Arnab Kumar Paul

Ph.D. Candidate in Computer Science · Distributed Systems
Virginia Tech, Blacksburg, Virginia, USA
+1 (540) 998-1480 | akpaul@vt.edu | https://arnabkrpaul.github.io/ | arnabkrpaul

"Your best teacher is your last mistake." - Dr. A. P. J. Abdul Kalam

Personal Statement

I am a fifth year Ph.D. candidate in the Department of Computer Science at Virginia Tech. I work in the Distributed Systems and Storage Laboratory headed by Dr. Ali R. Butt. My research interests include *virtualization, distributed systems, distributed file systems, Internet of Things and big data APIs.*

My current research focus is building scalable and fast file system indexing tools for large scale distributed file systems. This work is in collaboration with Cray and Los Alamos National Laboratory. My other research projects include developing novel research methods for collecting and analyzing file system performance metrics for better I/O performance on future HPC systems, predictive modeling of HPC I/O workloads for improved job scheduling to lower I/O contention at a finer granularity, home automation for research data management in distributed file systems, and creating a load balanced distributed file system.

I have also worked on optimizing the performance of Spark workloads. For my Masters degree, I worked on optimizing and improving performance of dynamic virtual machine placement in cloud computing using cooperative and non-cooperative game theory.

Apart from research, I enjoy playing and watching cricket. I love listening to songs and writing poems. I take active participation in 10k (best time - 58:33) and half marathon (best time - 2:13:04) runs. At Virginia Tech, I have also served as the President of Bengali Students' Association for the term 2017-18. I believe that a happy and balanced lifestyle yields positive research.

Research Experience

Cray Inc.
Los Alamos, NM, USA

Graduate Research Intern — *Jun. 2019 - Aug. 2019*

- **Scalable Metadata Indexing in Distributed File Systems** (June 2019 - February 2020)
  - Compare different indexing technologies used in large scale distributed file system.
  - Build a stream-based rules processing engine for Lustre File System.
  - Compare indexing system with stream-based rules processing approach.

Lawrence Livermore National Laboratory
Livermore, CA, USA

Graduate Student Summer Intern (Computation Scholar) — *May 2018 - Aug. 2018*

- **Analysis and Predictive Modeling of HPC I/O Workloads** (May 2018 - PRESENT)
  - Analyze the metadata and job statistics of HPC I/O workloads.
  - Build predictive models of I/O workloads based on the time series server data.
  - Form a design solution for job scheduling to lower I/O contention based on the predictive models.

Argonne National Laboratory
Lemont, IL, USA

Graduate Student Summer Intern (Research Aide) — *May 2017 - Aug. 2017*

- **Home automation of research data management** (May 2017 - May 2019)
  - Responsive storage architecture that allows users to express data management tasks via a rules notation for distributed file systems, such as Lustre.
  - The system monitors the storage system for events, evaluates rules and then uses serverless computing techniques to execute actions in response to the events.

Virginia Tech
Blacksburg, VA, USA

Ph.D. Student (Distributed SYstems and Storage Laboratory) — *Aug. 2015 - PRESENT*

- **Load balancing in large scale storage system** (May 2016 - PRESENT)
  - Most distributed file systems are hierarchical. Lustre distributed file system is taken as the use case to perform load balancing.
  - In order to have a global view of the whole system, a publisher-subscriber model is used to collect statistics of various components and pass them onto the Metadata Server.
  - In addition to statistics collection, machine learning is used to model the application behavior and predict future requests.
  - A list of Object Storage Targets is generated using the minimum cost maximum flow algorithm to have a load balanced setup.

- **Auto-tuning of parallelism in Spark** (Jan. 2016 - May 2016)
  - Sizes of partitions play a big role in determining the execution speed of stages in a Spark job.
  - For best performance, Spark should find optimal partition sizes between two stages.
  - At present, users optimize the number of partitions manually to increase performance.
  - The aim of this project is to modify the DAG scheduler to consider the number of partitions and their sizes before scheduling the next stage.

National Institute of Technology, Rourkela — *Odisha, India*

Masters Student (Information Security and Data Communication Laboratory) — *May 2014 - May 2015*

- **Dynamic Virtual Machine Placement in Cloud Computing** (May 2014 - May 2015)
  - A decentralized approach based on game theoretic method is used here in order to reach optimal solutions and also a list of executable live virtual machine migrations is provided to reach the optimal placement.
  - Both cooperative as well as non-cooperative game theoretic approaches have been used to find optimal solution to the dynamic virtual machine placement problem.

