Transparency and Reproducibility Practice in Large-Scale Computational Science: A Preface to the Special Section

Beth Plale, *Member, IEEE* and Stephen Lien Harrell, *Member, IEEE*

Abstract—With this special section we bring you a practice and experience effort in transparency and reproducibility for large-scale computational science. A unique section, it consists of a research work plus six critques, each by a student team that reproduced the work. The original research work has been expanded in its science and also in its contribution to open science with a discussion of the student effort. Our letter contemplates implications as well.

Index Terms—Open science, computational science, reproducibility, practice and experience

1 Introduction

Welcome to our special section on reproducibility in large-scale computational science. Science that is open and transparent is critical to maintaining the overall rigor and respect of the scientific enterprise. The sharing of research results beyond just the peer reviewed publication can, with proper accommodations for proprietary or protected information, facilitate reproducibility and accelerate science. The latter has been amply demonstrated by the rapid response by funders, publishers, industry members, and professional societies to making COVID-19 resources openly available . Transparency in computer and computational sciences has numerous benefits: it increases the probability that a research result can be successfully built upon – scientific results often require replication as a starting point for building on prior research. Transparency contributes positively to the incentives for research as it allows research results beyond the publication to be assigned a persistent ID (PID), so can be cited and counted. Finally, an individual researcher's commitment to transparency is a commitment to the greater good.

In this special section we are excited to bring the reader a practice and experience effort in transparency and reproducibility. This special section brings together in one place, for the first time, a multi-faceted activity in reproducibility held annually at the International Conference for High Performance Computing, Networking, Storage, and Analysis (SCxy) conference. This special section features a novel computational science paper together with six individual student team efforts to reproduce the science paper. While a controlled experiment in reproducibility, the collection taken together gives a meaningful and practical example of what reproducibility can mean for high performance computing.

2 Framework

The computer and computational sciences publish a large proportion of their research in selective peer reviewed conferences. When a conference has strong organizational continuity from year to year, as does SCxy, communities will organize themselves around these major venues thus positioning the conference to take a leading role in advancing a collective objective like open science and reproducibility. This has been the case for SCxy, which began taking steps towards improved reproducibility as early as 2015.

The SCx annual conference and IEEE TPDS have been leaders in open science for high performance computing. SCxy has approached this through a multi-pronged approach of improving the rigor of the science that it publishes while at the same time enhancing the open science skillset of the next generation of researchers.

The first SCxy reproducibility prong achieves greater transparency of research artifacts for the papers that SCxy accepts for publication. Through means of a peer reviewed appendix, authors provide information about the tools, data, libraries, and other software and hardware that is both used in, and produced by, the research. Artifacts information complements the experimental environment description that is routine in experimental evaluation papers. Review of the artifact description appendices is aligned with ACM's policy on Artifact Review and Badging where artifacts can be deemed to have at a minimum been evaluated according to well defined criteria of being complete, exercisable, consistent, and documented.

The second SCxy reproducibility prong complements artifacts transparency. SCxy holds an annual Student Cluster Challenge (SCC), first developed in 2007, to provide an immersive high performance computing experience to undergraduates. Student teams choose to participate in the

- *Beth Plale is with the Intelligent Systems Engineering Department, Indiana University Bloomington, Bloomington, IN 47408 USA. E-mail: plale@indiana.edu.*
- *Stephen Lien Harrell is with the Texas Advanced Computing Center, University of Texas at Austin, Austin, TX 78758 USA. E-mail: sharrell@tacc.utexas.edu.*

Date of current version 13 May 2021.
Digital Object Identifier no. 10.1109/TPDS.2021.3058393

reproducibility challenge wherein they reproduce results from an accepted paper from the prior year's Technical Program.

This special section brings together in one place an extended version of the SC18 research paper by Shia *et al.* [1] that served as the basis for SC'19 student teams' reproducibility efforts, and six of the best performing student team efforts. The student papers are published here in the form of critiques of the paper. Shi *et al.*, in this special section, extends prior work on a first-order mixed finite element method (FEM) by presenting the computational and numerical performance of a second-order mixed element FEM. Jia *et al.* additionally reflect on use of their first-order solver in the SC19 Student Cluster Challenge (SCC) as a case study in computational reproducibility.

