Artifact Evaluation at FPGA 2020: Lessons Learned and Moving Forward

Miriam Leeser and Suhaib Fahmy, Artifact Evaluation Chairs for FPGA 2020 and 2021

1. Introduction

The recent trend encouraging sharing of code, data, and results to foster open science and reproducibility has spread to computer science and computer architecture publication venues. The ACM now has an initiative to award badges for different levels of artifact availability and reproducibility:

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

FPGA 2020 was the first FPGA conference to evaluate artifacts and award badges according to the ACM policies. We were the artifact evaluation chairs for this effort. In this note we describe the process, some misconceptions, and ways we believe the process could be improved. We found some details of the ACM badges to be vague. Thus, in addition to following ACM badging processes, we also considered the efforts of the Supercomputing Conference (SC), and modelled some of our decisions, forms, etc. on the SC effort.

The FPGA community presents unique challenges when it comes to artifacts and their evaluation. An artifact in our community may be an application that maps to a commercial FPGA, a design of a new FPGA architecture or components of that architecture, tools that target a specific commercial FPGA, tools that target any reconfigurable hardware, tools that support developing new FPGA architectures, data associated with experiments, etc.

For the Supercomputing (SC) reproducibility effort, one code artifact was selected and then shared with the student cluster competition teams where they were asked to reproduce the results in the paper. The model was one set of code that should run on many different hardware platforms. While a good model for SC, it does not match the needs of the FPGA community.

We took the approach of matching each artifact with a single evaluator, such that they had access to the tools or hardware needed to replicate the design. In some cases the authors made available the necessary hardware, tools, or platform to perform the evaluation. This 1-to-1 approach is a better match to our community than the 1-to-many approach used by other computing conferences in light of the variation in evaluation requirements.

The process we followed was time constrained. For FPGA, matching of evaluators to artifacts started after paper acceptance decisions were made in mid November. Artifact evaluation results were due early January so that badges could appear in the proceedings. This short timeline affected other decisions regarding the AE process.

2. Badges

ACM recommends that up to three separate badges related to artifact review be associated with research articles in ACM publications: Artifacts Available, Artifacts Evaluated, and Results Validated.

Artifacts Available is a single badge based on the availability of artifacts to the public. We chose to follow the SC practice for Artifacts Available and require a permanent DOI. Note that the ACM badging process does not require such a DOI. However, we felt that such a DOI should be required if an artifact

is to be awarded the Artifacts Available badge. Github provides information on making code citable: https://guides.github.com/activities/citable-code/.

Note that artifacts do not have to be available to be evaluated. An author can have proprietary artifacts (for example if they plan to commercialize the artifact at a later date), which can be evaluated and granted an Artifacts Evaluated Badge, but not an Artifacts Available Badge.

For Artifacts Evaluated, there are two badges, Artifacts Evaluated—Functional, and Artifacts Evaluated—Reusable. A paper can be awarded none or one of these badges, but not both. Papers received the Functional badge if their artifact was documented, all components were available and the evaluator could run the code; this is the minimum standard. Artifacts Evaluated—Reusable is a higher standard, awarded if the evaluator felt the code could be used by others.

The third category of badges is Results Validated. ACM has two flavors of this badge, Results Replicated and Results Reproduced. The Results Reproduced badge requires that the results of the paper be independently obtained. This badge was *not* considered for the FPGA 2020 artifact evaluation process due to time constraints. Results Replicated badges were awarded if the evaluator could replicate the results using the tools and hardware used by the authors. Note that validating results is somewhat orthogonal to the Artifacts Evaluated badges. Artifacts can be poorly documented and not easy to reuse, but the evaluator could still replicate the results.

To summarize, authors who ask for their artifacts to be evaluated can receive between 0 and 3 badges. The 3 badges an FPGA 2020 paper could receive were Artifacts Available, Artifacts Evaluated (Functional or Reusable), and Results Validated (Results Replicated).

3. The process

For FPGA 2020, we asked authors to submit an Artifact Evaluation (AE) form with their paper at the time of submission. Note that the artifacts themselves were not due, just the form. The form asks whether authors have artifacts that they would like to be evaluated, and asks for details of those artifacts.

Collecting the forms at submission time worked well. Forms were not looked at until acceptances were determined, but having them early allowed us to start organizing the AE process as soon as the program committee determined acceptance outcomes, and before authors were notified, gaining a week in this time-constrained process and allowing us to match evaluators to artifacts based on their interests and ability to obtain the hardware or software needed.

It is important that the AE process *not* be double blind, since evaluators need to be able to contact authors, ask clarifying questions and get help in running code where needed. Authors can update details in the form during the evaluation process.

It is also important that paper reviewers *not* have access to the AE forms and that authors not provide identifying information such as a URLs in their paper submission. A link to anonymized code or to state that artifacts would be made available is acceptable, but the paper *review process* should remain double blind. Note that anonymized code is not an artifact. Artifacts need to be identified by their authors. This was perhaps the most misunderstood part of the process.

We recruited evaluators and matched them with artifacts to evaluate. In all 6 cases for FPGA 2020, papers that took part in the process were awarded at least one badge.

4. **Issues:**

Time in the process is very tight. Decisions regarding acceptance are made mid-November, and the results of AE are needed in early January. All of the evaluators adhered to the timeline for FPGA2020. Since there is only one evaluator per artifact it is important that the evaluators be responsible.

We ask for forms at the time of submission, but do not look at artifacts until after acceptance. The artifact to be evaluated was not always available even after paper acceptance. In at least one case for FPGA 2020, authors kept relating that they were not ready to release the code, and so their evaluation was cancelled when we felt the evaluator would not have enough time to do a reasonable job (mid-Dec.).

For FPGA2020 there were 6 papers with artifacts that were evaluated. Thus the matching process was straightforward. If submitting artifacts becomes more popular, the matching process may become more difficult. We propose to prioritize long papers with artifacts over short papers. For FPGA2020 we did not need to do so.

One paper submitted an artifact that was the data used to draw their graphs, and not code. The other five papers submitted code. We debated whether data should get the same badges as code. The rules do not say anything about this, so we decided it should, but this could be open to debate.

Due to the short time schedule, in some cases, evaluators could not evaluate all of the artifacts completely. We asked them to use their judgement. This could result in uneven standards for evaluation.

We would like to reward evaluators for their efforts. We discussed this with the FPGA organizing committee and agreed to offer students who participated in the process travel grants. This was not advertised ahead of time, and might result in more volunteers for evaluation in future. The student travel grant committee asked applicants to report whether or not they participated in the AE process. There was at least one case where someone who did not participate said that they had, and another where someone who was evaluated reported that they participated. If we are going to reward evaluators in this way in the future, we need to better inform evaluators and set some criteria.

We chose to follow the practice of SC to ask authors to use OSI approved licenses for software. https://opensource.org/licenses/alphabetical We also pointed them, on the form, to the open source hardware definition: https://www.oshwa.org/definition/. It is not entirely clear that this definition works well for the FPGA community. In particular, it is not clear if RTL or high level code that gets mapped to an FPGA is hardware or software. We decided to leave it up to the authors to determine if they considered their artifacts to be hardware or software.

5. **Summary**

In our opinion, evaluating artifacts and awarding badges ran smoothly for FPGA2020. The timing worked out well, and we plan to follow the same process for FPGA2021. There were no major issues, although a number of minor things can be improved. The process as it is set up may not scale well if the number of papers with artifacts grows significantly. We view that as a sign of success if more authors

are interested in participating. We welcome input from the community regarding ways to improve the process.