Bin Dong, PhD

Lawrence Berkeley National Laboratory One Cyclotron Road
MS 50B 3238E
Berkeley, California 94720 US

Phone: (510) 779-8060
Email: goon1983@gmail.com
Web: http://crd.lbl.gov//dongbin

Profile Overview

Passionate computer scientist and engineer with over ten years of R&D experience in large-scale data storage, index, and analysis. Significant experience with applying statistics or machine learning techniques (e.g. regression, clustering, time series, CNN) in mathematical modeling, system design and optimizations toward resolving real-world problems. Team player with solid written and oral communication skills anchored by a strong publication and professional service record in conferences and journals, e.g., *SIGMOD, JPDC, HPDC and PAAA*. Accomplished projects and software in past years are highlighted as below:

- ***SDS***, an autonomous data management framework for optimally selecting data index/organization to accelerate terabyte or even exascale-byte data analysis and mining on supercomputer. `https://code.lbl.gov/svn/sds/` (/w permission)
- ***SALB***, a distributed intelligent based self-acting load balancing system for large-scale data servers at supercomputing centers. `https://github.com/Goon83/SALB`
- ***MinCPP***, a Bayes's theorem based file assignment algorithm towards minimizing parallel data access contention.
- ***SDS-Sort***, the first parallel sorting algorithm with upper bound of $O(4\frac{N}{p})$ on load balancing, scaling to *130*K CPU cores and having sorting throughput of *117*TB/min.
- ***ArrayUDF***, a structure locality-aware parallel data processing system towards replacing MapReduce paradigms for multi-dimensional array processing, such as convolution used by CNN. It is *2000*X faster than Spark. `https://bitbucket.org/arrayudf/`
- ***DataElevator***, an ad-hoc software to perform low-contention data movement in deep-hierarchical storage system. `https://bitbucket.org/sbyna/dataelevator`

Computer Skills

- Program languages: C/C++, Python, R, SQL, MPI, Java, Lisp, HTML, JavaScript, Shell, etc.
- Development tools: GNU Make, Automake, Subversion, Git, Gcc, Gdb, LaTeX, etc.
- Data management & analysis tools: R, TensorFlow, Spark, SciDB, Lingo, Matlab, Berkeley DB, MySQL, etc.

Professional Experience

- **Research Scientists** — Lawrence Berkeley National Laboratory, CA
  Scientific Data Management (SDM) Group — *Feb 2016 – Present*
- **Postdoctoral Research Fellow** — Lawrence Berkeley National Laboratory, CA
  Scientific Data Management (SDM) Group — *Feb 2013 – Feb 2016*
- **Research Intern** — SINA Corporation. Beijing. China
  Architecture Group — *Jun – Oct 2012*

  - Profiling and tuning FlashCache (SSD based) for Sina Weibo [1] Storage System

- **Research Assistant** Beihang University. Beijing. China
  The School of Computer Science *2008 – 2012*
  - Research projects are Load Balancing of Parallel File System, SSD-HDD Hybrid Storage System, Small File Accesses, Metadata-intensive I/O, Cloud Computing Model for Immersive Environment, and Key Technology Research for Petaflops HPC
- **Intern** Intel's SSG Dept. Shanghai, China
  Software Test & Quality Assurance Group *Jan 2008 – May 2008*
  - Developed a UEFI based Diagnostic Software for Computer System
- **Research Intern** UESTC. Chengdu. China
  UESTC-Intel Lab *May 2007 – Dec 2007*
  - Port Smart Common Input Method (SCIM) to UMPC, Data based Computer System Forensic

Education

- **Beihang University** Beijing, China
  Ph.D. Computer System Architecture *2008 – 2013*
  - Advisor: Professor Xiao Li-Ming
  - Thesis: Performance Optimization for Parallel Data Access in High Performance Computing
- **University of Electronic and Science Technology of China (UESTC)** Chengdu, China
  B.S. Computer Science and Technology (software engineering oriented) *2004 – 2008*
  - Graduation Project: A UEFI based Diagnostic Software for Computer System

Academic Service and Contributions

- Publications chair of SSDBM 2017
- Program committee member: *9th IEEE ICOSST, SSDBM('2016, '2017), IEEE CIT'2017*
- Invited paper reviewers for journals: *PLOS ONE, Neurocomputing, PAAA, IEEE Transactions on Big Data, ACM ToMPECS, IEEE TPDS, Elsevier JSA('2016, '2017), JAAUBAS.*
- Invited paper reviewers for conferences: *IEEE Cluster('2017), IEEE BigData('2015,'2016), SSDBM'2015, NVMSA('2015,'2016), CIT'2015, CCGrid'2015, NAS'2015, NPC'2015, SDPS'2015, DISCS'2014*
- Coordinator of LBNL CRD Postdoc Program (2015-2016)

Presentations & Invited Talks

- Poster presentation, "Scientific Data Service-Autonomous Data Management on Exascale Infrastructure", UC Berkely, BIDS Spring 2017 Data Science Faire, 05/01/2017
- Invited Talks at UC Merced, USA, Sept. 2, 2015

