Expanding the Scope of Artifact Evaluation at HPC Conferences: Experience of SC21

Tanu Malik
DePaul University
Chicago, Illinois, USA
tanu.malik@depaul.edu

Anjo Vahldiek-Oberwagner
Intel Labs
Portland, USA
anjo.lucas.vahldiek-oberwagner@intel.com

Ivo Jimenez
Redpanda Data Inc
San Francisco, California, USA
ivo@redpanda.com

Carlos Maltzahn
UC Santa Cruz
Santa Cruz, California, USA
carlosm@ucsc.edu

ABSTRACT

A scientific paper consists of a constellation of artifacts that extend beyond the document itself: software, hardware, evaluation data and documentation, raw survey results, mechanized proofs, models, test suites, benchmarks, and so on. In some cases, the quality of these artifacts is as important as that of the document itself. Based on the success of the Artifact Evaluation efforts at other systems conferences, the 2021 International Conference for High Performance Computing, Networking, Storage, and Analysis (SC21) organized a comprehensive Artifact Description/Artifact Evaluation (AD/AE) review and competition as part of the SC21 Reproducibility Initiative. This paper summarizes the key findings of the AD/AE effort.

CCS CONCEPTS

• **General and Reference → Cross-computing tools and techniques**.

KEYWORDS

Research objects; computational reproducibility, replicability; cloud infrastructure; supercomputing;

ACM Reference Format:
Tanu Malik, Anjo Vahldiek-Oberwagner, Ivo Jimenez, and Carlos Maltzahn. 2022. Expanding the Scope of Artifact Evaluation at HPC Conferences: Experience of SC21. In *Proceedings of the 5th Int'l Workshop on Practical Reproducible Evaluation of Computer Systems (P-RECS '22), June 30, 2022, Minneapolis, MN, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3526062.3536354

1 INTRODUCTION

The objective of the reproducibility initiative at SC is to advance scientific rigor. Rigor is defined as "the strict application of the scientific method to ensure robust and unbiased experimental design" [5]. SC achieves this objective by accepting high-quality, peer-reviewed technical contributions from authors but, more recently, also allowing technical contributions to engage in enhanced reproducibility. For last several years, the enhanced reproducibility comprised of authors submitting *appendices* with their technical contribution. The appendices described the contents of the artifact in terms of hardware/ software requirements, and textual description of how the artifact can generate results and establish claims mentioned in the paper.

The 33^{rd} edition of SC, *i.e.*, SC21 continued the practice of achieving enhanced reproducibility via author-contributed appendices but further developed on the structure of appendices by conducting a comprehensive peer-review *evaluation* of appendices. This added evaluation step comprised of evaluating the availability, functionality, and reproducibility of the artifact. The objective was to assess how accessible is the artifact, how functional it is in terms of reuse, and how reproducible is the artifact to reproduce the paper's key results and claims, and make the assessments available for the community. This added artifact evaluation step, beyond the artifact description via the appendices, was considered necessary and timely for a variety of reasons:

- Publishers, such as Association of Computing Machinery (ACM), Institute of Electrical and Electronics Engineers (IEEE) and the National Institute of Standards and Technology (NIST) developed a common badging standard for computational artifacts [1]. These badges have the potential to guide users of the artifact about the current state of the artifact, and thus set the expectation level accordingly.
- Roughly 19 other systems conferences, such as EuroSys, OSDI, SIGMOD and ASPLOS conduct a thorough evaluation of the computational artifacts [4].
- Reproducibility practices within scientific communities are evolving rapidly. A more concerted evaluation practice, if rolled out, has the potential to provide evidence for the readiness of the community in terms of adoption and also inform about bottlenecks.

P-RECS '22, June 30, 2022, Minneapolis, MN, USA.
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9313-3/22/06…$15.00
https://doi.org/10.1145/3526062.3536354

To extend these benefits, the SC21 Reproducibility Initiative [3] brought appendix description and evaluation under the sole oversight of the Artifact Description/Artifact Evaluation (AD/AE) Committee. This paper describes how the committee organized, prepared, and conducted the AD/AE process over the span of one year. The evolution of the process was widely recognized by the SC21 Steering Committee insomuch that it also lead to the introduction of the SC Best Reproducibility Advancement Award, a new SC award to recognize outstanding efforts in improving transparency and reproducibility of methods for high performance computing, storage, networking and analysis.

