Implementing a Reproducibility Initiative in HPC: Experiences from SC'24

Guillaume Pallez
Inria Rennes
Rennes, France
guillaume.pallez@inria.fr

Judith Hill
Lawrence Livermore National Laboratory
Livermore, CA, USA
hill134@llnl.gov

Sascha Hunold
TU Wien
Vienna, Austria
hunold@par.tuwien.ac.at

Abstract

Reproducibility is fundamental to scientific research, but can be particularly challenging in research that involves High Performance Computing (HPC) due to the unique characteristics of supercomputers. Performance-based metrics such as execution time, energy consumption, and throughput further complicate reproducibility, especially on shared systems.

In this paper, we present our experience implementing a reproducibility initiative at SC24, with particular emphasis on changes made compared to prior SC conferences. We outline HPC-specific challenges, describe the measures adopted to address them, and reflect on the limitations of reproducibility badges. Faced with the constraints of the existing badging nomenclature, we discuss our implementation of a reproducibility report, which aims to provide more context about the reproducibility of each paper. We conclude by recommending that the "Artifact Replicable" badge be dropped by HPC conferences at this time, and discuss alternate ways of ensuring replicability evaluation.

CCS Concepts

• **General and reference → Validation**.

Keywords

Reproducibility, HPC, SC, Badges, Reviewing, Best Practices

ACM Reference Format:
Guillaume Pallez, Judith Hill, and Sascha Hunold. 2025. Implementing a Reproducibility Initiative in HPC: Experiences from SC'24. In *ACM Conference on Reproducibility and Replicability (ACM REP '25), July 29–31, 2025, Vancouver, BC, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3736731.3746148

1 Introduction

Reproducibility is a key element in furthering scientific research. Reproducing scientific studies is essential to ensure that findings are reliable and to build confidence within the broader community. In the context of High Performance Computing (HPC), reproducing results can be particularly challenging due to the individual properties of supercomputers, which can make it difficult to replicate results on different machines because of their varying architectures and software environments.

One important aspect of HPC research is often the target metrics, which are commonly performance-based. Papers and the methods that are developed are evaluated based on their performance on a specific machine, which could be the time to solution, the energy consumption, or the throughput. Solving a problem fast with less energy is as important as solving it at all. This makes it difficult to reproduce results as performance reproducibility is inherently challenging [11, 13], especially on multi-user supercomputers managed by batch scheduling systems such as Slurm [29].

In this paper, we present our experience in implementing a reproducibility initiative at the 2024 Edition of the *IEEE/ACM International Conference for High Performance Computing, Networking, Storage, and Analysis* (SC24). We discuss the challenges specific to HPC and SC, and the solutions we implemented to address them and reflect on the lessons learned from our experience.

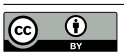
Our main contribution is to raise awareness on the limits of reproducibility badges within the HPC community, discuss our experience addressing these limits with a reproducibility report.

In Section 2, we present related work on Reproducibility badges, along with the history of reproducibility in the SC conference series. Then we provide a brief overview of the challenge of reproducibility for High-Performance Computing (HPC) in Section 3 from a practical perspective and also when considering their evaluation. Next in Section 4, motivated by these challenges we discuss the changes that we made by introducing the reproducibility report and the limits of the report. Finally, based on this experience we provide concluding remarks in Section 6.

2 Reproducibility and Recognition Badges

In this paper, we use the term "reproducibility" to mean that the results of a study can be reproduced subsequently by other researchers. Thus, we use the term "reproducibility" interchangeably with "replicability" and "repeatability", as our primary interest is in the ability to reproduce scientific results. However, upon closer examination of the literature, there are important distinctions among these terms. About a decade ago, ACM defined the Artifact Review and Badging process [1]. In particular, ACM clearly distinguishes between the terms "repeatability", "reproducibility", and "replicability". In the latest version of the ACM Review and Badging guidelines [1], the concept of reproducibility is refined by distinguishing

This manuscript has been authored by Lawrence Livermore National Security, LLC under Contract No. DE-AC52-07NA2 7344 with the US. Department of Energy. The United States Government retains, and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.



This work is licensed under a Creative Commons Attribution 4.0 International License.
ACM REP '25, Vancouver, BC, Canada
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1958-5/25/07
https://doi.org/10.1145/3736731.3746148

who is conducting the experiments and what type of experimental setup is used. In particular, the ACM defines these three terms as follows:

- **Repeatability**: Same team, same experimental setup.
- **Reproducibility**: Different team, different experimental setup.
- **Replicability**: Different team, same experimental setup.

With these definitions, ACM follows the suggestions of National Information Standards Organization (NISO) [21].

The IEEE has also been an early adopter of reproducibility badges. For example, the IEEE Transactions on Parallel and Distributed Systems (TPDS) introduced a reproducibility badge in 2019 [22], where authors can "earn a reproducibility badge by submitting their associated code for post-publication peer review."

Reproducibility badges are a way to recognize the effort of authors to make their work reproducible. Conferences and journals can use badges to indicate the level of reproducibility of a paper. For the following discussion, authors of accepted papers at SC24 could apply for three different reproducibility badges:

(1) **Artifact Available**: The artifact is permanently archived in a public repository.
(2) **Artifact Functional**: The artifact provides enough information to exercise the artifact's components.
(3) **Artifact Replicable**: The artifact can be used by other researchers to reproduce the key results of the paper.

Authors could apply for all three badges; badges were ultimately awarded based on evaluations by reviewers from the reproducibility committee who assessed the quality and executability of the artifacts provided by the authors. Therefore, an accepted paper could receive one, two, or three badges. For SC24, the badges were incrementally awarded, i.e., if a paper received the Artifact Replicable badge, it also received the Artifact Functional and Artifact Available badges.

