Dong Dai, Ph.D

Assistant Professor
Department of Computer Science
University of North Carolina at Charlotte
9201 University City Blvd, Charlotte, NC

Phone: 704-687-1978
Email: ddai@uncc.edu
Homepage: http://webpages.uncc.edu/ddai/

Research Interests

- I am interested in a wide range of topics related to build and optimize high-performance data-intensive systems, such as job and I/O scheduler, parallel file systems, metadata management, graph storage, and machine learning infrastructure.

Education

- Ph.D. Computer Science, University of Science and Technology of China, 2013.

  Thesis: *Research and Implementation on Cloud Software Infrastructure*. Advisor: Prof. Xuehai Zhou.
- B.S. Computer Science, University of Science and Technology of China, 2006.

Professional Experience

- **Assistant Professor** 2018 – Current

  Computer Science Department, University of North Carolina at Charlotte
- **Research Assistant Professor** 2016 - 2018

  Computer Science Department, Texas Tech University
- **Post-doctoral Researcher** 2013 - 2016

  Texas Tech University and Argonne National Lab.

Research Projects

- **Active Projects**

  PI, National Science Foundation, CNS 2020 – 2023
  Moving Machine Learning into the Next-Generation Cloud Flexibly, Agilely and Efficiently $471,808

  PI, National Science Foundation, SHF 2019 – 2023
  A Hybrid NVM based Computing Architecture for Machine Learning Applications $497,576

  PI, National Science Foundation, SHF 2019 – 2022
  A Parallel Graph-Based Paradigm for HPC Parallel File System Checkers $307,682
  Collaborator: Mai Zheng (Iowa State University)

  Co-PI, National Science Foundation, OAC 2018 – 2021
  Empowering Data-driven Discovery with a Provenance Collection, Management, and Analysis Software Infrastructure $599,982

PI: Yong Chen, Co-PI: William Hase, Brian Ancell (Texas Tech University),
sub-contract pending from Texas Tech University

- **Past Projects**

Single PI, National Science Foundation, CRII 2018 – 2022
Partitioning Large Graphs in Deep Storage Architecture $168,201

Co-PI, National Science Foundation, CNS 2018 – 2022
Tuning Extreme-scale Storage Stack through Deep Reinforcement Learning $240,026
PI: Forrest Sheng Bao(Iowa State University), Yong Chen (Texas Tech University),
$30,000 subcontracted from Iowa State University

Co-PI, National Science Foundation, CCF 2017 – 2022
Uncovering Vulnerabilities in Parallel File Systems for Reliable HPC $233,000
PI: Yong Chen (Texas Tech University),
sub-contract pending from Texas Tech University

Selected Publications

Names with (*) are the Ph.D. students I mentored; (†) are master or undergraduate students I mentored.

[1] Di Zhang*, Chris Egersdoerfer†, Tabassum Mahmud, Mai Zheng, **Dong Dai**. Drill: Log-based Anomaly Detection for Large-scale Storage Systems Using Source Code Analysis. *Accepted to appear in 37th IEEE International Parallel & Distributed Processing Symposium (**IPDPS'23**), 2023*

[2] Saisha Kamat*, Abdullah Al Raqibul Islam*, Mai Zheng, **Dong Dai**. FaultyRank: A Graph-based Parallel File System Checker. *Accepted to appear in 37th IEEE International Parallel & Distributed Processing Symposium (**IPDPS'23**), 2023*

[3] Abdullah Al Raqibul Islam*, **Dong Dai**, A Framework for Large Dynamic Graph Analysis on Persistent Memory, *Accepted to appear in 21st USENIX Conference on File and Storage Technologies (**FAST'23 WiP**), 2023*

[4] Duo Zhang, Om Rameshwar Gatla, Abdullah Al Raqibul Islam*, **Dong Dai**, Mai Zheng, On the Scalability of Testing the Crash Consistency of PM Systems, *Accepted to appear in 21st USENIX Conference on File and Storage Technologies (**FAST'23 WiP**), 2023*

[5] Chris Egersdoerfer†, Di Zhang*, **Dong Dai**. ClusterLog: Clustering Logs for Effective Log-based Anomaly Detection. *Accepted in Workshop on Fault Tolerance for HPC at eXtreme Scale at SC'22 (**FTXS'22**), 2022*

