Arpan Jain

Website: aj-prime.github.io/aj-prime/
Email: arpan.jain1405@gmail.com
LinkedIn: aj-prime
GitHub: github.com/aj-prime

June, 2025

Research Interests

My research interests lie at the intersection of Deep Learning (DL) and High Performance Computing (HPC). I am actively contributing to distributed DNN training framework - LBANN, developing novel parallelization strategies to accelerate the DNN training on HPC systems, and optimizing inference methods.

Education

The Ohio State University Columbus, USA
Ph.D. in High Performance Deep Learning, Advisor: Prof. D. K. Panda 2018–2022

- GPA: 4.00/4.00
- Received Graduate Research Award from CSE Department in Apr 2022
- Thesis: “Novel Parallelization Strategies for High-Performance DNN Training on HPC Systems”

ABV-Indian Institute of Information Technology and Management Gwalior, India
Integrated Post Graduate 2013–2018

- GPA: 8.83/10.00
- Master Thesis: “Designing of Hybrid Machine Learning Model Based on Deep Learning and Its Performance Comparison”
- Bachelor Thesis: “VOP Detection and Significance in Recognition”

Select Publications

I am the lead author of the following publications

1. **A. Jain**, A. Shafi, Q. Anthony, P. Kousha, H. Subramoni, and D. Panda, “Hy-Fi: Hybrid Five-Dimensional Parallel DNN Training on High-Performance GPU Clusters,” in ISC High Performance 2022, May 2022.

2. **A. Jain**, N. Alnaasan, A. Shafi, H. Subramoni, and D. Panda, “Optimizing Distributed DNN Training using CPUs and BlueField-2 DPUs,” in IEEE Micro, doi: 10.1109/MM.2021.3139027.

3. **A. Jain**, T. Moon, T. Benson, H. Subramoni. S. Jacobs , DK Panda, and B. Essen. SUPER: SUb-Graph Parallelism for TransformERs, 35th IEEE International Parallel & Distributed Processing Symposium (IPDPS ’21) , May 2021.

4. **A. Jain**, A. A. Awan, A. Aljuhani , J. Hashmi , Q. Anthony , H. Subramoni , DK Panda , R. Machiraju , and A. Parwani GEMS: GPU Enabled Memory Aware Model Parallelism System for Distributed DNN Training, The International Conference for High Performance Computing, Networking, Storage, and Analysis (SC ’20), Jun 2020.

5. **A. Jain**, A. A. Awan, H. Subramoni, D. Panda, Scaling TensorFlow, PyTorch, and MXNet using MVAPICH2 for High-Performance Deep Learning on Frontera, 3rd Deep Learning on Supercomputers Workshop at SC19 (DLS), Nov 2019.

6. **A. Jain**, A. A. Awan, Q. Anthony, H. Subramoni, D. Panda, Performance Characterization of DNN Training using TensorFlow and PyTorch on Modern Clusters, 21st IEEE International Conference on Cluster Computing (Cluster '19), Sep 2019.

7. **A. Jain**, A. Mishra, A. Shukla, R. Tiwari, A novel genetically optimized convolutional neural network for traffic sign recognition: A new benchmark on Belgium and Chinese traffic sign datasets, Neural Processing Letters, 50(3), 2019.

8. **A. Jain**, A. Singh, A. Shukla, Vowel Onset Point Detection in Hindi Language Using Long Short-Term Memory, Information Systems Design and Intelligent Applications. Springer, Singapore, 2019.

