Longevity of Artifacts in Leading Parallel and Distributed Systems Conferences: A Review of The State of the Practice in 2023

Quentin Guilloteau[1], Florina M. Ciorba[1], Millian Poquet[2], Dorian Goepp[3], Olivier Richard[3]

June 19th 2024 - *ACM Conference on Reproducibility and Replicability 2024*

[1]University of Basel, Switzerland
[2]Univ. Toulouse, CNRS, IRIT
[3]Univ. Grenoble Alpes, Inria, CNRS, LIG

DAPHNE

Reproducibility Crisis (in Parallel/Distributed Computing)

Reproducibility Crisis (in Parallel/Distributed Computing)

Q4.2.1 Do you think the state of reproducibility for articles in our research domain (Parallel Computing/HPC) needs to be improved?

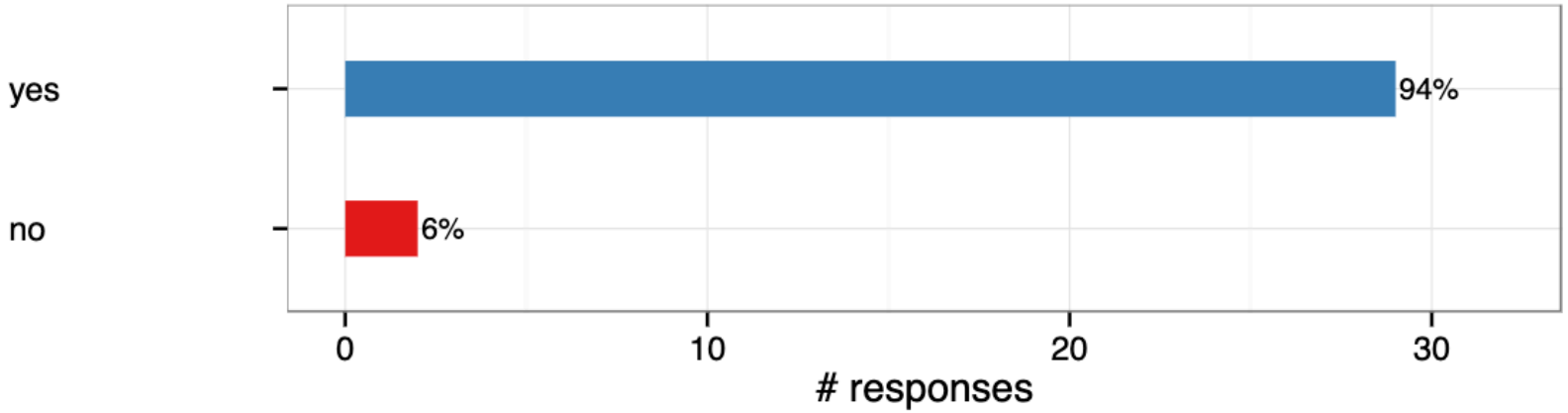


Figure: From Hunold 2015 [1]

Reproducibility Crisis (in Parallel/Distributed Computing)

Q4.3.8 What are the main reasons for NOT making the source code/raw data/data analysis procedure available?

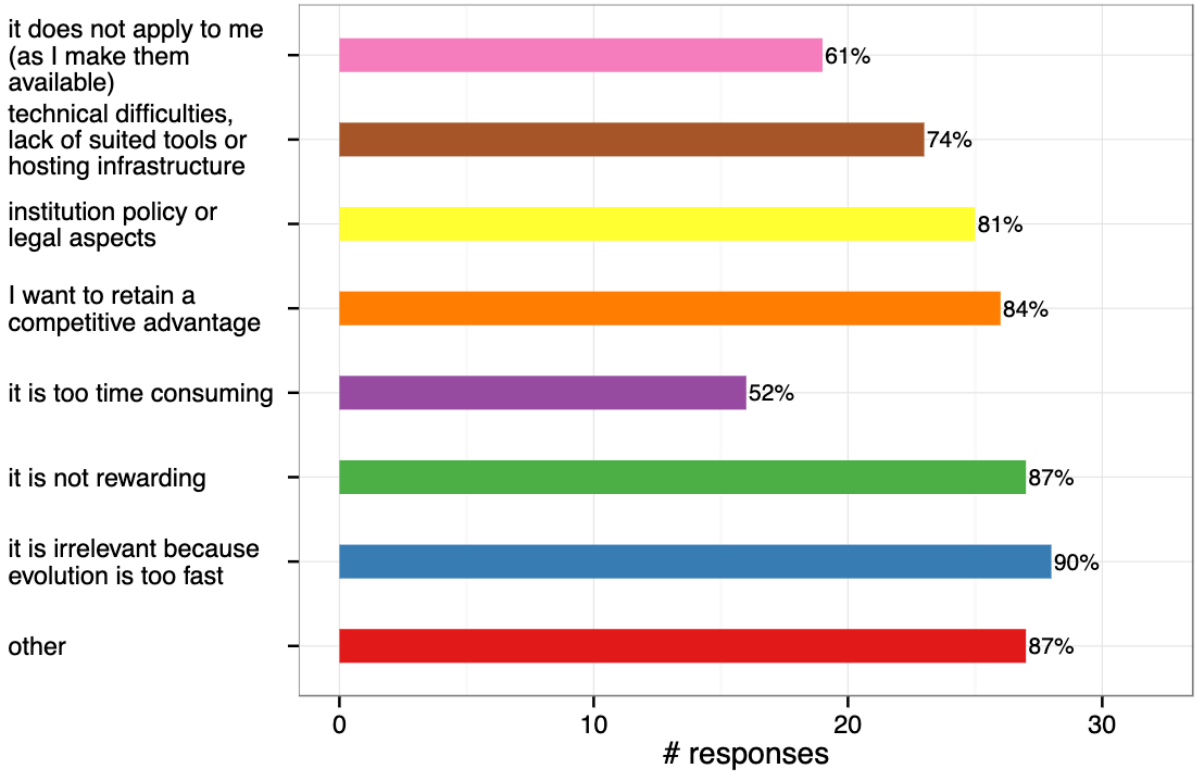


Figure: From Hunold 2015 [1]

Reproducibility Crisis (in Parallel/Distributed Computing)

Q4.3.8 What are the main reasons for NOT making the source code/raw data/data analysis procedure available?