[6] Abdullah Al Raqibul Islam*, Christopher York†, **Dong Dai**. A performance study of Optane persistent memory: from storage data structures' perspective. *Accepted in CCF Transactions on High Performance Computing (**THPC'22**), 2022*

[7] Di Zhang*, **Dong Dai**, Bing Xie. SchedInspector: A Batch Job Scheduling Inspector Using Reinforcement Learning. *Accepted to appear in the 31st International ACM Symposium on High-Performance Parallel and Distributed Computing (**HPDC'22**), 2022*

[8] Abdullah Al Raqibul Islam*, **Dong Dai**, Dazhao Cheng. VCSR: Mutable CSR Graph Format Using Vertex-Centric Packed Memory Array. *Accepted to appear in The 22nd IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (**CCGrid'22**), 2022*

[9] Runzhou Han, Om Rameshwar Gatla, Mai Zheng, Jinrui Cao, Di Zhang*, **Dong Dai**, Yong Chen, Jonathan Cook. A Study of Failure Recovery and Logging of High-Performance Parallel File Systems. *Accepted to appear in ACM Transactions on Storage (**TOS'22**), 2022*

[10] Dazhao Cheng, Yu Wang, **Dong Dai**. Dynamic Resource Provisioning for Iterative Workloads on Apache Spark. *IEEE Transactions on Cloud Computing (**TCC'21**), 2021*

[11] Di Zhang*, **Dong Dai**, Runzhou Han, Mai Zheng. SentiLog: Anomaly Detecting on Parallel File Systems via Log-based Sentiment Analysis. *Accepted to apear in 13th ACM Workshop on Hot Topics in Storage and File Systems (**HotStorage'21**), 2021* **Best Paper Nominee**

[12] Jiang Zhou, Yong Chen, **Dong Dai**, Yu Zhuang, Weiping Wang. I/O characteristic discovery for storage system optimizations. *Journal of Parallel and Distributed Computing (**JPDC'21**), Vol 148, Pages 1-13, 2021*

[13] Di Zhang*, **Dong Dai**, Youbiao He, Forrest Sheng Bao, and Bing Xie. RLScheduler: An Automated HPC Batch Job Scheduler Using Reinforcement Learning. *in the proceeding of the International Conference for High Performance Computing, Networking, Storage and Analysis (**SC'20**), 2020.* (acceptance rate: 22.3%).

[14] Abdullah Al Raqibul Islam*, Anirudh Narayanan, Christopher York†, and **Dong Dai**. A Performance Study of Optane Persistent Memory: From Indexing Data Structures' Perspective. *Accepted to appear in the 36th International Conference on Massive Storage Systems and Technology (**MSST'20**), 2020.*

[15] Abdullah Al Raqibul Islam*, and **Dong Dai**. Understand the Overheads of Storage Data Structures on Persistent Memory (Poster). *In 25th ACM SIGPLAN Symposium on Principle and Practice of Parallel Programming (**PPoPP'20**), 2020.*

[16] Jiang Zhou, Yong Chen, Wei Xie, **Dong Dai**, Shuibing He, and Weiping Wang. PRS: A Pattern-Directed Replication Scheme for Heterogeneous Object-Based Storage. *IEEE Transactions on Computers (**TC'19**), 2019.*

[17] **Dong Dai**, Om Rameshwar Gatla, and Mai Zheng. A Performance Study of Lustre File System Checker: Bottlenecks and Potentials. *Accepted to appear in the proceedings of the 35th International Conference on Massive Storage Systems and Technology (**MSST'19**), 2019.*

[18] **Dong Dai**, Yong Chen, Philip Carns, John Jenkins, Wei Zhang, and Robert Ross. Managing Rich Metadata in High-Performance Computing Systems Using a Graph Model. *Accepted to appear in the IEEE Transactions on Parallel and Distributed Systems (**TPDS'18**), 2018.*

[19] Jinrui Cao, Om Rameshwar Gatla, Mai Zheng, **Dong Dai**, Vidya Eswarappa, Yan Mu and Yong Chen. PFault: A General Framework for Analyzing the Reliability of High-Performance Parallel File Systems. *Accepted to appear in the proceedings of the 32nd ACM/SIGARCH International Conference on Supercomputing (**ICS'18**), 2018.* (acceptance rate: 18.7%).

