RESEARCH STATEMENT

1 Overview

Over the past five years (2018-2023) at Indiana University (IU), Washington State University (WSU), and the University of Alabama (UA) as a tenure-track faculty, I've established myself as a world-renowned leader in **high-performance computing (HPC)** research. I've developed a series of parallel algorithms and software to address critical performance, scalability, and reliability challenges in supercomputing systems and applications. In particular, the award-winning data compression software effectively reduces data size while maintaining high data fidelity for subsequent analysis, catering to a wide variety of use-cases in scientific simulations and instrument facilities. Building on this progress, I am now pursuing more ambitious goals at the intersection of HPC and artificial intelligence (AI), aiming to create HPC infrastructure that supports large-scale AI tasks in alignment with the emerging national-level "AI for Science" initiative.

1.1 Overall goals

HPC plays a vital role in addressing pressing societal challenges such as vaccine and drug design, climate change, water management, and advanced manufacturing. However, scientific simulations running on these systems generate extremely large volumes of data, leading to challenges such as storage burden, I/O bottlenecks, and insufficient memory. Thus, efficiently storing, transferring, and analyzing such vast amounts of data is a significant and demanding issue. Additionally, with the substantial increase in the scale of HPC systems and the constraints of power consumption, the reliability of HPC systems has rapidly declined. This problem makes it difficult for large-scale applications to run stably over extended periods. Consequently, designing efficient techniques to improve the reliability of HPC systems is crucial for ensuring the long-term stable execution of HPC applications. My goal is to explore novel **software solutions** that provide **efficient**, **scalable**, and **reliable** computing for HPC applications. Central to my solutions is the development of innovative **data reduction** techniques and cyber-infrastructures.

1.2 Key statistics

During my four years at UA and WSU as a tenure-track assistant professor before joining IU (August 2018-August 2022), I accomplished the following:

- I authored 40 peer-reviewed conference papers and 9 peer-reviewed journal papers (see my CV).
- I co-designed, developed, and released SZ, an open-source lossy compression framework for scientific data, which won the **2021 R&D 100 Award** (often referred to as "The Oscars of Innovation" and the "Nobel Prize of Engineering"). The SZ compression framework effectively reduces data size while maintaining high data fidelity for subsequent analysis. It showcases the broadest range of use-cases and delivers superior performance in compression ratios, speed, and accuracy compared to competing products. SZ is suitable for a wide array of scientific simulations and instrument facilities.
- I launched cuSZ, an open-source GPU-accelerated error-bounded lossy compression for scientific data. As the first error-bounded lossy compressor (circa 2020) on GPU for scientific data, it aims to significantly improve SZ's throughput on heterogeneous HPC systems.
- I secured 5 grants from the National Science Foundation (NSF) as the PI, totaling $1.46 million, including the prestigious **NSF CAREER Award** and **NSF CRII Award**.
- I obtained 4 grants from Argonne National Lab (ANL) as the PI, totaling $419K.
- I was awarded 1 grant from NSF and 1 grant from the National Oceanic and Atmospheric Administration (NOAA) as a co-PI, totaling $10.4 million.
- I received **1 Best Paper Award** (1 out of 154) and **2 Best Area Paper Awards** (4 out of 154) from the 2022 IEEE International Conference on Cluster Computing (CLUSTER).
- I earned the **IEEE Computer Society TCHPC Early Career Researchers Award for Excellence in High Performance Computing**. This award recognizes up to 3 individuals who have made outstanding, influential, and potentially long-lasting contributions in the field of HPC within 5 years of receiving PhD.
- I delivered 22 invited and contributed talks internationally.

During my one year at Indiana University (August 2022-July 2023), I achieved the following:

- I authored 7 top-tier conference papers, 2 peer-reviewed journal papers, 4 accepted paper, and 2 papers under review (see my CV for the full publication list).
- I hold the **No. 1 position in the U.S.** and the **No. 2 position globally** alongside Prof. Ian Foster from the University of Chicago, in terms of the publications in prestigious HPC conferences (i.e., SC, HPDC, ICS) within the past five years (2018-2023), according to the widely recognized *CSRankings*.
- I co-designed and released SZ3, a more flexible and modularized lossy compression framework for scientific data compared to its previous version, SZ2.
- I integrated the SZ software into the Extreme-scale Scientific Software Stack (E4S), whose docker image has **over 1M downloads**, and deployed it in the first U.S. exascale supercomputer, "Frontier".
- I received consistent invitations to participate in NSF panels (once in 2022 and five times in 2023), including the NSF CSSI, OAC Core, and DESC programs.
- I transferred 5 previously awarded NSF grants to IU, totaling $1.29 million.
- I secured 3 new grants from ANL as the PI, amounting to $305K.
- I obtained 1 new grant from NSF, totaling $3.6 million (my share is $680K).
- I received a research award of $50K from Meta Platforms.
- I delivered 5 invited and contributed talks internationally.
- I expanded my research lab to include a team of **7 funded** PhD students (with another student starting in Fall 2023) and 2 undergraduate students (see Teaching Statement for students mentoring).

2 Progress and ongoing work

2.1 Novel lossy data compression framework and its efficient parallel I/O system

As HPC systems' computing power and AI algorithms advance, HPC-driven simulations, modeling, big data analysis, and AI are converging. This convergence leads to an exponential increase in scientific data volume, driving the demand for I/O bandwidth and network communication in HPC systems. However, current memory read/write speeds lag behind processor computation speeds, causing processors to spend considerable time idle while waiting for data in applications requiring massive data movement (i.e., the memory wall problem). This prevents processors from realizing their full potential and hinders HPC and AI development. Furthermore, as discussed in my prior work [1], I/O bandwidth development in HPC systems lags behind peak speed and memory capacity growth. Thus, enhancing I/O performance and storage capacity in HPC systems is crucial.

