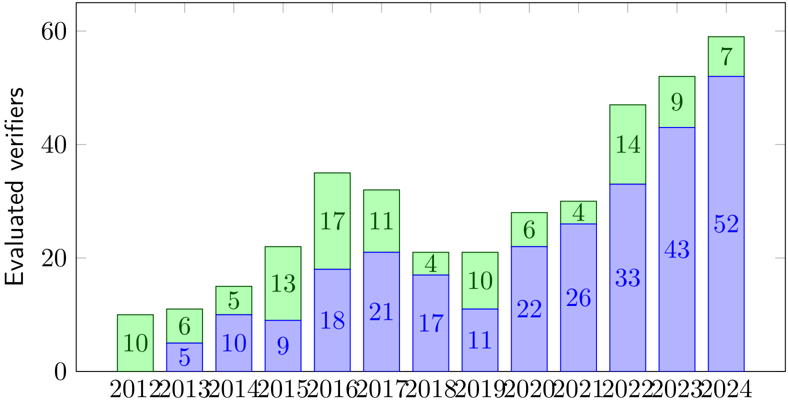
13th Competition on Software Verification

Dirk Beyer (Competition Chair)

Proc. TACAS 2024, doi:10.1007/978-3-031-57256-2_15

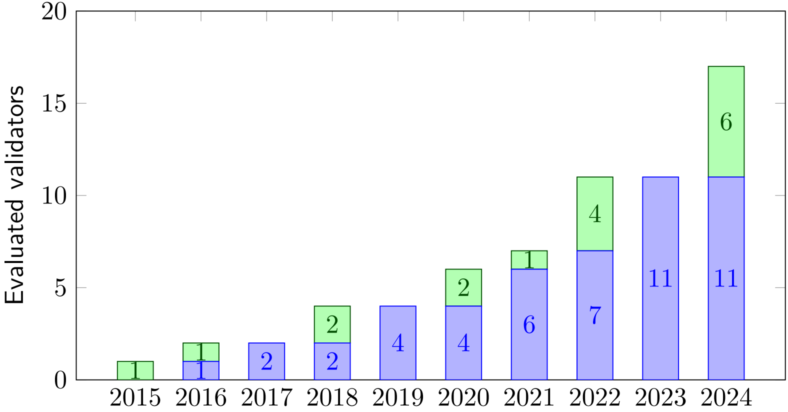
Number of Participants — Verification

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



Number of Participants — Validation

Number of evaluated validators for each year
(first-time participants on top)



Motivation - Goals

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

Schedule of Sessions

Session 1:

- Competition Report, by organizer
- System Presentations, 4 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 → GNU/ANSI C standard
- Property
  → Reachability of error label, of overflows
  → Memory safety (inv-deref, inv-free, memleak)
  → 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

Volume: $787\,779$ verification runs, 13.6 million validation runs (training pre-runs not included)

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
  → Integrity and reproducibility guaranteed

76 Competition Candidates

Qualification:

- 59 verification track
- 17 in validation track
- 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

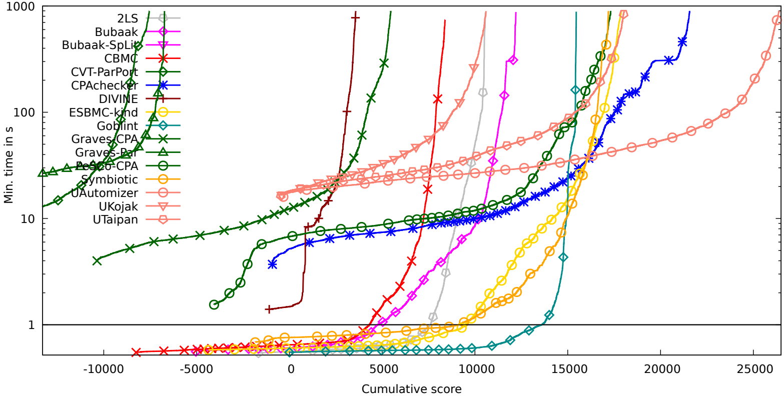
Benchmark Sets

- Everybody can submit benchmarks (conditions apply)
- Eight categories when closed (scores normalized):
    - Reachability: $11\,222$ tasks
    - Memory Safety: $2\,080$ tasks
    - Concurrency: $3\,129$ tasks
    - NoOverflows: $8\,113$ tasks
    - Termination: $2\,298$ tasks
    - Software Systems: $3\,458$ tasks
    - Overall: $30\,300$ tasks
    - Java: 587 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/benchmarking/fm-tools
- Witnesses:
  https://doi.org/10.5281/zenodo.10669737

Results – Example: Overall



Impact / Achievements

- Large benchmark set of verification tasks
  → established and used in many papers
  for experimental evaluation
- Good overview over state-of-the art
  → 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 2024]
https://doi.org/10.1007/978-3-031-57256-2_15

New Developments

New 2024:

- Tools are submitted via DOIs from now on
- Validation Track was established 2023
- Now with more witnesses
  (classification by definition and majority)
- New witness format 2.0 for correctness, violation
- Benchmark extensions

New 2025 (Hopefully not much, consolidation phase):

- Benchmark restructuring
- Complete adoption of witnesses of version 2.0
  (but still keep 1.0)

Thanks to:

- TACAS (PC Chairs + TACAS SC, thanks!)
- Jury and program committee (∼40 people)
- Participants (203 people)
- Next we celebrate the winners

Report:

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

ETAPS EUROPEAN JOINT CONFERENCES ON THEORY & PRACTICE OF SOFTWARE