2009 Sixth International Conference on Quantitative Evaluation of Systems

Simulation-Based CTMC Model Checking: An Empirical Evaluation

Joost-Pieter Katoen*† and Ivan S. Zapreev‡
*RWTH Aachen University, D-52056 Aachen, Germany
†University of Twente, 7500 AE Enschede, The Netherlands
‡CWI, 1098 XG Amsterdam, The Netherlands

Abstract

This paper provides an experimental study of the efficiency of simulation-based model-checking algorithms for continuous-time Markov chains by comparing: MRMC – the only tool that implements (new) confidence-interval-based algorithms for verification of all main CSL formulae; Ymer – that allows for verification of time-bounded and time-interval until using sequential acceptance sampling; and VESTA – that can verify time-bounded and unbounded until by means of simple hypothesis testing. The study shows that MRMC provides the most accurate verification results. Ymer and VESTA, unlike MRMC, have almost constant memory consumption. Ymer requires the least number of observations to assess the model-checking problem, but MRMC is mostly the fastest. This indicates that the tools' efficiency does not so much depend on sampling but is rather determined by extra computations.

1. Introduction

The applicability of probabilistic model checking ranges from areas such as randomised distributed algorithms to planning and AI, security [1], and even biological process modelling [2]. Probabilistic model-checking engines have been integrated in existing tool chains for widely used formalisms such as stochastic Petri nets [3], Statemate [4], the stochastic process algebra PEPA [5], and a probabilistic variant of Promela [6]. Popular logics are Probabilistic CTL (PCTL) [7] and Continuous Stochastic Logic (CSL) [8]. At present, there are several model checkers, such as PRISM [9], MRMC [10], VESTA [11], Ymer [12], and APMC [13], that support verification of finite-state continuous-time Markov chains. The typical kind of properties that they can verify are time-bounded until properties–"Does the probability to reach a certain set of goal states (by avoiding bad states) within a maximal time span exceed 0.5?", unbounded until which is similar to the previous one, but with no time bound, and steady-state–"In equilibrium, does the likelihood to leak confidential information remain below 10^{-4}?"

Probabilistic model checking can be done employing numerical or statistical approaches. The former one, is carried out by symbolic and numerical methods. It typically guarantees a high degree of accuracy, but often requires a lot of intricate computations. The latter approach, is based on sampling and Monte Carlo simulation and allows for much simpler algorithms. Being statistical in nature, simulations can not guarantee that the verification result is 100% correct, but the approach allows to bound the probability of generating an incorrect answer.

Like in the traditional setting, probabilistic model checking suffers from the state-space explosion: the number of states grows exponentially in the number of system components and cardinality of data domains. This brings a great deal of inefficiency when using numerical model-checking algorithms. Fortunately, the simulation-based approach often overcomes this problem due to simpler algorithms and "on-the-fly" state-space generation.

In this paper we provide a comparative experimental study of the simulation-based model-checking techniques for CSL. This study (empirically) evaluates three distinct approaches: one, implemented in MRMC, and based on confidence intervals (*c. i.*) [14]; another, realised in Ymer, and based on sequential acceptance sampling [15]; and the third one, supported by VESTA, and based on simple hypothesis testing [16].

Let us note that both PRISM and APMC allow for statistical model checking of DTMCs and CTMCs, based on the theoretical results of [17]. These algorithms allow for model-checking until formulae, by considering the *finite* path prefixes and computing the probability estimates by means of the Chernoff-Hoeffding bounds. The reasons why we did not consider PRISM and APMC in our experiments are as follows: *(i)* when using its simulation engine, PRISM allows to compute the probability estimates, but

978-0-7695-3808-2/09 $25.00 © 2009 IEEE
DOI 10.1109/QEST.2009.25

does not support probability bounds in the formulae; *(ii)* the algorithms implemented in MRMC, although not using Chernoff-Hoeffding bounds, seem to be a generalisation that of [17] for the case of knowing structural properties of the Markov chain.

Our experiments are aimed at the following main points: *(i)* the verification time – the required time to verify a formula; *(ii)* the confidence levels – the match between the theoretically guaranteed confidence and the one obtained in practice; *(iii)* the peak memory usage (VSZ) – the maximal amount of virtual memory (RAM + swap) needed by the tools during the verification; *(iv)* the required number of observations – an indicator of the simulation effort.

For our experiments, we have chosen two case studies (CTMCs): Cyclic Server Polling System (CPS) and Tandem Queueing Network (TQN), also used in [18] for performance evaluation of probabilistic model checkers. The experimental results reported in this paper can be reproduced since the necessary models along with the test scripts are part of the MRMC distribution [10].

The rest of the paper is organised as follows: Section 2 gives a description of the considered case studies and Section 3 provides further details about the employed statistical model checkers and outlines their differences. Further, in Section 4 we discuss and match the tool's parameters, justifying their values selected in Section 5. The latter considers the experimental setup. Section 6 discusses the experimental results and Section 7 concludes.

2. Case Studies

Cyclic Server Polling System (CPS). A cyclic polling system [19] consists of N stations and a server. Each station has a buffer of capacity 1 and the stations are attended by a single server in cyclic order. The server starts by polling the first station. If this station has a message in its buffer, the server serves it. Once the station has been served, or if its buffer was empty, the server moves to the next station cyclically. The polling and service times are exponentially distributed with rates $\gamma = 200$ and $\mu = 1$, respectively. The arrival rate of messages at each station is exponentially distributed with rate $\lambda = \frac{\mu}{N}$. Applications of this case study can be found in e. g. [12], [20], [11], [21].

Tandem Queueing Network (TQN). The Tandem Queueing Network [22] (see also [20], [21], [11]) consists of two queues of capacity N in sequence. Messages arrive at the first queue; when they get served, they are routed to the second queue, from where they leave the system. The message arrivals are exponentially distributed with rate $\lambda = 4 \cdot N$. The server handles messages from the first queue according to a two-phase Coxian [23] distribution. The time between departures from the second queue is exponentially distributed with rate $\kappa = 4$.

3. Tools

MRMC [10] (version 1.4.1, September 2009) is a command-line tool, written in C. The tool implements numerical model-checking techniques for DTMC and CTMC models, and reward extensions thereof. Since $v1.4.1$, it has a full support for the statistical model checking of CSL properties on CTMC models. For time-interval until formulae the tool employs simple terminating simulation. For unbounded-until formulae, the Markov chain model is divided into transient and absorbing states and then the long-run reachability probability is bounded by transient probabilities. For the steady-state formulae the probability is estimated based on steady-state simulation of bottom strongly connected components (BSCCs) and estimates for the probabilities to reach those BSCCs. The tool's distinguishing features are that: *(i)* to verify a formula it estimates its probability using a *c. i.* of desired width and then compares it against the formula's probability bound; *(ii)* MRMC does not employ standard sequential *c. i.* but rather emulates it by gradually increasing the sample size; *(iii)* the tool requires two independent samples when model-checking unbounded-until formulae.

Ymer [12] (version 3.0, February 2005) is a command-line tool, written in C and C++, for verifying transient properties of CTMCs and generalizations. Ymer implements statistical CSL model checking techniques based on discrete event simulation [24] and sequential acceptance sampling [25]. It also incorporates simple acceptance sampling and a numerical engine adopted from PRISM [26]. For time-interval formulae Ymer uses terminating simulations but instead of *c. i.* employs sequential acceptance sampling. The latter minimizes the number of required observations by rejecting/accepting the verified property at early stages, when the simulations show that the formula is clearly satisfied/violated. The procedure has the advantage of requiring fewer observations, on average, than fixed sample size tests, e. g. *c. i.*, for similar levels of accuracy. Ymer has a special option that allows to request probability estimates (Ymer P). In this case, results are computed using sequential confidence-interval based approach [27].