Data reduction technology is key to addressing I/O performance and storage capacity issues. I innovatively developed the scientific data compression software, SZ (Figure 1), which is highly modular, parameterized, and valuable academically. Widely used in fields like climate, cosmology, molecular dynamics, and more, it won the **2021 R&D 100 Award**. SZ substantially reduces data volume while providing precise compression-error control strategies tailored to the application's needs. As foundational HPC software, SZ leverages surplus computing power, helping applications minimize the gap between system computing speed and I/O and communication speeds. The core algorithm of SZ was published in IEEE IPDPS 2017 (I am the first author) and has received **over 220 Google Scholar citations** since its publication (top three Supercomputing conference papers averaged 180 citations in the past five years).

My team effectively implemented SZ on parallel hardware like GPUs and FPGAs, integrating it into widely-used parallel I/O libraries (e.g., HDF5 and ADIOS). It has been deployed on the University of Maryland's flagship cluster and the Exascale supercomputer Frontier. I conducted high-accuracy theoretical modeling and precise control for metrics such as compression ratio, throughput, and rate distortion, enabling asynchronous execution of compression and I/O operations to significantly improve large-scale parallel I/O performance. The core technology was published in IEEE TPDS 2019 and ACM/IEEE SC 2022 and IEEE ICDE 2022. These results attracted attention from top research teams: (1) University of Tennessee Distinguished Prof. Jack Dongarra (Turing Award winner) praised SZ in his IEEE Cluster 2022 paper [2];

User Institution	Application Domain
University of Chicago	Quantum Circuit Simulation (Wu et al. [10])
Argonne National Lab	Quantum Chemistry Computation (Gok et al. [5]
National Center for Atmospheric Research	Climate and Environmental Simulation (Poppick et al. [9])
Lawrence Berkeley National Lab	Cosmological Simulation (Jin et al. [7]
Los Alamos National Lab	Molecular Dynamics Simulation (Zhao et al. [13]
Saudi Aramco	Seismic Tomography Imaging (Zhao et al. [12])
Stanford University	Combustion Numerical Simulation (Chung et al. [3])

Table 1: Examples of SZ's Usage in Various Application Domains

(2) Stanford University Prof. Gianluca Iaccarino commended SZ in his IJHPCA 2022 paper [4]; and (3) University of Utah Chair Prof. Valerio Pascucci lauded SZ in his IEEE TVCG 2018 paper [6].

My open-source scientific data compression software, SZ, developed with Argonne National Lab scientists, is widely adopted by international researchers across various fields (Table 1). SZ has been integrated into the Exascale Scientific Software Stack (E4S). Users can install it using "spack install sz". I actively collaborate with user teams to optimize SZ for different applications. For example, I partnered with Argonne to propose a data prediction method for double-electron integrals in quantum chemistry simulations. This optimized SZ version compresses data by 16.8 times with high accuracy and improves the performance of the GAMESS application by approximately 3.7 times. The core method was published in IEEE Cluster 2018, winning the **Best Paper Award** (1/154). Additionally, I collaborated with Los Alamos National Lab to propose an in-situ compression error control theory for large-scale structure numerical simulations in cosmology. This lossy compression method allows accurate dark matter halo and spectral analysis while increasing the compression ratio by 73%. The core method was published in ACM HPDC 2021. We recently extended this research to adaptive mesh refinement applications and published the results in HPDC 2022.

Ongoing and future work. I am currently researching and developing novel algorithms and software to improve the **efficacy**, **usability**, **performance**, and **scalability** of scientific data reduction through **holistic analytical modeling** and **deep architectural optimization** as part of my NSF CAREER project. This research will contribute to the cyberinfrastructure of big data management in HPC environments. The framework is applicable to HPC applications in various domains.

Current error-bounded lossy compression methods have four significant drawbacks: (1) **Efficacy**: Existing methods use static designs without considering dataset characteristics, limiting dynamic optimization of compression ratio/quality. (2) **Usability**: Scientists must conduct tedious trial-and-error experiments to configure compression parameters, ensuring no invalid results in post-analyses. (3) **Performance**: Existing methods with low GPU compression throughput or ratio are limited in in-memory compression scenarios, especially compared to high-bandwidth CPU-GPU interconnects. (4) **Scalability**: Current parallel I/O and communication libraries cannot fully leverage lossy compression to reduce data-movement time.

My current research addresses four critical data reduction issues in scientific applications on advanced CI: efficacy, usability, performance, and scalability. *High efficacy* achieves maximum data reduction, while *high usability* ensures post-analysis quality. *High performance* optimizes application performance, and *high scalability* supports complex HPC data movements. My research consists of the following thrusts:

- **Thrust 1:** I'm developing lightweight models to estimate compression ratio and quality, optimizing

