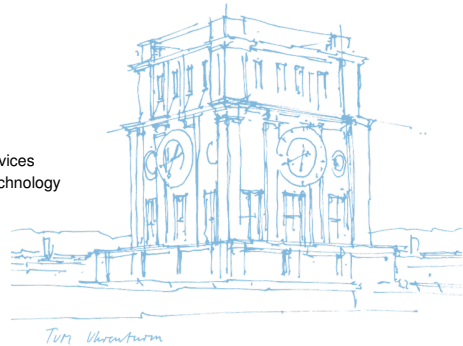
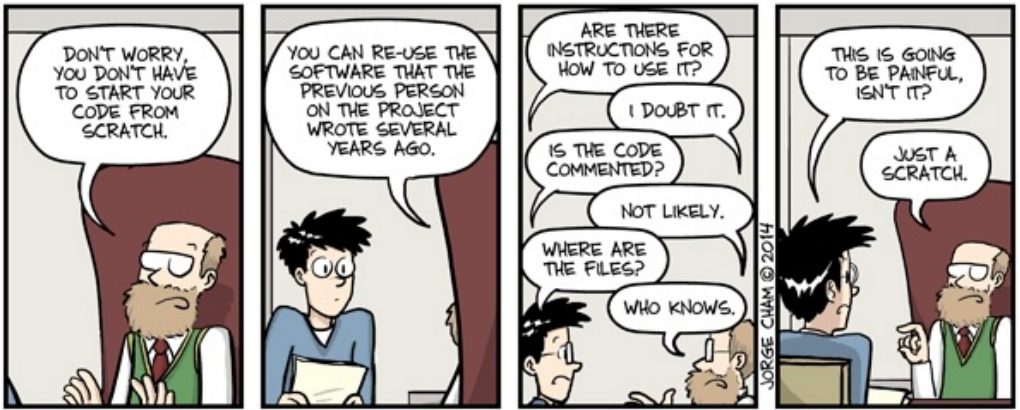
Reproducible Experiments, SIGCOMM and CoNEXT Artifact Evaluation and Infrastructure Needs

Sebastian Gallenmüller

Chair of Network Architectures and Services
School of Computation, Information and Technology
Technical University of Munich



Reproducibility in Computer Science



[1]

[1] Jorge Cham, https://phdcomics.com/comics/archive.php?comicid=1689, Last accessed: Nov. 26, 2023

Artifact Evaluation

Goals

- Raising awareness for reproducibility in our community
- Rewarding papers and authors that make their research reproducible

Means

- ACM definition for reproducible research:
  1. **Repeatability:** Same team executes experiment using same setup
  2. **Reproducibility:** Different team executes experiment using same setup
  3. **Replicability:** Different team executes experiment using different setup
- In 2015, ACM started its initiative to introduce badges to reflect different qualities of reproducibility

ACM reproducibility badge

Artifact Evaluation — The Process

Preconditions for artifact evaluation

- Conference offers artifact evaluation (AE)
- Enough people willing & capable to evaluate the artifacts

Typical process of AE

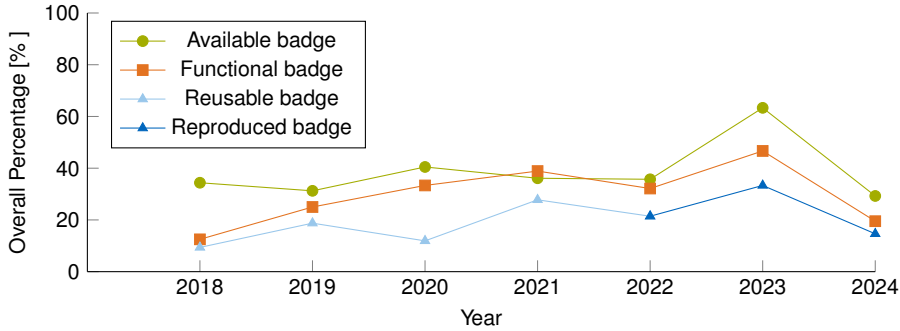
- Authors of papers are invited to AE after paper has been accepted
- Authors prepare AE of their paper for evaluation (typically a few weeks after acceptance)
- Artifact evaluators evaluate artifacts (approx. 4 weeks)
  - Authors and artifact evaluators communicate regularly & anonymously
  - Authors improve artifacts during process
- Papers are awarded badges based on AE reviews

