8th Competition on Software Verification

Dirk Beyer (Competition Chair)

Supported by:

References

The content of this presentation is avaliable 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

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 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

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 (2019)

(from 2020 onwards: only confirmed results count)

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 unconfirmed	+1	Proof found but unconfirmed
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

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:
  https://gitlab.com/sosy-lab/sv-comp/archives-2019
- - Witnesses:
  https://sv-comp.sosy-lab.org/2019/results/
  results-verified

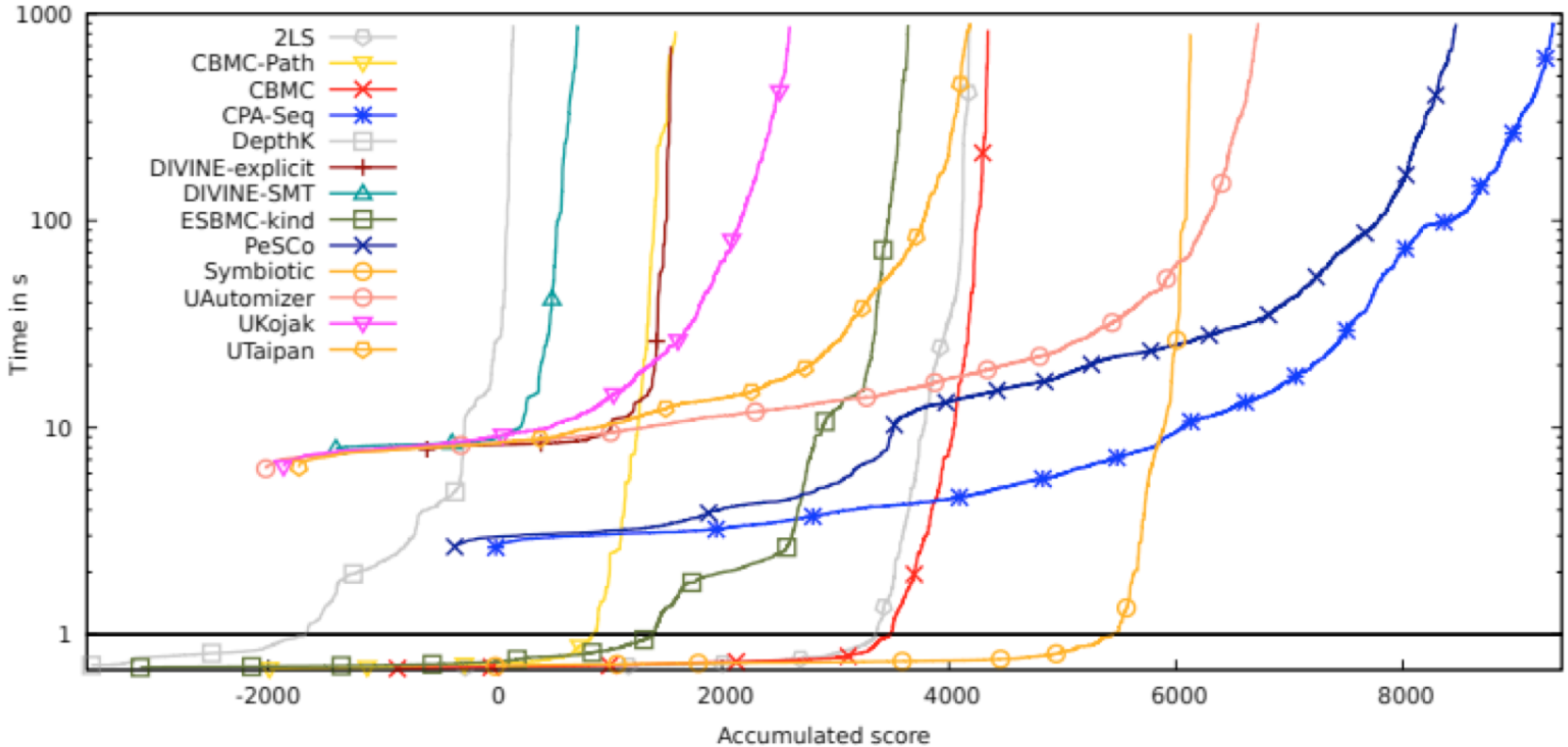
Results – Details at: sv-comp.sosy-lab.org

Select Columns | Filter Rows | Quantile Plot | Scatter Plot | Shrink Header | Generated with BenchExec