2.1 History of Reproducibility at SC

The *IEEE/ACM International Conference for High Performance Computing, Networking, Storage, and Analysis* series, a.k.a *SC* is one of the leading conference series in the HPC community. The 2024 edition attracted more than 18, 000 people. In the rest of this work, when we talk about SCXY, we mean the 20XY edition of the SC conference.

Historically, as a flagship conference in HPC, SC has been open to modifying and updating its processes, such as the one we describe here, for the betterment of the broader field of HPC. Often those modifications are implemented on a trial basis and evaluated to determine whether the change achieved its desired objective. As a result, many other conferences often follow the lead of SC and implement similar process improvements after SC tries and evaluates them.

The Reproducibility Initiative for the SC conference series has its roots in the SC15 conference, where authors were invited to submit an *Artifact Description Appendix (AD)* after the conference.[1] Since then, the Reproducibility Initiative has evolved significantly, where detailed information about the respective initiatives can be found on the SC conference webpages. At SC16, the AD appendix was introduced as a way to provide additional information about the artifacts used in the paper. At this time, the appendix was limited to two pages and could only be used positively to favor a paper if it was needed to distinguish between two potentially accepted papers during the review process. Only nine papers included an artifact description in the proceedings [20].

The initiative for SC17 went one step further, as the Artifact Description Appendix was required to be considered for best paper. In addition, another appendix was introduced at this conference, which was the *Computational Results Analysis Appendix*, which later became the *Artifact Evaluation Appendix (AE)*. At SC19, the Artifact Description Appendix was mandatory for all submitted papers, while the Artifact Evaluation Appendix was optional. A special AD/AE evaluation committee was introduced to review the AD/AE appendices, which was composed of 11 members. This committee increased to 14 members for SC20, and the number of members was further increased to 48 for SC21. SC21 also introduced the first reproducibility badges. At SC23, the number of members was further increased to 60, to accommodate the increasing number of artifacts.

At SC24, the SC24 Reproducibility Initiative first required the authors to clearly state the key results of their paper in the AD appendix, and which paper elements reflect these key results (e.g., figures, tables, etc.). The SC24 Technical Paper Reviewers were also asked to confirm or identify which results should be attempted to be reproduced by the Reproducibility committee. Finally, a new reproducibility report, which is a non-anonymous and public report written by the reproducibility committee members to document which key results could be reproduced and why certain badges were awarded, was also introduced.. The number of reproducibility committee members was substantially increased to 101 owing to the significant amount of work required during the reproducibility review.

Separately and in addition to the development of appendices, another major part of the reproducibility initiative at SC has been the reproducibility of a selected paper from the previous year by the student cluster competition teams[2]. The inclusion in the competition started in SC16 based on a SC15 paper. This aspect of reproducibility at SC will not be discussed further in this paper.

2.2 Related Work

Reproducibility has been an important topic in the Computer Science community for many years [6]. With a focus on the area of Parallel Computing, Hunold and Träff [14] discuss the lack of reproducibility and the need for improving the reproducibility in this field. In this report, they highlight four key challenges: clearly defining reproducibility, determining whether (and how) reproducibility impacts scientific progress, enforcing stricter publication rules, and improving software tools to better support reproducible research.

The SC conference series was among the first to adopt appendices for detailing artifact descriptions and evaluations. In 2021, Plale et al. [24] presented survey results from the SC community on reproducibility. The survey collected data from individuals who had participated in the SC17, SC18, or SC19 technical programs. One of

[1] https://github.com/SC-Tech-Program/SCreproducibility

[2] https://sc16.supercomputing.org/2016/03/16/sc16-explores-reproducibility-advanced-computing-student-competition-michela-taufer/index.html

the main findings was that the "SC community reproducibility effort has contributed to higher levels of awareness" of reproducibility within the HPC community.

A recent study by Guillauteau et al. [10] surveyed almost 300 articles from five leading conferences on parallel and distributed systems held in 2023 (CCGrid, EuroSys, OSDI, PPoPP, and SC). In particular, they focused on the longevity of artifacts. For example, the authors investigated how many artifact URLs are still functional, only one year after the publication of the articles. With respect to the reproducibility of the artifacts, the authors observed challenges with properly specifying software dependencies and relying on classical package managers. The authors also stated that the "community could greatly benefit by adopting workflow managers." Finally, they proposed a "new badge to reward artifacts that will withstand the test of time." [10].

Krassnitzer [18] made similar observations during a reproducibility study of SC24 papers that applied for the Artifact Replicated badge and were potentially reproducible on Chameleon Cloud [17]. After a detailed investigation of 18 papers, Krassnitzer compiled a list of suggestions to improve artifact reproducibility, including enhancing the documentation for artifact evaluation, clearly specifying software dependencies, improving access to experimental data, and explicitly stating the expected results of the experiments.

In a 2020 blog entry, Beller[3] clearly explains why he will no longer review artifacts. One of the major challenges he identifies is the significant time required to reproduce a paper's results, as many artifacts do not work out of the box and often lack a "step-by-step walkthrough on how to get the paper's results."

3 Challenges of Reproducibility for HPC

To provide background on the decisions that were made for the Reproducibility track of SC24, we start by discussing challenges of reproducing results in the HPC community and constraints specific to the SC conference series.

3.1 General Challenges to Reproducibility

We have seen some of the challenges related to reproducibility which were well summarized by Beller (see Section 2.2).