[20] **Dong Dai**, Yong Chen, Philip Carns, John Jenkins, and Robert Ross. Lightweight Provenance Service for High Performance Computing. *In Proceedings of the 26th International Conference on Parallel Architectures and Compilation Techniques (**PACT'17**), 2017.* (acceptance rate: 23%).

[21] **Dong Dai**, Wei Zhang, and Yong Chen. IOGP: An Incremental Online Graph Partitioning Algorithm for Distributed Graph Databases. *In Proceedings of the 26th ACM International Symposium on High Performance Parallel and Distributed Computing (**HPDC'17**), 2017.* (acceptance rate: 19%).

[22] **Dong Dai**, Yong Chen, Dries Kimpe, and Robert Ross. Two-Choice Randomized Dynamic I/O Scheduler for Object Storage Systems. *In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (**SC'14**), 2014.* (acceptance rate: 82/394=20.8%).

[23] **Dong Dai**, Yong Chen, Dries Kimpe, and Robert Ross. Trigger-based Incremental Data Processing with Unified Sync and Async Model. *Accepted to appear in Transaction on Cloud Computing (**TCC'18**)*

[24] **Dong Dai**, Forrest Sheng Bao, Jiang Zhou, Xuanhua Shi, and Yong Chen. Vectorizing Disk Blocks for Efficient Storage Systems via Deep Learning. *International Journal of Parallel Computing (**ParCo'18**).*

[25] **Dong Dai**, Phil Carns, Robert Ross, John Jenkins, Nicholas Muirhead, and Yong Chen. An Asynchronous Traversal Engine for Graph-Based Rich Metadata Management. *International Journal of Parallel Computing (**ParCo'16**)*.

Other Conference/ Journal Publications

[26] Jiang Zhou, **Dong Dai**, Yu Mao, Xin Chen, Yu Zhuang, and Yong Chen. I/O Characteristics Discovery in Cloud Storage Systems. *In Proceedings of the 11th International Conference on Cloud Computing (**CLOUD'18**), 2018*.

[27] Wei Zhang, **Dong Dai**, and Yong Chen. AKIN: A Streaming Graph Partitioning Algorithm for Distributed Graph Storage Systems. *In Proceedings of the 18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (**CCGrid'18**), 2018*. (acceptance rate: 20.8%).

[28] **Dong Dai**, Wei Zhang, and Yong Chen. IOGP: An Incremental Online Graph Partitioning for Large-Scale Distributed Graph Databases. *In Proceedings of the 22nd ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (**PPoPP'17**), 2017. (Short Paper)*.

[29] Jiang Zhou, Wei Xie, **Dong Dai**, and Yong Chen. Pattern-Directed Replication Scheme for Heterogeneous Object-based Storage. *In Proceedings of the 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (**CCGrid'17**), 2017*.

[30] Chao Wang, **Dong Dai**, Xi Li, Aili Wang, and Xuehai Zhou. SuperMIC: Analyzing Large Biological Datasets in Bioinformatics with Maximal Information Coefficient. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (**TCBB'17**), 2017*.

[31] **Dong Dai**, Yong Chen, Phil Carns, John Jenkins, Wei Zhang, and Robert Ross. GraphMeta: A Graph-based Engine for Managing Large-Scale HPC Rich Metadata. *In Proceedings of the IEEE International Conference on Cluster Computing (**CLUSTER'16**), 2016*. (acceptance rate: 39/162=24%).

[32] Jinrui Cao, Simeng Wang, **Dong Dai**, Mai Zheng, and Yong Chen. A Generic Framework for Testing Parallel File Systems. *In Proceedings of the Joint International Workshop on Parallel Data Storage and Data Intensive Scalable Computing Systems held in conjunction with SC'16 (**PDSW-DISCS'16**), 2016*.

[33] **Dong Dai**, Forrest Sheng Bao, Jiang Zhou, and Yong Chen. Block2Vec: A Deep Learning Strategy on Mining Block Correlations in Storage Systems. *In Proceedings of the 9th International Workshop on Parallel Programming Models and Systems Software for High-End Computing held in conjunction with ICPP'16 (**P2S2'16**), 2016*.