In this paper we provide a finer-grained data analysis of the work of this committee, comparing reviewer and author effort, and comparing author engagement in artifact evaluation within different SC sub-areas. We hope that future SC21 committees can use this analysis for efforts, and use the processes adopted here to provide guidance to both authors and reviewers.

The rest of the paper is organized as follows: Section 2 describes a brief history of the Reproducibility Initiative at SC. Section 3 describes how we organized and prepared the AD/AE committee for artifact evaluation. Section 4 describes the results of the AD/AE effort, and finally in Section 5 we highlight some of the achievements of this committee, describing its scale, complexity, and the dedicated effort of several members that improved the practice and brought it to successful fruition. Finally, we conclude in Section 6.

2 HISTORY OF THE AD/AE AT SC

The SC21 AD/AE built upon several past AD/AE processes conducted at several systems and programming language conferences [4],[2]. While each AD/AE process is unique, the part common to each effort are the reproducibility badges [1]. Each community chooses badges relevant to their process, and then refines the process. In this section we describe the history of the Reproducibility Initiative at SC and the community perception.

The Supercomputing (SC) series of conferences has taken the lead in community efforts in reproducibility through its SC Reproducibility Initiative of which each of us authors has had a leadership role at one time or another. The International Conference for High Performance Computing, Networking, Storage, and Analysis (SC) attracts over 10,000 participants annually to an event that features breaking news, a expansive exhibit floor, and a technical program for high-quality original research, groundbreaking ideas, and compelling insights on future trends in high performance computing, networking, storage, and analysis. The technical program receives over 300 submissions annually, and after an extensive peer review process, selects about 20% for presentation and publication in the SC proceedings which are archived in the ACM Digital Library and IEEE Xplore.

The SC conference began its Reproducibility Initiative in 2015 primarily as an optional practice for authors of accepted papers to describe their experimental framework and results in more detail. In 2019, the Artifact Description (AD) appendix became mandatory. The initial objective of the AD appendix was to provide transparency and sufficient detail to support an independent audit. In 2015, authors of only one paper responded to the initiative, and that paper became the source for the SC16 Student Cluster Competition Reproducibility Challenge; it is also the first SC paper to display an ACM badge. By 2017, 39 papers had an AD. In 2019, the AD became mandatory [6]. While established papers became part of student cluster competitions [7], however, a comprehensive evaluation of appendices, *i.e.,* extensive Artifact Evaluation (AE) of the appendices was never conducted.

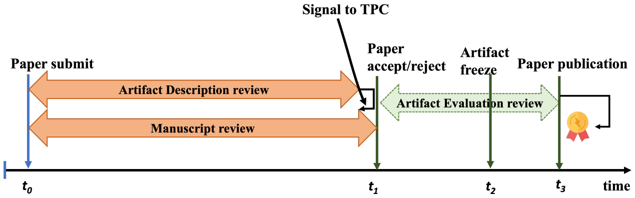

Figure 1: AD and AE are independent evaluation phases. AD ends a bit early to provide a signal to TPC about the quality of the artifact. The AE phase culminates with a badge.

A survey [8] collected information about the SC Reproducibility Initiative practices from the attendees of the SC conference in years 2017, 2018, and 2019. The results of this survey specifically showed that the reproducibility initiative practices have contributed to higher levels of awareness on the part of SC conference technical program participants, and hint at contributing to greater scientific impact for the published papers of the SC conference series. The survey argued against any *stringent* point-of-manuscript-submission verification. The survey authors highlight computational reproducibility challenges in HPC and propose for an artifact evaluation that is an indication and not a measure of the state of the research.

