12th Competition on Software Verification

Dirk Beyer (Competition Chair)

Motivation - Goals

1. Community suffers from unreproducible results
   $\rightarrow$ Establish set of benchmarks
2. Publicity for tools that are available
   $\rightarrow$ Provide state-of-the-art overview
3. Support the development of verification tools
   $\rightarrow$ Give credits and visibility to developers
4. Establish standards
   $\rightarrow$ Specification language, Witnesses,
   Benchmark definitions, Validators

Schedule of Sessions

Session 1:

- ▶ Competition Report, by organizer
- ▶ System Presentations, 6 min by each team
- ▶ Short discussion

Session 2:

- ▶ Open Jury Meeting, Community Discussion, moderated by organizer

Procedure – Time Line

Three Steps – Three Deadlines:

- ▶ Benchmark submission deadline
- ▶ System submission
- ▶ Notification of results (approved by teams)

Verification Problem

Input:

- ▶ C program $\rightarrow$ GNU/ANSI C standard
- ▶ Property
  - $\rightarrow$ Reachability of error label, of overflows
  - $\rightarrow$ Memory safety (inv-deref, inv-free, memleak)
  - $\rightarrow$ Termination

Output:

- ▶ TRUE + Witness   (property holds)
- ▶ FALSE + Witness   (property does not hold)
- ▶ UNKNOWN   (failed to compute result)

Environment

Machines (1000 \$ consumer machines):
- ▶ CPU: 3.4 GHz 64-bit Quad-Core CPU
- ▶ RAM: 33 GB
- ▶ OS: GNU/Linux (Ubuntu 22.04)

Resource limits:
- ▶ 15 GB memory
- ▶ 15 min CPU time (consumed 470 days)

Volume: $490\,858$ verification runs, 1.4 million validation runs
Incl. preruns: 2.8 million verification runs using 14 years, and
36 million validation runs using 18 years of CPU time

Scoring Schema

Common principles: Ranking measure should be

- ► easy to understand
- ► reproducible
- ► computable in isolation for one tool

SV-COMP:

- ► Ranking measure is the quality of verification work
- ► Expressed by a community-agreed score
- ► Tie-breaker is CPU time

Scoring Schema (2023, unchanged)

Reported result	Points	Description
UNKNOWN	0	Failure, out of ressources
FALSE correct	+1	Error found and confirmed
FALSE incorrect	−16	False alarm (imprecise analysis)
TRUE correct	+2	Proof found and confirmed
TRUE incorrect	−32	Missed bug (unsound analysis)

Fair and Transparent

Jury:
- ▶ Team: one member of each participating candidate
- ▶ Term: one year (until next participants are determined)

Systems:
- ▶ All systems are available in open GitLab repo
- ▶ Configurations and Setup in GitHub repository
  $\rightarrow$ Integrity and reproducibility guaranteed

52 Competition Candidates

Qualification:
- ▶ 34 qualified, additional 18 hors concours
- ▶ 10 results validator, 1 witness linter
- ▶ One person can participate with different tools
- ▶ One tool can participate with several configurations (frameworks, no tool-name inflation)

Benchmark quality:
- ▶ Community effort, documented on GitHub

Role of organizer:
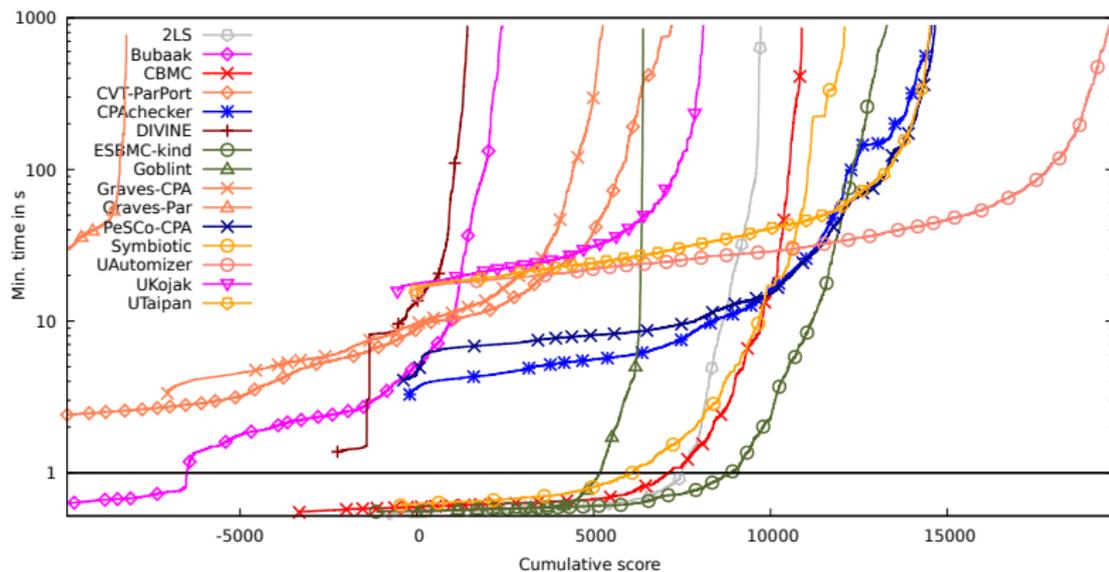- ▶ Just service: Advice, Technical Help, Executing Runs

