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

2024 Community Workshop on Practical Reproducibility in HPC

Quentin Guilloteau[1], Millian Poquet[2], Jonas H. Müller Korndörfer[1], Florina M. Ciorba[1]

2024-11-18

[1]University of Basel, DMI, HPC, Basel, Switzerland
[2]University of Toulouse, IRIT, Toulouse, France

Who are we?

Research Topics

- HPC
- Scheduling (OMP, MPI, RJMS, etc.)
- **Reproducible Research**!

Recent Activities and AE Experiences

- Artifact Reviewers (SC24, EuroSys'25)
- Artifact Authors (Euro-Par24, TPDS'22)
- Attendees and Organizers of Reproducibility Hackathons (`https://www.reprohack.org/`)
- **Study of ADs in HPC (ACM REP24)**

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

Quentin Guilloteau
Florina M. Ciorba
Quentin.Guilloteau@unibas.ch
Florina.Ciorba@unibas.ch
University of Basel
Basel, Switzerland

Millian Poquet
Millian.Poquet@prt.fr
Univ. Toulouse, CNRS, IRIT
Toulouse, France

Dorian Goepp
Olivier Richard
Dorian.Goepp@inria.fr
Olivier.Richard@inria.fr
Univ. Grenoble Alpes, Inria, CNRS, LIG
Grenoble, France

June 18–20, 2024, Rennes, France. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3641525.3663631

ABSTRACT

Reproducibility is the cornerstone of science. Many scientific communities have been struck by the reproducibility crisis, and computer science is no exception. Its answer has been to require artifact evaluations along with accepted articles and award badges to reward authors for their efforts to support reproducibility. Authors voluntarily submit artifacts associated with a submission to reviewers who decide their 'reproducibility' properties. We argue that the notion of 'reproducibility' considered by such badges is limited and misses important aspects of the reproducibility crisis. In this article, we survey almost 300 articles from five leading conferences on parallel and distributed computing held in 2023 (CCGrid, EuroSys, OSDI, PPoPP, and SC). For each article, we gather information about its artifacts (how it was shared, under which experimental setup, and how the software environment was generated and shared), as well as the reproducibility badges awarded. By reviewing the methods and tools used to create and share artifacts in a technical, in-depth, and article-content-agnostic manner, we found that the state of practice does not address reproducibility in terms of artifact *longevity* and we expose eight observations that support this finding. To address the longevity of artifacts, we propose a new badge based on source code, experimental setup, and software environment. These criteria will allow rewarding artifacts expected to withstan test of time. This work aims to shed light on the issue of long reproducibility in parallel and distributed systems and to st discussion in the community towards addressing the issue.

CCS CONCEPTS

• **General and reference** → **Empirical studies**.

KEYWORDS

Reproducibility, Artifact Evaluation, Badges, Longevity

ACM Reference Format:

Quentin Guilloteau, Florina M. Ciorba, Millian Poquet, Dorian G and Olivier Richard. 2024. Longevity of Artifacts in Leading Parallel a Distributed Systems Conferences: a Review of the State of the Practice in In *ACM Conference in Reproducibility and Replicability (ACM PR*

ACM REP '24, June 18–20, 2024, Rennes, France

ACM ISBN 979-8-4007-0539-1/24/06
https://doi.org/10.1145/3641525.3663631

Who are we?

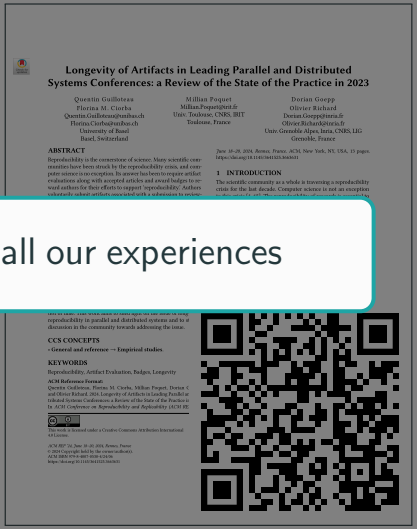
Research Topics

- HPC
- Scheduling (OMP, MPI, RJMS, etc.)
- **Reproducible Research**!