The survey also points out that a full 90% of the respondents are aware of issues related to reproducibility in computational and computer sciences, and only 15% think that the concerns about reproducibility in science are exaggerated. But more importantly from an evaluation perspective, the survey indicated that a full thirty five percent (35%) of the respondents were affirmative in their use of appendix information. This was indicative that an formal artifact evaluation phase will specifically strengthen a research work's potential for long-term impact through reuse 5-10 years down the road.

3 AD/AE AT SC21

To establish a formal evaluation process that is single-blind, peer-reviewed, we recruited 48 AD/AE committee members (early career researchers, postdocs, and graduate students) to engage with each phase of the process. The two phases of Artifact Description (AD) and Artifact Evaluation (AE) were clearly distinguished on a timeline (Figure 1). Authors will be required to describe their artifacts as part of the AD phase ($t_0 \rightarrow t_1$), and submission of appendices will be along with technical contribution submission. AE will be pursued after AD ($t_1 \rightarrow t_3$), and be available only for papers accepted by the technical program committee (the greenline).

A primary issue in drafting a timeline was determining how much reproducibility should be a requirement and how much should it act as an incentive to improve the current state. Our philosophy was to take a hybrid approach emphasizing that reproducibility is a carrot instead of a stick. However, we did put in one minimal form of a stick: a bit signal that indicates to the SC21 Technical Program Committee (TPC), whether the paper passes the Artifact Description reproducibility requirement. This rule was chosen so that authors are engaged to minimally think about reproducibility before paper submission and to provide guidance to the technical program committee, especially if reproducibility of results becomes a critical factor in experimental results.

Following the carrot/stick philosophy, another issue was to not to assume artifacts to be ready for evaluation after their corresponding technical contributions are accepted. For this we divided the evaluation phase into finer intervals so that authors have sufficient time to improve their guidance post notification of their accepted paper to SC. More specifically, we created an Artifact Freeze point (t_2), 3 weeks later to the notification of the accepted paper by which authors of accepted paper must have their artifacts ready for evaluation.

We decided the final outcome of the AD/AE process to be the publisher provided badges. To receive a badge, authors were asked to apply for publisher-provided, *i.e.,* ACM provided badges at the time of submission of the Artifact Description appendix. These basdges were themselves chosen per the NISO Reproducibility Badging and Definitions Standard [1]. These badges included:

Open Research Objects (ORO) Badge: Receiving this artifact meant that an author-created artifact of the paper is accessible via a persistent, shareable URI, preferably associated via standard open licenses. To receive this badge, the following was deemed necessary from authors: (i) An pre-assigned DOI from research object repositories such as Zenodo, FigShare, Dryad, Software Heritage to the research object by the Article Freeze deadline, and (ii) Links to code and data repositories on a hosting platform that supports versioning such as GitHub, or GitLab. In other words, the badge prevented use of DropBox links or gzipped files hosted through personal or lab webpages, which had been typical in the submissions of previous SC.

Reviewers took the AD Appendices, which describes the metadata of the research artifact, as a guide to check for the extent of the accessibility criteria of this badge. We defined accessibility as those artifacts used in the research (including data and code) that are permanently archived in a public repository and are assigned a global identifier and guarantees persistence, and are made available via standard open licenses that maximize artifact availability.

Research Objects Reviewed (ROR) Receiving this badge meant that an artifact, during the peer-review process, was exercisable. By exercisable, we implied being able to answer one or more questions such as (i) Is it possible to compile the artifact, use a Makefile, or perform a small run?, (ii) If the artifact runs on a large cluster—can it be compiled on a single machine?, (iii) Can analysis be run on a small scale? and (iv) Does the artifact describe the components to nurture future use of this artifact?

To receive the badge, authors were supposed to provide sufficient details to build the artifact as part of the AD form or as part of an accompanying documentation within a version-controlled repository. We encouraged authors to describe their (i) workflow underlying the paper, (ii) describing some of the black boxes, or a white box (source, configuration files, build environment), (iii) input data: either the process to generate the input data should be made available, or when the data is not generated, the actual data itself or a link to the data should be provided, (iv) environment (system configuration and initialization, scripts, workload, measurement protocol) used to produce the raw experimental data, and (v) the scripts needed to transform the raw data into the graphs included in the paper.