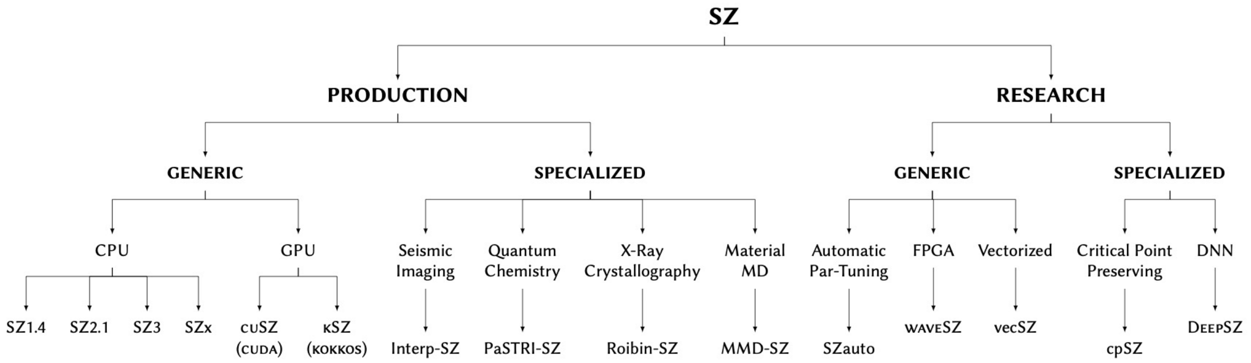

Figure 1: Overview of SZ compression framework.

configurations for maximum efficacy and usability under quality constraints.

- **Thrust 2:** To tackle performance issues, I'm creating efficient predictors and encoding methods for GPU-based lossy compression, aiming for high throughput and ratio.
- **Thrust 3:** Addressing scalability, I'm integrating optimized compression with parallel I/O and MPI communication to improve complex data movement performance and HPC application scalability.

In addition, taking into account that data movement represents the largest performance and energy consumption overhead in future HPC systems, I plan to further explore the use of approximate computing techniques (encompassing both low-precision and lossy compression methods) to substantially enhance the performance of next-generation high-performance computers, particularly in areas such as efficient communication and I/O layers.

2.2 Efficient and scalable deep learning framework for large-scale model training

AI models, particularly deep learning models, are evolving towards ultra-large scales, with distributed model training on HPC systems as the future trend. Large-scale model training consumes significant computing resources and time due to high communication, data loading, and memory overheads, limiting scalability and training speed. Maximizing communication efficiency and minimizing memory overhead are crucial for integrating future HPC and AI and developing AI for Science.

First, I innovatively proposed a dynamic compression method for activation data in deep learning training and conducted a comprehensive analysis of training accuracy under the influence of compression errors from both theoretical and experimental perspectives. By integrating activation CPU-GPU data migration strategies, I've significantly reduced memory overhead and improved training performance. This research was published in a top database conference, i.e., VLDB 2022.

Second, to further reduce the training cost, I proposed an efficient and accurate pruning-during-training framework for CNNs. Different from the existing pruning-during-training work, my new framework provides higher model accuracy and compression ratio via fine-grained architecture-preserving pruning. It can generate highly accurate and fast pruned CNN models for direct deployment without any time overhead. This research was published in a top HPC conference, i.e., ACM ICS 2021.

Third, to address the random data access problem caused by data shuffling during training, I proposed optimization strategies such as pre-generating data shuffling sequences, optimizing epoch and sample orders based on pre-generated shuffling sequences to enhance data reuse, trading computational balance for I/O balance, and aggregating small chunk reads to maximize I/O bandwidth utilization. This research has been submitted to a top parallel computing conference, i.e., ACM PPoPP 2024 (the pre-print version).

Ongoing and future work. My team is currently collaborating with Meta through my Meta Research Award to apply lossy compression techniques to deep learning recommendation model (DLRM) [8] training to reduce high all-to-all communication cost. DLRM is an emerging method for tasks in personalization and recommender systems, such as click-through rate (CTR) prediction. State-of-the-art DLRMs often employ large embedding tables to map high-dimensional sparse vectors from raw categorical features to low-dimensional dense vector representations. These tables are typically distributed across nodes (i.e., model parallelism) and require collective communication primitives (such as alltoall) to aggregate looked-up vectors and corresponding gradients during minibatch iterations. Existing work [11] has demonstrated that this synchronization places significant overhead on training latency and stresses network bandwidth, even with high-bandwidth interconnects. To address this, my team is working with Meta research scientists to reduce communication volume and increase throughput in DLRM using error-bounded lossy compression techniques. We are developing an offline analysis tool to configure error bounds for different embeddings by analyzing their lookup frequencies and the required reduction size to reduce communication latency based on the network environment. Additionally, we are developing new runtime APIs for collective communications with lossy compression. Based on our current preliminary results, this method can potentially provide a 5-10 times communication speedup while maintaining training accuracy and convergence speed.

My team is also collaborating with Prof. Tong Geng from the University of Rochester and Dr. Ang Li from Pacific Northwest National Lab to accelerate Graph Neural Networks (GNNs) training and inference using graph compression methods. GNNs have gained attention due to their ability to extract latent information from graph data and extend neural network approaches to various applications, such as drug and vaccine discovery, supply network crisis prediction, misinformation detection, and power grid failure prevention. However, a performance gap, primarily resulting from communication bottlenecks due to ultra-sparse, large, and irregular graphs, hinders GNNs' potential in real-world applications. To address these bottlenecks, my collaborators and I are developing a research project leveraging graph locality enhancement and compression techniques. Our primary objective is to create a collaborative locality-enhancement and compression co-design framework for GNN acceleration. The research comprises three thrusts:

- **Thrust 1:** Develop an on-the-fly graph locality enhancer through hardware-software co-design, significantly improving versatility and further reducing communication demands compared to current state-of-the-art methods.
- **Thrust 2:** Create an efficient lossy compressor that provides high-ratio, error-bounded compression for graph data, encompassing both graph embedding and topology information.
- **Thrust 3:** Investigate methods to effectively integrate the graph locality enhancer and graph compressor, allowing them to be mutually aware and benefit each other. These approaches can jointly and fundamentally address the persistent communication bottlenecks in GNNs, unleashing their potential for society.