This presentation: Feedback from all our experiences

Rec

- Artifact Reviewers (SC24, EuroSys'25)
- Artifact Authors (Euro-Par24, TPDS'22)
- Attendees and Organizers of Reproducibility Hackathons (`https://www.reprohack.org/`)
- **Study of ADs in HPC (ACM REP24)**



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

Quentin Guilloteau, Florina M. Ciorba, Millian Poquet, Dorian Goepp, Olivier Richard

ABSTRACT

1 INTRODUCTION

CCS CONCEPTS

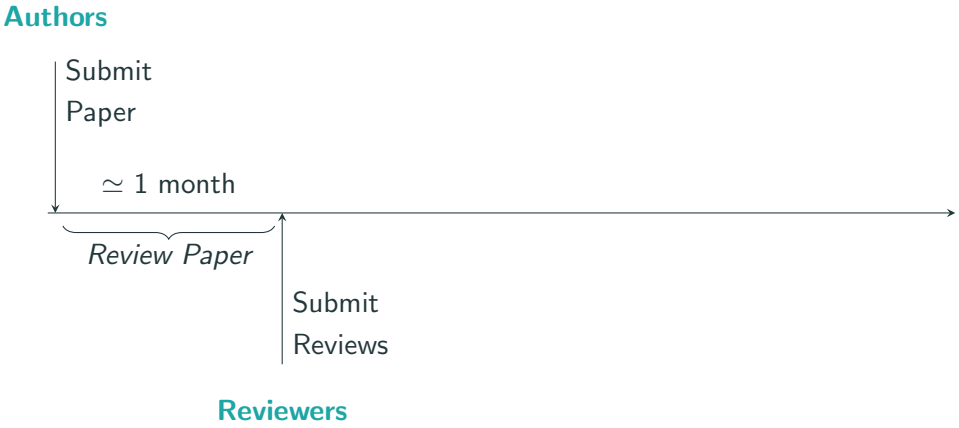
KEYWORDS

The AE Process in a Nutshell

Authors

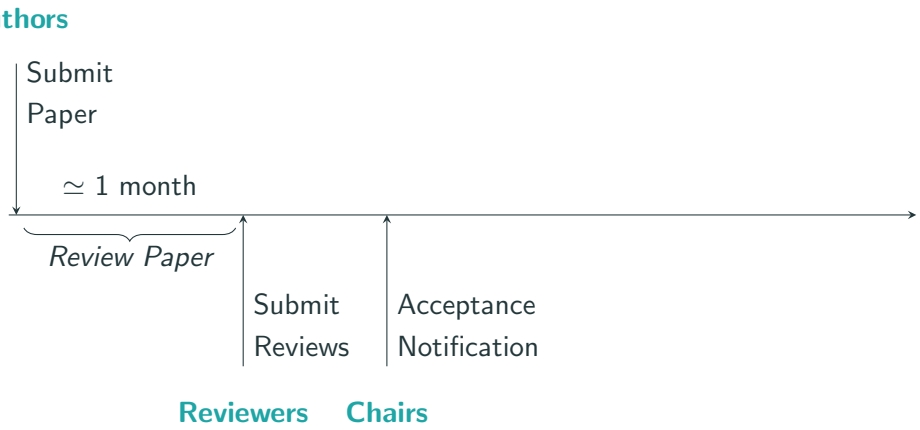
Timeline not to scale!

The AE Process in a Nutshell



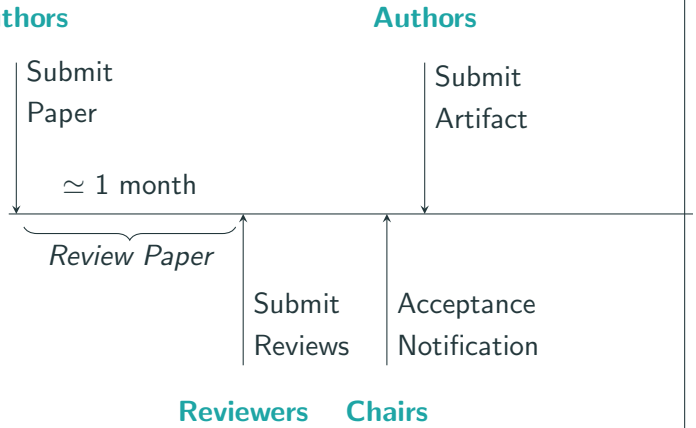
Timeline not to scale!