VESTA [11] (version 2.0, 2005) is a Java-based tool for statistical analysis of probabilistic systems. In particular, VESTA allows to verify CSL (PCTL) properties on CTMC (DTMC) models. The tool implements model-checking techniques, based on simple hypothesis testing [28], discussed in [25] and [16]. Simple hypothesis testing is a simplified version of a sequential acceptance sampling that uses fixed sample sizes. For until formulae the tool uses terminating simulations. For unbounded until, a *terminal* state $\perp$ is added to the model. Every state of the original

model is then extended with a transition to this state (taken with some fixed probability $p_\perp$), and the existing transition probabilities are renormalized to form proper probability distributions. This allows to avoid infinite simulation runs, but at the same time requires an extra condition for guarantying confidence levels, see Section 4.

Tool differences: Before we proceed let us overview the differences between the considered tools, and the techniques they implement, and try to forecast their possible influence on the experimental results. *(i)* VESTA is implemented in Java and thus, can be slower than the other tools. Also, its VSZ values should mostly reflect the total memory allocated by JVM. *(ii)* VESTA uses simple hypothesis testing whereas Ymer uses sequential. Therefore, we expect VESTA to be slower than Ymer, since to achieve the same level of confidence, sequential hypothesis testing requires fewer observations than simple hypothesis testing [29]. *(iii)* Ymer and VESTA, unlike MRMC: *(iii.a) have on-the-fly model generation:* MRMC accepts pre-generated CTMCs, and thus the tool's VSZ values should depend on the model size; *(iii.b) can only verify properties in the initial state of the model:* Thus, our results correspond to model checking formulae in the initial state; *(iii.c) do not provide the probability estimates:* The exception is Ymer P which is discussed above.

4. Tool Parameters

For a fair experimental comparison of model-checking algorithms it is vital to have their input parameters matching each other in the best possible way. Further, we consider the main simulation parameters of MRMC, Ymer, and VESTA. We will assume that $\widetilde{p} := \mathit{Prob}\,(s_0,\ \Phi\ \mathcal{U}\ \Psi)$, $\mathit{Prob}\left(s_0,\ \Phi\ \mathcal{U}^{[t_1,t_2]}\ \Psi\right)$, or $\mathit{Prob}^{\infty}\,(s,\ \Psi)$, and b is the probability bound of the formulae, e. g. when we want to verify $\mathcal{P}_{\geq b}\,(\Phi\ \mathcal{U}\ \Psi)$. Note that, since we consider formulae without nested probabilistic operators, the correctness conditions of the algorithms of Ymer and VESTA are the relaxed versions thereof given in [30]. An extended discussion about the tool's simulation parameters can be found in Section 7.1 of [14].

MRMC has two parameters: ξ – the desired confidence of the result; δ' – the upper bound on the width of the considered *c. i. Confidence-level guarantees*: if δ' is such that $\delta' \leq |b - \widetilde{p}|$, then the probability of getting the correct answer to the verified problem is guaranteed to be $\geq \xi$.

Ymer has three parameters: α – the desired probability of the false-positive answer; β – the desired probability of the false-negative answer; δ – the half width of the indifference region. *Confidence-level guarantees*: if δ is such that $\widetilde{p} \notin (b-\delta,\ b+\delta)$, then the probability of getting the correct answer to the model-checking problem is guaranteed to be $1-\alpha$. Here take $\alpha = \beta$ as we do not want do distinguish between false- positive and negative error probabilities.

VESTA inherits the parameters and the error-level guarantees of Ymer. In addition it has two parameters and one condition specific for unbounded-until formulae: $p_\perp > 0$ – the stopping probability; δ_1 – the width of the indifference region for the problem $\mathcal{P}_{=0}\,(\mathcal{A}\,\mathcal{U}\,\mathcal{G})$. *Confidence-level guarantees*: if $p_\perp$ and δ_1 are such that: $\widetilde{p} \notin \left(0,\ \frac{\delta_1}{p_m^{(|S|-1)}\cdot(1-p_\perp)^{(|S|-1)}}\right]$, where p_m is the smallest non-zero transition probability in the model, then the probability of getting the correct answer for the unbounded-until formulae is guaranteed to be $1-\alpha$ (here we take $\alpha = \beta$).

4.1. Relating parameters

To match the parameters of Ymer, VESTA and MRMC, we take $1-\xi = \alpha = \beta$, because we want to have equal bounds on probabilities of having incorrect answers. In addition, we take $\delta' = \delta$ since then fulfilling $\delta' \leq |b - \widetilde{p}|$ is equivalent to choosing δ such that $\widetilde{p} \notin (b-\delta,\ b+\delta)$.

The extra condition of VESTA, required for the unbounded-until operator, does not have analogs in MRMC and Ymer. Therefore, in our experiments we use the default tool values for $p_\perp$ and δ_1. Note that, trying to satisfy this condition can cause serious problems when model checking large models due to the exponentials in the divider of the interval's right border. Moreover, according to [16], the decrease of $p_\perp$ dramatically increases the model-checking times. The same increase of verification time is likely to happen when δ_1 is decreased.

5. Experimental setup

Every experiment, unless stated otherwise, was repeated 100 times. Average verification times (milliseconds) and number of used observations, have logarithmic scale and are based on tool's statistics [1]. Peak memory usage of the tools was collected by sampling process-memory consumption (approximately) every 100 msec. The (actual) confidence levels are computed as the average number of successful model-checking runs on each experiment. The experiments were performed on a cluster-computer node with two 2.33 GHz Intel Dual-Core Xeon processors (64-bit) and 16 GB of RAM (time bounded- and unbounded-until formulae) and an Intel® Core™2 Quad 2.40 GHz processor (64-bit), 8 GB of RAM (steady-state formulae). The operating system was Linux, because it is supported by all the tools. Considering the discussion in Section 4, the main tool parameters were set as follows: $1-\xi = \alpha = \beta = 0.05$, $\delta' = \delta = 0.01$, $p_\perp = 0.01$ and $\delta_1 = 0.1$.

These tool settings are expected to guarantee the 95% accuracy of the verification results. The accuracy can be

1. A minor output modification was introduced into Ymer, see [31].

lower if the conditions specified in Section 4 are violated. Also, when verifying the unbounded-until formulae with VESTA, we use the default tool's settings. For the steady-sate formulae we have chosen the minimal sample size and the sample-size step to be $1,000$. The latter was done because for smaller model sizes we had premature *c. i.* convergence that resulted in low confidence levels.

6. Experimental Results

Note that, Ymer does not support unbounded-until and VESTA cannot verify interval-until properties. Thus, our experimental results do not always include all the tools. MRMC is the only tool that supports verification of the steady-state operator, which can be verified in a pure simulation (*P*) or hybrid (*H*) mode. In the latter case the probabilities of reaching bottom strongly connected components are computed by means of numerical computations. Also, the regeneration method, used in steady-state simulations, can be run in the original setting (*O*), when the regeneration point is chosen arbitrarily, or using the heuristic (*H*), when it is chosen to be the most recurring state in a test run preceding the verification. Moreover, the sample-size step for sequential *c. i.* computations can be chosen to be fixed (*C*) or dynamic (*A*). In the latter case, the tool exponentially increases the sample-size step during the simulations. Therefore, for each steady-state formula, we have MRMC curves with names formed as $MRMC_{TMS}$ where $T \in \{P, H\}$, $M \in \{O, H\}$, and $S \in \{C, A\}$.