[34] **Dong Dai**, Phil Carns, Robert Ross, John Jenkins, Kyle Blauer†, and Yong Chen. GraphTrek: Asynchronous Graph Traversal for Property Graph Based Metadata Management. *In Proceedings of the IEEE International Conference on Cluster Computing (**CLUSTER'15**), 2015*. (acceptance rate: 38/157=24.2%).

[35] **Dong Dai**, Yong Chen, Dries Kimpe, and Robert Ross. Provenance-Based Object Storage Prediction Scheme for Scientific Big Data Applications. *In Proceedings of the 2014 IEEE International Conference on Big Data (**BigData'14**), 2014*. (acceptance rate: 49/264=18.6%).

[36] **Dong Dai**, Robert Ross, Philip Carns, Dries Kimpe, and Yong Chen. Using Property Graphs for Rich Metadata Management in HPC Systems. *In Proceedings of the 9th Parallel Data Storage Workshop held in conjunction with SC'14 (**PDSW'14**), 2014*.

[37] **Dong Dai**, Xuehai Zhou, Dries Kimpe, Robert Ross, and Yong Chen. Domino: An Incremental Computing Framework in Cloud with Eventual Synchronization. *In Proceedings of the 23rd ACM International Symposium on High-Performance Parallel and Distributed Computing (**HPDC'14**), 2014. (Short Paper)*.

[38] **Dong Dai**, Xi Li, Chao Wang, Junneng Zhang, and Xuehai Zhou. Detecting Associations in Large Dataset on MapReduce. *In Proceedings of the 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (**TrustCom'13**), 2013.*

[39] **Dong Dai**, Xi Li, Chao Wang, Mingming Sun, and Xuehai Zhou. Sedna: A Memory Based Key-Value Storage System for Realtime Processing in Cloud. *In Proceedings of the 2012 IEEE International Conference on Cluster Computing Workshops (**CLUSTER WORKSHOPS'12**), 2012.*

[40] **Dong Dai**, Xuehai Zhou, Feng Yang, and Chao Wang. An Auto-configuration Tool for Heterogeneous Hadoop Cluster. *Journal of the Graduate School of Chinese Academy of Sciences, 2012.* (Chinese)

Conference Posters

[41] Neda Tavakoli, **Dong Dai**, John Jenkins, Philip Carns, Robert Ross, and Yong Chen. A Software-Defined Approach for QoS Control in High-Performance Computing Storage Systems. *SC'16, Salt Lake City, UT, 2016.*

[42] **Dong Dai**, Robert Ross, Dounia Khaldi, Yonghong Yan, Matthieu Dorier, Neda Tavakoli, and Yong Chen. Exploiting Locality in Scientific Workflow System: A Cross-Layer Solution. *SC'16, Salt Lake City, UT, 2016.*

[43] **Dong Dai**, Yong Chen, Dries Kimpe, and Robert Ross. Provenance-Based Prediction Scheme for Object Storage System in HPC. *CCGrid'14, Chicago, IL, 2014.*

[44] Kun Lu, **Dong Dai**, and Mingming Sun. HDFS+: Concurrent Writes Improvements for HDFS. *CCGrid'13, Delft, Netherlands, 2013.*

[45] **Dong Dai**, Xi Li, Chao Wang, and Xuehai Zhou. Cloud Based Short Read Mapping Service. *CLUSTER'12, Beijing, China, 2012.*

Teaching

University of North Carolina at Charlotte, Charlotte, NC

- ITSC 3181 Introduction to Computer Architecture — 2020 Fall, 2021 Spring, 2021 Fall, 2022 Spring
- ITCS 5145 Parallel Computing — 2019-22 Spring & Fall
- ITCS 6144/8144 Operating Systems Design — 2018-19 Fall

Texas Tech University

- CS4352: Operating Systems — 2016
- CS5352: Advanced Operating Systems Design — 2017, 2015
- CS5331: Big Data Infrastructure and Data Management — 2014

University of Science and Technology of China

- Parallel Algorithm — 2012
- Practical Optimization Algorithm Design — 2011
- Principles of Computer Organization — 2011, 2010

Mentoring Experience

- **REU Undergraduate Students**:
  - At UNCC, mentor UG student Christopher York in a REU supplement award in 2019. Outcomes include a conference publication in MSST'20.
  - At TTU, participate REU'15 and mentor one undergraduate student (Nicholas Muirheada) on HPC metadata management system. Outcomes include a poster, a technical report, and a journal publication in ParCo'16.
- **Graduate Students**:
  - At UNCC, mentor four Ph.D students: Abdullah Al Raqibul Islam, Di Zhang, Md. Hasanur Rashid, and Saisha Kamat. Among them, Saisha Kamat is from under-representative group.
  - At TTU, Mentor two Ph.D. students (Neda Tavakoli and Wei Zhang) on HPC job scheduling and distributed graph storage systems. Outcomes include a poster (SC'16), a workshop paper (P2S2'16), conference papers (HPDC'17, CLUSTER'16), and a journal publication (ParCo'18).