- **Application of Greedy Algorithms to Virtual Machine Distribution across Data Centers** (May 2014 - Aug. 2014)
  - Two parameters namely; minimization of energy consumption and minimization of prices to distribute virtual machines over data centers form the objective function.
  - Analysis is done by applying greedy algorithms to solve the proposed cloud model.

Education

Ph.D. in Computer Science — *Blacksburg, Virginia, USA*

Virginia Tech, USA — *Aug. 2015 - Aug. 2020 (Expected)*

GPA: 4.0/4.0

- **Advisor**: Dr. Ali R. Butt

Master of Science (M.S.) in Computer Science and Applications — *Blacksburg, Virginia, USA*

Virginia Tech, USA — *Aug. 2015 - May 2018*

GPA: 3.85/4.0

Master of Technology (M.Tech.) in Computer Science and Engineering — *Odisha, India*

National Institute of Technology, Rourkela — *Aug. 2013 - Jun. 2015*

GPA: 9.56/10.0

- **Specialization**: Software Engineering
- **Thesis**: Dynamic Virtual Machine Placement in Cloud Computing
- **Advisor**: Dr. Bibhudatta Sahoo
- Gold medalist in Software Engineering.

Bachelor of Technology (B.Tech.) in Computer Science and Engineering — *West Bengal, India*

West Bengal University of Technology — *Aug. 2009 - Jun. 2013*

GPA: 9.02/10.0

Employment Experience

Cray Inc. — *Los Alamos, NM, USA*

Graduate Research Intern — *June 2019 - Aug. 2019*

- **Mentor:** Cory Spitz, Nathan Rutman (Cray Inc.) and Scott White (Los Alamos National Laboratory)

Lawrence Livermore National Laboratory — *Livermore, CA, USA*

Graduate Student Summer Intern (Computation Scholar) — *May 2018 - Aug. 2018*

- **Advisor:** Dr. Kathryn Mohror

Argonne National Laboratory — *Lemont, IL, USA*

Graduate Student Summer Intern (Research Aide) — *May 2017 - Aug. 2017*

- **Advisor:** Dr. Ian Foster

Virginia Tech — *Blacksburg, VA, USA*

Graduate Research Assistant — *May 2016 - Aug. 2016*

- **Advisor:** Dr. Ali R. Butt
- Distributed Systems and Storage Laboratory

National Institute of Technology, Rourkela — *Odisha, India*

Teaching Assistant — *Jul. 2014 - May 2015*

- **CS 171: Computing Lab** - *Autumn 2014, Spring 2015*
  - Taught lectures.
  - Graded assignments and projects.
  - Aided in forming homework problems.
- **CS 670: Data Mining Lab** - *Spring 2015*
  - Graded assignments and projects.
  - Aided in forming homework problems.

Virginia Tech *Blacksburg, VA, USA*

Instructor *Aug. 2019 - Dec. 2019*

- **CS 2505: Intro Computer Organization** - *Fall 2019*
  - An introduction to the design and operation of digital computers.
  - Taught lectures.
  - Prepared assignments and projects.

Virginia Tech *Blacksburg, VA, USA*

Graduate Teaching Assistant *May 2015 - PRESENT*

- **CS 3214: Computer Systems** - *Spring 2019*
  - Held recitation sessions.
  - Conducted lectures for 200 students.
  - Graded assignments and projects.
  - Held office hours.
- **CS 5584: Network Security** - *Fall 2018*
  - Helped 30 graduate students with research ideas
  - Graded assignments and projects.
  - Held office hours.
- **CS 3114: Data Structures and Algorithms** - *Spring 2018*
  - Graded assignments and projects.
  - Held office hours.
- **CS 2506: Computer Organization II** - *Fall 2017*
  - Graded assignments and projects.
  - Held office hours.
- **CS 2114: Software Design and Data Structures** - *Fall 2016, Spring 2017*
  - Conducted lab sessions for 60 students.
  - Designed assignments.
  - Held practice sessions.
  - Graded assignments and projects.
  - Held office hours.
- **CS 1054: Introduction to Programming in Java** - *Fall 2015, Spring 2016*
  - Conducted lab sessions for 60 students.
  - Graded assignments and projects.
  - Held office hours.