	2LS 0.5.0
	timelimit: 900 s, memlimit: 15000 MB, CPU core limit: 8
	apollon*
	Linux 4.4.0-57-generic
	CPU: Intel Xeon E3-1230 v5 @ 3.40 GHz, cores: 8, frequency: 3.8 GHz, Turbo Boost: disabled; RAM: 335
	2017-01-10 17:21:21 CET [[2017-01-14 18:00:17 CET]] [[2017-01-14 20:02:31 CET]] [[2017-01-14 18:18:08 CET]] [[20
	sv-comp17.ReachSafety-ControlFlow
	--graphml-witness witness.graphml [[-witnessValidation -setprop witness.checkProgramHash=false -disable-java-assertions -heap 10000m -witness ../../results-verified/2ls.2017-01-10_1721.logfiles/sv-comp17.${inputfile_name}.files/witness.graphml]] [[--validate ../../results-verified/2ls.2017-01-10_1721.logfiles/sv-comp17.${inputfile_name}.files/witness.graphml]] [[-witnessValidation -setprop witness.checkProgramHash=false -disable-java-assertions -heap 10000m -witness ../../results-verified/2ls.2017-01-10_1721.logfiles/sv-comp17.${inputfile_name}.files/witness.graphml]] [[--validate ../../results-verified/2ls.2017-01-10_1721.logfiles/sv-comp17.${inputfile_name}.files/witness.graphml]]

	verifier status	score	witness	inspect witness	cpu (s)	wall (s)	energy (J)	mem (MB)	blkio-w (MB)	blkio-r (MB)	validator cpachecker violation t<90s status	cpu (s)	wall (s)	energy (J)	mem (MB)	validator uautomizer violation t<90s status	cpu (s)	wall (s)	energy (J)	mem (MB)	c
ination.cil.c	false(unreach-call)	1	wit	inspect	1.3	1.3	13	370	.0041	0	false(unreach-call)	8.2	4.4	120	320	false(unreach-call)	17	9.1	320	520	
nation.cil.c	false(unreach-call)	1	wit	inspect	.35	.34	3.3	60	.0041	0	false(unreach-call)	8.1	4.3	170	310	false(unreach-call)	13	6.6	240	450	
nation.cil.c	false(unreach-call)	1	wit	inspect	.55	.53	4.9	120	.0041	0	false(unreach-call)	8.3	4.4	100	330	false(unreach-call)	12	6.6	210	500	
ination.cil.c	false(unreach-call)	1	wit	inspect	.29	.28	2.4	39	.0041	0	false(unreach-call)	7.8	4.2	80	380	false(unreach-call)	13	6.7	180	410	
nation.cil.c	true	2	wit	inspect	1.4	1.4	14	410	.0041	12	-	–	–	–	–	-	–	–	–	–	
ination.cil.c	true	2	wit	inspect	.53	.53	4.9	110	.0041	0	-	–	–	–	–	-	–	–	–	–	
tion.cil.c	true	2	wit	inspect	.36	.35	3.3	67	.0041	0	-	–	–	–	–	-	–	–	–	–	
tion.cil.c	true	2	wit	inspect	.59	.58	5.6	130	.0041	0	-	–	–	–	–	-	–	–	–	–	
nation.cil.c	true	2	wit	inspect	.19	.19	1.4	26	.0041	0	-	–	–	–	–	-	–	–	–	–	
nation.cil.c	true	2	wit	inspect	.29	.29	2.3	45	.0041	0	-	–	–	–	–	-	–	–	–	–	
	false(unreach-call)	1	wit	inspect	21	21	200	200	.0041	0	false(unreach-call)	7.4	3.9	170	310	false(unreach-call)	12	6.8	210	460	
	false(unreach-call)	1	wit	inspect	23	23	200	200	.0041	0	false(unreach-call)	7.4	3.9	130	310	false(unreach-call)	14	7.1	200	480	

Results – Example: Overall



'Value' of result is defined by Scoring Schema

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 '19]
https://doi.org/10.1007/978-3-030-17502-3_9

Alternative Rankings — Definitions

- Correct Verifiers — Low Failure Rate:

$$\frac{\text{number of incorrect results}}{\text{total score}}$$

with unit E/sp.

- Green Verifiers — Low Energy Consumption:

$$\frac{\text{total CPU energy}}{\text{total score}}$$

with the unit J/sp.

- New Verifiers — High Quality:
quality in score points as rank measure.