For both case studies, each tool and each model size (CPS: $N \geq 15$, TQN $N \geq 511$), the VSZ did not show any significant correlation (not more than a 2% difference) with the verified until formulae. This means, that in case of MRMC, the tool's memory consumption was mostly caused by storing large state spaces in RAM. Also, the memory consumption of Ymer and VESTA were practically constant. Due to these facts, we do not provide the VSZ plots of the until formulae, except for the first one verified on the CPS case study.

6.1. Cyclic Server Polling System (CPS)

For this case study we verified a bounded-until, an interval-until, two unbounded-until, and a steady-state formulae on the models with number of stations N ranging from 3 to 18 and the corresponding state-space sizes ranging from 36 to $7,077,888$. With the increase of N, the numerically-computed probabilities for the considered properties change as follows: for $\mathit{Prob}\left(true\ \mathcal{U}^{[0,80]}\ busy_1\right)$: from 1.0000 to 0.9882; for $\mathit{Prob}\left(true\ \mathcal{U}^{[40,80]}\ serve_1\right)$: from 0.9999 to 0.8944; for $\mathit{Prob}\left(poll_1\ \mathcal{U}\ serve_1\right)$: from 0.0016 to 0.0002; for $\mathit{Prob}\left(\neg serve_2\ \mathcal{U}\ serve_1\right)$: from 0.5213 to 0.5386; for $\mathcal{S}\left(busy_1\right)$: from 0.3481 to 0.1717.

$\mathcal{P}_{\geq 0.95}\left(true\ \mathcal{U}^{[0,80]}\ busy_1\right)$ – the probability that station 1 becomes busy (full) within 80 time units is at least 0.95. With increase of N, all the tools show increase of the model-checking times (cf. Fig. 1) and the number of observations (cf. Fig. 2). This is because: *(a)* $\mathit{Prob}\left(true\ \mathcal{U}^{[0,80]}\ busy_1\right)$ decreases and approaches the probability constraint (0.95); *(b)* the model state space grows, requiring for more and longer simulation paths. For the largest model size ($N = 18$), MRMC uses (respectively) 1.2, 3.6 and 10.2 times more observations than VESTA, Ymer P and Ymer. Yet, MRMC is 1.5, 3.2, and 4.4 times faster than (respectively) Ymer, VESTA, and Ymer P. For Ymer it means that either the tool does not have a sufficiently efficient implementation or that sampling does not have a significant impact on verification times, when compared to the effort needed for, e. g., performing hypothesis testing. Still, the verification times of MRMC are growing faster than that of the other tools.

According to Fig. 3, Ymer and VESTA use constant

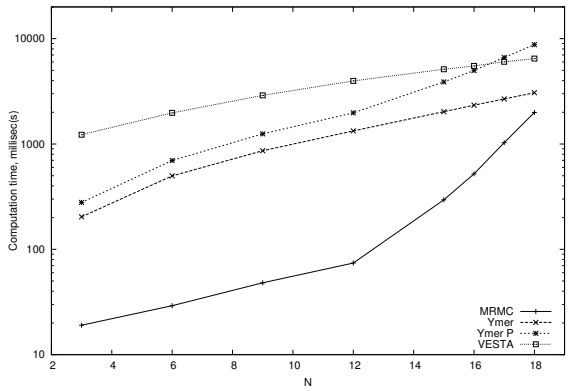

Fig. 1: CPS : $\mathcal{P}_{\geq 0.95}\left(true\ \mathcal{U}^{[0,80]}\ busy_1\right)$ (time)

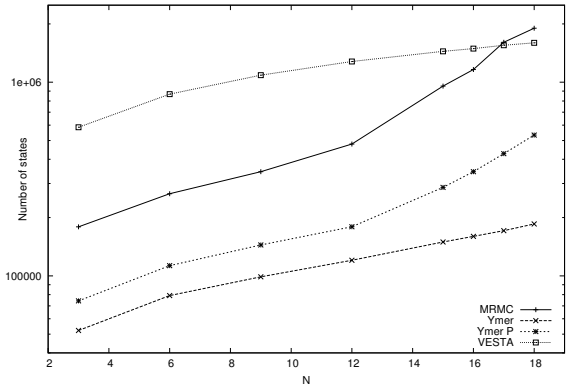

Fig. 2: CPS : $\mathcal{P}_{\geq 0.95}\left(true\ \mathcal{U}^{[0,80]}\ busy_1\right)$ (# observations)

memory and the VSZ of MRMC, as predicted, grows with the model size. This implies that, since both Ymer and VESTA do not generate the model's state space, the memory consumption for sampling is insignificant.

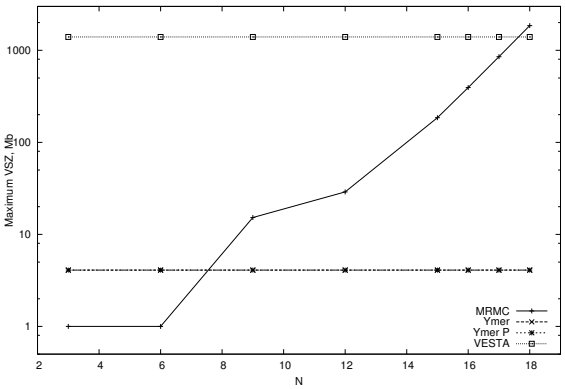

Fig. 3: CPS: $\mathcal{P}_{\geq 0.95}\left(true\ \mathcal{U}^{[0,80]}\ busy_1\right)$ (VSZ)

The large memory usage of VESTA is dominated by the amount of memory acquired by the JVM.

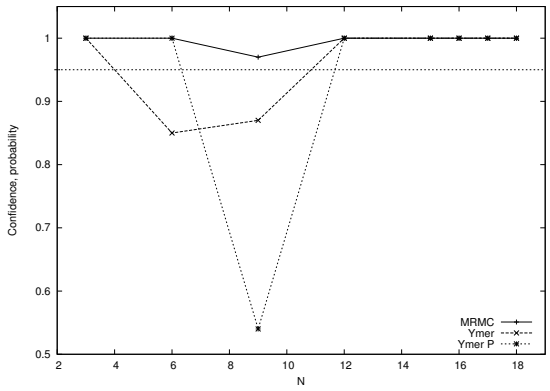

Fig. 4: CPS: $\mathcal{P}_{\geq 0.99}\left(true\ \mathcal{U}^{[40,80]}\ serve_1\right)$ (confidence)

$\mathcal{P}_{\geq 0.99}\left(true\ \mathcal{U}^{[40,80]}\ serve_1\right)$ – the probability that station 1 is served within the time interval $[40,\ 80]$ is at least 0.99. The confidence levels for $N \in \{6,\ 9\}$ (cf.

