Ammar Ahmad Awan

2015 Neil Ave. • Columbus • OH 43210 • USA
+1 614 360 8349 • ammar.ahmad.awan@gmail.com
http://cse.osu.edu/~awan.10

Research Interests

[red] obvious High Performance Computing, Deep Learning, Scalable AI Systems, Parallel Computing, and Performance Evaluation and Characterization of High-Performance Systems

Education

The Ohio State University (OSU), Columbus, Ohio, USA
Ph.D. in Computer Science and Engineering, Aug 2014—May 2020 (Expected)
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

Research and Development Experience

- Network Based Computing Lab (http://nbcl.cse.ohio-state.edu) at The Ohio State University, Columbus, OH
    - Graduate Research Assistant (Aug '14 – present)
        - Investigate Collective Communication Designs and Implementations for CUDA-Aware MPI libraries like 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.
- X-Scale Solutions (http://x-scalesolutions.com), Columbus, OH
    - Research Intern (May '19—Aug '19)

        - Conduct in-depth performance characterization of TensorFlow/Horovod on Large Scale HPC Systems like Summit (#1 on Top500) and Sierra (#2 on Top500).
        - Implementation of user-friendly installers for X-Scale products including X-ScaleAI and X-ScaleHPC.
- Microsoft Research, Redmond, WA
    - Research Intern (May '18 – Aug'18)
        - Assisted in design and evaluation of semantics-preserving SGD codes for CPUs in Azure cloud.
        - Developed code to evaluate Criteo's Ad clicks 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 to overcome 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)

Awards and Distinctions

1. "IEEE TCHPC Travel Award" for attending Supercomputing, SC '19.
2. "ACM Student Travel Award" for participating in ACM Student Research Competition at SC '17.
3. "NSF Student Travel Award" for attending ACM PPoPP '17.
4. "Student Travel Award" for attending HotI '17.
5. "Best Student Poster Award" at ISC High-Performance Event (ISC '19).
6. "Best Paper Runner-up" EuroMPI 2016, Edinburgh, UK.
7. "O'Donnell Fellowship" for Ph.D. in Computer Science and Engineering, OSU, USA (2014 - 2015).
8. "Global IT Talents Program Scholarship" for Masters Degree in South Korea (2011 - 2013).

9. **“President’s Gold Medal”** for highest CGPA in Bachelors Degree (NUST - 2008).
10. **“Rector’s Gold Medal”** for Best Final Year Project (NUST - 2008).
11. **“Best Industry Project”** award for the Final Year Project at NUST-SEECS Open House '08.
12. **“Merit Scholarship”** for 7 out of 8 semesters at NUST. (Awarded to students with 3.5 and above GPA).
13. **“Third Prize”** for presenting Project: Constella Platinum at All Pakistan software competition - Softcom '06.
14. **“Student Volunteer”** for SC ‘08, USA. (Selected but couldn’t travel).

Publications

Select

1. [PARCO '19] 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**, Volume 85, Jul '19, Pages 141-152, https://doi.org/10.1016/j.parco.2019.03.005.
2. [HiPC '18] 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, Dec '18.
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 '18. **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 ’17.
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 '16. **Best Paper Runner-Up**.

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 ’19 (to appear in).
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, Volume 85, July 2019, Pages 141-152, https://doi.org/10.1016/j.parco.2019.03.005.
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 '18), vol. 30, no. 3.
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, Volume 58, October 2016, Pages 27-36.
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 (2013).

Refereed Conference/Workshop Papers

1. 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, to be held in conjunction with SC ‘19.

2. 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).
3. 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).
4. 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”, in Proceedings of IEEE/ACM CCGrid '19.
5. 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 with ASPLOS '19.
6. 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”, in Proceedings of IEEE HiPC '18.
7. 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?”, in Proceedings of EuroMPI ’18.
8. 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 with SC ’17.
9. 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”, International Conference on Parallel Processing (ICPP), Aug ’17.
10. 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, Feb ’17.
11. 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 ’16.
12. 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). Best Paper Runner-Up.
13. 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).
14. 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”, EuroMPI ’15.
15. 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 Cluster 2015, Sep ’15.
16. 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”, OpenSHMEM 2015 for PGAS Programming in the Exascale Era, Aug ‘15.
17. 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 2015 (ISC '15), Jul ’15.
18. 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 ’15.
19. A. A. Awan, M. S. Ayub, A. Shafi and S. Lee, "Towards Efficient Support for Parallel I/O in Java HPC," 2012 13th International Conference on Parallel and Distributed Computing, Applications and Technologies, Beijing, Dec ’12.
20. 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).

