Model Checking Contest

Results for 2014

Fabrice Kordon - LIP6, Univ. P. & M. Curie, France
Hubert Garavel - Inria/Laboratoire d’Informatique de Grenoble, France
Lom-Messan Hillah, LIP6, Univ. P. & M. Curie & Univ. Paris Ouest, France
Alban Linard - LSV, Inria/École Normale Supérieure de Cachan, France
Francis Hulin-Hubard - LSV, CNRS/École Normale Supérieure de Cachan, France

MCC
2014

Model Checking Contest @
Objectives

Evaluation procedure

The models

Participating tools

Analysis of the results

Concluding remarks

Special thanks for those who helped to organize this MCC, in particular Fabrice Legond-Aubry (Nanterre)
Objectives
Lots of questions are raised...

- To verify highly concurrent systems, should we use a symmetry-based or a partial order-based model checker?
- For models with large variable domains, should we use decision diagram-based, or a symmetry-based model checker?
- Can we combine structural reductions techniques with partial-order ones or symmetry-based ones?
- How do tools evolve in the community?
...

A large variety of model checking techniques and their potential combination

A large variety of model categories

A challenge with large scale specifications and variety of models

A need to evaluate in the fairest way current MC implementations
**The Objectives...**

**MCC is intended to:**
- Exchange experience between tool programmers,
- Imagine some association of techniques, and thus better tools
- Stimulate development of tools
- Provide visibility to these tools

**MCC can also be of great help for the PN community (and users):**
- Define a common set of models for benchmarks
- Identify experimentally classes of problems (in models)
  - identify the techniques able to cope with a given class of problems...
- Improve communication between tools (and PNML :-)
- Provides raw data for comparison

**This is the fourth edition**
- Stabilized evaluation procedure (BenchKit) + better reproducibility
- Enriched Benchmark...
- ...still elements to be improved
Evaluation Procedure
What to be measured?

The «enemies» of model checking
- Memory consumption
- CPU consumption

13 classes of «Examinations» to be processed
- State space generation
- Formula evaluation
  - Reachability Formulas (6 subclasses)
  - CTL formulas (3 subclasses)
  - LTL formulas (3 subclasses)

Sub-categories in formulas

<table>
<thead>
<tr>
<th></th>
<th>Bounds</th>
<th>Compute bounds</th>
<th>Deadlock</th>
<th>Cardinality</th>
<th>Fireability simple</th>
<th>Fireability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reachability</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>CTL</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>LTL</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>
### Facts about the MCC

<table>
<thead>
<tr>
<th></th>
<th>Quadhexa-2</th>
<th>Rented machine</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cores</strong></td>
<td>24 @ 2.4GHz</td>
<td>40 @ 2.5GHz</td>
<td></td>
</tr>
<tr>
<td><strong>Memory</strong></td>
<td>128 GB</td>
<td>256 GB</td>
<td></td>
</tr>
<tr>
<td><strong>Used Cores</strong></td>
<td>21</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td><strong>number of runs</strong></td>
<td>48 395</td>
<td>34 913</td>
<td>83 308</td>
</tr>
<tr>
<td><strong>total computation time</strong></td>
<td>157 days 0 hours 10 minutes 51 seconds</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>time spend</strong></td>
<td>8 days 21 hours 50 minutes 29 seconds</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Boot time cumulated</strong></td>
<td>Approx 19 days 4 hours of CPU (for 20s boot-time)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**BenchKit-2**

- Powerful tool to operate a large
- See &ACSD net friday

**A Big Benchmark**
# Facts about the MCC

<table>
<thead>
<tr>
<th></th>
<th>Quadhexa-2</th>
<th>Rented machine</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cores</strong></td>
<td>24 @ 2.4GHz</td>
<td>40 @ 2.5GHz</td>
<td></td>
</tr>
<tr>
<td><strong>Memory</strong></td>
<td>128 GB</td>
<td>256 GB</td>
<td></td>
</tr>
<tr>
<td><strong>Used Cores</strong></td>
<td>21</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td><strong>number of runs</strong></td>
<td>48 395</td>
<td>34 913</td>
<td>83 308</td>
</tr>
<tr>
<td><strong>total computation time</strong></td>
<td>157 days 0 hours 10 minutes 51 seconds</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>time spend</strong></td>
<td>8 days 21 hours 50 minutes 29 seconds</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Boot time</strong></td>
<td>Approx 19 days 4 hours of CPU (for 20s boot-time)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Thanks!**

