The increasing gap between DRAM capacity and computation power emphasizes the importance of cache efficiency. Meanwhile, the increasing number of cores per CPU makes thread scalability a critical requirement for designing software caches.

This thesis explores different approaches to improving the efficiency and scalability of software caches. We first perform a large-scale cache workload analysis to advance the understanding of in-memory key-value caches. In addition to confirming previous observations, we identify several new patterns in cache workloads, e.g., extensive use of time-to-live (TTL), and the prevalence of write-heavy workloads.

Inspired by the workload analysis, we design Segcache, a TTL-indexed segment-structured cache. Segcache (1) removes expired objects proactively; (2) minimizes per-object metadata via approximation and sharing; and (3) reduces critical section size using macro management. These properties enable more objects to be stored with a close-to-linear thread scalability.

We also design C2DN, a fault-tolerant CDN cache cluster, which leverages erasure coding for low-overhead redundancy. Naive use of erasure coding in a CDN cache cluster is ineffective because of independent evictions. We introduce parity rebalance to balance the write load among caches so that C2DN provides efficient and effective fault tolerance.

Existing learned caches often incur a significant overhead. We propose a new approach to low-overhead learned cache called GL-Cache, which learns the usefulness of object groups rather than individual objects. Group-level learning amortizes the learning overheads and accumulates more information for learning. While GL-Cache improves efficiency, using machine learning as a black box increases complexity and debugging difficulty. I plan to explore the design of (1) an interpretable learned cache and (2) simple yet efficient cache eviction algorithms by leveraging observations in our recent workload study.

DISSERTATION ABSTRACT: Beyond Model Efficiency: Data Optimizations for Machine Learning Systems

Michael Kuchnik
Carnegie Mellon University, SCS

PhD Defense — May 1, 2023

The field of machine learning, particularly deep learning, has witnessed tremendous recent advances due to improvements in algorithms, compute, and datasets. Systems built to support deep learning have primarily targeted computations used to produce the learned model.

This thesis proposes to instead focus on the role of data in both training and validation. For the first part of the thesis, we focus on training data, demonstrating that the data pipeline responsible for training data is a prime target for performance considerations. To aid in addressing performance issues, we introduce a form of data subsampling in the space of data transformations, a reduced fidelity I/O format, and a system for automatically tuning data pipeline performance knobs.

Jekyeom Jeon describing his research to PDL Alum Qing Zheng (LANL) at one of the retreat poster sessions.

In the second part of the thesis, motivated by the trend toward increasingly large and expressive models, we turn to the validation setting, developing a system for automatically querying and validating a large language model's behavior with off-the-shelf regular expressions. We conclude with future work in the space of data systems for machine learning.

DISSERTATION ABSTRACT: Efficient Loss Recovery for Videoconferencing via Streaming Codes and Machine Learning

Michael Harrison Rudow
Carnegie Mellon University, SCS

PhD Defense — April 28, 2023

Packet loss degrades the quality of experience (QoE) of live communication. However, conventional methods for loss recovery are inefficient at protecting against the bursty losses that arise in practice. Instead, a new class of theoretical erasure codes, called "streaming codes," efficiently communicates a sequence of frames over a bursty packet loss channel. Existing streaming codes apply when all frames are of the same

continued on page 14

continued from page 13

fixed size, but many applications like videoconferencing involve sending frames of varying sizes. This thesis presents a generalized model for streaming codes that incorporates frames of variable sizes, studies the fundamental limits on the optimal rate for the new model, designs new high-rate streaming codes using machine learning and coding theory, and integrates streaming codes into a videoconferencing application to assess their positive impact on the QoE.

We start by examining the fundamental limits on the "offline" communication rate, wherein the sizes of all future frames are known. We show that the variability in the sizes of frames (a) induces a new trade-off between the rate and the decoding delay under lossless transmission and (b) impacts the optimal rate of transmission. We then design rate-optimal streaming codes for the practically relevant "online" setting-- without access to the sizes of the future frames—when each frame is sent immediately. We then use a learning-augmented algorithm to spread frame symbols over one extra frame to design approximately rate-optimal streaming codes.

However, many real-world applications experience what we dub "partial burst" losses of only some packets per frame, unlike the existing model, which assumes all or no packets are lost for each frame. To address this gap, we introduce a new streaming-codes-based approach to videoconferencing called Tambur. When assessed over emulated networks, Tambur improves several key metrics of QoE compared to conventional methods (e.g., it reduces the frequency of freezes by 26%). We then extend the theoretical streaming model to accommodate partial bursts and design an online approximately rate-optimal streaming code. The code combines (a) a building block construction given any choice of how much parity to allocate per frame with (b) a learning-augmented algorithm to allocate parity per frame.

DISSERTATION ABSTRACT: Design Principles for Replicated Storage Systems Built On Emerging Storage Technologies

Thomas Kim
Carnegie Mellon University, SCS

PhD Defense — March 15, 2023

With the slowing down of Moore's law, persistent storage hardware has continued to scale at the cost of exposing hardware-level write idiosyncrasies to the software. Thus, a key challenge for systems developers is to reason about and design around these idiosyncrasies to create replicated storage systems that can effectively leverage these new technologies. Two examples of such new and emerging persistent storage technologies are Intel Optane non-volatile main memory and Zoned Namespace (ZNS) solid-state drives.

Through our experiences and setbacks when designing, implementing, and evaluating systems based on Optane and ZNS, we propose three guidelines to assist developers in designing storage systems on new and emerging persistent storage technologies: (1) systems, even those expected to serve read-heavy workloads, should be optimized for write performance, (2) set and fulfill performance, durability, and fault tolerance guarantees, but do not exceed them as that may result in excessive write overheads, and (3) systems can overcome limitations of write-constrained persistent hardware by optimizing data placement and internal data flows based on assumptions about temporal and spatial locality of the expected client workload.

The first system we present is CAND-Store, a highly-available, cost-effective, replicated key-value store that uses Intel Optane for primary storage, and solves the challenge of bottlenecked data ingestion during primary failure recovery through a novel online workload-guided recovery protocol. The second system we present is RAIZN, which is a system that provides RAID-like striping and redundancy for arrays of ZNS SSDs, and solves the various challenges that arise as a result of the lack of overwrite semantics in ZNS. We describe how the above guidelines arose from the setbacks and successes during the development of the above two systems, then apply these guidelines to extend the functionality of RAIZN to create RAIZN+.

