8th Competition on Software Verification

Dirk Beyer (Competition Chair)

Supported by:
The content of this presentation is available in the SV-COMP 2019 report: https://doi.org/10.1007/978-3-030-17502-3_9
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
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 18.04)

Resource limits:

- 15 GB memory
- 15 min CPU time (consumed 461 days)

Volume: 178 674 ver. runs, 517 175 val. runs
Total: 5 880 071 runs using 15 years and 182 days of CPU time
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 (2019)
(from 2020 onwards: only confirmed results count)

<table>
<thead>
<tr>
<th>Reported result</th>
<th>Points</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNKNOWN</td>
<td>0</td>
<td>Failure, out of ressources</td>
</tr>
<tr>
<td>FALSE correct</td>
<td>+1</td>
<td>Error found and confirmed</td>
</tr>
<tr>
<td>FALSE incorrect</td>
<td>-16</td>
<td>False alarm (imprecise analysis)</td>
</tr>
<tr>
<td>TRUE correct</td>
<td>+2</td>
<td>Proof found and confirmed</td>
</tr>
<tr>
<td>TRUE unconfirmed</td>
<td>+1</td>
<td>Proof found but unconfirmed</td>
</tr>
<tr>
<td>TRUE incorrect</td>
<td>-32</td>
<td>Missed bug (unsound analysis)</td>
</tr>
</tbody>
</table>
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
31 Competition Candidates

Qualification:

- 31 Qualified (out of 31 Submitted)
  1 verifier disqualified from several categories (rule viol.)
- 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: 3831 tasks
  - Memory Safety: 434 tasks
  - Concurrency: 1082 tasks
  - NoOverflows: 359 tasks
  - Termination: 2007 tasks
  - Software Systems: 2809 tasks
  - Overall: 10522 tasks
  - Java: 368 tasks
Replicability

- SV-Benchmarks:
  https://github.com/sosy-lab/sv-benchmarks
- SV-COMP Setup:
  https://github.com/sosy-lab/sv-comp
- Resource Measurement and Process Control:
  https://github.com/sosy-lab/benchexec
- Archives:
- Witnesses:
  https://sv-comp.sosy-lab.org/2019/results/results-verified
# Results – Details at: sv-comp.sosy-lab.org

<table>
<thead>
<tr>
<th>verifier status</th>
<th>score</th>
<th>witness</th>
<th>inspect</th>
<th>cpu (s)</th>
<th>wall (s)</th>
<th>energy (J)</th>
<th>mem (MB)</th>
<th>blkio-w (MB)</th>
<th>blkio-r (MB)</th>
<th>validator cpachecker violation</th>
<th>t&lt;90s status</th>
<th>validator vaautimizer violation</th>
<th>t&lt;90s status</th>
<th>cpu (s)</th>
<th>wall (s)</th>
<th>energy (J)</th>
<th>mem (MB)</th>
<th>blkio-w (MB)</th>
<th>blkio-r (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ination.cil.c</td>
<td>false</td>
<td>unreach-call</td>
<td>true</td>
<td>1.3</td>
<td>1.3</td>
<td>18</td>
<td>370</td>
<td>0.0041</td>
<td>0</td>
<td>false(unreach-call)</td>
<td>8.2</td>
<td>4.4</td>
<td>120</td>
<td>320</td>
<td>17</td>
<td>9.1</td>
<td>320</td>
<td>520</td>
<td></td>
</tr>
<tr>
<td>ation.cil.c</td>
<td>false</td>
<td>unreach-call</td>
<td>true</td>
<td>0.35</td>
<td>0.34</td>
<td>3.3</td>
<td>60</td>
<td>0.0041</td>
<td>0</td>
<td>false(unreach-call)</td>
<td>8.1</td>
<td>4.3</td>
<td>170</td>
<td>310</td>
<td>13</td>
<td>6.6</td>
<td>240</td>
<td>450</td>
<td></td>
</tr>
<tr>
<td>ation.cil.c</td>
<td>false</td>
<td>unreach-call</td>
<td>true</td>
<td>0.5</td>
<td>0.53</td>
<td>4.9</td>
<td>120</td>
<td>0.0041</td>
<td>0</td>
<td>false(unreach-call)</td>
<td>8.3</td>
<td>4.4</td>
<td>100</td>
<td>330</td>
<td>12</td>
<td>6.6</td>
<td>210</td>
<td>500</td>
<td></td>
</tr>
<tr>
<td>ation.cil.c</td>
<td>false</td>
<td>unreach-call</td>
<td>true</td>
<td>0.29</td>
<td>0.28</td>
<td>2.4</td>
<td>39</td>
<td>0.0041</td>
<td>0</td>
<td>false(unreach-call)</td>
<td>7.8</td>
<td>4.2</td>
<td>80</td>
<td>380</td>
<td>13</td>
<td>6.7</td>
<td>180</td>
<td>410</td>
<td></td>
</tr>
<tr>
<td>ation.cil.c</td>
<td>true</td>
<td>1</td>
<td>true</td>
<td>1.4</td>
<td>1.4</td>
<td>14</td>
<td>410</td>
<td>0.0041</td>
<td>12</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>ation.cil.c</td>
<td>true</td>
<td>1</td>
<td>true</td>
<td>0.53</td>
<td>0.53</td>
<td>4.9</td>
<td>110</td>
<td>0.0041</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>ation.cil.c</td>
<td>true</td>
<td>1</td>
<td>true</td>
<td>0.35</td>
<td>0.35</td>
<td>3.3</td>
<td>67</td>
<td>0.0041</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>ation.cil.c</td>
<td>true</td>
<td>1</td>
<td>true</td>
<td>0.58</td>
<td>0.58</td>
<td>5.6</td>
<td>130</td>
<td>0.0041</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>ation.cil.c</td>
<td>true</td>
<td>1</td>
<td>true</td>
<td>0.19</td>
<td>0.19</td>
<td>1.4</td>
<td>26</td>
<td>0.0041</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>ation.cil.c</td>
<td>true</td>
<td>1</td>
<td>true</td>
<td>0.29</td>
<td>0.29</td>
<td>2.3</td>
<td>45</td>
<td>0.0041</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>false(unreach-call)</td>
<td>false</td>
<td>unreach-call</td>
<td>true</td>
<td>21</td>
<td>21</td>
<td>200</td>
<td>200</td>
<td>0.0041</td>
<td>0</td>
<td>false(unreach-call)</td>
<td>7.4</td>
<td>3.9</td>
<td>170</td>
<td>310</td>
<td>12</td>
<td>6.8</td>
<td>210</td>
<td>460</td>
<td></td>
</tr>
<tr>
<td>false(unreach-call)</td>
<td>false</td>
<td>unreach-call</td>
<td>true</td>
<td>23</td>
<td>23</td>
<td>200</td>
<td>200</td>
<td>0.0041</td>
<td>0</td>
<td>false(unreach-call)</td>
<td>7.4</td>
<td>3.9</td>
<td>130</td>
<td>310</td>
<td>14</td>
<td>7.1</td>
<td>200</td>
<td>480</td>
<td></td>
</tr>
</tbody>
</table>
Results – Example: Overall