Alternative Rankings — Results

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)

Rank	Verifier	Quality (sp)	CPU Time (h)	CPU Energy (kJ)	Solved Tasks	Wrong Results (E)	Rank Measure
Correct Verifiers							(E/sp)
1	**CPA-Seq**	9 329	120	4 300	2 811	0	.0000
2	Symbiotic	6 129	9.7	390	2 519	0	.0000
3	PeSCo	8 466	120	3 900	2 431	9	.0011
worst							.3836
Green Verifiers							(J/sp)
1	**Symbiotic**	6 129	9.7	390	299	0	64
2	CBMC	4 341	11	380	296	14	88
3	DIVINE-explicit	1 547	4.4	180	256	10	120
worst							4 200
New Verifiers							(sp)
1	**PeSCo**	8 466	120	3 900	1 026	9	8 466
2	CBMC-Path	1 587	8.9	380	1 006	69	1 587

Witness-Based Result Validation

Table 10: Confirmation rate of verification witnesses in SV-COMP 2019

Result	True				False			
	Total	Confirmed		Unconf.	Total	Confirmed		Unconf.
CPA-Seq	4 417	3 968	90 %	449	2 859	2 686	94 %	173
PeSCo	4 176	3 814	91 %	362	2 823	2 652	94 %	171
UAutomizer	4 244	4 199	99 %	45	1 523	1 255	82 %	268
Symbiotic	2 430	2 381	98 %	49	1 451	1 214	84 %	237
CBMC	1 813	1 702	94 %	111	1 975	1 248	63 %	727
UTaipan	3 015	2 936	97 %	79	915	653	71 %	262
2LS	2 072	2 045	99 %	27	1 419	945	67 %	474
ESBMC-kind	3 679	3 556	97 %	123	2 141	1 753	82 %	388
UKojak	2 070	2 038	98 %	32	553	548	99 %	5
CBMC-Path	1 206	1 162	96 %	44	897	670	75 %	727
DIVINE-explicit	693	673	97 %	20	768	353	46 %	415
DIVINE-SMT	645	626	97 %	19	943	601	64 %	342
DepthK	612	602	98 %	10	1 938	1 370	71 %	568