The final part of this thesis details exactly how we applied these principles to RAIZN+, and demonstrates the efficacy of these design guidelines through the evaluation of RAIZN+, which is able to achieve near-zero write amplification when serving RocksDB workloads.

DISSERTATION ABSTRACT: Practical Coding-Theoretic Tools for Machine Learning Systems and by Machine Learning Systems

Jack Kosaian
Carnegie Mellon University, SCS

PhD Defense — March 14, 2023

The use of machine learning (ML) in many domains has led to the development of many ML systems for deploying and training ML models. Beyond achieving high accuracy, ML systems must also use computing infrastructure efficiently and tolerate unreliable infrastructure.

Coding-theoretic tools enable many systems to operate both reliably and efficiently. These tools are used in production storage and communication systems, and there is growing interest in their use for distributed computing.

This thesis explores the interplay between ML systems and practical applications of coding-theoretic tools. Specifically, we show how ML systems can be made more reliable and efficient via novel uses of coding-theoretic tools, and how coding-theoretic tools can be expanded in reach and be made more efficient through techniques from ML and systems. We illustrate this via multiple thrusts:

continued on page 15

continued from page 14

(1) Properties unique to ML systems can be exploited to efficiently integrate coding-theoretic tools into ML systems. First, we reduce the execution-time overhead of fault-tolerant inference on GPUs by exploiting trends in neural network design and GPU hardware. Second, we show how coding-theoretic tools can be coupled with the unique properties of recommendation models to enable low-overhead fault tolerance in training.

(2) Co-designing coding-theoretic tools with ML systems offers new opportunities to extend the reach of these tools. Specifically, we enable resource-efficient fault tolerance in distributed prediction serving systems by using ML to overcome a key barrier in prior coding-theoretic tools.

(3) Ideas inspired by coding theory can be used to improve the performance of ML systems even when reliability is not a concern. We increase the throughput and GPU utilization of specialized convolutional neural network inference by inferring over images in a coding-theory-inspired manner and making small modifications to the model.

(4) Coding-theoretic tools can operate at higher throughput with little developer effort via advancements in ML systems. We exploit similarities between operations in erasure codes, a popular coding-theoretic tool, and those in ML libraries to enable erasure codes to be easily represented via ML libraries, and thus allow erasure-coding libraries to immediately adopt the many optimizations that have gone into ML libraries.

PDL grauate students Wan Shen Lim, Sam Arch, Mike Xu and Deepayan Patra at the PDL 2022 Retreat Industry Poster Session.

DISSERTATION ABSTRACT: Large-scale Machine Learning over Streaming Data

Ellango Jothimurugesan
Carnegie Mellon University, SCS

PhD Defense — November 30, 2022

This thesis introduces new techniques for efficiently training machine learning models over continuously arriving data to achieve high accuracy, even under changes in the data distribution over time, known as concept drift. First, we address the case of IID data with STRSAGA, an optimization algorithm based on variance-reduced stochastic gradient descent that can incorporate incrementally arriving data and efficiently converges to statistical accuracy. Second, we address the case of non-IID data over time with DriftSurf. Previous work on drift detection generally rely on magic thresholds, making them less practical without prior knowledge of the magnitude and rate of change. DriftSurf improves the robustness of traditional change detection tests through a stable-state/reactive-state process, and attains higher statistical accuracy whenever an efficient optimizer like STRSAGA is used. Third, we address the case of non-IID data both over time and distributed in space in the federated learning setting with FedDrift. Previous centralized drift adaptation and previous personalized federated learning methods are ill-suited for staggered drifts. FedDrift is the first algorithm explicitly designed for both dimensions of heterogeneity, and identifies distinct concepts by learning a time-varying clustering, which enables accurate collaborative training despite drifts. We show the presented algorithms are effective through theoretical competitive analyses and experimental studies that demonstrate higher accuracy on benchmark datasets over the prior state-of-the-art.

DISSERTATION ABSTRACT: Auto-batching Techniques for Dynamic Deep Learning Computations

Pratik Pramod Fegade
Carnegie Mellon University, SCS

PhD Defense — November 30, 2022

Deep learning is increasingly used across a range of domains. Dynamism---where the execution of a computation differs across different inputs---has been shown to be important in enabling deep learning models to effectively model the varying structure of input data in these domains, thereby achieving high accuracy. However, dynamism often makes batching, an important performance optimization, difficult to apply. This thesis presents techniques to enable efficient auto-batching—automatically enabling batched execution for a computation—for dynamic deep learning computations. We consider two kinds of dynamism commonly exhibited by deep learning computations---control flow dynamism, where the computation involves control flow structures such as conditional statements and recursion, and shape dynamism, where the computation involves tensors of different shapes across different inputs.

Past work has proposed a variety of approaches for solving this problem. However, past work is often characterized by significant fragmentation from a compilation and execution point of view. Techniques often target individual components of the compilation/runtime stack without taking a holistic view of the entire stack, and hence the entire computation, into account. For instance, tensor kernels are often optimized in isolation, independent of the surrounding computation, while auto-batching techniques often primarily rely either on compile-time, or on runtime approaches, rather than an end-to-end approach.

Considering these limitations, this thesis attempts to remove the afore-

continued on page 16

continued from page 15

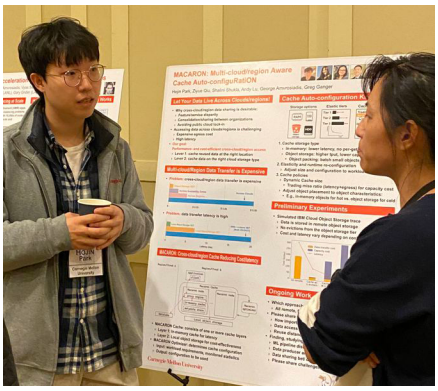
mentioned fragmentation to enable efficient auto-batching. We rely on two insights (1) hybrid static+dynamic analysis to exploit available parallelism while keeping the runtime overheads low and (2) allowing the flow of information across the compilation and execution of tensor operators and the surrounding computation. These insights enable us to obtain significant gains over past work. For instance, Cortex, a compiler specialized for recursive deep learning computations achieves up to 14x faster inference over past work, while ACRoBat, an auto-batching framework that can handle unrestricted control flow is up to 8.5x faster. Further, CoRa, a tensor compiler we designed for batched execution in the presence of shape dynamism performs on-par with highly hand-optimized implementations of the transformer model.

