IACR Transactions on Symmetric Cryptology
ISSN XXXX-XXXX, Vol. 0, No. 0, pp. 1–5. DOI:XXXXXXXX

CHES 2021 Artifact Review

Douglas Stebila[1]

University of Waterloo, Waterloo, Ontario, Canada, dstebila@uwaterloo.ca

For the 2021 edition of the Cryptographic Hardware and Embedded Systems (CHES) conference and its accompanying journal Transactions on Cryptographic Hardware and Embedded Systems (TCHES), the program co-chairs elected to add an artifact review and archiving component for accepted papers, the first of its kind for an IACR event. I was very pleased to be invited to develop and lead the artifact review process.

Full confidence in an experimental result cannot be achieved without the ability to independently reproduce the result. While the physical sciences have a long tradition of publishing protocols and datasets to enable reproducibility, the reproducibility of experimental computer science research has not been as well developed. Increasingly, many authors do post datasets or source code for hardware or software implementations to personal, institutional, or public repositories. Recently, some conferences and journals in computer science, software engineering, and mathematics have started to formally incorporate review and archiving of these artifacts into their processes (see [Kri, Hau] for a partial list).

There are a variety of goals that can be achieved by artifact review. For example, after establishing a task force on reproducibility [Rou17], the Association for Computing Machinery (ACM), developed a "badging" system [Ass20] with several levels of assessment: whether the artifacts are functional, reusable, and publicly available, and whether the results have been reproduced (using the original artifact) or replicated (using independent tools). Several similar classifications exists [Hau, Nat21].

While most would agree that artifact review and archiving is valuable, it is not without its challenges. As Beller [Bel20] writes about his experience serving on the artifact review committee of a major software engineering conference: "the idea of an artifact track sounds simple enough: You install a bit of well-documented software, you let it run for a bit, and you check whether the results in the paper match those produced on your computer. However, in practice, my experience was less than ideal."[1]

Many artifacts are created to demonstrate something for a particular paper, and are thus custom tools written to be used solely by their original authors in a very specific context. As a result, they may have little documentation, have never been tested on another computer, have very specific dependencies, lack an automated build process, may produce output that requires substantial post-processing, and may not have a modular design enabling reuse. The prospect of artifact review and archiving can inspire to authors to think about these factors prior to submission, but artifact reviewers may still struggle with reviewing, installing, running, and interpreting an artifact. Thus, a major outcome of the artifact review process is the improvement of artifacts with respect to these factors.

For TCHES 2021 artifact review, the CHES program committee co-chairs and I decided that our artifact review process should focus on improving the *functionality* and *reusability* of artifacts to enable reproducibility and extension by the scientific community. *Reproducibility* means that the scientific results claimed can be obtained by a different team using the original authors' artifacts. However, the artifact review process for TCHES

[1] Leading to Beller titling his article: "Why I will never join an Artifacts Evaluation Committee Again".

2021 explicitly *did not* include attempting to reproduce the experiment and verify the scientific claims in the accepted paper. (Since TCHES 2021 artifact review took place only for papers that had already been accepted by the program committee, we could proceed under the assumption that PC had accepted the claimed results as scientifically sound. We did of course try to run the software and see if it produced reasonable output, but did not aim to validate it scientifically.) Rather, the artifact review process for TCHES 2021 aimed at ensuring sufficient functionality of the artifact to enable other research teams to attempt to reproduce the results. The artifact review process for TCHES 2021 also aimed to improve reusability, meaning that the artifacts are not just functional, but of sufficient quality that they could be extended and reused by others.

We expected to receive between 5 and 10 artifacts for each of the 4 issues of TCHES 2021. With the help of the CHES PC chairs, I assembled a 33-member artifact review committee, aiming to include committee members with a range of expertises and tools. We included a number of junior researchers: these are often the people tasked with creating artifacts or extending others' artifacts, so they are well-suited to comment on reusability. Having a large committee also helped manage load: each artifact could receive 3 reviews while still assigning each committee member only one artifact per issue, given the potentially greater time required to properly review an artifact.