Additionally, my team is also collaborating with Prof. Bo Yuan from Rutgers University to leverage tensor decomposition approaches for accelerating DNN training and inference. Tensor decomposition techniques, such as Tucker decomposition, are state-of-the-art CNN model compression methods. However, we observed only limited inference time reduction with Tucker-compressed models using existing GPU software like cuDNN. To address this, we proposed an efficient end-to-end framework generating highly accurate, compact CNN models through Tucker decomposition and optimized GPU inference code, with part of this research published in ACM PPoPP 2023. In addition, building on our prior VLDB work, Prof. Yuan and I are now developing a novel method combining various compression techniques, including Tucker decomposition, to achieve efficient GPU-based deep learning training through algorithm and software co-design.

2.3 Efficient and effective online algorithm-based fault tolerance techniques

As HPC systems scale and power consumption constraints become critical, their reliability declines. For example, the mean time between failures (MTBF) for the US Department of Energy's supercomputers has decreased significantly. This challenge makes it difficult for large-scale applications to maintain stable, long-term operation. Designing low-overhead, high-coverage fault-tolerant technologies to enhance HPC systems' reliability is essential.

To address this, I proposed a novel checksum encoding scheme for matrix/tensor multiplication, capable of detecting and correcting software errors. I developed online algorithm-based fault-tolerance (ABFT) techniques with low overhead and high coverage based on this novel checksum, exhibiting minimal time overhead for various scientific computing routines. The core technology has been published in top HPC conferences, including SC 2018, SC 2017, PPoPP 2017, and HPDC 2016.

Additionally, I applied lossy compression techniques to checkpoint/restart mechanisms, introducing the "lossy checkpoint/restart" concept and model. This technology leverages lossy compression to reduce checkpoint data access time overhead while controlling the impact of compression errors on overall performance. This research was published in a top HPC conference, i.e., HPDC 2018, and has attracted attention from the NWChem computational chemistry software team, who are attempting to apply this technology to their next-generation exascale application software, NWChemEX.

Ongoing and future work. Lossy compression effectively reduces data size, but introduces errors that can lead to incorrect computation results. Existing studies have not investigated error propagation in HPC applications or how to minimize error impact while maintaining computation validity. This has made scientists hesitant to use lossy compression, including the "lossy checkpoint/restart" approach. There is a critical need to develop a method for identifying a compression strategy with a high compression ratio and low error impact for HPC applications. I am collaborating with Prof. Guanpeng Li from the University of Iowa to model compression error propagation in HPC applications by integrating program analysis and machine learning (ML) techniques through my NSF OAC Core Award. Our primary goal is to automatically select the best-fit lossy compression strategy that minimizes error impact based on the target compression ratio. This ongoing research project includes four critical thrusts:

- **Thrust 1:** Develop an accurate, efficient fault injection infrastructure integrated with common lossy compression algorithm fault models.
- **Thrust 2:** Design a fine-grained approach to characterize error propagation in HPC programs through program analysis and deposition based on compressed data dependencies and life cycles.
- **Thrust 3:** Develop an ML-based predictive model to select a compression strategy that minimizes error impact on a given program and compression ratio.
- **Thrust 4:** Integrate our technique with domain-specific error impact metrics in real-world applications and demonstrate effectiveness by selecting low error impact compression strategies for the same ratios.

In my future work, I plan to conduct research on energy-efficient and fault-tolerant computing, specifically exploring the interaction and integration strategies among approximate computing, fault-tolerant computing, and energy-efficient computing technologies. For example: (1) I'll investigate mixed-precision checkpoint/restart techniques, the impact of precision loss on applications, error propagation rules, and control strategies. I aim to expand this technology's application in HPC and cloud computing scenarios and optimize it for specific applications. (2) Reducing data movement can significantly decrease power and energy consumption. While my prior data compression research primarily focused on improving program performance, I'll further explore the combination of approximate computing technology and energy-efficient computing. (3) I'll establish a power consumption and fault-tolerance model for large-scale systems to deeply understand their interaction, develop a method to collaboratively manage system power consumption and fault tolerance, and conduct load-aware driven power management and fault-tolerance technology research based on application characteristics. I strive to cooperate with university HPC centers or DOE leadership computing facilities to deploy these research achievements to multiple supercomputers and deeply optimize large-scale applications with national strategic significance and extensive economic benefits. My long-term exploration in data compression, fault-tolerant computing, and energy-efficient heterogeneous computing technologies has laid a solid foundation for in-depth research in this field.

3 Future funding opportunities

Federal funding: I haveextensive NSF and DOE funding through 2027 and will continue seeking suitable opportunities in the Office of Advanced Cyberinfrastructure (OAC) and the Advanced Scientific Computing Research (ASCR). I'll also partner with collaborators for multiple-PI projects via CSSI, PPoSS, POSE, and other relevant programs. I'll continue refining SZ software, supporting users, and developing training materials, positioning my lab for federal funding as national software infrastructure. In addition, I'll seek funding for AI-accelerated scientific research and computing software and systems.