DISSERTATION ABSTRACT: Realizing value in shared compute infrastructures

Andrew F. Chung
Carnegie Mellon University, SCS

PhD Defense — November 18, 2022

As company operations become increasingly digitized, the demand to process data efficiently and cost-effectively has been ever-growing. More and more companies are therefore moving their workloads off of dedicated, silo-ed clusters in favor of more cost-efficient, shared data infrastructures, e.g., public and private clouds. These shared data infrastructures are often deployed on highly heterogeneous servers, are multi-tenant with server resources shared across multiple organizations, and serve widely diverse workloads ranging from batch analytics jobs to consumer-facing services with stringent service level objectives (SLOs). Both users and operators of such shared data infrastructures strive to optimize for value. Users look to complete their tasks in an efficient and timely manner without having to pay large amounts of money, while operators seek to satisfy the demands



Hojin Park talks about his work on "MACARON: Multi-cloud/region Aware Cache Auto-configuRatiON" with Yao Yue, Twitter.

of their customers to increase adoption and lower turnover, all the while without sacrificing cluster operation costs and overhead.

This dissertation presents two case studies that allows users to improve value attainment when running their workloads in shared data infrastructures in Tributary and Stratus. Tributary is an elastic control system that embraces the uncertain nature of transient cloud resources to manage elastic long-running services with latency SLOs more robustly and more cost-effectively. Stratus is a cluster scheduler specialized for orchestrating batch job execution on virtual clusters focusing primarily on dollar cost considerations: since resources in virtual clusters are charged-for while allocated, Stratus aggressively packs tasks onto machines, guided by job run time estimates, such that allocated resources remain highly utilized.

This dissertation presents two more case studies that allow cluster operators to attain value in Wing and Talon. Inter-job dependencies pervade today's shared data infrastructures, yet are often invisible to cluster schedulers. The Wing dependency profiler analyzes job and data provenance logs to find hidden inter-job dependencies, characterizes them, and provides improved guidance to cluster schedulers and workflow managers to help users attain more value. Talon is one such workflow manager that uses information provided by Wing to load-shift batch analytics jobs to off-peak hours, thereby allowing cluster operators to save on infrastructure operation costs through reduced machines managed and usage of lower-cost, transient resources from the cloud.

THESIS PROPOSAL: Efficient Loss Recovery for Videoconferencing via Streaming Codes and Machine Learning

Michael Rudow, CSD
November 4, 2022

Packet loss degrades the quality of experience (QoE) of videoconferencing. For long-distance communication, the lost packets cannot be retransmitted in real-time, necessitating forward error correction (FEC). Conventional approaches for FEC for real-time applications are highly inefficient at protecting against bursts of losses. Yet such bursts frequently arise in practice and can be better tamed with a new class of theoretical FEC schemes called "streaming codes." Streaming codes require significantly less redundancy to recover bursts, enabling more bandwidth for data or better loss recovery for the same redundancy. While existing streaming codes are well-suited to applications such as VoIP that send a sequence of frames of the same fixed size, the codes cannot handle the variability in the sizes of compressed video frames in videoconferencing. Furthermore, streaming codes' potential to improve the QoE for videoconferencing is largely untested.

This thesis presents a generalized theoretical model for streaming codes that incorporates frames of variable sizes, fundamental limits on the minimum bandwidth-overhead of redundancy needed for the new model, and new bandwidth-efficient streaming codes using machine learning and coding theory. We then present new streaming codes that address all practical limitations of these theoretical results, integrate them

continued on page 17

continued from page 16

into a videoconferencing application, and evaluate them over a simulated network over key metrics of the QoE. Our streaming codes outperform state-of-the-art baselines, such as by reducing the frequency of video freezes by 26% and the frequency of failing to render frames by 28%. Finally, we propose designing a new class of theoretical streaming codes incorporating all the nuances required for integration with videoconferencing applications.

THESIS PROPOSAL: Erasure Codes for Time and Space Heterogeneity in Storage Systems

Francisco Jose Maturana Sanguineti, CSD
October 20, 2022

Distributed storage systems play an essential role in supporting many applications that society relies on. Hence, these systems need to be highly reliable. To achieve this goal in a resource-efficient way, most storage systems use erasure codes. The degree of reliability and the resource overhead are determined by the erasure code parameters, which are set based on the expected operating conditions. However, operating conditions vary significantly across time and across the system. For example, failure rates, workloads, and density of devices can change with time and in different parts of the system. The existing approaches to handle these variations are costly and inefficient. My thesis will focus on making distributed storage systems more robust by enabling them to automatically adapt to these variations efficiently. To make progress towards this goal, we develop and use tools from both the Coding Theory and Computer Systems research.

The first part will focus on variations across time and their effects on the system. Our main contribution here is the "convertible codes" framework, designed to study and construct erasure codes that can efficiently change their parameters over time. We propose the framework, derive the fundamental limits of this problem and design optimal codes. Additionally, we propose two system designs for deciding when and how to change the erasure code.

The second part will focus on heterogeneity across the system. Specifically, we will discuss the setting of a geo-distributed storage system, where the density of nodes and latencies between nodes vary significantly and the cost of sending data across the wide-area network (WAN) is crucial. Our main contribution here is a new class of codes that optimizes for both storage overhead and WAN bandwidth for a given instance. We propose a new strongly-consistent geo-distributed storage system that enhances a state-of-the-art consensus protocol using this new class of codes.

MASTERS THESIS: High Performance DBMS Design for Intelligent Query Scheduling

Deepayan Patra, CSD
December 13, 2022

Decades of research in the field of database management systems (DBMSs) have focused on improving system performance with impressive results. Modern analytical databases take advantage of innovative methods such as vectorization and compilation to improve single query performance, use supporting data structures such as indexes or views to reduce data access requirements, and support the execution of multiple queries in parallel while maintaining necessary isolation guarantees.

We propose a new line of work with workload and architecture-aware scheduling algorithms to optimize system performance beyond the now limited incremental gains beyond the aforementioned approaches. In a modern execution environment with heterogeneous query performance and parallelism characteristics and with datasets predominantly residing in memory, resource allocation and system efficiency become paramount. Our proposed scheduling approaches take advantage of known query characteristics to intelligently order query sub-tasks in our execution environment.