We used the HotCRP system to manage the submission and review process.[2] Authors submitted an archive of their artifact, a copy of the original paper, information about the tools required to evaluate the artifact, and classified their artifact according to both artifact type (hardware implementation, software implementation, physical attack, etc.) and cryptographic topic. Authors could also link to a GitHub or other repository. (The review process was single-blinded: reviewers were anonymous to authors, but authors' names were known to reviewers.) Assigning reviewers to artifacts is especially important since specialized tools may be required; HotCRP's bidding process was used to assist. In almost all cases at least one committee member had the tools available to be able to successfully run the artifact, although in a few cases we required the assistance of external reviewers.

The review form included questions guiding the feedback from the reviewers, including whether they were able to run the artifact, whether its output was in a form that corresponded to the type of data reported in the paper, whether it was favourable to reuse, and whether the artifact was ready for archiving. I asked the reviewers to think about the artifact from the perspective of a first-year grad student, 5 to 10 years from now, being asked by their supervisor to try to run this artifact and then extend it. Are they likely to succeed? How can we make their job easier?

Given the potential challenges facing reviewers, we used the comment feature of HotCRP to permit reviewers to interact with the authors throughout the entire review process (not just during a rebuttal phase), so that the reviewers could (anonymously on the reviewers' part) ask the authors for help if they were unable to install or run the artifact.

Of the 24 artifact submissions we received across the four issues of TCHES 2021, 22 were accepted for archiving, 4 after shepherding. The remaining 2 were given the chance to revise and resubmit (with requested revisions primarily focused around improving documentation and output), but the authors declined to carry that out, so the artifacts were not accepted for archiving. I do not believe "acceptance rate" is a meaningful metric in this context, given that the artifacts being reviewed are for papers that have already been accepted. Rather, success in artifact review comes from having improved the quality of artifacts, increasing the chance that the results may be reproduced or extended, and ensuring the artifacts are available in the long-term. If all of the submissions are high-quality artifacts ready for use and archiving, so much the better!

The artifact review committee selected one artifact to receive a best artifact award:

[2] The same HotCRP instance was used for all 4 submission dates.

"NTT Multiplication for NTT-unfriendly Rings: New Speed Records for Saber and NTRU on Cortex-M4 and AVX2", by Chi-Ming Marvin Chung, Vincent Hwang, Matthias J. Kannwischer, Gregor Seiler, Cheng-Jhih Shih, and Bo-Yin Yang.

After acceptance, authors were asked to upload their source code, which has been archived in perpetuity on the new IACR artifacts website (https://artifacts.iacr.org/). The TCHES website (where the paper PDFs are archived) also includes prominent links highlighting papers with associated artifacts.

We debated whether to ask authors for source code or an executable (such as a Docker container or virtual machine image), concerned about the likelihood of being able to run the artifact in 5, 10, or 20+ years. We decided to always archive the source code, and consider a pre-built executable format in cases where there are very complex dependencies. This was the case for one artifact, for which we asked authors to also prepare a Linux virtual machine image with all dependencies already installed; their 14 MB source code archive is directly linked from the IACR artifacts website, while their 11 GB virtual machine is archived on an IACR file server and is available upon request.

Reviewing and archiving artifacts raises some intellectual property issues. During the review process, we asked authors if they had patent protection on any of the technology in the artifact, in case some artifact reviewers' employers are sensitive to reviewing the source code of patented technology. For the IACR to distribute an archive of the artifact, appropriate copyright licensing must be ensured. The copyright situation for artifacts is more complicated than for papers, as an artifact may include source code (or binaries) from a variety of sources with different licenses. Authors were required to submit a copyright form that indicated the copyright license under which the authors release the code they wrote, as well as indicating what third-party materials are present in the archive and the corresponding licenses. Given the variety of open source licenses, and to also permit inclusion of artifacts from authors that choose to require a paid license for commercial use, we decided not to mandate the use of a particular license. Indeed we saw a variety of licenses used by authors, including public domain / Creative Commons Zero, the MIT, GPL, and Affero GPL licenses, and software that requires an additional license for commercial use. Licensing information is documented on the webpage for each artifact and included in each artifact's archive.

