CTA: A Correlation-Tolerant Analysis of the
Deadline-Failure Probability of Dependent Tasks

Filip Marković, Pierre Roux, Sergey Bozhko, Alessandro V. Papadopoulos, Björn B. Brandenburg

1 Max Planck Institute for Software Systems (MPI-SWS), Germany
2 ONERA/DTIS, Université de Toulouse, France
3 Saarbrücken Graduate School of Computer Science, Saarland University, Germany
4 Mälardalen University (MDU), Sweden

Abstract—Estimating the worst-case deadline failure probability (WCDFP) of a real-time task is notoriously difficult, primarily because a task’s execution time typically depends on prior activations (i.e., history dependence) and the execution of other tasks (e.g., via shared inputs). Previous analyses have either assumed that execution times are probabilistically independent (which is unrealistic and unsafe), or relied on complex upper-bounding abstractions such as probabilistic worst-case execution time (pWCET), which mask dependencies with pessimism. Exploring an analytically novel direction, this paper proposes the first closed-form upper bound on WCDFP that accounts for dependent execution times. The proposed correlation-tolerant analysis (CTA), based on Cantelli’s inequality, targets fixed-priority scheduling and requires only two basic summary statistics of each task’s ground-truth execution time distribution: upper bounds on the mean and standard deviation (for any possible job-arrival sequence). Notably, CTA does not use pWCET, nor does it require the full execution-time distribution to be known. Core parts of the analysis have been verified with the Coq proof assistant. Empirical comparison with state-of-the-art WCDFP analyses reveals that CTA can yield significantly improved bounds (e.g., a lower WCDFP than any pWCET-based method for ≥70% of the workloads tested at 90% pWCET utilization and 60% average utilization). Beyond accuracy gains, the favorable results highlight the potential of the previously unexplored analytical direction underlying CTA.

I. INTRODUCTION

Probabilistic analysis of real-time systems holds the promise of addressing the central challenge of modern hardware and software architectures: unavoidable uncertainty in the execution behavior of real-time tasks. Such uncertainty, deeply embedded in the fabric of modern computing systems, more often than not precludes meaningful (classical) worst-case analysis, leaving a stochastic perspective as the only viable option.

One of the most pressing open problems in this space is the issue of dependent execution times (also referred to as execution-time correlation). Specifically, when bounding a task’s worst-case deadline-failure probability (WCDFP), it is crucial to account for possible dependencies on both previous activations (intra-task dependence) and other tasks in the system (inter-task dependence). If such dependencies are ignored, the WCDFP may be severely under-approximated.

These observations are not new: the lack of independence in practice was recognized as a safety problem already more than 25 years ago by Tia et al. [49] in one of the first works on probabilistic schedulability analysis. Unfortunately, only little progress has been made on this issue since Tia et al.’s observation, with Davis and Cucu-Grosjean noting in the closing remarks of their recent survey [19]: “Issues of dependence are of great importance in probabilistic schedulability analysis […] Analyses are needed that can address dependencies”.

Prior attempts at tackling dependence in state-of-the-art WCDFP analyses have relied on over-approximation. The common idea in this line of work is to “pad” the ground-truth execution-time distributions with “sufficient pessimism,” to the point that task behavior can be safely assumed to be independent. The primary mechanism for realizing such an analysis in a sound manner is the concept of a probabilistic worst-case execution time (pWCET) distribution [5, 8, 14, 17, 18], which can be determined for each task either via static analyses [e.g., 4, 6, 16, 31] or with measurement-based techniques such as extreme value theory (EVT) [e.g., 32, 33, 46, 47].

Specifically, the pWCET approach promises that the analysis may model execution times with independent random variables following the pWCET distribution, provided the pWCET distribution is suitably determined [19]. However, a significant limitation of such independence-assuming analysis (IAA) lies in its inherent over-approximation of the ground truth, which can lead to considerable pessimism compared to actual behavior.

This paper. Exploiting a fundamentally different direction, we propose a novel correlation-tolerant analysis (CTA) of WCDFP under fixed-priority scheduling. CTA is based on Cantelli’s inequality [9] and departs from the state of the art in three major ways: first, CTA does not use pWCET; nor does it otherwise require ground-truth distributions to be pessimistically padded; second, unlike traditional methods, CTA does not require full knowledge of the ground-truth distributions, as it uses only bounds on their means and standard deviations (under any possible job-arrival sequence); and last but not least, CTA is safe in the presence of arbitrarily dependent execution times. Notably, CTA also does not require the degree of inter- or intra-task correlation to be quantified, which is desirable in practice.

In developing CTA, we make the following contributions:

• We convey the core idea with a simple example (Sec. II).
• From Cantelli’s inequality [9], we derive, and verify with Coq [13, 41], an upper bound on the sum of random variables with unknown degrees of correlation (Sec. IV).
• We formally model the execution of a stochastic sporadic real-time workload under preemptive uniprocessor fixed-
priority scheduling with a job-abortion policy that discards incomplete jobs at their deadline (Sec. V).

- Connecting Sec. IV and Sec. V, we obtain CTA (Sec. VI).
- Finally, we report on an empirical evaluation that reveals CTA to be effective at reclaiming pessimism relative to pWCET-based baselines in many (but not all) scenarios, thereby showing CTA to be a promising addition to the existing portfolio of WCDFP analysis methods.

II. Motivating Example

Fig. 1 shows an illustrative example comprised of two tasks \( \tau_1 \) and \( \tau_2 \) executing on a uniprocessor. Both tasks have identical periods and deadlines of 10 time units (TUs). Additionally, we assume their arrivals to be aligned, and that jobs are aborted upon reaching their deadlines. Task \( \tau_1 \) has higher priority than task \( \tau_2 \) and thus always executes first.

Let us assume that each job of \( \tau_1 \) executes for

- 1 TU with probability 0.965,
- 3 TUs with probability 0.015, or
- 5 TUs with probability 0.02.

Assume that \( \tau_2 \)’s execution depends on the execution of the previously executed job of \( \tau_1 \) as illustrated in Fig. 1. From the six depicted scenarios, we can infer that a job of \( \tau_2 \) requires

- 2 TUs with probability 0.975, or
- 8 TUs with probability 0.025.

We refer to these distributions as ground-truth distributions. As shown in Fig. 1, there are only two scenarios in which \( \tau_2 \)’s job misses its deadline. Consequently, the ground-truth WCDFP of \( \tau_2 \) is \( 0.02 = 0.01 + 0.01 \), denoted \( WCDFP_2 = 0.02 \).

Now, let us explore three different approaches for calculating an upper bound on WCDFP.

A. Assuming Independence When There is None

Already in one of the earliest papers on the stochastic analysis of real-time systems [49], Tia et al. observed that assuming random variables that model ground-truth behavior to be independent when they are not may cause the WCDFP to be under-approximated. This is indeed the case with \( \tau_2 \), since under an (incorrect) independence assumption, the response-time distribution of \( \tau_2 \) would be:

- \( 3 = 1 + 2 \), with probability 0.940875 = 0.965 \cdot 0.975
- \( 5 = 3 + 2 \), with probability 0.014625 = 0.015 \cdot 0.975
- \( 7 = 5 + 2 \), with probability 0.0195 = 0.02 \cdot 0.975
- \( 9 = 1 + 8 \), with probability 0.024125 = 0.965 \cdot 0.025

While the only two cases resulting in deadline failure are:

- \( 11 = 3 + 8 \), with probability 0.000375 = 0.015 \cdot 0.025
- \( 13 = 5 + 8 \), with probability 0.0005 = 0.02 \cdot 0.025

Consequently, the independence-assuming WCDFP bound 0.000375 + 0.0005 = 0.000875 under-approximates the ground-truth WCDFP (0.02). As this example shows again, to ignore correlations is to risks unsound WCDFP estimates.

B. Safe Over-Approximation with pWCET

The widely studied pWCET approach [17, 19] promises sound results without having to forego the analytical conveniences afforded by independence. Let us next sketch how to safely upper-bound \( WCDFP_2 \) in this manner.

The essence of the pWCET idea is to come up with one execution-time distribution for each task that is sufficiently “padded” to over-approximate the task’s actual execution-time distribution in any scenario, even if job execution times are assumed to be independent. In other words, by injecting “sufficient pessimism” into each task’s pWCET distribution, it becomes possible to introduce independence as a simplifying assumption without jeopardizing soundness.

When deriving pWCET distributions, there is some degree of freedom due to the interplay of the “padding.” Multiple valid pWCET distributions can hence be derived for both tasks; the following two yield the least pessimistic WCDFP for \( \tau_2 \).

According to \( pWCET_1 \), any job of \( \tau_1 \) executes for at most

- 1 TU with probability 0.2,
- 3 TUs with probability 0.4, or
- 5 TUs with probability 0.4.

According to \( pWCET_2 \), any job of \( \tau_2 \) executes for at most

- 2 TUs with probability 0.3, or
- 8 TUs with probability 0.5.