In this work, we discuss modifications to a highly optimized execution engine supporting both vectorization and compilation to support newly proposed scheduling algorithms with minimal overhead. Changes to the execution architecture and in-memory data layout mitigate access pattern and function invocation overheads on the path to support NUMA-aware execution. These improvements enable the performance benefits of more intelligent scheduling approaches, which, when implemented, result in average query latency decreases of over 30%.

MASTERS THESIS: Extendable Rule-Based Action Generation for Self-Driving Database Systems

Peijing Xu, CSD
December 13, 2022

Database management systems (DBMSs) have become more complex to meet increasingly demanding usage. To owners and operators, the need for a self-driving DBMS that can automatically tune and optimize itself without human intervention is apparent now more than ever. Such a self-driving DBMS considers a set of candidate actions to apply to reach a configuration that improves performance for a given workload. Furthermore, the DBMS would continuously adjust the configuration in anticipation of changing workloads and data distributions.

Efforts to architect self-driving DBMSs suffer from the engineering overhead of combining different tuning subtasks, such as index tuning and knob tuning. These individual subtasks have a vast candidate action space, requiring the tuning algorithms to reduce the action space before searching. However, each subtask and algorithm has its own representation of the action space and

continued on page 18

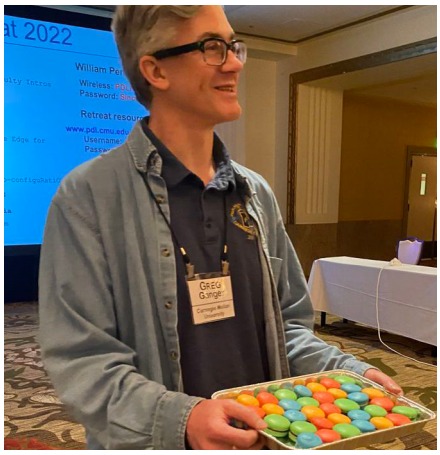
DEFENSES & PROPOSALS

continued from page 17

methods for obtaining requisite inputs for the algorithm, including a representative workload and the database schema.

This thesis presents an extensible framework for defining the action space of tuning subtasks. The framework allows the self-driving DBMS engineer to define action types, rules for constructing the action space, and input information for the tuning algorithm – all in a standard interface shared across subtasks. The framework reduces the overhead of developing and evaluating new tuning algorithms and allows the DBMS to dynamically define the subset of the action space to explore in any given tuning task. By restricting the search space before executing the tuning algorithm, the framework reduces the time expended on evaluating suboptimal configurations. It thereby improves the speed at which the algorithm converges on a solution.

We then use this framework to alter and restrict the action spaces of existing algorithms for index tuning and knob tuning. We demonstrate that by filtering the search space against low-quality candidate actions, the framework enables tuning algorithms to converge more quickly on tuning actions that can match or outperform the baseline solution.



Greg offers macarons to retreat attendees (homemade by George Amvrosiadis) in honor of the Macaron project.

RECENT PUBLICATIONS

continued from page 7

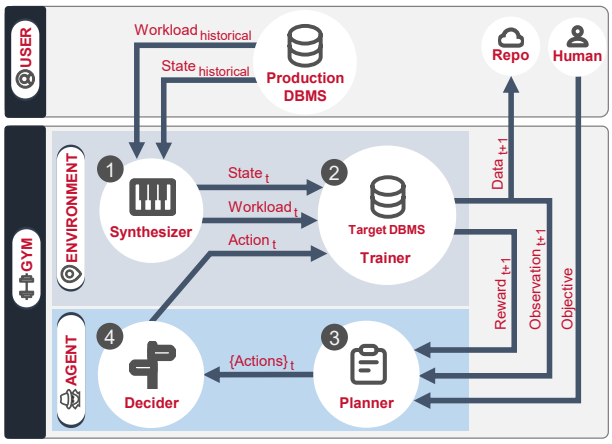
ZNS's opportunities for increased application performance, allowing higher-level software (e.g., F2FS or RocksDB) to carefully control garbage collection. This allows, for example, RAIZN to maintain consistent performance under scenarios where conventional SSD arrays experience up to 87.5% throughput drop due to device-level garbage collection.

Database Gyms

Wan Shen Lim, Matthew Butrovich, William Zhang, Andrew Crotty, Lin Ma, Peijing Xu, Johannes Gehrke, Andrew Pavlo

CIDR 2023. 13th Annual Conference on Innovative Data Systems Research (CIDR '23). January 8-11, 2023, Amsterdam, The Netherlands.

In the past decade, academia and industry have embraced machine learning (ML) for database management system (DBMS) automation. These efforts have focused on designing ML models that predict DBMS behavior to support picking actions (e.g., building indexes) that improve the system's performance. Recent developments in ML have created automated methods for finding good models. Such advances shift the bottleneck from DBMS model design to obtaining the training data necessary for building these models. But generating good training data is challenging and requires encoding subject matter expertise into DBMS instrumentation.



Database Gym Architecture – An overview of the database gym's internals. The t subscript refers to iterations inside the Gym

Existing methods for training data collection are bespoke to individual DBMS components and do not account for (1) how workload trends affect the system and (2) the subtle interactions between internal system components. Consequently, the models created from this data do not support holistic tuning across subsystems and require frequent retraining to boost their accuracy.

This paper presents the architecture of a database gym, an integrated environment that provides a unified API of pluggable components for obtaining high-quality training data. The goal of a database gym is to simplify ML model training and evaluation to accelerate autonomous DBMS research. But unlike gyms in other domains that rely on custom simulators, a database gym uses the DBMS itself to create simulation environments for ML training. Thus, we discuss and prescribe methods for overcoming challenges in DBMS simulation, which include demanding

continued on page 19

continued from page 18

requirements for performance, simulation fidelity, and DBMS-generated hints for guiding training processes.

PIM-tree: A Skew-resistant Index for Processing-in-Memory

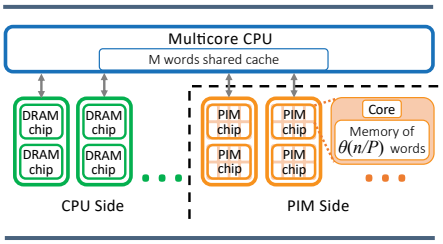
Hongbo Kang, Yiwei Zhao, Guy E. Blelloch, Laxman Dhulipala, Yan Gu, Charles McGuffey, Phillip B. Gibbons

VLDB Endow. 16(4): 946-958 (2022).

