Ammar Ahmad Awan

555 110th Ave NE • Bellevue • WA 98004
+1 614 360 8349 • ammar.ahmad.awan@gmail.com
https://awan-10.github.io

Research Interests

My broad interests lie at the interesection of High Performance Computing (HPC) and Machine Learning (ML). I am actively investigating new approaches to improve performance and productivity of scalable software systems for HPC and ML.

Education

The Ohio State University (OSU), Columbus, Ohio, USA
Ph.D. in Computer Science and Engineering, Aug 2014—May 2020
Advisor: D.K. Panda • CGPA: 3.68/4.0
Thesis: Co-designing MPI Middleware and DL Frameworks for High-Performance DNN Training on HPC Systems

Kyung Hee University (KHU), Suwon, South Korea
Master of Computer Engineering, 2011—2013
Advisor: Sungyoung Lee • CGPA: 4.22/4.3
Thesis: Efficient Support for Parallel File Access in Java HPC

National University of Sciences and Technology (NUST), Islamabad, Pakistan
Bachelor of Information Technology, 2004—2008
Advisor: Aamir Shafi • CGPA: 3.71/4.0
Final Project: Optimizing N-body Simulations for Multicore Compute Clusters

Select Publications

I am the lead author of the following publications.

1. **A. A. Awan**, A. Jain, Q. Anthony, H. Subramoni, and DK Panda,*HyPar-Flow: Exploiting MPI and Keras for Scalable Hybrid-Parallel DNN Training using TensorFlow*, ISC High-Performance (**ISC '20**), June 2020.
2. **A. A. Awan**, C-H Chu, X. Lu, H. Subramoni, and D. K. Panda,*OC-DNN: Exploiting Advanced Unified Memory Capabilities in CUDA 9 and Volta GPUs for Out-of-Core DNN Training*, 25th IEEE International Conference on High-Performance Computing, Data, Analytics, and Data Science (**HiPC '18**) '18, Dec 2018.
3. **A. A. Awan**, C-H Chu, X. Lu, H. Subramoni, and DK Panda,*Can Unified-Memory support on Pascal and Volta GPUs enable Out-of-Core DNN Training?*, ISC High-Performance (**ISC '18**), June 2018. **Best Student Poster Award**.
4. **A. A. Awan**, K. Hamidouche, J. Hashmi, and D. K. Panda,*S-Caffe: Co-designing MPI Runtimes and Caffe for Scalable Deep Learning on Modern GPU Clusters*, 22nd ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (**PPoPP '17**), Feb 2017.
5. **A. A. Awan**, K. Hamidouche, A. Venkatesh, and D. K. Panda,*Efficient Large Message Broadcast using NCCL and CUDA-Aware MPI for Deep Learning*, 23rd European MPI Users' Group Meeting (**EuroMPI '16**), Sep 2016. **Best Paper Runner-Up**.

Awards and Distinctions

1. *Graduate Research Award (2020).* Awarded by CSE Department at OSU to Top 3 PhD Graduates every year. Value: USD 1,000.
2. *O'Donnell Fellowship (2014—2015).* Awarded to Top 5 (out of > 1,000 applicants) for first year of Ph.D. studies at OSU. Value: USD 54,685.

3. *Global IT Talents Scholarship (2011—2013)*. Awarded to top candidates for pursuing Masters in Computer Science at South Korean universities.
4. *President's Gold Medal (2008)*. Highest CGPA in Bachelors Degree at NUST, Pakistan.
5. *Rector's Gold Medal (2008)*. Best Final Year Project at NUST, Pakistan.
6. *Best Industry Project Award (2008)*. Best Final Year Project at NUST-SEECS Open House '08.
7. *Merit Scholarship for 7/8 semesters (2004—2008)*. Awarded to students at NUST with GPA above 3.5.
8. *IEEE TCHPC Travel Award* for presenting Doctoral Showcase at SC '19.
9. *ACM Student Travel Award* for participating in ACM Student Research Competition at SC '17.
10. *NSF Student Travel Award* for presenting S-Caffe at ACM PPoPP '17.
11. *Student Travel Award* for presenting Tutorial at HotI '17.
12. *Best Student Poster Award* at ISC High-Performance Event (ISC '19).
13. *Best Paper Runner-up* at EuroMPI 2016, Edinburgh, UK.
14. *Third Prize* for presenting Project: Constella Platinum at All Pakistan software competition - Softcom '06.
15. *Student Volunteer* for SC '08, USA. (Selected but couldn't travel).