The AE Process in a Nutshell

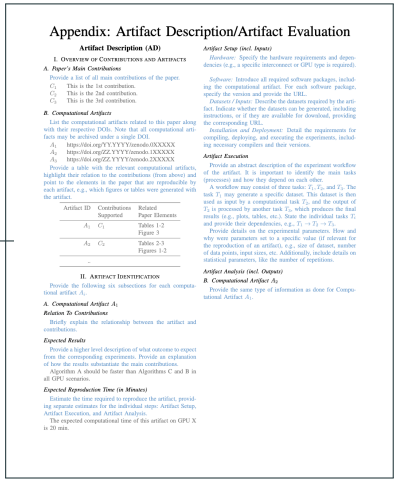


Timeline not to scale!

The AE Process in a Nutshell



Timeline not to scale!



Appendix: Artifact Description/Artifact Evaluation

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/ZZ.YYYY/zenodo.1XXXXXX
A_2 : https://doi.org/ZZ.YYYY/zenodo.1XXXXXX
A_3 : https://doi.org/ZZ.YYYY/zenodo.2XXXXXX

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_i

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., $T_1 \rightarrow T_2 \rightarrow T_3$.

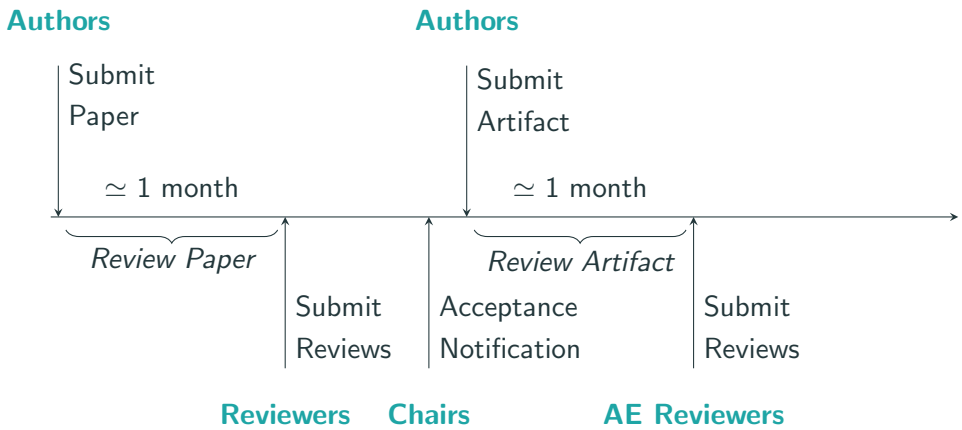
Provide details on the experimental parameters. How and why were parameters set to a specific value (if relevant for the reproduction of an artifact), e.g., size of dataset, number of data points, input sizes, etc. Additionally, include details on statistical parameters, like the number of repetitions.

Artifact Analysis (incl. Outputs)

B. Computational Artifact A_2

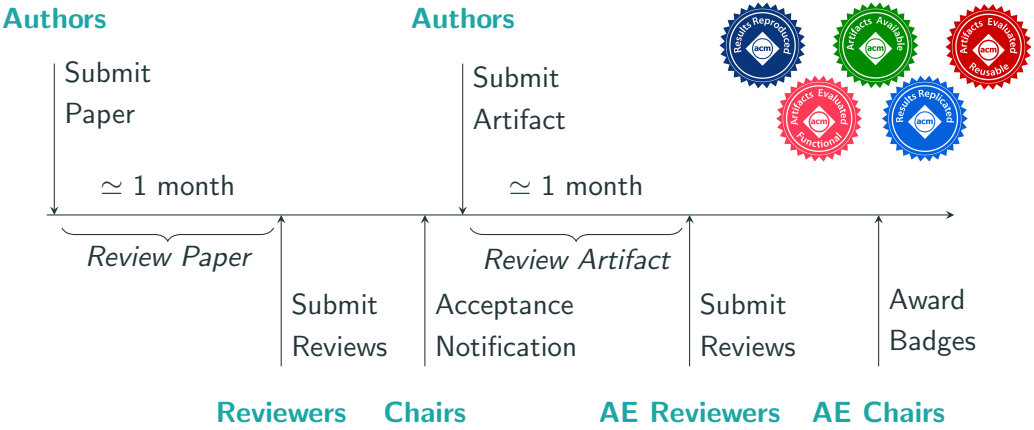
Provide the same type of information as done for Computational Artifact A_1.