BEST PAPER RUNNER-UP

The performance of today's in-memory indexes is bottlenecked by the memory latency/bandwidth wall. Processing-in-memory (PIM) is an emerging approach that potentially mitigates this bottleneck, by enabling low-latency memory access whose aggregate memory bandwidth scales with the number of PIM nodes. There is an inherent tension, however, between minimizing inter-node communication and achieving load balance in PIM systems, in the presence of workload skew. This paper presents PIM-tree, an ordered index for PIM systems that achieves both low communication and high load balance, regardless of the degree of skew in the data and the queries. Our skew-resistant index is based on a novel division of labor between the multi-core host CPU and the PIM nodes, which leverages the strengths of each. We introduce push-pull search, which dynamically decides whether to push queries to a PIM-tree node (CPU→PIM-node) or pull the node's keys back to the CPU (PIM-node→CPU) based on workload skew. Combined with other PIM-friendly optimizations (shadow subtrees and chunked skip lists), our PIM-tree provides high-throughput, (guaranteed) low communication, and (guaranteed) high load balance, for batches of point queries, updates, and range scans.

We implement the PIM-tree structure, in addition to prior proposed PIM indexes, on the latest PIM system from UPMEM, with 32 CPU cores and 2048 PIM nodes. On workloads with 500 million keys and batches of 1 million queries, the throughput using PIM-trees is up to 69.7× and 59.1× higher than the two best prior PIM-based methods. As far as we know these are the first implementations of an ordered index on a real PIM system.



The architecture for the UPMEM PIM system, a specific example of our generic PIM system architecture. PIM modules are packed into memory DIMMs connected to the host CPU via normal memory channels. The CPU side also includes traditional DRAM modules, which are not part of the PIM model.

Extending and Programming the NVMe I/O Determinism Interface for Flash Arrays

Huaicheng Li, Martin L Putra, Ronald Shi, Fadhil I Kurnia, Xing Lin, Jaeyoung Do, Achmad Imam Kistijantoro, Gregory R Ganger, Haryadi S Gunawi

ACM Transactions on Storage, Vol. 19, No. 1, Article 5. January 2023.

Predictable latency on flash storage is a long-pursuit goal, yet, unpredictability stays due to the unavoidable disturbance from many well-known SSD internal activities. To combat this issue, the recent NVMe IO Determinism (IOD) interface advocates host-level controls to SSD internal management tasks. While promising, challenges remain on how to exploit it for truly predictable performance.

We present IODA, an I/O deterministic flash array design built on top of small but powerful extensions to the IOD interface for easy deployment. IODA exploits data redundancy in the context of IOD for a strong latency predictability contract. In IODA, SSDs are expected to quickly fail an I/O on purpose to allow predictable I/Os through proactive data reconstruction. In the case of concurrent internal operations, IODA introduces busy remaining time exposure and predictable-latency-window formulation to guarantee predictable data reconstructions. Overall, IODA only adds 5 new fields to the NVMe interface and a small modification in the flash firmware, while keeping most of the complexity in the host OS. Our evaluation shows that IODA improves the 95–99.99th latencies by up to 75×. IODA is also the nearest to the ideal, no disturbance case compared to 7 state-of-the-art preemption, suspension, GC coordination, partitioning, tiny-tail flash controller, prediction, and proactive approaches.

Rateless Sum-Recovery Codes For Distributed Non-Linear Computations

Ankur Mallick, Gauri Joshi

Information Theory Workshop (ITW), November 6-9, 2022. Mumbai, India.

We address the problem of slowdown caused by straggling nodes in distributed non-linear computations. Many common non-linear computations can be written as a sum of inexpensive non-linear functions (for e.g. Taylor series). Based on this observation, we propose a new class of rateless codes called rateless sum-recovery codes whose aim is to recover the sum of source symbols, without necessarily recovering individual symbols. Source symbols correspond to individual inexpensive functions and each encoded symbol is the sum of a subset of source symbols. Encoded symbols are computed in a distributed fashion and for a computation that can be written as a sum of m inexpensive functions, successful sum-recovery is possible with high probability as long as slightly more than m encoded symbols are received. Our code is rateless, systematic and has sparse parities. Moreover,

continued on page 20

continued from page 19

encoded symbols are constructed by sampling without replacement at individual nodes, thereby making decoding superfluous if the encoded symbols from any node cover all source symbols. We validate our claims through a range of simulations and also discuss open questions for future works.

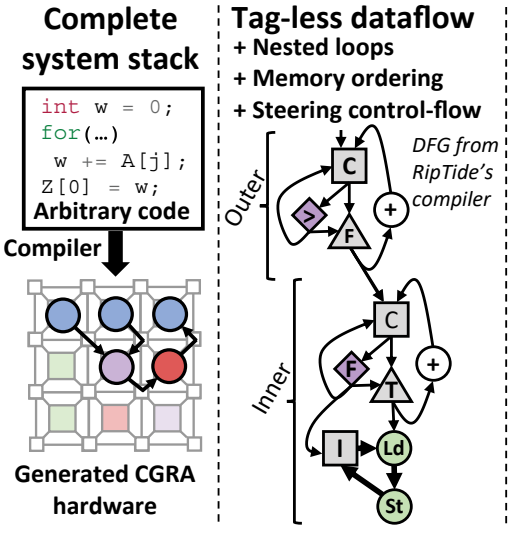
RipTide: A Programmable, Energy-minimal Dataflow Compiler and Architecture

Graham Gobieski, Souradip Ghosh, Marijn Heule, Todd Mowry, Tony Nowatzki, Nathan Beckmann, Brandon Lucia

MICRO 2022 - 55th IEEE/ACM International Symposium on Microarchitecture, October 1–5, 2022 Chicago, Illinois, USA.

Emerging sensing applications create an unprecedented need for energy efficiency in programmable processors. To achieve useful multi-year deployments on a small battery or energy harvester, these applications must avoid off-device communication and instead process most data locally. Recent work has proven coarse-grained reconfigurable arrays (CGRAs) as a promising architecture for this domain. Unfortunately, nearly all prior CGRAs support only computations with simple control flow and no memory aliasing (e.g., affine inner loops), causing an Amdahl efficiency bottleneck as non-trivial fractions of programs must run on an inefficient von Neumann core.