Publications

Masters Dissertation

- **Arnab Kumar Paul**. *Dynamic Virtual Machine Placement in Cloud Computing.* National Institute of Technology, Rourkela, 2015.

Book Chapters

- **Arnab Kumar Paul** and Bibhudatta Sahoo. *Dynamic Virtual Machine Placement in Cloud Computing.* Resource Management and Efficiency in Cloud Computing Environments, IGI Global 2016.

Peer-Reviewed Conferences

- **Arnab K. Paul**, Brian Wang, Nathan Rutman, Cory Spitz, and Ali R. Butt. *Efficient Metadata Indexing for HPC Storage Systems.* Proceedings of the 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGrid), Melbourne, Australia, pages 10, May 2020. (Acceptance Rate - 66/221)
- **Arnab K. Paul**, Ryan Chard, Kyle Chard, Steven Tuecke, Ali R. Butt, Ian Foster. *FSMonitor: Scalable File System Monitoring for Arbitrary Storage Systems.* IEEE Cluster 2019 (Acceptance Rate - 22%)
- Bharti Wadhwa, **Arnab K. Paul**, Sarah Neurith, Feiyi Wang, Sarp Oral, Ali R. Butt, Jon Bernard, Kirk W. Cameron. *Resource Contention Aware Load Balancing for Large-Scale Parallel File Systems.* IEEE IPDPS 2019 (Acceptance Rate - 25%)
- **Arnab K. Paul**, Arpit Goyal, Feiyi Wang, Sarp Oral, Ali R. Butt, Michael J. Brim, Sangeetha B. Srinivasa. *I/O Load Balancing for Big Data HPC Applications.* IEEE International Conference on Big Data (Big Data '17), pp. 233-242 (Acceptance Rate - 18%)
- **Arnab Kumar Paul**, Wenjie Zuang, Luna Xu, Min Li, M. Mustafa Rafique, Ali R. Butt. *CHOPPER: Optimizing Data Partitioning for In-Memory Data Analytics Frameworks.* IEEE International Conference on Cluster Computing (Cluster '16), pp. 110-119 (Acceptance Rate - 24%)

- **Arnab Kumar Paul**, Sourav Kanti Addya, Bibhudatta Sahoo and Ashok Kumar Turuk. *Application of Greedy Algorithms to Virtual Machine Distribution across Data Centers.* 11^{th} IEEE India Conference INDICON 2014, Emerging Trends and Innovation of Technology, pp. 1-6.
- Arjun Datta and **Arnab Kumar Paul**. *Online Compiler as a Cloud Service.* IEEE International Conference on Advanced Communication Control and Computing Technologies, 2014 IEEE Computer Society, pp. 1798-1801

Workshops and Posters

- **Arnab K. Paul**, Olaf Faaland, Adam Moody, Elsa Gonsiorowski, Kathryn Mohror, Ali R. Butt. *Improving I/O Performance of HPC Application Using Intra-Job Scheduling.* Work-In-Progress in Proceedings of the 4^{th} Joint International Workshop on Parallel Data Storage Data Intensive Scalable Computing Systems (PDSW-DISC'19) in conjunction with SC'19, Denver, CO.
- **Arnab K. Paul**, Olaf Faaland, Adam Moody, Elsa Gonsiorowski, Kathryn Mohror, Ali R. Butt. *Understanding HPC Application I/O Behavior Using System Level Statistics.* Poster In SC 2019, Denver, CO.
- Hyogi Sim, **Arnab K. Paul** , Eli Tilevich, Ali R. Butt, Muhammad Shahzad. *Cslim: Automated Extraction of IoT Functionalities from Legacy C Codebases.* In 2019 8th International Workshop on Computing and Networking for IoT and Beyond (ComNet-IoT' 19) in conjunction with 20th International Conference on Distributed Computing and Networking, Bangalore, India, pp. 421-426. ACM, 2019.
- **Arnab K. Paul** , Ryan Chard, Kyle Chard, Steven Tuecke, Ali R. Butt, Ian Foster. *Toward Scalable Monitoring on Large-Scale Storage for Software Defined Cyberinfrastructure.* In Proceedings of the 2^{nd} Joint International Workshop on Parallel Data Storage & Data Intensive Scalable Computing Systems (PDSW-DISC '17) in conjunction with SC, Denver, CO, pp. 49-54
- Sangeetha B. Srinivasa, **Arnab K. Paul** , Arpit Goyal, Feiyi Wang, Sarp Oral, Ali R. Butt. *I/O Load Balancing for Lustre Distributed File System.* WHPC, SC 2016.