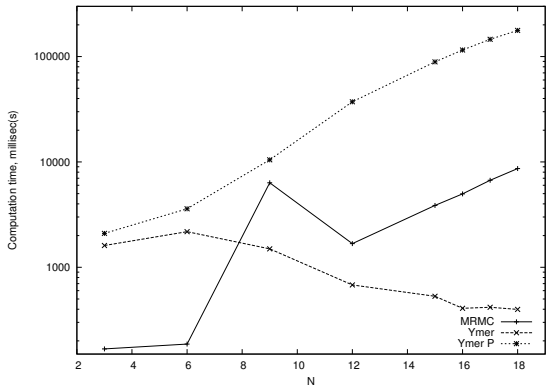

Fig. 5: CPS : $\mathcal{P}_{\geq 0.99}\left(true\ \mathcal{U}^{[40,80]}\ serve_1\right)$ (time)

Fig. 4) are compromised, especially in case of Ymer and Ymer P. This happens because the corresponding probabilities *Prob* $\left(true\ \mathcal{U}^{[40,80]}\ serve_1\right)$ are 0.9988 and 0.9888, i. e., they fall in the indifference region. Moreover, the condition $\delta' \leq |b - \widetilde{p}|$, required by MRMC for ensuring the 95% confidence, is also violated. MRMC provides more accurate answers as: *(i)* the specified confidence

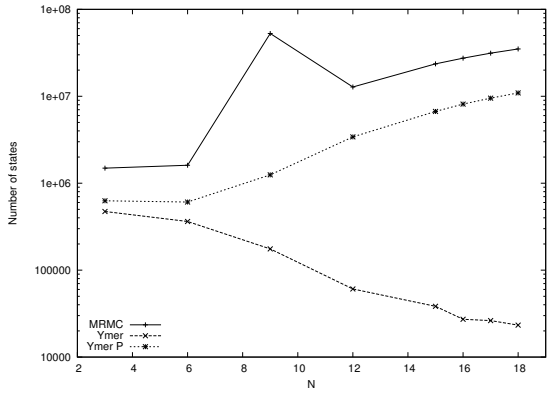

Fig. 6: CPS : $\mathcal{P}_{\geq 0.99}\left(true\ \mathcal{U}^{[40,80]}\ serve_1\right)$ (# observations)

($\xi = 0.95$) only defines the lower bound on the actual confidence; *(ii)* the tool uses the Agresti-Coull *c. i.* that is known to have a coverage probability that exceeds the specified confidence; *(iii)* the tool first simulates until the *c. i.* is tighter than δ' and then until it reaches the definite answer to the problem. The latter improves the resulting confidence by considering more observations, cf. Fig. 6.

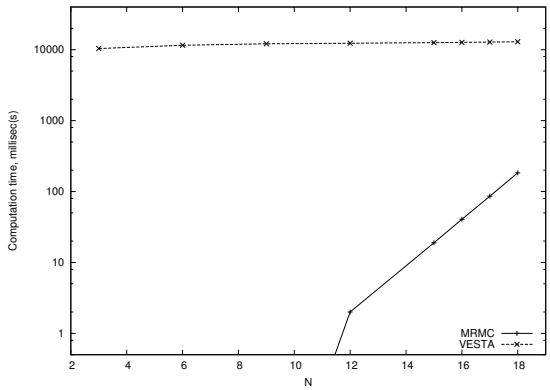

Fig. 7: CPS: $\mathcal{P}_{\geq 0.2}\left(poll_1\ \mathcal{U}\ serve_1\right)$ (time)

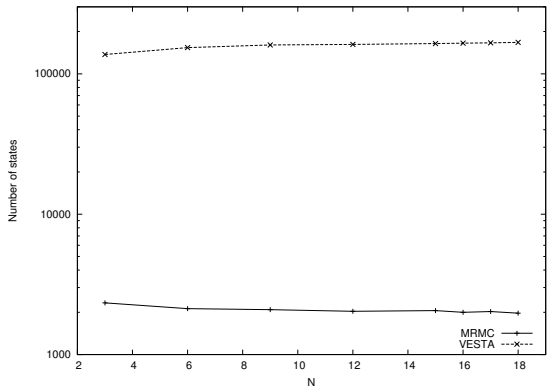

Fig. 8: CPS: $\mathcal{P}_{\geq 0.2}\left(poll_1\ \mathcal{U}\ serve_1\right)$ (# observations)

The model-checking times (cf. Fig. 5) and the number of observations (cf. Fig. 6) indicate that the accuracy of MRMC comes at a price, as witnessed by the peaks for $N = 9$. In general ($N = 18$), MRMC is up to 8 times faster than Ymer P, but is up to 21.7 times slower than

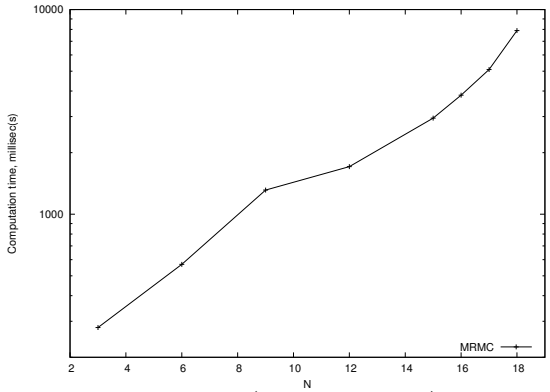

Fig. 9: CPS: $\mathcal{P}_{\geq 0.5}\left(\neg serve_2\ \mathcal{U}\ serve_1\right)$ (time)

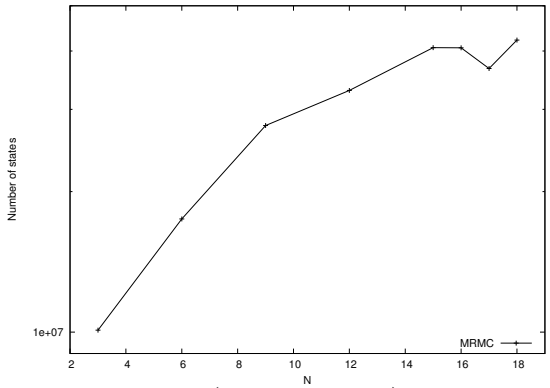

Fig. 10: CPS: $\mathcal{P}_{\geq 0.5}\left(\neg serve_2\ \mathcal{U}\ serve_1\right)$ (# observations)

Ymer. The performance of the latter one is improving with the growth of N. The reason for that is likely to be the rapid increase of distance between the values of *Prob* $\left(true\ \mathcal{U}^{[40,80]}\ serve_1\right)$ and the probability bound of the formula. In this case, Ymer P and MRMC continue simulations until they reach the *c. i.* of the desired width but Ymer, that uses sequential hypothesis testing, does not need that, so it stops much earlier.

$\mathcal{P}_{\geq 0.2}\left(poll_1\ \mathcal{U}\ serve_1\right)$ – the probability that station 1 is served after being polled is at least 0.2. Both MRMC and VESTA showed 100% accuracy when model checking this property. The performance results given in Fig. 7 indicate that the time required by VESTA is almost constant for all model sizes and for MRMC it is insignificantly small in the beginning (up to $N = 12$) and then starts growing. This contradicts to Fig. 8. One can notice that the number of observations required by VESTA is growing whereas for MRMC it is decreasing. Putting these facts together, we conclude that the increase of verification time for MRMC might be caused by: *(i)* the effort required for traversing the large (up to about $7 \cdot 10^6$ states) Markov chain stored in RAM; *(ii)* the need to search for BSCCs. Still, MRMC is at least 10 times faster than VESTA.

$\mathcal{P}_{\geq 0.5}\left(\neg serve_2\ \mathcal{U}\ serve_1\right)$ – the probability that station 1 is served before station 2 is at least 0.5. Fig. 9 provides the model-checking times for MRMC which again showed $> 95\%$ accuracy. The plots for VESTA are not present because it did not terminate within the 15 minutes time-out (compared to seconds required by MRMC). Fig. 10 shows the number of required observations. Notice that, there is a significant drop for $N = 17$ and also the values for $N = 15$ and 16 are almost equal. At the same time, the model-checking times for these values of N show a stable and continuous increase. This strengthens our belief in that supplementary computations, such as traversal through a large pre-generated Markov chain, stored in RAM, give a much stronger influence on the model-checking time than the increase in the number of required observations.

