Jahanzeb Maqbool Hashmi

https://jahanzeb-hashmi.github.io

Objective

I am interested to work with collaborative research and development teams and design large-scale parallel and distributed systems to scale state-of-the-art computational and deep learning workloads on HPC and Cloud systems.

Education

Ph.D. in Computer Science and Engineering

The Ohio State University, Columbus, Ohio, USA 2015 – 2020
Thesis: Designing High Performance Shared-Address-Space and Adaptive Communication Middlewares for Next-Generation HPC Systems
Advisor: Dhabaleswar K. (DK) Panda

M.S. in Computer Engineering

Ajou University, Suwon, South Korea 2012 – 2014
Thesis: Exploring Performance and Energy Efficiency of ARM Multicore Cluster for High Performance Scientific Computing
Advisor: Sangyoon Oh

B.S. Information Technology

National University of Science and Technology, Islamabad, Pakistan 2007 – 2011
Thesis: Implementation and Evaluation of Scientific Simulations on HPC Architectures
Advisor: Aamir Shafi

Professional Experience

- **Senior High Performance Compute Architect,** NVIDIA Corporation, USA. March 2021 – Present
  - Working in the GPU Architecture End-to-End HPC team to guide the direction of HPC hardware and software by working with architecture, software, and product teams.
  - Identifying prioritized list of HPC application, capturing configuration specific traces, maintaining workloads and benchmark suites, and analyzing top hardware/software performance limiters.
  - inventing new compute and programming models to mitigate performance limiters
  - Generating projections and analysis reports for internal/external consumption and assisting leadership in decision making.
- **Senior Research Associate,** The Ohio State University, USA. June 2020 – March 2021
  - Worked on the design and development of high-performance MPI library for next-generation HPC and Cloud systems with multi-core CPUs (AMD Rome, Intel Xeon, IBM POWER9, ARM A64FX) and many-core GPUs (NVIDIA, AMD)
  - Led the design and development of a generalized hierarchical MPI collective communications framework for modern CPU and GPU systems
  - Led the design and development of MVAPICH2-GDR, a high-performance GPU-aware MPI library, for NVIDIA and AMD based multi-GPU systems
  - Mentoring and guiding Ph.D. and Masters students on various areas of research and development. This includes novel algorithms and designs for MPI communication protocols, software best practices, debugging and performance characterizations of distributed software stacks
  - Participating in NSF and industry grant proposal writing and acquisition
- **Graduate Research Associate,** The Ohio State University, USA. August 2015 – May 2020
  - Designed an adaptive and topology aware algorithm for mapping of MPI processes to hardware cores by capturing the communication-patterns of AI and HPC applications
  - Worked collaboratively on efficient parallelization of large-scale distributed DNN training (data and model parallel) on CPU and GPU systems
  - Designed and developed a truly zero-copy XPMEM-based inter-process (IPC) communication with shared address space principle targeting manycore architectures
  - Designed a novel algorithm to cache data layouts to help mitigate the performance costs of layout translation of MPI derived datatypes. This work was nominated for the Best Paper Award at IPDPS'19

- Worked on PGAS libraries e.g., OpenSHMEM, UPC++ and task-based programming models e.g., Kokkos with MPI backend

Technical Skills

- Parallel Programming Systems – MPI, OpenMP, CUDA, HIP, OpenSHMEM, UPC++
- Deep Learning – Tensorflow, PyTorch, Horovod
- Programming Languages – C, C++, Java
- Scripting Languages – Bash, Python
- Tools – Git, LaTeX, GDB, PerfAPI (PAPI), mpiP, Valgrind
- Familiarity with developing Linux Kernel modules
- Strong programming, debugging, and problem solving skills.
- Experienced with large-scale software design, development, and release life-cycle.
- Experienced with performance benchmarking of parallel hardware with focus on scientific and AI applications
- Strong communication and presentation skills.

Select Publications

For complete list of publications, please refer to my Google Scholar.

1. **J. Hashmi**, C. Chu, S. Chakraborty, M. Bayatpour, H. Subramoni, and DK Panda. "FALCON-X: Zero-copy MPI Derived Datatype Processing on Modern CPU and GPU Architectures", Journal of Parallel and Distributed Computing (JPDC), Volume 144, October 2020, Pages 1-13, doi.org/10.1016/j.jpdc.2020.05.008

2. **J. Hashmi**, S. Xu, B. Ramesh, M. Bayatpour, H. Subramoni, and D. K. Panda. "Machine-agnostic and Communication-aware Designs for MPI on Emerging Architectures", in proceeding of *34th IEEE International Parallel and Distributed Processing Symposium* (IPDPS '20) , May 2020

3. **J. Hashmi**, S. Chakraborty, M. Bayatpour, H. Subramoni, D. K. Panda. "FALCON: Efficient Designs for Zero-copy MPI Datatype Processing on Emerging Architectures", in proceeding of *33rd IEEE International Parallel and Distributed Processing Symposium* (IPDPS '19) , May 2019 [Best Paper Finalist]

4. **J. Hashmi**, S. Chakraborty, M. Bayatpour, H. Subramoni, D. K. Panda. "Designing Efficient Shared Address Space Reduction Collectives for Multi-/Many-cores", in proceeding of *32nd IEEE International Parallel and Distributed Processing Symposium* (IPDPS '18) , May 2018

5. **J. Hashmi**, S. Chakraborty, M. Bayatpour, H. Subramoni, D. K. Panda. "Design and Characterization of Shared Address Space MPI Collectives on Modern Architectures", in proceeding of *The 19th Annual IEEE/ACM International Symposium in Cluster, Cloud, and Grid Computing* (CCGRID '19) , May 2019

6. S. Chakraborty, M. Bayatpour, **J. Hashmi**, H. Subramoni, D. K. Panda. "Cooperative Rendezvous Protocols for Improved Performance and Overlap", in proceeding of *IEEE/ACM International Conference for High Performance Computing, Networking, Storage, and Analysis* (SC '18) , Nov 2018 [Best Paper Finalist]

7. M. Bayatpour, **J. Hashmi**, S. Chakraborty, H. Subramoni, P. Kousha, D. K. Panda. "SALaR: Scalable and Adaptive Designs for Large Message Reduction Collectives", in proceeding of *IEEE International Conference on Cluster Computing* (CLUSTER 2018) , Sep 2018 [Best Paper Award]

8. 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", in proceeding of *22nd ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming* (PPoPP '17) , February 2017

9. C. Chu, **J. Hashmi**, K. S. Khorassani, H. Subramoni, and D. K. Panda. "High-Performance Adaptive MPI Derived Datatype Communication for Modern Multi-GPU Systems", in proceeding of *26th IEEE International Conference on High Performance Computing, Data, Analytics and Data Science* (HiPC '19) , Dec. 2019