"Value' of result is defined by Scoring Schema
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 ’19]
https://doi.org/10.1007/978-3-030-17502-3_9
Correct Verifiers — Low Failure Rate:

\[
\frac{\text{number of incorrect results}}{\text{total score}} \quad \text{with unit } E/sp.
\]

Green Verifiers — Low Energy Consumption:

\[
\frac{\text{total CPU energy}}{\text{total score}} \quad \text{with the unit } J/sp.
\]

New Verifiers — High Quality:
quality in score points as rank measure.
Table 9: Alternative rankings; quality is given in score points (sp), CPU time in hours (h), energy in kilojoule (kJ), wrong results in errors (E), rank measures in errors per score point (E/sp), joule per score point (J/sp), and score points (sp)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Verifier</th>
<th>Quality (sp)</th>
<th>CPU Time (h)</th>
<th>CPU Energy (kJ)</th>
<th>Solved Tasks</th>
<th>Wrong Results (E)</th>
<th>Rank Measure (J/sp)</th>
<th>Rank Measure (E/sp)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correct Verifiers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>CPA-SEQ</td>
<td>9329</td>
<td>120</td>
<td>4300</td>
<td>2811</td>
<td>0</td>
<td>.0000</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>SYMBIOTIC</td>
<td>6129</td>
<td>9.7</td>
<td>390</td>
<td>2519</td>
<td>0</td>
<td>.0000</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>PeSCo</td>
<td>8466</td>
<td>120</td>
<td>3900</td>
<td>2431</td>
<td>9</td>
<td>.0011</td>
<td></td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.3836</td>
</tr>
<tr>
<td></td>
<td>Green Verifiers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>SYMBIOTIC</td>
<td>6129</td>
<td>9.7</td>
<td>390</td>
<td>299</td>
<td>0</td>
<td>64</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>CBMC</td>
<td>4341</td>
<td>11</td>
<td>380</td>
<td>296</td>
<td>14</td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>DIVINE-EXPLICIT</td>
<td>1547</td>
<td>4.4</td>
<td>180</td>
<td>256</td>
<td>10</td>
<td>120</td>
<td></td>
</tr>
<tr>
<td></td>
<td>worst</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4200</td>
</tr>
<tr>
<td></td>
<td>New Verifiers</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>PeSCo</td>
<td>8466</td>
<td>120</td>
<td>3900</td>
<td>1026</td>
<td>9</td>
<td>8466</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>CBMC-PATH</td>
<td>1587</td>
<td>8.9</td>
<td>380</td>
<td>1006</td>
<td>69</td>
<td>1587</td>
<td></td>
</tr>
</tbody>
</table>
## Table 10: Confirmation rate of verification witnesses in SV-COMP 2019