[1] A Chinese microblogging (weibo) website. Akin to a hybrid of Twitter and Facebook

- PDSW'2015, Austin, TX, USA, Nov. 16, 2015
- IEEE Cluster 2013, Indianapolis, USA, Sept. 23, 2013
- The 2011 Conference on Grid and Distributed Computing, Jeju Island, Korea, Dec. 8, 2011
- The Conference on Computational Sciences & Optimization, Yunan, China, April 15, 2011
- IEEE International Workshop on HPC and Grid Applications, Anhui, China, May 28, 2010

Awards & Honors

Spot Recognition Award of LBNL CRD	2016
Best Paper Award of The 24th High Performance Computing Symposium	2016
Guang Hua Scholarship	2011
Outstanding Graduate Student of BeiHang University	2009 – 2010
Inter-class Basketball Match of BeiHang University (third place)	2009 – 2010
Excellent Graduate of UESTC	2009 – 2010
National Encouragement Scholarship of China	2008 – 2009
Intel Platform Administration Technology based Software Design Contest (second place)	2007
National Scholarship of China	2007 – 2008
National Mathematical Modeling Contest (third place)	2006
National Scholarship of China	2006 – 2007
The 6th Mathematical Modeling Contest of UESTC (third place)	2006
Higher-level Mathematics Competition of UESTC (third place)	2005
National Scholarship of China	2004 – 2005

Selected Publications

Journal Articles

1) **Bin Dong**, Xiuqiao Li, Limin Xiao, Li Ruan, *Towards minimizing disk I/O contention: A partitioned file assignment approach*, Future Generation Computer Systems (FGCS), Volume 37, July 2014, Pages 178-190, July 1, 2014

2) **Bin Dong**, Xiuqiao Li, Qimeng Wu, Limin Xiao, Li Ruan, *A dynamic and adaptive load balancing strategy for parallel file system with large-scale I/O servers*, Journal of Parallel and Distributed Computing (JPDC), Volume 72, Issue 10, October 2012, Pages 1254-1268

3) **Bin Dong**, Xiuqiao Li, Limin Xiao, Li Ruan, *A Non-Partitioned File Assignment Strategy for Parallel I/O System with Minimum Disk I/O Contention Probability*, Information Journal, 2013

4) **Bin Dong**, Xiuqiao Li, Limin Xiao, Li Ruan, *An optimal candidate selection model for self-acting load balancing of parallel file system*, International Journal of High Performance Computing and Networking (IJHPCN), Vol. 7, No. 2, pp.123-128, 2012

5) Xiuqiao Li, **Bin Dong**, Limin Xiao, Li Ruan, *Adaptive Tradeoff in Metadata-Based Small File optimizations for a Cluster File System*, International Journal of Numerical Analysis and Modeling (IJNAM), Volume 9, Number 2, Pages 289303, 2012

Conference Articles

1) **Bin Dong**, Kesheng Wu, Suren Byna, Jialin Liu, Weijie Zhao, and Florin Rusu, *ArrayUDF: User-Defined Scientific Data Analysis on Arrays*, The ACM International Symposium on High-Performance Parallel and Distributed Computing (HPDC) 2017 (Acceptance rate:19%)

2) Weijie Zhao, Florin Rusu, **Bin Dong**, Kesheng Wu, and Peter Nugent. *Incremental View Maintenance over Array Data.* In Proceedings of the 2017 ACM International Conference on Management of Data (SIGMOD '17). ACM, New York, NY, USA, 139-154. DOI: https://doi.org/10.1145/3035918.3064041 (Acceptance rate: 20%)

3) **Bin Dong**, Suren Byna, Kesheng Wu, Prabhat, Hans Johansen, Jeffrey N. Johnson, and Noel Keen, *Data Elevator: Low-contention Data Movement in Hierarchical Storage System*, The 23rd annual IEEE International Conference on High Performance Computing, Data, and Analytics (HiPC), 2016 (Acceptance rate: 25%)

4) Houjun Tang, Suren Byna, Steve Harenberg, Wenzhao Zhang, Xiaocheng Zou, Daniel F Martin, **Bin Dong**, Dharshi Devendran, Kesheng Wu, David Trebotich, *In Situ Storage Layout Optimization for AMR Spatio-temporal Read Accesses*, 45th International Conference on Parallel Processing (ICPP), 2016, 406–415, DOI: 10.1109/ICPP.2016.53 (Acceptance rate: 21.1%)

5) Wenzhao Zhang, Houjun Tang, Stephen Ranshous, Surendra Byna, Daniel F Martn, Kesheng Wu, **Bin Dong**, Scott Klasky, Nagiza F Samatova, *Exploring memory hierarchy and network topology for runtime AMR data sharing across scientific applications*, 2016 IEEE Big Data (Acceptance rate: 19.39% as short papers.)