I am grateful to the authors and artifact review committee members for being flexible and accommodating as we created this process, and for the diligent work done by our review committee members, as well as the efforts by authors to respond and revise. As our community is new to artifact review, both authors and reviewers were asking the question: what is a good artifact? While I cannot answer this question definitively, I do believe that every artifact was improved during this process.

It is too soon to say whether the benefits outweigh the costs and whether artifact review should be widely adopted by the IACR, but the process went much more smoothly than I had feared [Bel20], and we have received positive feedback from many involved. I believe there are clear benefits to artifact review, and that long-term archiving of artifacts improves the reproducibility called for in experimental computer science.

It is heartening to see that TCHES 2022 will continue with artifact review, and it is in excellent hands with Martin Albrecht as the artifact review chair. It is especially promising to see that many members of this year's artifact review committee, as well as authors of artifacts accepted this year, have agreed to serve on next year's committee.

Thank you again to the CHES program committee co-chairs, Elke De Mulder and Peter Schwabe; the review committee members and external reviewers, acknowledged below; and the authors who submitted.

Douglas Stebila
September 2021

Artifact Review Committee

- Andreas Abel, Saarland University
- Martin R. Albrecht, Royal Holloway, University of London
- Alejandro Cabrera Aldaya, Tampere University
- Erdem Alkim, Ondokuz Mayis University
- Estuardo Alpirez Bock, Aalto University
- Pedro G. M. R. Alves, University of Campinas
- Gustavo Banegas, Chalmers University of Technology
- Shivam Bhasin, Temasek Labs, Nanyang Technological University
- Cecylia Bocovich, Tor Project
- Joppe Bos, NXP Semiconductors
- Olivier Bronchain, UCLouvain
- Lauren De Meyer, Rambus
- Cesar Pereida Garcia, Tampere University
- François Gérard, University of Luxembourg
- Sohaib ul Hassan, Tampere University
- James Howe, PQShield
- Jan Jancar, Masaryk University
- Natalia Kulatova, Inria Paris and ENS Ulm
- Kris Kwiatkowski, PQShield
- Norman Lahr, Fraunhofer SIT
- Ben Marshall, University of Bristol
- Guilherme Perin, Delft University of Technology
- Richard Petri, Fraunhofer SIT
- Duy-Phuc Pham, Inria, CNRS, IRISA
- Robert Primas, Graz University of Technology
- Joost Renes, NXP Semiconductors
- Raghvendra Rohit, University of Rennes, CNRS, IRISA
- Pranesh Santikellur, Indian Institute of Technology, Kharagpur
- Nigel Smart, KU Leuven
- Akira Takahashi, Aarhus University
- Pepe Vila, Arm
- Junwei Wang, CryptoExperts
- Lennert Wouters, KU Leuven

External Reviewers

- Kartik Nayak, Università della Svizzera italiana
- Francesco Regazzoni, Universiteit van Amsterdam
- Devanshi Upadhyaya, Universität Stuttgart

References

[Ass20] Association for Computing Machinery. Artifact review and badging version 1.1, August 2020. https://www.acm.org/publications/policies/artifact-review-and-badging-current.

[Bel20] Moritz Beller. Why I will never join an artifacts evaluation committee again, June 2020. https://inventitech.com/blog/why-i-will-never-review-artifacts-again/.

[Hau] Matthias Hauswirth. Experimental evaluation of software and systems in computer science – artifact evaluation. http://evaluate.inf.usi.ch/artifacts.

[Kri] Shriram Krishnamurthi. Artifact evaluation for software conferences. https://artifact-eval.org.

[Nat21] National Information Standards Organization. Reproducibility badging and definitions. Technical Report NISO RP-31-2021, January 2021. https://groups.niso.org/apps/group_public/download.php/24810/RP-31-2021_Reproducibility_Badging_and_Definitions.pdf.

[Rou17] Bernard Rous. The ACM task force on data, software, and reproducibility in publication, January 2017. https://www.acm.org/publications/task-force-on-data-software-and-reproducibility.