$\mathcal{S}_{>0.19}\left(busy_1\right)$ – the steady-state probability of station 1 being busy is greater than 0.19. Fig. 11 provides the

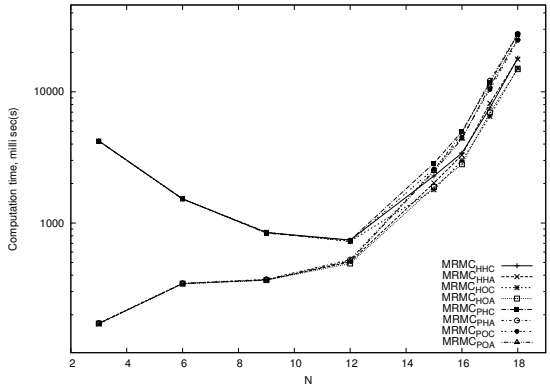

Fig. 11: CPS : $\mathcal{S}_{>0.19}\left(busy_1\right)$ (time)

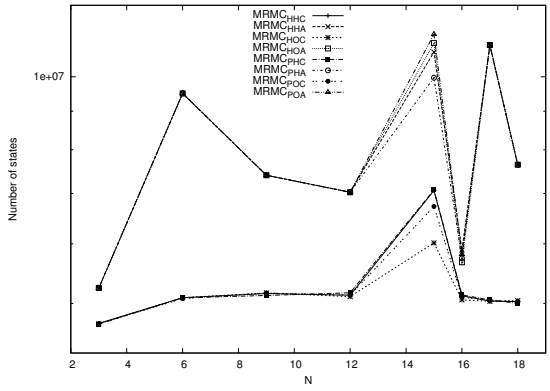

Fig. 12: CPS : $\mathcal{S}_{>0.19}\left(busy_1\right)$ (# observations)

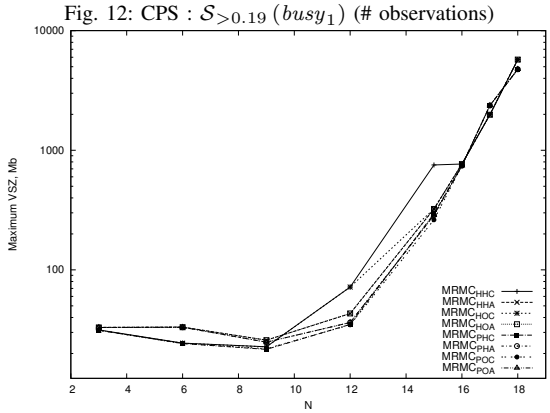

Fig. 13: CPS: $\mathcal{S}_{>0.19}\left(busy_1\right)$ (VSZ)

model-checking times for MRMC which showed 100% accuracy. Notice that up to $N = 12$ there is a significant difference between MRMC runs with the constant sample-size step, the upper bunch of curves, and the dynamic step, the lower bunch. Also, the latter ones are using significantly more observations (cf. Fig. 12), from which we conclude that the re-computation of *c. i.*, on small models, can significantly slow down model checking. Besides, for $N = 15, 16$ the estimated probabilities fall into the indifference interval. The corresponding peaks are especially distinctive for the $MRMC_{**C}$ curves of Fig. 12. Yet, this does not have any drastic affect on the corresponding model-checking times. This is because for $N >= 12$ the effort needed for additional computations still exceeds the effort required for simulating a large model. The memory consumption curves, cf. Fig. 13, are all showing similar behavior. Yet, memory required to store samples, starts playing an important role. Notice that, the VSZ values of MRMC are significantly higher (up to 3.1 times for $N = 18$) than in Fig. 3.

6.2. Tandem Queueing Network (TQN)

Here we verified two bounded-until, one interval-until, one unbounded-until, and one steady-state formulae on the models with the queue capacities N ranging from 2 to 1023 and the corresponding state-space sizes are ranging from 15 to $2,096,128$. With the increase of N, the numerically-computed probabilities for the considered properties change as follows: for *Prob* $\left(true\ \mathcal{U}^{[0,2]}\ full\right)$: from 0.0262 to 0.0000; for *Prob* $\left(true\ \mathcal{U}^{[0.5,2]}\ full\right)$: from 0.0225 to 0.0000; for *Prob* $\left(true\ \mathcal{U}^{[0,10]}\ full_1\right)$: it is constantly 1.0000; for *Prob* $(\neg full_1\ \mathcal{U}\ full_2)$: from 0.0177 to 0.0000; for $\mathcal{S}\ (full_1)$: from 0.8032 to 0.9995. Since the value of N is changed in a non-linear manner, the horizontal axis of the plots given in this section is *logarithmic*.

$\mathcal{P}_{\leq 0.01}\left(true\ \mathcal{U}^{[0,2]}\ full\right)$ – the probability that both queues become full within 2 time units is at most 0.01. There are no results for Ymer P because it was not terminating within the 15 minutes time-out. The confidence estimates in Fig. 14 exhibit a slight decrease of confidence for Ymer and VESTA at $N = 2$. This is due to the fact that in this case *Prob* $\left(true\ \mathcal{U}^{[0,2]}\ full\right) = 0.0262$ is relatively close to the probability bound. Still, the confidence levels stay above the theoretically predicted one. As before, MRMC is generally faster than the other tools (cf. Fig. 15) but levels out with Ymer at $N = 1023$. Also, its model checking times grow faster that that of the other tools. The peaks in MRMC plots for $N = 2$ (see also Fig. 16) is the price it pays for being 100% accurate. Notice that, for $N \geq 10$ the number of observations grows uniformly for all tools. E. g., MRMC requires from 3 to 3.2 times more observations than Ymer, and VESTA needs about

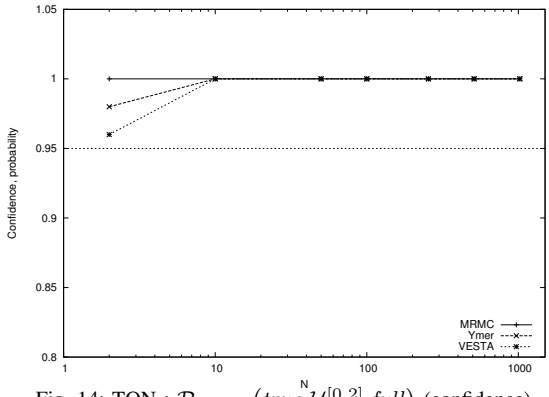

Fig. 14: TQN : $\mathcal{P}_{\leq 0.01}\left(true\ \mathcal{U}^{[0,2]}\ full\right)$ (confidence)

12% more samples than MRMC. The verification times show a different behavior. VESTA has the slowest increase of time, Ymer's times grow a bit faster, and MRMC has the fastest time increase. In the worst case ($N = 1023$), Ymer is only 2.3 times faster than VESTA, and about 7% faster than MRMC. Considering the corresponding increase in the number of observations, this might mean that the Ymer's implementation is either not very efficient or that sampling does not have a sufficient effect on model checking times, compared to supplementary computations. MRMC most likely suffers from the need to store and traverse the complete CTMC.