RipTide is a co-designed compiler and CGRA architecture that achieves both high programmability and extreme energy efficiency, eliminating this bottleneck. RipTide provides a rich set of control-flow operators that support arbitrary control flow and memory access on the CGRA fabric. RipTide implements these primitives without tagged tokens to save energy; this requires careful ordering analysis in the compiler to guarantee correctness. RipTide further saves energy and area by offloading most control operations into its programmable on-chip network, where they can re-use existing network switches. RipTide's compiler is implemented in LLVM, and its hardware is synthesized in Intel 22FFL. RipTide compiles applications written in C while saving 25% energy v. the state-of-the-art energy-minimal CGRA and 6.6x energy v. a von Neumann core.



RipTide is a co-designed compiler and CGRA architecture that executes programs written in a high-level language with minimal energy and high performance. RipTide introduces new control-flow primitives to support common programming idioms, like deeply nested loops and irregular memory accesses, while minimizing the energy overhead. RipTide implements control flow in the NoC to increase utilization and ease compilation.

MATCHA: A Matching-Based Link Scheduling Strategy to Speed up Distributed Optimization

Jianyu Wang, Anit Sahu, Gauri Joshi, Soummya Kar

IEEE Transactions on Signal Processing, Oct 2022.

In this paper, we study the problem of distributed optimization using an arbitrary network of lightweight computing nodes, where each node can only send/receive information to/from its direct neighbors. Decentralized stochastic gradient descent (SGD) has been shown to be an effective method to train machine learning models in this setting. Although decentralized SGD has been extensively studied, most prior works focus on the error-versus-iterations convergence, without taking into account how the topology affects the communication delay per iteration. For example, a denser (sparser) network topology results in faster (slower) error convergence in terms of iterations, but it incurs more (less) communication time per iteration. We propose MATCHA, an algorithm that can achieve a win-win in this error-runtime trade-off for any arbitrary network topology. The main idea of MATCHA is to communicate more frequently over connectivity-critical links in order to ensure fast convergence, and at the same time minimize the communication delay per iteration by using other links less frequently. It strikes this balance by decomposing the topology into matchings and then optimizing the set of matchings that are activated in each iteration. Experiments on a suite of datasets and deep neural networks validate the theoretical analyses and

continued on page 21

continued from page 20

demonstrate that MATCHA takes up to 5x less time than vanilla decentralized SGD to reach the same training loss. The idea of MATCHA can be applied to any decentralized algorithm that involves a communication step with neighbors in a graph.

Kangaroo: Theory and Practice of Caching Billions of Tiny Objects on Flash

Sara McAllister, Benjamin Berg, Julian Tutuncu-Macias, Juncheng Yang, Sathya Gunasekar, Jimmy Lu, Daniel S Berger, Nathan Beckmann, Gregory R Ganger

ACM Transactions on Storage, Vol. 18, No. 3, Article 21. August 2022.

Many social-media and IoT services have very large working sets consisting of billions of tiny (≈100 B) objects. Large, flash-based caches are important to serving these working sets at acceptable monetary cost. However, caching tiny objects on flash is challenging for two reasons: (i) SSDs can read/write data only in multi-KB "pages" that are much larger than a single object, stressing the limited number of times flash can be written; and (ii) very few bits per cached object can be kept in DRAM without losing flash's cost advantage. Unfortunately, existing flash-cache designs fall short of addressing these challenges: write-optimized designs require too much DRAM, and DRAM-optimized designs require too many flash writes.

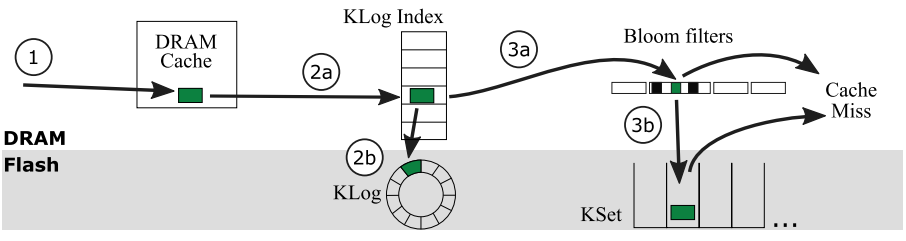
We present Kangaroo, a new flash-cache design that optimizes both DRAM usage and flash writes to maximize cache performance while minimizing cost. Kangaroo combines a large, set-associative cache with a small, log-structured cache. The set-associative cache requires minimal DRAM, while the log-structured cache minimizes Kangaroo's flash writes. Experiments using traces from Meta and Twitter show that Kangaroo achieves DRAM usage close to the best prior DRAM-optimized design, flash writes close to the best prior write-optimized design, and miss ratios better than both. Kangaroo's design is Pareto-optimal across a range of allowed write rates, DRAM sizes, and flash sizes, reducing misses by 29% over the state of the art. These results are corroborated by analytical models presented herein and with a test deployment of Kangaroo in a production flash cache at Meta.

Plumber: Diagnosing and Removing Performance Bottlenecks in Machine Learning Data Pipelines

Michael Kuchnik, Ana Klimovic, Jiri Simsa, Virginia Smith, George Amvrosiadis

5th MLSys Conference, Santa Clara, CA, USA, August 2022.

Input pipelines, which ingest and transform input data, are an essential part of training Machine Learning (ML) models. However, it is challenging to implement efficient input pipelines, as it requires reasoning about parallelism, asynchrony, and variability in fine-grained profiling information. Our analysis of over two million ML jobs in Google datacenters reveals that a significant fraction of model training jobs could benefit from faster input data pipelines. At the same time, our analysis indicates that most jobs do not saturate host hardware, pointing in the direction of software-based bottlenecks. Motivated by these findings, we propose Plumber, a tool for finding bottlenecks in ML input pipelines. Plumber uses an extensible and interpretable operational analysis analytical model to automatically tune parallelism, prefetching, and caching under host resource constraints. Across five representative ML pipelines, Plumber obtains speedups of up to 47 for misconfigured pipelines. By automating caching, Plumber obtains end-to-end speedups of over 50% compared to state-of-the-art tuners.