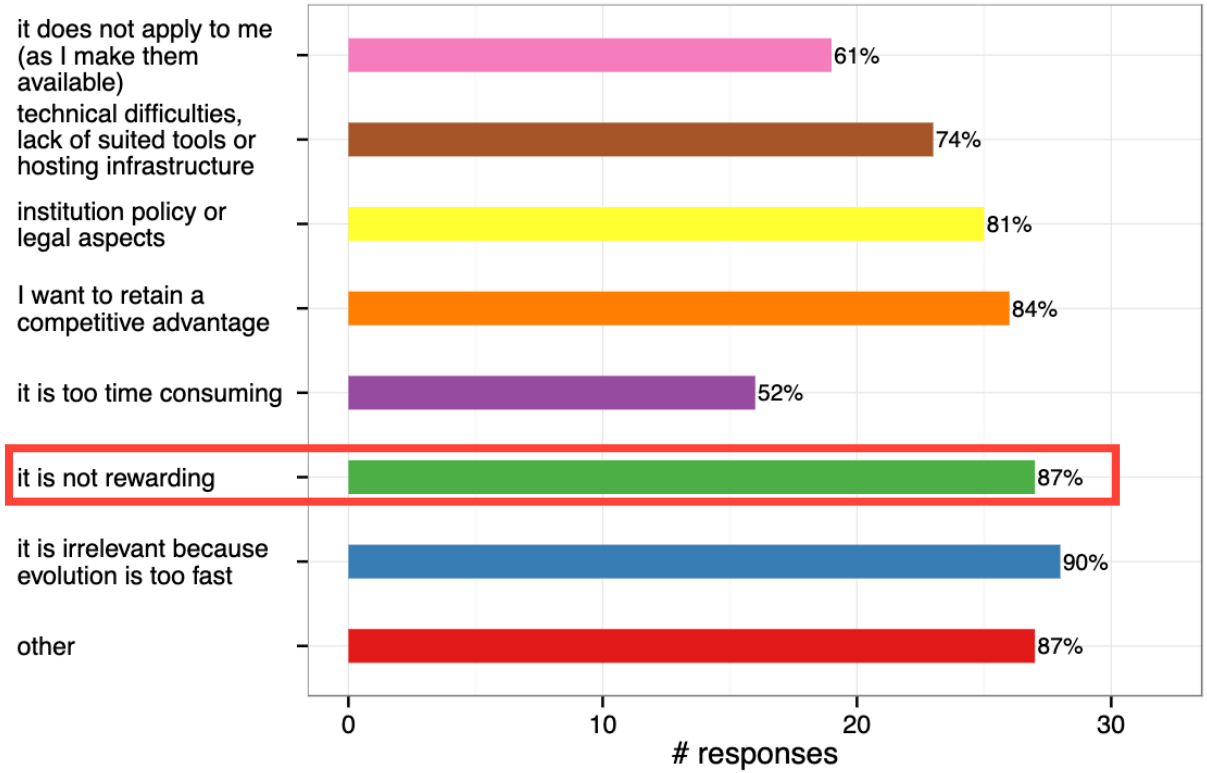
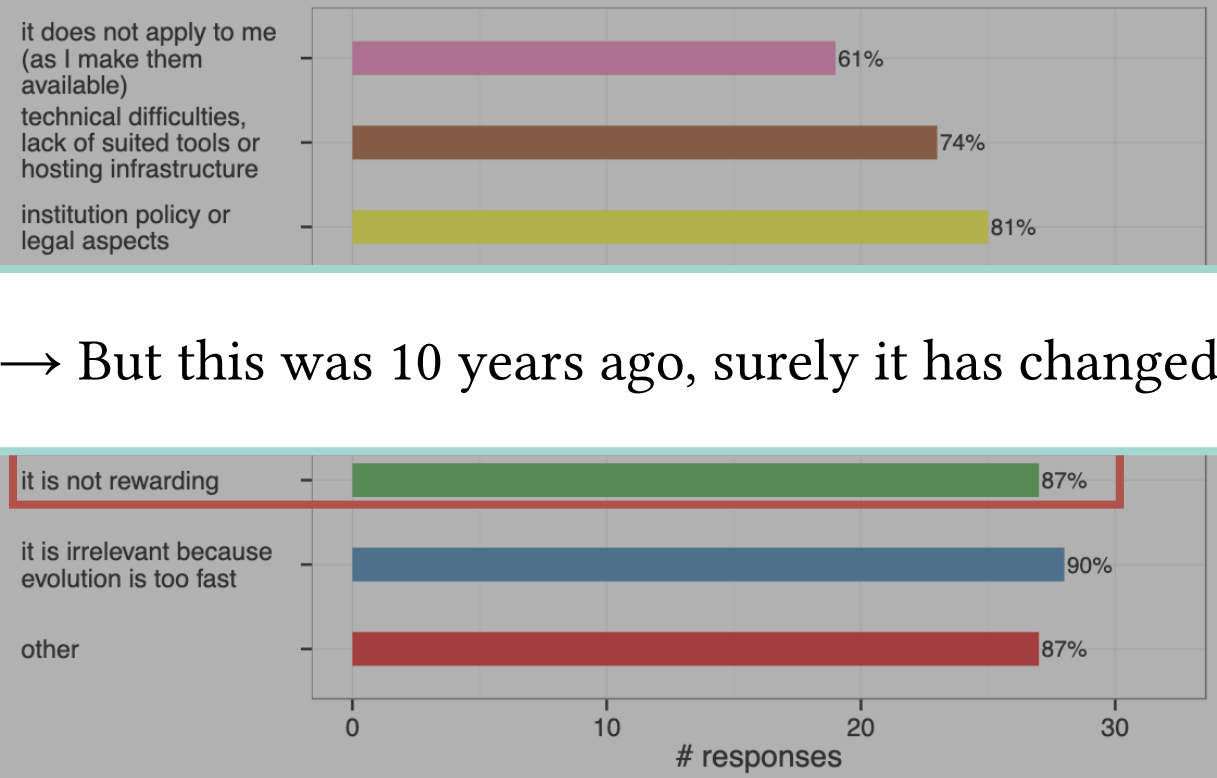


Figure: From Hunold 2015 [1]

Reproducibility Crisis (in Parallel/Distributed Computing)

Q4.3.8 What are the main reasons for NOT making the source code/raw data/data analysis procedure available?



→ But this was 10 years ago, surely it has changed

Figure: From Hunold 2015 [1]

Community Answer: Artifact Description/Evaluation and Badges

- Validate/Promote/Reward
- First: 2011 at the ESC/FSE conference
- In computer science: ACM gave definitions [2]
  - 3 *Repeatability* (Same team, same setup)
  - 2 *Reproducibility* (Different team, same setup)
  - 1 *Replicability* (Different team, different setup)

Artifact Description (AD)

I. Overview of Contributions and Artifacts

A. Paper's Main Contributions

Provide a list of all main contributions of the paper.

C_1 This is the 1st contribution.
C_2 This is the 2nd contribution.
C_3 This is the 3rd contribution.

B. Computational Artifacts

List the computational artifacts related to this paper along with their respective DOIs. Note that all computational artifacts may be archived under a single DOI.

A_1 https://doi.org/YY.YYYY/zenodo.0XXXXX
A_2 https://doi.org/ZZ.YYYY/zenodo.1XXXXX
A_3 https://doi.org/ZZ.YYYY/zenodo.2XXXXX

Provide a table with the relevant computational artifacts, highlight their relation to the contributions (from above) and point to the elements in the paper that are reproducible by each artifact, e.g., which figures or tables were generated with the artifact.

Artifact ID	Contributions Supported	Related Paper Elements
A_1	C_1	Tables 1-2 Figure 3
A_2	C_2	Tables 2-3 Figures 1-2
..		

II. Artifact Identification

Provide the following six subsections for each computational artifact A_i.

A. Computational Artifact A_1

Relation To Contributions

Briefly explain the relationship between the artifact and contributions.

Expected Results

Provide a higher level description of what outcome to expect from the corresponding experiments. Provide an explanation of how the results substantiate the main contributions.