Given the two pWCETs, we may derive a bound on the sum \( pWCET_1 + pWCET_2 assuming independence (i.e., using convolution), and thereby obtain as a response-time estimate

- \( 1 + 2 = 3 \), with probability 0.06 = 0.2 \cdot 0.3
- \( 3 + 2 = 5 \), with probability 0.13 = 0.4 \cdot 0.3
- \( 5 + 2 = 7 \), with probability 0.13 = 0.4 \cdot 0.3
- \( 1 + 8 = 9 \), with probability 0.13 = 0.2 \cdot 0.5

While the only two cases resulting in deadline failure are:

- \( 3 + 8 = 11 \), with probability 0.26 = 0.4 \cdot 0.5
- \( 5 + 8 = 13 \), with probability 0.26 = 0.4 \cdot 0.5

In concurrent work [8], a rigorous, axiomatic definition of pWCET has been formalized using the Coq proof assistant and formally verified to enable independence-based reasoning. The pWCET distributions obtained in the presented example are consistent with the Coq-verified definition [8].

2The entire motivating example, including the complete derivation of the pWCET distributions, is available online as a Python Jupyter notebook file [42].
Thus, any pWCET-based analysis has no choice but to vastly over-approximate the ground-truth WCDFP (0.02) with 0.53 = 0.26 + 0.26. While independence is a convenient simplifying assumption, it requires all dependencies to be “masked” by padding, which results in prohibitive pessimism in this case.

C. Embracing Correlation with CTA

In this paper, we explore a different approach. Suppose that from measurements we can determine upper bounds on the expected values and standard deviations of the execution-time distributions of the two tasks. For example, let us assume the upper bounds given Table I.

<table>
<thead>
<tr>
<th>Task</th>
<th>Ground-Truth Statistics and Upper Bounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected Value</td>
<td>$e_1 = 1.11$ TUs, $e_2 = 2.15$ TUs</td>
</tr>
<tr>
<td>Upper Bound on Expected Value</td>
<td>$\hat{e}_1 = 1.12$ TUs, $\hat{e}_2 = 2.16$ TUs</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>$s_1 \approx 0.606$ TUs, $s_2 \approx 0.937$ TUs</td>
</tr>
<tr>
<td>Upper Bound on Standard Deviation</td>
<td>$\hat{s}_1 = 0.61$ TUs, $\hat{s}_2 = 0.94$ TUs</td>
</tr>
</tbody>
</table>

As we show later, given these upper bounds $\hat{e}_1, \hat{e}_2, \hat{s}_1,$ and $\hat{s}_2$ on the respective ground-truth statistics, we can upper-bound the ground-truth WCDFP for the deadline $D_2 = 10$ as follows.

$$WCDFP_2 \leq \frac{(\hat{s}_1 + \hat{s}_2)^2}{(\hat{s}_1 + \hat{s}_2)^2 + (D_2 - (\hat{e}_1 + \hat{e}_2))^2} \approx 0.05$$

As the example demonstrates, an upper bound on the ground-truth WCDFP of $\tau_2$ derived in this way (i.e., using CTA), which avoids the “pWCET detour,” can be markedly less pessimistic than the pWCET-based WCDFP bound (0.53) derived in Sec. II-B, i.e., $0.02 < 0.05 < 0.53.$ Importantly, the CTA bound is safe as opposed to the optimistic WCDFP (0.000875) derived in Sec. II-A, i.e., $0.000875 < 0.02 < 0.05.$

In the remainder of this paper, we justify Inequality 1.

III. Probability Theory Background

We briefly review the needed probability theory background. Let $(\Omega, F, P)$ be a probability space, with sample space $\Omega$ being the set of all possible outcomes, $F \subseteq 2^\Omega$ the event space, where an event is a set of outcomes in the sample space, and $P : F \rightarrow [0, 1]$ a probability function. In the following, we let $\mathbb{R}$ denote the reals and $\mathbb{R}$ the extended reals ($\mathbb{R} \triangleq \mathbb{R} \cup \{\pm \infty\}$).

**Def. 1 (Random Variable).** A random variable $X$ on the probability space $(\Omega, F, P)$ is a measurable function $X : \Omega \rightarrow \mathbb{R}$ such that $\{\omega \in \Omega : X(\omega) = x\} \in F$ for all $x \in \mathbb{R}.$

We denote the probability of a random variable $X$ taking a value $x$ with $P[\omega \in \Omega | X(\omega) = x]$ or, more briefly, $P[X = x].$

We use the following definitions and theorems on random variables.

**Def. 2 (Expected value).** Given a random variable $X,$ its expected value $\mathbb{E}[X] \in \mathbb{R}$ is a measure of the central tendency or average value of the possible outcomes of $X.$

$$\mathbb{E}[X] \triangleq \int_{\omega \in \Omega} X(\omega) \, dP.$$  

The expectation operator $\mathbb{E}[-]$ acts linearly on sums of random variables, which is known as linearity of expectation.

**Fact 1 (Linearity of expectation, [e.g., 25, p. 40]).** Let $X$ and $Y$ be two (possibly dependent) random variables. If $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ are finite, then

$$\forall a, b \in \mathbb{R}, \mathbb{E}[a \cdot X + b \cdot Y] = a \cdot \mathbb{E}[X] + b \cdot \mathbb{E}[Y].$$

**Def. 3 (Covariance).** Given two random variables $X$ and $Y,$ their covariance, denoted by $\text{Cov}[X, Y],$ is a measure of the degree to which $X$ and $Y$ fluctuate in similar ways.

$$\text{Cov}[X, Y] \triangleq \mathbb{E}[(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[Y])].$$

**Def. 4 (Variance).** Given a random variable $X,$ its variance, denoted by $\sqrt{\mathbb{V}[X]},$ is a measure of the dispersion or spread of the possible outcomes of $X.$

$$\sqrt{\mathbb{V}[X]} \triangleq \sqrt{\text{Cov}[X, X]}$$

**Fact 2.** Given two random variables $X$ and $Y,$ if $\mathbb{V}[X], \mathbb{V}[Y]$ and $\text{Cov}[X, Y]$ are finite, then

$$\mathbb{V}[X + Y] = \mathbb{V}[X] + \mathbb{V}[Y] + 2 \cdot \text{Cov}[X, Y].$$

In the following, we write $\mathbb{V}[X] < \infty$ and $\mathbb{E}[X] < \infty$ to denote that a random variable $X$ has finite variance and mean.

**Def. 5 (Standard deviation).** Given a random variable $X,$ its standard deviation $\sigma[X]$ is the square root of its variance:

$$\sigma[X] \triangleq \sqrt{\mathbb{V}[X]}.$$

**Def. 6 (Conditional Probability).** Let $A$ and $B$ be two events in the probability space $(\Omega, F, P)$ with $P[B] > 0.$ The conditional probability of event $A$ given event $B$ is denoted $P[A|B],$ where

$$P[A|B] \triangleq \frac{P[A \cap B]}{P[B]}.$$

With all basic definitions in place, we next derive an upper bound on the sum of potentially correlated random variables.

IV. Concentration Inequality

To lay the foundations for CTA, we first obtain a concentration inequality on the sum of dependent random variables from a well-known result in probability theory and statistics—Cantelli’s inequality [9]. The result we prove (Theorem 2) is a closed-form expression, depending only on the means and standard deviations of the random variables forming the sum. All proofs in this section have been verified with the Coq proof assistant [13], using the MathComp Analysis library [2, 3]. The Coq development is available online [41].

The main problem to be solved is the following.

**Problem 1.** Given a sum $X_1 + X_2 + \ldots + X_n$ of $n$ potentially correlated random variables, derive an upper bound $B$ on the
probability that the sum exceeds a given value \( t \in \mathbb{R} \), i.e.,
\[
\mathbb{P}[\sum_{i=1}^{n} X_i > t],
\]
using only the expected values and standard deviations of \( X_1, X_2, \ldots, X_n \):
\[
\mathbb{P}\left[ \sum_{i=1}^{n} X_i > t \right] \leq B(t, \mathbb{E}[X_1], \sigma[X_1], \ldots, \mathbb{E}[X_n], \sigma[X_n]).
\]

To define such a function \( B \), we start with Cantelli’s inequality, which provides a bound on the probability of a random variable deviating from its mean.

**Theorem 1** (Cantelli’s inequality [9]). For an arbitrary random variable \( S \) such that \( \mathbb{V}[S] < \infty \) and \( \mathbb{E}[S] < \infty \), and any \( \lambda > 0 \),
\[
\mathbb{P}[S - \mathbb{E}[S] \geq \lambda] \leq \frac{\mathbb{V}[S]}{\mathbb{V}[S] + \lambda^2}.
\]

It will be useful to restate Theorem 1 as follows.