BenchKit

- Powerful enough to operate a large
- See & ACS Friday

quadhexa-2 lent by
Univ. Paris Ouest Nanterre

A Big Benchmark
# Facts about the MCC

<table>
<thead>
<tr>
<th></th>
<th>Quadhexa-2</th>
<th>Rented machine</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cores</strong></td>
<td>24 @ 2.4GHz</td>
<td>40 @ 2.5GHz</td>
<td></td>
</tr>
<tr>
<td><strong>Memory</strong></td>
<td>128 GB</td>
<td>256 GB</td>
<td></td>
</tr>
<tr>
<td><strong>Used Cores</strong></td>
<td>21</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>number of runs</td>
<td>48 395</td>
<td>34 913</td>
<td>83 308</td>
</tr>
<tr>
<td>total computation time</td>
<td>157 days 0 hours 10 minutes 51 seconds</td>
<td></td>
<td></td>
</tr>
<tr>
<td>time spend</td>
<td>8 days 21 hours 50 minutes 29 seconds</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Boot time: Approx 19 days 4 hours of CPU (for 20s boot-time)

---

Thanks!

BenchKit

- Powerful to operate a large
- See & ACS Friday

server rented thanks to the ImPro ANR project

A Big Benchmark
First of all **correct results !!!**

- How to be sure?
- Compare results (we had time this year)

**Basic algorithm «per line»**

- One line = all results for a model/scale value
- The good one has the majority
- Undecided when no majority

**Comparison between «known» and «scrambled» models**

**Points are counted only if the result is true!**

**Interesting side effect**

- Developers have a detailed feed-back on their tool capabilities and
The Models
**The models**

28 «known» models (7 from 2011 + 12 from 2012 + 9 from 2013)
- 28 presented in the original disk image for submission
- 28 «scrambled» (exactly the same but without tool specific information)

15 new «Surprise models»
- ENS de Cachan
- Inria/LIG
- ST microelectronics & Inria
- Univ. Cottbus
- Univ. Paris & M. Curie
- Univ. Saarland
- Univ. Utha

Each tool confronted to
670 execution
for each examination
Participating Tools
9 tools as for last year

*Only one version per tool*

<table>
<thead>
<tr>
<th>Tool</th>
<th>Origin</th>
<th>VM published</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cunf</td>
<td>ENS de Cachan/ Univ. Oxford (France/UK)</td>
<td>✔</td>
</tr>
<tr>
<td>GreatSPN</td>
<td>Univ. Torino (Italy)</td>
<td>✔</td>
</tr>
<tr>
<td>Helena</td>
<td>Univ. Paris 13 (France)</td>
<td>✔</td>
</tr>
<tr>
<td>IoLA</td>
<td>Univ. Rostock (Germany)</td>
<td>✔</td>
</tr>
<tr>
<td>Marcie</td>
<td>Univ. Cottbus (Germany)</td>
<td>✔</td>
</tr>
<tr>
<td>PNMC</td>
<td>ISAE (Toulouse, France)</td>
<td>✔</td>
</tr>
<tr>
<td>PNXDD</td>
<td>Univ P. &amp; M. Curie (France)</td>
<td>✔</td>
</tr>
<tr>
<td>Stratagem</td>
<td>Univ. Geneva (Switzerland)</td>
<td>✔</td>
</tr>
<tr>
<td>Tapaal</td>
<td>Univ. Aalborg (Denmark)</td>
<td>✔</td>
</tr>
</tbody>
</table>
May not be activated simultaneously

<table>
<thead>
<tr>
<th>Tool</th>
<th>Reported Techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cunf</td>
<td>NET_UNFOLDING SAT_SMT</td>
</tr>
<tr>
<td>GreatSPN</td>
<td>DECISION_DIAGRAMS SYMMETRIES</td>
</tr>
<tr>
<td>Helena</td>
<td>EXPLICIT STUBBORN_SETS</td>
</tr>
<tr>
<td>IoLA</td>
<td>EXPLICIT STATE_COMPRESSION STUBBORN_SETS SYMMETRIES TOPOLOGICAL UNFOLDING_TO_PT</td>
</tr>
<tr>
<td>Marcie</td>
<td>DECISION_DIAGRAMS</td>
</tr>
<tr>
<td>PNMC</td>
<td>DECISION_DIAGRAMS</td>
</tr>
<tr>
<td>PNXDD</td>
<td>DECISION_DIAGRAMS TOPOLOGICAL</td>
</tr>
<tr>
<td>Stratagem</td>
<td>DECISION_DIAGRAMS TOPOLOGICAL</td>
</tr>
<tr>
<td>Tapaal</td>
<td>EXPLICIT STRUCTURAL_REDUCTION</td>
</tr>
</tbody>
</table>
Analysis of the Results
# Considerations About The Outputs from the Executions

