International Competition on Runtime Verification (CRV)*

Ezio Bartocci[1], Yliès Falcone[2], Giles Reger[3]

[1] TU Wien, Austria
[2] Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France
[3] Univ. of Manchester, UK

Abstract. We review the first five years of the international Competition on Runtime Verification (CRV), which began in 2014. Runtime verification focuses on verifying system executions directly and is a useful lightweight technique to complement static verification techniques. The competition has gone through a number of changes since its introduction, which we highlight in this paper.

1 Introduction

Runtime verification (RV) is a class of lightweight scalable techniques for the analysis of system executions [16, 17, 6, 4]. The field of RV is broad and encompasses many techniques. The competition has considered a significant subset of techniques concerned with the analysis of user-provided specifications, where executions are checked against a property expressed in a formal specification language. The core idea of RV is to instrument a software/hardware system so that it can emit events during its execution. The sequence of such events (the so-called trace) is then processed by a monitor that is automatically generated from the specification. One usually distinguishes online from offline monitoring depending on whether the monitor runs with the system or post-mortem (and thus collects events from a trace).

In 2014, we observed that, in spite of the growing number of RV tools developed over the previous decade, there was a lack of standard benchmark suites as well as scientific evaluation methods to validate and test new techniques. This observation motivated the promotion of a venue[4] dedicated to comparing and evaluating RV tools in the form of a competition. The Competition on Runtime Verification (CRV) was established as a yearly event in 2014 and has been organized as a satellite event of the RV conference since then [3, 5, 18, 31, 32].

Over the last five years, the competition has helped to shape the development of new tools and evaluation methods but the broad objective of the competitions remain the same. CRV aims to:

* This work was partially supported by the Austrian FWF-funded National Research Network RiSE/SHiNE S11405-N23 and ADynNet project (P28182).

[4] `https://www.rv-competition.org/`

– stimulate the development of new efficient and practical runtime verification tools and the maintenance of the already developed ones;
– produce benchmark suites for runtime verification tools, by sharing case studies and programs that researchers and developers can use in the future to test and to validate their prototypes;
– discuss the metrics employed for comparing the tools;
– compare different aspects of the tools running with different benchmarks and evaluating them using different criteria;
– enhance the visibility of presented tools among different communities (verification, software engineering, distributed computing and cyber-security) involved in monitoring.

Related Work. Over the last two decades, we have witnessed the establishment of several software tool competitions [33, 1, 21, 23, 22, 8] with the goal of advancing the state-of-the-art in the computer-aided verification technology.

In particular, in the area of software verification, there are three related competitions: SV-COMP [8], VerifyThis [22] and the RERS Challenge [21].

SV-COMP targets tools for software model checking, while CRV is dedicated to monitoring tools analyzing only a single program's execution using runtime and offline verification techniques. While in software model checking the verification process is separated from the program execution, runtime verification tools introduce instead an overhead for the monitored program and they consume memory resources affecting the execution of the program itself. As a consequence CRV assigns a score to both the overhead and the memory utilization. Another related series of competitions are VerifyThis [22] and the *Rigorous Examination of Reactive Systems (RERS)* challenge [21] that provide to the participants verification problems to be solved. On the contrary of CRV format, these competitions are problem centred and focus on the problem solving skills of the participants rather than on the tool characteristics and performance.

In the remainder of this paper, we discuss the early years of the competition during 2014-2016 (Section 2), the activities held in 2017 and 2018 that have shifted the focus of the competition (Section 3), and what the future holds for the competition in 2019 and beyond (Section 4).

2 The Early Years: 2014-2016

The early competition was organized into three different tracks: (1) offline monitoring, (2) online monitoring of C programs, and (3) online monitoring of Java programs. The competition spanned over several months before the announcement of results during the conference. The competition consisted of the following steps:

1. **Registration** collected information about participants.
2. **Benchmark Phase.** In this phase, participants submitted benchmarks to be considered for inclusion in the competition.

Table 1. Participants in CRV between 2014 and 2016.