Research and Development Experience

- Microsoft
  - Researcher (Jun '20 — present)
    - Investigate Large-scale Communication and Computation challenges in the context of Machine and Deep Learning
    - Contribute to design and implementation of an open-source and scalable software called DeepSpeed (http://www.deepspeed.ai). DeepSpeed is being developed by the team led by Dr. Yuxiong He.
- Network Based Computing Lab (http://nbcl.cse.ohio-state.edu) at The Ohio State University
  - Graduate Research Assistant (Aug '14 – May '20)
    - Investigate Collective Communication Designs and Implementations for CUDA-Aware MPI libraries like *MVAPICH2* and *MVAPICH2-GDR*.
    - Co-design Deep Learning frameworks like Caffe and MPI runtimes like *MVAPICH2* to enable efficient distributed Deep Learning on modern GPU clusters.
    - Utilize existing benchmark suites like OSU Microbenchmarks (OMB), Intel MPI benchmarks (IMB) and test suites like MPICH tests, Intel tests, etc. to rigorously test and evaluate new designs on multiple HPC systems with diverse set of CPU and GPU architectures.
    - Perform regression/sanity testing on software stacks that are released periodically as new features from several research students and staff are developed and pushed to the main *MVAPICH2* codebase.
    - Design new benchmarks to evaluate the capabilities of the MVAPICH2 MPI library as well as OSU-Caffe and other DL stacks like Horovod for TensorFlow and PyTorch on large-scale HPC systems.

Note: *MVAPICH2* is a popular and open-source MPI Library being used by more than 3,000 organizations around the world. It has been downloaded 613,000 times directly from the project site (http://mvapich.cse.ohio-state.edu).

- X-Scale Solutions (http://x-scalesolutions.com), Columbus, OH
  - Research Intern (May '19—Aug '19)
    - Conducted in-depth performance characterization of TensorFlow/Horovod on Large Scale HPC Systems like Summit (#1 on Top500) and Sierra (#2 on Top500) using X-ScaleAI.
    - Implemented one-click installers for X-Scale products (X-ScaleAI and X-ScaleHPC).
- Microsoft Research (MSR), Redmond, WA
  - Research Intern with the RiSE Group at MSR (May '18 – Aug'18)
    - Mentors: Madan Musuvathi, Todd Mytkowicz, and Saeed Maleki.
    - Assisted in design and evaluation of semantics-preserving SGD codes that scale to hundreds of CPUs.
    - Designed and developed code/experiments to evaluate Criteo's Ad-click prediction at scale using TensorFlow on Cloud-based systems like Google Cloud ML, Amazon SageMaker, and Azure BatchAI.

- iFaST Solutions Pvt. Ltd, Peshawar, Pakistan
  - Vice President: Innovation (Jun ‘13 – Jun ‘14)
    - Developed tutorials and delivered talks on Version Control (Git) and use of PHP frameworks (CodeIgniter) to transform internal processes. This helped to avoid software development delays faced by the company.
- Ubiquitous Computing Laboratory, Kyung Hee University, South Korea
  - Graduate Research Assistant (Aug ‘11 – Jun ‘13)
    - Co-founded the HPC over Cloud (HPCoC) project for the team.
    - Published two papers on Parallel I/O for Java HPC project.
- Skylight Software Inc., CA and Islamabad, Pakistan
  - Principle Software Engineer (Apr ‘11 – Jul ‘11)
    - Designed and implemented a state-charts based approach for developing efficient custom controls for a new document format proposed by Skylight.
- NUST-SEECS, Pakistan (Feb ‘08 – Nov ‘09) / University of Reading, UK (Feb ‘09 – Jun ‘09)
  - Research Assistant
    - Analyzed and profiled performance of Gadget-2 code and proposed hybrid-parallelism to speed-up the simulations on multi-core clusters.