Research and Development Experience

Microsoft Redmond, USA

Senior Researcher at AI Frameworks (CoreAI) Aug '24 - present

- Working on optimizing online inference for Large Language Models (LLMs), including OpenAI models.
- Developed internal inference stack based on SGLang
- Characterize and Evaluate the performance of state-of-the-art LLMs.
- Research and develop state-of-the-art solutions for LLM servings and optimize the performance for Microsoft use cases
- Collaborate with other relevant researchers or research groups to contribute to and advance research in Deep Learning inference

Applied Scientist 2 at Bing Ads Jan '23 - Aug '24

- Worked on optimizing online inference for DNN models.
- Saved over $1.5 million annually by optimizing GPU usage in online deployments through techniques such as dynamic batching, multi-GPU utilization, and model-specific optimizations.
- Characterize and Evaluate the performance of state-of-the-art DNNs.
- Explore new architectures like AMD GPUs and pipelines for Deep Learning Inference
- Explored Approximate Nearest Neighbors for Bing Ads use cases on NVIDIA GPUs and presented a talk at NVIDIA GTC 2024

Network Based Computing Lab at The Ohio State University Columbus, USA

Graduate Research Assistant Dec '18 - Dec '22

- Leading High-Performance Deep Learning project created by Prof. Panda.
- Leading newly created conversational AI for HPC project.
- Characterize and Evaluate the performance of distributed DNN training for Deep Learning frameworks like TensorFlow, PyTorch, MXNet on GPUs as well as CPUs.
- Explore new parallelization strategies for distributed training and co-design it with Deep Learning Frameworks like TensorFlow and communication middleware MVAPICH2
- Performance Regression and sanity testing software stacks that are released periodically (MVAPICH2, MVAPICH2-X, MVAPICH2-GDR, and OMB).

Microsoft Bellevue, USA

Researcher Intern Summer 2022

- Worked on DeepSpeed distributed DNN training middleware.
- Topic: Elastic Fault-Tolerant DNN Training for large-scale Transformer models

Lawrence Livermore National Laboratory Livermore, USA
Computational Research Student Intern in LBANN Group Summer 2020 & 2021

- Implemented sub-graph parallelism for multi-branch deep neural networks like Transformers and ResNeXt.
- Proposed novel design for sub-graph parallelism and optimized data parallelism for Transformers
- Worked on sub-grid parallelism for 2nd order optimizations (KFAC)
- Optimized communication in Hydrogen library for sub-graph parallelism and realized inter-grid communication for distributed matrices.
- Achieved better performance than Data Parallelism for in-core Transformer models.
- Paper accepted at IPDPS '21

CSE Department at The Ohio State University Columbus, USA
Graduate Teaching Assistant Aug '18 - Dec '18

- Grader for Artificial Intelligence 1 Course

ABV- Indian Institute of Information Technology and Management Gwalior, IND
Machine Learning Project Associate at Sponsored Research Consultancy Cell May '18 - Jul '18

- Worked on Automatic Speech Recognition Project sponsored by DEIT, Government of India
- Trained Convolutional Neural Network for traffic signs and tuned hyper-parameters using modified genetic algorithm.
- Co-mentor for B.Tech final year projects

Speech Markers Pune, IND
Machine Learning Trainee Jul '17 - May '18

- Worked on Speech Analysis, Speech Processing, Voice Activity Detection, and Deep Learning for speech.
- Developed speech indicators for Health Research using Deep Neural Networks

Busigence Bangalore, IND
Data Science Associate May '17 - Jul '17

- Developed automated solution for time series modeling using machine learning and deep learning libraries
- Worked on Prescriptive Analytics and Deep Learning models
- Separated speakers audio over phone call using unsupervised learning

Teaching and Mentoring

- **Instructor** at Ohio Super Computer Spring 21 & 22 and Autumn 22
  OSC AI Bootcamp for Professionals
- **Lab Assistant** at The Ohio State University Autumn 2020, 2021, & 2022
  Introduction to High-Performance Deep Learning (CSE 5449)
- **Teaching Assistant** at The Ohio State University Autumn 2018
  Introduction to Artificial Intelligence (CSE 3521/5521)
- **Lab Assistant** at ABV-IIITM Gwalior Spring 2017
  Advance Topics in Speech Processing
- **Research Project Lead in CSE5249** at The Ohio State University Autumn 2020
  Capstone course for research projects under Prof. Ramnath
  Members