Professional Service

Grant Panel Service

- **Panelist**: National Science Foundation, Computer Systems Research (CSR) 2017-2019, 2021, 2022
- **Panelist**: National Science Foundation, Office of Advanced Cyberinfrastructure (OAC) 2022
- **Panelist**: National Science Foundation, Small Business Innovation Research (SBIR) 2021, 2022

Conference Service

- Program Chair/Co-Chair
  - The 4th Industry/University Joint International Workshop on Data Center Automation, Analytics, and Control, held in conjunction with UCC 2021.
  - The 3rd International Industry/University Workshop on Data-center Automation, Analytic, and Control, held in conjunction with SC'19 (DAAC'19)
  - The 2nd International Industry/University Workshop on Data-center Automation, Analytic, and Control, held in conjunction with SC'18 (DAAC'18)
  - The 5th IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT'18) Big Data and HPC Track
  - The 10th IEEE/ACM International Conference on Utility and Cloud Computing (UCC'17) Poster Program
  - The 1st International Industry/University Workshop on Data-center Automation, Analytic, and Control, held in conjunction with UCC'17 (DAAC'17)
- Program Committee Member
  - IEEE International Conference on Big Data (IEEE BigData 2021, 2022)
  - IEEE International Symposium on Parallel and Distributed Processing with Applications (IEEE ISPA 2021, 2022)
  - International Conference on Parallel Processing (IEEE ICPP 2020, 2022)

  - IEEE International Parallel & Distributed Processing Symposium (IEEE IPDPS 2020, 2022, 2023)
  - IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (IEEE/ACM CCGrid 2020, 2022)
  - International Conference for High Performance Computing, Networking, Storage and Analysis (IEEE/ACM SC'20 Poster, SC'22 Workshop)
  - The 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications (TrustCom'18)
  - The National Cyber Security Summit (NCS'18)
  - IEEE International Conference on Big Data (BigData Congress'18)
  - Workshop on Advances in Software and Hardware for Big Data to Knowledge Discovery held in conjunction with BigData'14 (ASH'14)
- Session Chair
  - The 35th IEEE International Parallel & Distributed Processing Symposium (IPDPS'20)
  - The 10th IEEE/ACM International Conference on Utility and Cloud Computing (UCC'17)
  - The 4th IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT'17)
  - International Workshop on Parallel Programming Models and Systems Software for High-End Computing held in conjunction with ICPP'16 (P2S2'16)
  - International Workshop on Data Intensive Scalable Computing Systems held in conjunction with SC'14 (DISCS'14)

Journal Editorial Service

- Editor, *Cluster Computing Special Issue*
- Guest Editor, *Applied Soft Computing (BigData Special Issue)*

Journal Reviewing Service

- Reviewer for IEEE Transactions on Computers 2022
- Reviewer for IEEE Transactions on Storage 2020, 2021
- Reviewer for IEEE Transactions on Parallel and Distributed Systems 2015-2022
- Reviewer for IEEE Transactions on Cloud Computing 2015,2017
- Reviewer for IEEE Transactions on Industrial Informatics 2017, 2016
- Reviewer for Applied Soft Computing 2017, 2015
- Reviewer for International Journal of Parallel Programming 2015
- Reviewer for International Journal of High Performance Systems Architecture 2015

Department Service

- Search Committee member for ADR, Computer Science Department, UNC-Charlotte, 2021-2023.
- Faculty Search Committee, Computer Science Department, UNC-Charlotte, 2020-21.

Professional Memberships

- Institute for Electrical and Electronics Engineers (IEEE), Member, 2012–present
- Association for Computing Machinery (ACM), Member, 2014–present

Last updated: February 2, 2023