# Diagnosing Performance Differences in Model Checkers via Runtime-Guided Problem Generation

**Yibo Dong** [1,3] Yicong Xu [3] Wenjing Deng [3] Yu Chen [4,3] Xiaoyu Zhang [3] Jianwen Li[3] **Chengyu Zhang** [2] **Geguang Pu**[3]

[1]National University of Singapore [2]Loughborough University [3]East China Normal University [4]Chuzhou University

## Background

Safety model checking could be reduced to reachability analysis.

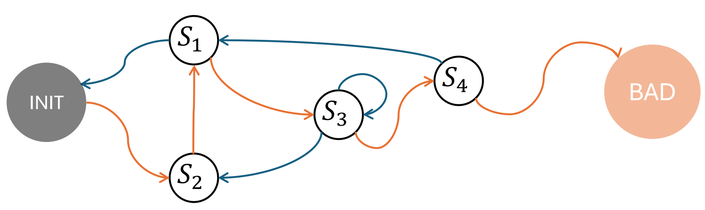



Figure 1. Simple Safety Model Checking

In practical designs, each *state* is a complete assignment to all variables, leading to a state space that grows exponentially with the number of variables.


Figure 2. Practical Safety Model Checking

## Motivation

SOTA model checkers are COMPLEX

- Accumulation of intricate optimizations
- Fine-tuned implementation details
- ...

> It's just a minor fix!
> Why is there such a huge change in performance!
> How could we diagnose?

Compact yet Challenging problems could help!

