Arnab K. Paul

Postdoctoral Research Associate,Computing & Computational Sciences
Oak Ridge National Laboratory, U.S.A.

Oak Ridge, Tennessee
akpaul@vt.edu
arnabkrpaul.github.io/

Research Interest

My interests lie in various domains of computer systems including **distributed systems**, **parallel file systems**, and **high performance computing**. Recently, my interest has also piqued in the direction of **big data analysis** to optimize system performance.

Education

2015–2020 **Ph.D., Computer Science and Applications**, *Virginia Polytechnic Institute and State University (Virginia Tech).*
Dissertation: *An application-attuned framework for optimizing HPC storage systems.*
Advisor: *Ali R. Butt.* ***GPA:*** *4.0/4.0*

2015–2018 **M.S., Computer Science and Applications**, *Virginia Tech.*
GPA: *3.85/4.0*

2013–2015 **M.Tech., Computer Science and Engineering**, *National Institute of Technology, Rourkela.*
Thesis: *Dynamic virtual machine placement in cloud computing.*
Advisor: *Bibhudatta Sahoo.* ***GPA:*** *9.56/10.0 (Gold Medalist - First position in the department)*

2009–2013 **B.Tech., Computer Science and Engineering**, *West Bengal University of Technology.*
GPA: *9.02/10.0 (First position in the department)*

2009 **Class XII - I.S.C. Examination**, *Council for The Indian School Certificate Examination.*
Hill Top School, Jamshedpur
Marks: *96.25%*

2007 **Class X - I.C.S.E**, *Council for The Indian School Certificate Examination.*
Hill Top School, Jamshedpur
Marks: *95.8%*

Research Experience

09/20–present **Oak Ridge National Laboratory - Analytics & AI Methods at Scale Group**, *Postdoctoral Research Associate.*
- Analyze the I/O patterns of emerging data science applications, and identify performance bottlenecks.
- ML techniques to predict I/O patterns to help job scheduling in large-scale supercomputers, like Summit.

08/15–08/20 **Virginia Tech - Distributed Systems and Storage Laboratory**, *Ph.D. Student in Dept. of CS.*
- Conducted an empirical study on the use of containers in HPC platforms.
- Developed an approach for estimating the performance of edge-based clustering applications.
- Built an I/O framework for load balancing storage servers in HPC parallel file systems, like Lustre.
- Developed a model to optimize data partitioning for in-memory data analytics platforms, like Spark.

06/19–08/19 **Cray Inc.**, *Graduate Research Intern.*
Mentors: Cory Spitz, Nathan Rutman (Cray Inc.), and Scott White (Los Alamos National Laboratory)
- Built a scalable re-indexer for `BRINDEXER` - a metadata indexing tool used in Cray.

05/18–08/18 **Lawrence Livermore National Laboratory**, *Graduate Student Summer Intern (Computation Scholar).*
Mentor: Kathryn Mohror
- Analyzed the characteristics of metadata and I/O for jobs running on two supercomputers at LLNL.

05/17–08/17 **Argonne National Laboratory**, *Graduate Student Summer Intern (Research Aide).*
Mentor: Ian Foster
- Created `FSMonitor` - a tool for scalable file system event monitoring for arbitrary file systems.

01/14–05/15 **NIT Rourkela - Information Security and Data Communication Laboratory**, *M.Tech. Student.*
- Proposed an approach for dynamic virtual machine placement in the cloud using game theory.
- Applied and analyzed greedy algorithms on virtual machine distribution across data centers.

Teaching Experience

Fall 2019 **Instructor**, *Virginia Tech* courses.cs.vt.edu/cs2505/fall2019/.
CS2505: Introduction to Computer Organization - I: Prepared and gave lectures to two sections (~150 students), prepared assignments and examinations, awarded final grades, mentored graduate teaching assistants.

2015–2019 **Graduate Teaching Assistant**, *Department of Computer Science, Virginia Tech*.

Spring 2019	CS 3214: Operating Systems	*Recitation sessions, grading, guest lectures, office hours*
Fall 2018	CS 5584: Network Security	*Project ideas with 15 groups, grading, office hours*
Spring 2018	CS 3114: Data Structures and Algorithms	*Grading, office hours*
Fall 2017	CS 2506: Introduction to Computer Organization - II	*Grading, office hours*
Spring 2017, Fall 2016	CS 2114: Software Design and Data Structures	*Lab sessions for 60 students, practice sessions, designing and grading assignments, office hours*
Spring 2016, Fall 2015	CS 1054: Introduction to Programming in Java	*Lab sessions for 60 students, grading, office hours*

2014–2015 **Graduate Teaching Assistant**, *Department of Computer Science, NIT Rourkela*.