**Corollary 1.** For an arbitrary random variable \( S \) such that \( \mathbb{V}[S] < \infty \) and \( \mathbb{E}[S] < \infty \), and any \( t > \mathbb{E}[S] \),
\[
\mathbb{P}[S \geq t] \leq \frac{\mathbb{V}[S]}{\mathbb{V}[S] + (t - \mathbb{E}[S])^2}. \tag{2}
\]

**Proof.** Making the variable change \( t = \lambda + \mathbb{E}[S] \), we get \( \lambda = t - \mathbb{E}[S] > 0 \) for \( t > \mathbb{E}[S] \). Hence, according to Theorem 1:
\[
\mathbb{P}[S \geq t] = \mathbb{P}[S - \mathbb{E}[S] \geq \lambda] \leq \frac{\mathbb{V}[S]}{\mathbb{V}[S] + \lambda^2} = \frac{\mathbb{V}[S]}{\mathbb{V}[S] + (t - \mathbb{E}[S])^2},
\]
which concludes the proof. \( \square \)

Next, we show that the right-hand side of Inequality (2) is a non-decreasing function w.r.t. \( \mathbb{E}[S] \) and \( \mathbb{V}[S] \). For brevity, we abbreviate \( \mathbb{E}[S] \) as \( e \) and \( \mathbb{V}[S] \) as \( v \).

**Lemma 1.** The function \( f(e, v) = \frac{v}{v + (t - e)^2} \) is non-decreasing w.r.t. both \( e \in \mathbb{R} \) and \( v \in \mathbb{R} \) if \( e < t \) and \( v \geq 0 \).

**Proof.** We must show \( f(e_1, v_1) \leq f(e_2, v_2) \) for \( 0 \leq v_1 \leq v_2 \) and \( e_1 \leq e_2 < t \). Rewrite \( f \) as \( f(e, v) = v \cdot (v + (t - e)^2)^{-1} \):
\[
v_1 \cdot (v_1 + (t - e_1)^2)^{-1} \leq v_2 \cdot (v_2 + (t - e_2)^2)^{-1}
\]
Multiply by \( (v_1 + (t - e_1)^2) \cdot (v_2 + (t - e_2)^2) \):
\[
v_1 \cdot (v_2 + (t - e_2)^2) \leq v_2 \cdot (v_1 + (t - e_1)^2)
\]
Subtract \( v_1 \cdot v_2 \) and rearrange:
\[
v_1 \cdot (t - e_2)^2 \leq v_2 \cdot (t - e_1)^2
\]
Since \( 0 \leq v_1 \leq v_2 \) and \( 0 \leq (t - e_2)^2 \leq (t - e_1)^2 \) given \( e_1 \leq e_2 < t \), the final inequality holds. \( \square \)

In other words, by Lemma 1, it is safe to use Corollary 1 even if \( \mathbb{E}[S] \) and \( \mathbb{V}[S] \) are over-approximated with upper bounds.

Recall from Problem 1 that we seek a bound on the sum of possibly dependent random variables. Thus let us now consider \( S \) to be that sum, i.e., \( S = \sum_{i=1}^{n} X_i \). By Corollary 1 and Lemma 1, we can use Cantelli’s inequality to bound \( \mathbb{P}[S \geq t] \) despite unknown correlations among the terms, provided we can upper-bound \( \mathbb{E}[S] \) and \( \mathbb{V}[S] \). Let us hence turn our attention to the problem of finding suitable upper bounds on \( \mathbb{E}[S] = \mathbb{E}[\sum_{i=1}^{n} X_i] \) and \( \mathbb{V}[S] = \mathbb{V}[\sum_{i=1}^{n} X_i] \).

The former is trivial: we can efficiently compute \( \mathbb{E}[\sum_{i=1}^{n} X_i] \) thanks to the linearity of expectation (Fact 1). Importantly, this holds even for correlated random variables.

The latter requires more elaboration, as the sum of variances depends on the covariance (recall Fact 2), which in the context of execution times is difficult to obtain in practice. Therefore, the next step is to find an upper bound that avoids this dependency. To this end, we first recall a useful fact.

**Fact 3.** For two possibly dependent random variables \( X \) and \( Y \), if \( \mathbb{V}[X] < \infty \), \( \mathbb{V}[Y] < \infty \), and \( \text{Cov}[X, Y] < \infty \), then
\[
\mathbb{V}[X + Y] \leq (\mathbb{V}[X] + \mathbb{V}[Y])^2.
\]

Two of possibly many proofs of Fact 3 can be found in textbooks by Keener [28, Inequality (4.11), p. 71] and Mukhopadhyay [45, Inequality (3.9.13), p. 150]. From Fact 3, we obtain the desired bound on \( \mathbb{V}[S] \).

**Lemma 2.** Let \( X_1, X_2, \ldots, X_n \) be \( n \) possibly dependent random variables. If \( \text{Cov}[X_i, X_j] < \infty \) for all pairs \( X_i, X_j \), then
\[
\mathbb{V}\left[ \sum_{i=1}^{n} X_i \right] \leq \left( \sum_{i=1}^{n} \mathbb{V}[X_i] \right)^2.
\]

**Proof.** By induction on \( n \).

**Base case \( n = 0 \):** trivially, \( \mathbb{V}[0] = 0 \leq 0^2 \).

**Induction step:** for any arbitrary \( n \in \mathbb{N} \), assume that:
\[
\mathbb{V}\left[ \sum_{i=1}^{n} X_i \right] \leq \left( \sum_{i=1}^{n} \mathbb{V}[X_i] \right)^2
\]

Then:
\[
\mathbb{V}\left[ \sum_{i=1}^{n+1} X_i \right] \leq \left( \mathbb{V}\left[ \sum_{i=1}^{n} X_i \right] + \sqrt{\mathbb{V}[X_{n+1}]} \right)^2 \leq \left( \left( \sum_{i=1}^{n} \mathbb{V}[X_i] \right)^2 + \mathbb{V}[X_{n+1}] \right)^2 \leq \left( \sum_{i=1}^{n+1} \mathbb{V}[X_i] \right)^2 \leq \left( \sum_{i=1}^{n+1} \mathbb{V}[X_i] \right)^2,
\]
where Inequality (i) follows from Fact 3, and Inequality (ii) from the induction hypothesis. \( \square \)

We now have everything in place to state the main concentration inequality, thus solving Problem 1.

**Theorem 2.** Let \( X_1, X_2, \ldots, X_n \) be \( n \) possibly dependent random variables with finite covariances (i.e., \( \text{Cov}[X_i, X_j] < \infty \) for all \( i, j \)).
for all pairs $X_i$, $X_j$). Define $b \triangleq \sum_{i=1}^{n} \mathbb{E}[X_i]$ and $a \triangleq \left(\sum_{i=1}^{n} \sigma[X_i]\right)^2$. Then, for any $t > b$, we have
\[
P \left[ \sum_{i=1}^{n} X_i \geq t \right] \leq \frac{a}{a + (t - b)^2}.
\]

Proof. We have:
\[
P \left[ \sum_{i=1}^{n} X_i \geq t \right] \leq \frac{\mathbb{V}[\sum_{i=1}^{n} X_i] + (t - \mathbb{E}\sum_{i=1}^{n} X_i)^2}{a + (t - E\sum_{i=1}^{n} X_i)^2} \leq \frac{a}{a + (t - b)^2},
\]
where Inequality (i) follows from Corollary 1, Inequality (ii) from Lemmas 1 and 2, and Equality (iii) from Fact 1. \hfill \square

Finally, we observe that the bound is robust with regard to over-approximation, which is key to making it practical.

**Corollary 2.** Let $X_1, X_2, \ldots, X_n$ be $n$ possibly dependent random variables with finite covariances (i.e., Cov$[X_i, X_j] < \infty$ for all pairs $X_i, X_j$). For any $\bar{\epsilon}_i \geq \mathbb{E}[X_i]$ and $\bar{s}_i \geq \sigma[X_i]$, define $b \triangleq \sum_{i=1}^{n} \bar{\epsilon}_i$ and $a \triangleq \left(\sum_{i=1}^{n} \bar{s}_i\right)^2$. Then, for any $t > b$, we have
\[
P \left[ \sum_{i=1}^{n} X_i \geq t \right] \leq \frac{a}{a + (t - b)^2}.
\]

Proof. From the assumption that $\bar{\epsilon}_i$ and $\bar{s}_i$ are upper bounds, we have $\sum_{i=1}^{n} \bar{\epsilon}_i \geq \sum_{i=1}^{n} \mathbb{E}[X_i]$, and $\sum_{i=1}^{n} \bar{s}_i \geq \sum_{i=1}^{n} \sigma[X_i]$. Then, by Theorem 2 and Lemma 1, the claim follows. \hfill \square

In Sec. VI, we will use Corollary 2 as the core of CTA.