Goal: AE improves/ensures the quality of artifacts

Artifact Evaluation — The Badges

- **Available:** Relevant artifacts of the paper are publicly available
- **Functional:** Artifacts are documented, consistent, complete, and exercisable
- **Reusable:** Artifacts have quality that exceeds minimal functionality
- **Reproduced:** Main results of paper have been independently obtained by subsequent study by persons other than the authors, using, in part, author-supplied artifacts
- **Replicated:** Main results of paper have been independently obtained by subsequent study by persons other than the authors, without author-supplied artifacts

Development of the CoNEXT Artifact Evaluation over Time



- CoNEXT is one of the leading conferences in information and communications technology (ICT) research
- Graph shows the percentage of papers that received one of the respective badges
- Adoption stagnates over the years

Results of Artifact Evaluation: CoNEXT '23

Artifacts

- papers accepted: 30
- papers that handed in artifacts: 19 (63%)

Awards applied	19	17	17
Awards awarded	18	14	10
Overall award rate	60 %	47 %	33 %

Analysis of AE — Requirements

Examples of hardware requirements for reviews

- 3 × artifacts require Nvidia GPUs
- 3 × artifacts require Intel Tofino switch(es)
- 1 × artifact requires Intel SGX-capable CPUs
- RAM requirements:
  - Most demanding artifact required 512 GB in one machine
  - Another artifact requires several machines with at least 64 GB
- SIGCOMM'23: Large AWS instance (>1000 USD costs for reviewing)

Prices for Nvidia GPUs (June 2025)

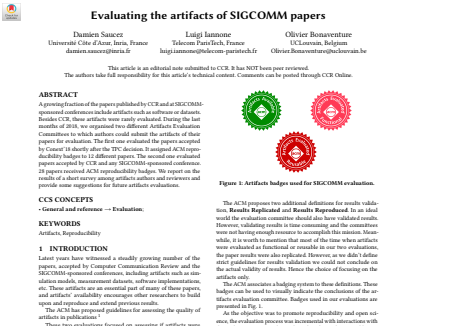
Analysis of AE — Benefits & Challenges

Observations

- AE participation stagnates
- Hardware requirements may prevent effective reproduction

Survey[2] among authors and artifact evaluators

- Main message:
  - \+ AE is useful and interesting
  - \- AE is time consuming for authors and evaluators



Evaluating the artifacts of SIGCOMM papers

Damien Saucez, Université Côte d'Azur, Inria, France, damien.saucez@inria.fr
Luigi Iannone, Telecom ParisTech, France, luigi.iannone@telecom-paristech.fr
Olivier Bonaventure, UCLouvain, Belgium, Olivier.Bonaventure@uclouvain.be

This article is an editorial note submitted to CCR. It has NOT been peer reviewed.
The authors take full responsibility for this article's technical content. Comments can be posted through CCR Online.

ABSTRACT

A growing fraction of the papers published by CCR and at SIGCOMM-sponsored conferences include artifacts such as software or datasets. Besides CCR, these artifacts were rarely evaluated. During the last months of 2018, we organized two different Artifacts Evaluation Committees to which authors could submit the artifacts of their papers for evaluation. The first one evaluated the papers accepted by Conext'18 shortly after the TPC decision. It assigned ACM reproducibility badges to 12 different papers. The second one evaluated papers accepted by CCR and any SIGCOMM-sponsored conference. 28 papers received ACM reproducibility badges. We report on the results of a short survey among artifacts authors and reviewers and provide some suggestions for future artifacts evaluations.

CCS CONCEPTS

• General and reference → Evaluation;

KEYWORDS

Artifacts, Reproducibility

1 INTRODUCTION

Latest years have witnessed a steadily growing number of the papers, accepted by Computer Communication Review and the SIGCOMM-sponsored conferences, including artifacts such as simulation models, measurement datasets, software implementations, etc. These artifacts are an essential part of many of these papers, and artifacts' availability encourages other researchers to build upon and reproduce and extend previous results.