Autumn 2014, Spring 2015	CS 171: Computing Lab	*Prepared and gave lectures, held lab sessions for 220 students, preparing and grading assignments*
Spring 2015	CS 670: Data Mining Lab	*Lab sessions for 30 students, grading*

Publications (Google Scholar scholar.google.co.in/citations?user=az8MAG0AAAAJ&hl=en)

Book Chapters

CRC Press '20 **Arnab K. Paul**. Edge or Cloud: What to Choose?. In Cloud Network Management: An IoT based Framework, CRC Press, Taylor & Francis Group, pages 14, 2020. doi.org/10.1201/9780429288630

IGI Global '17 **Arnab Kumar Paul**, and Bibhudatta Sahoo. Dynamic virtual machine placement in cloud computing. In Resource Management and Efficiency in Cloud Computing Environments, pp. 136-167, IGI Global, 2017. doi.org/10.4018/978-1-5225-1721-4.ch006

Journal Publications

TPDS '21 **[Core Rank: A*]** Nannan Zhao, Vasily Tarasov, Hadeel Albahar, Ali Anwar, Lukas Rupprecht, Dimitrios Skourtis, **Arnab K. Paul**, Keren Chen, and Ali R. Butt. Large-Scale Analysis of Docker Images and Performance Implications for Container Storage Systems. IEEE Transactions on Parallel and Distributed Systems (TPDS), pages 13, April 2021. doi.org/10.1109/TPDS.2020.3034517

Conference and Workshop Publications

INDIS '21 @ SC '21 **[Core Rank: A]** Debasmita Biswas, Sarah Neuwirth, **Arnab K. Paul**, and Ali R. Butt. Bridging Network and Parallel I/O Research for Improving Data-Intensive Distributed Applications. In Proceedings of the 8th Annual International Workshop on Innovating the Network for Data-Intensive Science (INDIS) in conjunction with SC'21, pages 7, November 2021. Accepted

MASCOTS '21 **[Core Rank: B]** **Arnab K. Paul**, Ahmad Maroof Karimi, and Feiyi Wang. Characterizing Machine Learning I/O Workloads on Leadership Scale HPC Systems. In Proceedings of the 29th IEEE International Symposium on the Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, pages 8, November 2021. Accepted

REX-IO '21 @ Cluster '21 **[Core Rank: A]** Sarah Neuwirth and **Arnab K. Paul**. Parallel I/O Evaluation Techniques and Emerging HPC Workloads: A Perspective. In Proceedings of the 1st Workshop on Re-envisioning Extreme-Scale I/O for Emerging Hybrid HPC Workloads (REX-IO) in conjunction with IEEE Cluster'21, pages 8, September 2021. Accepted

HiPC '20 **[Core Rank: National - India]** **Arnab K. Paul**, Olaf Faaland, Adam Moody, Elsa Gonsiorowski, Kathryn Mohror, and Ali R. Butt. Understanding HPC Application I/O Behavior Using System Level Statistics. In Proceedings of the 27th IEEE International Conference on High Performance Computing, Data, and Analytics, pages 10, December 2020. (AR: 23%). doi.org/10.1109/HiPC50609.2020.00034

SMDS '20 **[Core Rank: B]** Breno Dantas Cruz, **Arnab K. Paul**, Zheng Song, and Eli Tilevich. STARGAZER: A Deep Learning Approach for Estimating the Performance of Edge-Based Clustering Applications. In Proceedings of the IEEE International Conference on Smart Data Services, pages 9, October 2020. (AR: 17%). doi.org/10.1109/SMDS49396.2020.00009 **(Awarded the YESC award for the most innovative student paper at IEEE Services 2020!)**

Cloud '20 [Core Rank: B] Subil Abraham[1], **Arnab K. Paul**[1], Redwan Ibne Seraj Khan, and Ali R. Butt. On the Use of Containers in High Performance Computing Environments. In Proceedings of the IEEE International Conference on Cloud Computing, pages 9, October 2020. (AR: 17%). doi.org/10.1109/CLOUD49709.2020.00048