## Approach

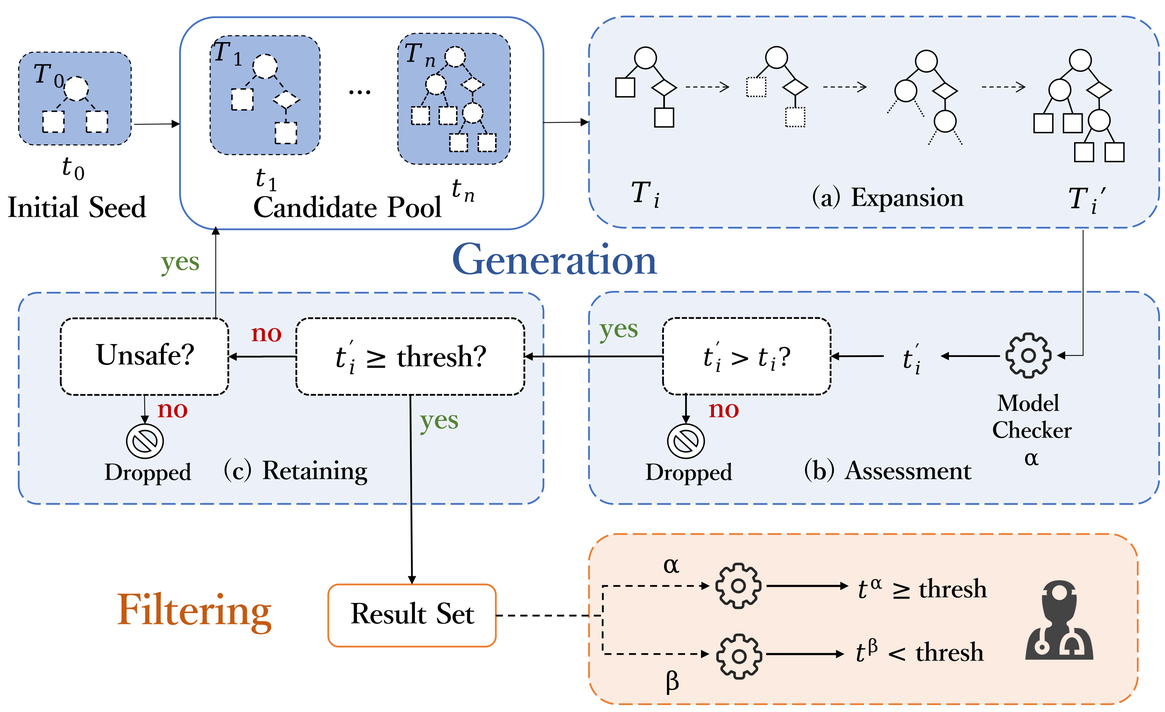



Figure 3. Workflow: Feedback-guided Generation

## Key insight: Safe Remaining

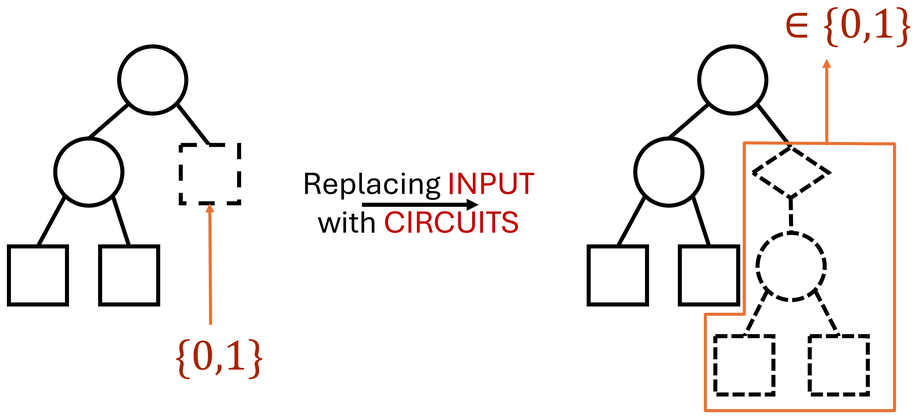



Figure 4. After expansion, safe cases remain safe.