As already mentioned, all results in this section have been formalized and verified with the Coq proof assistant~[13], using the MathComp Analysis library~[2, 3]. The initial development encompassed 750 lines of code (statements and proofs) and took an experienced Coq user about a week of work to complete. The results of general interest (e.g., Corollary 1) have been upstreamed into the MathComp Analysis library, leaving only about 200 lines of code specific to this paper~[41]. The formalization effort was helpful in generalizing arguments and making all assumptions explicit.

**V. GROUND-TRUTH SYSTEM MODEL**

We consider a set $\tau \triangleq \{\tau_1, \tau_2, \ldots, \tau_n\}$ of $n$ sporadic tasks running on a uniprocessor under fixed-priority preemptive scheduling. Tasks are indexed in order of decreasing priority, i.e., $\tau_1$ has the highest priority, and no two tasks have equal priority. We assume that, for each task $\tau_i$, the minimum inter-arrival time $T_i$ between any two consecutive jobs is known, as well as its relative deadline $D_i$. We focus on constrained deadlines, i.e., $\forall t \in \mathbb{N}, 1 \leq i \leq n, D_i \leq T_i$. When a job misses its deadline, it is aborted, i.e., cut off from service and discarded.

We assume discrete time in the following, i.e., the set of natural numbers $\mathbb{N}$ is the time domain (e.g., processor cycles).

**Ground-truth behavior.** Recall from Sec. III that $\Omega$ is the set of all possible outcomes, i.e., system state evolutions, and that $\omega \in \Omega$ represents a single evolution. In the context of a given evolution $\omega \in \Omega$, the $j$-th job of $\tau_i$ arriving in $\omega$ is denoted with $J_{\omega,j}(\tau_i)$, its arrival time with $a_{\omega,j}(\tau_i)$, absolute deadline with $d_{\omega,j}(\tau_i) \triangleq a_{\omega,j}(\tau_i) + D_i$, and execution time with $C_{\omega,j}(\tau_i)$.

Following Bozhko et al.~[7], we use the notion of an arrival sequence $\zeta(t, \omega) \triangleq \{J_{\omega,j}(\tau_i) | d_{\omega,j}(\tau_i) = t\}$, which for a given $\omega \in \Omega$ maps each $t \in \mathbb{N}$ to the jobs that arrive at time $t$ in $\omega$.

Recall that $F \subseteq 2^{\Omega}$ denotes the event space of $\Omega$. Building on $\zeta(t, \omega)$, we define $\Xi \subseteq F$ to be the set of all possible disjoint events of $\Omega$ such that, for each event $\xi \in \Xi$, $\xi$ encompasses all evolutions in $\Omega$ with identical arrival sequence, that is, $\forall \omega, \omega' \in \Xi, \forall t \in \mathbb{N}, \zeta(t, \omega) = \zeta(t, \omega')$, and $\mathbb{P}[\xi] > 0$. In the context of a fixed event $\xi \in \Xi$, we can drop $\omega$ for brevity and simply write $J_{\xi,j}, a_{\xi,j}, d_{\xi,j}$ since they are the same for all $\omega \in \xi$. In contrast, $C_{\xi,j}(\omega)$ may vary for different $\omega \in \xi$. For notational convenience, we define $C_{\xi,j}(\omega)$ to be zero in the case that fewer than $j$ jobs of $\tau_i$ arrive in evolution $\omega$.

Table II summarizes the adopted notation.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau$</td>
<td>A task set.</td>
</tr>
<tr>
<td>$\tau_i$</td>
<td>A task from $\tau$ with index $i$.</td>
</tr>
<tr>
<td>$T_i$</td>
<td>The minimum inter-arrival time of $\tau_i$.</td>
</tr>
<tr>
<td>$D_i$</td>
<td>Relative deadline of $\tau_i$.</td>
</tr>
<tr>
<td>$t \in \mathbb{N}$</td>
<td>A point in time.</td>
</tr>
<tr>
<td>$\omega \in \Omega$</td>
<td>Sample space of system evolutions.</td>
</tr>
<tr>
<td>$\xi \subseteq \Omega$</td>
<td>Event encompassing all evolutions exhibiting an identical arrival sequence.</td>
</tr>
<tr>
<td>$J_{\xi,j}$</td>
<td>The $j$-th job of $\tau_i$ arriving in $\xi$.</td>
</tr>
<tr>
<td>$a_{\xi,j}$</td>
<td>Arrival time of $J_{\xi,j}$ in $\xi$.</td>
</tr>
<tr>
<td>$d_{\xi,j}$</td>
<td>Absolute deadline of $J_{\xi,j}$ in $\xi$.</td>
</tr>
<tr>
<td>$C_{\xi,j}(\omega)$</td>
<td>Execution time of $J_{\xi,j}$ in $\omega$ (specific to each $\omega \in \xi$).</td>
</tr>
<tr>
<td>$TC\xi_j(\omega)$</td>
<td>Carry-in workload at time $t$ in $\omega$, of jobs of $\tau_i$.</td>
</tr>
<tr>
<td>$CT_{\xi,j}(\omega)$</td>
<td>Carry-in workload at time $t$ in $\omega$, of jobs of $\tau_i$ and higher-priority jobs.</td>
</tr>
<tr>
<td>$TWi_{t_1,t_2}(\omega)$</td>
<td>Workload of $\tau_i$ arriving within $[t_1, t_2]$ in $\omega$.</td>
</tr>
<tr>
<td>$Wi_{t_1,t_2}(\omega)$</td>
<td>Workload of jobs of $\tau_i$ and higher-priority jobs in interval $[t_1, t_2]$ in $\omega$.</td>
</tr>
<tr>
<td>$\xi \vee \Delta(\omega)$</td>
<td>Processor demand w.r.t. $\tau_i$, in interval $[t, t + \Delta]$ in $\omega$.</td>
</tr>
<tr>
<td>$Ra_{\xi,j}(\omega)$</td>
<td>Truncated response time of $J_{\xi,j}$ in evolution $\omega$.</td>
</tr>
<tr>
<td>$WCDFP_i$</td>
<td>Worst-case deadline failure probability of $\tau_i$.</td>
</tr>
<tr>
<td>$\tilde{f}(\xi)$</td>
<td>Upper bound on the given function $f(\xi)$.</td>
</tr>
</tbody>
</table>

**Workload characterization.** To formalize the stochastic execution behavior while acknowledging potential dependencies among jobs, we next introduce several well-known concepts commonly encountered in the schedulability analysis literature, adapted to our context and notation.

For simplicity, we define the following functions in the context of a fixed, individual possible event $\xi \in \Xi$, thus fixing the arrival times and limiting randomness to execution costs.

**Def. 7.** The cumulative demand of jobs of $\tau_i$ issued within the time interval $[t_1, t_2] \subseteq \mathbb{N}$ in evolution $\omega \in \xi$ is defined as
\[
TWi_{t_1,t_2}(\omega) \triangleq \sum_{j:a_{\xi,j} \in [t_1,t_2]} C_{\xi,j}(\omega).
\]
Def. 8. The \textit{workload} of jobs of $\tau_i$ and higher-priority jobs in an interval $[t_1, t_2) \subseteq \mathbb{N}$ in evolution $\omega \in \xi$ is given by

$$W_{i,|t_1, t_2)}(\omega) \triangleq \sum_{1 \leq k \leq i} T W_{i,k,|t_1, t_2)}(\omega).$$

Let $\sigma(\omega, t)$ denote the job (if any) scheduled at time $t$ in evolution $\omega \in \xi$, and $|\tau|$ the cardinality of the enclosed set.

Def. 9. Let $S_{i,j}(\omega, t)$ denote the \textit{total service} received by a job $J_{i,j}^\xi$ up to (but not including) time $t$, in evolution $\omega \in \xi$.

$$S_{i,j}(\omega, t) \triangleq \left\{ t' \in [0, t) \mid \sigma(\omega, t') = J_{i,j}^\xi \right\}$$

Next, we define the carry-in workload of jobs of $\tau_i$, i.e., the remaining execution time of jobs of $\tau_i$ at time instant $t$. Recall that incomplete jobs are aborted at their deadline and $D_i \leq T_i$ for each $\tau_i \in \tau$. Thus, at most one job of each higher-priority task contributes carry-in workload at any time.

Def. 10. The \textit{carry-in workload} at time $t$ due to task $\tau_i$ in evolution $\omega \in \xi$ is:

$$\mathcal{TCl}_{i,t}(\omega) \triangleq \begin{cases} C_{i,j}(\omega) - S_{i,j}(\omega, t), & \text{if } \exists j \in \mathbb{N} : a_{i,j}^\xi \leq t < d_{i,j}^\xi \\
0, & \text{otherwise.} \end{cases}$$

Def. 11. The \textit{total carry-in workload} of higher-priority jobs affecting $\tau_i$ at time $t \in \mathbb{N}$ in evolution $\omega \in \xi$ is:

$$\mathcal{Cl}_{i,t}(\omega) \triangleq \sum_{1 \leq k \leq i} \mathcal{TCl}_{k,t}(\omega).$$

To complement the concept of carry-in work, we analogously characterize the amount of work that has been discarded.