[1] **Both authors contributed equally.**

CCGrid '20 [Core Rank: A] **Arnab K. Paul**, Brian Wang, Nathan Rutman, Cory Spitz, and Ali R. Butt. Efficient Metadata Indexing for HPC Storage Systems. In Proceedings of the 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing, Australia, pages 10, May 2020. (AR: 23%). doi.org/10.1109/CCGrid49817.2020.00-77

PDSW '19 @ SC '19 [Core Rank: A] **Arnab K. Paul**, Olaf Faaland, Adam Moody, Elsa Gonsiorowski, Kathryn Mohror, and Ali R. Butt. Improving I/O Performance of HPC Application Using Intra-Job Scheduling. Work-In-Progress in Proceedings of the 4th Joint International Workshop on Parallel Data Storage & Data Intensive Scalable Computing Systems (PDSW-DISC'19) in conjunction with SC'19, Denver, CO, pages 1, November 2019. www.pdsw.org/pdsw19/wips/ArnabPaul-pdswWIP.pdf

Cluster '19 [Core Rank: A] **Arnab K. Paul**, Ryan Chard, Kyle Chard, Steven Tuecke, Ali R. Butt, and Ian Foster. FSMonitor: Scalable File System Monitoring for Arbitrary Storage Systems. In Proceedings of the IEEE International Conference on Cluster Computing, Albuquerque, NM, pages 11, September 2019. (AR: 22%). doi.org/10.1109/CLUSTER.2019.8891045

IPDPS '19 [Core Rank: A] Bharti Wadhwa, **Arnab K. Paul**, Sarah Neuwirth, Feiyi Wang, Sarp Oral, Ali R. Butt, Jon Bernard, and Kirk W. Cameron. iez: Resource Contention Aware Load Balancing for Large-Scale Parallel File Systems. In Proceedings of the IEEE International Parallel and Distributed Processing Symposium, Rio de Janeiro, Brazil, pages 11, May 2019. (AR: 25%). doi.org/10.1109/IPDPS.2019.00070

ComNet-IoT @ ICDCN '19 [National - India] Hyogi Sim, **Arnab K. Paul**, Eli Tilevich, Ali R. Butt, and Muhammad Shahzad. CSLIM: Automated Extraction of IoT Functionalities from Legacy C Codebases. In Proceedings of the 8th International Workshop on Computing and Networking for IoT and Beyond in conjunction with ICDCN '19, Bangalore, India, pages 6, January 2019. doi.org/10.1145/3288599.3296013

BigData '17 **Arnab K. Paul**, Arpit Goyal, Feiyi Wang, Sarp Oral, Ali R. Butt, Michael J. Brim, and Sangeetha B. Srinivasa. I/O Load Balancing for Big Data HPC Applications. In Proceedings of the IEEE International Conference on Big Data, Boston, MA, pages 10, December 2017. (AR: 18%). doi.org/10.1109/BigData.2017.8257931

PDSW '17 @ SC '17 [Core Rank: A] **Arnab K. Paul**, Ryan Chard, Kyle Chard, Steven Tuecke, Ali R. Butt, and Ian Foster. Toward Scalable Monitoring on Large-Scale Storage for Software Defined Cyberinfrastructure. In Proceedings of the 2nd Joint International Workshop on Parallel Data Storage & Data Intensive Scalable Computing Systems (PDSW-DISC'17) in conjunction with SC'17, Denver, Colorado, pages 6, November 2017. doi.org/10.1145/3149393.3149402

WHPC '16 @ SC '16 [Core Rank: A] Sangeetha B. Srinivasa, **Arnab K. Paul**, Arpit Goyal, Feiyi Wang, Sarp Oral, and Ali R. Butt. I/O Load Balancing for Lustre Distributed File System. In Women in HPC in conjunction with SC'16, Salt Lake City, Utah, November 2016.

Cluster '16 [Core Rank: A] **Arnab Kumar Paul**, Wenjie Zhuang, Luna Xu, Min Li, Mustafa Rafique, and Ali R. Butt. CHOPPER: Optimizing Data Partitioning for In-Memory Data Analytics Frameworks. In Proceedings of the IEEE International Conference on Cluster Computing, Taiwan, pages 10, September 2016. (AR: 24%). doi.org/10.1109/CLUSTER.2016.41

INDICON '14 **Arnab Kumar Paul**, Sourav Kanti Addya, Bibhudatta Sahoo, and Ashok Kumar Turuk. Application of greedy algorithms to virtual machine distribution across data centers. In Proceedings of 2014 Annual IEEE India Conference, pp. 1-6. IEEE, 2014. doi.org/10.1109/INDICON.2014.7030633