Teaching and Mentoring Experience

- Mentored undergradute and graduate students at The Ohio State University to work on various research and development projects.
  - Arpan Jain, Ph.D. student at OSU
  - Quentin Anthony, Ph.D. Student at OSU
  - Vardaan Gangal, B.S Student at OSU
- Mentored seven prospective M.S and Ph.D. students for GradAppLab (http://gradapplab.pk)
- Developed and designed the overall curriculum, lectures, homework assignments, and labs for special-topic graduate course at OSU: *CSE 5194.01: Introduction to High Performance Deep Learning* (Autumn '18 and Autumn '19)

All Publications

Most updated list of publications is available from my Google Scholar page.

Journal Articles

1. **A. A. Awan**, A. Jain, C-H Chu, H. Subramoni, and DK Panda, *Communication Profiling and Characterization of Deep Learning Workloads on Clusters with High-Performance Interconnects*, IEEE Micro (Early Access: doi: 10.1109/MM.2019.2949986).
2. **A. A. Awan**, K. V. Manian, C-H Chu, H. Subramoni, and DK Panda, *Optimized Large-Message Broadcast for Deep Learning Workloads: MPI, MPI+NCCL, or NCCL2?*, Parallel Computing (PARCO '19), Vol. 85, Pages 141-152, July 2019.
3. C-H Chu, X. Lu, **A. A. Awan**, H. Subramoni, Bracy Elton, and DK Panda, *Exploiting Hardware Multicast and GPUDirect RDMA for Efficient Broadcast*, IEEE Transactions on Parallel and Distributed Systems (TPDS '19), Vol. 30, No. 3, Pages 575-588, Mar 2019.
4. K. Hamidouche, A. Venkatesh, **A. A. Awan**, H. Subramoni, and D. K. Panda, *CUDA-Aware OpenSHMEM: Extensions and Designs for High Performance OpenSHMEM on GPU Clusters*, Parallel Computing (PARCO '16), Vol. 58, Pages 27-36, Oct 2016.
5. Z. Pervez, **A. A. Awan**, A. M. Khattak, S. Y. Lee, and Eui-Nam Huh, *Privacy-aware searching with oblivious term matching for cloud storage*, Journal of Supercomputing, Vol. 63, Issue 2, Pages 538–560, Feb 2013.