Strategic partnerships with federal agencies: I havestrong collaborations with multiple DOE national labs, including Argonne, Los Alamos, Oak Ridge, and Pacific Northwest. I've published 50+ collaborative papers and secured over $700K in funding from these labs. My six Ph.D. students have interned or worked with these labs on funded projects. I'm also part of a new IU-ORNL initiative to increase institutional collaboration. I anticipate ongoing and additional funding opportunities with these partners in the future.

Industry and foundation funding: My lab has successfully obtained industry funding. I'll maintain and build relationships with companies like Microsoft Research and Meta, focusing on AI software and systems. One student, Chengming Zhang, worked at Microsoft Research on high-performance recommender systems and efficient training frameworks for scientific ML models. I also collaborate with the IU Corporate and Foundation Relations Office to identify funding opportunities and engage with major donors (e.g., Lilly Endowment, Amazon Web Services) to benefit my lab.

4 Closing thoughts

In the past five years, my research program has rapidly expanded in scope and impact. I've developed innovative open-source HPC system software and tackled fundamental AI problems, especially in large-scale deep learning. My work, widely adopted by the community, has received new capabilities, data, and expertise. Collaborating with my extensive network, I'm working to launch the "systems for scientific ML" initiative alongside DOE and other partners, aiming to create emerging computer systems for AI and ML to advance scientific understanding.

TEACHING STATEMENT

1 Core philosophy and approach

Over the past five years, as a tenure-track faculty member at three U.S. R1 land-grant research universities (IU, WSU, and UA), I've dedicated myself to integrating my research with course development. My teaching efforts are carried out in a **symbiotic** fashion with my research activities. I use my research outcomes to develop novel educational materials, while my teaching outcomes provide invaluable feedback and prepare a significant workforce for further research innovation.

I've successfully integrated advanced cyberinfrastructure (CI) concepts and usage into the computing curriculum. Specifically, in the HPC course, I covered data storage infrastructures and created a project on testing and analyzing the performance of popular parallel I/O libraries (e.g., HDF5, ADIOS). Based on course project performance, I found that students gained a deeper understanding of utilizing CI tools to improve large-scale application performance. Motivated by this success, I plan to extend and evaluate the "integration of concepts and use of advanced CI" in my future graduate and undergraduate education.

2 Key statistics

In my one year at IU (August 2022-July 2023):

- I've revised and taught a course, "Engineering Cloud Computing," for both undergraduate and graduate students in the Intelligent Systems Engineering (ISE), Computer Science (CS), and Data Science (DS) programs at IU, with a total enrollment of over **150 students**.
- I'm graduating my first PhD student, Sian Jin, in Fall 2023. He received **tenure-track Assistant Professor** offers from six U.S. R1 universities and will start his position at Temple University in 01/2024.
- I'm currently the primary advisor of **7 PhD students** in ISE, with an additional PhD student set to start in Fall 2023, all of whom are fully funded by my federal grants.
- I've served as a PhD dissertation research committee member for one ISE student.
- I'm a member of the PhD advisory committee (a.k.a, qualifying exam committee) for two ISE students.
- I've mentored early career researchers, such as Dr. Guanpeng Li at the University of Iowa, Dr. Tong Teng at University of Rochester, Dr. Xin Liang at University of Kentucky, and Dr. Kai Zhao at University of Alabama at Birmingham, through research collaborations and proposal development efforts.

In my 4 years at WSU and UA (August 2018-August 2022):

- I revised and taught two undergraduate computer science (CS) courses, "Introduction to Computer Networks" and "Advanced Data Structures," at WSU.
- I developed a new CS course, "High Performance Computing," and taught another course, "Computer Algorithms," for senior undergraduate and graduate students at UA.
- I recruited and advised 3 PhD students at UA and 4 PhD students at WSU.
- I regularly served on PhD dissertation and qualifying exam committees.

3 Teaching (including new courses)

Over the past five years, I've developed and taught five CS/CE courses for (under)graduate students:

- **ENGR 516: Engineering Cloud Computing**, IU (Fall 2022, Spring 2023)
  To accommodate students with diverse backgrounds in CS, ISE, and DS, I've significantly revised the curriculum to focus on hands-on experience and practical skills. The course covers traditional cloud computing topics like hardware virtualization and distributed resource management and includes four programming labs, allowing students to gain proficiency in popular big data programming models like MapReduce, Hadoop, and Spark. Students also explore efficiency and scalability on platforms like AWS and IU Jetstream2. Additionally, there's a semester-long collaborative exploration project where groups investigate distributed systems issues or design new applications using popular cloud platforms. Students submit proposals, mid-term reports, and final reports, and present their final project in class.
- **CS 233 Advanced Data Structures**, WSU (Spring 2021, Fall 2021)

Besides basic data structures, this course covers algorithm design using various data structures and algorithmic techniques, as well as algorithm analysis employing asymptotic notation and complexity analysis. Assignments and projects are designed to encourage students to devise multiple algorithmic solutions to a single problem using different combinations of data structures learned in class.

- **CS 455: Introduction to Computer Networks**, WSU (Fall 2020, Fall 2021)
  I've significantly revised this course to incorporate materials on popular interconnect technologies (e.g., interconnect topology, Ethernet, InfiniBand) and a project focused on benchmarking and analyzing the communication/networking performance with interconnects used in various HPC systems (such as lab cluster, university cluster, and supercomputer).
- **CS 470/570: Computer Algorithms**, UA (Fall 2019)
  This course covers the design and analysis of parallel algorithms, as well as data compression algorithms. It also covers topics in divide-and-conquer, dynamic programming, greedy method, max flow and matchings, string matching, computational geometry, NP-completeness, approximation algorithms.
- **CS 481/581: High Performance Computing**, UA (Spring 2019, Spring 2020)
  In this course, students learn principles of parallel algorithm design, shared-memory systems programming (OpenMP), parallel computer architectures, distributed-memory systems programming (MPI), analytical modeling of program performance, performance analysis of parallel programs, and GPU programming (CUDA). The course can also be referred to as parallel computing.

