

---

---

# Benchmarking in Datacenter: Expanding to the Cloud

---

---

PPoPP 21 (Virtual)  
Feb 28th, 2021

Ammar Ahmad Awan (Microsoft)  
Shahzeb Siddiqui (Lawrence Berkeley National Laboratory)

# Logistics

- Moderator and Panel Introduction
- Q/A with Panelist
- Audience can post questions in Zoom chat



## Moderators

### Ammar Ahmad Awan (Microsoft)

Bio: Ammar Ahmad Awan is a Researcher at Microsoft. He received his PhD in Computer Science in May, 2020 from The Ohio State University. He received his B.S. and M.S. degrees in Computer Science and Engineering from National University of Science and Technology (NUST), Pakistan and Kyung Hee University (KHU), South Korea, respectively. His current research focus lies at the intersection of High Performance Computing (HPC) libraries and Deep Learning (DL) frameworks. He previously worked on a Java-based Message Passing Interface (MPI) and nested parallelism with OpenMP and MPI for scientific applications. He has published 20 papers in conferences and journals related to these research areas. He actively contributes to the DeepSpeed project at Microsoft. Before that, he has contributed to various projects like MVAPICH2-GDR (High Performance MPI for GPU clusters, OMB (OSU Micro Benchmarks), and HiDL (High Performance Deep Learning). He is the lead author of the OSU-Caffe framework (part of HiDL project) that allows efficient distributed training of Deep Neural Networks.



### Shahzeb Siddiqui (LBNL)

Bio: Shahzeb Siddiqui is a HPC Consultant/Software Integration Specialist at Lawrence Berkeley National Laboratory/NERSC. He spends 50% of his time on Consulting where he helps address any incoming issues reported by NERSC users. The rest of his time is dedicated to the **Exascale Computing Project** (ECP). Shahzeb is part of the **Software Deployment** (SD) group where he is responsible for installing Spack E4S stack for DOE machines. Shahzeb is involved in several open-source projects including spack, singularity, easybuild. He is the creator of [buildtest](#) and [lmodule](#) used for testing HPC systems and module stack. His experience is in DevOps, installing scientific software, cluster administration, containers, configuration management. Shahzeb has a Masters in Computer Science from KAUST and Bachelors in Computer Engineer from Penn State University. Shahzeb has a certificate in Red Hat Certified System Administrator (RHCSA)

## Panelists



### **Verónica G. Melesse Vergara (ORNL)**

Bio: Verónica G. Melesse Vergara (Vergara Larrea) is originally from Quito, Ecuador. Verónica earned a B.A. in Mathematics/Physics at Reed College and a M.S. in Computational Science at Florida State University. Verónica has over a decade of experience in the high performance computing field and is currently Group Leader of the User Assistance — Pre-production Systems Group at the Oak Ridge Leadership Computing Facility. In addition to providing assistance to OLCF users, Verónica is part of the systems testing team, led acceptance for Summit, and is leading acceptance for Frontier, ORNL's exascale supercomputer. Her research interests include high performance computing, large-scale system testing, and performance evaluation and optimization of scientific applications. Verónica is a member of both IEEE and ACM and serves in the ACM SIGHPC Executive Committee.



### **Jithin Jose (Microsoft)**

Bio: Dr. Jose's work is focused on co-design of software and hardware building blocks for high performance computing platforms, and designing communication runtimes that seamlessly expose hardware capabilities to programming models and middleware. His research interests include high performance interconnects and protocols, parallel programming models, and cloud computing. Before joining Microsoft, he worked at Intel and IBM Research. He has published more than 25 papers in major conferences and journals. Dr. Jose received his Ph.D. degree from The Ohio State University.



### **Sreeram Potluri (NVIDIA)**

Bio: Sreeram Potluri is a system software manager at NVIDIA. He leads the GPU Communications Group, which provides network and runtime solutions that enable high-performance and scalable communication on clusters with NVIDIA GPUs. Sreeram received a Ph.D. in computer science from Ohio State University. His research interests include high-performance interconnects, heterogeneous architectures, parallel programming models and high-end computing applications. He has published over 30 papers in major peer-reviewed journals and international conferences related to these research areas.

**What are some of the major benchmarks that we use in data centers commercial cloud environments like Azure, AWS, etc. for various workloads including machine/deep learning?**

**Do we need to develop any new benchmarks for existing or new workloads?**

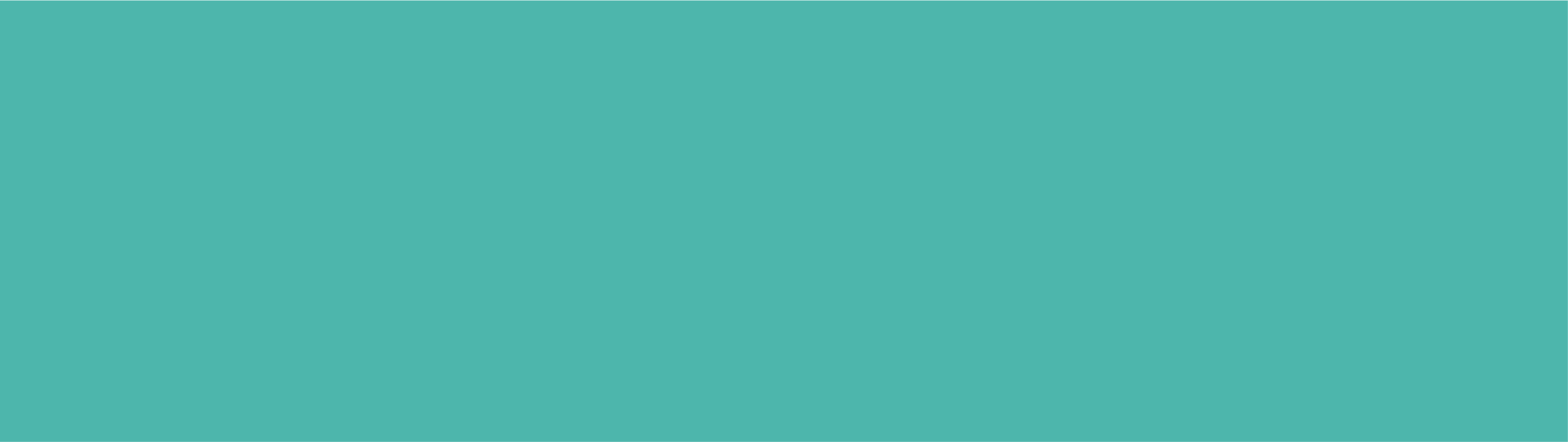
**How does the data center differ from the cloud in terms of benchmarking tools and workloads?**



**If you had to pick 3 benchmarks to run for your next system acceptance test, what will they be and why would you pick them? Will you submit your results?**



**Please describe your process for running benchmarks at your data center? How do you evaluate benchmark results between runs? How many runs? Do you keep track of historical results when running benchmarks?**



# Q/A with Audience