In practice, from a conference perspective, one of the challenges is determining the appropriate size of the reproducibility committee. Too few reviewers impacts their individual review workload and hence the quality of their reviews given the limited time available for the review. Conversely, recruiting a large number of reviewers to manage the individual workload can also be challenging for the committee chair(s). In addition, with a large committee, creating a unified strategy for the review, communicating their expectations, and ensuring sufficient review quality is a challenge for the committee chair(s).

The reproducibility review process also continues to be a work-in-progress under constant evolution. As a result, and likely due to the changing requirements, our observation is that many authors and reviewers alike do not follow the instructions given, which further complicates the process. This may not be surprising: for the past eight years, the reproducibility process at SC has been evolving rapidly with significant changes nearly every year, compared to the Technical Paper review which has largely remained constant with only minor changes each year. As a result, paper authors who have submitted previously may incorrectly expect that both processes remain unchanged year over year.

[3]https://inventitech.com/blog/why-i-will-never-review-artifacts-again/

Table 1: SC badge statistics from SC21 to SC24, showing the number and percentage of papers that received each badge.

	No badge	Available	Functional	Replicable	Num. Papers
SC24	21 (21%)	81 (79%)	47 (46%)	25 (25%)	102
SC23	30 (33%)	60 (67%)	46 (51%)	33 (37%)	90
SC22	19 (23%)	62 (77%)	49 (60%)	36 (44%)	81
SC21	32 (33%)	66 (67%)	52 (53%)	38 (39%)	98

3.2 Challenges specific to HPC

There are several challenges to reproducibility that are unique to very large HPC environments. One of the obvious limitation is the scale of the experiments done for some of the work presented. In extreme cases, such as papers that are contenders for the Gordon Bell prize, [4] reproducing experiments is simply not possible (e.g., no other machine exists of the equivalent capability to run the simulation), or reproducing them is deemed too difficult or expensive (e.g., another comparable machine exists, but the reproducibility reviewer cannot gain access to it in a timely fashion, or the computational requirements of the simulation outweigh the value of the reproducibility experiment).

In addition, there is a fundamental challenge related to the main metrics of HPC: performance, or time to solution [11, 13]. Applications are often tailored to the use of their specific HPC resources in order to reach the peak performance of a particular machine. Hence reproducing the numerical results on a slightly different machine (even one with the same underlying compute hardware but a different type of network, for example) will not give similar performance.

These challenges show the struggle of awarding reproducibility badges to HPC papers; most HPC papers submitted to SC fall into one of these two categories. Yet, the HPC papers that pose significant reproducibility challenges are often those for which demonstrating the reproducibility of results is crucial to the research.

Limits of badging system for HPC. We discussed in Section 2.2 the limits of the *Artifact Available* badge as shown by Guillauteau et al. [10] and the fact that reproducibility artifacts often do not survive the test of time. With the challenges articulated above, we believe that these also demonstrate the primary limit of the badging system for our field: the *Artifact Replicable* badge is too *binary*. Within the HPC community, it is our impression that very few papers would be considered *replicated* by the ACM definition of Reproducibility in Section 2.

We also studied badge statistics since SC started to award them to its accepted papers. (see Table 1). Prior to SC24, almost 40% of the accepted papers received the badge *Artifact Replicable*. This seems to

[4]The Gordon Bell prize, https://awards.acm.org/bell, is a prize that often rewards peak performance or special achievements in scalability and time-to-solution on some of the largest HPC machines.

contradict our position that reproducibility in HPC is particularly hard. This number went down to 25% in SC24. We believe, and informal discussions with reviewers seems to confirm, that this decrease was due to the communication with reviewers about the expectations required to receive an *Artifact Replicable* badge.

Finally, another limit of the binary aspect of badges is the expectation from authors that they should receive them. Indeed, given the evolution of the evaluation of science, there is an important risk that they become "indicators", creating well-known perverse effects that risk making them useless [2, 4, 15]

3.3 SC-Specific Challenges

In addition to the general challenges associated with reproducing research done on or with HPC computers, there are a handful of challenges associated with the SC Technical Paper review process that may not be generalizable to other HPC venues.

Review committe organization. Within the SC conference organization, the merit review of a technical paper is conducted by the Technical Paper committee, and the evaluation and awarding of the Reproducibility Badges is conducted by the Reproducibility Committee. The primary benefit of separating these two committees is that it reduces the workload for all persons. The Technical Papers committee already must thoroughly review seven to eight papers per person within a timeframe of approximately 45 days. The Reproducibility evaluation is also on a fairly short timeframe and can often require significant time on the part of the Reproducibility reviewer if technical challenges are encountered or if the artifact is not easily reproduced. A secondary benefit of separating these two communities is that it allows the Technical Paper review to be double-anonymous; whereas there are benefits to having the reproducibility evaluation be open (discussed in Section 4.4).

One drawback of this separation is that often the Reproducibility reviewers are often not as familiar with the paper and may not fully understand the main results of the paper that should be reproduced without investing additional time.

Timeline. A second challenge within the constraints of the SC conference organization is the review and publication timelines. The SC conferences prefer to have the conference proceedings available by the start of the conference. As a result, the reproducibility evaluation must occur after the papers have been reviewed (ending in mid-June) and before the proceedings deadline (usually early August), during a time period when many potential reviewers may have other commitments. Because of the compressed timelines, the number of papers that are submitted versus the number ultimately accepted, and the time required for the reproducibility review, only accepted papers are evaluated for replicability.

Unique hardware requirements. One of the central challenges of the reproducibility initiative at SC is hardware access. Many papers report performance results as their primary contribution, often obtained on highly specific hardware configurations. To mitigate this issue, the Chameleon Cloud [17] offers a valuable platform for sharing computational artifacts between authors and reviewers. Chameleon has previously supported SC reproducibility efforts and provides a broad range of computational resources to the community. However, specialized hardware—such as high-end GPUs—is typically in high demand. Because these resources are shared, they are not readily available on short notice. Reviewers must often reserve such nodes weeks in advance. As a result, reproducibility reviewers must adapt their schedules and plan significantly ahead to ensure resource availability for conducting timely evaluations.