$\mathcal{P}_{\leq 0.1}\left(true\ \mathcal{U}^{[0.5,2]}\ full\right)$ – the probability that both queues become full within time interval $[0.5,\ 2]$ is at most 0.1. For this property all the tools showed 100% accuracy. Once again, Ymer P was not able to finish verification within 15 minutes. The performance results given in Fig. 17 and 18 show the behavior similar to the one for $\mathcal{P}_{\leq 0.01}\left(true\ \mathcal{U}^{[0,2]}\ full\right)$.

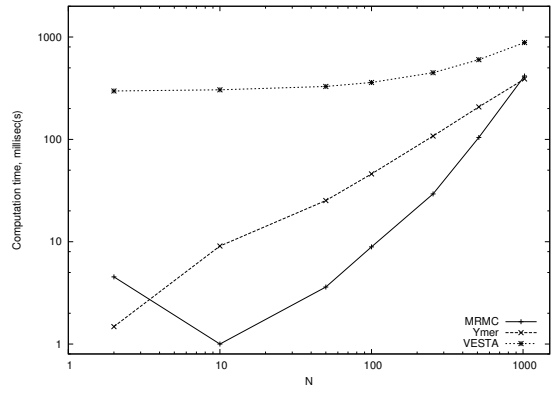

Fig. 15: TQN: $\mathcal{P}_{\leq 0.01}\left(true\ \mathcal{U}^{[0,2]}\ full\right)$ (time)

$\mathcal{P}_{\leq 0.98}\left(true\ \mathcal{U}^{[0,10]}\ full_1\right)$ – the probability that the first queue becomes full within 10 time units is at most 0.98. In this case Ymer P successfully verified the formula, and all the tools were 100% accurate. The performance displayed in Fig. 19 and Fig. 20, are similar to the ones for the previous two properties.

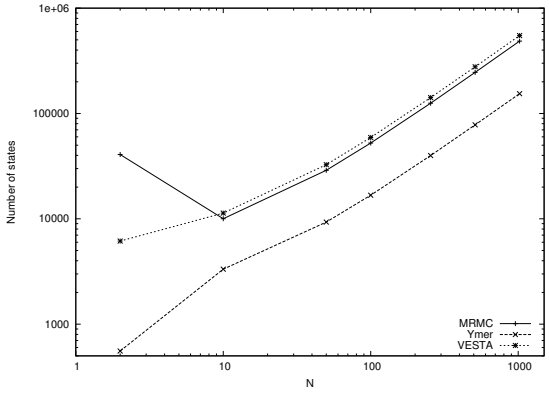

Fig. 16: TQN: $\mathcal{P}_{\leq 0.01}\left(true\ \mathcal{U}^{[0,2]}\ full\right)$ (# observations)

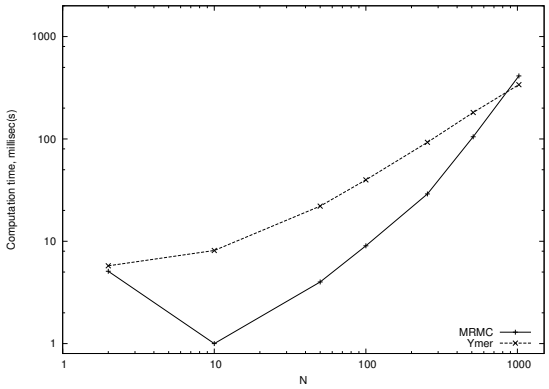

Fig. 17: TQN: $\mathcal{P}_{\leq 0.1}\left(true\ \mathcal{U}^{[0.5,2]}\ full\right)$ (time)

$\mathcal{P}_{\leq 0.03}\left(\neg full_1\ \mathcal{U}\ full_2\right)$ – the probability that the second queue becomes full before the first queue is at most 0.03. Both, VESTA and MRMC were completely accurate in their model-checking results. The verification times and the number of observations in Fig. 21 and 22 reflect that, since $\mathit{Prob}\left(\neg full_1\ \mathcal{U}\ full_2\right) = 0.000$ for all $N \geq 10$, VESTA needs an almost constant amount of observations to decide on the property. This can be because it uses hypothesis testing and that the distance between the probability bound 0.03 and the true value of $\mathit{Prob}\left(\neg full_1\ \mathcal{U}\ full_2\right)$ stays constant. Still, MRMC is at

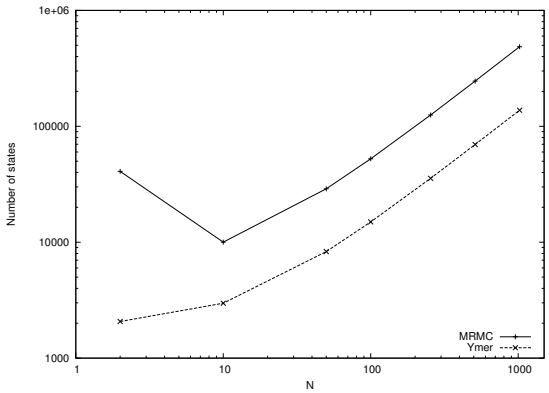

Fig. 18: TQN: $\mathcal{P}_{\leq 0.1}\left(true\ \mathcal{U}^{[0.5,2]}\ full\right)$ (# observations)

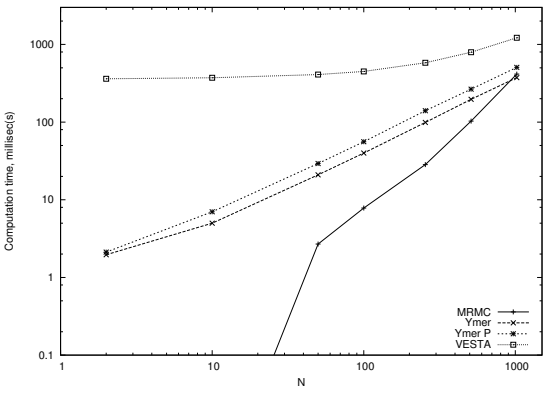

Fig. 19: TQN: $\mathcal{P}_{\leq 0.98}\left(true\ \mathcal{U}^{[0,10]}\ full_1\right)$ (time)

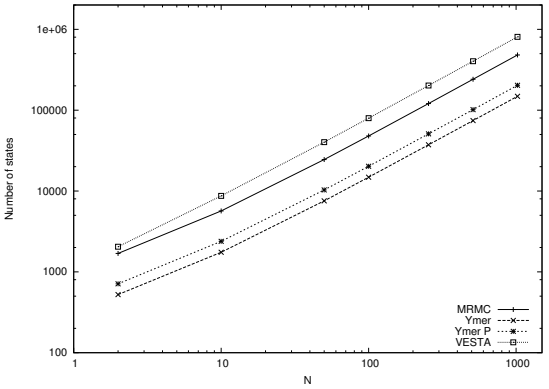

Fig. 20: TQN: $\mathcal{P}_{\leq 0.98}\left(true\ \mathcal{U}^{[0,10]}\ full_1\right)$ (# observations)

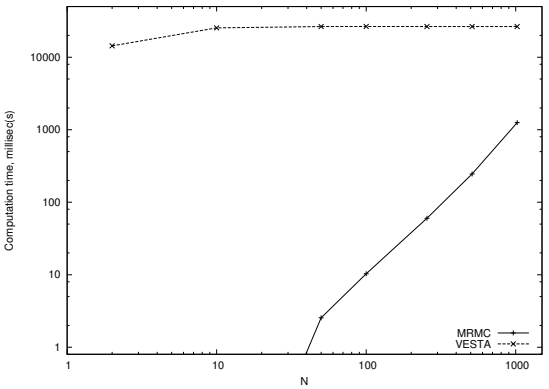

Fig. 21: TQN: $\mathcal{P}_{\leq 0.03}\left(\neg full_1\ \mathcal{U}\ full_2\right)$ (time)

