Advancing Adoption of Reproducibility in HPC: A Preface to the Special Section

Stephen Lien Harrell, *Member, IEEE*, Scott Michael, *Member, IEEE*, and Carlos Maltzahn, *Member, IEEE*

Abstract—In this special section we bring you a practice and experience effort in reproducibility for large-scale computational science at SC20. This section includes nine critiques, each by a student team that reproduced results from a paper published at SC19, during the following year's Student Cluster Competition. The paper is also included in this section and has been expanded upon, now including an analysis of the outcomes of the students' reproducibility experiments. Lastly, this special section encapsulates a variety of advances in reproducibility in the SC conference series technical program.

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

1 Introduction

Welcome to our special section on reproducibility in large-scale computational science. The idea of research being reproducible underpins the entire notion of the scientific method and is the bedrock of truth in many domains. Reproducibility in HPC has made great strides over the past several years with a host of new technologies being produced or coming to greater prominence to make reproducible computational research more easily attainable.

Although research in HPC does have many of the challenges faced in other domains such as proprietary data, code, or hardware, limited access to emerging technologies, and sometimes difficult to use tooling for reproducibility; one of the unique aspects of reproducibility in HPC is the scale of the research being conducted. Often research is presented where fundamental findings have been uncovered with hundreds of thousands of computational cores or hundreds of terabytes of memory.

Despite these challenges, there are many benefits to reproducible research. Perhaps the greatest benefit being the ability to build upon the research of others and having one's own research extended in new and exciting ways. Reproducible research allows HPC researchers to collaborate more effectively and build a community approach to scientific advancement. This special section presents an update to an paper from SC19 with nine critiques created by undergraduate student teams from the Student Cluster Competition (SCC) at SC20. The SC20 competition gives insight on how larger scale experiments (both in core count and memory footprint) can be scaled to the student cluster scale, and some potential issues in such an endeavor. We encourage the reader to explore all the papers in this special section.

- *Stephen Lien Harrell is with Texas Advanced Computing Center, University of Texas at Austin, Austin, TX 78758 USA. E-mail: sharrell@tacc.utexas.edu.*
- *Scott Michael is with Research Software, Indiana University, Bloomington, IN 47408 USA. E-mail: scamicha@iu.edu.*
- *Carlos Maltzahn is with the Center for Research in Open Source Software, University of California Santa Cruz, Santa Cruz, CA 95064 USA. E-mail: carlosm@ucsc.edu.*

Digital Object Identifier no. 10.1109/TPDS.2021.3128796

2 Progress in Reproducibility

Over the past several years, the International Conference for High Performance Computing, Networking, Storage, and Analysis (SC) conference has made progress in enhancing the reproducibility of the technical content presented at the conference. The process of students replicating research from a paper of the prior year's SC has been in place since 2015. In addition to working with a team of authors to design an appropriate challenge for the SCC, the SC Reproducibility Initiative guides the process for the collection of reproducibility artifacts for all technical paper submissions and works with the most highly rated of the student teams from the SCC to present their competition reports as the reproducibility critiques collected in this, and previous years', special sections.

Each year there are new aspects to reproducibility that the Initiative focuses on and different areas that are emphasized. For SC21, the technical paper artifact collection, evaluation, and badging process saw some major improvements (see Section 2.1). This special issue has some new and novel aspects as well. For the first time in the history of the Student Cluster Competition, the SC20 SCC, from which these critiques come from, was a completely virtual event. Instead of a standard power limit of 3,000 watts, as had been used in previous competitions, student teams were limited to a standard budget of Microsoft Azure credits that they could spend on their clusters over the course of the competition. While this may have somewhat limited the heterogeneity of hardware seen at the SCC, it provided many new opportunities for students to experiment with scale and virtualization. Another change to the process of publishing this special section is that student teams have included artifacts that encapsulate the experiment that they ran in the 48 hour SCC.

2.1 Strengthening Reproducibility at SC

The SC21 Reproducibility Initiative has moved SC substantially forward by bringing it in line with best practices used by other conferences: (1) authors applied for up to three reproducibility badges (available, functional, reproducible); (2) the application process cleanly separated the responsibilities of the paper reviewers of technical program committees and the Initiative's reviewers of badge applications (of accepted papers only); (3) the Initiative collaborated with NSF Cloud infrastructure providers for preparing and examining artifacts; (4) accepted papers were listed with their awarded badges in the SC21 Schedule prior to paper presentations; and (5) the SC Steering Committee approved the SC Best Reproducibility Advancement Award, and members of the technical program and the reproducibility initiative committees selected a winner of the Award.