Lookups in Kangaroo first check a tiny DRAM cache; then KLog, a small on-flash log-structured cache with an in-DRAM index; and finally KSet, a large on-flash set-associative cache. Kangaroo uses little DRAM because KLog is small and KSet has no DRAM index.

SurgeProtector: Mitigating Temporal Algorithmic Complexity Attacks using Adversarial Scheduling

Nirav Atre, Hugo Sadok, Erica Chiang, Weina Wang, Justine Sherry

SIGCOMM '22, August 22–26, 2022, Amsterdam, Netherlands.

Denial-of-Service (DoS) attacks are the bane of public-facing network deployments. Algorithmic complexity attacks (ACAs) are a class of DoS attacks where an attacker uses a small amount of adversarial traffic to induce a large amount of work in the target system, pushing the system into overload and causing it to drop packets from innocent users. ACAs are particularly dangerous because, unlike volumetric DoS attacks, ACAs don't require a significant network bandwidth investment from the attacker. Today, network functions (NFs) on the Internet must be designed and engineered on a case-by-case basis to mitigate the debilitating impact of ACAs. Further,

continued on page 22

continued from page 21

the resulting designs tend to be overly conservative in their attack mitigation strategy, limiting the innocent traffic that the NF can serve under common-case operation.

In this work, we propose a more general framework to make NFs resilient to ACAs. Our framework, SurgeProtector, uses the NF's scheduler to mitigate the impact of ACAs using a very traditional scheduling algorithm: Weighted Shortest Job First (WSJF). To evaluate SurgeProtector, we propose a new metric of vulnerability called the Displacement Factor (DF), which quantifies the 'harm per unit effort' that an adversary can inflict on the system. We provide novel, adversarial analysis of WSJF and show that any system using this policy has a worst-case DF of only a small constant, where traditional schedulers place no upper bound on the DF. Illustrating that SurgeProtector is not only theoretically, but practically robust, we integrate SurgeProtector into an open source intrusion detection system (IDS). Under simulated attack, the SurgeProtector- augmented IDS suffers 90-99% lower innocent traffic loss than the original system.

Extending and Programming the NVMe I/O Determinism Interface for Flash Arrays

Huaicheng Li, Martin L Putra, Ronald Shi, Fadhil I Kurnia, Xing Lin, Jaeyoung Do, Achmad Imam Kistijantoro, Gregory R Ganger, Haryadi S Gunawi

ACM Transactions on Storage, 2022.

Predictable latency on flash storage is a long-pursuit goal, yet, unpredictability stays due to the unavoidable disturbance from many well-known SSD internal activities. To combat this issue, the recent NVMe IO Determinism (IOD) interface advocates host-level controls to SSD internal management tasks. While promising, challenges remain on how to exploit it for truly predictable performance.

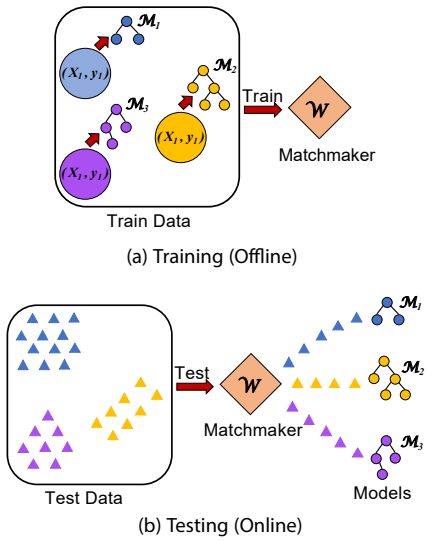
We present IODA , an I/O deterministic flash array design built on top of small but powerful extensions to the IOD interface for easy deployment. IODA exploits data redundancy in the context of IOD for a strong latency predictability contract. In IODA, SSDs are expected to quickly fail an I/O on purpose to allow predictable I/Os through proactive data reconstruction. In the case of concurrent internal operations, IODA introduces busy remaining time exposure and predictable-latency-window formulation to guarantee predictable data reconstructions. Overall, IODA only adds 5 new fields to the NVMe interface and a small modification in the flash firmware, while keeping most of the complexity in the host OS. Our evaluation shows that IODA improves the 95–99.99th latencies by up to 75×. IODA is also the nearest to the ideal, no disturbance case compared to 7 state-of-the-art preemption, suspension, GC coordination, partitioning, tiny-tail flash controller, prediction, and proactive approaches.

Matchmaker: Data Drift Mitigation in Machine Learning for Large-scale Systems

Ankur Mallick, Kevin Hsieh, Behnaz Arzani, Gauri Joshi

Proceedings of the 5th MLSys Conference, Santa Clara, CA, USA, August, 2022.

Today's data centers rely more heavily on machine learning (ML) in their deployed systems. However, these systems are vulnerable to the data drift problem, that is, a mismatch between training data in the past and test data in the future, which can lead to significant performance degradation and system inefficiencies. In this paper, we demonstrate the impact of data drift in production by studying two real-world deployments in a leading cloud provider. Our study shows that, despite frequent model retraining, these deployed models experience



Overview of Matchmaker with 3 training batches. Predictive models (M_1,M_2,M_3) are trained offline along with Matchmaker, W (a). At test time (b) W assigns each test point to the model from the most similar batch (same color) for prediction.

major accuracy drops (up to 40%) and high accuracy variation, which lead to significant increase in operational costs. Existing solutions to the data drift problem are not designed for large-scale deployments, which need to address real-world issues such as scalability, ground truth latency, and mixed types of data drift. We propose Matchmaker, the first scalable, adaptive, and flexible solution to the data drift problem in large-scale production systems. Matchmaker finds the most similar training data batch and uses the corresponding ML model for inference on each test point. As part of Matchmaker, we introduce a novel similarity metric to address multiple types of data drifts while only incurring limited overhead. Experiments on our two real-world ML deployments show that Matchmaker significantly improves model accuracy (up to 14% and 2%), which saves 18% and 1% in the operational costs. At the same time, Matchmaker provides 8x and 4x faster predictions than a state-of-the-art ML data drift solution, AUE.

continued from page 4

January 2023

- Wan Shen Lim presented "Database Gyms" at CIDR 2023 in Amsterdam, The Netherlands.
- The paper "Extending and Programming the NVMe I/O Determinism Interface for Flash Arrays" by Huaicheng Li, Martin L Putra, Ronald Shi, Fadhil I Kurnia, Xing Lin, Jaeyoung Do, Achmad Imam Kistijantoro, Gregory R Ganger, Haryadi S Gunawi appeared in ACM Transactions on Storage.