Refereed Conference/Workshop Papers

1. **A. A. Awan**, A. Jain, Q. Anthony, H. Subramoni, and DK Panda,*HyPar-Flow: Exploiting MPI and Keras for Scalable Hybrid-Parallel DNN Training using TensorFlow*, ISC High-Performance (**ISC '20**), June 2020 (Accepted to be presented).
2. A. Jain, **A. A. Awan**, H. Subramoni, and DK Panda,*Scaling TensorFlow, PyTorch, and MXNet using MVAPICH2 for High-Performance Deep Learning on Frontera*, 3rd Deep Learning on Supercomputers Workshop, held in conjunction with SC '19, Nov 2019.
3. A. Jain, **A. A. Awan**, Q. Anthony, H. Subramoni, and DK Panda,*Performance Characterization of DNN Training using TensorFlow and PyTorch on Modern Clusters*, 21st IEEE International Conference on Cluster Computing, (Cluster '19), Sep 2019.
4. **A. A. Awan**, A. Jain, C-H Chu, H. Subramoni, and D. K. Panda,*Communication Profiling and Characterization of Deep Learning Workloads on Clusters with High-Performance Interconnects*, 26th Symposium on High-Performance Interconnects (HotI '19), Aug 2019.
5. **A. A. Awan**, J. Bedorf, C-H Chu, H. Subramoni, and D. K. Panda,*Scalable Distributed DNN Training using TensorFlow and CUDA-Aware MPI: Characterization, Designs, and Performance Evaluation*, 19th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid '19), May 2019.
6. K. Vadambacheri Manian, **A. A. Awan**, A. Ruhela, C. Chu, and D. K. Panda,*Characterizing CUDA Unified Memory (UM)-Aware MPI Designs on Modern GPU Architectures*, 12th Workshop on General Purpose Processing Using GPU (GPGPU '19), held in conjunction with ASPLOS '19, Apr 2019.
7. **A. A. Awan**, C-H Chu, X. Lu, H. Subramoni, and D. K. Panda,*OC-DNN: Exploiting Advanced Unified Memory Capabilities in CUDA 9 and Volta GPUs for Out-of-Core DNN Training*, IEEE 25th International Conference on High Performance Computing (HiPC '18), Dec 2018.
8. **A. A. Awan**, C-H Chu, H. Subramoni, D. K. Panda, *Optimized Broadcast for Deep Learning Workloads on Dense-GPU InfiniBand Clusters: MPI or NCCL?,* 25th European MPI Users' Group Meeting (EuroMPI '18), Sep 2018.
9. **A. A. Awan**, H. Subramoni, D. K. Panda,*An In-depth Performance Characterization of CPU- and GPU-based DNN Training on Modern Architectures*, 3rd Workshop on Machine Learning in HPC Environments (MLHPC '17), held in conjunction with SC '17, Nov 2017.
10. C-H Chu, X. Lu, **A. A. Awan**, H. Subramoni, J. Hashmi, Bracy Elton, and DK Panda, *Efficient and Scalable Multi-Source Streaming Broadcast on GPU Clusters for Deep Learning*, 46th International Conference on Parallel Processing (ICPP '17), Aug 2017.
11. **A. A. Awan**, K. Hamidouche, J. Hashmi, and D. K. Panda,*S-Caffe: Co-designing MPI Runtimes and Caffe for Scalable Deep Learning on Modern GPU Clusters*, 22nd ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP '17), Feb 2017.
12. K. Hamidouche, **A. A. Awan**, A. Venkatesh, and D. K. Panda,*CUDA M3: Designing Efficient CUDA Managed Memory-aware MPI by Exploiting GDR and IPC*, 23rd IEEE International Conference on High Performance Computing, Data, and Analytics, Dec 2016.
13. **A. A. Awan**, K. Hamidouche, A. Venkatesh, and D. K. Panda,*Efficient Large Message Broadcast using NCCL and CUDA-Aware MPI for Deep Learning*, 23rd European MPI Users' Group Meeting (EuroMPI '16), Sep 2016.**Best Paper Runner-Up**.
14. C. Chu, K. Hamidouche, A. Venkatesh, **A. A. Awan**, and D. K. Panda, *CUDA Kernel based Collective Reduction Operations on Large-scale GPU Clusters*, 16th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid '16), May 2016.
15. **A. A. Awan**, K. Hamidouche, A. Venkatesh, J. Perkins, H. Subramoni, and D. K. Panda,*GPU-Aware Design, Implementation, and Evaluation of Non-blocking Collective Benchmark*, 22nd European MPI Users' Group Meeting (EuroMPI '15), Sep 2015.
16. K. Hamidouche, A. Venkatesh, **A. A. Awan**, H. Subramoni, and D. K. Panda,*Exploiting GPUDirect RDMA in Designing High Performance OpenSHMEM for NVIDIA GPU Clusters*, IEEE International Conference on Cluster Computing (Cluster '15), Sep 2015.
17. **A. A. Awan**, K. Hamidouche, C. Chu, and D. K. Panda,*A Case for Non-Blocking Collectives in OpenSHMEM: Design, Implementation, and Performance Evaluation using MVAPICH2-X*, Workshop on OpenSHMEM and Related Technologies (OpenSHMEM '15), Aug 2015.
18. H. Subramoni, **A. A. Awan**, K. Hamidouche, D. Pekurovsky, A. Venkatesh, S. Chakraborty, K. Tomko, and D. K. Panda, *Designing Non-Blocking Personalized Collectives with Near Perfect Overlap for RDMA-Enabled Clusters*, ISC High Performance (ISC '15), Jul 2015.
19. S. Chakraborty, H. Subramoni, J. Perkins, **A. A. Awan**, and D. K. Panda, *On-demand Connection Management for OpenSHMEM and OpenSHMEM+MPI* (HIPS '15), IPDPS Workshop, May 2015.