Offline Track	Online C Track	Online Java Track
AgMon [25]	E-ACSL [13]	JavaMOP [24]
BeepBeep3 [19]	MarQ [2]	JUnitRV [12]
Breach	RiTHM-1 [28]	Larva [9]
CRL [30]	RTC [27]	MarQ [2]
LogFire [20]	RV-Monitor [26]	Mufin [11]
MonPoly [7]	TimeSquare [10]	RV-Monitor [26]
MarQ [2]		
OCLR-Check [15]		
OptySim [14]		
RiTHM-2 [28]		
RV-Monitor [26]		

3. **Clarification Phase.** The benchmarks resulting from the previous phase were made available to participants. This phase gave participants an opportunity to seek clarifications from the authors of each benchmark. Only benchmarks that had all clarifications dealt with by the end of this phase were eligible for the next phase.
4. **Monitor Phase.** In this phase, participants were asked to produce monitors for the eligible benchmarks. Monitors had to be runnable via a script on a Linux system. Monitor code should be generated from the participant's tool (therefore the tool had to be installable on a Linux system).
5. **Evaluation Phase.** Submissions from the previous phase were collected and executed, with relevant data collected to compute scores as described later. Participants were given an opportunity to test their submissions on the evaluation system. The outputs produced during the evaluation phase were made available after the competition.

Input Formats. The competition organizers fixed input formats for traces in the offline track. These were based on XML, JSON, and CSV and evolved between the first and second years of the competition based on feedback from participants. The CSV format proved the most popular for its simplicity and is now used by many RV tools. See the competition report from 2015 [18] for details.

Participants. Over the first three years of the competition 14 different RV tools competed in the competition in the different tracks. These are summarized in Table 1. One of these tools, Mufin, was written specifically in response to the competition and all tools were extended or modified to handle the challenges introduced by the competition.

Benchmarks. Benchmarks, as submitted by the participants, should adhere to requirements that ensured compliance with the later phases of the competition. This also ensured uniformity between benchmarks and was also the first

step in building a benchmark repository dedicated to Runtime Verification. A benchmark contains two packages: a program/source package and a specification package. The program/source package includes the traces or the source of the program as well as scripts to compile and run it. In these early years of the competition, we chose to focus on closed, terminating and deterministic programs. The specification package includes an informal and a formal description (in some logical formalism), the instrumentation information (i.e., what in the program influences the truth-value of the specification), and the verdict (i.e., how the specification evaluates w.r.t. the program or trace).

In these three competitions, over 100 benchmarks were submitted and evaluated. All benchmarks are available from the competition website[5] organized in a repository for each year.

Evaluation criteria/scores. Submissions from the participants were evaluated on correctness and performance. For this purpose, we designed an algorithm that uses as inputs (i) the verdicts produced by each tool over each benchmark (ii) the execution time and memory consumption in doing so, and produces as output a score reflecting the evaluation of the tool regarding correctness and performance (the higher, the better). Correctness criteria included (i) finding the expected verdict, absence of crash, and the possibility of expressing the benchmark specification in the tool formalism. Performance criteria were based on the classical time and memory overheads (lower is better) with the addition that the score of a participant accounts for the performance of the other participants (e.g., given the execution time of a participant, more points would be given if the other participants performed poorly) using the harmonic mean. Tools were evaluated against performance, only when they produced a correct result (negative points were given to incorrect results). A benchmark score was assigned for each tool against each submitted benchmark, and the tool final score was the sum of all its benchmark scores. A participant could decide not to compete on a benchmark and would get a zero score for this benchmark.

Experimental environment, availability, reproducibility, quality. Git-based repositories and wiki pages were provided to the participants to share their benchmarks and submissions. This facilitated the communication and ensured transparency. To run the experiments, we used DataMill [29], to ensure robust and reproducible experiments. We selected the most powerful and general-purpose machine and evaluated all submissions on this machine. DataMill ensured flexibility and fairness in the experiments. Benchmarks could be setup and submitted via a Web interface and then be scheduled for execution. DataMill ensured that only one monitor was running on the machine at a time, in addition to a minimalist operating system, cleaned between each experiments. Execution times and memory consumption measures were obtained by averaging 10 executions. Results were available through the Web interface.

[5] `https://www.rv-competition.org/benchmarks/`

Table 2. Winners of CRV between 2014 and 2016.