The reviewer was to assess the details of the research artifact based on the following criteria:

Documentation: Are the artifacts sufficiently documented to enable them to be exercised by readers of the paper?

Completeness: Do the submitted artifacts include all of the key components described in the paper?

Exercisability: Do the submitted artifacts include the scripts and data needed to run the experiments described in the paper, and can the software be successfully executed.

Results Reproduced (ROR-R) Receiving the final and the highest badge meant the peer-review successfully reproduced the key computational results using the author-created research objects, methods, code, and conditions of analysis. The objective of this badge is not bit-wise reproducibility, especially since we expected many hardware-based artifacts. The aim was to reproduce behavior. For example, if we get access to the same hardware as used by experiments, we will aim to reproduce the results on that hardware. If not, the objective of this badge was to work with authors to determine the equivalent or approximate behavior on available hardware. If results-to-be-reproduced were latency and performance-oriented, our objective will be to check if a given algorithm is significantly faster than another one, or that a given parameter affects negatively or positively the behavior of a system.

To receive the badge, the peer-review process must be able to reproduce the central results and claims of the paper, i.e., the objective was not to to reproduce all the results and claims of the paper, but to let the peer-review process decide the central results of the accepted paper, and work with authors to confirm it. Once confirmed, the badge will be assigned based on the peer-review process being able to reproduce behavior of these central results.

Authors were also encouraged to apply for the following combinations of badges, namely (i) Open Research Object (ORO), (ii) Research Objects Reviewed (ROR), (iii) Open Research Object and Research Objects Reviewed (ORO+ROR), (iv) Research Objects Reviewed and Results Reproduced (ROR+ROR-R), (v) Open Research Object and Research Objects Reviewed and Results Reproduced (ORO+ROR+ROR-R), and (vi) No badge. In other words accommodations were made for papers with proprietary code, so as to not dissuade them from the peer-review badging process.

The AD/AE co-chairs conducted 3 webinars in total: 2 were for reviewers to understand AD and AE process, and 1 was for authors and reviewers combined to understand the computing infrastructure to be used for evaluation. Authors were provided concrete guidance in terms of a variety of packaging methods that can make the evaluation task easier.

4 RESULTS OF AD/AE

300 papers underwent AD and 69 papers underwent AE. Each committee member had 7-8 papers to review in the AD phase and

3-4 artifacts to evaluate in the AE phase. In this section, we describe the outcome from each individual phase.

4.1 Artifact Description

The AD requirements, from the author's side, consisted of filling a form asking to report HW, SW requirements, environment configuration, source code links, and details of experiments.

Figure 2 shows the success of AD reporting after each AD form was reviewed by 2 reviewers. Out of 361 papers, 64 were desk rejects by the TPC. Out of the remaining 101 were marked inadequate ADs, 100 were marked OK, and 96 were marked Excellent. After 131 discussions, 87 of the ADs were updated. The TPC committee was informed that 7 out of all ADs are inadequate. Amongst the 99 papers that got accepted to SC signal, only 1 out of the 7 non-OK ADs was accepted by the TPC and was marked as major revision. The TPC was concerned about the artifacts description quality of this paper, and an acceptable AD was considered as a requirement for the TPC to accept the major revision. As the numbers report, the quality of AD after the discussion improved, and 55 were OK and 36 were marked excellent.

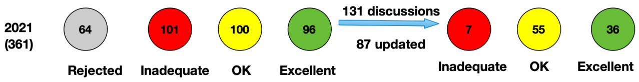

Figure 2: AD initial review Vs final signal to TPC

The following Figure (Figure 3) compares the average AD evaluation of accepted and rejected papers across different areas. In general, as we see the evaluation score is higher for accepted papers. We believe that mere participation does not translate to improved reproducibility—reproducibility is improved by independent verification and the review process introducing near-term deadlines in the process so authors can continue to enrich their artifacts.