Algorithm A should be faster than Algorithms C and B in all GPU scenarios.

Expected Reproduction Time (in Minutes)

Estimate the time required to reproduce the artifact, providing separate estimates for the individual steps: Artifact Setup, Artifact Execution, and Artifact Analysis.

The expected computational time of this artifact on GPU X is 20 min.

Artifact Setup (incl. Inputs)

Hardware: Specify the hardware requirements and dependencies (e.g., a specific interconnect or GPU type is required).

Software: Introduce all required software packages, including the computational artifact. For each software package, specify the version and provide the URL.

Datasets / Inputs: Describe the datasets required by the artifact. Indicate whether the datasets can be generated, including instructions, or if they are available for download, providing the corresponding URL.

Installation and Deployment: Detail the requirements for compiling, deploying, and executing the experiments, including necessary compilers and their versions.

Artifact Execution

Provide an abstract description of the experiment workflow of the artifact. It is important to identify the main tasks (processes) and how they depend on each other.

A workflow may consist of three tasks: T_1, T_2, and T_3. The task T_1 may generate a specific dataset. This dataset is then used as input by a computational task T_2, and the output of T_2 is processed by another task T_3, which produces the final results (e.g., plots, tables, etc.). State the individual tasks T_i and provide their dependencies, e.g., $T_1 \rightarrow T_2 \rightarrow T_3$.

Provide details on the experimental parameters. How and why were parameters set to a specific value (if relevant for the reproduction of an artifact), e.g., size of dataset, number of data points, input sizes, etc. Additionally, include details on statistical parameters, like the number of repetitions.

Artifact Analysis (incl. Outputs)

B. Computational Artifact A_2

Provide the same type of information as done for Computational Artifact A_1.

Figure: Artifact description template (SC24)

Benefits of the Artifact Evaluation

- Authors of the article? ⟶ Reward, visibility
- Publication venue (Journals/Conferences)? ⟶ Advertisment/Promotion (?)
- Future researchers? ⟶ Easier access to artifact, can audit/reproduce/extend

Benefits of the Artifact Evaluation

- Authors of the article? ⟶ Reward, visibility
- Publication venue (Journals/Conferences)? ⟶ Advertisment/Promotion (?)
- Future researchers? ⟶ Easier access to artifact, can audit/reproduce/extend

Our claim

All of the above, but mainly for future researchers (including oneself)

- Science: self-correcting process, *"standing on the shoulders of giants"*
- This requires **Longevous Artifacts**
- The dream: ✨ precise introduction of Variation ✨

Research Questions (in the context of Parallel/Distributed Computing)

RQ1: What are the current practices in research artifacts?

RQ2: Is the reproducibility of the current practices satisfactory?

→ Let's review of the state of the practice!

Study Design

Study Design

- **Leading** Parallel and Distributed systems conferences
- 5 conferences of 2023 (CORE ranking):
  - CCGrid (A), EuroSys (A), OSDI (A*), PPoPP (A), SC (A)
  - with a Artifact Description (AD) / Artifact Evaluation (AE) process 👏
- 4 dimensions
  - AD and badges (available & reproduced)
  - Artifact availability
  - Software environment
  - Experimental platforms

Study Questions

Study Questions

1. Artifact Badges:

- How many badges?
- Which badges?
- How many AD sections?

Study Questions

1. Artifact Badges:

- How many badges?
- Which badges?
- How many AD sections?

2. Artifact Availability:

- URL available? Valid?
- GitHub, Zenodo, ...?
- Fixed commit hash?

Study Questions

1. Artifact Badges:

- How many badges?
- Which badges?
- How many AD sections?

2. Artifact Availability:

- URL available? Valid?
- GitHub, Zenodo, ...?
- Fixed commit hash?

3. Software Environment:

- How was the software environment described and shared?

Study Questions

1. Artifact Badges:

- How many badges?
- Which badges?
- How many AD sections?

2. Artifact Availability:

- URL available? Valid?
- GitHub, Zenodo, ...?
- Fixed commit hash?

3. Software Environment:

- How was the software environment described and shared?

4. Experimental Platform:

- Which machines/platforms were used?

Observations and Findings

1. Artifact Descriptions and Badges

- 296 papers
- 157 Artifact Descriptions
  - 53% of papers
- 168 artifact links, 154 valid at the time of the study
- 161 “Artifacts Available” Badges
  - 54% of papers, 102% of ADs 🤔
- 101 got the top badge 🥇
  - 34% of papers, 64% of ADs

B.2 Description & Requirements

B.2.1 How to access. Source code, datasets, instructions for building the software, and scripts to run the experiments are available in our git repository: <link-removed>.

Figure: Retracted link

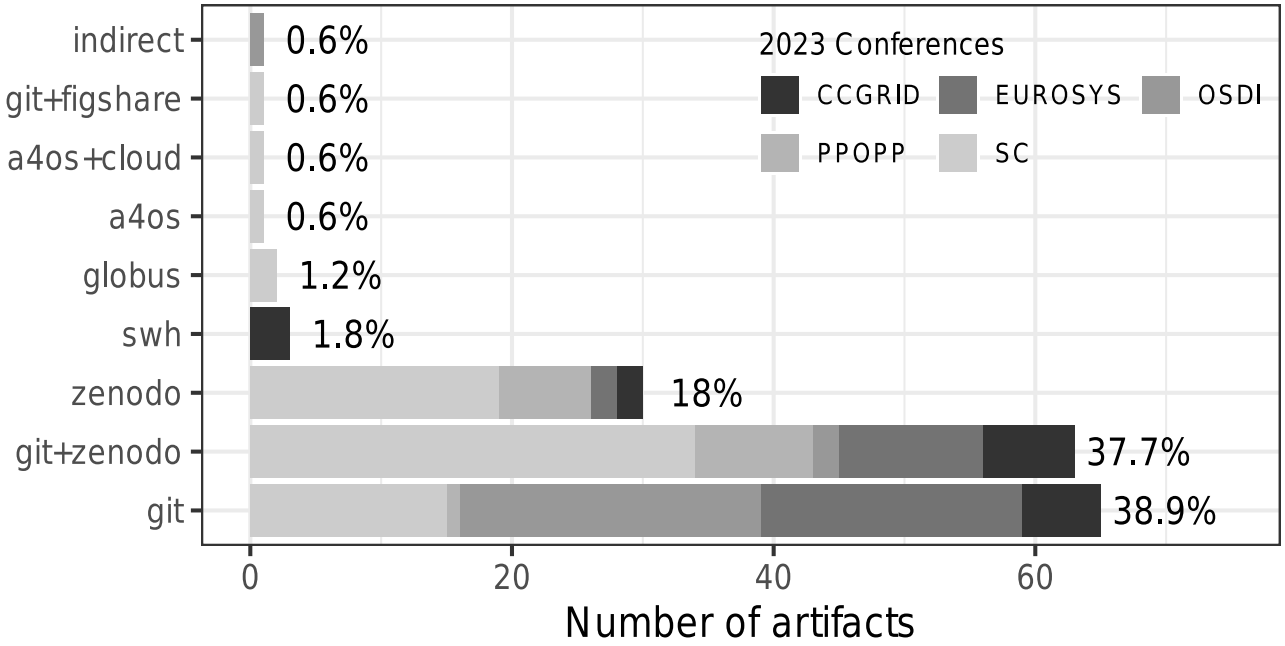
```
[cfs@cfs-client ~]$ df -hT | grep -E "ext4|xfs|nfs" | cut -c 15-55
ext4    387G  108G  279G  28% /home
ext4    485G   70M  460G   1% /mnt/ext4
 xfs    500G   33M  500G   1% /mnt/xfs
nfs4     10P     0   10P   0% /mnt/cfs
[cfs@cfs-client ~]$ cd /mnt/ext4
[cfs@cfs-client ext4]$ sudo prove -rQ /home/cfs/pjdfstest/tests/ -j64
All tests successful.
Files=237, Tests=8861, 80 wallclock secs ( 1.54 usr  0.40 sys + 14.64 cusr 60.20 csys = 76.78 CPU)
Result: PASS
[cfs@cfs-client ext4]$ cd /mnt/xfs
[cfs@cfs-client xfs]$ sudo prove -rQ /home/cfs/pjdfstest/tests/ -j64

Test Summary Report
-------------------
/home/cfs/pjdfstest/tests/symlink/03.t        (Wstat: 0 Tests: 6 Failed: 2)
  Failed tests:  1-2
/home/cfs/pjdfstest/tests/chown/00.t          (Wstat: 0 Tests: 1323 Failed: 0)
  TODO passed:   693, 697, 708-709, 714-715, 729, 733
Files=237, Tests=8832, 81 wallclock secs ( 1.53 usr  0.40 sys + 14.69 cusr 60.19 csys = 76.81 CPU)
Result: FAIL
[cfs@cfs-client xfs]$ cd /mnt/cfs/
[cfs@cfs-client cfs]$ sudo prove -rQ /home/cfs/pjdfstest/tests/ -j64
All tests successful.

Test Summary Report
-------------------
/home/cfs/pjdfstest/tests/chown/00.t          (Wstat: 0 Tests: 1323 Failed: 0)
  TODO passed:   693, 697, 708-709, 714-715, 729, 733, 1097
                 1101, 1107, 1112, 1116, 1122, 1127, 1131
                 1137, 1142, 1146, 1152, 1157, 1161, 1167
                 1172, 1176, 1182, 1187
Files=237, Tests=8832, 123 wallclock secs ( 1.73 usr  0.53 sys + 14.20 cusr 56.32 csys = 72.78 CPU)
Result: PASS
```