Number of Participants

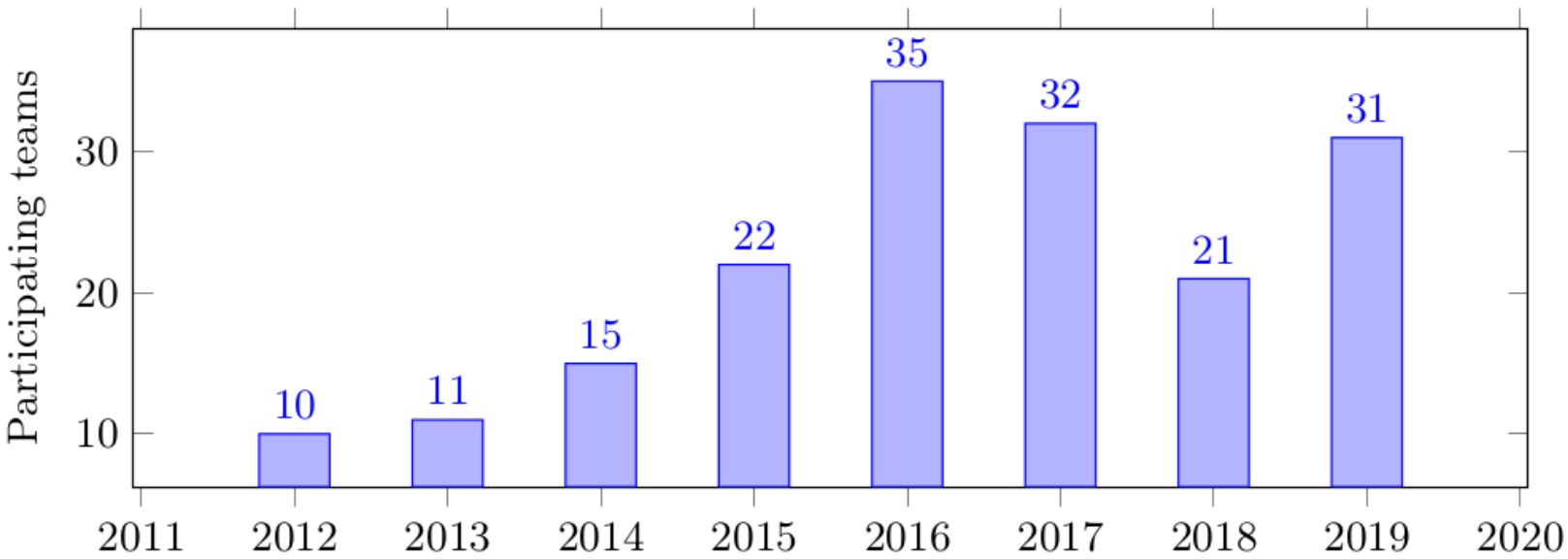


Fig. 7: Number of participating teams for each year

Open Issues for SV-COMP 2020

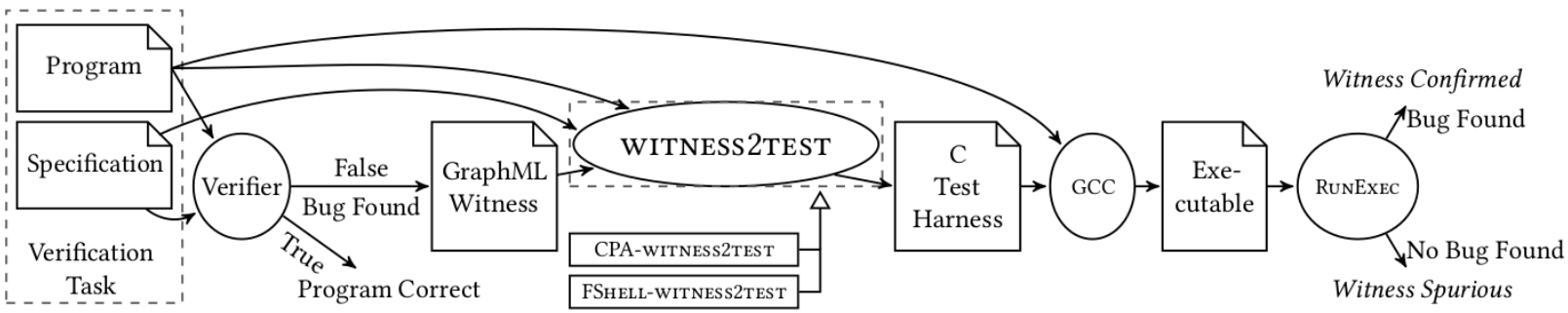
- Task definition
- License
- More programs
- LTL properties
- Eliminate pre-processing
- Undefined behavior of C programs
- Witnesses in all categories
- Tests as Witnesses

Task Definition

```
1  format_version: '1.0'
2
3  # old file name: floppy_true-unreach-call_true-valid-memsafety.i.cil.c
4  input_files: 'floppy.i.cil-3.c'
5
6  properties:
7    - property_file: ../properties/unreach-call.prp
8      expected_verdict: true
9    - property_file: ../properties/valid-memsafety.prp