The AD reviewers checked for completion of the AD form: hardware, software requirements, compilers, documentation, and validity of Github repositories. A text analysis of the AD requirements showed the following: Out of the 331 AD descriptions that we analyzed, the primary operating system of choice is Ubuntu (113), followed by CentOS (85), Red Hat (53), Suse (18), Mac OSX (8), Fedora (6), Windows (5), Cray (5), and others (38). The primary compiler used is GCC with 192 mentioning it. 41 also mentioned Python. CUDA and MPI were the common libraries that were mentioned.

The authors also mentioned some prominent computing platforms in their AD descriptions: ORNL/Summit(20), NERSC/Cori (15), ALCF/Theta(11), TACC/Frontera (10), Chameleon (3), ALCF/Mira (2), CloudLab(1). The following is the list of all supercomputers that were mentioned: Lassen, ABCI, Bebop, Sunway Taihulight, Catalyst, Blue Waters, Andes, Daint (Swiss National Supercomputing Centre), AIST, Frontera, Shaheen-II, Fugaku, Tianhe-3, Cheyenne, Theta.

4.2 Artifact Evaluation

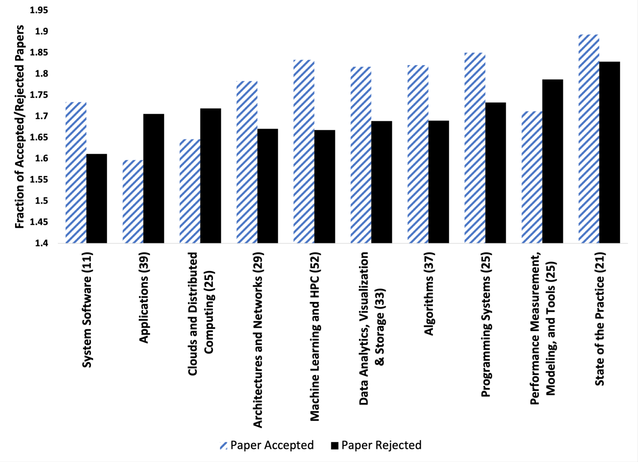

Figure 3: Average AD Evaluation within SC21 Areas

The AE process for SC21 consists of reviewing artifacts for AA (Artifact Available), AF (Artifact Functional), RR (Results Reproduced) badges. Authors were asked to apply for the badges at the time of submission of Artifact Description submission. The following Table (Table 1) provides how many authors of paper submissions and accepted papers indicated badge interest at the time of submission. The table indicates that for some authors', confidence in artifact evaluation increases once the paper is accepted.

	At Paper Submission	After Paper Acceptance
Total number	336	99
ORO	188 (51.9%)	68 (68.6%)
ROR	178 (49.2%)	59 (59.5%)
ROR-R	165 (45.6%)	52 (52.5%)

Table 1: Badge Interest at the Time of AD submission Vs After Paper was Accepted

Out of 99 accepted or major revision papers 69 applied for at least 1 badge (one paper did not apply for AA, but applied for AF). This number includes 2 major revision papers which were later on removed from the program.

The following figure (Figure 4) presents in which areas authors were more inclined to go for AE badges. Clearly, the winner is data analysis and visualization in which 9 out of 10 papers wished to be badged vs on the opposite spectrum was Performance, Measurement, and Modeling in which 4 out of 7 did not wish to be badged. The trend is clear–more systems and tools-based papers wished to be badged than papers in applications, state of practice, and performance measurement which are traditionally known to be hard to reproduce areas. Even in these difficult areas, a majority decided to apply for badges.

The AE review discussion process consisted of 133 assignments and a total of 240 reviews with an average of 1.92 out of 2 on reviews. The AE review process was an interactive, anonymous collaboration between reviewers and authors. A total of 879 committee-author comments were generated with the highest being 50 comments on a single paper. This data does not include interactions via Slack and Email.