Six student teams, out of sixteen, completed the SC19 Student Cluster Challenge (SCC) with quality scores that merited an invitation to this special section. An shepherding process strengthen student reports. Student teams use a completely different dataset from Shia *et al.* [1] (a Mars-model dataset) and carried out their experimentation on different clusters although with similar architecture. It is important to note that students are restricted to three thousand watts of power and 48 hours of non-stop competition for the completion of all four competition applications, including the reproducibility challenge.

Shi *et al.* aided the the student team efforts by providing MATLAB scripts for visualization and for generating various planetary models with multi-resolution. They further shared their code for computing the normal modes and shared the corresponding eigensolver, pEVSL.

The SCC committee designed the tasks that student teams undertake, drawing from Shi *et al.* results. Students were asked to i) design and perform weak scalability of sparse matrix-vector multiplication (SpMV), ii) design and perform scaling of the model size task, and iii) design and perform a strong scalability task.

The six teams represented in this special section all brought their own cluster for the SCC challenge (which took place on the floor of the convention center). Each cluster had between 2-5 nodes with Intel Xeon CPUs and NVIDIA V100 GPUs, although no GPUs were used for the reproducibility challenge. The 6 top teams are: • National Tsing Hua University team (Sun *et al.*) • ETH Zurich team (Burger *et al.*) • Tsinghua University team (Zhang *et al.*) • University of Warsaw team (Masiak *et al.*) • University of Washington team (Liu *et al.*) • Peking University team (Cheng *et al.*)

Jia *et al.* Section 4 summarizes the results of each of the student team efforts, and points out the unique approaches taken by the different teams. The student teams discuss distinct and creative approaches to achieving scalability in the resource-limited environments within which they were constrained. Their effort, and the reflections of Jia *et al.* on the student papers, is captured in this special section.

3 Observations

We encourage the reader to explore all the papers in this special section. We offer for consideration a couple of observations.

The first observation is of who is it that defines when a reproducibility effort meets the bar of obtaining the main results of the paper? ACM definition of results reproduced is that "the main results of the paper have been obtained in a subsequent study by a person or team other than the authors, using, in part, artifacts provided by the author" . All six teams met this bar of "results reproduced." But how? While there is independence between the authors of the primary paper and the teams attempting to reproduce the results, the very structure of the SCC brings a third party into the reproducibility effort who played a crucial role.

Those artifacts required by the student teams, and the quality of the artifacts, is determined by a third party (SCC organizers). This third party has a vested interest in student teams having a successful experience. The SCC organizers additionally define for the teams the criteria by which reproducibility is assessed. That is, in identifying three discrete tasks, the SCC organizers are themselves defining when a team meets of the bar of results reproducibility. Does this presence of a third party corrupt the outcome? We think not. The SCC organizers work carefully to fully grasp the goals of the original paper; this interpretation step is needed for undergraduate and high school student teams for whom many of the concepts are new. But in other settings it could be the authors themselves who offer more guidance on when reproduction of the main results of a paper have been met.

The second observation is that cluster size matters. For both weak and strong scalability tests, Jia *et al.* used up to 256 Intel Skylake nodes with an Intel Omni-Path interconnect. In the reproducibility challenge, the participant teams have 2 to 5 nodes with InfiniBand as their interconnect. Most of their CPUs are Intel Skylake or Cascade. The difference in cluster size is compensated for, in part, by the nature of scalability measures. The slope of a scalability line can be determined with fewer data points and then scaled. Jia *et al.* concludes that both weak and strong scaling performances of the reproducibility by the participant teams were impressive and encouraging.

In artificial intelligence research, however, scalability is not the entire picture and the significant difference in cluster size matters far more in results accuracy especially for deep learning which requires significant resources. Future reproducibility efforts of SCxy and beyond should grapple with the question of how to assess when the "main results of the paper have been obtained" in cases of large disparity in computational resources between the original group and reproducing team.

We thank Irene Qualters (LANL), Michela Taufer (U Tenessee), John Linford (ARM), Scott Michael (Indiana University) all of the SC20 organizing team. We thank Manish Parashar, IEEE TPDS editor-in-chief, for his support and creativity.

Beth Plale
Stephen Lien Harrell
Guest Editors

References

[1] J. Shi, R. Li, Y. Xi, Y. Saad, and M. V. de Hoop, "Computing planetary interior normal modes with a highly parallel polynomial filtering eigensolver," in *Proc. Int. Conf. High Perform. Comput. Netw. Storage Anal.*, 2018, pp. 894–906.