Skills & Expertise

- **Programming Languages:** C, C++, Python, JAVA, C#, SCALA
- **Parallel & Distributed File Systems:** Lustre
- **Databases:** MySQL, SQLite, MS SQLServer
- **I/O Benchmarks:** IOR, HACC-IO, FileBench
- **Tools:** gcc, git, svn, eclipse, visual studio, latex, gnuplot, Open ESB, IBM RQM
- **Web Development:** HTML, CSS, Javascript, JQuery, ASP.NET

Honors & Awards

2020	**Awardee**, BitShares Fellowship by Dept. of Computer Science	*Virginia Tech, U.S.A*
2018-20	**Member**, Deans Graduate Team	*Virginia Tech, U.S.A*
2019-20	**Member**, Association for India's Development, Blacksburg Chapter	*Virginia Tech, U.S.A*
2019-20	**Member**, Computer Science Graduate Students Council	*Virginia Tech, U.S.A*
2019	**Travel Grant Recipient**, IEEE Cluster 2017	*Albuquerque, U.S.A*
2019	**Student Volunteer**, SC'19	*Denver, CO, U.S.A*
2018	**Awardee**, BitShares Fellowship by Dept. of Computer Science	*Virginia Tech, U.S.A*
2018	**Student Volunteer**, SCiNet Team for SC'18	*Dallas, TX, U.S.A*
2017-18	**President**, Bengali Students' Association	*Virginia Tech, U.S.A*
2017	**Travel Grant Recipient**, IEEE Big Data 2017	*Boston, MA, U.S.A*
2016	**Travel Grant Recipient**, IEEE Cluster 2016	*Taipei, Taiwan*
2016	**Student Volunteer**, SC'16	*Salt Lake City, U.S.A*
2015	**Gold Medalist**, Dept. of Computer Science, National Institute of Technology, Rourkela	*Odisha, India*
2013	**Rank 1**, Dept. of Computer Science, West Bengal University of Technology	*West Bengal, India*
2009	**Recognition**, 1^{st} Rank from Kindergarten to Std. XII (all 15 years), Hill Top School	*Jamshedpur, India*

Professional Services

Reviewer

- IEEE Transactions on Parallel and Distributed Systems - 2019, 2020
- Cluster Computing Journal - 2019, 2020
- Advances in Science, Technology and Engineering Systems Journal (ASTESJ) - 2018

- International Journal of Grid and High Performance Computing (IJGHPC) - 2018, 2019
- Intelligent Automation & Soft Computing (AUTOSOFT) Journal - 2018
- Multiagent and Grid Systems (MGS) - An International Journal of Cloud Computing and Artificial Intelligence - 2017

External Reviewer

- IEEE Transactions on Services Computing '18, BigData '17/'18, Cluster '17/'18, ECOOP '20, HPDC '17/'18/'20, IC2E '17, ICDCS '17/'18/'19, ICS '17/'18, IPDPS '18/'19/'20, ICCD '19

Relevant Coursework

Virginia Tech

- Operating Systems, Research Methods in Computer Science, Cloud Computing, Software Refactoring, Statistics in Research, Data Analytics, Models of HCI

National Institute of Technology, Rourkela

- Cluster and Grid Computing, Advanced Computer Architecture, Design of Computer Networks, Software Design, Software Testing, Software Project Process and Quality Management, Software Engineering Requirement & Modeling, Software Architecture

References

Dr. Ali R. Butt - Virginia Tech, USA *<butta@cs.vt.edu>*

Dr. Ian T. Foster - Argonne National Laboratory, USA & University of Chicago, USA *<foster@anl.gov>*

Dr. Eli Tilevich - Virginia Tech, USA *<tilevich@cs.vt.edu>*

Dr. Bibhudatta Sahoo - National Institute of Technology, Rourkela, India *<bdsahu@nitrkl.ac.in>*