ICACCCT '14 Arjun Datta, and **Arnab Kumar Paul**. Online compiler as a cloud service. In Proceedings of 2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies, pp. 1783-1786. IEEE, 2014. doi.org/10.1109/ICACCCT.2014.7019416

Posters

SC '19 [Core Rank: A] **Arnab K. Paul**, Olaf Faaland, Adam Moody, Elsa Gonsiorowski, Kathryn Mohror, and Ali R. Butt. Understanding HPC Application I/O Behavior Using System Level Statistics. In Proceedings of The International Conference for High Performance Computing, Networking, Storage, and Analysis (SC) 2019, Denver, CO, pages 3, November 2019. sc19.supercomputing.org/proceedings/tech_poster/tech_poster_pages/rpost157.html

Cray '19 **Arnab K. Paul**, Nathan Rutman, Cory Spitz, Brian Wang, Peter Bojanic, and Ali R. Butt. Analyzing the performance of file system indexing tools. In Cray Inc. Summer Student Poster Session, Minneapolis, MN, August 2019.

LLNL '18 **Arnab K. Paul**, Olaf Faaland, Adam Moody, Elsa Gonsiorowski, Kathryn Mohror, and Ali R. Butt. Analysis and predictive modeling of HPC I/O workloads. In LLNL Computation Summer Student Poster Session, Livermore, CA, August 2018.

Theses

Ph.D. '20 **Arnab Kumar Paul**. An application-attuned framework for optimizing HPC storage systems. Ph.D. dissertation, Department of Computer Science and Applications, Virginia Tech, U.S.A., 2020. hdl.handle.net/10919/99793

M.Tech. '15 **Arnab Kumar Paul**. Dynamic virtual machine placement in cloud computing. Master's thesis, Department of Computer Science and Engineering, National Institute of Technology, Rourkela, 2015. ethesis.nitrkl.ac.in/6811/

Others

SMC '21 Sajal Dash, **Arnab Kumar Paul**, Sarp Oral, Feiyi Wang. SMC Data Challenge 2021: Analyzing Resource Utilization and User Behavior on Titan Supercomputer. In Smoky Mountains Computational Sciences & Engineering Conference.

Talks and Presentations

2021 **Oak Ridge National Laboratory**, *Analyzing Machine Learning Workloads on Leadership Scale HPC Storage Systems*, Oak Ridge Postdoctoral Association Research Symposium.

2021 **Netaji Subhash Engineering College**, *Decoding Process Management in Operating Systems*, Special Lecture.

2020 **HiPC**, *Understanding HPC Application I/O Behavior Using System Level Statistics*, Paper Presentation.

2020 **Oak Ridge National Laboratory**, *A Framework for Whole Stack Optimization of Distributed Storage Systems*, Job talk.

2020 **Lawrence Berkeley National Laboratory**, *A Framework for Whole Stack Optimization of Distributed Storage Systems*, Job talk.

2019 **SC**, *Understanding HPC Application I/O Behavior Using System Level Statistics*, Poster presentation.

2019 **SC**, *Improving I/O Performance of HPC Application Using Intra-Job Scheduling*, WIP presentation.

2019 **Cluster**, *Scalable File System Monitoring for Arbitrary Storage Systems*, Paper presentation.

2019 **ICDCN**, *Automated extraction of IoT functionalities from legacy C codebases*, Paper presentation.

2017 **BigData**, *I/O Load Balancing for Big Data HPC Applications*, Paper presentation.

2017 **PDSW @ SC**, *Toward Scalable Monitoring on Large-Scale Storage for Software Defined Cyberinfrastructure*, Paper presentation.

2016 **Cluster**, *Optimizing Data Partitioning for In-Memory Data Analytics Frameworks*, Paper presentation.