Ongoing and future work. I'm currently developing the following two new courses:

- The first course under development is a new graduate-level course, **Big Data Reduction**, closely tied to my research expertise. It focuses on data reduction techniques and tools for scientific applications, covering classic compression algorithms, dimensionality reduction methods, and emerging scientific data compressors. Additionally, it addresses the design and use of data reduction and data quality assessment methods in real-world applications across various domains. The course will appeal to STEM students seeking solutions to big data challenges. Teams of two will choose a final project related to data reduction, enhancing their understanding and hands-on experience with related tools. High-quality research from projects may be submitted to top-tier systems or big data conferences.
- The second course under development is an enhanced undergraduate-level course, **Parallel Computing**. While current syllabi focus on topics like parallel algorithm design, MPI programming, and parallel operations, I plan to incorporate GPU programming and GPU library usage in scientific computing and machine learning, eliminating detailed OpenMP discussions. Furthermore, I'll add popular topics like parallel I/O, GPU compression, and direct storage, emphasizing efficient CI tool usage.

4 Graduate student mentoring

At IU (August 2022-present): I'm currently supervising 7 funded PhD students at IU ISE (listed below). Each student typically starts training on an existing project before transitioning to lead a new project. All students regularly present their work at lab meetings, IU research events, and scientific meetings.

- **Sian Jin** (my first PhD student) has developed efficient data reduction techniques for HPC and ML applications, co-authored 20 publications, completed his PhD dissertation proposal, and received tenure-track Assistant Professor offers from six U.S. R1 universities. He'll join Temple University from 01/2024.
- **Jiannan Tian** has collaborated with Argonne since 2019, leading GPU compression R&D for HPC applications. He has co-authored 19 publications and plans to submit his dissertation proposal in 06/2023.
- **Chengming Zhang** has focused on algorithm-hardware co-design for efficient ML systems. He has co-authored 10 publications and plans to complete his PhD dissertation proposal in 08/2023.
- **Daoce Wang** designs data reduction algorithms and software for scientific applications, leading the NSF CDS&E project. He has published two top-tier conference papers and has a manuscript under review.
- **Baixi Sun** studies the performance of scalable DNN training on HPC systems. He is now exploring lossy compression techniques for distributed 2nd-order optimizations in DNN training.
- **Boyuan Zhang** trained in GPU programming and leads the NSF CSSI project. He has published two papers on GPU compression and is working on efficient quantum simulations, in partnership with PNNL.
- **Hao Feng** leads the Meta research project to reduce communication volume in training DLRMs. He is in the process of implementing a compression solution and will begin working on his first research paper.

At UA and WSU (through August 2022): At UA, I advised three PhD students: Sian Jin, Jiannan Tian, and Chengming Zhang at UA. At WSU, I continued to advise Sian Jin, Jiannan Tian, Chengming Zhang, and also took on Daoce Wang, Baixi Sun, Boyuan Zhang, and Xinyu Chen. Additionally, I served as a dissertation committee member for two EECS students (Xiaoqin Fu, Devjeet Roy) at WSU, one Civil Engineering student (Peyman Abbaszadeh) at UA, and two CS students (Nasir U. Eisty, Tasnuva Mahjabin) at UA.

Ongoing and future work. An additional PhD student, Jiangfan Ye (graduated from University of Science and Technology of China), will join my lab in Fall 2023. My PhD students Sian Jin and Jiannan Tian are expected to defend their PhD theses in Fall 2023. My other two PhD students (Baixi Sun and Boyuan Zhang) will work on their qualifying exams in Fall 2023. Although Baixi has already passed their exams at WSU, he needs to retake them at IU. Hao Feng will concentrate on his first-authored paper in Summer 2023 and aims to pass his qualifying exam in Fall 2023 as well. Each student should have at least 3-4 first-authored high-quality papers published or completed before defending their PhD theses.

5 Undergraduate student mentoring

Since 2018, I've advised 13 undergraduates, including two female students and a Hispanic student. They conducted research under my mentorship, resulting in several high-quality publications, with three featuring undergraduates as first authors. I've recruited undergraduates through various programs and social media. I partnered with UA's RRSP program and WSU's LSAMP and TMP programs, and now work with IU's UROC program, engaging two undergraduate students in my NSF-funded projects.

Engagement with underrepresented minority groups. I have extensive experience in mentoring undergraduate students from underrepresented minority groups. For instance, I mentored Cody Rivera, a Hispanic undergraduate, on two GPU-related projects, helping him establish a strong research and publication record. Cody's TSM2X library drew attention from ORNL scientists and is being ported for the first U.S. exascale supercomputer. Cody was selected as a DOE SULI intern for Summer 2021 and has one publication in JPDC and one in IEEE IPDPS. Due to his exceptional performance, Cody received multiple offers from prestigious PhD programs and began his studies at UIUC in Fall 2022. I collaborate with Cody to share his success story and attract more URM students to join my group and engage in HPC research.