4 SC24 Reproducibility Initiative: Vision, Implementation, and Lessons Learned

In this section, we start by discussing our vision of the role of reproducibility evaluation in the SC24 conference and the concrete actions that were taken towards realizing our stated goals. We conclude with a handful of lessons learned and recommendations for others who may be considering implementing a reproducibility initiative.

4.1 Reproducibility: Ideals versus Practicalities

Reproducibility is at the core of the scientific process: for a result to be accepted, other scientists should be able to replicate it. This is the goal that we aim for: ideally, we would like all published results to be reproducible.

However, a reproducibility track should be evaluated also considering the effort required on the part of both the authors and the evaluators, and compared to how well the articulated goal is or is not achieved. When we implemented the reproducibility track in the SC24 conference, our goal was *not* to validate that all the accepted work was reproducible. If that had been the primary goal, all papers that could not be fully reproduced would have been rejected, and we have already stated in Section 3 many of the reasons that reproducibility is particularly challenging within HPC.

This is the difference between ideal and practical reproducibility. For SC24, our primary goals within the Reproducibility track were to *encourage and promote reproducible research within the SC community.* This is why the SC reproducibility process was designed with incentives such as badges and a requirement of an artifact appendix for best paper finalists rather than the binary output of a reproducibility process.

4.2 Simplifying the Authors' Job

Although reproducibility of scientific research is widely spoken of within the community in positive terms, we conjecture that it is often an after-thought during the preparation of a scientific paper. One explanation for this may be that the creation of the artifact appendices is still very much in the development and debugging phases.

In past SC conferences, several solutions have been tried to improve the quality of the reproducibility appendices, such as delaying the time at which the authors submit their appendices so it does not compete with the main paper (starting in SC22), or including fill-in boxes directly in the submission system to automatically generate it (SC22).

For SC24, we maintained the delay in the submission of both artifact description appendix (two weeks after paper submission) and the optional artifact evaluation appendix (only optionally submitted after paper acceptance). However, we abandoned the automatic generation of the appendices due to the challenges of writing a LaTeX webcompiler.

Artifact ID	Contributions Supported	Reproduced Paper Elements
A_1	C_2	Figure 7a–c
A_2	C_3	Figure 8

Figure 1: Example of an artifact identification table in paper [19].

Artifact ID	Available	Functional	Replicated
A_1	•	•	•
A_2	•	•	•
Badge awarded	yes	yes	yes

Figure 2: Example of an artifact review table in the Artifact Evaluation Report [25] for paper [19].

The SC23 reproducibility chairs began standardizing the reproducibility appendices to help authors understand the expectations. For SC24, we required authors to explicitly identify provided artifacts, specify which contributions these artifacts supported, and indicate the exact results in the paper that could be reproduced (see Figure 1). Additionally, we provided a detailed description of the reproducibility process on the SC public website[5].

Finally, one of the challenge was the lack of clear information shared from one year to the other. In SC24 the reproducibility chair documented the entire reproducibility evaluation process so that future chairs can refer to and modify it: https://github.com/hunsa/sc24-repro.

4.3 Simplifying the Reviewers' Job

One of the main challenges of the reproducibility evaluation committee is to recruit committee members due to the time requirements necessary for a thorough review and the timeframe under which the review must occur (see Section 2.2). In the past at SC, these reviewers have mostly been early career researchers (PhD students in their final years, postdocs, etc.), with the idea that this gives them a first experience into reviewing papers.

In order to make this simpler, we have implemented several modifications to the artifact evaluation process in an effort to reduce the overall time requirement of the evaluation process.

Non anonymity of reviewers with respect to authors. Historically, researchers are accustomed to at least a single-anonymous review process (peer reviewers are not known to the authors). The question as to whether to maintain single-anonymous (or double-anonymous) reviews in reproducibility has been an important question [27]. At least single-anonymous reviews has the benefit of giving the reviewer the ability to be frank and open in their reviews without fear of retribution.

Anonymity in the reproducibility track raises several challenges, but most importantly it makes the job of the reproducibility committee much harder. While the job of the reviewer could stop once they hit a blockade, it would go against our goal of improving the scientific quality and reproducibility of the papers (Section 4.1). In past SC conferences, the overall program and reproducibility chairs wanted to keep this anonymity, hence reviewers would ask their questions to the reproducibility chair who would then transfer the question to the authors.

For SC24, our vision however was different. Because the only papers that are reproduced are already accepted papers, we believed that the possible conflict related to non-anonymity of reviewers is less present. Hence, we determined that the cost of anonymity compared to the drawbacks was not worth it. The main advantages of breaking the anonymity of the reviewers is that reviewers could have the opportunity to gain access to specialized HPC hardware (e.g., supercomputers or novel architectures) if needed for the evaluation. This opportunity was used at least once [5]. In addition, it shortened the iterative question-and-answer exchanges between the reviewers and authors when technical obstacles in the evaluation were encountered [23].

Reducing the review load. In past SC conferences, qualitative feedback from many of the reproducibility reviewers was that the time requirement to conduct a single evaluation was large, with some reviewers taking days or weeks to complete the review or to eventually give up. Because the reproducibility reviewers are often very early career researchers, this could be a detriment to their career, particularly when there is little external recognition for this service. To this end, we implemented several practices with the goal of reducing the time commitment required by each reproducibility reviewer.

