Report on "Analytical Approximations in Probabilistic Analysis of Real Time System"

Group L – Moiz Ahmed Mateen Ahmed, Narender Krishna Rapolu, Shijie Zhou
St. Francis Xavier University

Note on Authors – This is a report for the Real Time Systems Course, the report summarizes the finding of the original authors Markovic et. Al from Malardalen University and Max Planck Institute for Software Systems.

I. INTRODUCTION

The paper focuses on the challenges in analyzing hard real-time systems, specifically on the evolution of bound-based analysis. Real-time systems often exhibit execution times lower than estimated worst-case scenarios, resulting in pessimistic resource provisioning. To address this, researchers have worked on analytical bounds for deadline miss probability using probabilistic inequalities, such as Hoeffding, Bernstein, and Chernoff bounds.

Three key problems outlined in this paper are:

1) Problem 1 - "Efficiently and accurately deriving an upper bound on the probability that a distribution (e.g., execution workload) exceeds a given value (e.g., a deadline)."
2) Problem 2 - "Efficiently, and accurately approximating a probability distribution (e.g., of an execution workload or a response-time) whose exact computation is intractable."
3) Problem 3 - "Efficiently and accurately deriving the least value x (e.g., a time point within some interval) such that the probability that a distribution (e.g., execution workload) precedes x is greater than or equal to some probability threshold p."

These problems are relevant in various areas, including probabilistic cache and Worst-Case Execution Time (WCET) analysis. Solving these problems efficiently and accurately would lead to significant improvements in the analysis and control of real-time systems.

II. CONTRIBUTIONS OF THIS PAPER AND RELATED WORK

This paper proposes and investigates the effectiveness of solutions to Problems 1, 2, and 3 using the Lyapunov Central Limit Theorem (CLT). It adjusts and improves bounds for the probabilistic analysis of real-time systems, with the main contributions being:

1) Proving safe bounds on the asymptotic behavior of the probability distribution of preemptive independent tasks.
2) Adjusting the Berry-Esseen inequality to safely approximate entire distributions in the probabilistic analysis of real-time systems, addressing Problem 2.
3) Using the Berry-Esseen inequality and Lyapunov's CLT to address Problem 3.

The evaluation part of the study shows that using the Berry-Esseen theorem helps to estimate complex calculations accurately and efficiently as the problem gets bigger. This method is better than the circular convolution approach, which is the best current method for solving Problems 2 and 3. The two methods work well together. The circular convolution works best for a small number of items being analyzed, while the Berry-Esseen method gets better and faster as more items are analyzed. This improvement is due to the central limit theorem, which is part of the proposed method. The paper is divided into sections that discuss related research, math concepts, safety, and usefulness of the methods, how they behave over time, solutions to Problems 2 and 3, an evaluation, and a conclusion.

Finally, the related works in this field are discussed on the background of analytical bounds as efficient solutions for probabilistic timing and schedulability analysis in real-time systems. The work of Chen et al. adapted the use of Chernoff bound, and other researchers proposed using Bernstein and Hoeffding bounds. Non-analytical contributions to Problems 1 and 2 include improvements for computing convolution and the use of circular convolution.

Various non-analytical methods exist for periodic and sporadic task models, with most of them relying on convolutions. Problem 3's formulation is similar to the quantile function, which has been used in the analysis and control-optimization problems of real-time systems. Examples include Short and Proenza's stochastic error model, Bertini et al.'s work on improving predictability, and Marković et al.'s method for controlling preemption overheads.

This paper aims to provide an efficient and accurate approximation that address the second problem. To recall the 2nd problem mentioned in the earlier section was How to efficiently and accurately approximate a probability distribution whose exact computation is not possible.

Another Aim of the paper is providing an efficient and accurate way for computing quantiles, given a general random variable, not constraining the problem only to Poisson and Binomial distributions (since those were addressed by other authors like Short and Proenza).

III. NOTATIONS USED IN THE PAPER

Due to the complexity of the proofs and mathematics involved the Authors described notations and terminology used throughout the paper. The equations consider a real time task model, while some equations have been generalized so that they are easier to understand.