Year	Offline Track	Online C Track	Online Java Track
2014	MarQ	RiTHM-1	MarQ & JavaMOP (joint)
2015	LogFire	E-ACSL	Mufin
2016	MarQ	-	Mufin

Winners. Table 2 indicates the winners in each track in each year. The detailed results are available from the competition website and associated reports [3, 5, 18]. In 2014, the scores in the Online Java track were so close that a joint winner was announced. In 2016, only one participant entered the C track and the track was not run. (We note that, more tools have been developed for monitoring Java programs thanks to the AspectJ support for instrumentation.)

Issues. The early years of the competition were successful in encouraging RV tool developers to agree on common formats but the number of participants dropped in each year with two main issues identified:

1. The amount of work required to enter was high. This was mainly due to the need to translate each benchmark into the specification language of the entered tool. Common specification languages would address there was no agreement on such languages at the time.
2. It was not clear how good the benchmarks were at differentiating tools. More work was required to understand which benchmarks were useful for evaluating RV tools.

The next two years of activities addressed these issues as described below.

3 Shifting Focus: 2017-2018

In 2017, the competition was replaced by a workshop (called RV-CuBES) [32] aimed at reflecting on the experiences of the last three years and discussing future directions. A workshop was chosen over a competition as there was strong feedback from participants in 2016 that the format of the competition should be revised (mainly to reduce the amount of work required by participants). It was decided that this was a good opportunity to reassess the format of the competition in an open setting. The workshop attracted 12 tool description papers and 5 position papers and led to useful discussion at the 2017 RV conference. A full account can be found in the associated report.

A suggestion of the workshop was to hold a benchmark challenge focusing on collecting relevant new benchmarks. Therefore, in 2018 a benchmark challenge was held with a track for Metric Temporal Logic (MTL) properties and an Open track. The purpose of the MTL track was to see what happened when participants were restricted to a single input language whilst the Open track gave full freedom on the choice of the specification language.

There were two submissions in the MTL track and seven in the Open track. The submissions in the Open track were generally in much more expressive languages than MTL and no two submissions used the same specification language. All submissions were evaluated by a panel of experts and awarded on qualities in three categories: (1) correctness and reliability, (2) realism and challenge, and (3) utility in evaluation. As a result of the evaluation two benchmark sets were identified for use in future competitions (see below).

4 Back to the Future

The 2019 competition is now in its initial stages and will return to a competition comparing tools, using the benchmarks from the 2018 challenge. The competition will use two specification languages: MTL and a future-time first-order temporal logic. We have chosen to fix two specification languages (with differing levels of expressiveness) to reduce the overall work for participants. Standardising the specification language of the competition has been a goal of the competition from the start and the benchmark challenge has allowed us to pick two good candidates. MTL was chosen as it can be considered a 'smallest shared' specification language in terms of expressiveness and usage. Similarly, the future-time first-order temporal logic was chosen as it can be considered a 'largest shared' specification language in terms of expressiveness and usage.

Beyond 2019, there are many opportunities to take the competition in different directions. For example, a key issue in RV is that of specifications. Thus, when organizing a competition, one may wonder whether a competition could also focus on evaluating aspects related to specifications (e.g., expressiveness, succinctness and elegance of specifications). Moreover, in so far, the competition has neglected the area of hardware monitoring, and the comparison of tools in such domains remains an open question. We note that there have been less research efforts on monitoring hardware where instrumentation aspects are more challenging. The main reasons for common specification languages not being used in the early years stemmed from two facts: (i) a main research activity in RV consists in developing new languages to have alternative representation of problems (ii) the monitoring algorithm of an RV tool is often closely coupled to the input language. Hence, a challenge is to rely on a shared specification language whilst encouraging research that explores the relationship between input language and performance or usability.

Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful comments and all the colleagues involved in CRV over the last years (B. Bonakdarpour, D. Thoma, D. Nickovic, S. Halle, K. Y. Rozier, and V. Stolz).