**Outputs are HUGE**
- 2.7 Gbyte of csv + ASCII (306+ Kfiles)
- Need for automated analysis

**Rating based on correctness of results**
- Computation of a «result mask»

<table>
<thead>
<tr>
<th>model</th>
<th>class</th>
<th>parameter</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>T5</th>
<th>T6</th>
<th>T7</th>
<th>T8</th>
<th>T9</th>
</tr>
</thead>
<tbody>
<tr>
<td>model1</td>
<td>PT</td>
<td>3</td>
<td>F??T??T??</td>
<td>FX??F??F??</td>
<td>----</td>
<td>--X--T-X-??</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>model1</td>
<td>PT</td>
<td>4</td>
<td>DNF</td>
<td>MOVF</td>
<td>-----</td>
<td>-?--T-??</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>model1</td>
<td>PT</td>
<td>5</td>
<td>CC</td>
<td>MOVF</td>
<td>-----</td>
<td>-------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>model1</td>
<td>PT</td>
<td>6</td>
<td>CC</td>
<td>MOVF</td>
<td>-----</td>
<td>T--T--T--</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>model1</td>
<td>PT</td>
<td>7</td>
<td>CC</td>
<td>DNF</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>model1</td>
<td>PT</td>
<td>8</td>
<td>CC</td>
<td>DNF</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>model2</td>
<td>PT</td>
<td>10</td>
<td>T?????T????</td>
<td>T?????T????</td>
<td>----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>model2</td>
<td>PT</td>
<td>15</td>
<td>T?????T????</td>
<td>T?????T????</td>
<td>----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>model2</td>
<td>PT</td>
<td>20</td>
<td>T?????T????</td>
<td>T?????T????</td>
<td>----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>model2</td>
<td>PT</td>
<td>50</td>
<td>DNF</td>
<td>DNF</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>model2</td>
<td>PT</td>
<td>#</td>
<td>CC</td>
<td>MOVF</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
</tbody>
</table>

- A tool may be wrong between «known» and «scrambled»

- All points (may be honor)
- Only a part of the points
- No point

The HTML report is being mostly generated automatically!
Points and Honor points

Point gained by a tool

- Those for the examinations where
  - You get the good result (even partially)
  - Partially = you get a part of the points ;-)
- The result can be confirmed
  - Your tool is not the only one to compete on this specific one
- No point provided when we cannot state on a given result
  - The algorithm could be improved
  - Consider a «memory» of «right tools»

Honor points

- Those gained when a tool provided a result...
- The result was not confirmed by another tool

Ranking does not include honor points

- These are indicative
Decided before computation was done

- To avoid any influence

Applies when results are OK

- An full successful emanation = 10 pts
  - State space → size = 4pts, others = opts
  - Deadlock Formulas → 5pts per result
  - Other formulas → 10pts per result

- Best tool (goes rarest when model have scaling parameter) — model basis
  - +2

- Last value of scaling parameter for a model — model basis
  - +3

- Fastest on a «line»
  - +1

- Smallest memory foot print on a «line»
  - +1

- Multiplicator coefficient
  - «known» = x1 — «scrambled» = x2 — «surprise» = x3
Las year, too many examination classes
- Not readable in general

2014, covering some use cases
- State Space
  - Gathers StatsSpace
- Reachability
  - Gathers ReachabilityComputeBounds ReachabilityBounds ReachabilityDeadlock ReachabilityCardinality ReachabilityFireabilitySimple ReachabilityFireability
- CTL
  - Gathers CTLCardinality CTLFireabilitySimple CTLFireability
- LTL
  - Gathers LTLCardinality LTLFireabilitySimple LTLFireability