The AE Process in a Nutshell



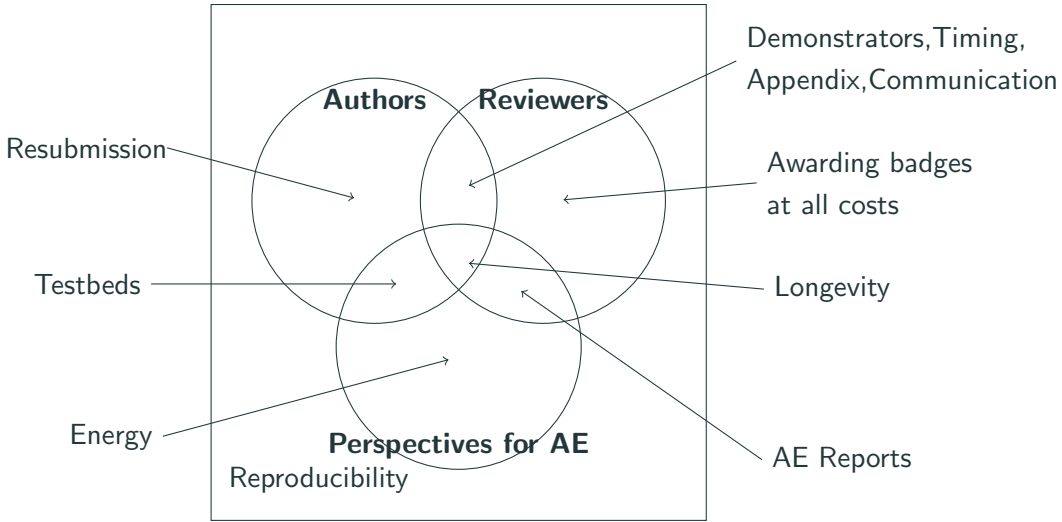
Timeline not to scale!

The AE Process in a Nutshell



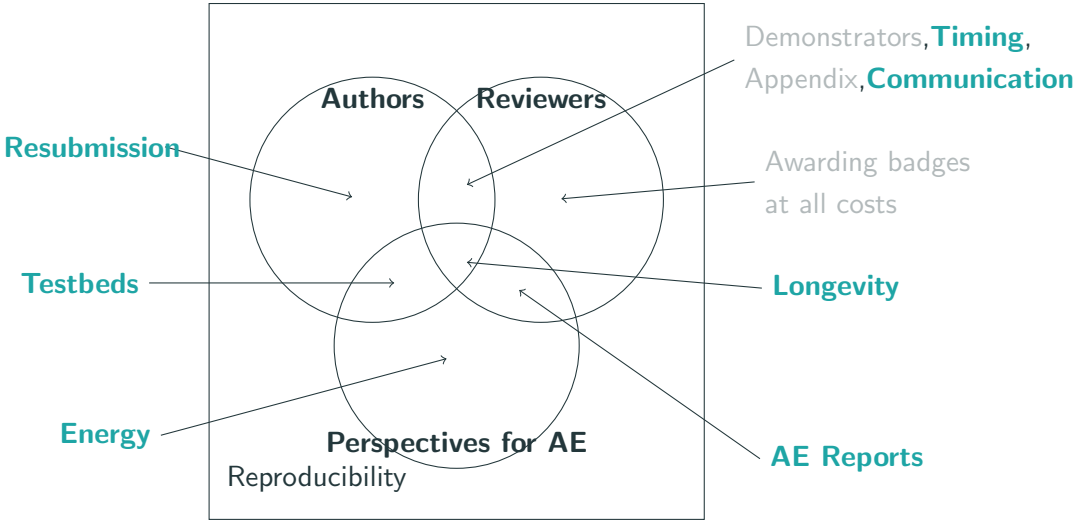
Timeline not to scale!

Topics Addressed in this Presentation



(If you were/are an AE Chair: share your perspective!)

Topics Addressed in this Presentation



(If you were/are an AE Chair: share your perspective!)