First, we set an upper bound on the amount of time for each paper evaluation, limiting it to eight hours or a single working day. This time did not include the time for the actual experiments (wallclock time on a computer), but did include the time required to understand what should be reproduced, to set up the experiments on the appropriate hardware including any needed software environment, and to write the reproducibility report. Second, we reduced the number of papers each reviewer should evaluate to two papers per reviewer, which implied increasing the overall size of the committee. And finally, we provided a "Reproducibility report template" (described in more detail below) that the reproducibility reviewers used to document their efforts in a consistent and concise way.

Partnering with Chameleon Cloud. Chameleon Cloud is a cloud-based infrastructure that has supported the SC Reproducibility Initiative for several years. It provides access to real HPC hardware hosted at institutions such as the University of Chicago and the Texas Advanced Computing Center (TACC).

A key advantage of Chameleon for the SC Reproducibility Initiative is that its computational infrastructure can be shared between authors and reviewers. Chameleon allows authors to package their artifacts in a reproducible manner, either by saving and restoring full OS images or by programmatically configuring hardware and software environments using tools such as Jupyter notebooks.

While Chameleon Cloud is a powerful platform for reproducibility, it requires users to learn how to use it effectively. Reproducibility is not automatic and comes with a learning curve that both authors and reviewers must overcome. To support this, the Chameleon team

[5]https://sc24.supercomputing.org/program/papers/ad-ae-appendices/, accessed June 2025.

hosted two webinars prior to the review period to explain how to use the system and how to package artifacts properly.

While Chameleon Cloud facilitates the setup of hardware and software environments [16], the actual software artifact must still be packaged by the authors. One of the main challenges in this process is managing software dependencies. Several HPC-specific package managers are available to assist with this task, including Spack [9], EasyBuild [12], and Guix [7]. For SC24, several members of the Reproducibility Committee with experience using Guix provided a detailed guide to support its use.[6]

4.4 Introduction of the Reproducibility Report

The largest change that we introduced in SC24 was the introduction of the reproducibility report. This goal of this report is to describe the reproducibility process, what was reproduced and to what extent, what was not reproduced, and why. This report was motivated by our concerns about some of the limits of badges, particularly in terms of interpretability: If you do not have access to the full scale required to reproduce a result, but are able to reproduce it at the scale available to you, should this paper receive a "result replicated" badge? Would two different reviewers award the same badge? We wanted to provide more information about what was and what was not evaluated when the badge was or was not awarded. We provided authors and reviewers a template to ensure that all reports were consistent and to simplify the report-writing process for the reviewers.

Non-anonymous report. After deciding that reviewers could share their identity with authors (see Section 4.3), we also encouraged reviewers to include their names on the public report. It was our belief that this would showcase their contribution, particularly given the fact that most reproducibility reviewers are early career.

Open reviews had been shown to make reviewer recruitment harder [28], although this data is quite old and may not be still applicable given the evolution of science and new journals appearing with open reviews (e.g., Journal of Open Science Software [26]). Hence we had expected the same risk with the introduction of the reproducibility report. To mitigate this concern, all reproducibility reports were co-authored by two reviewers, and we added a disclaimer giving the context of the reproducibility evaluation (such as limited time), see Figure 3.

The introduction of this (optionally) non-anonymous report did not impact reviewer recruitment. At the time of the invitation, committee members were informed about the report and the fact that they could not be anonymous. In 2024, recruitment acceptance ratio was similar to the ones of 2023 and 2022. Of the 82 Reproducibility Reports published in the SC24 proceedings, 6 reports were anonymized or partially anonymized (i.e., one reviewer out of two remained anonymous).

Publication. The Reproducibility reports were published as supplementary material on IEEE's webpage, associated with each of the accepted papers (that requested badges). This additional effort did create a small amount of overhead on the SC24 proceedings chair.

[6] https://gitlab.inria.fr/guix-hpc/sc24-reproducibility-initiative

Disclaimer: This Artifact Evaluation Report was crafted by volunteers with the goal of enhancing reproducibility in our research domain. The time period allocated for the reproducibility analysis was constrained by paper notification deadlines and camera-ready submission dates. Furthermore, the compute hours in the shared infrastructure (e.g., Chameleon Cloud) available to the authors of this report were limited and restricted the scope and quantity of experiments in the review phase. Consequently, the inability to reproduce certain artifacts within this evaluation should not be interpreted as definitive evidence of their irreproducibility. Limitations in the time allocated to this review and the compute resources available to the reviewers may have prevented a positive outcome. Furthermore, reviewers assess the reproducibility of the artifacts provided by the authors; however, they are not accountable for verifying that the artifacts support the main claims of the paper.

Figure 3: Disclaimer to reproducibility report

4.5 Identifying Key Results to be Reproduced

Finally, the last question that we asked ourselves was about the result that should be reproduced. The ACM definition[1] of the "Artifact Replicable" badge states: *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.*

However, it raises the question on how to identify what the reproducibility committee should evaluate. As described earlier, the authors provided a clear and concise table in the Artifact Description detailing the artifacts provided and the results that they were to reproduce.

To corroborate the information provided by the authors, we asked a new question of the paper reviewers:

> Assuming the paper is accepted, and assuming unlimited resources: what is/are the key result(s) of this paper that the reproducibility committee should focus on (it can be a Figure, a Table or other artifact in the paper). This comment will not be shared with the authors.

