Message from the Artifact Evaluation Co-Chairs

Dear systems community:

In systems research, artifacts play an important role since the results are often tied to the produced artifact. This not only includes software systems but also datasets, benchmarks, models and test suites. In many cases, it is impossible to reproduce the results without the artifact. Yet, as a community, we offered no formal means to submit and evaluate anything but the paper. This year, we took a small step in this direction. In this letter, we share our experiences of organizing the first artifact evaluation (AE) process at SOSP 2019.

As this was the inaugural year, artifact evaluation was limited only to accepted papers. We were encouraged to see that 23 of the 38 accpeted papers (i.e., 61%) applied to participate. We reached out to broader systems community via Twitter and Slack (systems-research.slack.com) to put together an artifact evaluation committee (AEC). This helped us bring together a team of 42 researchers, postdocs, and graduate students hailing from 11 different countries, who volunteered to read papers and evaluate the systems. We decided on three badges that made sense for systems research (from out of the six that ACM has proposed): *Artifact Available*, when the artifact is made available for retrieval, publicly and permanently; *Artifact Functional*, when the artifact has gone through an independent audit, and it functions as described; *Results Replicated*, when the main results of the paper are obtained independently using the supplied artifacts.

We designed the evaluation process to be single-blind (i.e., evaluator identities were not revealed to the authors). Every artifact was evaluated by at least two members of the AEC. We advised the AEC members to work with the authors to help them achieve the badges they sought. This required a significant amount of communication between evaluators and authors as well as accepting several revisions of artifacts and instructions. Due to the single-blind nature of the process, communications were via HotCRP comments. Average length of communications including reviews, comments, and troubleshooting help was 3456 words per paper. The whole process starting from authors registering their artifacts, to evaluators familiarizing themselves with the underlying research papers, to verifying the artifact functionality, to reproducing the results, to writing the reviews, and to awarding the badges was completed in the span of 28 days.

We aimed to recognize authors who had gone the extra mile and put in work to produce artifacts that were of high quality. We formed an independent committee consisting of two reviewers from the AEC. The committee focused on papers which obtained the Results Reproduced Badge and made their artifact publicly available. We believe this extra recognition will motivate authors to produce high-quality artifacts.

Here are the key results from the artifact evaluation process.

- Of the 23 submissions, 21 got *Artifact Available*, 19 got *Artifact Functional*, and 12 got *Results Replicated* badges.
- 96% (22/23) of the submissions earned at least one badge, and 48% (11/23) earned all three badges.
- We saw a strong preference for open-sourcing. In fact, all of the papers that have industry affiliations or industry-based collaborators sought *Artifact Available* badge.
- Despite having to meet rigorous standards in a short evaluation timeframe, 82% artifacts earned the *Artifact Functional* badge, and 52% earned the *Results Replicated* badge.

Given this is our first time, we would be remiss to not share our key takeaways: (i) *AE increased the paper quality.* For instance, when AEC members identified a performance mismatch, the authors worked with them to root cause it to usage of older versions of external dependencies (TensorFlow, in this case). In response, the authors revised the numbers in their camera ready paper. Also, for the first time, a large number of SOSP papers will release artifacts that have been externally validated. (ii) *Specialized hardware is not a hindrance for AE.* Our effort dispels the conventional wisdom that projects involving custom hardware or expansive clusters cannot be evaluated. We observed that in such cases, the authors allowed AEC members to access their resources (via ssh) to perform evaluations. This was the case for 6 (out of 23) evaluations. (iii) *Interest in AE is not limited to academic projects.* Nearly 40% of the submitted artifacts either originated from industry, or had industrial collaborators. Half the industry papers got all three badges. Even when business concerns did not allow open-sourcing the artifacts, the authors were happy to let AEC members access artifacts privately.

We sincerely thank the conference chairs for trusting us with this effort, and hope that these results serve as a catalyst in making artifact evaluation more common at future systems conferences.

Thank you,
Baris Kasikci, Supreeth Shastri, and Vijay Chidambaram.
https://sysartifacts.github.io/