Timing, Communication, and Resubmission

Timing of AE

- Artifacts **quickly** created by the authors close to the deadline
- AE Reviewers have a short time to evaluate in a **busy** schedule
- **Available** physical resources might be hard to find during evaluation process

Timing, Communication, and Resubmission

Timing of AE

- Artifacts **quickly** created by the authors close to the deadline
- AE Reviewers have a short time to evaluate in a **busy** schedule
- **Available** physical resources might be hard to find during evaluation process

$\hookrightarrow$ Is the AE process **too short / rushed**?

Timing, Communication, and Resubmission

Timing of AE

$\hookrightarrow$ Is the AE process **too short / rushed**?

Communication Between Authors and AE Reviewers

Timing, Communication, and Resubmission

Timing of AE

↪ Is the AE process **too short / rushed**?

Communication Between Authors and AE Reviewers

↪ What is the role of AE Reviewers: **debug or evaluate** artifacts?

Timing, Communication, and Resubmission

Timing of AE

↪ Is the AE process **too short / rushed**?

Communication Between Authors and AE Reviewers

↪ What is the role of AE Reviewers: **debug or evaluate** artifacts?

Resubmission Not Possible

Timing, Communication, and Resubmission

Timing of AE

↪ Is the AE process **too short / rushed**?

Communication Between Authors and AE Reviewers

↪ What is the role of AE Reviewers: **debug or evaluate** artifacts?

Resubmission Not Possible

↪ If no badge, then "**artifact rejection**" ⇝ no resubmission

Timing of AE

↪ Is the AE process **too short / rushed**?

Communication Between Authors and AE Reviewers

Does the AE process need to be **redesigned**?
Only for journals? Required **before** submission?

Resubmission Not Possible

↪ If no badge, then "**artifact rejection**" ⇝ no resubmission

Longevity

- Science: **Iterative Self-Correcting Process**
  ("standing on the shoulders of giants" – Isaac Newton)

Longevity and AE Reports

Longevity

- Science: **Iterative Self-Correcting Process**
  ("standing on the shoulders of giants" – Isaac Newton)
- Recent Findings:
  - Poor control of artifact sources
    (dead links, commit not fixed/specified)
    $\rightsquigarrow$ **Zenodo, Software-Heritage**

Longevity and AE Reports

Longevity

- Science: **Iterative Self-Correcting Process**
  ("standing on the shoulders of giants" – Isaac Newton)
- Recent Findings:
  - Poor control of artifact sources
    (dead links, commit not fixed/specified)
    $\rightsquigarrow$ **Zenodo, Software-Heritage**
  - Poor control of software environments
    (package list, sometimes versions, `apt`, `pip`, userspace only)
    $\rightsquigarrow$ **Nix(OS)/Guix(System)**

Longevity

- Science: **Iterative Self-Correcting Process**
  ("standing on the shoulders of giants" – Isaac Newton)
- Recent Findings:
  - Poor control of artifact sources
    (dead links, commit not fixed/specified)
    $\rightsquigarrow$ **Zenodo, Software-Heritage**
  - Poor control of software environments
    (package list, sometimes versions, `apt`, `pip`, userspace only)
    $\rightsquigarrow$ **Nix(OS)/Guix(System)**
  - Hardware not easy to access
    $\rightsquigarrow$ **Experimental testbeds**

Longevity

- Science: **Iterative Self-Correcting Process**
  ("standing on the shoulders of giants" – Isaac Newton)
- Recent Findings:
  - Poor control of artifact sources
    (dead links, commit not fixed/specified)
    $\rightsquigarrow$ **Zenodo, Software-Heritage**
  - Poor control of software environments
    (package list, sometimes versions, `apt`, `pip`, userspace only)
    $\rightsquigarrow$ **Nix(OS)/Guix(System)**
  - Hardware not easy to access
    $\rightsquigarrow$ **Experimental testbeds**

New Badge?

Longevity

- Science: **Iterative Self-Correcting Process**
  ("standing on the shoulders of giants" – Isaac Newton)
- Recent Findings:
  - Poor control of artifact sources
    (dead links, commit not fixed/specified)
    $\rightsquigarrow$ **Zenodo, Software-Heritage**
  - Poor control of software environments
    (package list, sometimes versions, `apt`, `pip`, userspace only)
    $\rightsquigarrow$ **Nix(OS)/Guix(System)**
  - Hardware not easy to access
    $\rightsquigarrow$ **Experimental testbeds**

New Badge?