Ongoing and future work. I'll persist in engaging with URM groups in STEM, especially women, African American, and Hispanic students, as many undergraduates transfer from community colleges to research universities for computer science studies. I plan to apply for REU supplements for my NSF projects and recruit 2-3 URM students annually through LSAMP and UROC programs at IU. My focus will be on those interested in HPC research and considering graduate school. The Luddy School's recognition for its commitment to recruiting and retaining women in undergraduate computing programs will help me attract more female students to my research group.

6 Other education activities

Engagement with K-12 studnets. I've consistently reached out to K-12 students to increase interest in computing and STEM disciplines. At UA CS, I organized seminars on "Introduction to High-Performance Computing" for over 80 high school students attending the 2019 Alabama Summer Computer Science Camps. I recruited three high school students and mentored them for a year on research projects. My long-term vision includes organizing HPC seminars in Summer Camps, educating diverse high school students, and selecting a subset for year-round mentoring in my research group.

Training program for scientists & engineers. I've actively collaborated with scientists from multiple national labs, resulting in over 40 publications. I plan to organize a training program for researchers from universities and national labs, teaching them to use software products from my research projects. This will help advance data management in advanced CI and drive R&D in data reduction methods and tools.

Ongoing and future work. I plan to collaborate with IU Pervasive Technology Institute (PTI) to apply for NSF CyberTraining program funding. Our objective is to develop innovative, scalable training and education materials for the research workforce in advanced CI-enabled research. This goal aligns with PTI's workforce development initiatives, supporting various current and future scholarly and industry needs.

SERVICE STATEMENT

Core philosophy. In my opinion, service is an opportunity to build robust relationships and partnerships - both within the department and between the department and the university and research community. Service is a vehicle for collaboration, learning, knowledge sharing and creativity. I strongly believe that my commitment and willingness for service were established and developed due to my cultural background and through my extensive experience in volunteer and professional work. In that respect, service comes natural within my personality – especially when it comes to introducing initiatives, leading change and projects and collaborating within teams. I believe that any committee should be focused on achieving incremental targets that lead to its goal. Moreover, committees or teams should be viewed as vicinities for building networks and increasing personal development and exposure.

Departmental service. I have held multiple service roles in the departments with which I have been affiliated, including UA CS, WSU EECS, and IU ISE. For instance, I served as the **faculty search committee chair** at IU ISE in 2022-2023. As the chair, I was responsible for coordinating the search process, screening candidates, and organizing interviews to ensure the selection of the most suitable faculty members for the department. The committee helped shortlist 10 faculty candidates for two assistant professor positions. I hosted four candidate visits, which involved coordinating their talks, arranging meetings with faculty and students, and organizing campus tours. I also actively participated in the CS faculty hiring process in 2022-2023, meeting five CS faculty candidates and providing my professional feedback, particularly related to computer systems research. In addition, I also served the **faculty mentor** for the newly established HPC Club for IU ISE/CS undergraduate students.

I served as a member of faculty search committees at UA in 2019 and WSU in 2021. Furthermore, I was a member of the **CS curriculum committee** at WSU in 2021-2022, reviewing curriculum-related requests from undergraduate students, particularly those from transfer students. Moreover, I was a member of the **graduate student committee**, discussing graduate study-related topics such as recommending publication venues for WSU CS students.

School/University service. I view university service as an opportunity to build networks, increase department visibility, advocate for needs, and foster appreciation for other disciplines. I've represented my department in school/university-level services, such as collaborating with IU PTI and establishing the WSU Quantum Initiative. I participated in the multi-college effort to create the iSciMath program, training students in mathematical and computational tools for data-intensive problems. Additionally, I was involved in discussing the "One WSU" principles and development plan, aiming to deliver educational, research, and outreach benefits statewide.

Professional Service. I have actively participated in and contributed to the HPC community by organizing and serving in various roles at numerous international conferences and workshops, as detailed below.

- I served as the *program co-chair* for the IEEE ScalCom 2021 conference, DRBSD 2023 workshop, IWBDR 2022, 2021, 2020 workshops. I acted as the session chair for HPDC 2022, ICS 2022, IPDPS 2021, ICS 2021, and ICPP 2020, and was a member of the organizing committee for SC 2021. I participated in the program committees of SC 2023, 2022, 2020, HPDC 2023, 2022, ICS 2023, IPDPS 2023, 2021, HiPC 2022, 2021, 2019, 2018, Cluster 2023, 2021, 2020, BigData 2021-2023, ICMLA 2021-2023, NPC 2018-2021, and ICPP 2023, 2020, among others (see my CV for the full list).
- I have been serving on the *Technical Review Board* of IEEE Transactions on Parallel and Distributed Systems (TPDS), which complements its editorial board. The goals of the TRB are to provide high-quality and timely reviews for TPDS submissions and to build, support, and grow the next generation of leaders, such as editors and associate editors.
- I served as a *reviewer* for numerous impactful journals, such as IEEE TPDS, TDSC, TKDE, TCC, TBD, TSG, TETC, TSUSC, IEM, Access, Journal of Supercomputing, SIAM SISC, Journal of Systems Architecture, and more. Detaileds about my journal review record can be found at this website.

Supporting junior faculty. In addition, I have actively supported junior faculty members (e.g., Dr. Guanpeng Li from the University of Iowa, Dr. Tong Geng from the University of Rochester, Dr. Xin Liang from the University of Kentucky, Dr. Kai Zhao from the University of Alabama at Birmingham) by advising them on proposal writing, teaching, and student mentoring, and involving them in my research collaborations, such as those with DOE national labs.