10     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 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>
<rundefinition name="sv-comp19_prop-reachsafety">
<tasks name="ReachSafety-Arrays">
<includesfile>../sv-benchmarks/c/ReachSafety-Arrays.set</includesfile>
<propertyfile>../sv-benchmarks/c/properties/unreach-call.prp</propertyfile>
</tasks>
<tasks name="ReachSafety-BitVectors">
<includesfile>../sv-benchmarks/c/ReachSafety-BitVectors.set</includesfile>
<propertyfile>../sv-benchmarks/c/properties/unreach-call.prp</propertyfile>
</tasks>
```

Competition License

LICENSE FOR RESEARCH AND EVALUATION

[...]

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

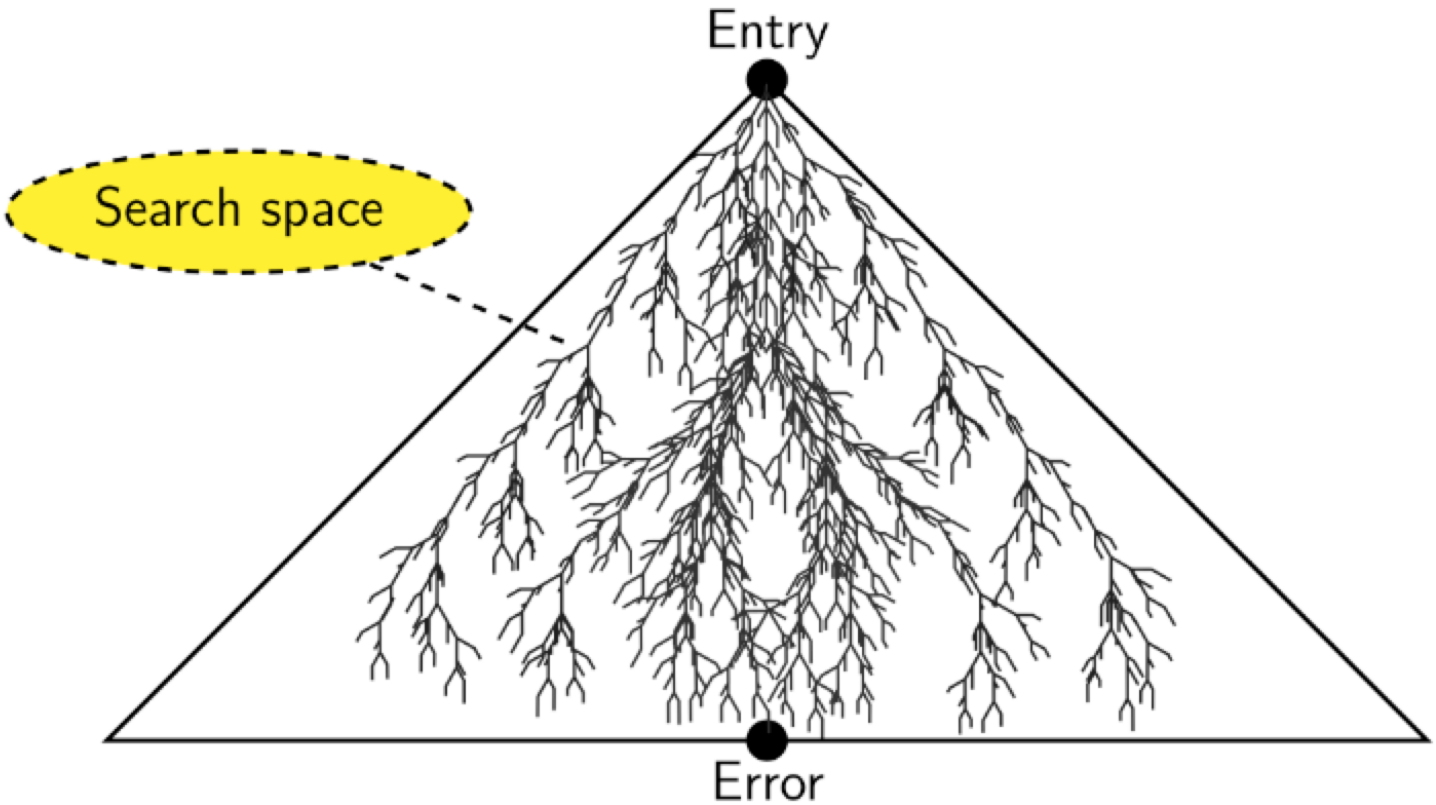
GUARANTEE OF RIGHTS FOR RESEARCH AND EVALUATION