Awards and Honors

2020 Awarded the YESC award for most innovative student paper at IEEE Services 2020, Beijing, China

2019-2020 BitShares Fellowship, Department of Computer Science, Virginia Tech

2019 Travel Grant Recipient, IEEE Cluster, Albuquerque, NM, USA

2019 Student Volunteer, SC, Denver, CO, USA

2018, '19, '20 Member of the Dean's Graduate Team & Ambassador to the College of Engineering, Virginia Tech

2018 Student Volunteer, SCiNet @ SC, Dallas, TX, USA

2017-2018 President, Bengali Students' Association, Virginia Tech

2017-2018 BitShares Fellowship, Department of Computer Science, Virginia Tech

2017 Travel Grant Recipient, IEEE BigData, Boston, MA, USA

2016 Travel Grant Recipient, IEEE Cluster, Taipei, Taiwan

2016 Student Volunteer, SC, Salt Lake City, Utah, USA

2015 Gold Medalist, Department of Computer Science, National Institute of Technology - Rourkela
2015 Recognition for building a university website for project and advisor allocation at NIT - Rourkela
2015 Recognition for building a website for CWS hospital in Rourkela ☞ www.cwshospital.org
2013 Recognition for Rank 1, Department of Computer Science, West Bengal University of Technology
2010 Golden Jubilee Scholarship, Highest Marks in Class XII, Tata Motors Limited, Jamshedpur
2009 Recognition for holding 1^{st} rank from Kindergarten to Std. XII (all 15 years), Hill Top School
2009 Tata Hitachi Award, Tata Motors Limited, Jamshedpur
2008 Golden Jubilee Scholarship, Highest Marks in Class X, Tata Motors Limited, Jamshedpur
2007 Young Achiever Award, Highest Marks, Tata Cummins, Jamshedpur

Professional Service

Workshop Co-Chair: REX-IO '21 (1^{st} Workshop on Re-envisioning Extreme-Scale I/O for Emerging Hybrid HPC Workloads) in conjunction with IEEE Cluster 2021

TPC Member: AHPC '22 (Advances in High-Performance Computing), SC '21, AHPC '21, Cloud Computing '21, Book on Convergence of Deep Learning in Cyber-IoT Systems and Security '21, ICDCS '20

Reviewer: IEEE Transactions on Parallel and Distributed Systems (TPDS) '19 '20, Neural Processing Letters (NEPL) '20 '21, Cluster Computing Journal '19 '20 '21, IJGHPC '18 '19 '20, ASTESJ '18, AUTOSOFT Journal '18, MGS Journal '17

External Reviewer: IEEE TSC Journal '18, BigData '17 '18, Cluster '17 '18 '20, ECOOP '20, HPDC '17 '18 '20, IC2E '17, ICCD '19, ICDCS '17 '18 '19, ICS '17 '18, IPDPS '18 '19 '20

Facilitator: SC '20, HPDC '19

Mentor: SC '20

Mentoring Experience

Graduate Students: Redwan Ibne Seraj Khan (Ph.D., Viriginia Tech, 2019 -)

Debasmita Biswas (Ph.D., Virginia Tech, 2020 -)

Subil Abraham (MS, Virginia Tech, 2019 - 2020) (Thesis: On the Use of Containers in High Performance Computing Environments)

Arpit Goyal (MS, Virginia Tech, 2016 - 2017) (Thesis: I/O Load Balancing for Lustre Distributed File System)

Undergraduate Students: Subrat Dhal (B.Tech., NIT Rourkela, 2014 - 2015) (Project: Performance of bin-packing algorithms for virtual machine placement in the cloud)

Harshit Verma (B.Tech., NIT Rourkela, 2014 - 2015) (Project: Performance of bin-packing algorithms for virtual machine placement in the cloud)

Skills

General: C, C++, Python, Java, UNIX, git, svn, latex, gnuplot.

Analytics: Apache Spark, pandas, matplotlib, bigdata analysis, applied machine learning, federated learning.

File Systems: Lustre file system, Ceph object store, HDFS, IBM Spectrum Scale (GPFS).

Distributed Computing: Containers, cloud computing, key-value stores, edge computing, IoT, map-reduce.

Memberships

2019 – present Association for India's Development, Blacksburg Chapter.
2018 – 2020 Graduate Students' Council, Department of Computer Science, Virginia Tech.
2016 – present Institute of Electrical and Electronics Engineers (IEEE), Student member.

References

1. Dr. Ali R. Butt <butta@cs.vt.edu>, *Professor, Virginia Tech.*
2. Dr. Ian Foster <foster@anl.gov>, *Senior Scientist - Argonne National Laboratory, Professor - University of Chicago.*

3. Dr. Eli Tilevich <tilevich@cs.vt.edu>, *Professor, Virginia Tech*.
4. Dr. Feiyi Wang <fwang2@ornl.gov>, *Group Leader, Analytics & AI Methods at Scale Group, Oak Ridge National Laboratory*.