20. **A. A. Awan**, M. S. Ayub, A. Shafi and S. Lee, *Towards Efficient Support for Parallel I/O in Java HPC*, 13th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT '12), Dec 2012.
21. M. B. Amin, W. A. Khan, **A. A. Awan**, and S. Y. Lee, "Intercloud Message Exchange Middleware", 6th International Conference on Ubiquitous Information Management and Communication (ICUIMC '12), Sep 2012.

Posters

1. **A. A. Awan** and DK Panda, *Co-designing Communication Middleware and Deep Learning Frameworks for High-Performance DNN Training on HPC Systems*, Doctoral Showcase at SC '19, Nov 2019.
2. **A. A. Awan**, H. Subramoni, and DK Panda, *Exploiting CUDA Unified Memory for Efficient Out-of-Core DNN Training*, Poster at NVIDIA GTC '19, April 2019.
3. **A. A. Awan**, C-H Chu, X. Lu, H. Subramoni, and DK Panda, *Can Unified-Memory support on Pascal and Volta GPUs enable Out-of-Core DNN Training?*, ISC High-Performance (ISC '18), Jun 2018. **Best Student Poster Award**.
4. **A. A. Awan** and DK Panda, *Co-designing MPI Runtimes and Deep Learning Frameworks for Scalable Distributed Training on GPU Clusters*, ACM Student Research Competition (SRC) poster at SC '17, Nov 2017.
5. **A. A. Awan**, M. B. Amin, S. Hussain, A. Shafi, S. Y. Lee, *An MPI-IO Compliant Java based Parallel I/O Library*, Poster at 13th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid '13), May 2013.

Talks

1. *Benchmarking Deep Learning Workloads on Large-scale HPC Systems* (Invited Talk), Benchmarking in the Data Center Workshop, PPoPP '20, Feb 2020.
2. *Co-designing Communication Middleware and Deep Learning Frameworks for High-Performance DNN Training on HPC Systems*, Doctoral Showcase Presentation at SC '19, Nov 2019.
3. *An In-depth Performance Characterization of CPU- and GPU-based DNN Training on Modern Architectures*, MLHPC '17, SC '17 Workshop, Nov 2017.
4. *S-Caffe: Co-designing MPI Runtimes and Caffe for Scalable Deep Learning on Modern GPU Clusters*, PPoPP '17, Feb 2017.
5. *Efficient Large Message Broadcast using NCCL and CUDA-Aware MPI for Deep Learning*, Best Paper Runner-up Session, EuroMPI '16 @ EPCC Edinburgh UK, Sep 2016.
6. *Why Execution is more important than Ideas*, Invited Talk at CECOS University, Peshawar, Pakistan, Feb 2014.

Invited Tutorials

Number of Attendees are in parentheses.

1. *High Performance Distributed Deep Learning: A Beginner's Guide*, HotI '20 (Accepted; to be presented).
2. *High Performance Distributed Deep Learning: A Beginner's Guide*, ISCA '20, May '20 (35).
3. *High Performance Distributed Deep Learning: A Beginner's Guide*, NVIDIA GTC '20.
4. *High Performance Distributed Deep Learning*, PPoPP '20, Feb 2020. (25)
5. *High Performance Distributed Deep Learning: A Beginner's Guide*, SC '19, Nov 2019. (120)
6. *High Performance Architectures for Distributed Deep Learning*, MICRO '19, Oct 13, 2019. (60)
7. *HPC Meets Distributed Deep Learning*, Hot Interconnects (HotI '19), Aug 14, 2019. (50)
8. *High-Performance Distributed Deep Learning: A Beginner's Guide*, PEARC '19, Jul 29, 2019. (80)
9. *High-Performance Distributed Deep Learning: A Beginner's Guide*, ISCA '19, Jun 22, 2019. (40)
10. *High-Performance Distributed Deep Learning: A Beginner's Guide*, ISC '19, Jun 16, 2019. (40)
11. *High-Performance Distributed Deep Learning: A Beginner's Guide*, CCGrid '19, May 15, 2019. (40)
12. *High-Performance Distributed Deep Learning: A Beginner's Guide*, NCAR SEA '19, Apr 12, 2019. (10)
13. *How to Boost the Performance of HPC/AI Applications Using MVAPICH2 Library* NVIDIA GTC '19, Mar 20, 2019. (50)
14. *High-Performance Distributed Deep Learning: A Beginner's Guide*, NVIDIA GTC '19, Mar 18, 2019. (100)
15. *High-Performance Distributed Deep Learning: A Beginner's Guide*, PPoPP '19, Feb 17, 2019. (15)
16. *High-Performance Distributed Deep Learning: A Beginner's Guide*, DOD-PETTT '18, May 15, 2018. (25)