December 2022

- Ziqi Wang gave his speaking skills talk on "Memento: Architectural Support for Ephemeral Memory Management."
- Deepayan Patra presented his MS Thesis on " High Performance DBMS Design for Intelligent Query Scheduling."
- Peijing Xu presented his MS Thesis on "Extendable Rule-Based Action Generation for Self-Driving Database Systems."

November 2022

- The 28th PDL Retreat was held at the Omni William Penn hotel in Pittsbugh.
- Ankur Mallick presented "Rateless Sum-Recovery Codes For Distributed Non-Linear Computations" at ITW '22 in Mumbai, India.
- Majd Sakr was awarded an NSF Future of Work Grant.
- David Andersen won the ACM SIGOPS Mark Weiser Award.
- Akshitha Sriraman was a runner-up for the 2022 Dennis M. Ritchie Thesis Award for her research on "Enabling Hyperscale Web Services."
- Andrew Chung defended his dissertation on "Realizing Value in Shared Compute Infrastructures."
- Pratik Pramod Fegade defended his reasearch on "Auto-batching Techniques for Dynamic Deep Learning Computations."
- Ellango Jothimurugesan defended his research on "Large-scale Machine Learning over Streaming Data."
- Michael Rudow proposed his PhD research on "Efficient Loss Recovery for Videoconferencing via Streaming Codes and Machine Learning."

October 2022

- Graham Gobieski presented "RipTide: A Programmable, Energy-minimal Dataflow Compiler and Architecture" at MICRO 2022 in Chicago, Illinois, USA.
- The paper "MATCHA: A Matching-Based Link Scheduling Strategy to Speed up Distributed Optimization" by Jianyu Wang Anit Sahu, Gauri Joshi, and Soummya Kar appeared in the IEEE Transactions on Signal Processing.
- Zhihao Jia, an assistant professor of CS/ECE, won an Amazon Research Award.
- Thomas Kim gave his speaking skills talk on "Reliability and Availability for Arrays of Zoned Namespace SSDs."
- Francisco Jose Maturana Sanguineti proposed his PhD research topic "Erasure Codes for Time and Space Heterogeneity in Storage Systems."

PDL ALUMNI NEWS

John Linwood Griffin & Jiri Schindler

(both PDL 1998-2004)

Both John and Jiri still enjoy a yearly ski trip with retired PDC member Paul Massiglia, who was with Veritas when he joined the PDL at retreats and visit days for many years.

Niraj Tolia, Eno Thereska & Rajat Kateja

(PDL 2002-07, 2002-07 & 2015-22)

Niraj has embarked on another startup adventure, this time with Alcion — an AI-driven approach to securely protect data from ransomware, malware, accidents, and outages (www.alcion.ai). Joining him are two other PDL alumni, Eno Thereska, and Rajat Kateja. There are several other CMU/ECE people involved with Alcion too, including Vaibhav Kamra, the startup's co-founder.

Raja Sambasivan

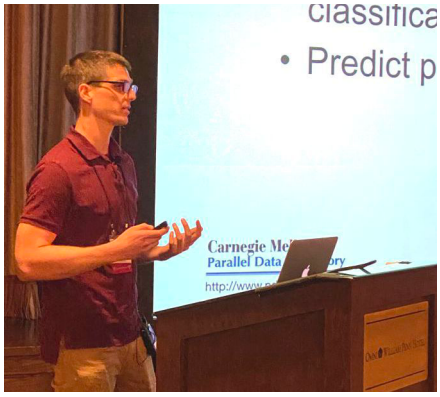
(PDL 2002-2007)

Raja sent a photo from his research lab's "start-of-semester dinner" last year. He said it reminded him of the camaraderie and fun times he enjoyed as a PDL student years ago, and wanted to share it. Raja currently holds the Ankur and Mari Sahu Professorship, in the Computer Science Department of the Tufts University School of Engineering.

NEW PDL FACULTY

Jignesh Patel
Professor, CSD

We would like to welcome Jignesh Patel to the PDL. Jignesh is an incoming professor in the Computer Science Department at Carnegie Mellon University. His research focuses on data management, emphasizing both system efficiency (e.g., making data platforms run faster) and human efficiency (e.g., designing LLM-based query interfaces). His papers have been recognized as the best papers at top database conferences, including SIGMOD and VLDB. He is a fellow of the AAAS, ACM, and IEEE organizations. He has also received teaching awards at the U. of Wisconsin and the U. of Michigan. Jignesh has a strong interest in technology transfer. He has launched four startups and has also made key contributions to product improvements for industry sponsors.



Michael Kuchnik tells retreat attendees about his research on "Validating Large Language Models with ReLM."

PDL NEWS & AWARDS

continued from page 10

November 2022
Akshitha Sriraman is a Runner-up for the 2022 Dennis M. Ritchie Thesis Award

Congratulations to Akshitha! Her dissertation "Enabling Hyperscale Web Services, presentedat the University of Michigan under the advisorship of Tomas Wenisch has been noted as a runner-up to the The Dennis M. Ritchie Doctoral Dissertation Award. The award was created in 2013 by ACM SIGOPS to recognize research in software systems and to encourage the creativity that Dennis Ritchie embodied, providing a reminder of Ritchie's legacy and what a difference one person can make in the field of software systems research.

October 2022
Zhihao Jia Wins Amazon Research Award

Congratulations to Zhihao on receiving an Amazon Research Award funded under the fall 2021 AWS AI and winter 2022 Alexa: Fairness in AI call for proposals for his ideas on "Towards Affordable and Accessible ML by Leveraging Heterogeneous Spot Instances. "Proposals were reviewed for the quality of their scientific content, their creativity, and their potential to impact both the research community and society more generally. Theoretical advances, creative new ideas, and practical applications were all considered. The Amazon Research Awards program provides unrestricted funds and AWS Promotional Credits to academic researchers investigating research topics across a number of disciplines.

Attendees of the 28th Almost Annual PDL Retreat held at the Omni William Penn in Pittsburgh, PA, November 7-9, 2022.