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:

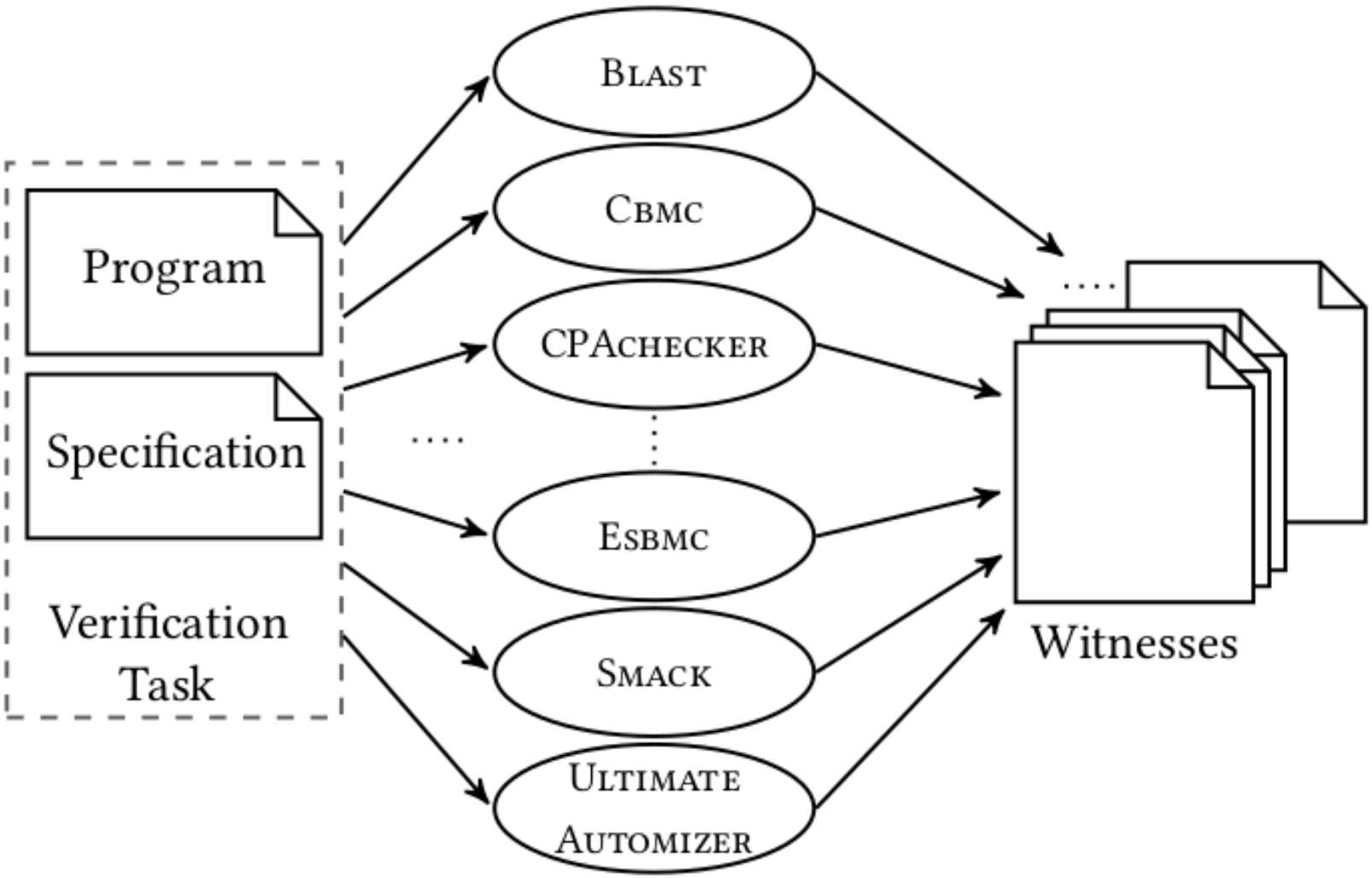
1. Redistribution of exact copies of this archive.
2. Execution of this software for research and evaluation purposes.
3. Publication of the output and measurement results obtained from executing this software.