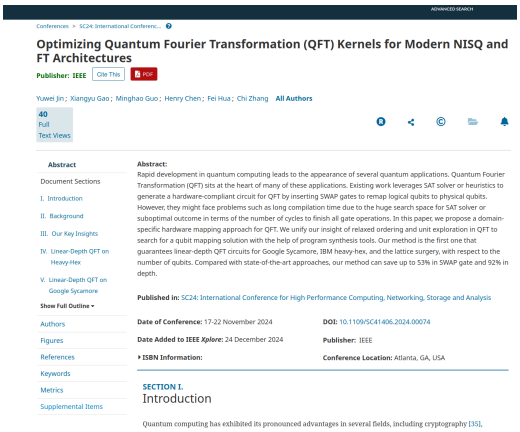
Together, these two pieces of information were available to the reproducibility committee to give them direction on what to focus on in their evaluations.

4.6 Lessons Learned and Recommendations

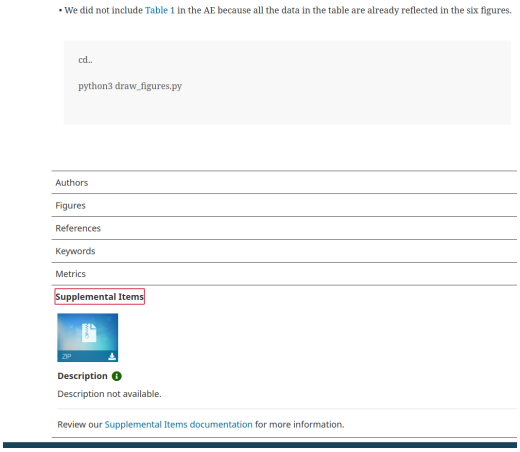
The SC24 Reproducibility Initiative was another step, building on the work of several prior years, to encourage the SC community and the HPC community writ large to further embrace reproducible science. The process continues to evolve and will continue to need further contributions from the community. While we are satisfied with the progress and improvements made this year, we have several recommendations for the community to consider for future instantiations.

Completeness and availability of the reproducibility report. For SC'24, due to our own workload and for acceptability, we reduced the expectations on the reproducibility reports.

If the report is independent of the paper as we chose to make them, then they should ideally be thorough, reminding the reader of the paper claims, and with performance figures. Currently the



(a) The report is a the very bottom of the paper, in the section "Supplemental Item"



(b) Until we download the supplemental item, we cannot know that this is the reproducibility report

Figure 4: Report for the paper "Optimizing quantum fourier transformation (qft) kernels for modern nisq and ft architectures" on IEEE's website: https://ieeexplore.ieee.org/abstract/document/10793127 (accessed Apr. 2025).

performance can be well described, see for instance the "results replicated" report by Exposito and Zhou [8] which discusses the fact that *"The results obtained for the baseline (...) vary very slightly with those shown in the paper. According to authors, this behavior is expected due to the use of random seeds"*, but producing the actual data would likely help.

Another concern is their visibility. Currently the reports are extremely hard to find: they appear in a supplemental section on the IEEE website (see Figure 4).

Recommendation: If conference organizers choose to include a Reproducibility Report that documents the results of a reproducibility evaluation and chooses to keep them as an independent publication separate from the paper itself, then they should have their own DOIs and proceedings.

However, our recommendation is to include them as supplementary material to increase their visibility, particularly to the readership of the papers.

Artifact Replicable badge. Even with limiting the time commitment for each reproducibility review, awarding the Artifact Replicable badge is a substantial amount of work. In addition, increasing the number of reviewers creates a major load on the chair of the committee, as they need to deal with the *human* challenges, such as making sure that people have read and understood the instructions, that they attend any information sessions or seminars, keeping track of late reviewers, etc.

The overall review and publication timelines makes it particularly challenging. After this experience, we recommend for conferences that offer these badges to drop the *Artifact Replicable* badge.

Recommendation: Conferences should drop the Artifact Replicable badge from their paper evaluation process in the conference year, and instead focus on the two other badges. This Artifact Replicable badge should be an independent work, reviewed, and should be awarded separately and much later in the process by the societies.

An alternative solution would be for conference of year $N + X$ to be in charge of reproducing their work from year N that have had the most influences, for instance the 10% most cited work of year N, with a report appended to the paper discussing the process.

Dropping the Artifact Replicable badge would allow the conference to go back to a smaller reproducibility committee because of the considerably reduced workload. The HPC community could have a journal dedicated to work reproducing (or not reproducing) previous results, which could also generate more scientific discussion [3]. This process could be author-based, i.e., authors submitting their work to be reproduced to obtain the badge, or reviewer-based, i.e., other contributors deciding to reproducing a work. The journal could publish report similar to the one described in the previous section.

5 Post-badge Reproducibility

Badges were a good first step towards encouraging authors and our community to think about reproducibility in HPC. But as the reproducibility in HPC evolves and matures, we suggest that we want to seek a solution that values the importance of all the aspects of reproducibility, including

- The considerable effort of the person reproducing the work;
- Reproducibility of the reproducibility process itself;
- Documenting the reasons for non-reproducibility (e.g., was the work too computationally intensive to be reproduced);
- The unique and differing challenges different reproducers can encounter.

For papers where the main contribution relies strongly on computational results or on a concrete practical implementation, we recommend a process that separates functionality from reproducibility:

- **Functionality of artifacts:** At the time of the acceptance of the work, reproducibility reviewers only check whether artifacts are available and functional. These elements should take a greater importance in the acceptance of the paper, i.e., when artifacts are not functional but the reviewers believe that they should be, this could be grounds for rejection. It also means that this step may need to go back to paper reviewer, or there should be a discussion between artifact reviewers and papers reviewers around which results need to be functional.
- **Reproducibility of artifacts** Editors should encourage the publication of reproducibility reports with at least the following properties:
  - Brevity: Preparing the report itself does not need to take more than one hour, and the relation between the reproduction and the initial work should be clear. The submission could be based on the template we have designed for the reproducibility report;
  - Availability: The accepted (non-)reproduction reports should be paired somehow to the original work, keeping in mind that several groups may want to reproduce the same work and may have different results. Hence we do not recommend for these reproduction reports to be appended to the paper. A solution would be to have a dedicated section on each paper webpage about reproduction. Similarly, paper aggregators such as Scholar[7] could make these reports more visible in dedicated sections when providing results.