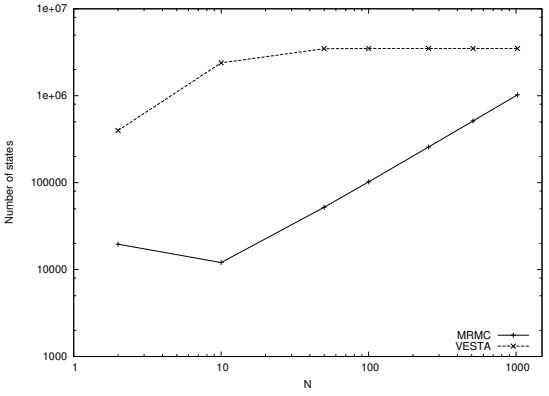

Fig. 22: TQN: $\mathcal{P}_{\leq 0.03}\left(\neg full_1\ \mathcal{U}\ full_2\right)$ (# observations)

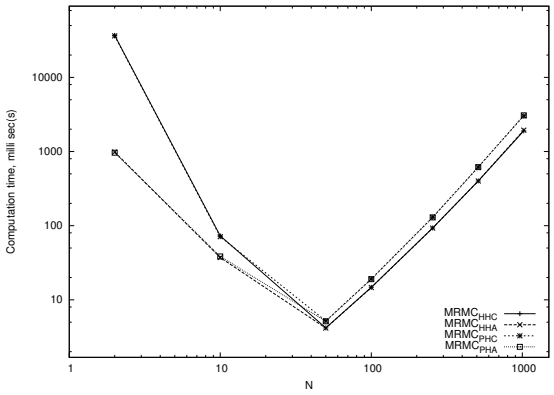

Fig. 23: TQN: $\mathcal{S}_{>0.999}\left(full_1\right)$ (time)

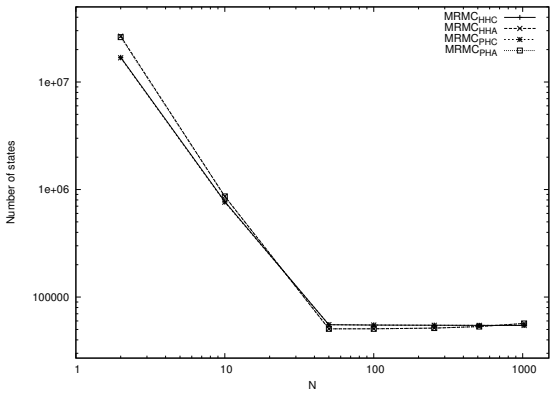

Fig. 24: TQN: $\mathcal{S}_{>0.999}\left(full_1\right)$ (# observations)

least 6 times faster than VESTA.

$\mathcal{S}_{>0.999}\left(full_1\right)$ – the steady-state probability of the first queue being full is greater than 0.999. For this property we set the width of the indifference region to be 0.0003. Although MRMC was 100% accurate, we should note that in case of $N = 511$ the estimated probability falls in the indifference region. Also, using the pure regeneration method, without the heuristic for choosing the regeneration point, failed. MRMC was unable to finish simulations within 15 minutes timeout. The reason for that is that the TQN's model is an ergodic Markov chain. The latter, especially for larger models, causes most of the regeneration cycles to be enormously large. Once again, cf. Fig. 23, we see that for smaller model sizes ($N < 50$) having a dynamic sample-size increase saves a lot of effort needed for re-computation of the *c. i.* For $N \geq 50$ the pure simulation method requires more time. This is because, although for an ergodic CTMC there is no need to compute reachability probabilities, the pure and hybrid simulation methods have two different implementations and the former has a higher complexity. The required observation in Fig. 24 show that MRMC with the dynamic sample-size increase needs more observations for $N < 50$, and for $N \geq 50$ all curves exhibit comparable behavior. At the moment, we do not have any good explanation for the decrease in the number of needed observations and the

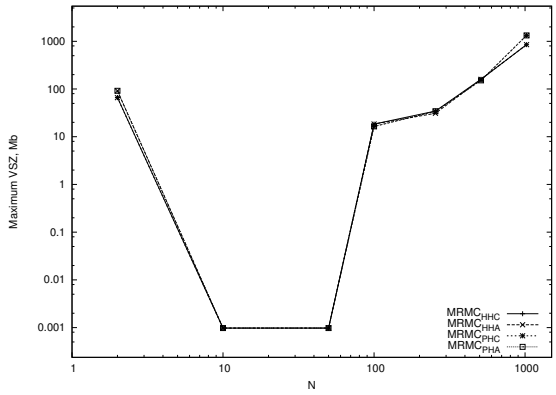

Fig. 25: TQN: $\mathcal{S}_{>0.999}\left(full_1\right)$ (VSZ)

almost constant values for $N \geq 50$. Note that, verifying the given steady-state property in the worst case ($N = 1023$, $MRMC_{PHA}$) requires about 4.2 times more memory (cf. Fig. 25[2]) than verifying, e. g., $\mathcal{P}_{\leq 0.01}\left(true\ \mathcal{U}^{[0,2]}\ full\right)$ for $N = 1023$.

7. Conclusions

Our analysis showed that for until formulae the peak-memory consumption (VSZ) of MRMC grows in accordance with the growth of the model sizes. This is due to using the pre-generated Markov chain, as opposed to the on-demand state-space generation implemented in Ymer and VESTA. The latter tools show (almost) constant memory consumption. For the steady-state operator the situation is different. When model checked with MRMC, for the same model size, VSZ values can be up to 4.5 (TQN, $N = 1023$) times larger than the ones for the until formula. This means that memory needed for storing sampled data is almost negligible when verifying until, and is significant when verifying steady-state formulae.

The actual confidence levels of all tools were within theoretically predicted bounds. At the same time MRMC showed high accuracy even in cases when the sufficient conditions for providing these bounds were violated.

Ymer P and VESTA were not always able to provide model checking times within the 15 minutes time out. In all other cases, the model-checking times for all the tools were within seconds. The exception is Ymer P, cf. Fig. 5. On the considered models, verification times of MRMC were mostly several times (up to 10) smaller than that of Ymer and VESTA, but the performance of MRMC was rapidly decreasing with the growth of the model sizes[3]. This might be because, e. g., generating random paths through a large Markov chain requires addressing far distant blocks of RAM. Another observation is that, for steady-state simulations on smaller models ($N \leq 12$ for CPS, and $N \leq 511$

2. $N = 10, 50$: An inadequate statistics due to small verification times.
3. For larger model sizes, the trend is expected to persist.

for TQN) computation of confidence intervals requires much more effort than doing sampling. To conclude, we must admit that Ymer showed an excellent performance on larger models, where in one case it was 21.7 times faster than MRMC (cf. Fig. 5). Also, Ymer always needed fewer observations to provide correct model-checking results than other tools. This means that its algorithms are more efficient from the simulation point of view. Considering its performance on smaller models, we must conclude that either its implementation is not very efficient or that the sampling effort does not play a significant role when compared to supplementary computations. Last, but not least is VESTA which, considering that it is implemented in Java, showed a reasonably good performance. The tool typically required more observations, but, with the growth of the model sizes, the increase in their numbers was not as significant as in case of MRMC. In general, to have an efficient simulation-based algorithm and its implementation we suggest to: *(i)* use structural information of the Markov chain; *(ii)* use sequential hypothesis testing or confidence intervals; *(iii)* employ on-the-fly model generation; *(iv)* remember that the low simulation effort can be suppressed by the algorithm's supplementary computations.