  - *Tom Ballas*
  - *Zenqui Dong*
- **Research Project Lead in CSE5194.01** at The Ohio State University — Spring 2021
  Independent Group Studies under Prof. Ramnath
  Members
  - *Nawras Alnaasan*
  - *Rayan Hamza*
  - *Zenqui Dong*
- **Mentored PhD Students** at The Ohio State University
  - Nawras Alnaasan
  - Hyunho Ahn
  - Lang Xu
- **Mentored Graduate Students** at The Ohio State University
  - Saisree Reddy Miriyala
  - Sainath Prasanna
  - Ayyappa Kolli
  - Mingzhe Han
- **Mentored Undergraduate Students** at ABV-Indian Institute of Information Technology and Management
  - Hardik Khandelwal
  - Saloni Jain
  - Gatij Jain
  - Aditi Agarwal
  - Aishwarya Selvam

Professional Service

Memberships

1. IEEE Student Member
2. ACM Student Member
3. MLPerf HPC

Invited Tutorials (Introduction to High Performance Machine Learning and Deep Learning)

1. Hot Interconnects 27 (Hoti 2020) (Attendees: 85)
2. Principles and Practice of Parallel Programming (PPoPP 2021) (Attendees: 100+)
3. Architectural Support for Programming Languages and Operating Systems (ASPLOS 2021) (30+)
4. International Symposium on Computer Architecture (ISCA 2021) (25)
5. ISC High Performance 2021 (40+)

6. Hot Interconnects 28 (Hoti 2021) (55+)
7. Practice & Experience in Advanced Research Computing (PEARC 2021) (50+)
8. SuperComputing 2021 (SC21) (40+)
9. Principles and Practice of Parallel Programming (PPoPP 2022) (60+)
10. International Symposium on Computer Architecture (ISCA 2022) (30+)
11. ISC High Performance 2022 (20+)
12. MVAPICH User Group Conference (60+)
13. OFA Virtual Workshop (25+)
14. Hot Interconnects 29 (Hoti 2022) (45+)
15. IRDTA 8th International School on Deep Learning (DeepLearn 2023 Winter) (30+)
16. Principles and Practice of Parallel Programming (PPoPP 2023) (10+)

Reviewer

1. Hot Interconnects 30 (Hoti 2023) (Technical Program Committee member)
2. Transactions on Parallel and Distributed Systems 2023
3. Concurrency and Computation: Practice and Experience 2023
4. 35th IEEE International Parallel & Distributed Processing Symposium (IPDPS 2021)
5. 35th ACM International Conference on Supercomputing (ICS 2021)
6. HPCS 2021
7. Transactions on Parallel and Distributed Systems (Special Section on AI/ML/DL) 2020
8. 34th ACM International Conference on Supercomputing (ICS 2020)
9. 34th IEEE International Parallel and Distributed Processing Symposium (IPDPS 2020)
10. 37th IEEE International Conference on Computer Design (ICCD 2019)
11. 26th European MPI Users' Group Meeting (EuroMPI 2019)

Volunteer

1. First Midwestern Consortium Workshop for Computational Pathology (MCCP) 2021
2. SuperComputing (SC) '19, '20, '21, and '22
3. OSU Booth, SuperComputing (SC) '19 and '22
4. MVAPICH Users Group Meeting (MUG) '19, '20, '21, and '22

Student Mentoring Program:

1. SuperComputing (SC) 2021
2. SuperComputing (SC) 2020
3. ISCA 2020
4. SuperComputing (SC) 2019
5. IEEE Cluster 2019

Skills

- **DL Frameworks:** TensorFlow, PyTorch, LBANN
- **Computer Design:** Photoshop, Premiere
- **Concepts:** Nature-inspired algorithms, Fuzzy logic, Time Series modeling, Data Mining, Parallel Computing
- **Spoken Languages:** English, Hindi

Languages

- **Python:** Expert
- **MPI:** Intermediate
- **C:** Intermediate
- **C++:** Beginner
- **Matlab:** Beginner

