CGO and PPoPP (2015, 2016)

In 2015 and 2016, CGO and PPoPP decided to adopt the Artifact Evaluation process (described below for SIGPLAN and SIGSOFT) to encourage authors to improve their experimental methodologies and to make their experimental artifacts available to other researchers. We adopted the AE process in its basic form with a small change for a clarification period; i.e., the process used is the same as described on the AE web site (http://www.artifact-eval.org). The basic process reviews artifacts for accepted papers to ensure the artifacts and experiments are consistent with the results reported in the published paper. Papers that were successfully validated had a validation seal attached to their camera-ready PDF before inclusion in the ACM Digital Library. Upon successful validation, authors are also encouraged to make their artifact and experimental materials publicly available, e.g., using ACM's supplementary material in the Digital Library, although it is not required.

To form the AE, we invited the program committees of the two conferences to nominate evaluators for the AEC. The evaluators were a mix of senior PhD students, post doctoral fellows, research scientists, and a few faculty. In 2016, we added more members to the AEC and invited the past AEC members to also participate, creating a range of experience with artifact evaluation.

In both years, we made a small change to the basic AE process. The change was to incorporate a clarification step. After an initial review, the evaluators were given a means to ask authors to clarify issues with the artifacts.

In 2016, we made a further change to add sheperding for artifacts. The AEC, operating through the AE chairs, worked with authors to clarify

concerns and fix as many bugs and other issues as possible to get the artifact to the point where it could be successfully validated.

Experience

We found the AE process to work quite smoothly for both conferences. Both conferences are systems oriented; i.e., prototype implementations of compilers, run-time systems, analyzers, debuggers, etc., are used in experiments. The authors generally packaged these prototypes in virtual machines that could be downloaded and run by the evaluators to replicate experimental results. A few cases, as noted below, required special handling.

During evaluation, bugs were found in the documentation, packaging and scripts to run artifacts. The clarification step proved important to allow authors to fix smaller concerns and to clarify documentation. During 2016, the sheperding process permitted authors to fix more complex issues and to clarify how to work with their tools. Only a few artifacts required sheperding. In the end, all submitted artifacts were successfully valdiated! Note the sheperding represents a slight shift in philosophy of AE. In 2015, the view was to reward authors for doing an exemplary job by awarding the seal. In 2016, our view changed to treat the process as a “certification” or “validation” step, where the goal is to usher all submitted artifacts successfully through the process. In this way, we hope the artifacts were improved enough that authors would be more likely to publicly release their materials (when possible).

In both years, a few artifacts required special handling. These artifacts generally revolved around two issues: 1) the artifacts relied on proprietary and/or commercial tools (where licensing was needed), or 2) the artifacts relied on special hardware access. Our belief is everyone should be encouraged and allowed to participate in AE, regardless of whether they

face proprietary issues (i.e., the authors are industrial researchers where their materials cannot be released) or used special machines. To this end, we arranged for authors to provide anonymous remote access to evaluators to get access to the artifacts. The evaluators were able to use this remote access to replicate experiments.

In a couple cases, we were unable to resolve the concerns. In one situation, the industrial researchers could not provide remote access due to company policies. It may have been possible to sign a nondisclosure agreement for access, but this would have violated anaonymity in reviewing and seemed relatively heavyweight for AE. In another situation, the authors used a large-scale parallel machine that did not permit anonymous remote access. Again, this would have violated anonymity of the evaluators and we were unable to accommodate the artifact. In a third situation, the authors could not provide access to the hardware because it was highly specialized (instrumented for power measurement). Instead, the authors provided access to raw data files from their instrument as a proxy to access to the real hardware for replication. In a final situation, the authors used a high-end graphics processing card. For this case, we asked evaluators whether they had the necessary hardware, and we were able to find a couple evaluators with the card. These evaluators reviewed that artifact.

For evaluation, we felt that being as inclusive and flexible as possible, without sacrificing anonymity or imposing a high burden on evaluators/authors, was critical to achieving a high participation rate. It is particularly important for industrial researchers. In communities like PPoPP and CGO, there are many industrial researchers and allowing them to participate in AE is critical to the ultimate goal of improving accountability in reserach methods and tools.

Another issue we faced is the definition of “validation”. That is, what are the criteria that reviews should use? In 2015, authors and evaluators

interpreted the term slightly differently in some cases. For example, some evaluators believed reviewing source code was important for validation, but this was not always made available due to proprietary concerns. Thus, there were sometimes a mismatch in expectation of the evaluation. In 2016, we addressed by this by clarifying that "validation" meant "are the artifact materials consistent with the results reported in the paper?". We left it up to reviewers to interpret and decide how to arrive at the answer to this question. Generally, evaluators took the packaged mateirals, read the documentation and re-run scripts/experiments to replicate results. Some evaluators tried alternative data sets, alternative setups and so forth. We also provide clarification on several dimensions of criteria, where evaluators evaluated specific aspects and were asked to provide written feedback on certain issues. We used a standardize form for this purpose. We also provide authors and evaluators guidelines and steps for evaluation. These materials are available:

1. http://ctuning.org/ae/submission_extra.html
2. http://cTuning.org/ae/submission.html
3. http://ctuning.org/ae/templates/ae–20151015.tex

Process vs. Mechanism

During our experience with CGO and PPoPP, it became apparent that it is important to distinguish between process and mechanism. AE is a process by which evaluation can be carried out (with many options). Equally important is the mechanism used for evaluation. We found that AE as a process is very smooth and navigates through many thorny issues, e.g., industrial participation and specialized hardware. It leaves open specific choices, such as evaluation criteria and how closely evaluation outcomes are tied to paper acceptance. We believe this is the right choice since each community is different. However, we encountered numerous issues with mechanisms. As noted above, authors had to worked around several

concerns in providing access to their materials. Additionally, authors all packaged their materials differently, even if they used the same virtual machine (e.g., VirtualBox). It would greatly help both authors and evaluators to have a set of possible tools available to ease AE, and ultimately, the archiving and distribution of artifact materials.