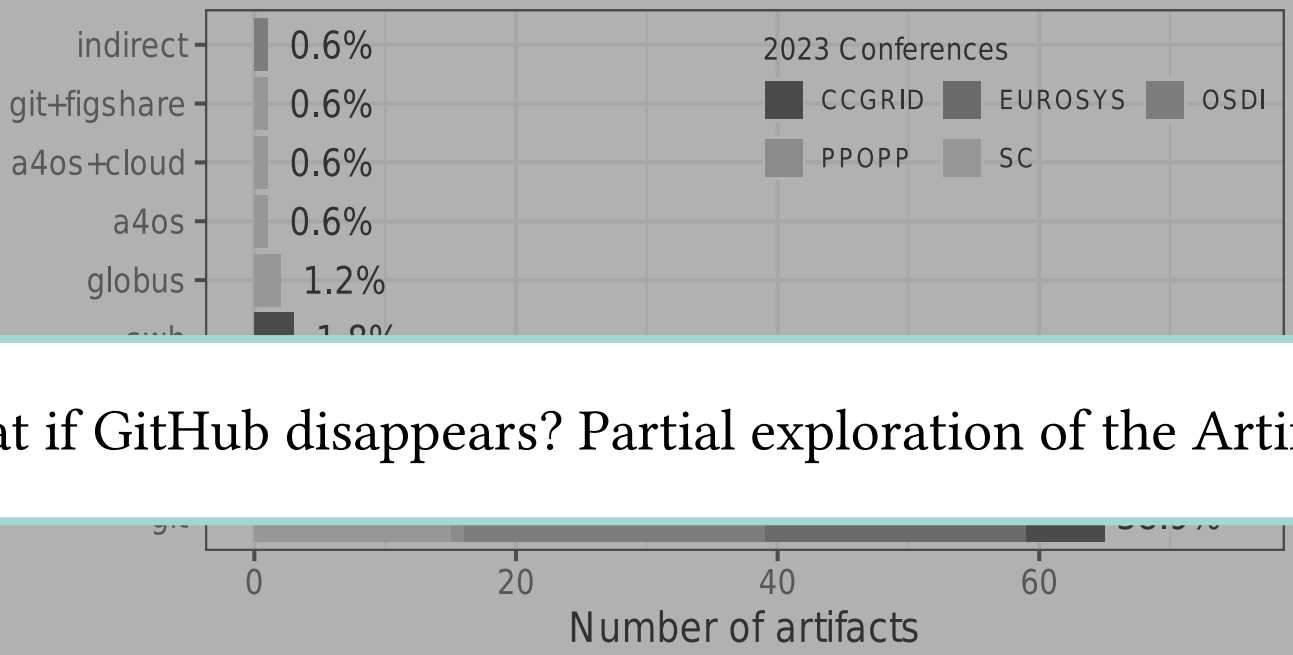
Figure: Screenshot as proof

2. Artifacts Sharing



- mostly a Git(Hub|Lab) URL and/or a Zenodo archive
- when only using **git**, 93% do not report the commit

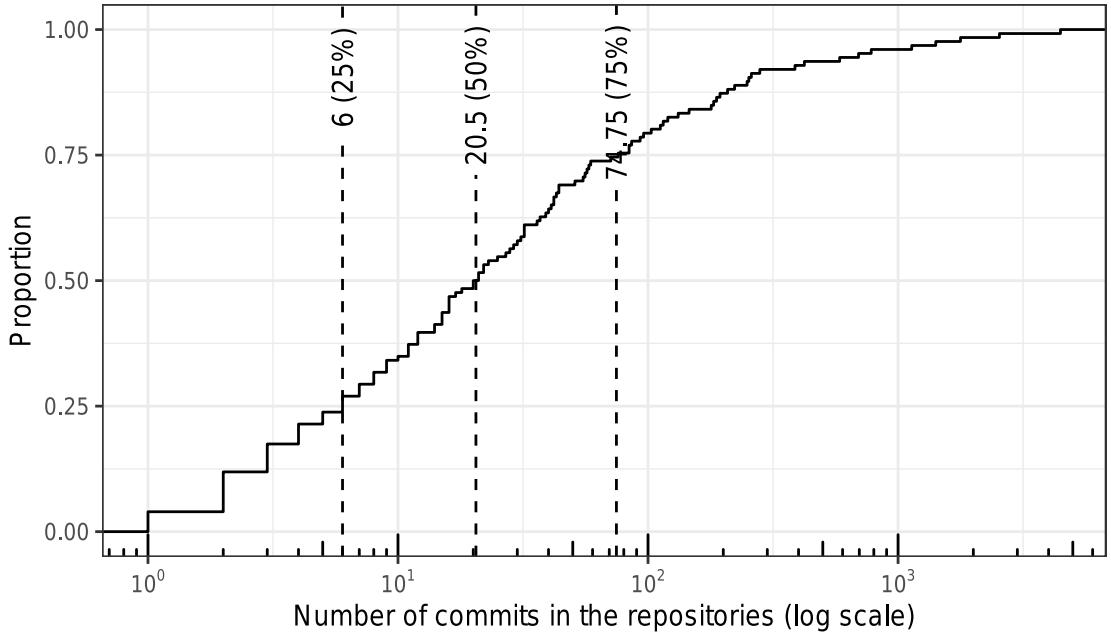
2. Artifacts Sharing



→ What if GitHub disappears? Partial exploration of the Artifacts?