References

1. C. Barrett, M. Deters, L. de Moura, A. Oliveras, and A. Stump. 6 Years of SMT-COMP. *Journal of Automated Reasoning*, pages 1–35, 2012.
2. H. Barringer, Y. Falcone, K. Havelund, G. Reger, and D. E. Rydeheard. Quantified Event Automata: Towards Expressive and Efficient Runtime Monitors. In *Proc. of FM 2012: the 18th International Symposium on Formal Methods*, volume 7436 of *LNCS*, pages 68–84. Springer, 2012.
3. E. Bartocci, B. Bonakdarpour, and Y. Falcone. First international competition on software for runtime verification. In B. Bonakdarpour and S. A. Smolka, editors, *Proc. of RV 2014: the 5th International Conference on Runtime Verification*, volume 8734 of *LNCS*, pages 1–9. Springer, 2014.
4. E. Bartocci and Y. Falcone, editors. *Lectures on Runtime Verification - Introductory and Advanced Topics*, volume 10457 of *LNCS*. Springer, 2018.
5. E. Bartocci, Y. Falcone, B. Bonakdarpour, C. Colombo, N. Decker, K. Havelund, Y. Joshi, F. Klaedtke, R. Milewicz, G. Reger, G. Rosu, J. Signoles, D. Thoma, E. Zalinescu, and Y. Zhang. First International Competition on Runtime Verification: rules, benchmarks, tools, and final results of CRV 2014. *International Journal on Software Tools for Technology Transfer*, 21:31–70, Apr 2019.
6. E. Bartocci, Y. Falcone, A. Francalanza, and G. Reger. Introduction to runtime verification. In *Lectures on Runtime Verification - Introductory and Advanced Topics*, volume 10457 of *LNCS*, pages 1–33. Springer, 2018.
7. D. Basin, M. Harvan, F. Klaedtke, and E. Zălinescu. MONPOLY: Monitoring usage-control policies. In *Proc. of RV 2011: the 2nd Internat. Conference on Runtime Verification*, volume 7186 of *LNCS*, pages 360–364. Springer, 2012.
8. D. Beyer. Software verification and verifiable witnesses - (report on SV-COMP 2015). In *Proc. of TACAS 2015: the 21st International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, volume 9035, pages 401–416. Springer, 2015.
9. C. Colombo, G. J. Pace, and G. Schneider. Larva — safer monitoring of real-time java programs (tool paper). In *Proceedings of the 2009 Seventh IEEE International Conference on Software Engineering and Formal Methods*, SEFM '09, pages 33–37, Washington, DC, USA, 2009. IEEE Computer Society.
10. J. Deantoni and F. Mallet. TimeSquare: Treat your Models with Logical Time. In S. N. Carlo A. Furia, editor, *TOOLS - 50th International Conference on Objects, Models, Components, Patterns - 2012*, volume 7304, pages 34–41, Prague, Czech Republic, May 2012. Czech Technical University in Prague, in co-operation with ETH Zurich, Springer.
11. N. Decker, J. Harder, T. Scheffel, M. Schmitz, and D. Thoma. Runtime monitoring with union-find structures. In *Proc. of TACAS 2016: the 22nd International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, volume 9636 of *LNCS*. Springer, 2016.
12. N. Decker, M. Leucker, and D. Thoma. jUnit$^{\mathrm{RV}}$—adding runtime verification to jUnit. In G. Brat, N. Rungta, and A. Venet, editors, *NASA Formal Methods, 5th International Symposium, NFM 2013, Moffett Field, CA, USA, May 14-16, 2013. Proceedings*, volume 7871 of *Lecture Notes in Computer Science*, pages 459–464. Springer, 2013.
13. M. Delahaye, N. Kosmatov, and J. Signoles. Common Specification Language for Static and Dynamic Analysis of C Programs. In *Proceedings of SAC '13: the 28th Annual ACM Symposium on Applied Computing*, pages 1230–1235. ACM, Mar. 2013.

