Artifact Evaluations as Authors and Reviewers: Lessons, Questions, and Frustrations

Quentin Guilloteau*, Millian Poquet†, Jonas H. Müller Korndörfer*, Florina M. Ciorba*

In recent years, we were involved in various reproducibility activities: Artifact Evaluation (AE) committee members (SC24 and EuroSys'25), Artifact authors (Euro-Par 2024[a], TPDS[b]), as well as organizers and attendees of Reproducibility Hackathons. We also surveyed 296 Artifact Descriptions (AD) from papers accepted in leading parallel and distributed systems conferences to gauge their *longevity*[c]. A presentation at the 2024 Community Workshop on Practical Reproducibility in HPC will be a great opportunity to disseminate our varied experiences, *lessons learned*, and make *recommendations*, as well as raise new *questions* to be discussed at the event and addressed in the future by the community.

1 Lessons Learned

1.1 Artifact Creation

Timing of Artifact Creation From our observations, *most authors create their artifact only for the AE submission*. But this should not be a last minute activity. Artifacts should be shareable at any point during the research process. *Reviewers should be critical of `git` repositories containing a single very large commit made right before the Artifact submission*, as there is no evidence of the *provenance* of the data presented in the paper. Moreover, to ensure that the results obtained by the reviewers are the same as the authors', the authors should to provide the cryptographic hash of the produced objects. Yet this approach has limitations in terms of reproducibility: measurement variations, non-determinism of tools (*e.g.,* compression, plotting), etc.

Automation Most of the artifacts we encountered required to either copy-paste commands from the AD appendix, the documentation, or to execute large, complex, and often poorly described bash scripts. *The community would benefit from adopting Workflow Managers* (WMs) to automate the process of engaging with the artifacts. WMs will also support providing information about the *provenance* of the results. However, WMs should be used with care, as they do not focus on reproducibility and their additional complexity can limit the artifact's understandability.

Longevity The AE takes place shortly after the authors created the artifacts. Hence, even if the artifact did not fully capture the required software environment, the mirrors of the package managers will, most likely, be for reviewers in the same or similar state as they were for the authors, thereby masking any reproducibility issues that may arise in the long term. *Authors and reviewers should prioritize* long-term artifact reproducibility.

1.2 No Opportunity for Artifact Resubmission

If the quality of the artifact does not meet the requirements of the venue, then either no/not all reproducibility badges will be awarded, which can be seen as a "*permanent artifact rejection*". Unlike a paper rejection, which authors can revise based on reviewers' feedback and resubmit to the same/other venues, there is no opportunity to revise the artifact of an accepted paper that was initially "rejected". Moreover, as conferences rightly insist on DOIs for artifacts, the accepted papers will forever point to an artifact that has no reproducibility badge. *Would evaluating artifacts* before *the paper submission benefit the quality of the artifacts?* Venues could then require the artifact badges/reports (see §AE Report) along with the paper submission. This would require a trusted process/entity, independent of the venue, that could evaluate the artifacts (*e.g.,* linked to the publisher?).

*University of Basel, DMI, HPC, Basel, Switzerland
†University of Toulouse, IRIT, Toulouse, France

[a]Poquet, M., Carastan-Santos, D., Da Costa, G., Stolf, P., & Trystram, D. "Artifact of the paper: Light-weight prediction for improving energy consumption in HPC platforms", Euro-Par 2024, `https://doi.org/10.5281/zenodo.11208389`

[b]Korndörfer, J. H. M., Eleliemy, A., Mohammed, A., Ciorba, F. M., "LB4OMP: A Dynamic Load Balancing Library for Multithreaded Applications", IEEE TPDS, `https://doi.org/10.1109/TPDS.2021.3107775`

[c]Guilloteau, Q., Ciorba, F.M., Poquet, M., Goepp, D., & Richard, O. "Longevity of Artifacts in Leading Parallel and Distributed Systems Conferences: a Review of the State of the Practice in 2023", ACM-REP 2024, `https://hal.science/hal-04562691`

1.3 Artifact Reviewing Process

Artifact Description and Artifact Evaluation Appendix The SC's efforts to structure the AD and AE appendices forced authors to explicitly provide important information about their artifact. Nevertheless, the requirements do not seem to have been clear to all authors. *SC24's two-stage evaluation process was able to filter out obvious incomplete artifacts* without having to score them. However, this process is not "fail proof" and incomplete artifacts have passed through this filter. EuroSys'25 had no such filtering, and reviewers had to bid on and evaluate artifacts without prior information and completeness verification.