- mostly a Git(Hub|Lab) URL and/or a Zenodo archive
- when only using **git**, 93% do not report the commit

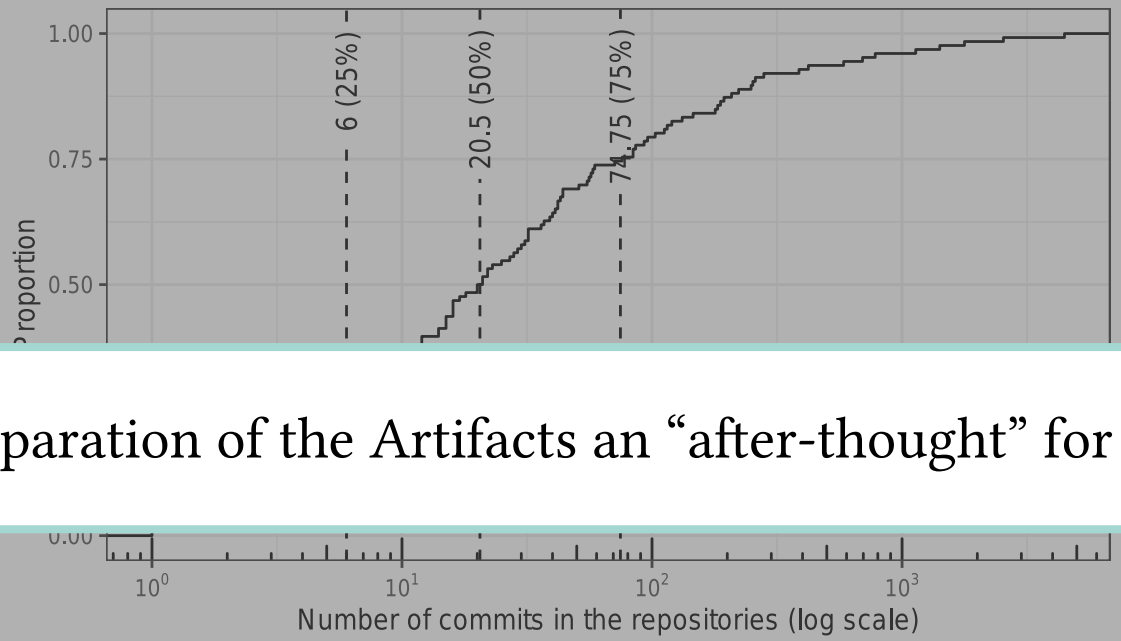
Number of commits in the shared repository



- A lot of repositories are a “dump” of the artifact ⟶ no history / transparency?
- **git** archive in Zenodo?

Showing **15,115 changed files** with **6,755,080 additions** and **1 deletion**. 😬

Number of commits in the shared repository

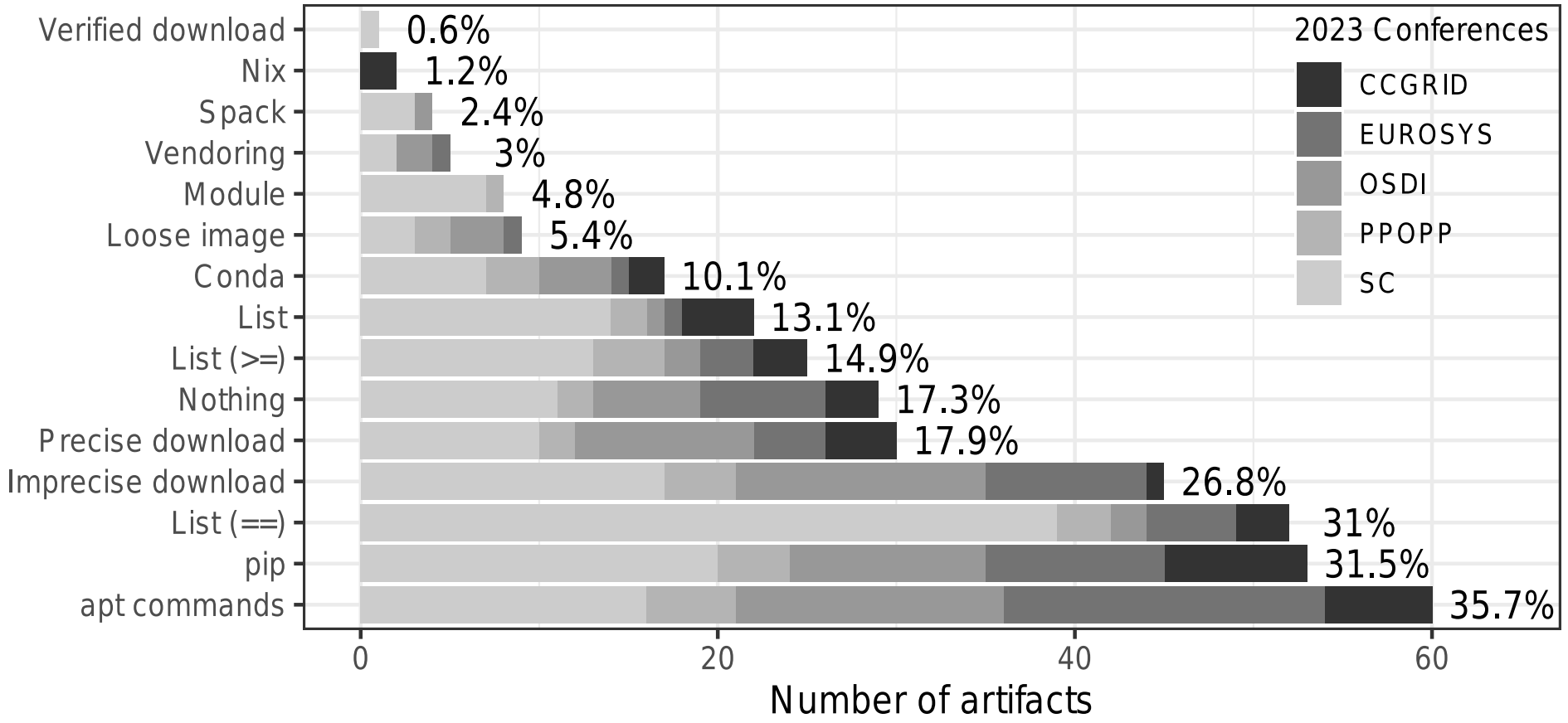


→ Is the preparation of the Artifacts an “after-thought” for the authors?