So, 4 trophies
- We will see in fact 😊
### Various information on Colored nets

<table>
<thead>
<tr>
<th></th>
<th>cunf</th>
<th>greatspn</th>
<th>helena</th>
<th>lola</th>
<th>marcie</th>
<th>pnmc</th>
<th>pnxdd</th>
<th>stratagem</th>
<th>tapaal</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of errors</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Alone to compute</td>
<td>0</td>
<td>7</td>
<td>48</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Computed correctly</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>computed / models</td>
<td>0.00%</td>
<td>4.98%</td>
<td>21.99%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>(Colored nets)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>computed correctly /</td>
<td>0.00%</td>
<td>2.08%</td>
<td>2.08%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>models (Colored nets)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>computed alone /</td>
<td>0.00%</td>
<td>2.91%</td>
<td>19.91%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>models (Colored nets)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>computed erroneously /</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>models (Colored nets)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Various information on P/T nets

<table>
<thead>
<tr>
<th></th>
<th>cunf</th>
<th>greatspn</th>
<th>helena</th>
<th>lola</th>
<th>marcie</th>
<th>pnmc</th>
<th>pnxdd</th>
<th>stratagem</th>
<th>tapaal</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of errors</td>
<td>0</td>
<td>54</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Alone to compute</td>
<td>0</td>
<td>44</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>22</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Computed correctly</td>
<td>0</td>
<td>83</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>211</td>
<td>218</td>
<td>191</td>
<td>119</td>
</tr>
<tr>
<td>computed / models</td>
<td>0.00%</td>
<td>43.83%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>52.05%</td>
<td>59.08%</td>
<td>47.22%</td>
<td>28.81%</td>
</tr>
<tr>
<td>(P/T nets)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>computed correctly /</td>
<td>0.00%</td>
<td>20.09%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>51.09%</td>
<td>52.78%</td>
<td>46.25%</td>
<td>28.81%</td>
</tr>
<tr>
<td>models (P/T nets)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>computed alone /</td>
<td>0.00%</td>
<td>10.66%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.97%</td>
<td>5.33%</td>
<td>0.97%</td>
<td>0.00%</td>
</tr>
<tr>
<td>models (P/T nets)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>computed erroneously /</td>
<td>0.00%</td>
<td>13.08%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.97%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>models (P/T nets)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Tool classification for StateSpace examinations

- Marcie
- PNMC
- Tapaal

Computed Correct (P/T)  Computed Correct (Colored)
Reachability examinations

Tool classification for Reachability examinations

Points

1. LoLA
2. Tapaal
3. GreatSPN

18/08/2014, 09:00

Computed Correct (P/T)  Computed Correct (Colored)
Honor points for Reachability examinations

LoLA
Tapaal
GreatSPN

18/08/2014, 09:00
Only honor points!!!

Honor points for CTL examinations

18/08/2014, 09:00

Honor (P/T)  Honor (Colored)
Only one tool participating

Honor points for LTL examinations

points

helena

18/08/2014, 09:00

Honor (P/T)  Honor (Colored)
Globally

These podiums do not show all aspects of tools
- Expressivity
- Formula translation problems

There is probably a lot of useful information in the detailed report
- To appear soon (generator to be enhanced)
- HTML (online) and pdf (CoRR — Cornell Research Repository)

So this should not be considered as an absolute classification!
- See it as outputs for tool developers
- They can focus their effort and evaluate some strategies
Concluding Remarks
On the models
- Huge effort to increase consistency
  - Hubert Garavel + Lom Hillah

On the formulas
- Huge effort to generate better formulas
  - Alban Linard

On the execution environment
- Huge effort to increase «ease to use» and efficiency
  - Fabrice Kordon + Francis Hulin-Hubard

On the results
- Much more efficient analysis
  - 12h down to 35mn from raw data
  - Evaluation of correctness

To come
- Diffusion of results + VM and environment
  - Reproducibility of results if needed (training for next year?)
The team is ready

Challenge, the machines
- Quadhexa from Nanterre
- Rented OVH server? (contract used in 2014 finished)
- BenchKit 2 shows efficiency = gain of time

Challenge, increase detection of «good results»
- Weight tools according to their detected accuracy

Challenge, increase the generation of formulas
- 50 000+ formula needed

And also...
- «live event» (quite complex)

We really need help!
Thank you for your attention

Ready for discussion?