2.2 Reproducibility Through Undergraduate Engagement

During SC20 the students participated in a directed computational reproducibility challenge. They were asked to reproduce specific sections from a paper accepted at SC19 [1]. Here we share the results of this experiment from the students as well as an additional analysis from the original paper authors about the students' results.

Of the 19 teams that participated in the SC20 SCC, nine teams were invited to improve their initial competition reports to the critiques that are included in this special section. Hidayetoğlu *et al.* have extended their work in memory bound X-ray tomographic reconstruction to address the findings of the student teams. The nine critiques in this special section were chosen based on the strength of their scores on the Reproducibility Challenge and in the competition overall.

The competition data sets were constructed by the Hidayetoğlu *et al.* author team in conjunction with the SC Reproducibility Challenge committee to allow the challenge experiment to be conducted on the student team virtual clusters and still be representative of the experiments from the original paper, which were performed on much larger computational resources like Stampede and Theta. Student teams were asked to reproduce the results seen in the paper on a single CPU and single GPU, as well as demonstrate the strong scaling relationship seen by Hidayetoğlu *et al.* The teams were provided with test data sets that were scaled down versions of the paper data sets to experiment with before the competition.

For the SCC each of the nine teams in this special section had access to a wide variety of virtual machine (VM) types in the Microsoft Azure cloud. The extended version of the SC19 paper presented in this special section includes a comparison of the VM instances used by the teams and presents some insights into the different approaches and results. Of note is the fact that all teams that were able to measure GPU performance found strikingly similar results to those presented in the original paper (see Section 6.4 and Fig. 13 of the extended paper). However, the reproducibilty effort did not come without challenges. Student teams were challenged to reproduce results seen on Intel's Knight's Landing architecture used by Stampede. This is illustrative of a more general challenge when trying to reproduce results on hardware that the reproduction effort does not have access to. Another challenge, that of scale, is a perennial problem seen in the SCC Reproduciblity Challenge and reproducibility efforts more broadly. The scaling studies of the original paper ranged in size from a single node to 4096 nodes, with the strong scaling studies starting at 32 or 128 nodes. Student teams were able to scale to *at most* 4 nodes and stay within time and budget of the competition. Although the same general trends were observed, some teams were challenged to explain how the details of the behavior seen at these small scales translated to the larger scaled results.

The reproducibility challenge is a unique way to engage undergraduates in current research, but is also a sandbox for experimenting with techniques and guidelines that can inform the computational reproducibility verification processes for SC and beyond. Each year the reproducibility challenge tries new techniques to collect ideas of what could work and what might not in the operational aspects of computational reproducibility.

This year we asked the students to create digital artifacts from the runs of their experiments. The students recorded the run, compile, result visualization and scaling plot scripts from their experiment. Additionally, they include environmental variables and documentation about how the experiment was performed. The critiques in this special section each correspond to an artifact which further explains the outcomes from the reproducibility experiment as well as the student team's assessment of whether the code is reproducible. The artifacts have been made publicly available via Zenodo and are linked on each critique.

3 Future Work for Reproducible Research

As one of the leading conferences in high performance computing SC has had an impact in many different areas over the years. With its intentional emphasis on reproducible research starting in 2015 with the SC Reproducibility Initiative, we have seen many advances in reproducible research come from the conference series. Over the past six years the idea of providing a basic set of information, via artifact descriptors for technical papers, has become normalized and is now a standard part of the paper review process. A future step is to find a way to gauge the benefit of these activities. As we stated in this introduction, one of the greatest benefits of reproducible research is the ability to extend and build upon already existing discoveries and methodology. While we are certain that the efforts by the SC reproducibility community have helped to make this easier, the ability to quantify the benefit that researchers have experienced due to the reproducibility initiative remains elusive.

In future years the SC Reproducibility Initiative may find ways to draw a more direct line from their efforts to continuation and expansion of research presented at the conference. The initiative allowing paper submissions to have their claims of reproducibility tested prior to publication will continue and may be expanded in the future. Lastly, the Student Cluster Competition remains a valuable testbed for new ideas in computational reproducibility and a training ground for reproducible research with the HPC professionals of tomorrow.

We thank the entire SC21 reprodcubility team, with special thanks to Le Mai Weakley, Michela Taufer, and Abhinav Thota. We also thank Manish Parashar, IEEE TPDS editor-in-chief and Pavan Balaji, IEEE TPDS associate editor-in-chief of special sections for their expert support and advice.

REFERENCE

[1] M. Hidayetoğlu *et al.*, "MemXCT: Memory-centric X-ray CT reconstruction with massive parallelization," in *Proc. Int. Conf. High Perform. Comput. Netw. StorageAnal.*, 2019, Art. no. 85.