Practical implementation of reproducibility journals. There are many important questions to answer when it comes to the practical implementation of journals dedicated to reproducibility reports. For example, who determines who submits or authors the report? Are reproducibility reports peer-reviewed, and if so, what are the peer-review criterion?

In our opinion, these reports, and thus these potential journals, are not evaluating the science or methodology itself because that has already been peer-reviewed through the original publication. We assert that these reproducibility reports should only require minor editing and oversight. If the report makes a claim about reproducibility or non-reproducibility of an artifact, subsequent reproducibility reports will confirm or refute that claim. Thus our recommendation is that these contributions should only receive very light review, i.e., copy-editing. This is also why, whenever using a template whenever possible is recommended.

Another natural question is the relationship with the initial authors of the papers. Should they be able to provide input to the reproducibility report? There are several answers to these. A natural answer is to say that they do not need to. However, it may make it more acceptable to our community to give an opportunity for authors to rebut the reproduction.

[7] https://scholar.archive.org/

To conclude, from a practical perspective there are of course more practical corner cases to solve, but if we want to go forward we want to try and see those corner cases and solve them as they come.

6 Conclusions

Reproducibility badges have long shown their limits. This could have been expected since we have known for a long time that a quantitative, objective-based evaluation alone of science does not work. Science also needs a qualitative evaluation. In addition, reproducibility is much harder than what one may think. Yet, without reproducibility, we are not doing science.

At SC24, we have experimented with a reproducibility report to propose a more qualitative alternative to badges. We have analyzed its limits, mainly its lack of visibility and the fact that it did not solve the limited time that reviewer should be expected to dedicate to reproducing results. Based on these observations, we have made recommendations, including that conferences should not be offering the *Artifact Replicable* badge for papers that have just been accepted, The process for this last badge should be separate, and we should accept as a community that it takes a lot of time, and that maybe it should be done only for a subset of papers.

Acknowledgments

The authors are grateful to Phil Roth (SC24 General chair) and the SC steering committee for giving them the opportunity to try this experiment. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

References

[1] Association for Computing Machinery. 2020. Artifact Review and Badging Version 1.1. https://www.acm.org/publications/policies/artifact-review-and-badging-current Accessed: 2025-03-30.

[2] Alberto Baccini, Giuseppe De Nicolao, and Eugenio Petrovich. 2019. Citation gaming induced by bibliometric evaluation: A country-level comparative analysis. *PLoS One* 14, 9 (2019), e0221212.

[3] Robin Boëzennec, Fanny Dufossé, and Guillaume Pallez. 2024. Qualitatively analyzing optimization objectives in the design of hpc resource manager. *ACM Transactions on Modeling and Performance Evaluation of Computing Systems* 9, 4 (2024), 1–28.

[4] Luce Brotcorne, Anne Canteaut, Aline Carneiro Viana, Céline Grandmont, Benjamin Guedj, Stéphane Huot, Valérie Issarny, Guillaume Pallez, Valérie Perrier, Vivien Quema, Jean-Baptiste Pomet, Xavier Rival, Sylvain Salvati, and Emmanuel Thomé. 2020. *Indicateurs de suivi de l'activité scientifique de l'Inria.* Research Report. Inria. https://inria.hal.science/hal-03033764

[5] Jan Ciesko. 2024. Artifact Evaluation Report of: MIXQ: Taming Dynamic Outliers in Mixed-Precision Quantization by Online Prediction. Reproducibility report@SC24.

[6] Christian Collberg and Todd A. Proebsting. 2016. Repeatability in computer systems research. *Commun. ACM* 59, 3 (Feb. 2016), 62–69. doi:10.1145/2812803

[7] Ludovic Courtès and Ricardo Wurmus. 2015. Reproducible and User-Controlled Software Environments in HPC with Guix. In *Euro-Par 2015: Parallel Processing Workshops - Euro-Par 2015 International Workshops, Vienna, Austria, August 24-25, 2015, Revised Selected Papers (Lecture Notes in Computer Science, Vol. 9523)*, Sascha Hunold, Alexandru Costan, Domingo Giménez, Alexandru Iosup, Laura Ricci, María Engracia Gómez Requena, Vittorio Scarano, Ana Lucia Varbanescu, Stephen L. Scott, Stefan Lankes, Josef Weidendorfer, and Michael Alexander (Eds.). Springer, 579–591. doi:10.1007/978-3-319-27308-2_47

[8] Roberto R. Expósito and Keren Zhou. 2024. Artifact Evaluation Report of: Optimizing quantum fourier transformation (qft) kernels for modern nisq and ft architectures. Reproducibility report@SC24.

[9] Todd Gamblin, Matthew P. LeGendre, Michael R. Collette, Gregory L. Lee, Adam Moody, Bronis R. de Supinski, and Scott Futral. 2015. The Spack package manager: bringing order to HPC software chaos. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis,*

SC 2015, Austin, TX, USA, November 15-20, 2015, Jackie Kern and Jeffrey S. Vetter (Eds.). ACM, 40:1–40:12. doi:10.1145/2807591.2807623

[10] Quentin Guilloteau, Florina Ciorba, Millian Poquet, Dorian Goepp, and Olivier Richard. 2024. Longevity of Artifacts in Leading Parallel and Distributed Systems Conferences: a Review of the State of the Practice in 2023. In *Proceedings of the 2nd ACM Conference on Reproducibility and Replicability*. 121–133.