1) Discrete Random Variable X on the probability space (Ω, F, P) where Ω is the sample space the set of all possible outcomes. F is an event space where an event is a set of outcomes in the sample space. P represents a probability function that assigns each event in the event space of probability.
2) Usual Stochastic order - Two random variables X and Y , with cumulative distribution functions F_X and F_Y, are said to be in the usual stochastic order, denoted as $X \succeq Y$, if and only if $\forall x$, $F_X(x) \leqslant F_Y(x)$.
3) Independence - Two (discrete) random variables X and Y are independent if the pair of events $\{X = x\}$ and $\{Y = y\}$ are independent for all x, y $\in$ R.
4) Convolution or sum of two random variables.
5) n^{th} moment of a random variable
6) Linearity of Expectations and Variance
7) Lyapunov Central Limit Theorem

In this paper, the authors assume a taskset Γ of g sporadic independent tasks, with each task τ_k defined by a tuple $\langle C_k, T_k, D_k\rangle$, representing the execution time, minimum inter-arrival time, and relative deadline of τ_k, respectively. C_k is a discrete random variable with a known distribution and can represent execution time modes. The random variables $C_1, C_2, \ldots, C_g$ are assumed to be independent.

The taskset follows a constrained deadline model ($D_k \leq T_k$) and is scheduled using a preemptive fixed-priority scheduling policy, where task priorities are distinct and represented by task indexes. Task τ_1 has the highest priority, while task τ_g has the lowest priority in Γ.

$$\mathbf{DMP}_k \leq \min_{0<t\leq \mathbf{D}_k} \mathbb{P}\left(S^{\diamond}_{k,t} > t\right) \leq \mathbb{P}\left(S^{\diamond}_{k,t} > \mathbf{D}_k\right).$$

(Image Taken from Markovic et al.)

Where DMP is the maximum possible deadline miss for all Jobs τ_k .

8) Definition of Sk,t – "The random variable Sk,t is the upper bound on the probabilistic workload accumulated over an interval of length t, for the jobs with a priority greater than or equal to the priority of τ_k, and it is defined as"

$$S_{k,t} \triangleq \sum_{i=1}^{k} \sum_{j=1}^{\alpha_{i,t}} C_i$$

A. *Applicability of Sk,t*

The Sk,t term can be used to derive the safe distribution approximation and deadline miss probabilities under two job-abortion policies:

(A) Incomplete jobs are aborted at their deadline, with t being arbitrarily long.

(B) Jobs run to completion despite the deadline being missed.

In both cases, it is assumed that there is a synchronous busy period.

Under policy (A), the safe distribution approximation and deadline miss probabilities can be derived by considering the probability that a job that arrives at time k will not be completed by its deadline if it is not completed by time t.

Under policy (B), the safe distribution approximation and deadline miss probabilities can be derived by considering the probability that a job that arrives at time k will be completed by its deadline even if it misses its original deadline.

In both cases, the $S_{k,t}$ term plays a crucial role in deriving the safe distribution approximation and deadline miss probabilities.

For this purpose, the paper offers several Corollaries along with their proof for things like Incomplete Jobs that are aborted at their deadline.

The Authors also provide other properties such as Property 3 which states that deadlines do not affect the scheduling decisions at any point since the jobs are not aborted.

We then approach the final Theorem i.e., Theorem 3 which states that "If no task has an accumulated workload at time 0 (e.g., at system startup), then P ($S_{k,t}$ > t) is a safe upper bound on the probability that the accumulated workload of $\tau_k \in \Gamma_{JC}$ exceeds t time units.

The proof of this is provided as well.

B. *Asymptotic behaviour of $S_{k,t}$*

Further it can also be proven that CDF has asymptotic behaviour when t approached infinity.

This is done using the Lyapunov Central Limit theorem, we know that as the number of addends grows the sum approaches a normal distribution.

So, all that remains is to show that the Lyapunov CLT holds for the assumed task model.

This is easily done using the following Lemmas.

Lemma 1: "The third centred moment of the execution time distribution C_i is a finite value."

Lemma 2: "The Variance V is a finite value."

Which leads us to theorem 4.

IV. APPROXIMATION OF CDF

We have discussed about Problem 2 know that the sum of a Random variable will converge to a normal distribution. For large values t we cannot compute the probabilistic workload accumulated over an interval of length t, for the jobs with a priority greater than or equal to the priority of τk.

So, the approximation proposed here works quite well as and the authors have shown through their tests. It can also be possible to calculate the accuracy of this approximation which as we will see is quite useful as well.