<table>
<thead>
<tr>
<th>Result</th>
<th>True</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Confirmed</td>
</tr>
<tr>
<td>CPA-Seq</td>
<td>4417</td>
<td>3968</td>
</tr>
<tr>
<td>PeSCo</td>
<td>4176</td>
<td>3814</td>
</tr>
<tr>
<td>UAutomizer</td>
<td>4244</td>
<td>4199</td>
</tr>
<tr>
<td>Symbiotic</td>
<td>2430</td>
<td>2381</td>
</tr>
<tr>
<td>CBMC</td>
<td>1813</td>
<td>1702</td>
</tr>
<tr>
<td>UTaipan</td>
<td>3015</td>
<td>2936</td>
</tr>
<tr>
<td>2LS</td>
<td>2072</td>
<td>2045</td>
</tr>
<tr>
<td>ESBMC-kind</td>
<td>3679</td>
<td>3556</td>
</tr>
<tr>
<td>UKojak</td>
<td>2070</td>
<td>2038</td>
</tr>
<tr>
<td>CBMC-Path</td>
<td>1206</td>
<td>1162</td>
</tr>
<tr>
<td>DIVINE-explicit</td>
<td>693</td>
<td>673</td>
</tr>
<tr>
<td>DIVINE-SMT</td>
<td>645</td>
<td>626</td>
</tr>
<tr>
<td>DepthK</td>
<td>612</td>
<td>602</td>
</tr>
</tbody>
</table>
Number of Participants

Fig. 7: Number of participating teams for each year
Open Issues for SV-COMP 2020

- Task definition
- License
- More programs
- LTL properties
- Eliminate pre-processing
- Undefined behavior of C programs
- Witnesses in all categories
- Tests as Witnesses
format_version: '1.0'

# old file name: floppy_true-unreach-call_true-valid-memsafety.i.cil.c
input_files: 'floppy.i.cil-3.c'

properties:
  - property_file: ../properties/unreach-call.prp
    expected_verdict: true
  - property_file: ../properties/valid-memsafety.prp
    expected_verdict: true

Fig. 3: Example task definition for program floppy.i.cil-3.c
Practical Impact: Get Tests from Verification Tools
Thanks to:

- TACAS (PC Chairs + TACAS SC, thanks!)
- Jury (32 people)
- Participants (177 people)
- Sponsors: Amazon Web Services and LMU Munich
Benchmark Definition

```xml
<?xml version="1.0"?>
<!DOCTYPE benchmark PUBLIC "+//IDN sosy-lab.org//DTD BenchExec benchmark 1.9//EN"
"http://www.sosy-lab.org/benchexec/benchmark-1.9.dtd">
<benchmark tool="cpachecker" timelimit="15 min" hardtimelimit="16 min"
memlimit="15 GB" cpuCores="8">
<require cpuModel="Intel Xeon E3-1230 v5 @ 3.40 GHz" cpuCores="8"/>
<resultfiles>**.graphml</resultfiles>
<option name="--svcomp19"/>
<option name="--heap">10000M</option>
<option name="--benchmark"/>
<option name="--timelimit">900 s</option>
<rundefi"
The licensor grants to the user of this software the following rights, irrevocably:

1. Redistribution of exact copies of this archive.
2. Execution of this software for research and evaluation purposes.
3. Publication of the output and measurement results obtained from executing this software.

The licensor confirms that the licenses of all components contained in this archive are compatible with the above requirements.
Competition Guarantee

GUARANTEE OF RIGHTS FOR RESEARCH AND EVALUATION

The licensor guarantees that the licenses of all components contained in this archive grant the following perpetual, worldwide, no-charge, royalty-free, irrevocable rights to everybody:

1. Redistribution of exact copies of this archive.
2. Execution of this software for research and evaluation purposes.
3. Publication of the output and measurement results obtained from executing this software.

For all parts of the software for which the licensor holds the copyright, the licensor grants to the user of this software the above rights. This takes precedence over any contradicting clauses in accompanying licenses.
Search-Space Reduction for Stepwise Testification

Search space

Entry

Error
Produce Witnesses

- Program
- Specification
- Verification Task

Tools:
- BLAST
- CBMC
- CPAchecker
- ESBMC
- SMACK
- ULTIMATE AutoMizer

Outputs: Witnesses
Search-Space Reduction for Stepwise Testification
Refine Witnesses
Search-Space Reduction for Stepwise Testification

Entry

Search space

Error

→ Stepwise Testification ←
Produce Unit Tests From Witnesses
Search-Space Reduction for Stepwise Testification

![Diagram showing search space and stepwise testification](image)

- **Search space**
- **Entry**
- **Error**
- **Stepwise Testification**