17. *High-Performance Distributed Deep Learning: A Beginner's Guide*, NCAR SEA '18, Apr 5, 2018. (30)
18. *High-Performance Distributed Deep Learning: A Beginner's Guide*, PPoPP '18, Feb 25, 2018. (20)
19. *High-Performance Distributed Deep Learning for Dummies*, IT4 Innovations (Austria), Jan 24, 2018. (35)
20. *High Performance Distributed Deep Learning for Dummies*, Hot Interconnects (HotI '17) Aug 28, 2017. (50)

Professional Service

Memberships

1. MLPerf - https://mlperf.org/
2. MLPerf HPC - https://groups.google.com/forum/#!forum/mlperf-hpc
3. ACM Student Member
4. IEEE Student Member
5. Message Passing Interface (MPI) Forum - https://www.mpi-forum.org/

Reviewer

1. 40th IEEE International Conference on Distributed Computing Systems (ICDCS '20).
2. Elsevier SoftwareX Journal (2020).
3. 34th IEEE International Parallel & Distributed Processing Symposium (IPDPS '20).
4. The FREE Python conference in Columbus (PyOhio '19).
5. 32nd ACM International Conference on Supercomputing (ICS '18).
6. Intl. Conference on High Performance Computing, Networking, Storage, and Analysis (SC '17).
7. 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID '17).
8. 26th International Conference on Parallel Architectures and Compilation Techniques (PACT '17).
9. 31st IEEE International Parallel & Distributed Processing Symposium (IPDPS '17).
10. IEEE Transactions on Parallel and Distributed Systems.
11. ISC High Performance 2016 (ISC '16).
12. Elsevier Journal of Parallel and Distributed Computing.

Volunteer

1. OSU Booth, Supercomputing (SC) '17, '18, and '19.
2. MVAPICH Users Group Meeting (MUG) '16, '17, and '19.
3. IEEE ICDCS 2015.

Technical Skills

- Strong programming skills in C and Java (SE)/Java for HPC.
- Development experience in C and interaction of C, C\, Python and MPI.
- Product-development experience (Skylight Software) using C and Win32 programming.
- Experience of developing parallel programs using OpenMP, MPI and MPJ Express.
- Familiar with C#, ASP.NET, Android SDK, PHP, MySQL, IBM Cell SDK, and PerfAPI (PAPI)/Perfex.
- Understanding of web technologies including HTML, DHTML, CSS, XML, XSLT and XPath.
- Strong communication and presentation skills
    - Delivered several elaborate presentations on technical projects like OSU-Caffe, High-Performance Deep Learning (HiDL), MVAPICH2, Constella, Gadget-2, Oil Reservoir Simulators, and MPJ-IO.

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. Yuxiong He, Research Manager

   Microsoft

Email: yuxhe@microsoft.com

2. 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

3. Gagan Agrawal, Professor.

School of Computer and Cyber Sciences
Augusta University
Augusta, GA 30912, USA
Email: gagrawal@augusta.edu

4. Radu Teodorescu, Associate Professor.

Dept. of Computer Science and Engineering
The Ohio State University
2015 Neil Avenue
Columbus, OH-43210, USA
Email: teodores@cse.ohio-state.edu
Website: http://web.cse.ohio-state.edu/~teodorescu.1/