V. HOW IS THE APPROXIMATION DONE

The Authors propose several theorems that allow of the estimation of the probability on both the sides.

The Theorem given here is called the Berry-Essen Theorem (It is applied to S(k,t))

The Berry-Esseen theorem is a mathematical theorem that helps us understand how closely a random variable is distributed around its mean. It tells us that for any value of x, the difference between the cumulative distribution function of a sum of random variables (called S_n) and the standard normal distribution ($\Phi(x)$) never exceeds a certain value (called A times ψ).

This means that we can estimate the distribution of $S_{k,t}$ (which is related to Sn) without actually calculating it, by using the Berry-Esseen inequality to bound the possible values of its cumulative distribution function. This makes the estimation process much faster and more efficient.

The value of A is important because it determines how accurate the estimate will be. The tighter the bound-on A, the more accurate the estimate will be. Irina Shevtsova has proved the tightest bound on A so far, which is 0.5583.

Theorem 6 is a mathematical tool that helps us estimate the distribution of a sum of random variables (that is $S_{k,t}$) by approximating its cumulative distribution function (CDF) from both sides. The theorem tells us that the CDF of $S_{k,t}$ converges to the CDF of a generalized normal distribution.

The benefit of Theorem 6 is that it allows us to approximate the entire cumulative distribution function of $S_{k,t}$ This is useful for many statistical applications.

Using the normal CDF approximation, we can efficiently compute the distribution of $S_{k,t}$ using an equation (Eq. 9) without having to compute the actual $S_{k,t}$ values, which can be computationally expensive.

VI. SOLVING PROBLEM 3

In this section we discuss problem 3 that is where we want to find the least value of x (for example, a time point within an interval) such that the probability that a random variable X is less than or equal to x, The probability of this is greater than or equal to some threshold p.

To solve this problem, we use a mathematical concept called the quantile function QX(p), which is the inverse of the cumulative distribution function (CDF) of X. The quantile function takes a probability value p as input and returns the smallest value of x such that the probability of X being less than or equal to x is greater than or equal to p.

By using the quantile function, we can efficiently and accurately find the least value of x that satisfies the probability threshold. This can be useful in many applications, such as scheduling tasks based on their execution workload.

Theorem 8 – "Used for efficiently and accurately derive the least value x (e.g., time point within some interval), such that the probability that a random variable X (e.g., of execution workload) precedes x is greater than or equal to some probability threshold p."

Using the result of QX(p) for the given distribution X and the predefined threshold p can be relevant to controlling the deadline-miss probability when assigning the releases of any new workload.

VII. EVALUATION

The assessment comprised two parts, A and B, which had the objective of estimating the cumulative distribution and quantile function of S(q,t). The evaluation was executed on a MacBook Pro machine equipped with a 2.6 GHz 6-Core Intel Core i7 processor and 16 GB of RAM, and MATLAB with the Advanpix Multiprecision Computing Toolbox was employed to implement all equations.

The objective of the research was to compare two methods for approximating the cumulative distribution function (CDF) of S(q,t), which measures the maximum accumulated probabilistic workload over a time interval of length t for jobs with a priority equal to or greater than the priority of task τq. The first method used Theorem 6 (BE), which is a theoretical formula for computing the CDF of the maximum of a set of random variables. The second method was the circular-convolution method (CC), which is a numerical technique that approximates the CDF of the maximum using a discrete convolution of the probability density functions.

To evaluate the performance of these methods, the experiment generated 1000 tasksets for each point on the graphs for taskset sizes ranging from 5 to 50. The task periods were randomly selected from the log-uniform distribution with a range of 10 to 1000 ms. The evaluation measured the accuracy, computation time, and memory footprint of both methods in approximating S(q,t) and the exact result.

The accuracy of the approximations was evaluated by comparing the CDFs of the approximations with the exact CDFs computed using Monte Carlo simulation. The computation time was measured by recording the time taken by each method to compute the approximations for each taskset. The memory footprint was measured by recording the memory used by each method to store the data structures required for the computation.

The results of the evaluation showed that the BE method was more computationally efficient than the CC method, with significantly lower memory usage and faster computation times. The BE method utilized the moments of the random variables and expectation-related values, which made it more efficient than the CC method. Moreover, the BE method was found to be more accurate than the CC method in approximating the CDF of S(q,t) for all analyzed combinations. Therefore, the BE method is recommended for approximating the CDF of S(q,t) in practice.