Ongoing and future work. For departmental services, I will continue to support ISE's growth in both faculty and student numbers. I plan to remain actively involved in faculty hiring committees to build a strong engineering program at IU and participate in undergraduate and graduate student recruitment events to attract more domestic and international students. For school/university services, I aim to contribute to the newly established Luddy AI Center by fostering more discussions and collaborations, such as inviting talks and organizing campus visits with prominent tech companies focused on building AI infrastructure. For professional services, I am currently seeking opportunities to serve as an associate editor for influential parallel and distributed computing journals such as IEEE TPDS, Parallel Computing, and Future Generation Computer Systems. Additionally, I am actively pursuing roles as a program committee chair or track chair at prestigious conferences such as ACM/IEEE Supercomputing (SC) and IEEE IPDPS. Furthermore, I am in discussions with collaborators at the University of Notre Dame (with Prof. Douglas Thain) and IUPUI (with Prof. Fengguang Song) to bring esteemed conferences, such as ACM HPDC and IEEE IPDPS, to Indianapolis in the coming years.

In conclusion, service should be holistic, encompassing departmental, organizational, communal, and professional aspects. It enables academics to significantly impact their environment. My active engagement in service consistently offers learning, knowledge-sharing, networking, and relationship-building opportunities, leading to overall job satisfaction. I'm eager to expand and improve my service contributions within IU, Luddy School, and ISE department and the broader research community.

References

[1] Franck Cappello, Sheng Di, Sihuan Li, Xin Liang, Ali Murat Gok, Dingwen Tao, Chun Hong Yoon, Xin-Chuan Wu, Yuri Alexeev, and Frederic T Chong. Use cases of lossy compression for floating-point data in scientific data sets. *The International Journal of High Performance Computing Applications*, 33(6):1201–1220, 2019.

[2] Sébastien Cayrols, Jiali Li, George Bosilca, Stanimire Tomov, Alan Ayala, and Jack Dongarra. Lossy all-to-all exchange for accelerating parallel 3-d ffts on hybrid architectures with gpus. In *2022 IEEE International Conference on Cluster Computing (CLUSTER)*, pages 152–160. IEEE, 2022.

[3] Wai Tong Chung, Ki Sung Jung, Jacqueline H Chen, and Matthias Ihme. Blastnet: A call for community-involved big data in combustion machine learning. *Applications in Energy and Combustion Science*, 12:100087, 2022.

[4] Alec M Dunton, Lluís Jofre, Gianluca Iaccarino, and Alireza Doostan. Pass-efficient methods for compression of high-dimensional turbulent flow data. *Journal of Computational Physics*, 423:109704, 2020.

[5] Ali Murat Gok, Sheng Di, Yuri Alexeev, Dingwen Tao, Vladimir Mironov, Xin Liang, and Franck Cappello. Pastri: Error-bounded lossy compression for two-electron integrals in quantum chemistry. In *2018 IEEE International Conference on Cluster Computing (CLUSTER)*, pages 1–11. IEEE, 2018.

[6] Duong Hoang, Pavol Klacansky, Harsh Bhatia, Peer-Timo Bremer, Peter Lindstrom, and Valerio Pascucci. A study of the tradeoff between reducing precision and reducing resolution for data analysis and visualization. *IEEE transactions on visualization and computer graphics*, 25(1):1193–1203, 2018.

[7] Sian Jin, Jesus Pulido, Pascal Grosset, Jiannan Tian, Dingwen Tao, and James Ahrens. Adaptive configuration of in situ lossy compression for cosmology simulations via fine-grained rate-quality modeling. In *Proceedings of the 30th International Symposium on High-Performance Parallel and Distributed Computing (HPDC)*, pages 45–56, 2021.

[8] Maxim Naumov, Dheevatsa Mudigere, Hao-Jun Michael Shi, Jianyu Huang, Narayanan Sundaraman, Jongsoo Park, Xiaodong Wang, Udit Gupta, Carole-Jean Wu, Alisson G Azzolini, et al. Deep learning recommendation model for personalization and recommendation systems. *arXiv preprint arXiv:1906.00091*, 2019.

[9] Andrew Poppick, Joseph Nardi, Noah Feldman, Allison H Baker, Alexander Pinard, and Dorit M Hammerling. A statistical analysis of lossily compressed climate model data. *Computers & Geosciences*, 145:104599, 2020.

[10] Xin-Chuan Wu, Sheng Di, Emma Maitreyee Dasgupta, Franck Cappello, Hal Finkel, Yuri Alexeev, and Frederic T Chong. Full-state quantum circuit simulation by using data compression. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–24, 2019.

[11] Jie Amy Yang, Jongsoo Park, Srinivas Sridharan, and Ping Tak Peter Tang. Training deep learning recommendation model with quantized collective communications. In *Conference on Knowledge Discovery and Data Mining (KDD)*, 2020.

[12] Kai Zhao, Sheng Di, Maxim Dmitriev, Thierry-Laurent D Tonellot, Zizhong Chen, and Franck Cappello. Optimizing error-bounded lossy compression for scientific data by dynamic spline interpolation. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pages 1643–1654. IEEE, 2021.

[13] Kai Zhao, Sheng Di, Xin Liang, Sihuan Li, Dingwen Tao, Zizhong Chen, and Franck Cappello. Significantly improving lossy compression for hpc datasets with second-order prediction and parameter optimization. In *Proceedings of the 29th International Symposium on High-Performance Parallel and Distributed Computing (HPDC)*, pages 89–100, 2020.