Def. 12. The \textit{aborted workload} of task $\tau_i$ at time $t$ in evolution $\omega \in \xi$ is:

$$\mathcal{KW}_i(\omega, t) \triangleq \begin{cases} C_{i,j}(\omega) - S_{i,j}(\omega, t), & \text{if } \exists j \in \mathbb{N} : a_{i,j}^\xi = t \\
0, & \text{otherwise.} \end{cases}$$

Def. 13. Let $\mathcal{K}_{i,|t_1, t_2)}(\omega)$ be the \textit{total unfinished execution time} of jobs of $\tau_i$ and higher-priority jobs that are aborted due to missing their deadline during $[t_1, t_2)$ in $\omega \in \xi$.

$$\mathcal{K}_{i,|t_1, t_2)}(\omega) \triangleq \sum_{t' \in [t_1, t_2)} \sum_{1 \leq k \leq i} \mathcal{KW}_k(\omega, t')$$

Taken together, we obtain the total processor use affecting $\tau_i$.

Def. 14. Let $\mathcal{E}_{i,t,\Delta}(\omega)$ be the \textit{processor demand} relevant to task $\tau_i$ in the time interval $[t, t+\Delta)$ in system evolution $\omega \in \xi$.

$$\mathcal{E}_{i,t,\Delta}(\omega) \triangleq \mathcal{Cl}_{i,t}(\omega) + W_{i,|t, t+\Delta)}(\omega) - \mathcal{K}_{i,|t, t+\Delta)}(\omega)$$

Def. 15. The \textit{ground-truth response time} $\mathcal{RT}_{i,j}(\omega)$ of $J_{i,j}^\xi$ in evolution $\omega \in \xi$, assuming $t = a_{i,j}^\xi$, is:

$$\mathcal{RT}_{i,j}(\omega) \triangleq \inf \{ \Delta \mid \Delta > 0 \land \mathcal{E}_{i,t,\Delta}(\omega) \leq \Delta \}.$$
Next we derive upper bounds on the workload accumulation functions \( \mathcal{T}_{W_i,\{t_1,t_2\}}(\omega) \) and \( \mathcal{W}_{i,\{t_1,t_2\}}(\omega) \). Given \( \omega \in \xi \), let \( \mathcal{T}_{W_i,\{t_1,t_2\}}(\omega) \) be the sum of the first \( \left[ \frac{t_2 - t_1}{T_i} \right] \) execution times of jobs of \( \tau_i \) arriving at or after \( t_1 \) in \( \omega \in \xi \):
\[
\mathcal{T}_{W_i,\{t_1,t_2\}}(\omega) \triangleq \sum_{k \leq l \leq k+\left[ \frac{t_2 - t_1}{T_i} \right]} C_{i,j}(\omega)
\]
where \( J^\xi_{i,k} \) is the first job arriving at or after \( t_1 \). The corresponding aggregate bound is
\[
\mathcal{W}_{i,\{t_1,t_2\}}(\omega) \triangleq \sum_{1 \leq k \leq i} \mathcal{T}_{W_{k,\{t_1,t_2\}}}(\omega).
\]

**Lemma 4.** For all \( t_1, t_2 \in \mathbb{N} \) such that \( t_1 < t_2 \) and all \( \omega \in \xi \):
\[
\mathcal{W}_{i,\{t_1,t_2\}}(\omega) \geq \mathcal{W}_{i,\{t_1,t_2\}}(\omega)
\]

**Proof.** From the assumption of sporadic tasks, it follows that a task \( \tau_i \) can release at most \( \left[ \frac{t_2 - t_1}{T_i} \right] \) jobs within a time interval \( [t_1,t_2) \), and hence \( \mathcal{T}_{W_{i,\{t_1,t_2\}}}(\omega) \geq \mathcal{T}_{W_{i,\{t_1,t_2\}}}(\omega) \) for each \( \tau_i \). Thus, \( \mathcal{W}_{i,\{t_1,t_2\}}(\omega) = \sum_{1 \leq k \leq i} \mathcal{T}_{W_{k,\{t_1,t_2\}}}(\omega) \) is the sum of the first \( \mathcal{T}_{W_{k,\{t_1,t_2\}}}(\omega) \) jobs. \( \square \)

Next, we define a simplified upper bound \( \bar{R}_{i,j}(\omega, \Delta) \) on the ground-truth response time for any \( \omega \in \xi \), assuming \( t = \alpha_{i,j}^\xi \).
\[
\bar{R}_{i,j}(\omega, \Delta) \triangleq \bar{C}_{i,t}(\omega) + \mathcal{W}_{i,\{t+t+\Delta\}}(\omega)
\]

(4)

Our next objective is to prove that \( \bar{R}_{i,j}(\omega, \Delta) \) implies a bound on the ground-truth DFP, which requires additional setup.

**Lemma 5.** For all \( \Delta \in (0, D_i] \) and \( \omega \in \xi \), if \( R_{i,j}(\omega) > D_i \), then \( \bar{R}_{i,j}(\omega, \Delta) > \Delta \).

**Proof.** By Def. 17, \( R_{i,j}(\omega) > D_i \) implies \( \Delta < \bar{C}_{i,t}(\omega) \) for all \( \Delta \in (0, D_i] \). From Lemmas 3 and 4 and since \( \forall t_1, t_2 \in \mathbb{N}, \mathcal{W}_{i,\{t_1,t_2\}}(\omega) \geq 0 \), we further have \( \Delta < \bar{C}_{i,t}(\omega) \leq \bar{C}_{i,t}(\omega) + \mathcal{W}_{i,\{t+t+\Delta\}}(\omega) \). Thus, by Eq. (4), \( \Delta < \bar{R}_{i,j}(\omega, \Delta) \). \( \square \)

As the next stepping stone, we relate the ground-truth DFP of any \( J^\xi_{i,j} \) to a simplified upper bound using \( \bar{R}_{i,j}(\omega, \Delta) \).

**Lemma 6.** For all \( \xi \in \Xi, \tau_i \in \tau, j \in \mathbb{N}, \) and \( \Delta \in (0, D_i] \):
\[
\mathbb{P}[R_{i,j} > D_i | \xi] \leq \mathbb{P}[\bar{R}_{i,j}(\Delta) > \Delta | \xi]
\]

**Proof.** From Def. 5, we have
\[
\forall \omega \in \xi, R_{i,j}(\omega) > D_i \Rightarrow \bar{R}_{i,j}(\omega, \Delta) > \Delta,
\]
which implies
\[
\{ \omega \in \xi | R_{i,j}(\omega) > D_i \} \subseteq \{ \omega \in \xi | \bar{R}_{i,j}(\omega, \Delta) > \Delta \},
\]
and
\[
\mathbb{P} \{ \omega \in \xi | R_{i,j}(\omega) > D_i \} \leq \mathbb{P} \{ \omega \in \xi | \bar{R}_{i,j}(\omega, \Delta) > \Delta \},
\]
and hence, by dividing both sides by \( \mathbb{P}[\xi] \), we obtain
\[
\mathbb{P}[R_{i,j} > D_i | \xi] \leq \mathbb{P}[\bar{R}_{i,j}(\Delta) > \Delta | \xi]
\]

Note that Lemma 6 holds for all \( \Delta \in (0, D_i] \), and thus in particular also for the \( \Delta \in (0, D_i] \) that minimizes the bound. Finally, we obtain an upper bound on \( WCDFP_i \).

**Def. 19.** Let \( WCDFP_i \) be defined as follows.
\[
WCDFP_i \triangleq \max_{\xi \in \Xi} \max_{j \in \mathbb{N}} \min_{\Delta \in (0, D_i]} \left\{ \mathbb{P} \left[ \bar{R}_{i,j}(\Delta) > \Delta | \xi \right] \right\}
\]

**Theorem 3.** \( \forall \tau_i \in \tau : WCDFP_i \leq WCDFP \)

**Proof.** Follows from Def. 18 and Lemma 6:
\[
WCDFP_i \triangleq \max_{\xi \in \Xi} \max_{j \in \mathbb{N}} \min_{\Delta \in (0, D_i]} \left\{ \mathbb{P} \left[ \bar{R}_{i,j}(\Delta) > \Delta | \xi \right] \right\}
\]
where (i) follows from Lemma 6. \( \square \)

**B. Applying the Correlation-Tolerant Concentration Inequality**

Theorem 2 lets us bound the probability of a sum of correlated random variables by only considering upper bounds on each random variable’s expected value and standard deviation. Theorem 3 shows that \( WCDFP_i \) is upper-bounded by
\[
\max_{\xi \in \Xi} \min_{j \in \mathbb{N}} \min_{\Delta \in (0, D_i]} \left\{ \mathbb{P} \left[ \bar{R}_{i,j}(\Delta) > \Delta | \xi \right] \right\},
\]
where \( \bar{R}_{i,j}(\Delta) \) is a finite sum of possibly dependent random variables. We now put these pieces together.