6) **Bin Dong**, Suren Byna, and Kesheng Wu, *SDS-Sort: Scalable Dynamic Skew-aware Parallel Sorting*, The ACM International Symposium on High-Performance Parallel and Distributed Computing (HPDC) 2016, July 1, 2016 (Acceptance Rate:15.5%)

7) Weijie Zhao, Florin Rusu, **Bin Dong**, and Kesheng Wu. 2016. *Similarity Join over Array Data.* The 2016 International Conference on Management of Data (SIGMOD 2016).

8) Houjun Tang, Suren Byna, Steven Harenberg, Xiaocheng Zou, Wenzhao Zhang, Kesheng Wu, **Bin Dong**, Oliver Rubel, Kristofer Bouchard, Scott Klasky and Nagiza Samatova, *Usage Pattern-Driven Dynamic Data Layout Reorganization*, CCGrid'16

9) Wenzhao Zhang, Houjun Tang, Steven Harenberg, Suren Byna, Xiaocheng Zou, Dharshi Devendran, Daniel Martin, Kesheng Wu, Bin Dong, Scott Klasky and Nagiza Samatova, *AMRZone: A Runtime AMR Data Sharing Framework For Scientific Applications*, CCGrid'16

10) Tzuhsien Wu, Shyng Hao, Jerry Chou, **Bin Dong** and Kesheng Wu, *Indexing Blocks to Reduce Space and Time Requirements for Searching Large Data Files*, CCGrid 2016, May 16, 2016

11) Xiaocheng Zou, David Boyuka, Dhara Desai, Daniel Martin, Suren Byna, Kesheng Wu, Kushal Bansal, **Bin Dong**, Wenzhao Zhang, Houjun Tang, Dharshi Devendran, David Trebotich, Scott Klasky, Hans Johansen, Nagiza Samatova.*AMR-aware In Situ Indexing and Scalable Querying*, The 24th High Performance Computing Symposium (HPC 2016), April 4, 2016 (**Best Paper**)

12) **Bin Dong**, Suren Byna, and Kesheng Wu, *Heavy-tailed Distribution of Parallel I/O System Response Time*, 10th Parallel Data Storage Workshop (PDSW) 2015, to be held in conjunction with SC15, 2015,

13) **Bin Dong**, Suren Byna, and Kesheng Wu, *Spatially Clustered Join on Heterogeneous Scientific Data Sets*, 2015 IEEE International Conference on Big Data (IEEE BigData 2015), IEEE, 2015,

14) **Bin Dong**, Surendra Byna, Kesheng Wu, *Parallel Query Evaluation as a Scientific Data Service*, 2014 IEEE International Conference on Cluster Computing (CLUSTER), 2014

15) **Bin Dong**, Surendra Byna, Kesheng Wu, *Expediting scientific data analysis with reorganization of data*, 2013 IEEE International Conference on Cluster Computing (CLUSTER), pp.1,8, 23-27 Sept. 2013

16) **Bin Dong**, Surendra Byna, and Kesheng Wu. 2013. *SDS: a framework for scientific data services*, The 8th Parallel Data Storage Workshop (PDSW '13). ACM, New York, NY, USA

17) Spyros Blanas, Kesheng Wu, Surendra Byna, **Bin Dong**, Arie Shoshani, *Parallel data analysis directly on scientific file formats*, The 2016 International Conference on Management of Data (SIGMOD 2014).

18) Jialin Liu, Surendra Byna, **Bin Dong**, Kesheng Wu, Yong Chen, *Model-Driven Data Layout Selection for Improving Read Performance*, 2014 IEEE International Parallel & Distributed Processing Symposium Workshops (IPDPSW), IEEE, 2014

19) **Bin Dong**, Xiuqiao Li, Limin Xiao, Li Ruan, *A New File-Specific Stripe Size Selection Method for Highly Concurrent Data Access*, The 13th ACM/IEEE International Conference on Grid Computing (Grid 2012), 2012

20) **Bin Dong**, Xiuqiao Li, Limin Xiao, Li Ruan, *A File Assignment Strategy for Parallel I/O System with Minimum I/O Contention Probability*, The 2011 Conference on Grid and Distributed Computing, pp. 445-454, 2011

21) **Bin Dong**, Xiuqiao Li, Limin Xiao, Li Ruan, *Exploring Storage Optimizations to Accelerate Parallel Out-of-Core Matrix Product*, 2011 Fourth International Joint Conference on Computational Sciences and Optimization (CSO), pp.1-2, 2011

22) **Bin Dong**, Xiuqiao Li, Limin Xiao, Li Ruan, Binbin Yu, *Self-Acting Load Balancing with Parallel Sub File Migration for Parallel File System*, vol. 2, pp.317-321, IEEE International Workshop on HPC and Grid Applications, 2010

23) Xiuqiao Li, **Bin Dong**, Limin Xiao, Li Ruan, *Performance Optimization of Small File I/O with Adaptive Migration Strategy in Cluster File System*, 2nd International Conference on High Performance Computing and Applications, 2009.

24) Xiuqiao Li, **Bin Dong**, Limin Xiao, Li Ruan,*Small Files Problem in Parallel File System*, The 2011 International Conference on Network Computing and Information Security