Benchmark Sets

- Everybody can submit benchmarks (conditions apply)
- Eight categories when closed (scores normalized):
  - Reachability: 9814 tasks
  - Memory Safety: 4543 tasks
  - Concurrency: 5295 tasks
  - NoOverflows: 10200 tasks
  - Termination: 3103 tasks
  - Software Systems: 5132 tasks
  - Overall: 38644 tasks
  - Java: 827 tasks

Reproducibility

- ▶ - SV-Benchmarks:
  https://gitlab.com/sosy-lab/benchmarking/sv-benchmarks
- ▶ - SV-COMP Setup:
  https://gitlab.com/sosy-lab/sv-comp/bench-defs
- ▶ - Resource Measurement and Process Control:
  https://github.com/sosy-lab/benchexec
- ▶ - Archives:
  https://gitlab.com/sosy-lab/sv-comp/archives-2023
- ▶ - Witnesses:
  https://doi.org/10.5281/zenodo.7627791

Results – Example: Overall

Impact / Achievements

- ▶ Large benchmark set of verification tasks
  $\rightarrow$ established and used in many papers
  for experimental evaluation
- ▶ Good overview over state-of-the art
  $\rightarrow$ covers model checking and program analysis
- ▶ Participants have an archived track record
  of their achievements
- ▶ Infrastructure and technology for
  controlling the benchmark runs (cf. StarExec)

[Competition Report and System Descriptions
are archived in Proceedings TACAS 2023]
https://doi.org/10.1007/978-3-031-30820-8_29

Number of Participants

Number of evaluated verifiers for each year
(first-time participants on top)
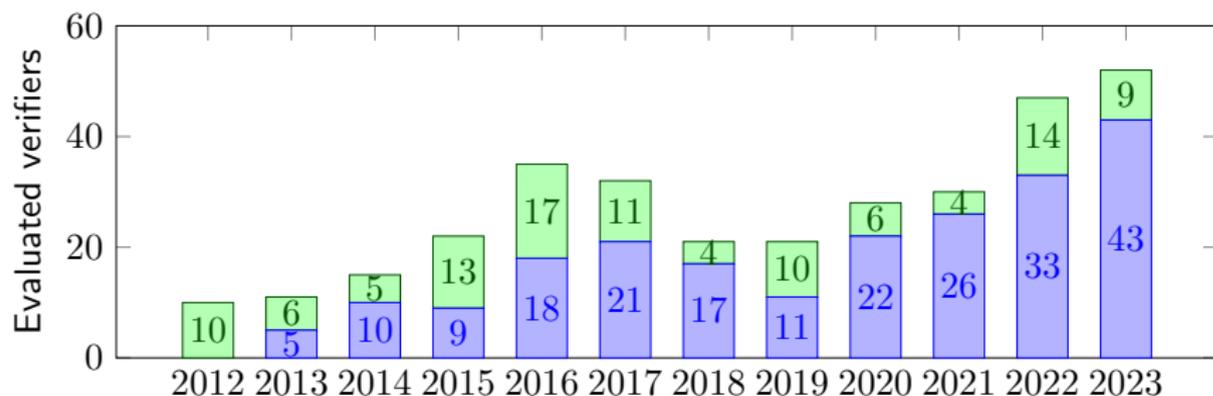
New Developments

New 2023:

- ▶ Validation Track was established 2023
- ▶ with witness refutation

New 2024 (TBD):

- ▶ DOIs from now one for verifier archives
- ▶ Extended scoring schema
- ▶ Extension of witness benchmark set
- ▶ New witness format for correctness, violation
- ▶ Benchmark extensions

Thanks to:

- TACAS (PC Chairs + TACAS SC, thanks!)
- Jury (32 people)
- Participants (184 people)
- Next we celebrate the winners
  (show slides and hand out awards)