Awarding Badges at All Costs Reproducibility Program Chairs should not try to "push" reviewers to award badges to artifacts. *Should the artifact be rewarded if only one reviewer was able to reproduce the results?*

Artifact Evaluation Report Badges are insufficient to express the nuances of artifact reproduction efforts (see §Awarding Badges at All Costs). SC24 allowed reviewers to write a small report to be included with the proceedings of the reviewed artifact. *AE Reports are an excellent way of informing future researchers about the challenges and pitfalls of using the artifact*. Note that the *longevity* of the artifact's reproducibility should be correctly assessed by the reviewers for this report to be helpful to future researchers in the long term (see §Longevity).

Communication Between Authors and Reviewers *How much communication should there be between authors and reviewers?* In the best case, none, if the artifact is created appropriately. But if a reviewer is blocked at some point in the reproduction process, should they ask the authors for help? *Is the role of the reviewers to debug and improve the authors' artifact, or to simply* evaluate *the artifact as is?* (see Section 1.2).

1.4 Hardware and Experimental Platforms

Authors' Demonstrators Access to the same hardware as the authors can be a challenge. We encountered artifacts where authors gave reviewers access to their own machines on which their artifacts are deployed. While HPC artifacts can be challenging to deploy, these *demonstrator solutions are by no means sufficient proof that the artifact can be successfully redeployed*. The use of *testbeds*, such as Chameleon, would facilitate artifacts evaluation, but is not yet common practice among authors (about 5% for 2023 conferences[c], and about 9% for SC24).

Dependencies to the Experimental Platform These testbeds greatly reduce the hassle of finding the necessary hardware, and may allow the authors' artifact to be easily redeployed. However, using such platforms can create a very tight coupling between the artifact and the platform (*e.g.,* dependence on platform-specific tools). The review process is short and reviewers are often busy. On these shared platforms, the waiting time to access specific resources might be longer than the time allocated to the artifact review process. *If the artifact is highly platform-dependent, this limits the ability of the reviewers to complete the review in a timely manner*. But then the *Replicability* is evaluated, not the *Reproducibility*[d], which could lead to an inconsistent evaluation across all evaluated artifacts.

1.5 Environmental Impact of Artifact Evaluation

HPC experiments often require significant time and *energy*. *Should each paper and all its experiments be reproduced by several reviewers?* If so, what is the added value, especially compared to the total time and energy required? If not, how can the reproducibility of the artifact of the paper be properly addressed? One solution would be for authors to provide minimally viable examples (which is already required by most AE guidelines). However, such minimally viable examples should be *clearly representative* of the full-scale results.

2 Impact and Future Directions

Based on the above, our impression is that *the AE process is rushed*. Authors create an artifact quickly once the paper is submitted. Reviewers try to find enough time in their busy schedules to properly evaluate the complex artifacts that our community produces. Reproducibility Program Chairs should allow enough time for AE, and should not "push" reviewers to award badges at all costs.
We would like to collect the community's opinion on the following questions:

1. *Who should benefit the most from Artifact Evaluations? The authors? Future researchers?*
2. *Should the community move away from the current approach of Artifact Evaluation?*
3. *Are Artifact Evaluations suited to the fast-paced review process of conferences in Computer Science?*
4. *What is the future of Artifact Evaluations in HPC in an energy-aware and energy-constrained world?*

[d] `https://www.acm.org/publications/policies/artifact-review-and-badging-current`