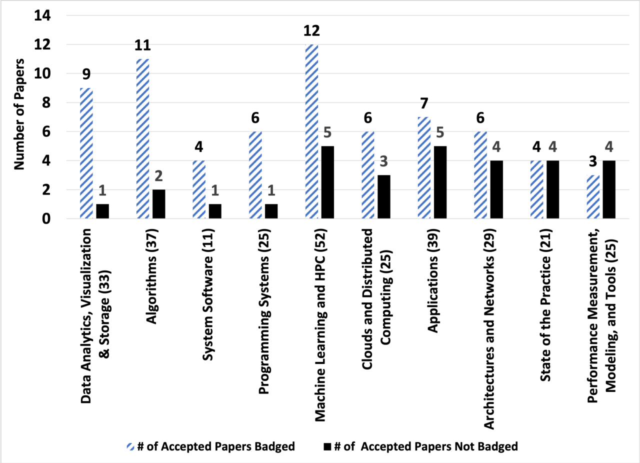

Figure 4: Author Interest in Badging grouped by Area

The reviewing lead to the following table (Table 2) of badge status:

	ORO*	ROR*	ROR-R*
Applied for	66	57	50
Granted	66	52	38
% Badged	100 %	91%	76%

Table 2: Badging results for accepted papers *(Excluding 2 (later rejected) major revisions which would have received all three badges)

Compared with all the papers, about 70% of accepted papers received at least one badge (67 out of 97). The percent of granted badges based on the applications is higher since we do not consider the full 98% of papers.

While we successfully completed the AE of all papers that applied for it in the given amount of time, there was always the overarching concern of the time and resources required to achieve AE. To assess AE, we gathered feedback from AE reviewers on a variety of questions. We summarize the salient points of this survey responded by 14 of the reviewers:

- 2 out of 14 did not think the evaluation criteria were entirely clear, but more than 92% say that the informational webinars helped.
- 75% spent 5-16 hours on the AD phase (reviewing 7-8 artifacts descriptions) and between 0-30 hours per artifact in the AE phase. The number of artifacts assigned in the AE phase was roughly 3-4 or 3.5 on average. This amounted to a weighted average of 13 hours per artifact or 40-50 hours overall artifacts per reviewer.
- More than 85% think that the number of assigned ADs was adequate. For AE, 71% would like 2 artifacts and 29% would like no more than 3 artifacts.
- The primary reason for marking a paper with inadequate ADs was that the AD description was too short for assessment, and either missed software requirements, hardware requirements or did not report version numbers. In particular, missing DOI/link was never the cause of rejection.
- 92% of reviewers used the comment mechanism to correspond with the authors. The three primary discussion topics were (i) build issues, (ii) insufficient documentation, and what are the primary results to be reproduced.
- 50% of the reviewers faced failures while evaluating, and 57% of the time the reviewers did not find adequate hardware.
- On the contrary, only 35% of the reviewers experienced authors not responding in time.

5 LESSONS LEARNED

We learned the following lessons based on conducting a comprehensive AD/AE process. We describe these lesson as means to sustain AE for similar peer-reviewed conferences.

Artifact Description is a necessary precursor, but one that is currently not self-contained. Collectively, the AD appendices are useful to build a comprehensive idea about the HW requirements for the AE phase. However, in its current form, the AD appendices do not provide details about (i) primary claims of the paper that the artifact supports, (ii) steps needed to go from claims to results, and (iii) results to be reproduced. Thus the form collects several setup details about the experiments but builds little understanding of the experiment from the reproducibility perspective.

Access to HW and community infrastructure requires more careful planning for successful AE. Our badging of papers indicates that reproducing a large subset of SC papers is possible even for ones with complex HW requirements (e.g., huge clusters, special HW), but, in our experience, more planning and committee organization is needed for a successful AE.

The AE window ranges from 3-4 weeks. Ideally, seamless access to proper HW resources should be possible during the AE reviewer assignment. However, currently, there is no such conference system that is linked to cloud or community providers. In fact, the current situation is very far from the ideal. We state steps that can help chairs be better prepared:

- Chairs must also hold webinars with community infrastructure providers well in advance so authors host their artifacts on these systems, to begin with. Authors must consider the use of containers and workflow systems to make it easy so different HW requirements do not pose build and environmental issues. This education and outreach must begin much earlier in the submission process since some experiments take long time to run.
- The process of obtaining access to supercomputing resources must be standardized and made easy. The access requires prior approval and is currently ties to the individual. This implies the chair changes the approvals must be reinstated. This is quite cumbersome and time-consuming. Instead, we hope that SC engages with the supercomputing facilities worldwide to come up with a reproducibility initiative policy that helps SC conduct AE in a seamless manner across years. This policy may include federated logins and passwords, liaisons, and automatic access approvals to SC AD/AE chairs independent of their personal details (such as visa/citizenship details etc.) We believe such an effort will make access toww resources much easier.