Posters

1. A. A. Awan and DK Panda, "Co-designing Communication Middleware and Deep Learning Frameworks for High-Performance DNN Training on HPC Systems", (To be presented), Doctoral Showcase @ SC '19, Denver, CO, Nov '19.
2. A. A. Awan, H. Subramoni, and DK Panda, "Exploiting CUDA Unified Memory for Efficient Out-of-Core DNN Training", NVIDIA GTC '19, San Jose, April '19.
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, Germany, June, 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, Denver, CO, Nov '17.
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)”, 13th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid '13), Delft, Netherlands, May ’13.

Talks/Invited Tutorials

1. “High Performance Distributed Deep Learning: A Beginner’s Guide”, NVIDIA GTC ’20 (Accepted; To be presented).
2. “High Performance Distributed Deep Learning: A Beginner’s Guide”, SC ’19
3. “High Performance Architectures for Distributed Deep Learning”, MICRO ’19, Oct 13, 2019.
4. “HPC Meets Distributed Deep Learning”, Hot Interconnects (HotI) ’19, Aug 14, 2019.
5. “High-Performance Distributed Deep Learning: A Beginner’s Guide”, PEARC '19, Jul 29, 2019.
6. “High-Performance Distributed Deep Learning: A Beginner’s Guide", ISCA '19, Jun 22, 2019.
7. “High-Performance Distributed Deep Learning: A Beginner’s Guide”, ISC '19, Jun 16, 2019.
8. “High-Performance Distributed Deep Learning: A Beginner’s Guide”, CCGrid '19, May 15, 2019.
9. “High-Performance Distributed Deep Learning: A Beginner’s Guide”, NCAR SEA '19, Apr 12, 2019.
10. “How to Boost the Performance of HPC/AI Applications Using MVAPICH2 Library” NVIDIA GTC '19, Mar 20, 2019.
11. “High-Performance Distributed Deep Learning: A Beginner’s Guide”, NVIDIA GTC '19, Mar 18, 2019.
12. “High-Performance Distributed Deep Learning: A Beginner’s Guide”, PPoPP '19, Feb 17, 2019.
13. “High-Performance Distributed Deep Learning: A Beginner’s Guide”, DOD-PETTT '18, May 15, 2018.
14. “High-Performance Distributed Deep Learning: A Beginner’s Guide”, NCAR SEA '18, Apr 5, 2018.
15. "High-Performance Distributed Deep Learning: A Beginner’s Guide”, PPoPP '18, Feb 25, 2018.
16. “High-Performance Distributed Deep Learning for Dummies”, IT4 Innovations (Austria), Jan 24, 2018.
17. “High Performance Distributed Deep Learning for Dummies”, Hot Interconnects (HotI) '17, Aug 28, 2017.
18. “S-Caffe: Co-designing MPI Runtimes and Caffe for Scalable Deep Learning on Modern GPU Clusters”, PPoPP ’17, Austin TX.
19. “Efficient Large Message Broadcast using NCCL and CUDA-Aware MPI for Deep Learning", Best Paper Runner-up Session, EuroMPI ’16 @ EPCC Edinburgh UK.
20. “Why Execution is more important than Ideas”*, Invited Talk at CECOS University, Peshawar, Pakistan.

Professional Service

Memberships

1. ACM Student Member
2. IEEE Student Member

Reviewer

1. PyOhio ’19.
2. 32nd ACM International Conference on Supercomputing (ICS ‘18).
3. Intl. Conference on High Performance Computing, Networking, Storage, and Analysis (SC ’17).
4. 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID ‘17).
5. 26th International Conference on Parallel Architectures and Compilation Techniques (PACT ‘17).
6. 31st IEEE International Parallel & Distributed Processing Symposium (IPDPS ‘17).
7. ISC High Performance 2016 (ISC ’16).
8. 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.

Forums Attended

1. Message Passing Interface (MPI) Forum (2015)

Technical Skills

- Strong programming skills in C and Java (SE)/Java for HPC.
- Development experience in C and interaction of C, C, and MPI (Caffe, OSU-Caffe, and OC-Caffe).
- Basic Python programming
- 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.