References

[1] G. Norman and V. Shmatikov, "Analysis of probabilistic contract signing," *Journal of Computer Security*, vol. 14, no. 6, pp. 561–589, 2006.

[2] P. Lecca and C. Priami, "Cell cycle control in eukaryotes: A BioSpi model," Informatica e Telecommunicazioni: University of Trento, Tech. Rep. DIT-03-045, 2003.

[3] D. D'Aprile, S. Donatelli, and J. Sproston, "CSL Model Checking for the GreatSPN Tool," in *Computer and Information Sciences*, ser. LNCS, vol. 3280. Springer, 2004, pp. 543–553.

[4] E. Bode, M. Herbstritt, H. Hermanns, S. Johr, T. Peikenkamp, R. Pulungan, R. Wimmer, and B. Becker, "Compositional Performability Evaluation for STATEMATE," in *Quantitative Evaluation of Systems (QEST)*. IEEE Computer Society, 2006, pp. 167–178.

[5] J. Hillston, *A Compositional Approach to Performance Modelling*, ser. Distinguished Dissertations Series. New York, NY, USA: Cambridge University Press, 1996.

[6] C. Baier, F. Ciesinski, and M. Größer, "ProbMela and verification of Markov decision processes," *ACM SIGMETRICS Performance Evaluation Review*, vol. 32, no. 4, pp. 22–27, 2005.

[7] H. Hannsson and B. Jonsson, "A logic for reasoning about time and reliability," *Formal Aspects of Computing*, vol. 6, no. 5, pp. 512–535, 1994.

[8] C. Baier, B. Haverkort, H. Hermanns, and J.-P. Katoen, "Model-Checking Algorithms for Continuous-Time Markov Chains," *IEEE Transactions on Software Engineering*, vol. 29, no. 6, pp. 524–541, 2003.

[9] M. Kwiatkowska, G. Norman, and D. Parker, "PRISM: Probabilistic Model Checking for Performance and Reliability Analysis," *ACM SIGMETRICS Performance Evaluation Review*, vol. 36, no. 4, pp. 40–45, 2009.

[10] J.-P. Katoen, I. S. Zapreev, E. M. Hahn, H. Hermanns, and D. N. Jansen, "The Ins and Outs of The Probabilistic Model Checker MRMC," in *Quantitative Evaluation of Systems (QEST)*. IEEE Computer Society, 2009, www.mrmc-tool.org.

[11] K. Sen, M. Viswanathan, and G. Agha, "Statistical Model Checking of Black-Box Probabilistic Systems," in *Computer Aided Verification (CAV)*, ser. LNCS, vol. 3114. Springer, 2004, pp. 202–215.

[12] H. Younes, "Ymer: A Statistical Model Checker," in *Computer Aided Verification (CAV)*, ser. LNCS, vol. 3576. Springer, 2005, pp. 429–433.

[13] R. Lassaigne and S. Peyronnet, "Approximate Verification of Probabilistic Systems," in *Process Algebra and Probabilistic Methods, Performance Modeling and Verification (PAPM/PROBMIV)*, ser. LNCS, vol. 2399. Springer, 2002, pp. 213–214.

[14] I. S. Zapreev, "Model Checking Markov Chains: Techniques and Tools," Ph.D. dissertation, University of Twente, Enschede, The Netherlands, 2008, http://doc.utwente.nl/58974/1/thesis_Zapreev.pdf.

[15] H. Younes and R. Simmons, "Statistical Probabilistic Model Checking with a Focus on Time-Bounded Properties," *Information and Computation*, vol. 204, no. 9, pp. 1368–1409, 2006.

[16] K. Sen, M. Viswanathan, and G. Agha, "On Statistical Model Checking of Stochastic Systems," in *Computer Aided Verification (CAV)*, ser. LNCS, vol. 3576. Springer, 2005, pp. 266–280.

[17] T. Hérault, R. Lassaigne, F. Magniette, and S. Peyronnet, "Approximate Probabilistic Model Checking," in *Verification, Model Checking, and Abstract Interpretation (VMCAI'04)*, ser. LNCS, vol. 2937. Springer-Verlag, 2004, pp. 73–84.

[18] D. N. Jansen, J.-P. Katoen, M. Oldenkamp, M. Stoelinga, and I. S. Zapreev, "How Fast and Fat Is Your Probabilistic Model Checker?" in *Haifa Verification Conference (HVC)*, ser. LNCS, vol. 4899. Springer, 2008, pp. 65 – 79.

[19] O. C. Ibe and K. S. Trivedi, "Stochastic Petri Net Models of Polling Systems," *Selected Areas in Communications*, vol. 8, no. 9, pp. 1649–1657, 1990.

[20] H. Hermanns, J.-P. Katoen, J. Meyer-Kayser, and M. Siegle, "A Markov Chain Model Checker," in *Tools and Algorithms for the Construction and Analysis of Systems (TACAS)*, ser. LNCS, vol. 1785. Springer, 2000, pp. 347–362.

[21] H. Younes, M. Kwiatkowska, G. Norman, and D. Parker, "Numerical vs. Statistical Probabilistic Model Checking," *Software Tools for Technology Transfer (STTT)*, vol. 8, no. 3, pp. 216–228, 2006.

[22] H. Hermanns, J. Meyer-Kayser, and M. Siegle, "Multi Terminal Binary Decision Diagrams to Represent and Analyse Continuous Time Markov Chains," in *Numerical Solutions of Markov Chains*. Prensas Universitarias, 1999, pp. 188–207.

[23] D. R. Cox, "A use of complex probabilities in the theory of stochastic processes," in *Cambridge Philosophical Society*, vol. 51, 1955, pp. 313–319.

[24] G. S. Shedler, *Regenerative Stochastic Simulation*. London, UK: Academic Press, 1993.

[25] H. Younes and R. Simmons, "Probabilistic Verification of Discrete Event Systems using Acceptance Sampling," in *Computer Aided Verification (CAV)*, ser. LNCS, vol. 2404. Springer, 2002, pp. 223–235.

[26] M. Kwiatkowska, G. Norman, and D. Parker, "PRISM: Probabilistic Symbolic Model Checker," in *Modelling Techniques and Tools for Computer Performance Evaluation (TOOLS)*, ser. LNCS, vol. 2324. Springer, 2002, pp. 200–204.

[27] A. Nadas, "An extension of a theorem of Chow and Robbins on sequential confidence intervals for the mean," *Annals of Mathematical Statistics*, vol. 40, no. 2, pp. 667–671, 1969.

[28] R. V. Hogg and A. T. Craig, *Introduction to Mathematical Statistics*, 4th ed. New York, NY, USA: MacMillan, 1978.

[29] A.Wald and J. Wolfowitz, "Optimum character of the sequential probability ratio test," *Annals of Mathematical Statistics*, vol. 19, pp. 326–339, 1948.

[30] K. Sen, M. Viswanathan, and G. Agha, "VESTA: A Statistical Model-checker and Analyzer for Probabilistic Systems," in *Quantitative Evaluation of Systems (QEST)*. IEEE Computer Society, 2005, pp. 251–252.

[31] I. S. Zapreev and C. Jansen, "MRMC: Test-suite manual," http://www.mrmc-tool.org/trac/wiki/Specifications, 2008.