- Supercomputing facilities, on the contrary, must work towards viable interfaces for conducting reproducibility. We found the Chameleon project has the most useful interface so far in terms of tracking reviewers, authors, and overall compute usage. However, it still does not associate a reviewer with the artifact that is being evaluated. For instance, initially, reviewers requested access to the community infrastructure for a single artifact evaluation. But when we surveyed the reviewers, they reported using community infrastructure for as many as 3 artifacts. Thus accounting usage becomes a challenge.

The cost of AE is reasonable with author engagement and support. Anecdotally, the experience was valued by reviewers and authors. A post-AD/AE survey confirmed it. However, the survey also highlighted that future AE's should however limit the time commitment of AD/AE reviewers by limiting to 2 artifact evaluations per reviewer.

As our survey results show currently there is a noticeable cost of conducting an AD/AE review. Contrasting AD/AE with paper reviewing, the paper reviewing load of a TPC is often known in the number of papers, the number of hours is never monitored. The quality of review is often an indication of the time spent in reviewing but is something very hard to measure. On the contrary in AD/AE, the objective is to fix the number of hours and then see how much evaluation can be performed in that fixed time. For SC21 we had estimated 10-15 hours for AD and 20 hours per artifact for 3-4 artifacts. Our numbers did not fall off significantly from this value; though reviewers mentioned that they would appreciate a lower number of total artifacts to be evaluated.

We believe that setting expectations right from the beginning leads to almost no complaints about the process. The authors were incentivized about the badges and the reviewers were aware of what they were in for. Consequently, we believe that similar to high-quality papers published at SC, the AD/AE process resulted in high-quality artifacts, which will advance future SC research.

Reproducibility is a spectrum that requires constant engagement—AD/AE is one milepost. As our process indicates, we never considered AD/AE as a filtering step that filters papers based on scientific merit. Instead, our philosophy was to help authors take advantage of their AD/AE process and improve their artifacts in a gradual manner. Our experience indicated that addressing the reproducibility of the artifact at the point of submission is the best way to improve artifact quality because that is a time when authors are engaged with the research process of the paper. It will be difficult to oversee the reproducibility of such a large scale of papers at any later point in time. Not all papers reached the highest artifact quality but many papers were made aware of the current state via reviewer comments. While an artifact standard is currently missing, the large number of comments by the AD/AE committee indicated to us that improving an artifact requires continuous engagement, and a standard could pose as a policing statement and be more detrimental to the process.

If embraced and sustained, the future opens up an exciting array of possibilities. The computational reproducibility process conducted via AD/AE provides a unique peek into the reproducibility process. As our survey and comment analysis indicates, build issues and lack of documentation are the most common issues that prevent an artifact from being reproduced. This opens up an interesting set of technical possibilities in the area of package managers, containers, and community infrastructure. Establishing standardization issues that can make it easier and more accessible for the purposes of achieving reproducibility. We also experienced in a few cases how screencasting makes it much easier to evaluate. However, it does not involve independent evaluation by a reviewer. We believe the screencasting calls for improved technology solutions combining augmented reality, screen sharing, and collaborative environments. Finally, we believe the right set of incentives will keep the practice sustainable. We were proud to instate the Best Reproducibility Advancement award as an incentive for authors to improve the quality of the artifact. Similar honorable mentions on the reviewer side will make this process worthy of their time.