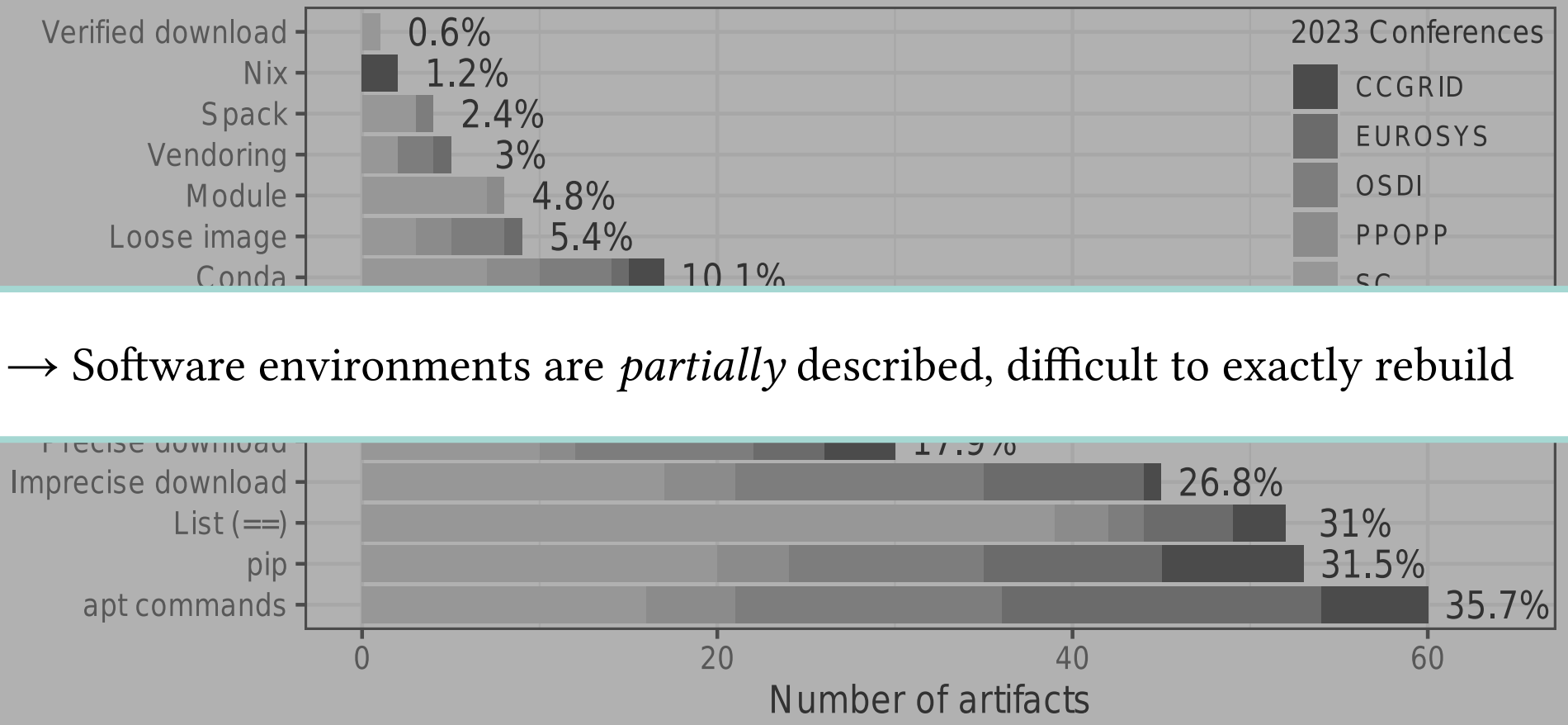
- A lot of repositories are a “dump” of the artifact → no history / transparency?
- **git** archive in Zenodo?

Showing **15,115 changed files** with **6,755,080 additions** and **1 deletion**. 😬

3. How are the software environments captured/described?

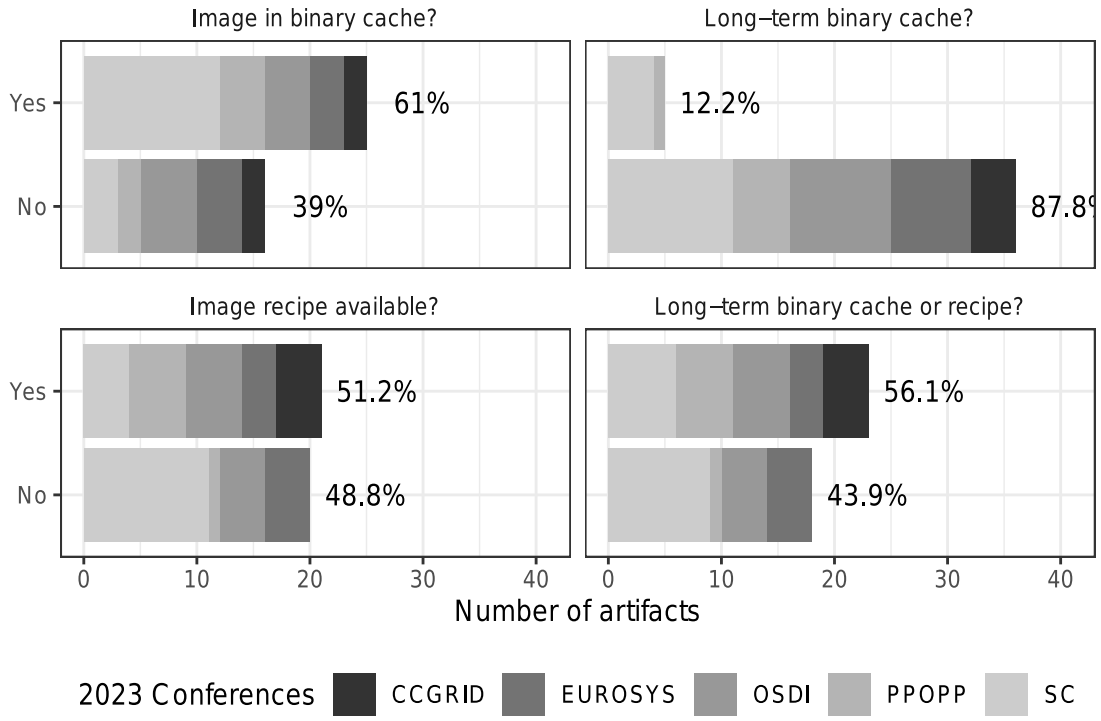


3. How are the software environments captured/described?



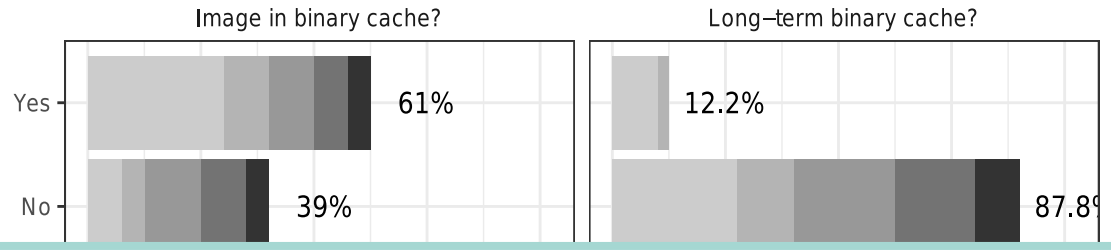
→ Software environments are *partially* described, difficult to exactly rebuild