$\hookrightarrow$ **Who is the target** of reproducible research?

Longevity

↪ **Who is the target** of reproducible research?

AE Reports

Longevity and AE Reports

Longevity

$\hookrightarrow$ **Who is the target** of reproducible research?

AE Reports

- Badges $\simeq$ pass/fail
- Review reports might contain complementary information to AD and badges
  $\rightsquigarrow$ **no access** for future researchers
- AE Reports allow for **more details** about the reviewers' reproduction attempts

Longevity and AE Reports

Longevity

$\hookrightarrow$ **Who is the target** of reproducible research?

AE Reports

- Badges $\simeq$ pass/fail
- Review reports might contain complementary information to AD and badges
  $\rightsquigarrow$ **no access** for future researchers
- AE Reports allow for **more details** about the reviewers' reproduction attempts

$\hookrightarrow$ Do badges carry **enough nuance/information**?

Longevity

↪ **Who is the target** of reproducible research?

AE Reports

Who should benefit from AE?
Authors and/or future researchers?

⇝ **no access** for future researchers

- AE Reports allow for **more details** about the reviewers' reproduction attempts

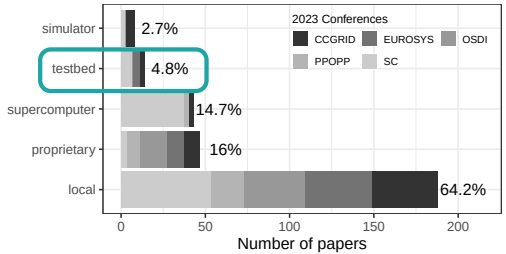
↪ Do badges carry **enough nuance/information**?

Testbeds (Chameleon, CloudLab, Grid'5000, etc.) are **amazing** to access hardware and reduce unknown/variability of deployment.

Testbeds (Chameleon, CloudLab, Grid'5000, etc.) are **amazing** to access hardware and reduce unknown/variability of deployment.

But...

- Waiting time > AE Reviewing time
- Not widely used in practice by authors
- Strong dependency on the testbed?

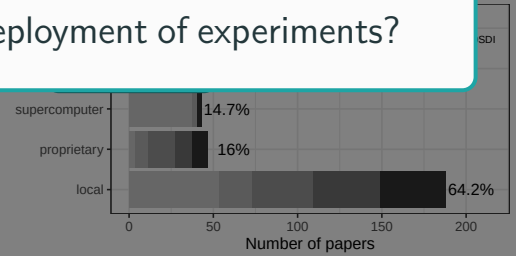


Testbeds (Chameleon, CloudLab, Grid'5000, etc.) are **amazing** to access hardware and reduce unknown/variability of deployment.

How to **"future-proof"** the deployment of experiments?

Bu

- Waiting time > AE Reviewing time
- Not widely used in practice by authors
- Strong dependency on the testbed?



HPC experiments consume **time and energy**

- (and increasingly more of each with AI in HPC conferences)
- Should we store **everything**?
  - why storing the result instead of the recipe?
  - $\hookrightarrow$ need reproducible/deterministic ways to produce research objects
  - Nix/Guix? but might recompile *a lot* $\rightsquigarrow$ Sustainable?
- Minimal viable example/experiment: but must be **representative** of the study
- How to create a valuable minimal viable example/experiment?
- How to reward a partially evaluated artifact?
- Is it worth to ask **several reviewers** to try to reproduce all or part of the study?

HPC experiments consume **time and energy**

- (and increasingly more of each with AI in HPC conferences)
- Should we store **everything**?

When do the **time and energy costs outweigh the value** of what is reproduced?

- Minimal viable example/experiment: but must be **representative** of the study
- How to create a valuable minimal viable example/experiment?
- How to reward a partially evaluated artifact?
- Is it worth to ask **several reviewers** to try to reproduce all or part of the study?

Is the Artifact Evaluation process (creation + evaluation) **rushed**?

Is **Artifact Evaluation** the path to **Reproducibility** in HPC?

Who is the target of reproducible research?

What is the future of AE in HPC in an **energy-constrained** world?