[11] Torsten Hoefler and Roberto Belli. 2015. Scientific benchmarking of parallel computing systems: twelve ways to tell the masses when reporting performance results. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, Jackie Kern and Jeffrey S. Vetter (Eds.). ACM, 73:1–73:12. doi:10.1145/2807591.2807644

[12] Kenneth Hoste, Jens Timmerman, Andy Georges, and Stijn De Weirdt. 2012. EasyBuild: Building Software with Ease. In *2012 SC Companion: High Performance Computing, Networking Storage and Analysis, Salt Lake City, UT, USA, November 10-16, 2012*. IEEE Computer Society, 572–582. doi:10.1109/SC.COMPANION.2012.81

[13] Sascha Hunold and Alexandra Carpen-Amarie. 2016. Reproducible MPI Benchmarking is Still Not as Easy as You Think. *IEEE Trans. Parallel Distributed Syst.* 27, 12 (2016), 3617–3630. doi:10.1109/TPDS.2016.2539167

[14] Sascha Hunold and Jesper Larsson Träff. 2013. On the State and Importance of Reproducible Experimental Research in Parallel Computing. *CoRR* abs/1308.3648 (2013). doi:10.48550/arxiv.1308.3648 arXiv:1308.3648

[15] Evaristo Jiménez-Contreras, Emilio Delgado López-Cózar, Rafael Ruiz-Pérez, and Víctor M Fernández. 2002. Impact-factor rewards affect Spanish research. *Nature* 417, 6892 (2002), 898–898.

[16] Kate Keahey, Jason Anderson, Mark Powers, and Adam Cooper. 2023. Three Pillars of Practical Reproducibility. In *19th IEEE International Conference on e-Science, e-Science 2023, Limassol, Cyprus, October 9-13, 2023*. IEEE, 1–6. doi:10.1109/E-SCIENCE58273.2023.10254846

[17] Kate Keahey, Joe Mambretti, Paul Ruth, and Dan Stanzione. 2019. Chameleon: A Large-Scale, Deeply Reconfigurable Testbed for Computer Science Research. In *27th IEEE International Conference on Network Protocols, ICNP 2019, Chicago, IL, USA, October 8-10, 2019*. IEEE, 1–2. doi:10.1109/ICNP.2019.8888067

[18] Klaus Kraßnitzer. 2025. AutoAppendix: Towards one-click Reproduction of Computational Artifacts. arXiv:2504.00876 [cs.DC] https://arxiv.org/abs/2504.00876

[19] Malgorzata Lazuka, Andreea Anghel, and Thomas P. Parnell. 2024. LLM-Pilot: Characterize and Optimize Performance of your LLM Inference Services. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis, SC 2024, Atlanta, GA, USA, November 17-22, 2024*. IEEE, 16. doi:10.1109/SC41406.2024.00022

[20] Arnaud Legrand. 2024. Reproducibility Issues and Publication Evolutions (in HPC). Slides available. https://orap.irisa.fr/52ieme-forum-reproductibilite/, Accessed March 2025..

[21] National Information Standards Organization (NISO). 2021. *NISO RP-31-2021, Reproducibility Badging and Definitions*. Technical Report. National Information Standards Organization. doi:10.3789/niso-rp-31-2021

[22] Manish Parashar. 2019. The Reproducibility Initiative. *Computer* 52, 11 (2019), 7–8. doi:10.1109/MC.2019.2935265

[23] Richard Pausch. 2024. Artifact Evaluation Report of: A High-Quality Workflow for Multi-Resolution Scientific Data Reduction and Visualization. Reproducibility report@SC24.

[24] Beth A. Plale, Tanu Malik, and Line C. Pouchard. 2021. Reproducibility Practice in High-Performance Computing: Community Survey Results. *Comput. Sci. Eng.* 23, 5 (2021), 55–60. doi:10.1109/MCSE.2021.3096678

[25] Benjamin Schwaller and Alessio Orsino. 2024. Artifact Evaluation Report of: LLM-Pilot: Characterize and Optimize Performance of your LLM Inference Services. Reproducibility report@SC24.

[26] Arfon M Smith, Kyle E Niemeyer, Daniel S Katz, Lorena A Barba, George Githinji, Melissa Gymrek, Kathryn D Huff, Christopher R Madan, Abigail Cabunoc Mayes, Kevin M Moerman, et al. 2018. Journal of Open Source Software (JOSS): design and first-year review. *PeerJ Computer Science* 4 (2018), e147.

[27] Mengyi Sun, Jainabou Barry Danfa, and Misha Teplitskiy. 2022. Does double-blind peer review reduce bias? Evidence from a top computer science conference. *J. Assoc. Inf. Sci. Technol.* 73, 6 (2022), 811–819. doi:10.1002/ASI.24582

[28] Susan Van Rooyen, Fiona Godlee, Stephen Evans, Nick Black, and Richard Smith. 1999. Effect of open peer review on quality of reviews and on reviewers' recommendations: a randomised trial. *Bmj* 318, 7175 (1999), 23–27.

[29] Andy B. Yoo, Morris A. Jette, and Mark Grondona. 2003. SLURM: Simple Linux Utility for Resource Management. In *Proceedings of the 9th International Workshop on Job Scheduling Strategies for Parallel Processing, (JSSPP) (Lecture Notes in Computer Science, Vol. 2862)*, Dror G. Feitelson, Larry Rudolph, and Uwe Schwiegelshohn (Eds.). Springer, 44–60. doi:10.1007/10968987_3