3. The case of Containers



- Binary cache → e.g., DockerHub; Long-term binary cache → e.g., Zenodo

3. The case of Containers

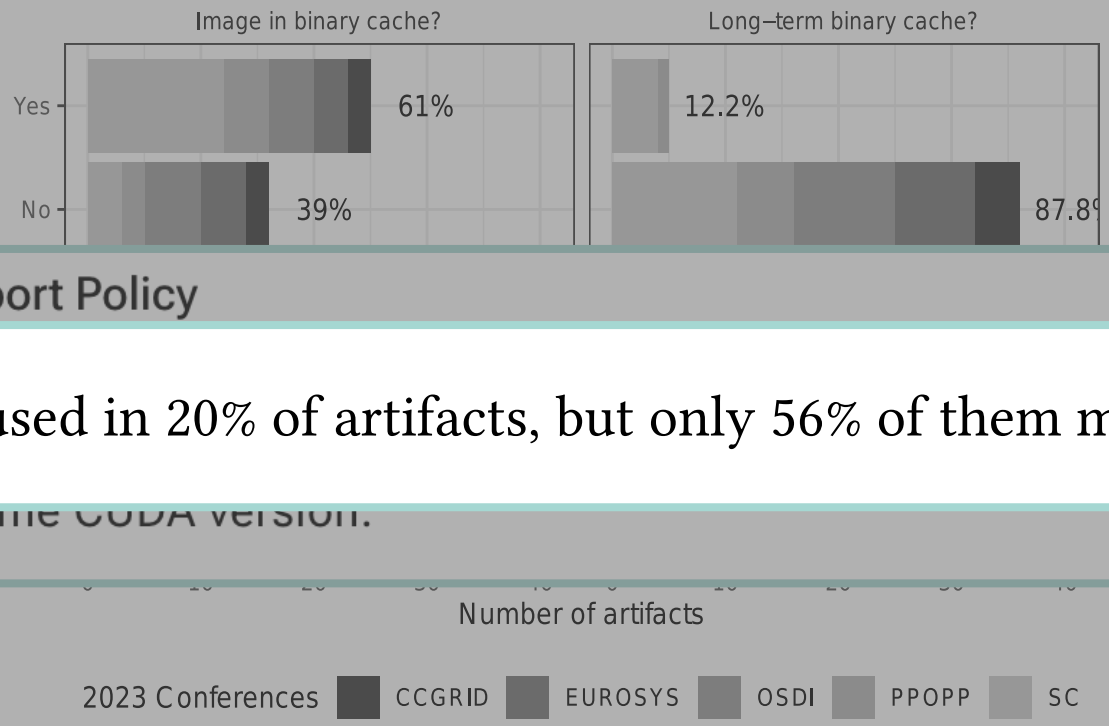


CUDA Container Support Policy

CUDA image container tags have a lifetime The tags will be deleted Six Months after the last supported "Tesla Recommended Driver" has gone end-of-life OR a newer update release has been made for the same CUDA version.

- Binary cache → e.g., DockerHub; Long-term binary cache → e.g., Zenodo

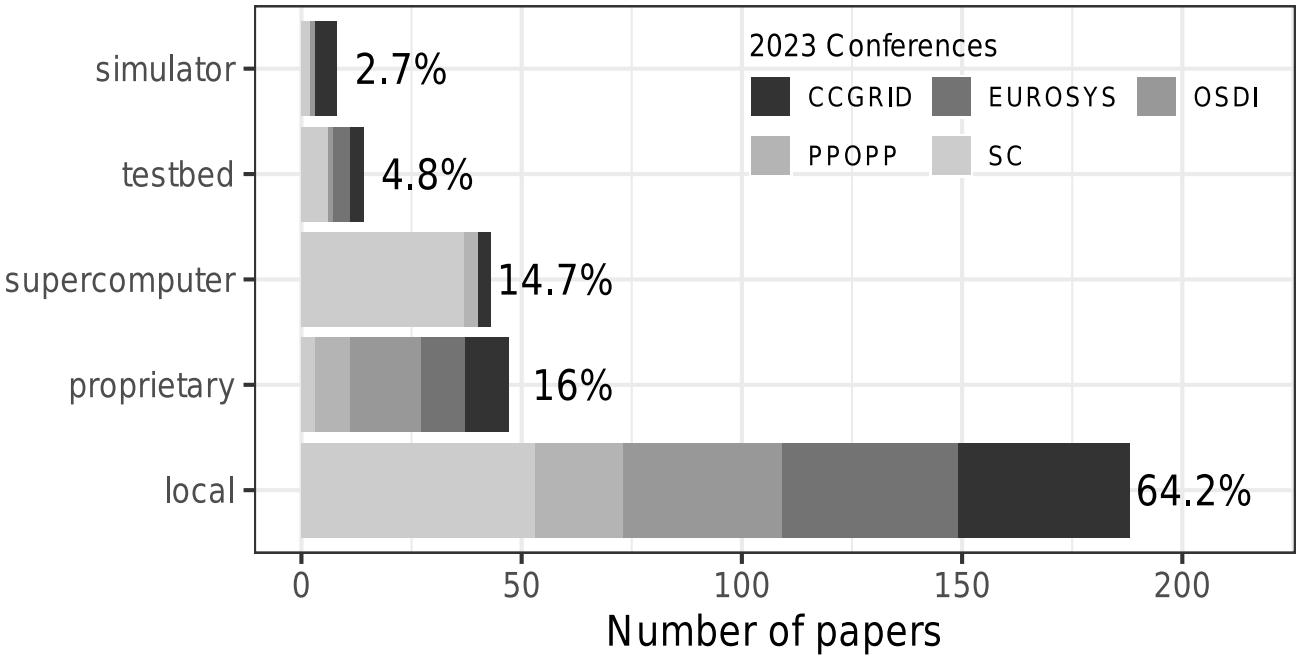
3. The case of Containers



→ Containers are used in 20% of artifacts, but only 56% of them might be reusable…