First, observe that, to safely bound \( \mathbb{P} \left[ \bar{R}_{i,j}(\Delta) > \Delta | \xi \right] \) for a particular \( \xi \in \Xi \), we require only the following bounds:
- \( \bar{c}_{i,t} \) — an upper bound on the expected execution time of any job of \( \tau_i \) in \( \xi \), i.e., \( \forall j \in \mathbb{N}, \bar{c}_{i,t} \geq \max_{j \in \mathbb{N}} E[C_{i,j} | \xi], \)
- \( \bar{c}_{i,t} \) — an upper bound on the standard deviation of \( \tau_i \)’s execution time in \( \xi \), i.e., \( \forall j \in \mathbb{N}, \bar{c}_{i,t} \geq \max_{j \in \mathbb{N}} \sigma[C_{i,j} | \xi], \)
where \( \sigma[C_{i,j} | \xi] \) is the standard deviation of \( C_{i,j} \) conditioned on \( \xi \), and \( \bar{c}_{i,t} \) denote vectors of per-task bounds, and define
\[
\alpha(i, x, \Delta) \triangleq x_i + \sum_{k=1}^{i-1} x_h \cdot \left( \frac{\Delta}{T_h} \right) + 1,
\]
where \( x \) is a vector such as \( \bar{c}_{i,t} \) or \( \bar{c}_{i,t} \).

Finally, we use the concentration inequality from Sec. IV.

**Lemma 7.** For all \( \tau_i \in \tau, \xi \in \Xi, \) and any \( \Delta \in (0, D_i] \), if \( 0 < \alpha(i, \bar{c}_{i,t}, \Delta) < \Delta \), then:
\[
\mathbb{P} \left[ \bar{R}_{i,j}(\Delta) > D_i | \xi \right] \leq \frac{\alpha(i, \bar{c}_{i,t}, \Delta)^2}{\alpha(i, \bar{c}_{i,t}, \Delta)^2 + (\Delta - \alpha(i, \bar{c}_{i,t}, \Delta))^2};
\]

**Proof.** Starting from Eq. (4), we obtain:
\[
\mathbb{P} \left[ \bar{R}_{i,j}(\Delta) > D_i | \xi \right] \leq \frac{\alpha(i, \bar{c}_{i,t}, \Delta)^2}{\alpha(i, \bar{c}_{i,t}, \Delta)^2 + (\Delta - \alpha(i, \bar{c}_{i,t}, \Delta))^2};
\]
where Inequality (i) follows from Corollary 2. \( \square \)
Lemma 7 is close to what we want, but still applies to individual arrival sequences. Next, we eliminate \( \xi \) from the inequality. To this end, suppose we are given the following bounds:

- \( \hat{c}_i \) — an upper bound on the expected execution time of any job of \( \tau_i \) in any \( \xi \in \Xi \), i.e., \( \hat{c}_i \geq \max_{\xi \in \Xi} \xi \), and
- \( \hat{s}_i \) — an upper bound on the standard deviation of the execution time of any job of \( \tau_i \) in any \( \xi \in \Xi \), i.e., \( \hat{s}_i \geq \max_{\xi \in \Xi} \xi \).

Again, we let \( \hat{c} \triangleq (\hat{c}_1, \ldots, \hat{c}_n) \) and \( \hat{s} \triangleq (\hat{s}_1, \ldots, \hat{s}_n) \) denote vectors of the just-defined per-task bounds.

**Lemma 8.** For all \( \tau_i \in \tau \), \( \xi \in \Xi \), and any \( \Delta \in (0, D_i) \), if \( 0 < \alpha(i, \hat{c}, \Delta) < \Delta \), then:

\[
\frac{\alpha(i, \hat{c}, \Delta)^2}{\alpha(i, \hat{s}, \Delta)^2 + (\Delta - \alpha(i, \hat{c}, \Delta))^2} \leq \frac{\alpha(i, \hat{c}, \Delta)^2}{\alpha(i, \hat{s}, \Delta)^2 + (\Delta - \alpha(i, \hat{c}, \Delta))^2}
\]

**Proof.** Note that \( \alpha(i, \hat{c}, \Delta) \leq \alpha(i, \hat{s}, \Delta) \) due to definition of \( \alpha(i, \cdot, \Delta) \) and since, by definition,

\[
\forall \tau_i \in \tau, \hat{s}_i \geq \max_{\xi \in \Xi} \xi \cdot \hat{s}_i.
\]

Similarly, \( \alpha(i, \hat{c}, \Delta) \leq \alpha(i, \hat{s}, \Delta) \) since

\[
\forall \tau_i \in \tau, \hat{c}_i \geq \max_{\xi \in \Xi} \xi \cdot \hat{c}_i.
\]

The claim then follows by Lemma 1. \( \square \)

**Theorem 4** (Correlation-Tolerant WCDFP Analysis). For all \( \tau_i \in \tau \) and any \( \Delta \in (0, D_i) \), if \( 0 < \alpha(i, \hat{c}, \Delta) < \Delta \), then:

\[
WCDFP_i \leq \frac{\alpha(i, \hat{c}, \Delta)^2}{\alpha(i, \hat{s}, \Delta)^2 + (\Delta - \alpha(i, \hat{c}, \Delta))^2}.
\]

**Proof.** Starting from Theorem 3, we have:

\[
WCDFP_i \leq \frac{\max_{\xi \in \Xi} \max_{j \in N} \min_{\tau \in \tau} \{ \mathbb{P}[\hat{R}_{i,j}(\Delta^*) > \Delta^* | \xi] \}}{\max_{\xi \in \Xi} \max_{j \in N} \{ \mathbb{P}[\hat{R}_{i,j}(\Delta) > \Delta | \xi] \}}
\]

Equality (i) is Def. 19. Inequality (ii) follows since \( \Delta \in (0, D_i) \) and from the definition of \( \min \). Inequality (iii) follows from Lemma 7 for \( \Delta > \alpha(i, \hat{c}, \Delta) \), which holds since we have \( \Delta > \alpha(i, \hat{c}, \Delta) \) as a premise and \( \forall \xi \in \Xi, \alpha(i, \hat{c}, \Delta) \geq \alpha(i, \hat{c}, \Delta) \). Similarly, Inequality (iv) holds by Lemma 8 for \( \Delta > \alpha(i, \hat{c}, \Delta) \), which is our premise. Finally, Equality (v) follows trivially since \( \xi \) and \( j \) no longer appear in the term being maximized. \( \square \)

Theorem 4 establishes the soundness of CTA, but it is not obvious from Theorem 4 that CTA offers any improvements over existing pWCET-based methods. To explore this aspect, we conducted an empirical evaluation comparing CTA with IAA.

**VII. Evaluation**

We report on experiments comparing our proposed method, referred to as CTA in the following, with two IAA baselines:

- **Berry-Esseen** — a lower bound on WCDFP derived by Marković et al. [40] from the Berry-Esseen theorem; and
- **Chernoff** — an upper bound on WCDFP proposed by Chen et al. [10] based on the Chernoff bound.

Recall from Sec. II that the baselines and CTA approach the analysis of each task \( \tau_i \) quite differently. The IAA baselines employ \( pWCET_i \), a distribution that over-approximates all conceivable scenarios of operation of \( \tau_i \) [8, 19]. In contrast, CTA relies only on \( \hat{c}_i \) and \( \hat{s}_i \), as established in Sec. VI-B.

The real-time systems literature presently offers no guidance on how, in real workloads, the summary statistics used by CTA relate to obtainable pWCET distributions. We, therefore, chose to investigate a broad spectrum of possible relationships in our study to give an account of how CTA might perform for many possible workload types. For this purpose, we designed a base setup that we refined into four experiments. Each of these experiments perturbs one of the key parameters used to generate synthetic workloads, as discussed next.

**A. Experimental Setup**

For each combination of the analyzed workload-generation parameters \( n, U^{wc}, U^{avg} \), and \( r^{max} \), all defined in the following, we randomly generated 500 sporadic task sets.

Each task set was generated as follows. Given the desired task set size \( n \), we randomly selected \( n \) periods \( T_1, T_2, \ldots, T_n \) from the set \{1, 2, 5, 10, 20, 50, 100, 200, 500, 1000\} (all in milliseconds), which are commonly found in automotive systems [29]. We assigned all tasks rate-monotonic priorities. Next, given a target utilization \( U^{wc} \), we used the Dirichlet-Rescale algorithm [24] to pick \( n \) random utilization values \( u_1^{wc}, u_2^{wc}, \ldots, u_n^{wc} \) summing to \( U^{wc} \). Mimicking the experimental setup of Bozhko et al. [7], we considered \( u_i^{wc} \) to be the expected utilization according to \( pWCET_i \). That is, \( \xi^{wc} = \mathbb{E}[pWCET_i]/T_i \) and \( U^{wc} = \sum_{i=1}^{n} \mathbb{E}[pWCET_i]/T_i \), where \( pWCET_i \) denotes \( \tau_i \)'s \( pWCET \) distribution.