The outcomes indicated that the proposed approximation method (BE) is computationally efficient, owing to its calculation method and usage of moments of random variables and expectation-related values. Additionally, BE had a significantly lower average memory footprint than CC for all analyzed combinations.

The second evaluation aimed to compare the accuracy and execution times of three different approximations for the quantile function of a system's response time. The three approximations that were compared were a Central Limit Theorem (CLT)-based approximation, an upper-bound approximation, and a lower-bound approximation.

The evaluation measured the accuracy of each approximation for different values of the response time and the computation time required for each approximation. The results of the evaluation indicated that the CLT-based approximation was more accurate as the response time increased but could be risky to use. This means that as the response time increases, the CLT-based approximation provides a more accurate estimate of the quantile function. However, there is a risk associated with using this approximation, as it may not provide reliable results for smaller response times.

On the other hand, the upper and lower-bound approximations were less accurate, particularly for small response times, but were faster to compute. This means that while these approximations may not provide the most precise estimates of the quantile function, they are still useful for faster computations and for systems with smaller response times.

The evaluation also demonstrated that the said ways are complemented with existing state-of-the-art methods. These approximations provide efficient ways of analyzing task sets and improving the accuracy of system response time predictions. By combining these methods with other existing techniques, it may be possible to obtain more accurate and efficient predictions of system response times.

VIII. CONCLUSION

The challenge of estimating analytically the challenging probabilistic execution demands in real-time systems is addressed in the research. The Lyapunov central limit theorem is applied and demonstrated to demonstrate that as time approaches infinity, the accumulated execution distribution can converge to a normal distribution. The accumulated execution distribution of a job over an arbitrarily long finite time interval is then approximated by the authors using two secure approximations. The proposed methods are shown to supplement the state-of-the-art of analytical approximations using synthetic tasksets to assess the computational effectiveness and approximation correctness.

IX. ADDRESSING THE QUESTIONS ASKED

After the presentation was completed, the students had the opportunity to field several questions to the presenters, here are some of the questions that were asked with a more detailed explanation.

Q.1) Explain what analytical approximations are in the context of probabilistic analysis of real-time systems?

Analytical approximations are mathematical methods used to derive approximate solutions to complex problems. In the context of probabilistic analysis of real-time systems, analytical approximations are used to estimate the performance metrics of a system, such as the probability of failure or the response time, by using simplified mathematical models that are easier to analyze.

Q.2) Why are analytical approximations important in probabilistic analysis of real-time systems?

Probabilistic analysis of real-time systems is often complex and computationally intensive. Analytical approximations provide a way to obtain estimates of system performance metrics quickly and efficiently, without the need for extensive computation.

Q.3) What are some common analytical approximations used in probabilistic analysis of real-time systems?

Common analytical approximations used in probabilistic analysis of real-time systems include the Mean Value Analysis (MVA) method, the Fluid Approximation (FA) method, and the Quasi-Birth-Death (QBD) process method.

Q.4) Are there any limitations of analytical approximations in probabilistic analysis of real-time systems?

The limitations of analytical approximations in probabilistic analysis of real-time systems include their reliance on simplified mathematical models, which may not accurately capture the complexity of real-world systems. Additionally, analytical approximations may not be able to capture the behaviour of systems with complex interactions between components or systems with non-linear behaviours.

Q.5) What are the benefits of using analytical approximations in real-time systems analysis, compared to other methods?

The potential benefits of using analytical approximations in real-time systems analysis include their speed and efficiency in providing estimates of system performance metrics. Additionally, analytical approximations can provide insights into system behavior and help identify potential bottlenecks or areas for improvement.

Q.6) What are any recent developments or research in the field of analytical approximations for probabilistic analysis of real-time systems?

Recent research in the field of analytical approximations for probabilistic analysis of real-time systems has focused on developing more accurate and efficient approximation methods. For example, researchers have developed hybrid approximation methods that combine analytical approximations with simulation-based methods to improve the accuracy of performance estimates.