6 CONCLUSIONS

The SC21 AD/AE advanced the current state of the reproducibility practice at SC in several ways: (i) adopting an incentive-based approach of engaging authors, requiring minimal involvement of the Technical Program Committee (TPC); (ii) laying out the rules for applying ACM Badges in a way that is consistent irrespective of the publisher; (iii) using community computing infrastructure for Artifact Evaluation; (iv) establishing the Best Reproducibility Advancement Award and choosing its first recipient; and finally (v) engaging with more than 50 members of AD/AE in a timely manner increasing awareness and enthusiasm.

The AD/AE process, which lasted an year, demonstrated that sustaining Artifact Evaluation at peer-reviewed conferences requires more than just planning: it requires sufficient representation and participation, encouraging incentives, and an enthusiastic community.

ACKNOWLEDGMENTS

We would like to thank Beth Plale and Andrew Younge for desccribing the history of Reproducibility Initiative, and more specifically on successfully handing over the past year SC effort. We would like to thank the SC21 steering and organizing committee members: Bronis R. de Supinski, Jeffrey Hollingsworth, David Abramson, Ilkay Altintas, Mary Hall, and Todd Gamblin for their continued support, discussions, and patience to develop an artifact description/evaluation process based on voluntary badging. We would like to thank Michela Taufer for championing the award. Finally, we would like to thank the artifact evaluation committee for their relentless effort to review and evaluate artifacts.

Tanu Malik acknowledges support from National Science Foundation under grants CNS-1846418, ICER-1639759, ICER-1661918 and a Department of Energy Fellowship.

REFERENCES

[1] NISO RP-31-2021, "Reproducibility Badging and Definitions", https://doi.org/10.3789/niso-rp-31-2021
[2] Programming Language and Software Engineering Artifact Evaluation. https://artifact-eval.org/
[3] SC21 Reproducibility Initiative. https://sc21.supercomputing.org/submit/reproducibility-initiative/
[4] Systems Research Artifacts. https://sysartifacts.github.io

[5] "Reproducibility and Replicability in Science: a Consensus Study Report", H. V. Fineberg, Committee Chair, National Academies of Science, Engineering, and Medicine, 2019

[6] Lorena A. Barba, "Trustworthy computational evidence through transparency and reproducibility", Computing in Science and Engineering, 23(1):58-64 (2021), IEEE Computer Society.

[7] B. Plale and S. L. Harrell, Transparency and Reproducibility Practice in Large-Scale Computational Science: A Preface to the Special Section, IEEE Trans. Parallel Distributed Systems, 32(11), pp. 2607–2608, 2021

[8] B. A. Plale, T. Malik and L. C. Pouchard, "Reproducibility Practice in High-Performance Computing: Community Survey Results," in Computing in Science & Engineering, vol. 23, no. 5, pp. 55-60, 1 Sept.-Oct. 2021, doi: 10.1109/MCSE.2021.3096678.

A PACKAGING METHODS

Authors were provided with guidance to consider one of the following methods to package the software components of their artifacts (although the AEC is open to other reasonable formats as well):

- Source Code: If your artifact has few dependencies and can be installed easily on several operating systems, you may submit source code and build scripts. However, if your artifact has a long list of dependencies, please use one of the other formats below.
- Virtual Machine/Container: A virtual machine or Docker image containing the software application already set up with the right toolchain and intended runtime environment. For example: For raw data, the VM would contain the data and the scripts used to analyze it.

  For a mobile phone application, the VM would have a phone emulator installed. For mechanized proofs, the VM would contain the right version of the relevant theorem prover. We recommend using a format that is easy for AEC members to work with, such as OVF or Docker images. An AWS EC2 instance is also possible.
- Binary Installer: Indicate exactly which platform and other runtime dependencies your artifact requires.
- Live Instance on the Web: Ensure that it is available for the duration of the artifact evaluation process.
- Internet-accessible Hardware: If your artifact requires special hardware (e.g., GPUs or clusters), or if your artifact is actually a piece of hardware, please make sure that AEC members can somehow access the device. VPN-based access to the device might be an option.

B ONLINE RESOURCES

- AD/E Committee Survey Results https://tinyurl.com/sc21adecommitteesurvey
- Conferences with artifact evaluation and awards https://tinyurl.com/confswithaeawards