Each task \( \tau_i \)'s \( pWCET \) was generated following a normal distribution with mean \( \mathbb{E}[pWCET_i] = u_i^{wc} \cdot T_i \), and a standard deviation selected uniformly at random from the interval \([0.01 \cdot \mathbb{E}[pWCET_i], 0.25 \cdot \mathbb{E}[pWCET_i]]\). To provide a rationale for the chosen maximum standard deviation, in a normal distribution with mean \( \mu = 100 \) and standard deviation \( \sigma = 0.25 \cdot 100,95\% \) of the samples fall within \([50,150]\). After generating the standard deviation and expected value, we discretized \( pWCET_i \) into four to eight discrete values, while preserving the targeted mean and standard deviation, to ensure that the Chernoff method can be computed efficiently.

As CTA relies on upper bounds on the true expected value and standard deviation of each task’s execution-time distribution (in any arrival sequence), we also defined an average utilization \( U^{avg} = \sum_{i=1}^{n} \frac{u_i^{avg}}{T_i} = \sum_{i=1}^{n} \frac{\hat{c}_i}{T_i} \), where \( \hat{c}_i \) is the upper
bound on the expected execution time of $\tau_i$ as specified in Sec. VI-B. Given a target $U^{\text{avg}}$, naturally it must hold that $U^{\text{avg}} \leq U^{\text{wc}}$ overall, and for each task $\tau_i$, $u_i^{\text{avg}} \leq u_i^{\text{wc}}$. Hence, we randomly generated $n$ individual $u_i^{\text{avg}}$ values summing to $\sum_i u_i^{\text{avg}} = U^{\text{avg}}$, once more using the Dirichlet-Rescale algorithm while setting the respective $u_i^{\text{wc}}$ as an upper bound for each $u_i^{\text{avg}}$. For each task, $\hat{e}_i$ was simply set to $e_i = u_i^{\text{avg}} \cdot T_i$.

The upper bound $\hat{s}_i$ on the ground-truth standard deviation was chosen uniformly at random from the interval $[0.01 \cdot \hat{e}_i, r_i^{\text{max}} \cdot \hat{e}_i]$, where $r_i^{\text{max}}$ denotes a configurable maximum ratio between the generated standard deviation and mean.

**Base setup.** The base configuration used as a starting point for all experiments consisted of $n = 25$ tasks per task set with $U^{\text{wc}} = 0.9$ and $U^{\text{avg}} = 0.2$. The gap between $U^{\text{wc}}$ and $U^{\text{avg}}$ matches the intuition sketched in Sec. If that pWCET distributions tend to significantly overestimate the average execution behavior of dependent tasks. The maximum ratio of the ground-truth standard deviation was $r_i^{\text{max}} = 0.25$, analogously to the generated pWCET distributions.

**B. Interpretation of Results**

The point of comparison is the WCDFP bound reported by each of the three methods for the lowest-priority task (i.e., the one subject to maximal interference). In the following discussion, and in particular in Figs. 2–5, three relationships among the bounds obtained with the CTA, Berry-Esseen, and Chernoff methods are of particular interest.

- **CTA < Berry-Esseen:** This condition indicates that CTA yields a better (i.e., lower) bound than any possible IAA method since it attains a WCDFP bound below the lower bound on WCDFP provided by Berry-Esseen.
- **Berry-Esseen < CTA < Chernoff:** In this case, CTA demonstrates the potential to deliver better results than some IAA methods—in particular, CTA attained a lower bound than the Chernoff method—but there might exist IAA methods more accurate than CTA since the bound reported by CTA exceeds the Berry-Esseen lower bound.
- **Chernoff < CTA:** This outcome indicates the case where CTA does not offer any advantages over the state of the art, since it provides a more conservative (pessimistic) WCDFP bound than the Chernoff method.

In addition to these summary categories, we also report scatter plots directly relating CTA and Chernoff, exhibiting the involved numerical magnitudes, as explained in more detail shortly. We next report on the results of the four experiments, focusing on high-level trends and major factors that affect CTA.

**C. Experiment 1: Influence of the Task Set Size**

In the first experiment, we varied the number of tasks $n$ from 5 to 50, in increments of 5. The results are shown in Fig. 2. Consider the plot at the top first. As the number of tasks in the task set increases, the relative advantage of the CTA method over the IAA baselines improves, going from 50% certainly better results ($n = 5$, case CTA < Berry-Esseen) to more than 80% certainly better results ($n = 50$).

The observed trend is explained by the linearity of expectation (Fact 1). Since the expected value of each task’s pWCET distribution exceeds that of its ground-truth distribution, IAA methods accumulate pessimism with each added higher-priority job that has to be considered. In contrast, the total average utilization is not impacted by $n$. This highlights a significant structural advantage of CTA, which benefits from the fact that expected values are unaffected by dependence, a factor that IAA methods are not equipped to consider.

Next, consider the scatter plot at the bottom of Fig. 2, which shows WCDFP estimates provided by the CTA and Chernoff methods. The results for all task set sizes ($n$) are combined and distinguished according to the color scale indicated in the legend. Points above the diagonal line represent instances where CTA provided a better WCDFP estimate than the Chernoff method. Conversely, points below the diagonal indicate cases where Chernoff is preferable.

While we observe that CTA yields better WCDFP estimates in a significant number of analyzed task sets, it is noteworthy...
that, when Chernoff outperforms CTA, its estimates seem to exhibit more breadth, covering a range of probabilities from $10^{-3}$ to $10^{-10}$, unlike CTA’s estimates, which fall within the narrower $10^{-4}$ to $10^{-3}$ range. This discrepancy arises because CTA, a simple closed-form bound, relies solely on $\hat{e}_i$ and $\hat{s}_i$, parameters that remain within a static range of values in this experiment. In contrast, Chernoff taps into the shape of pWCET distributions using a more intricate optimization process [10]. This result implies a potential area for future exploration: enhancing CTA by integrating more details about the underlying ground-truth distribution could be promising.

D. Experiment 2: Varying Total Expected pWCET Utilization

The second experiment varied the expected pWCET utilization, $U^{wc}$, from 0.8 to 1 in increments of 0.02. The results are shown in Fig. 3. The plot at the top of Fig. 3 reveals that the relative merit of CTA compared to the IAA baselines improves with the increase in $U^{wc}$. This trend is expected because, as the pessimism in the pWCET distributions progressively increases relative to the unchanging $U^{avg}$ from the base setup, CTA’s benefit becomes more pronounced. Conversely, when $U^{wc}$ is reduced, the analysis problem becomes simpler for IAA methods. As a result, for the lowest evaluated $U^{wc}$ value ($U^{wc} = 0.8$), Chernoff offers better WCDFP bounds for more than 50% of the tested task sets. Overall, Fig. 3 shows a clear trend: CTA is less attractive in settings where IAA methods are not challenged, and clearly preferable to any IAA method when pWCET distributions are subject to significant pessimism.

In the bottom plot of Fig. 3, we notice that, once again, due to the invariant $\tilde{e}_i$ and $\tilde{s}_i$ parameters, all WCDFP bounds provided by CTA fall within $[10^{-4}, 10^{-2}]$, while the Chernoff estimates converge to 1 as $U^{wc}$ increases.

E. Experiment 3: Influence of the Average Utilization

In the third experiment, we varied the average total utilization $U^{avg}$ from 0.05 to 0.9 in increments of 0.05. The results are shown in Fig. 4. In the plot at the top of the figure, the CTA method demonstrates a relatively consistent rate of success in identifying superior WCDFP estimates up to a utilization level of $\approx 0.65$. Beyond this point, however, the performance of CTA sharply declines until it is unable to identify a single better estimate at a utilization level of $\approx 0.85$.

This trend is not unexpected. As the average utilization increases, each task’s pWCET becomes more representative of the ground-truth distribution, implying greater task independence. Consequently, the CTA method, which tolerates correlation but
uses a comparably coarse concentration inequality, becomes increasingly pessimistic for the underlying system.

Interestingly, the fraction of workloads for which the Chernoff method outperforms CTA (i.e., the width of the “red band”) does not change substantially across the entire range. Instead, a new category emerges, namely workloads for which both Chernoff and CTA report a WCDFP of 1 (black), i.e., difficult workloads that defied effective analysis.

The bottom plot of Figure 4 provides further insight into the significant impact of average utilization on the WCDFP estimates provided by CTA. Lower $U_{avg}$ values entail smaller $\hat{e}_3$ values, favoring the concentration inequality at the heart of CTA. As $U_{avg}$ increases, $\hat{e}_3$ values also rise, resulting in estimated probabilities ranging from $10^{-5}$ to 1.