14. A. Díaz, P. Merino, and A. Salmeron. Obtaining models for realistic mobile network simulations using real traces. *IEEE Communications Letters*, 15(7):782–784, 2011.
15. W. Dou, D. Bianculli, and L. Briand. A model-driven approach to offline trace checking of temporal properties with ocl. Technical Report SnT-TR-2014-5, Interdisciplinary Centre for Security, Reliability and Trust, 2014.
16. Y. Falcone. You should better enforce than verify. In *Runtime Verification - First International Conference, RV 2010, St. Julians, Malta, November 1-4, 2010. Proceedings*, pages 89–105, 2010.
17. Y. Falcone, K. Havelund, and G. Reger. A tutorial on runtime verification. In M. Broy, D. A. Peled, and G. Kalus, editors, *Engineering Dependable Software Systems*, volume 34 of *NATO Science for Peace and Security Series, D: Information and Communication Security*, pages 141–175. IOS Press, 2013.
18. Y. Falcone, D. Nickovic, G. Reger, and D. Thoma. Second international competition on runtime verification CRV 2015. In E. Bartocci and R. Majumdar, editors, *Proc. of RV 2015: the 6th International Conference on Runtime Verification*, volume 9333 of *LNCS*, pages 405–422. Springer, 2015.
19. S. Hallé. When RV meets CEP. In Y. Falcone and C. Sánchez, editors, *Runtime Verification - 16th International Conference, RV 2016, Madrid, Spain, September 23-30, 2016, Proceedings*, volume 10012 of *LNCS*, pages 68–91. Springer, 2016.
20. K. Havelund. Rule-based Runtime Verification Revisited. *International Journal on Software Tools for Technology Transfer (STTT)*, page To appear, 2014.
21. F. Howar, M. Isberner, M. Merten, B. Steffen, D. Beyer, and C. S. Pasareanu. Rigorous examination of reactive systems - the RERS challenges 2012 and 2013. *STTT*, 16(5):457–464, 2014.
22. M. Huisman, V. Klebanov, and R. Monahan. Verifythis 2012 - A program verification competition. *STTT*, 17(6):647–657, 2015.
23. M. Järvisalo, D. L. Berre, O. Roussel, and L. Simon. The international SAT solver competitions. *AI Magazine*, 33(1), 2012.
24. D. Jin, P. O. Meredith, C. Lee, and G. Roşu. JavaMOP: Efficient Parametric Runtime Monitoring Framework. In *Proceedings of ICSE 2012: THE 34th International Conference on Software Engineering, Zurich, Switzerland, June 2-9*, pages 1427–1430. IEEE Press, 2012.
25. A. Kane, T. E. Fuhrman, and P. Koopman. Monitor based oracles for cyber-physical system testing: Practical experience report. In *44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, DSN 2014, Atlanta, GA, USA, June 23-26, 2014*, pages 148–155. IEEE, 2014.
26. Q. Luo, Y. Zhang, C. Lee, D. Jin, P. O. Meredith, T. Serbanuta, and G. Rosu. Rv-monitor: Efficient parametric runtime verification with simultaneous properties. In *Runtime Verification - 5th International Conference, RV 2014, Toronto, ON, Canada, September 22-25, 2014. Proceedings*, pages 285–300, 2014.
27. R. Milewicz, R. Vanka, J. Tuck, D. Quinlan, and P. Pirkelbauer. Runtime checking c programs. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, pages 2107–2114. ACM, 2015.
28. S. Navabpour, Y. Joshi, C. W. W. Wu, S. Berkovich, R. Medhat, B. Bonakdarpour, and S. Fischmeister. RiTHM: a tool for enabling time-triggered runtime verification for c programs. In *ACM Symposium on the Foundations of Software Engineering (FSE)*, pages 603–606, 2013.
29. J. Petkovich, A. B. de Oliveira, Y. Zhang, T. Reidemeister, and S. Fischmeister. Datamill: a distributed heterogeneous infrastructure for robust experimentation. *Softw., Pract. Exper.*, 46(10):1411–1440, 2016.

30. A. Piel. *Reconnaissance de comportements complexes par traitement en ligne de flux d'événements. (Online event flow processing for complex behaviour recognition).* PhD thesis, Paris 13 University, Villetaneuse, Saint-Denis, Bobigny, France, 2014.
31. G. Reger, S. Hallé, and Y. Falcone. Third international competition on runtime verification - CRV 2016. In Y. Falcone and C. Sánchez, editors, *Proc. of RV 2016: the 16th International Conference on Runtime Verification*, volume 10012 of *LNCS*, pages 21–37. Springer, 2016.
32. G. Reger and K. Havelund, editors. *RV-CuBES 2017. An International Workshop on Competitions, Usability, Benchmarks, Evaluation, and Standardisation for Runtime Verification Tools*, volume 3 of *Kalpa Publications in Computing*. EasyChair, 2017.
33. G. Sutcliffe. The 5th IJCAR automated theorem proving system competition - CASC-J5. *AI Commun.*, 24(1):75–89, 2011.