Awards and Recognitions

- Graduate Research Award, CSE Department, The Ohio State University 2022
- AICTE M.Tech Scholarship 2017–2018
- Student Editor - Hindi Magazine 2014–2015
- Chief Minister Merit Scholarship 2013

Extracurricular Activities

- Founder and Student Coordinator of Journalism Club (UTHAAN) 2014–2018
- Member of Rotary Club at ABV-IIITM Gwalior 2013–2015
- Member of Entrepreneur-cell 2014
- Student Organizer in various Institute-level events like Kavi Sammelan, Convocation, Alumni Meet, etc. 2014–2017
- Conducted a session on Competitive Programming at ABV-IIITM Gwalior 2014

All Publications

Journals

1. **A. Jain**, N. Alnaasan, A. Shafi, H. Subramoni, and D. Panda, "Optimizing Distributed DNN Training using CPUs and BlueField-2 DPUs," in IEEE Micro, doi: 10.1109/MM.2021.3139027.
2. A. A. Awan, **A. Jain**, C-H. Chu, H. Subramoni, D. Panda, Communication Profiling and Characterization of Deep Learning Workloads on Clusters with High-Performance Interconnects, IEEE Micro, 2019.

3. **A. Jain**, A. Mishra, A. Shukla, R. Tiwari, A novel genetically optimized convolutional neural network for traffic sign recognition: A new benchmark on Belgium and Chinese traffic sign datasets, Neural Processing Letters, 50(3), 2019.

Refereed Conference/Workshop Papers

[1] N. Alnaasan, **A. Jain**, A. Shafi, H. Subramoni, and D. Panda, “Accdp: Accelerated data-parallel distributed dnn training for modern gpu-based hpc clusters”, Bengaluru, India, Dec. 2022.

[2] N. Alnaasan, **A. Jain**, A. Shafi, H. Subramoni, and D. Panda, “Omb-py: Python micro-benchmarks for evaluating performance of mpi libraries on hpc systems”, in *23rd Parallel and Distributed Scientific and Engineering Computing Workshop (PDSEC) at IPDPS22*, Aug. 2022.

[3] **A. Jain**, A. Shafi, Q. Anthony, P. Kousha, H. Subramoni, and D. Panda, “Hy-fi: Hybrid five-dimensional parallel dnn training on high-performance gpu clusters”, in *ISC High Performance (ISC’22)*, May 2022.

[4] P. Kousha, **A. Jain**, A. Kolli, S. Miriyala, S. Sainath, H. Subramoni, A. Shafi, and D. Panda, “”hey cai” - enhancing user productivity through a conversational ai enabled user interface for hpc tools”, in *ISC High Performance (ISC’22)*, May 2022.

[5] **A. Jain**, N. Alnaasan, A. Shafi, H. Subramoni, and D. Panda, “Accelerating cpu-based distributed dnn training on modern hpc clusters using bluefield-2 dpus”, in *28th IEEE Hot Interconnects (HotI28)*, Aug. 2021.

[6] P. Kousha, K. R. Sankarapandian Dayala Ganesh Ram, M. Kedia, H. Subramoni, **A. Jain**, A. Shafi, D. Panda, T. Dockendorf, H. Na, and K. Tomko, “Inam: Cross-stack profiling and analysis of communication in mpi-based applications”, in *Practice and Experience in Advanced Research Computing*, ser. PEARC ’21, Boston, MA, USA: Association for Computing Machinery, 2021, ISBN: 9781450382922.

[7] **A. Jain**, T. Moon, T. Benson, H. Subramoni, S. Jacobs, D. Panda, and B. Essen, “Super: Sub-graph parallelism for transformers”, in *35th IEEE International Parallel & Distributed Processing Symposium (IPDPS)*, May 2021.

[8] **A. Jain**, A. Awan, A. Aljuhani, J. Hashmi, Q. Anthony, H. Subramoni, D. Panda, R. Machiraju, and A. Parwani, “Gems: Gpu-enabled memory-aware model-parallelism system for distributed dnn training”, in *2020 SC20: International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, Los Alamitos, CA, USA: IEEE Computer Society, Nov. 2020, pp. 621–635.