## Evalution: Efficiency

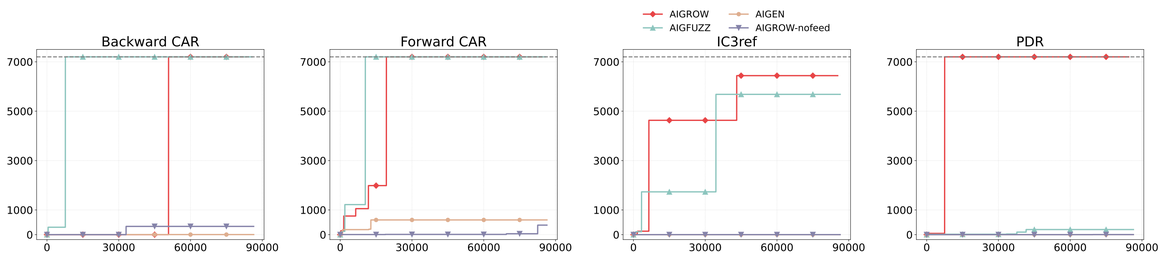


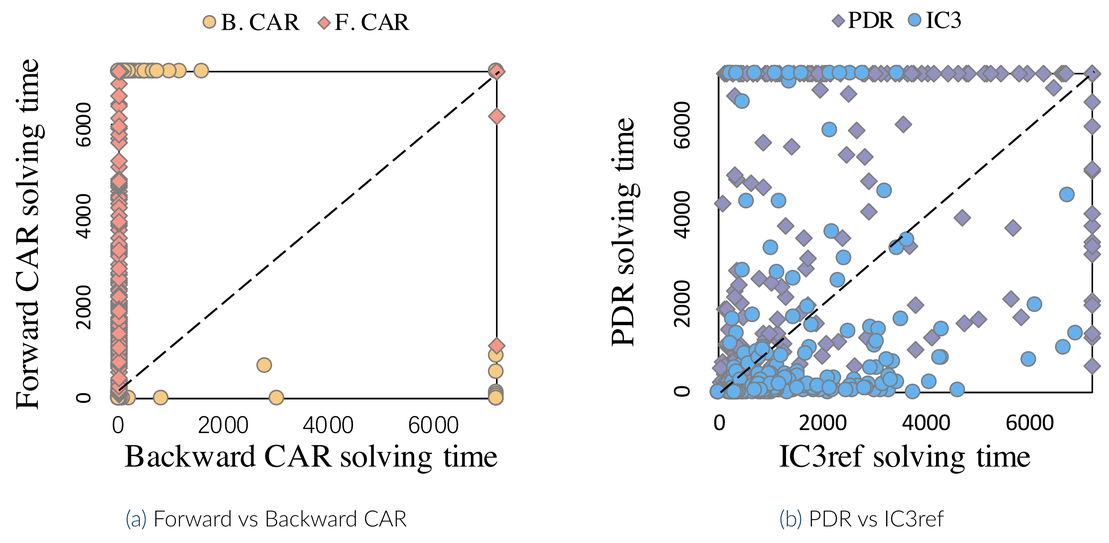

Figure 5. Comparison of the max solving time

## Evaluation: Compactness

Table 1. Difficulty and Size of Top-50 generated

| | AIGROW | | AIGFUZZ | | AIGEN | |
|---|---|---|---|---|---|---|
| | time(s) | size | time(s) | size | time(s) | size |
| PDR | 1532.59 | **219** | 11.96 | 22,790 | 0.79 | 180,120 |
| IC3ref | 371.37 | **224** | 562.75 | 29,036 | 1.81 | 180,186 |
| B. CAR | 775.94 | **51** | 1694.85 | 20,815 | 2.07 | 180,204 |
| F. CAR | 1599.03 | **159** | 1389.48 | 24,570 | 133.34 | 180,214 |

## Evalution: Effectiveness




(a) Forward vs Backward CAR (b) PDR vs IC3ref

Figure 6. Comparison of the solving time of different checkers

## Case Study: From a Tiny Generated Problem to an Algorithmic Breakthrough

Table 2. Detailed runtime comparison on one case, `pdr_101`.

| | Original CAR | CAR-DT |
|---|---|---|
| Runtime | > 3600 s (timeout) | 166 s |
| Final Frame Reached | 275 | 412 (proved safe) |
| Total # of UCs | 79,751 | 57,875 |

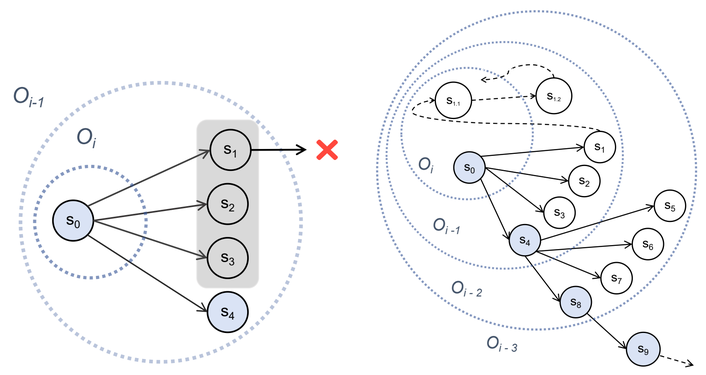



Figure 7. Inefficient patterns recognized.

Table 3. Impact of inspired optimization, *i.e.* CAR-DT

| Benchmark | #(And+Latch) | Original CAR (s) | CAR-DT (s) |
|---|---|---|---|
| `pdr_97` | 40 + 204 | timeout | **130.03** |
| `pdr_103` | 39 + 169 | timeout | **464.58** |
| `fcar_153` | 58 + 180 | 856.23 | **109.16** |
| `ic3ref_487` | 18 + 53 | 2333.53 | **633.04** |
| `ic3ref_560` | 16 + 55 | 3185.52 | **2.15** |

Table 4. Performance comparison on latest HWMCC benchmarks.

| HWMCC'24 Benchmark | Original CAR (s) | CAR-DT (s) |
|---|---|---|
| ILA-Piccolo-BEQ-sanity | > 3600 (timeout) | **3397.90** |
| ILA-Piccolo-BGEU-sanity | > 3600 (timeout) | **3519.51** |
| Problem05-label42+token-ring.08.cil-2 | > 3600 (timeout) | **2883.56** |
| Problem10-label08 | > 3600 (timeout) | **2812.81** |
| a16-p148 | > 3600 (timeout) | **1870.37** |
| brp2.3.prop2-func-interl | > 3600 (timeout) | **1310.36** |
| qspiflash-dualflexpress-divfive-p162 | > 3600 (timeout) | **1156.33** |
| yosyshq-appnote-123-cv32e40x-p500 | > 3600 (timeout) | **2099.85** |
| yosyshq-appnote-123-cv32e40x-p502 | > 3600 (timeout) | **2774.43** |
| yosyshq-appnote-123-veer-axi-p62 | > 3600 (timeout) | **121.03** |