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.

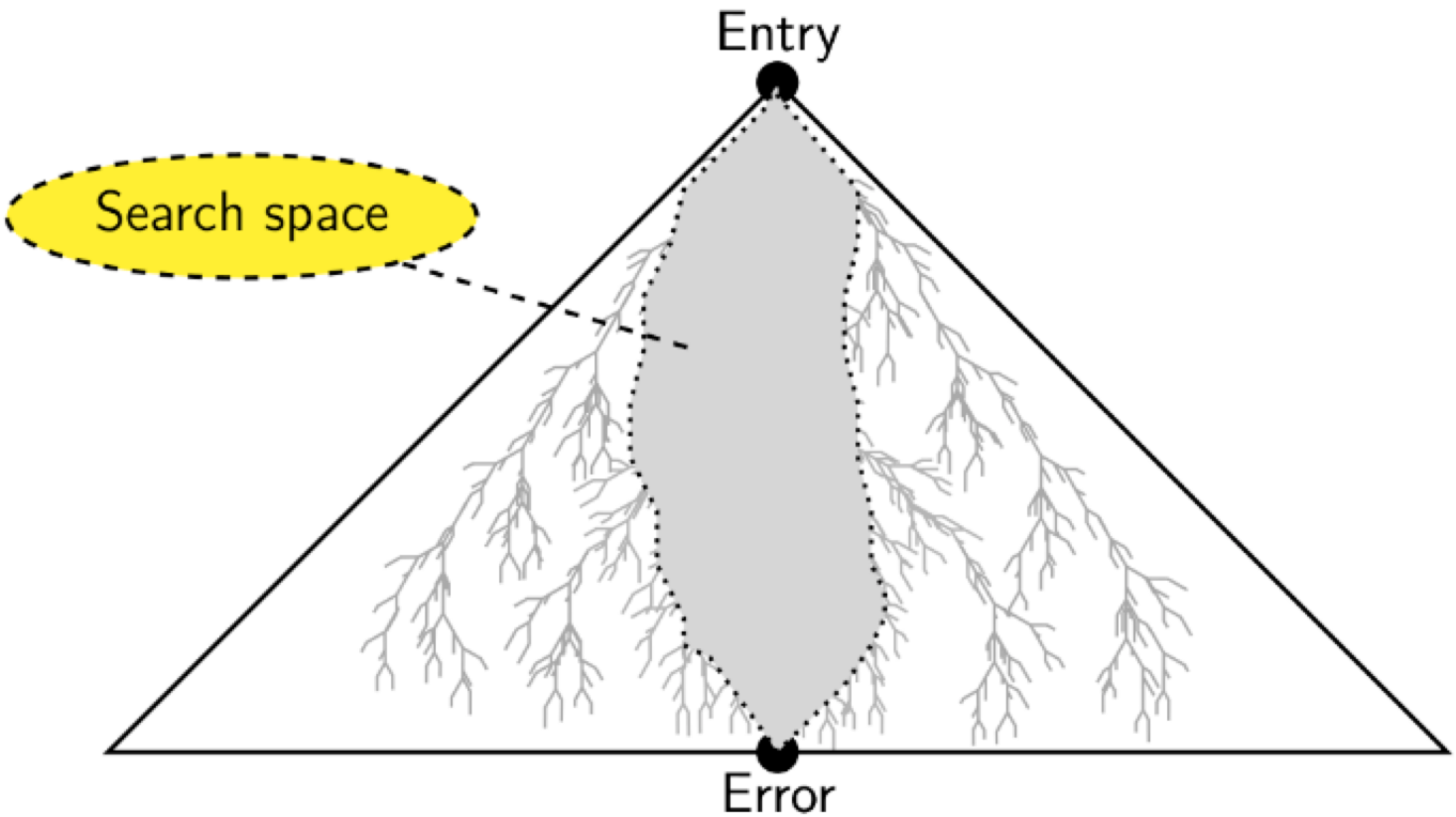
Search-Space Reduction for Stepwise Testification