[9] A. A. Awan, **A. Jain**, Q. Anthony, H. Subramoni, and D. K. Panda, “Hypar-flow: Exploiting mpi and keras for scalable hybrid-parallel dnn training with tensorflow”, in *High Performance Computing*, P. Sadayappan, B. L. Chamberlain, G. Juckeland, and H. Ltaief, Eds., Cham: Springer International Publishing, Jul. 2020, pp. 83–103, ISBN: 978-3-030-50743-5.

[10] Q. Anthony, A. A. Awan, **A. Jain**, H. Subramoni, and D. K. Panda, “Efficient training of semantic image segmentation on summit using horovod and mvapich2-gdr”, in *2020 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, 2020, pp. 1015–1023.

[11] P. Kousha, B. Ramesh, K. Kandadi Suresh, C. Chu, **A. Jain**, N. Sarkauskas, H. Subramoni, and D. K. Panda, “Designing a profiling and visualization tool for scalable and in-depth analysis of high-performance gpu clusters”, in *2019 IEEE 26th International Conference on High Performance Computing, Data, and Analytics (HiPC)*, 2019, pp. 93–102.

[12] **A. Jain**, A. A. Awan, H. Subramoni, and D. K. Panda, “Scaling tensorflow, pytorch, and mxnet using mvapich2 for high-performance deep learning on frontera”, in *2019 IEEE/ACM Third Workshop on Deep Learning on Supercomputers (DLS)*, 2019, pp. 76–83.

[13] **A. Jain**, A. A. Awan, Q. Anthony, H. Subramoni, and D. K. D. Panda, "Performance characterization of dnn training using tensorflow and pytorch on modern clusters", in *2019 IEEE International Conference on Cluster Computing (CLUSTER)*, 2019, pp. 1–11.

[14] A. A. Awan, **A. Jain**, C. Chu, H. Subramoni, and D. K. Panda, "Communication profiling and characterization of deep learning workloads on clusters with high-performance interconnects", in *2019 IEEE Symposium on High-Performance Interconnects (HOTI)*, 2019, pp. 49–53.

[15] **A. Jain**, A. Singh, and A. Shukla, "Vowel onset point detection in hindi language using long short-term memory", in *Information Systems Design and Intelligent Applications*, Springer, 2019, pp. 505–515.

[16] S. Sahu, **A. Jain**, R. Tiwari, and A. Shukla, "Application of Egyptian Vulture Optimization in Speech Emotion Recognition", in *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, 2018, pp. 230–234.

[17] **A. Jain** and A. Shukla, "Anomaly detection in speech signals using variational autoencoder", in *23rd International Symposium on Frontiers of Research in Speech and Music (FRSM '17)*, Rourkela, India, Dec. 2017.

References

I have collaborated with top researchers in the field and I can request reference/recommendation letters from them if needed. My most recent references are:

1. Dhabaleswar Kumar (DK) Panda, Professor.
   Dept. of Computer Science and Engineering
   The Ohio State University
   2015 Neil Avenue
   Columbus, OH-43210, USA
   Tel: (614) 292-5199
   Email: panda@cse.ohio-state.edu
   Website: http://web.cse.ohio-state.edu/ panda.2/
   Twitter: @dhabalkpanda

2. Brian Van Essen, Informatics Group leader.
   Computation/Center for Applied Scientific Computing
   Lawrence Livermore National Laboratory
   7000 East Avenue
   Livermore, CA-94550, USA
   Email: vanessen1@llnl.gov
   Tel: +1(925)-422-9300

3. Anupam Shukla, Director.
   Sardar Vallabhbhai National Institute of Technology,
   Surat, India
   Tel: +91-9575048000
   Email: dranupamshukla@gmail.com
   LinkedIn: www.linkedin.com/in/anupam-shukla-4a11a628/