The ACM has proposed guidelines for assessing the quality of artifacts in publications.[1]

These two evaluations focused on assessing if artifacts were **available**, **functional**, or **reusable**, which definitions are given by the ACM as follows.

- **Artifacts Available**: author-created artifacts relevant to this paper have been placed on a publicly accessible archival repository.
- **Artifacts Evaluated - Functional**: the artifacts associated with the research are found to be documented, consistent, complete, exercisable, and include appropriate evidence of verification and validation.
- **Artifacts Evaluated - Reusable**: the artifacts associated with the paper are of a quality that significantly exceeds minimal functionality.

[1] https://www.acm.org/publications/policies/artifact-review-badging

Figure 1: Artifacts badges used for SIGCOMM evaluation.

The ACM proposes two additional definitions for results validation, **Results Replicated** and **Results Reproduced**. In an ideal world the evaluation committee should also have validated results. However, validating results is time consuming and the committees were not having enough resource to accomplish this mission. Meanwhile, it is worth to mention that most of the time when artifacts were evaluated as functional or reusable in our two evaluations, the paper results were also replicated. However, as we didn't define strict guidelines for results validation we could not conclude on the actual validity of results. Hence the choice of focusing on the artifacts only.

The ACM associates a badging system to these definitions. These badges can be used to visually indicate the conclusions of the artifacts evaluation committee. Badges used in our evaluations are presented in Fig. 1.

As the objective was to promote reproducibility and open science, the evaluation process was incremental with interactions with the authors to improve the quality of artifacts when possible. For that reason, the artifacts study was optional and authors had to expressly apply in order to have their artifact evaluated. Therefore, the absence of badge on a 2018 SIGCOMM-sponsored venue paper doesn't indicate a lack of reproducibility of a paper.

2 CONEXT'18 ARTIFACTS EVALUATION RESULTS

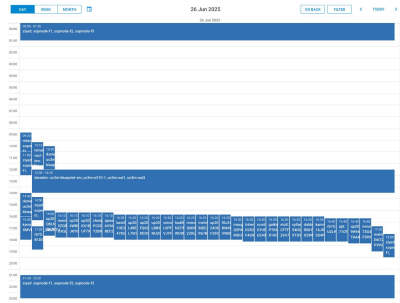
The evaluation of CoNEXT'18 papers' artifacts was carried out shortly after the acceptance notification. Out of 14 accepted papers proposing an artifact, 12 have been awarded a badge. Seven of them received the *Artifacts Available* badge.

- DeepTLC: A CoS-Free Massive MIMO System with Distributed LEDs [6]

[2] Damien Saucez, Luigi Iannone, Olivier Bonaventure: Evaluating the artifacts of SIGCOMM papers. Comput. Commun. Rev. 49(2): 44-47 (2019)

Analysis of AE — SLICES Testbeds to the Rescue

- Testbeds can provide access to a diverse set of hardware
  - Authors and artifact evaluators access the same platform
  - Same hardware and software through a shared access
- Limiting the effort for authors and reviewers
  - Shared platform accelerates and simplifies debugging of experimental code for authors and evaluators
  - Long-term availability of infrastructure ensured through the long-term ESFRI funding scheme of SLICES
- A framework enabling reproducibility by design:
  - Reproducibility through a structured experiment workflow
  - SLICES/pos framework[3] ensures its reproducibility



Calendar of the SLICES-RI (June 26, 2025)

[3] Sebastian Gallenmüller, Dominik Scholz, Henning Stubbe, Georg Carle: The pos framework: a methodology and toolchain for reproducible network experiments. CoNEXT 2021: 259-266

Conclusion — A (Subjectively) Ideal AE Process

- Enough time for artifact evaluation
  - A minimum of 2–3 weeks between paper & artifact submission
  - Artifacts available badges ready at the time of the conference
  - Further AE badging after the conference
- Conferences suggest and incentivize the use of testbeds:
  - Authors and reviewers have a common reference environment provided by the testbeds to run experiments
  - Testbeds will provide long-term availability of environment to run artifacts
- Testbeds can be easily accessed