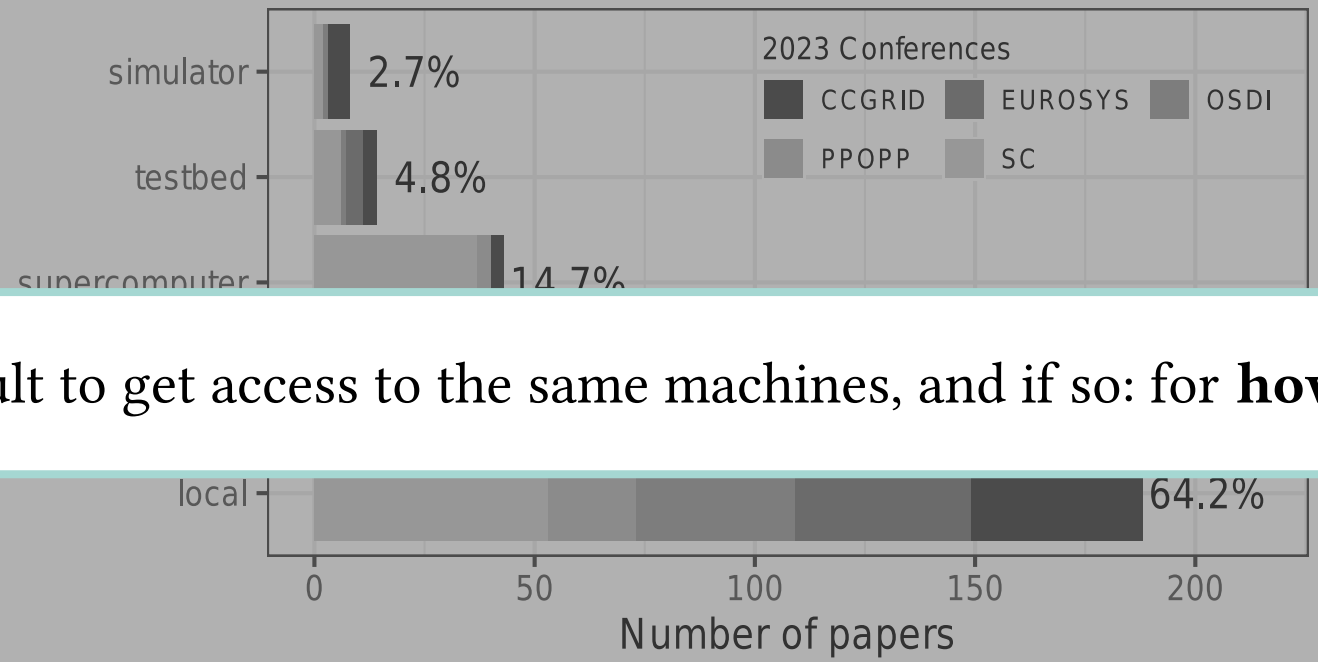
- Binary cache → e.g., DockerHub; Long-term binary cache → e.g., Zenodo

4. Where are the experiments executed?



- How to get access to **those** machines? → Azure/AWS/Google Cloud .... 💸

4. Where are the experiments executed?



→ Difficult to get access to the same machines, and if so: for **how long**?

How to get access to **those** machines? → Azure/AWS/Google Cloud .... 💸

Experiments and Workflow Managers

- **Not part of the study design**
- How is the execution of the experiments managed?
  - Large bash files
  - Copy-pasting commands from the README
- **No usage of Workflow managers**
  - (Snakemake, Nextflow, Luigi, etc.)

Proposal for Artifact Longevity and Recommendations

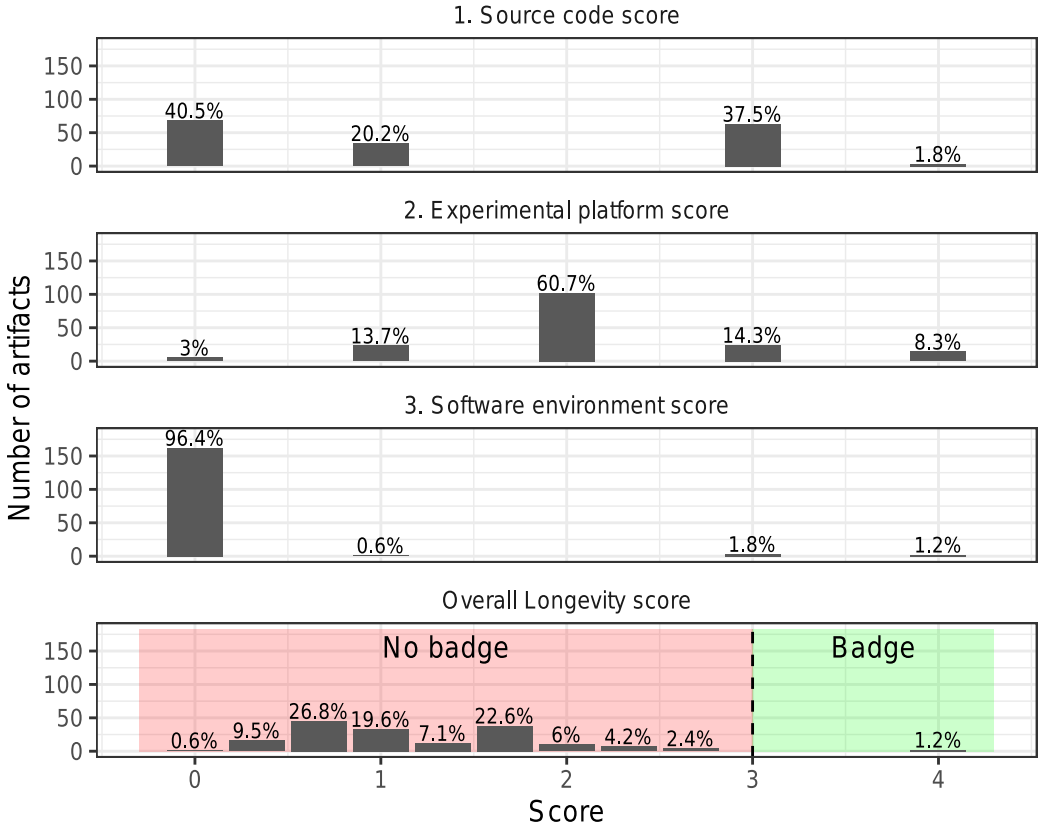
A Needed Badge?: ✨ Artifact Longevity ✨

Do you agree? Let's discuss!

What is the Artifact Longevity (AL) Badge?

- 3 dimensions of AD
  - Artifact availibility
  - Software environment
  - Experimental platform
- 0 to 4-point scale per dimension
- Overall score = avg. per dimension

Overall score ≥ 3 ⟹ Badge awarded



⟶ 2 out of 168 of the reviewed artifacts potentially awarded the *AL* badge

Recommendations to Improve Artifact Longevity

Source code availability [3]

- For source code: Software Heritage
- For data: Zenodo

Software environments

- Functional Package Managers
  - (Nix, Guix)

Experimental platforms

- Shared Testbeds [4]
  - (Grid'5000, Chameleon, CloudLab, etc.)

Conclusion and Perspectives

Conclusion and Perspectives

Conclusion

- AD/AE good for Science, but can be improved!
- State of the practice unsatisfactory ⟶ **Lacks *"Longevity"***
- Proposed a much **needed badge**

Perspectives

- Longitudinal study (from recent past to near term!)
  - **We need your help to re(de)fine the study questions!**
- Is the existing badging system *really* enough?
- Environmental cost of AE?

Take our survey! 😊

Bibliography

[1] S. Hunold, "A survey on reproducibility in parallel computing," *arXiv preprint arXiv:1511.04217*, 2015.

[2] ACM, "Artefact review badging."

[3] P. Alliez *et al.*, "Attributing and referencing (research) software: Best practices and outlook from Inria," *Computing in Science & Engineering*, vol. 22, no. 1, pp. 39–52, 2019.

[4] L. Nussbaum, "Testbeds support for reproducible research," in *Proceedings of the reproducibility workshop*, 2017, pp. 24–26.