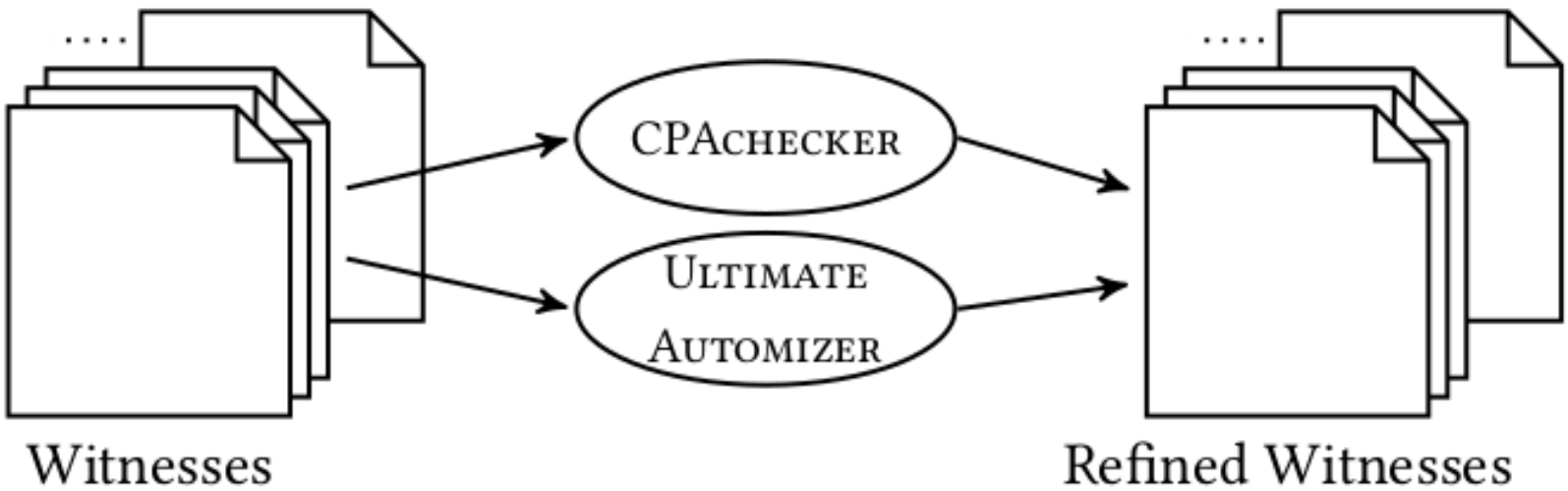
Produce Witnesses



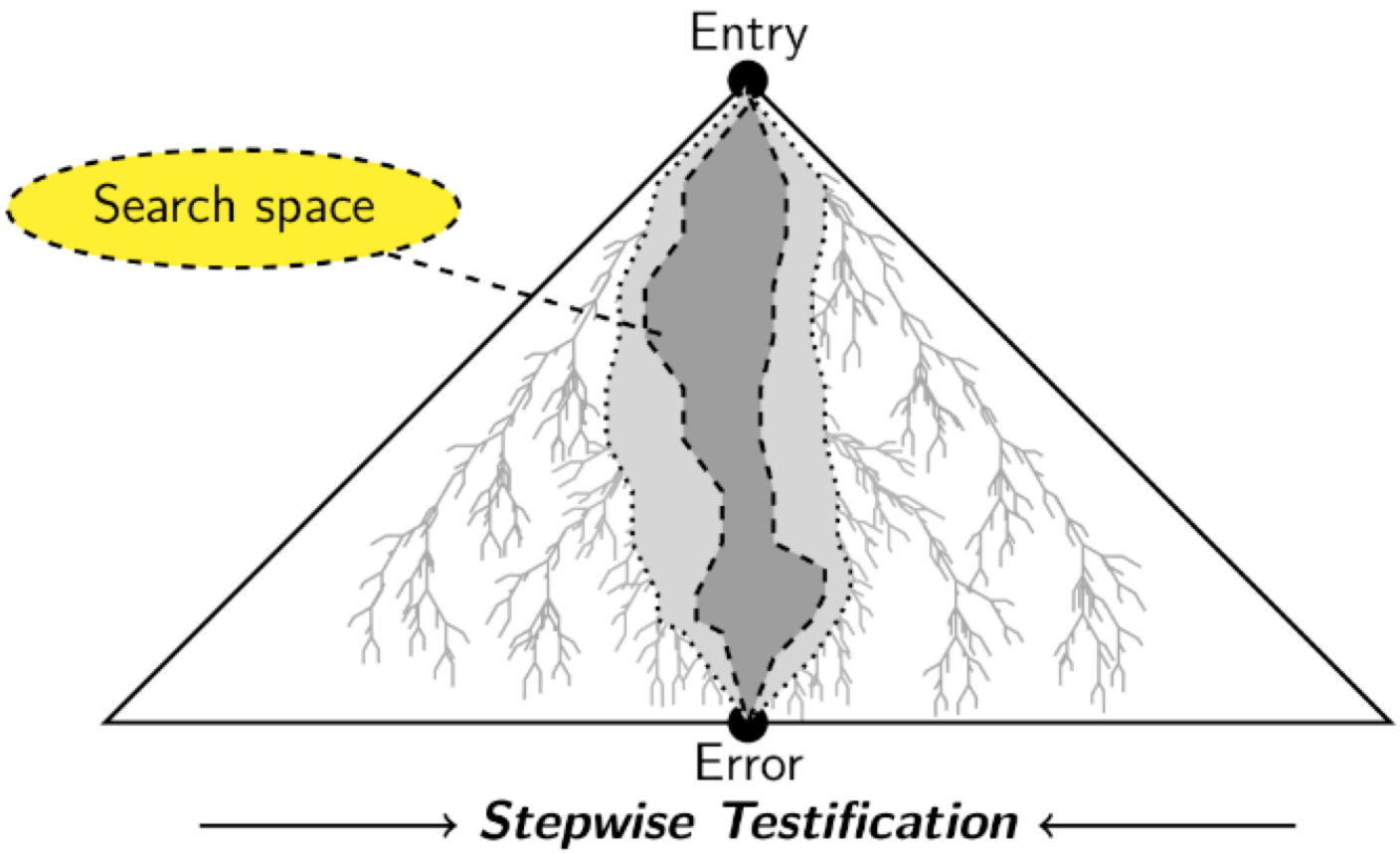
Search-Space Reduction for Stepwise Testification



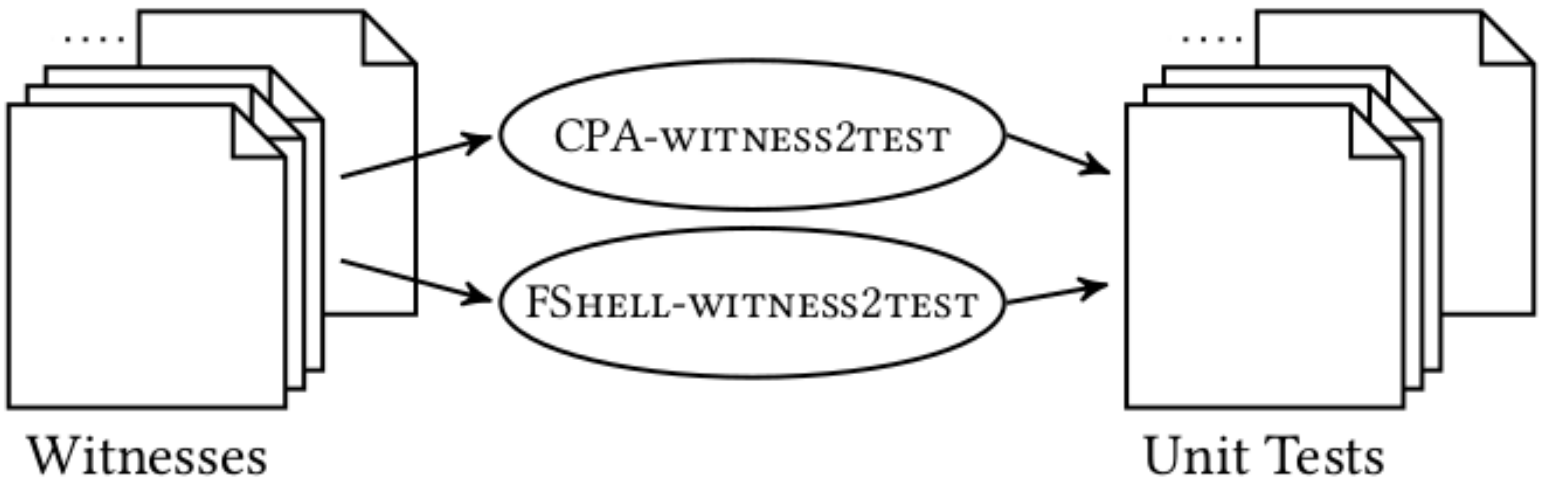
Refine Witnesses



Search-Space Reduction for Stepwise Testification



Produce Unit TestsFrom Witnesses



Search-Space Reduction for Stepwise Testification