### F. Experiment 4: Influence of the Maximum Standard Deviation

In the fourth experiment, we varied $r_{max}$ from 0.01 to 0.25 in steps of 0.01. To put this into perspective, consider a normal distribution with mean $\mu = 100$. For $\sigma = 0.1 \cdot 100$, 95% of the samples will fall within $[80, 120]$. For $\sigma = 0.01 \cdot 100$, it is $[98, 102]$, and as mentioned, for $\sigma = 0.25 \cdot 100$, it is $[50, 150]$.

In the top plot of Fig. 5, we observe that the number of superior WCDFP estimates produced by CTA remains consistent across the entire considered range. For each point, CTA yields a better WCDFP estimate than any possible IAA method for approximately 350 out of 500 analyzed task sets.

However, it would be wrong to conclude that $r_{max}$ has no impact on CTA. In the bottom plot, CTA’s estimated probabilities span from $10^{-5}$ to $10^{-3}$, increasing with the rising maximum standard deviation. This trend intuitively follows from the foundation of CTA, Cantelli’s inequality. Looking at Theorem 2, we observe that, by maintaining a constant expected value (as is the case with our baseline $U_{avg}$) and increasing variation (as performed in this experiment), the contribution of the term involving the expected execution cost diminishes. As a result, the bound tends towards 1 as $r_{max}$ increases.

These observations underscore that, unsurprisingly, the accuracy of the WCDFP bounds generated by CTA is quite sensitive to both $U_{avg}$ and $r_{max}$. Overall, our evaluation shows CTA to be complementary to existing IAA methods: in many cases, CTA can provide bounds better than any possible IAA method (i.e., when pWCETs are inherently pessimistic), but in settings inherently favoring IAA (i.e., pWCETs without much structural pessimism), CTA offers only limited improvements.

### VIII. Related Work

Comprehensive surveys by Davis and Cucu-Grosjean [17, 19] offer an extensive assessment of probabilistic schedulability and timing techniques, often highlighting two issues that are central to this paper: accounting for dependent tasks in probabilistic analysis, and effectively estimating the WCDFP.

Ivers and Ernst [27] investigated the problem of unknown dependencies among the execution times of jobs in a given, fixed arrival sequence, exploring the use of copulas, originally used in timing analysis by Bernat et al. [6]. Ivers and Ernst presented a solution for systems under fixed-priority preemptive scheduling, assuming availability of the entire probability distribution for each task. Their method uses copulas and Frechet bounds to model relationships among distributions, deriving probabilistic response-time bounds. In comparison, CTA operates under a different premise, assuming sporadic tasks, and uses only bounds on each task’s expected execution time and standard deviation in any arrival sequence (rather than full distributions, as in Ivers and Ernst’s approach [27]).

Markov models have also been employed to handle execution time dependencies in real-time systems. Frías et al. [20] and Abeni et al. [1] used hidden Markov models (HMMs) for periodic tasks with dependent execution times provisioned in constant bandwidth servers. Further, Friebe et al. [21–23] proposed the application of continuous Gaussian emission distributions in HMMs, and suggested an approach to bound the deadline-miss probability in a reservation-based system where each task is confined to a private reservation. The accuracy of Markov models depends heavily on the data used for model identification, and as observed by Friebe et al. [22, 23], such distributions are likely changing over time. We also note that the deadline-miss probability estimated in this line of work considers a long-frequency interpretation [19], which differs from the WCDFP metric addressed in this paper. Moreover,
while Markov models can capture intra-task dependencies well, inter-task dependencies remain a challenging problem.

The dependence problem has also been considered in the context of EVT and its application in measurement-based statistical analysis of execution times [15, 32, 33] and response times [35–37]. While EVT-based analyses have been extensively used in research and practice, they nonetheless are subject to some noteworthy limitations. EVT works under the assumption that the statistical limit laws hold for a given set of samples [12, Ch. 5, pp. 92–93]. The sample size required to obtain a good agreement between the empirical and theoretical distributions heavily depends on the degree of correlation, i.e., highly correlated sequences require a much larger dataset than weakly correlated ones. To analyze dependent tasks, assumptions of stationarity [30] or extremal independence [48] of the analyzed distributions must be met.

Several approaches tackled the dependency problem by introducing specific structural assumptions on how execution times may relate. Mills and Anderson [44] proposed a scheduling policy that allows for stochastic execution-time demands with arbitrary degrees of dependence limited to pre-specified time intervals of bounded length. von der Brüggen et al. [51] proposed an approach for approximating the WCDFP under EDF that allows for dependencies in a bounded number of subsequent jobs. Liu et al. [34] proposed a stochastic response-time analysis that introduces the concept of an independence threshold, i.e., a per-task threshold that splits each job’s execution cost into a dependent and a (presumed) independent part. In this paper, we do not impose any limitations or constraints on the nature or magnitudes of dependencies.

Except for work by von der Brüggen et al. [51], the just-cited approaches do not address WCDFP estimation, which has been primarily examined in the context of fixed-priority scheduling using pWCET-based IAA methods. There has been much progress in this direction in recent years: von der Brüggen et al. [50] adapted the Hoeffding and Bernstein inequalities for WCDFP estimation; Chen et al. [10] derived an IAA method from the Chernoff bound, which we compared CTA to in Sec. VII; Marković et al. [39] developed an optimal resampling and an efficient circular-convolution algorithm for IAA methods; and Bozhko et al. [7] proposed an approach rooted in Monte-Carlo sampling. Most recently, Chen et al. [11] rectified a mistaken critical-instant assumption found in several IAA methods, and Marković et al. [40] applied the Berry-Esseen theorem to estimate the range of a task’s response-time distribution, which we adopted as a baseline in Sec. VII. Complementing the studies cited so far, which for the most part consider fully-preemptive fixed-priority uniprocessor scheduling, there also exist a number of additional IAA methods applicable to other workload models [e.g., 26, 38, 43, 52].

CTA departs from the IAA tradition: instead of relying on pWCET distributions to mask any correlation with pessimism [8], CTA works directly with bounds on simple summary statistics of the ground-truth behavior such that arbitrary, unknown correlation is tolerated.

IX. Conclusion

We have proposed a new method, CTA, for safely estimating the WCDFP of sporadic real-time tasks under preemptive fixed-priority scheduling. CTA offers two major innovations: first, it is robust in the presence of arbitrary, unknown dependencies among execution times, and second, it does not rely on pWCET as a building block. Instead, CTA requires only bounds on the mean and standard deviation of each task’s ground-truth execution-time distribution (in any arrival sequence). Mathematically, CTA is a consequence of Cantelli’s Inequality, which previously had not been applied to probabilistic real-time systems. We have verified with Coq that the concentration inequality at the heart of CTA, Corollary 2, does indeed hold in the presence of dependent random variables, as claimed.

Empirically, our evaluation has shown that CTA effectively reduces analysis pessimism when pWCET distributions over-estimate the expected ground-truth execution time, which is generally impossible to avoid due to the guarantees that pWCET must provide for IAA [8]. Conversely, CTA becomes comparatively less effective as the difference between ground-truth distributions and pWCET diminishes. Overall, CTA complements existing methods by providing significantly improved bounds for many, but not all, tested workloads.

More generally, the CTA idea is broadly applicable beyond our setting since Corollary 2 is policy-agnostic and could be readily adapted to, for instance, global multiprocessor scheduling. Partitioned multicore scheduling, in particular, does not cause any conceptual issues: CTA applies as-is on each core. Cross-core interference (e.g., via shared caches or memory buses) will manifest in the ground-truth execution-time distributions of the workloads on each core. Our analysis remains sound, provided the bounds on the means and standard deviations correctly reflect the effects of cross-core interference. It is worth noting that CTA is particularly well suited for future extensions targeting locking-related delays, which are inherently not independently distributed (and thus challenging for IAA).

CTA opens the door to many interesting possibilities for future work. In practical terms, it is reasonable to expect that high confidence bounds on means and standard deviations will be much easier to obtain (and with substantially fewer samples) than full pWCET distributions. It will be interesting to validate CTA in conjunction with measurement-based approaches on a real hardware platform. Analytically, it is striking that CTA, by considering bounds on only two simple summary statistics (mean and standard deviation) and making no further assumptions about the ground-truth distributions, already manages to provide insights complementary to state-of-the-art pWCET-based methods. It stands to reason that substantially more advanced tools from probability theory can be brought to bear in a similar way, with the promise of even better bounds where the initial CTA, as developed in this paper, is still less effective (e.g., high average utilization). In general, methods that account for task dependencies warrant much further attention.
This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 803111). In addition, Alessandro Papadopoulos was supported by the Swedish Research Council (VR) with the PSI project (No. #2020-05094).

ACKNOWLEDGMENTS

REFERENCES


approximations in probabilistic analysis of real-time systems,” in 43rd IEEE Real-Time Systems Symposium (RTSS’22), December 5-8, Houston, TX, USA, 2022, pp. 158–171.