X. FUTURE WORK

The potential for future work when it comes probabilistic analysis of real time systems is huge, there have been several research papers that have talked about potential (and actual areas of research)

One major area of interest has been computing the convolution of between random variables in the context of probabilistic cache analysis, this has major applications for any real time system, in their paper titled "Speeding up static probabilistic timing analysis" Miltutinovic et al. have proposed many improvements for calculating the convolution between random variables for a real time system[1].

In another paper titled "On the convolution efficiency for probabilistic analysis of real-time systems" the author of this research paper Filip Markovic proposed the use of circular convolution to efficiently calculate the sum of random variables that represent the execution time mode[2].

As in any statistical analysis of a system, Monte Carlo Simulations are another useful area where a lot of research is possible and ongoing. Monte Carlo simulations are a computational tool used to model complex systems and evaluate their performance by generating multiple random scenarios. In real-time systems, Monte Carlo simulations can be used to estimate the likelihood of different outcomes or events, based on the variability of inputs and the uncertainties of the system.

In real-time systems, Monte Carlo simulations can be used to estimate the probability of different scenarios, such as the likelihood of a system failure or the impact of a specific event on the system's performance. By generating multiple random scenarios and analyzing their outcomes, Monte Carlo simulations can provide valuable insights into the behavior and performance of real-time systems, and help engineers and designers make informed decisions.

For this reason, Sergey Bozhko, Georg von der Bruggen, and Bjorn Brandenburg. In their paper "Monte Carlo response-time analysis" proposed the use of Monte Carlo simulation to calculate deadline-miss probability [3].

In this summary we have discussed the importance of the quantile function and there is potential for more work in this area as shown by Short and Proenza in their paper "Towards efficient probabilistic scheduling guarantees for real time systems. In this paper the authors make use of astochastic error modelling and one of the problems on which this is applied is the efficient and accurate calculation of the quantile function for quick control decisions in a real time system [4].

In another paper the same author focused on the efficient approximation of the upper-tail quantile functions for both the Poisson and Binomial Distributions [5].

Minimizing the deadline-miss probability using the quantile function is another area where research can be done, this was shown by Markovic et. Al when they made use of the quantile functions to control the distribution of pre-emption overheads [6].

Last but not the least there is a lot of potential for research to be done by applying Machine Learning techniques in predicting the scheduling of jobs as well as deadline miss prediction using advanced techniques such as Long Short-Term memory (LSTM) and Recurrent Convolution Neural Networks which work best for time-based date input. This area has seen a massive increase in interest with several important papers written such as Power Fluctuations Based on LSTM Recurrent Neural Network: A Case Study on Singapore Power System by Shuli Wen, Yu Wang, Yi Tang or another paper that uses LSTM for fault detection "LiReD: A Light-Weight Real-Time Fault Detection System for Edge Computing Using LSTM Recurrent Neural Networks" by Donghyun Park, Seulgi Kim, Yelin An and Jae-Yoon Jung [7].

REFERENCES

[1] Suzana Milutinovic, Jaume Abella, Damien Hardy, Eduardo Quiñones, Isabelle Puaut & Francisco J. Cazorla (2015). Speeding up Static Probabilistic Timing Analysis.

[2] Markovic, F., Papadopoulos, A. V., & Nolte, T. (2021). On the convolution efficiency for probabilistic analysis of real-time systems. In 33rd Euromicro Conference on Real-Time Systems (ECRTS2021). Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

[3] Bozhko, S., von der Bruggen, G., & Brandenburg, B. (2021). Monte Carlo response-time analysis. In IEEE 42nd Real-Time Systems Symposium (pp. 342-355). IEEE.

[4] Short, M., & Proenza, J. (2013). Towards efficient probabilistic scheduling guarantees for real-time systems subject to random errors and random bursts of errors. In 2013 25th Euromicro Conference on Real-Time Systems (pp. 259-268). IEEE.

[5] Short, M. (2013). Improved inequalities for the Poisson and binomial distribution and upper tail quantile functions. International Scholarly Research Notices.

[6] Markovic, F., Carlson, J., Dobrin, R., Lisper, B., & Thekkilakattil, A. (2018). Probabilistic response time analysis for fixed preemption point selection. In 2018 IEEE 13th International Symposium on Industrial Embedded Systems (SIES) (pp. 1-10). IEEE.

[7] Wen, S., Wang, Y., & Tang, Y. A Case Study on Singapore Power System.